## SPNHC 2022 DemoCamp: Using OpenRefine for natural history collections data

### Erica Krimmel<sup>1,2</sup> & Lindsay Walker<sup>1,3</sup>

<sup>1</sup> iDigBio (Integrated Digitized Biocollections)
 <sup>2</sup> Florida State University, Tallahassee, FL
 <sup>3</sup> Arizona State University, Tempe, AZ

See here for more details and links to resources



#### Why use OpenRefine?

- **Open-source** tool for manipulating small or large datasets in **numerous formats** (CSV, JSON, XML, etc.)
- Low barrier to entry with no prior programming knowledge needed
- Data transformations are **reversible** and **repeatable**, and original data are locally preserved
- Moderate learning curve with a large community of users and shared knowledge base for help
- Improvement and maintenance of data integrity for **best practices** in collections management

#### When to use?

- For **quality control**, e.g. to clean recent data entry prior to (or after) database ingestion, or to clean legacy data
- For combining and manipulating existing datasets, e.g. to transform or integrate your data with external resources like those in a taxonomic authority or Wikidata

#### When not to use?

- For adding new records individually to an existing dataset, e.g. when transcribing specimen labels
- For **text-heavy one-off data entry**, e.g. when typing a sentence in a notes field associated with each row
- For projects with multiple users on separate computers

uster & Ec	dit column "	formation"			
nis feature he ork" are very l ethod neare	lps you find gro likely to refer to est neighbor	ups of different cell values that might be al the same concept and just have capitalizat	ternative repretion difference Radius 1.5	esentations of the same thing. Fo is, and "Gödel" and "Godel" proba Block Chars 6	r example, the two strings "New York" and "new ably refer to the same person. Find out more… 14 clusters found
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	# Choices in Cluster
3	55	Budden Canyon Formation (34 rows)		Budden Canyon Formation	

OpenRe	efine SPNHC 2022 Demo	Camp Permalink Boforo
Facet / Filter	Undo / Redo 81 / 81	12 matching rows (2375 total)
Refresh	Reset All Remove All	Show as: rows records Show: 5 10 25 50 100 500 1000 rows Sort - « first < previous 1 2 of 1 page next > last »
× – phylum	change	

2	83	<ul><li> Chico Formation (72 rows)</li><li> Moreno Formation (11 rows)</li></ul>		Chico Formation	# Rows in Cluster
2	12	Deep Spring (9 rows)     Deep Springs (3 rows)		Deep Spring	0 — 230 Average Length of Choices
2	5	<ul> <li>San Lorenzo Fm (3 rows)</li> <li>San Lorenzo (2 rows)</li> </ul>		San Lorenzo Fm	
2	77	<ul> <li>Blakeley (73 rows)</li> <li>Blakeley Fm (4 rows)</li> </ul>		Blakeley	8 – 20 Length Variance of Choices
2	48	<ul> <li>Monterey Fm (43 rows)</li> <li>Monterey (5 rows)</li> </ul>		Monterey Fm	0.5 — 4.19
2	21	Budden Canyon Fm (12 rows)		Budden Canyon Fm	
Select Al	Unselect All		Expo	ort Clusters Merge Selected &	Re-Cluster Merge Selected & Close Close

# Make bulk edits to clean up similar values

1. FACET

2. CLUSTER

3. EDIT

W. P. Popenoe

Wilson

Popenoe (78)

