OpenRefine Basics

Metadata for Digital Humanities

Carolyn Hansen | Sean Crowe
@dhsi17 | June 2017
@meta_caro | @crowesn | #beyondTEI

OpenRefine: The Basics

- Powerful data cleaning and remediation tool
- Formerly GoogleRefine; became open source in 2012
- Code Base: https://github.com/OpenRefine/OpenRefine
- Java and Jetty tool that runs locally
- GUI runs in your browser of your choice (not IE)

OpenRefine: Functionality

- **Import/Export** (many options: CSV, Excel, JSON, XML, RDF/XML)
- **Cleans:** good UI for cleaning, normalizing and updating
- **Facets:** helpful for finding data value outliers, normalizing values, and getting a handle of what is the state of your data
- **Clusters:** when faceting, you can then access a number of grouping algorithms for clustering the values. This can help you normalize data and see what values probably should have the same label/datapoint.
- **GREL (Google Refine Expression Language):** This is a sort of Javascript-y, application-specific language for performing normalization and data munging work in OpenRefine.
- **Extensions:** these are written by folks who want to add a functionality to OpenRefine, such as with the DERI researchers and the DERI RDF Extension.

From: christinaharlow.com/openrefine-reconciliation-workshop-c4lmdc

OpenRefine: Limitations

Data Formats: You need to transform the original data to a tabular model for working with it as an OpenRefine project, then eventually getting that tabular model data back to or transformed to the data model you want. This is easy if working with CSV, Excel, or other files already in that tabular format. Working with JSON or XML, however, you'll need to do some data massaging to make OpenRefine work efficiently for you. For heavily-nested JSON or XML, this can be a problem, and OpenRefine may not always be the best option.

Broken/unsupported extensions: Many extensions for reconciliation use the Freebase API, which no longer works. The DERI RDF extension is also no longer actively supported.

From: christinaharlow.com/openrefine-reconciliation-workshop-c4lmdc

Open Refine:
Let’s work with some data!

Open Refine: Reconciliation

Definition: “Compare values in my dataset with values in an external dataset; if deemed a match, link and pull in external datapoint information”

From: https://github.com/cmh2166/c4lMDCpres/blob/master/slides/OpenRefineReconSlides.pdf

Hands on with OpenRefine

exercises/2_exercise_openRefine.docx

Reconciliation Options

Add column by fetching URL…

- HTTP requests to external data API in UI
- takes far longer to pull data
- requires parsing returned data with GREL

Standard Recon Service API

- RESTful API between OpenRefine and external data requires tinkering knowledge of API building can host for easier use

From: https://github.com/cmh2166/c4lMDCpres/blob/master/slides/OpenRefineReconSlides.pdf

Reconciliation Options (cont.)

DERI RDF Extension

- no longer actively supported
- Standard Recon Service API to work with RDF, SPARQL endpoints
- RDF docs held in memory
- SPARQL recon dependent on SPARQL server details

From: https://github.com/cmh2166/c4lMDCpres/blob/master/slides/OpenRefineReconSlides.pdf

Learn More

“DBpedia.” http://wiki.dbpedia.org/

“Linked Data Fragments.” http://linkeddatafragments.org/

“Linked Open Jazz.” https://linkedjazz.org/

“Linked Open Vocabularies.” http://lov.okfn.org/dataset/lov/

“New York Public Library Digital Collections API.” http://api.repo.nypl.org/

“Online Coins of the Roman Empire.” http://numismatics.org/ocre/

“Open Biogeographic Information System.” http://www.iobis.org/

“OpenRefine Codebase on Github.” https://github.com/OpenRefine/OpenRefine

“OpenRefine Reconciliation Workshop Cd4lmdc.” Christina Harlow. http://christinaharlow.com/openrefine-reconciliation-workshop-c4lmdc

“Protege.” http://protege.stanford.edu/

Riley, Jenn. “Seeing Standards: A Visualization of the Metadata Universe.” http://www.dlib.indiana.edu/~jenlrile/metadatamap/

“RDF Schema 1.1.” https://www.w3.org/TR/rdf-schema/

“SPARQL Tutorial.” https://jena.apache.org/tutorials/sparql.html

“Using OpenRefine.” Ruben Verborgh and Max De Wilde. https://www.amazon.com/Using-OpenRefine-Ruben-Verborgh-ebook/dp/B00F3VNPN0?ie=UTF8&btkr=1&redirect=true&ref_=dp-kindle-redirect