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

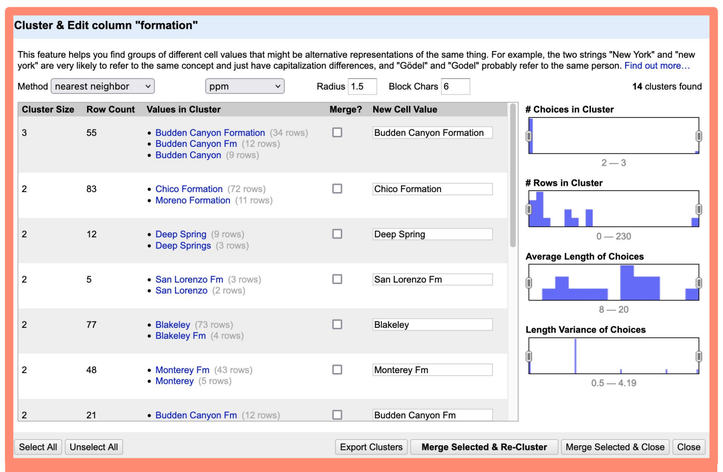
See here for more details and links to resources

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

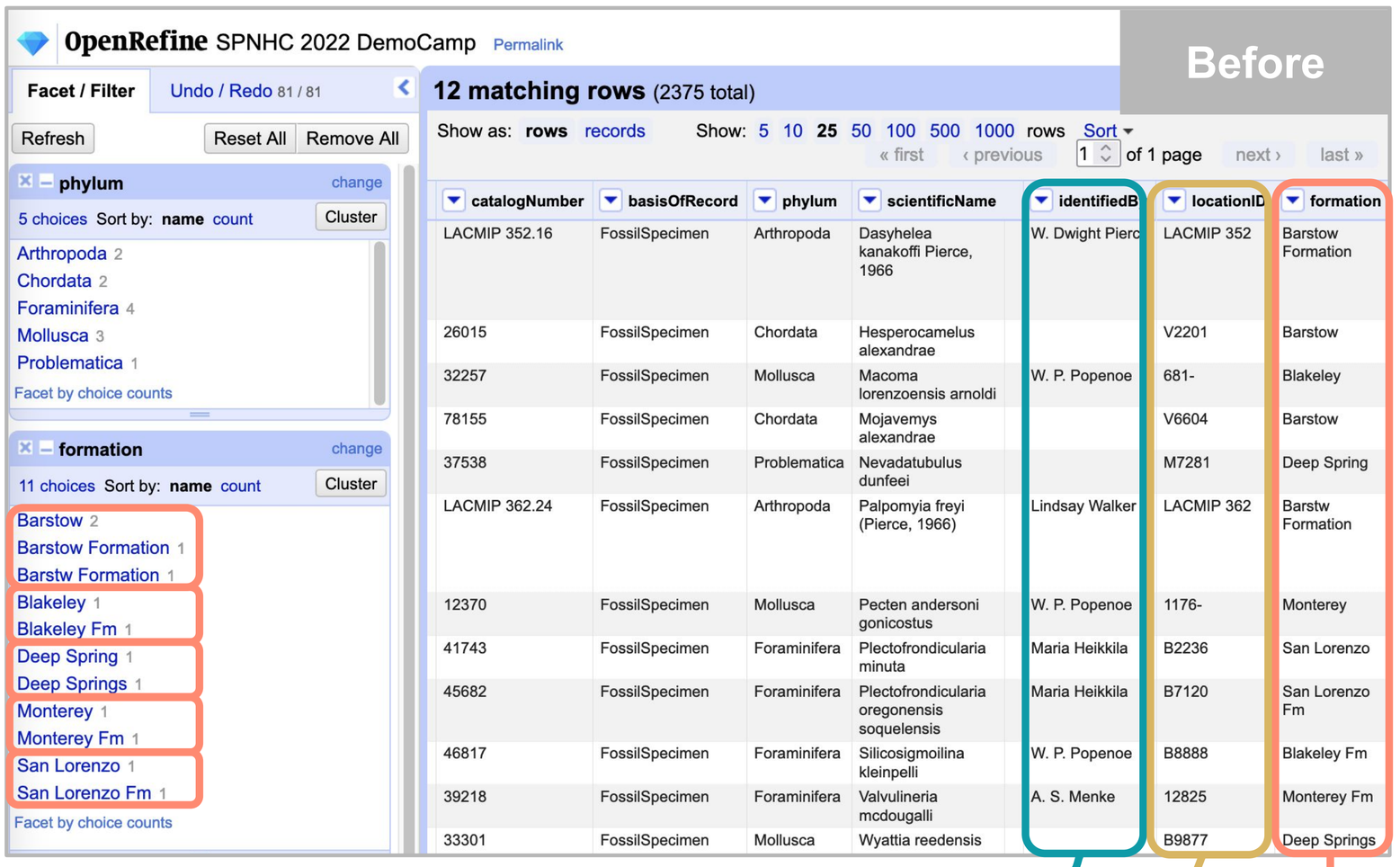
When not to use?

