

SPNHC 2022 DemoCamp: Using OpenRefine for natural history collections data

Erica Krimmel^{1,2} & Lindsay Walker^{1,3}

¹ iDigBio (Integrated Digitized Biocollections)

² Florida State University, Tallahassee, FL

³ Arizona State University, Tempe, AZ

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

See here for
more details and
links to resources



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**

Make bulk edits to
clean up similar values

1. FACET
2. CLUSTER
3. EDIT

OpenRefine SPNHC 2022 DemoCamp Permalink

12 matching rows (2375 total)

Facet / Filter: phylum (5 choices), formation (11 choices)

catalogNumber	basisOfRecord	phylum	scientificName	identifiedBy	locationID	formation
LACMIP 352.16	FossilSpecimen	Arthropoda	Dasyhelea kanakoffi Pierce, 1966	W. Dwight Pierce	LACMIP 352	Barstow Formation
26015	FossilSpecimen	Chordata	Hesperocamelus alexandrae		V2201	Barstow
32257	FossilSpecimen	Mollusca	Macoma lorenzoensis arnoldi	W. P. Popenoe	681-	Blakeley
78155	FossilSpecimen	Chordata	Mojavemys alexandrae		V6604	Barstow
37538	FossilSpecimen	Problematica	Nevadatulubulus dunfeeii		M7281	Deep Spring
LACMIP 362.24	FossilSpecimen	Arthropoda	Palpomyia freyi (Pierce, 1966)	Lindsay Walker	LACMIP 362	Barstow Formation
12370	FossilSpecimen	Mollusca	Pecten andersoni gonocostus	W. P. Popenoe	1176-	Monterey
41743	FossilSpecimen	Foraminifera	Plectofrondicularia minuta	Maria Heikkilä	B2236	San Lorenzo
45682	FossilSpecimen	Foraminifera	Plectofrondicularia oregonensis soquelensis	Maria Heikkilä	B7120	San Lorenzo Fm
46817	FossilSpecimen	Foraminifera	Silicosigmollina kleinpelli	W. P. Popenoe	B8888	Blakeley Fm
39218	FossilSpecimen	Foraminifera	Valvulineria mcdougalli	A. S. Menke	12825	Monterey Fm
33301	FossilSpecimen	Mollusca	Wyattia reedensis		B9877	Deep Springs

Disambiguate entities and
add information from online
data sources

e.g. RECONCILE values in *identifiedBy* to
populate *identifiedByID* from Wikidata

Add information from other
internal data sources

e.g. use CELL CROSS to populate *county*
based on *locationID* with values from
another spreadsheet

After

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



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.