

UC Irvine

Presentations

Title

Bibliographic Records as Data: Making research use of our shared collections

Permalink

<https://escholarship.org/uc/item/0xb8s204>

Authors

Kane, Danielle

Dickerson, Madelynn

Publication Date

2021-10-27

DOI

10.48448/3213-nc21

Supplemental Material

<https://escholarship.org/uc/item/0xb8s204#supplemental>

Data Availability

The data associated with this publication are in the supplemental files.

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at

<https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Bibliographic Records as Data: Making research use of our shared collections

UC Libraries Forum
October 27th-29th, 2021

Madelynn Dickerson

Head, Digital Scholarship Services, UCI Libraries

Danielle Kane

Computational Research Librarian, UCI Libraries



About the Project

The project started as a/an:

- Creative exploration!
- Demonstration to ourselves and others the ways that library bibliographic data can be useful in performing scholarly analysis of library collections
- Opportunity for collaboration between Digital Scholarship Services and Cataloging and Metadata Services departments



Collaborators

Madelynn Dickerson - Research Librarian for Digital Humanities and History

Joshua Hutchinson - Cataloging and Metadata Librarian

Danielle Kane - Computational Research Librarian

Sarah Wallbank - Electronic Resources and Serials Cataloging Librarian



Project Goals

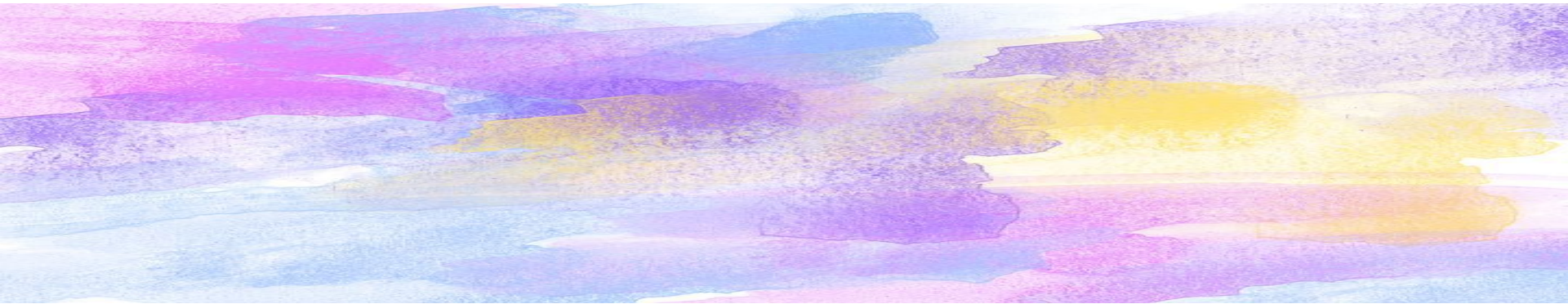
- Identify areas where bibliographic data might effectively answer scholarly research questions
- Demonstrate what skills are necessary to make use of this data
- Learn those skills and tools
- Perform a sample analysis on the UCI Libraries history monograph collection



Sample Research Questions

We imagined a digital humanities researcher could have questions like these:

- Of all the history monographs in our catalog (books with the call numbers C-F), how many were written by women?
- What topics are women historians writing about?



Research Question Implications

Heading into the sample research question, we wanted to explore the following:

- What processes and tools would be required to make this analysis?
- What are the challenges and pitfalls?
- Is it even possible to accurately and ethically identify an author as a woman based on their name alone? How would (and should?) one go about doing this for the purpose of scholarly analysis?
- What other data is available in the bibliographic record that could be of scholarly interest (or of value for collection analysis?)



Tools Used

- [C# MARC Editor](#): open source editor for Library of Congress MARC21 and MARCXML bibliography records.
- [OpenRefine](#): an open-source desktop application for data cleanup and transformation to other formats.
 - [GREL functions](#): (General Refine Expression Language) is designed to resemble Javascript. Formulas use variables and depend on data types to do things like string manipulation or mathematical calculations
- [Voyant Tools](#): a web-based reading and analysis environment for digital texts.



