OPEN Refine

A free, open source, powerful tool for working with messy data

ULS/iSchool Digital Scholarship

WORKSHOP & LECTURE SERIES

Mike Bolam

Metadata Librarian

Digital Scholarship Services

University Library System

michael.bolam@pitt.edu // 412-648-5908

Assessment Survey

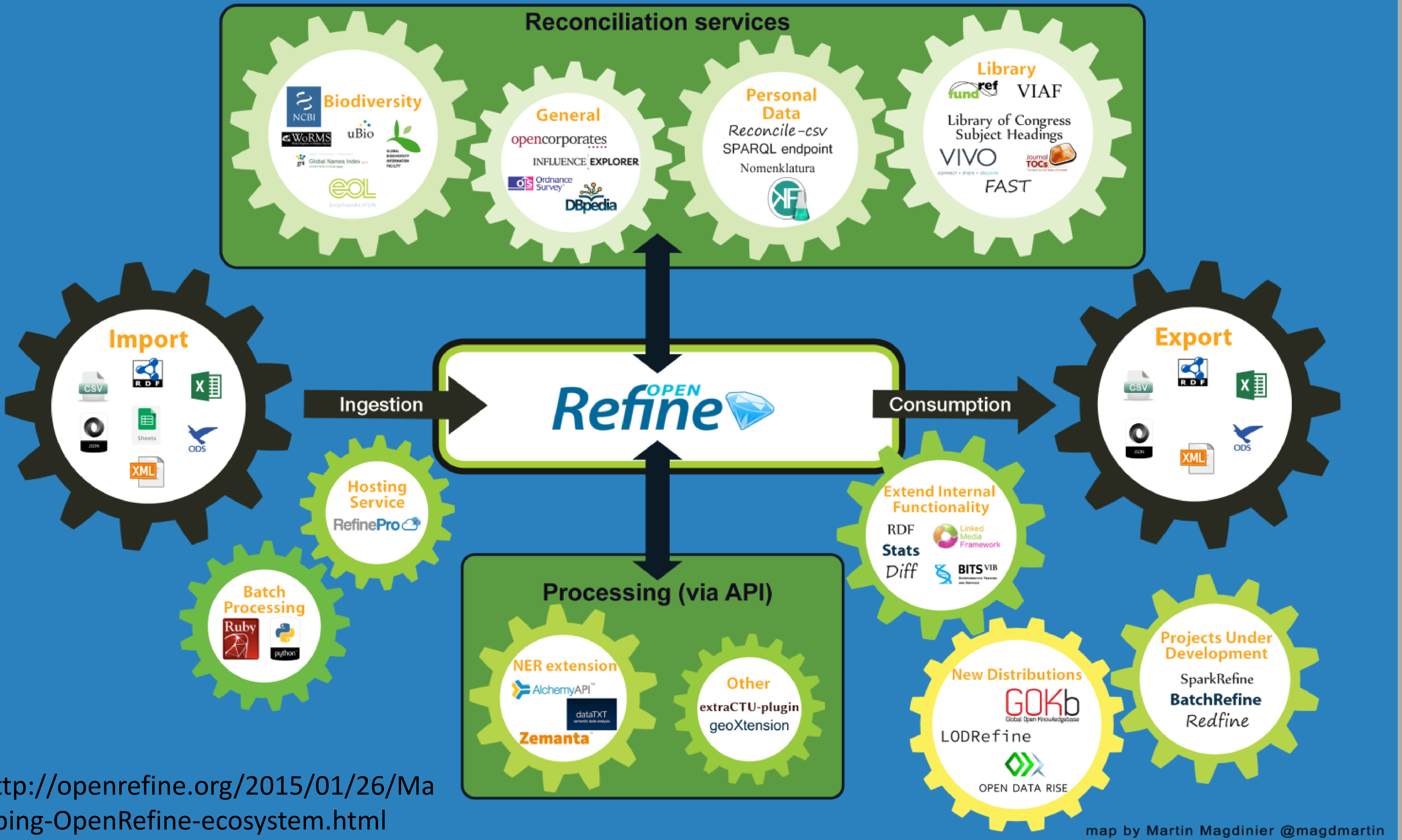
http://goo.gl/MiDZSm

Learning Objectives

- What is OpenRefine? What can I do with it?
- Installing OpenRefine
- Exploring data
- Analyzing and fixing data
- If we have time:
  - Some advance data operations
    - Splitting, clustering, transforming, adding derived columns
  - Installing extensions
  - Linking datasets & named-entity extraction

What is OpenRefine?

- Interactive Data Transformation (IDT) tool
- A tool for visualizing and manipulating data
- Not a good for creating new data
- Extremely powerful for exploring, cleaning, and linking data
- Open Source, free, and community supported
- Formerly known as Gridworks Freebase then GoogleRefine
  - OpenRefine 2.6 is still considered a beta release, so we’ll be using GoogleRefine 2.5.