- 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**



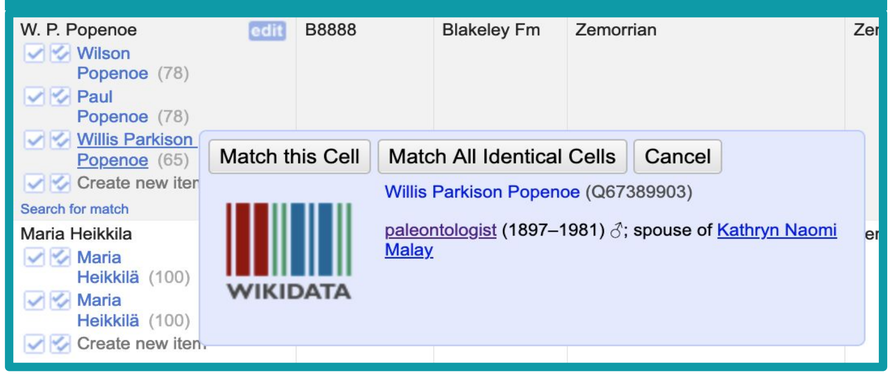
Make bulk edits to clean up similar values

1. FACET
2. CLUSTER
3. EDIT



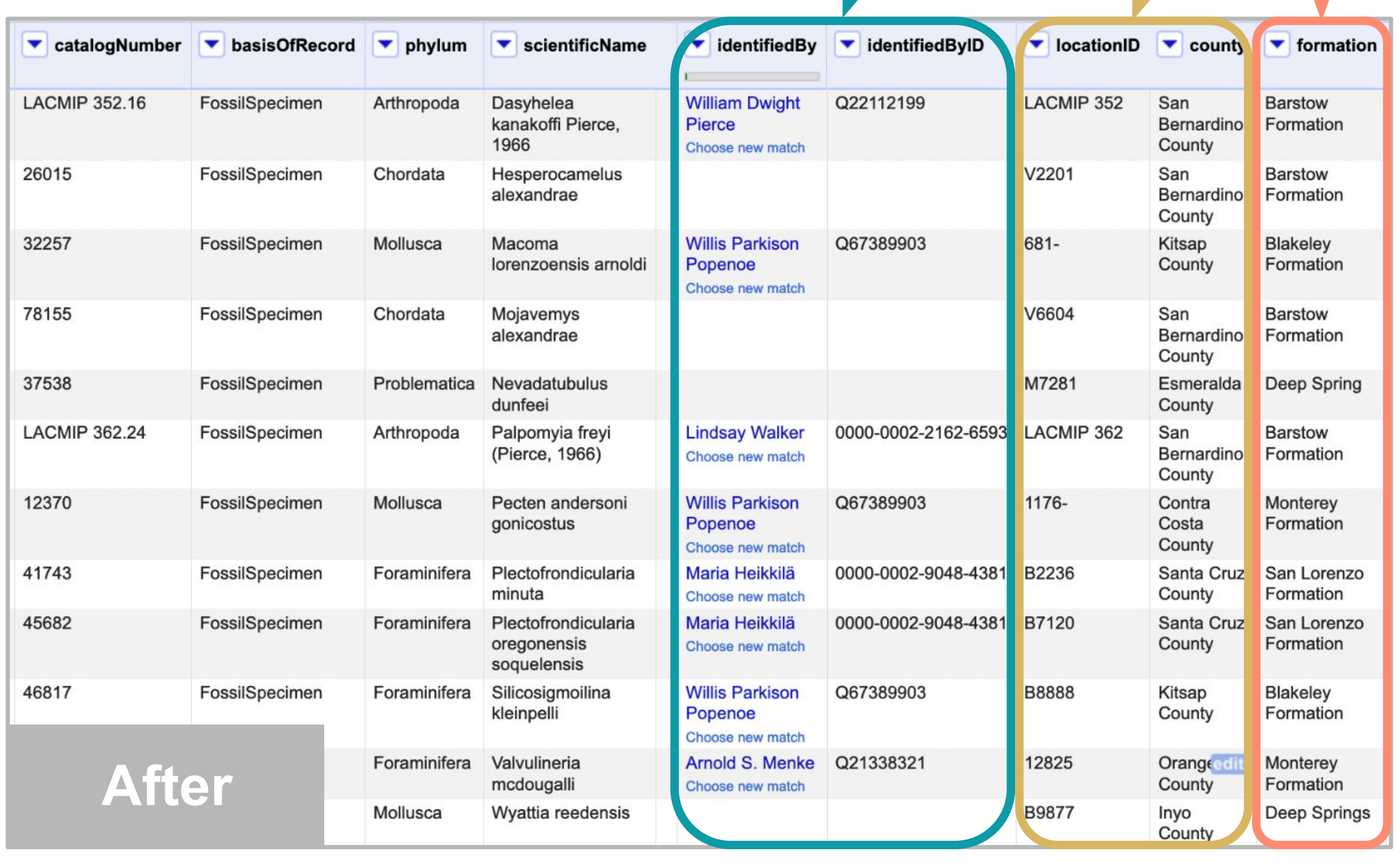
Disambiguate entities and add information from online data sources

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



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.