Methods

1. Exported MARC records from Alma
2. Popped raw data into Voyant Tools to see what came out
3. Strategized how to clean and organize the data
4. Prioritized fields we were interested in
5. Divided data by decade, focusing on 1970s, 1980s, 1990s, 2000s
6. Cleaned and parsed data
7. Assembled a preliminary list of baby names by gender



Data Collected

History monographs (call numbers C-F)

- 184,103 items
- File size: 292 MB
- Author and title
- OCLC number (marc_35)
- Publisher, year, place (marc_260a-c)
- Any fields with subject terms



A1	RecordID																			
	A	B	C	D	E	F	Q	R	S	V	Z	AA	AB	AC	AD	AE	AF	AG	A	
1	RecordID	DateAdded	DateChanged	Author	Title	Copyright	TagNumbe	Ind1	Ind2	LDR	7	8	9	10	19	15	21	22		
2	1	1/17/2019 14:55	8/14/2018 17:49	McFaul, Micha	Russia's unfinished	2001	913			01045pam	a2200301	a4010316s2001	nyu b 001 0 en	\$a2001001667		\$aGBA1-W1444				
3	2	1/17/2019 14:55	8/14/2018 17:48	Li, Liejun,	Li Liejun ji /	1996	987			01642cam	a22004331a4	970402s1996	cc a b 000 0 chi d							
4	3	1/17/2019 14:55	8/14/2018 17:48	Bayard de Volc	Mothers of hero	2001	913			01484pam	a2200337 a4	010212s2001	mdua b 001 0 e	\$a2001000239		\$aGBA1-V4272				
5	4	1/17/2019 14:55	8/14/2018 17:48	Li, Liangyu,	Li Liangyu shi xue	2010	946	1		01591cam	a22004091a4	100118s2010	cc b 000 0 chi d							
6	5	1/17/2019 14:55	8/14/2018 17:48	Shen, Zhihong,	Zhongguo di fanq	2002	987			01397nam	a22003851a4	021211s2002	cc bi 001 0 chi d							
7	6	1/17/2019 14:55	8/14/2018 17:48	Dombrowski, t	Against culture :	2001	913			01344pam	a22003854a	010222s2001	nbuab b s001 0 e	\$a2001027327						
8	7	1/17/2019 14:55	8/14/2018 17:49	Petra Ernst / G	Jewish Spaces : c	2010	946	1		01377cam	a22003611a4	101220s2010	au a 000 0 ger d							
9	8	1/17/2019 14:55	8/14/2018 17:49	Carile, Paolo,	Huguenots sans	2001	913			00864nam	a22002651i4	020211s2001	fr 000 0 fre d							
10	9	1/17/2019 14:55	8/14/2018 17:48	Qiu, Wei,	Wuxing Qian jia :	2009	946	1		02460cam	a22005531a4	090528s2009	cc a b 000 0 chi d							
11	10	1/17/2019 14:55	8/14/2018 17:48	Mao, Ce,	Xiao yi chuan jia	2009	946	1		02157cam	a22005171a4	090528s2009	cc a b 000 0 chi d							
12	11	1/17/2019 14:55	8/14/2018 17:49	Macdougall, N	An antidote to th	2001	913			01138nam	a2200325 a4	010730s2001	stkab b 001 0 er	\$a2001411268		\$aGBA1-Z8040				
13	12	1/17/2019 14:55	8/14/2018 17:49	Guston, Philip,	Philip Guston's P	2001	913			01118cam	a2200325 a4	010206s2001	ilua b 000 0 eng	\$a2001000741		\$aGBA1-V3618				
14	13	1/17/2019 14:55	8/14/2018 17:48	Jiang, Ruoshi,	Qin Han qian bi y	1997	987			01551cam	a2200421 a4	980211s1997	cc a b 000 0 chi	\$a98450298						
15	14	1/17/2019 14:55	8/14/2018 17:48		Assessors' handk	1999	946	1		01213cam	a2200373 a4	990512s1999	cau s000 0 eng	\$oRLINCCSD99-B235						