Reconciliation services

Biodiversity: NCBI, WoRMS, uBio, Global Names Index, Global Biodiversity Information Facility, eol Encyclopedia of Life

General: opencorporates, INFLUENCE EXPLORER, Ordnance Survey, DBpedia

Personal Data: Reconcile-csv, SPARQL endpoint, Nomenklatura

Library: fundref, VIAF, Library of Congress Subject Headings, VIVO, Journal TOCs, FAST

Import: CSV, RDF, X, JSON, Sheets, ODS, XML

Ingestion

OPEN Refine

Consumption

Export: CSV, RDF, X, JSON, XML, ODS

Hosting Service: RefinePro

Batch Processing: Ruby, python

Processing (via API)

NER extension: AlchemyAPI, dataTXT, Zemanta

Other: extraCTU-plugin, geoXtension

Extend Internal Functionality: RDF, Stats, Diff, Linked Media Framework, BITS VIB

New Distributions: GOKb Global Open Knowledgebase, LODRefine, OPEN DATA RISE

Projects Under Development: SparkRefine, BatchRefine, Redfine

map by Martin Magdinier @magdmartin

http://openrefine.org/2015/01/26/Mapping-OpenRefine-ecosystem.html

Why OpenRefine?

- Clean up data that is:
  - In a simple tabular format
  - Is inconsistently formatted
  - Has inconsistent terminology
- Get an overview of a data set
- Resolve inconsistencies
- Split data up into more granular parts
- Match local data up to other data sets
- Enhance a data set with data from other sources

Installing OpenRefine

- http://www.openrefine.org
- Direct link to the downloads
  - https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions
- Windows
  - Download the ZIP archive.
  - Unzip & extract the contents of the archive to a folder of your choice.
  - To launch OpenRefine, double-click on openrefine.exe.
- Mac
  - Download the DMG file.
  - Open the disk image & drag the OpenRefine icon into the Applications folder.
  - Double-click on the icon to start OpenRefine.

Installing OpenRefine

- OpenRefine runs locally on your computer. It does not require an internet connection, unless you want to reconcile your data with external sources.
  - If you close you browser, you can get back OpenRefine by pointing it here: http://127.0.0.1:3333/ or http://localhost:3333
- Your data is not stored online or shared with anyone.

Getting some data

- http://www.powerhousemuseum.com/collection/database/download.php
  - http://www.powerhousemuseum.com/collection/database/opensearch/phm_collection_mar12.zip
- Created from the Powerhouse Museum metadata which been released under a CC-BY-SA Creative Commons Attribution Share Alike license.

OpenRefine Demo

Creating a new project

- Left hand menu
  - Create project
  - Open project
  - Import project
- Mention File Formats
- Choose Create Project – This Computer
  - Choose PHM TSV file
  - Click next
- Give tour of upload page
  - Explain tick boxes – especially Parse Cell into Number, etc, and
  - Quotation Marks – Quotes have not syntactic meaning in this data
    - Uncheck this box. Some data may use quotes as a delimiter (especially common in comma separated data)

Exploring your data

- Total number of rows - 75,814
- Column Headers and Menus
  - All – Star, Flag, and ID – Can star/flag rows for later operations
  - Click one of the triangles – Look at menus
- Cell Contents
  - Quick review of the data to make sure it parsed correctly (change to view 50 rows)

Manipulating columns

- Collapsing columns
  - Dropdown, View, Collapse options, collapse all
  - All dropdown, view, expand all
- Rename & Remove
  - Dropdown, Edit, rename – remove period at end of Description header
- Moving columns around
  - Move a single column? From the column dropdown. Edit column. Describe options
  - Reorder or remove multiple columns? All dropdown, Reorder/Remove

Analyzing and fixing data

- Sorting
- Faceting
- Detecting duplications
- Using text filters
- Simple cell transformations

Sorting Data

- Can sort as a visual aid or make permanent.
- Record id - Sort
- Options (point out submenu smallest/earliest first,etc)
  - Text (case ignored)
  - Number
  - Dates
  - Boolean
  - Point out we can reorder blanks and errors (don’t match our sort type).
- Sort by number, smallest first & look at results.
- Right now, only effects display
- Let’s look at the new Sort menu at the top:
  - One click to reverse, remove, or make permanent.
  - Make permanent change the order of your data.

Faceting data

- Doesn’t affect the values, but is helpful for understanding data.
    - Find errors and blanks, find misspellings, other inconsistencies
- Can be used to perform operations on a subset of data

Text Facets

- Can give you a quick idea of the type of data in a field.
- Categories, Facet, Text Facet
  - 14,805 – uses too much memory
  - Click Facet by choice counts – brings up new menu
  - Move 0 to 1000 to show most common
  - Shows top 7 – Click “7 Choices” – opens a field you can copy/paste
  - Slide back to 100 – 89 categories show up in more >= 100 records
  - Click one to return the objects with that value in categories
  - Talk about multi-value cells. We’ll use GoogleRefine later to split
- Height, Facet, Text Facet
  - Note that we can’t do numeric here, because of units
  - Sort by count, scroll down towards the bottom. Learn that units weren’t used consistently and over 45,000 records don’t have height listed

