Using OpenRefine for Exploring Library Collections Metadata

Mary Wahl, Digital Services Librarian & Bibliographer/Librarian Liaison for Art
Oviatt Library, California State University, Northridge

Have data to explore? Try OpenRefine®!

- **What is it?**
  - Open-source tool
  - Data wrangling tool used for cleaning, manipulating, transforming, normalizing, and reconciling data
  - Uses your web browser as an interface

- **Why should I use it?**
  - Provides easy browsing of data
  - Provides previews of data manipulations
  - Great for programmers and non-programmers alike
  - Free!

- **What kind of data can I use?**
  - Allows for a variety of input formats such as Excel, CSV, XML, and Google data (e.g., Google Docs)
  - Get metadata in spreadsheet format? OpenRefine works well with data where one line = one record

- **Where can I get it?**
  - Check out openrefine.org for download, FAQs, GREL recipes and more
  - Check out the OpenRefine wiki on Github at github.com/OpenRefine/OpenRefine/wiki
  - OpenRefine.org includes list of tutorials around the web

- **How do I use it?**
  - Check out "Using OpenRefine" (Facts Publishing, 2013)
  - Just start playing with data!

*formerly Google Refine

Facets & Filters

- The most basic feature of OpenRefine: facets and filters are great for browsing and seeing a big picture of your dataset
- For each column, users can go to facet -> text facet, then sort by text or by number of cells included in the group
- Useful for finding duplicates in your data
- Can also perform a text filter to find cells that include a specific text string
- Combine with GREL to create custom facets and filters

GREL

- GREL = “Google Refine Expression Language”
- Think in patterns! If there’s a pattern in your data, you can likely write a GREL statement to transform it.
- Can start with a facet or filter to isolate data, then perform a GREL expression to transform that group
- Plenty of GREL recipes around the web to start with – Google around!
- Great for cleaning and parsing data in batches

Output

- Able to export to a variety of file types, including CSV, Excel and HTML
- Can also export as a project which will retain both the data and the history of changes

Fetching URLs

- OpenRefine can fetch data from a URL
- See the Add Column by Fetching URLs option on a column of data
- For example, have a column of locations in your data? Try geocoding them by collecting data from URLs from Google Maps

Reconciliation

- OpenRefine can connect data to a database/name registry/controlled vocabulary via reconciliation
- See Reconcile -> Start reconciling feature
- Options include pointing to SPARQL endpoints and uploading local RDF files
- For example, have a list of personal names you want to normalize? Try reconciling them against LC authority names
- Allows for "high confidence" and manual matching

Contact
Email: mary.wahl@csun.edu
Twitter: @marykwahl
Web: http://library.csun.edu/mwahl