16	15	1/17/2019 14:55	8/14/2018 17:48	Denham, Andri	Keith Joseph /	2001	913			01032cam	a2200313 a4	010320s2001	enkaf b 001 0 be	\$a2001411781		\$aGBA1-Y4507				
17	16	1/17/2019 14:55	8/14/2018 17:48	Coppieters, Bri	Federalism and c	2001	913			01244nam	a22003371a4	010918s2001	enk b 000 0 eng d							
18	17	1/17/2019 14:55	8/14/2018 17:49	Persico, Josepl	Roosevelt's secr	2001	913			01096cam	a22003134a4	010129s2001	nyuabf b 001 0 e	\$a2001019106						
19	18	1/17/2019 14:55	8/14/2018 17:48	Schwarz, Maur	Navajo lifeways	2001	913			01251pam	a22003014a	001101s2001	okua b s001 0 e	\$a00053276						
20	19	1/17/2019 14:55	8/14/2018 17:49	Wills, Brian Ste	The war hits hon	2001	913			01386cam	a2200349 a4	010313s2001	vauab b 001 0 e	\$a2001026035						
21	20	1/17/2019 14:55	8/14/2018 17:49	Hao Bin, Ouyai	Wu si yun dong y	2001	987			01924cam	a22004211a4	010906s2001	cc b 000 0 chi d							
22	21	1/17/2019 14:55	8/14/2018 17:49	Fahey, John,	Saving the reser	2001	913			01262cam	a22003494a4	010411s2001	waua b s001 0 b	\$a2001035249						
23	22	1/17/2019 14:55	8/14/2018 17:49	Sandes, Carolii	Archaeology, coi	2010	946	1		01673cam	a22004451a4	101027s2010	enkab b 000 0 eng d			\$aGBB0C9779S2bnb				
24	23	1/17/2019 14:55	8/14/2018 17:48	Zhang Zhongli :	Jin dai Shanghai	2008	946	1		02020cam	a2200421 a4	080628s2008	cc abf b 000 0 cf	\$a2009429712						
25	24	1/17/2019 14:55	8/14/2018 17:49	edited by Jaco	Recentring Asia :	2011	946	1		03365cam	a22003858a4	110428s2011	ne b 001 0 eng	\$a2011018365						
26	25	1/17/2019 14:55	8/14/2018 17:48	Chi, James San	Remembering th	2001	946	1		01022ntm	a22002891a4	101207s2001	xx rbm 000 0 eng d							
27	26	1/17/2019 14:55	8/14/2018 17:48	[edited by] Pau	White privilege :	2002	913			01020cam	a2200301 a4	001219s2002	nyu b 001 0 en	\$a00054644						
28	27	1/17/2019 14:55	8/14/2018 17:49	edited by Pauli	Patronage, cultu	2002	913			01939nam	a2200469 a4	010905s2002	ctuab b 001 0 ce	\$a2001095608		\$aGBA2-Z9901				
29	28	1/17/2019 14:55	8/14/2018 17:49	Otis, Eliza A.,	Architects of our	2001	913			01510cam	a2200361 a4	000313s2001	cauach b 000 0 c	\$a00035048						
30	29	1/17/2019 14:55	8/14/2018 17:49	Guo wu yuan S	Xia Jiang di qu ka	2009	946	1		02616cam	a22004811a4	090528s2009	cc ab b 000 0 chi d							
31	30	1/17/2019 14:55	8/14/2018 17:49	Boyle, Sheila T	Paul Robeson : ti	2001	913			01191cam	a2200337 a4	010122s2001	maua b s001 0 b	\$a2001017155		\$aGBA1-70574				
32	31	1/17/2019 14:55	8/14/2018 17:48	Dowek, Ephrai	Israeli-Egyptian r	2001	913			01130cam	a22003374a4	010316s2001	enk 001 0 eng	\$a2001028472		\$aGBA1-X4007				
33	32	1/17/2019 14:55	8/14/2018 17:49	edited by Amit	The post-colonie	2001	913			01008cam	a22003134a4	000504s2001	nyu b 001 0 en	\$a00034493						
34	33	1/17/2019 14:55	8/14/2018 17:49	Fagan, Brian M	Egypt of the pha	2001	913			01019pam	a2200325 a4	010201s2001	dcuab b 001 0 e	\$a2001030107		\$aGBA1-V9144				