5 choices Sort by: name count	Cluster	catalogivaliber	basison cecord	- phylum	Scientificitanie	dentineab	locationit	Iormation
S choices Soft by. Hame count	Ciuster	LACMIP 352.16	FossilSpecimen	Arthropoda	Dasyhelea	W. Dwight Pierc	LACMIP 352	Barstow
Arthropoda 2					kanakoffi Pierce,			Formation
Chordata 2					1900			
Foraminifera 4								
Mollusca 3		26015	FossilSpecimen	Chordata	Hesperocamelus		V2201	Barstow
Problematica 1		32257	FossilSpecimen	Mollusca	Macoma	W P Popence	681-	Blakeley
Facet by choice counts		52251	1 Ussilopecimen	Wollusca	lorenzoensis arnoldi	W. F. Fopence	001-	Diakeley
=	change	78155	FossilSpecimen	Chordata	Mojavemys alexandrae		V6604	Barstow
11 choices Sort by: name count	Cluster	37538	FossilSpecimen	Problematica	Nevadatubulus dunfeei		M7281	Deep Spring
Barstow 2		LACMIP 362.24	FossilSpecimen	Arthropoda	Palpomyia freyi	Lindsay Walker	LACMIP 362	Barstw
Barstow Formation 1					(Fierce, 1900)			Formation
Barstw Formation 1								
Blakeley 1		12370	FossilSpecimen	Mollusca	Pecten andersoni	W. P. Popenoe	1176-	Monterey
Blakeley Fm 1					gonicostus			200
Deep Spring 1		41743	FossilSpecimen	Foraminifera	Plectofrondicularia minuta	Maria Heikkila	B2236	San Lorenzo
Deep Springs 1		45682	FossilSpecimen	Foraminifera	Plectofrondicularia	Maria Heikkila	B7120	San Lorenzo
Monterey 1					oregonensis			Fm
Monterey Fm 1		46817	FossilSpecimen	Foraminifera	Silicosigmoilina	W P Popence	B8888	Blakeley Em
San Lorenzo 1		10017	1 Ossilopeointen	roraniiniicia	kleinpelli	W.I.I Opende	50000	Diakeley I III
San Lorenzo Fm 1		39218	FossilSpecimen	Foraminifera	Valvulineria	A. S. Menke	12825	Monterey Fm
Facet by choice counts		33301	FossilSpecimen	Mollusca	Wyattia reedensis		B9877	Deep Springs

Disambiguate entities and
add information from online
data sources
e.g. RECONCILE values in <i>identifiedBy</i> to
populate identifiedByID from Wikidata

Blakeley Fm

Zemorrian

Ze

catalogNumber	▼ basisOfRecord	💌 phylum	scientificName	identifiedBy	identifiedByID	locationID	county	formatio
LACMIP 352.16	FossilSpecimen	Arthropoda	Dasyhelea kanakoffi Pierce, 1966	William Dwight Pierce Choose new match	Q22112199	LACMIP 352	San Bernardino County	Barstow Formation
26015	FossilSpecimen	Chordata	Hesperocamelus alexandrae			V2201	San Bernardino County	Barstow Formation
32257	FossilSpecimen	Mollusca	Macoma Iorenzoensis arnoldi	Willis Parkison Popenoe Choose new match	Q67389903	681-	Kitsap County	Blakeley Formation
78155	FossilSpecimen	Chordata	Mojavemys alexandrae			V6604	San Bernardino County	Barstow Formation
37538	FossilSpecimen	Problematica	Nevadatubulus dunfeei			M7281	Esmeralda County	Deep Spring
LACMIP 362.24	FossilSpecimen	Arthropoda	Palpomyia freyi (Pierce, 1966)	Lindsay Walker Choose new match	0000-0002-2162-6593	LACMIP 362	San Bernardino County	Barstow Formation
12370	FossilSpecimen	Mollusca	Pecten andersoni gonicostus	Willis Parkison Popenoe Choose new match	Q67389903	1176-	Contra Costa County	Monterey Formation
41743	FossilSpecimen	Foraminifera	Plectofrondicularia minuta	Maria Heikkilä Choose new match	0000-0002-9048-4381	B2236	Santa Cruz County	San Lorenzo
45682	FossilSpecimen	Foraminifera	Plectofrondicularia oregonensis soquelensis	Maria Heikkilä Choose new match	0000-0002-9048-4381	B7120	Santa Cruz County	San Lorenzo
46817	FossilSpecimen	Foraminifera	Silicosigmoilina kleinpelli	Willis Parkison Popenoe Choose new match	Q67389903	B8888	Kitsap County	Blakeley Formation
Aft	er	Foraminifera	Valvulineria mcdougalli	Arnold S. Menke Choose new match	Q21338321	12825	Orangeodit County	Monterey Formation
		Mollusca	Wyattia reedensis			B9877	Inyo County	Deep Spring

Paul Popenoe (78)				
Willis Parkison Popenoe (65)	Match this Cell	Match All Identical Cells	Cancel	
Search for match		Willis Parkison Popenoe (Q673	389903)	
Maria Heikkila Maria Heikkilä (100) Maria Heikkilä (100)	WIKIDATA	<u>paleontologist</u> (1897–1981) ♂; <u>Malay</u>	spouse of <u>K</u>	<u>athryn Naomi</u>
Create new Item				

B8888

Add information from other internal data sources e.g. use CELL CROSS to populate *county* based on *locationID* with values from another spreadsheet



This project made possible by National Science Foundation Award 2027654. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.