Casual Findings

	Summary	Documents	Phrases			
	Term	Count ↓	Length	Trend		
<input checked="" type="checkbox"/>	world war	1743	2		^	
<input type="checkbox"/>	world the	433	2			
<input type="checkbox"/>	world of	274	2			
<input type="checkbox"/>	world history	157	2			
<input type="checkbox"/>	world a	142	2			
<input type="checkbox"/>	world order	121	2			
<input type="checkbox"/>	world politics	121	2			
<input type="checkbox"/>	world in	106	2			
<input type="checkbox"/>	world and	84	2			

OpenRefine (cleaning data)

- Isolating publication dates
- Removing a's and b's from start of lines
- Removing terminal punctuation marks and extra spaces
- Separating out information from cells - such as language and country of publication from the 008 field
- Splitting names - sending first and last names into separate cells
- Removing diacritics
- Using GREL to extract information
- Pulling and matching data from other datasets - combining datasets



OpenRefine BRRP_data_1980to1989 1 csv csv [Permalink](#)

Open... Export Help

Facet / Filter Undo / Redo 141 / 141

Refresh Reset All Remove All

country of publication change

160 choices Sort by: name count Cluster

- Afghanistan 1
- Alabama 50
- Alaska 14
- Albania 3
- Alberta 37
- Argentina 67
- Arizona 140
- Arkansas 16
- Armenia (Republic) 26
- Australia 200
- Austria 48
- Banladesh 2

25090 rows Extensions: Wikidata


Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 50 next > last »

	All	RecordID	Author	Name (first)	gender	Title	Codes (marc_008)
☆	8.	195	Hu Hua deng zhu			Zhou Enlai de si xiang ji li lun gong xian	830421s1982 cc t 000 0 chir
☆	9.	214	Guarducci, Margherita	Margherita	F	La cosiddetta Fibula Prenestina : elementi nuovi	860313s1984 it af b 001 0 ita d
☆	10.	219	Moscato, Sabatino	Sabatino	M	I gioielli di Tharros : origini, caratteri, confronti	880729s1988 it af b 000 0 ita d
☆	11.	229	[XII Settimana di studi aquileiesi, 30 aprile-5 maggio 1981			Aquileia nel IV secolo	830218s1982 it af b 100 0 ita
☆	12.	238	Eiseman, Cynthia Jones	Cynthia	F	The Porticello shipwreck : a Mediterranean merchant vessel of 415-385 B.C.	860722s1987 txuabe 001 0 eng
☆	13.	239	Baldi, Agnello	Agnello	NIL	L'anatema e la croce : Ebrei e Cristiani in Pompei antica	840904s1983 it a b 000 0 ita d
☆	14.	240	Guarducci, Margherita	Margherita	F	La cosiddetta Fibula Prenestina : antiquari, eruditi e falsari nella Roma dell'Ottocento	820405s1980 it af b 000 1 ita

Building a Name / Gender List

	List #1 (1970's)	List #2 (1970's)	List #3 (1970's)	List #3 (1980's)
Names (#)	3,196	4,818	7,502	7,502
Female (F)	1,342	1,639	2,794	2,290
Male (M)	11,939	13,379	17,997	12,022
Unisex (U)	81	346	676	583
Initial (I)	585	585	585	1,244
Blank	12,674	10,672	4397*	6497*
Not in List (NIL)	--	--	--	2,420
Title	--	--	--	1
Unidentified	--	--	172	33

Practical Takeaways

- Lots of bibliographic fields contain interesting information
 - Bibliographic data is nicely structured but requires significant cleaning and massaging
 - Collaborative data cleaning is hard!
 - Not enough data in our dataset to make any broad conclusions about our sample research question
 - No clear path forward on determining author gender unless authors self-identify
- 

Future Potential

- Thinking in a “collections as data” context, what would it mean to offer “library catalogs as data”?
 - Ex: Harvard Library API’s and Datasets
<https://library.harvard.edu/services-tools/harvard-library-apis-datasets>
- Potential to bring together more comprehensive data through UC’s combined catalog
- Opportunities to incorporate more nuanced search and analysis tools into our discovery systems, especially incorporating text analysis and data mining functionality



Examples of Bibliographic Data Analysis Platforms

- Harvard Library API's and Datasets
<https://library.harvard.edu/services-tools/harvard-library-apis-datasets>
- HathiTrust Research Center Analytics
<https://analytics.hathitrust.org/>
- JSTOR's Constellate Platform
<https://constellate.org/>
- MIT Libraries' List of Scholarly Publishing APIs
<https://libraries.mit.edu/scholarly/publishing/apis-for-scholarly-resources/>



Resources

- [Summary of transformations to decades](#)
- [First name and gender list](#)
- [LC countries list](#)

Publications

- Hutchinson, J., Wallbank, S., Dickerson, M., & Kane, D. (2019). Exploring Bibliographic Records as Research Data. *C&I*, 197. Retrieved from https://cdn.ymaws.com/www.cilip.org.uk/resource/collection/F71F19C3-49CF-462D-8165-B07967EE07F0/C&I_197.pdf
- Hutchinson, J. (2020). Collecting, Cleaning and Using Bibliographic Data to Perform a Large-Scale Assessment Project on University Library Collections. *UC Irvine: Libraries*. Retrieved from <https://escholarship.org/uc/item/7bb0g5np>



Contact

Madelynn Dickerson
mrosed@uci.edu

Danielle Kane
kaned@uci.edu