Numeric Facets

- Numbers are green, assuming you checked “parse dates, numbers…”
  - If not, reload data
- Record ID, Facet, Numeric
  - Get an overview of the number. Move slider – 0 to 500,000. See just the 536 with ids higher than 500,000
  - Check out the check boxes – 3 Non-Numeric?
  - Mouse over record id & click edit – See there is a space there?
  - Click it and apply to all matching cells
  - Just “cleaned up” some data

Timeline and Scatterplot

- Timeline
  - Production date, edit cells, common transformations, “to date” (only 72 converted)
  - Facet Timeline – Most don’t have dates, but we can use to see those that do
  - Unclick blank & Non-Time - Move slider
- Scatterplot
  - Good for grouping numeric data. This data set doesn’t really have a good data, but check it out

Customized Facets

- Many other options –
  - Customized using GREL – GoogleRefine Expression Language
  - Some pre-designed options
    - Word – lists all the different words (text facet lists strings). Data set a bit large for this
    - Duplicates – Find duplicate records (do more with this in a minute)
    - Numeric Log & 1-bounded numeric log – Can be used to help correlate dissimilar numeric sets.
    - Text length – number of characters field (try on description – Move slider around)
    - Unicode Char-code – less than 128 – English; accented & other characters up to 256, even higher for Arabic or Chinese, for example
    - Error – True/False
    - Blank – True/False – Try it on "Weight"

Faceting by Star/Flag

- Want to find all the rows that either have either diameter or weight
- Facet by blank on both and click “false”, but this isn’t right. Showing objects that have weight & diameter, not weight or diameter
  - Mouse over and exclude “weight”, so we’re just showing objects with diameter.
  - All – Edit Rows – Star
  - Mouse over and exclude “diameter”, and click false on weight.
  - All – edit rows – star
  - Close facets – All – Facet by star – Click True – 2256 rows that have weight, diameter or both!
  - Unstar the rows before moving on

Detecting Duplicates

- This only works on strings, so it won’t work with the numeric data in the first Record ID column
- Registration Number – Facets – Customized Facets – Duplicate (281)
  - Click true to display – A lot are blank – but some appear to be true dups.
- Registration Number – Facets – Customized Facets – Blank
  - Click false to show the 163 real duplicates
- Registration Number – Facets – Text
  - Sort by count – Most show up twice, with a few outliers
- Record ID – Sort – Numbers – Smallest first
  - This time, use sort menu to make it permanent – See the dups together now
- Record ID – Edit Cells – Blank Down (can be dangerous, but you can undo if something doesn’t work).
  - Blank down keeps the first instance, and deletes everything below it
- Remove all facets – Record id – facets- custom – facet by blank – to show the redundant rows.
  - We’ll learn how to remove them later.

Using the project history

- Click Undo/Redo tab
- Demonstrate earlier actions
    - Be careful – Going back will erase all subsequent steps. Can’t go from step 5 to step 2 without undoing 3 & 4.
    - Note – Can extract and edit JSON of actions for more control. Outside scope.

Removing matching rows

- Start with the baseline data again – Use undo/redo to go to step 0.
- Change data from records to rows
- Remove all rows without a record id
    - Record ID – Numeric Facet – uncheck numeric – show 3 non-numeric
    - All – Edit Rows – Remove all matching – Clear Facet – 75,811
- Remove all with no registration number (suspicious in museum collection)
    - Registration Number – Facets – Custom – Facet by blank – Click true (all blank)
    - All – Edit Rows – Remove all matching – Clear Facet – 75,696
- Remove duplicates (remember back with we detected them. A bit trickier)

Removing matching rows (cont.)

- Registration Number – Facet – Custom – duplicate – True (163)
  - If we suppress all of these, we lose the original and the duplicate data.
- Registration Number – Sort – text – a-z // Sort – Reorder Rows Permanently
- Registration Number – Edit Cells – Blank down (should effect 84)
- All – Edit Rows – Remove all matching – Clear Facet – 75,612
- First removed 3 blanks, then 115 further that only had record ID. Then 84 dups by reordering/blank down on the Reg. Number.
  - Cleaned up a total of 202 rows

Applying Text Filters

- Object Title, Text Filter – USA - 1866 matching rows
  - Might be getting words like Jerusalem or usable
  - Flip on case-sensitive – 1737 rows, but could still get JERUSALEM or USABLE
- Try adding a space in front and behind – Drops to 172– maybe missing examples because of commas or other characters at the front or end
- We can use a regular expression
  - \bU.?S.?A.?\b – 1983 Rows – Catches USA & U.S.A.
    - \b sets word boundary
    - .? – Match preceding character 0 or 1 times

Simple Cell Transformations

- Object Title, edit cells, common transforms
  - Trimming whitespace - at beginning or end of string
  - Collapse consecutive whitespace – remove multiple spaces
  - Unescape HTML Entities - &eacute; or &00010; - can convert that to appropriate text
  - To titlecase/To uppercase/to lowercase
  - Blank out cells – Clear everything from a column

Exporting a project

- Top left menu-
  - Export project lets you save everything, including history of edits (undo/redo). Share with others or move to another device
    - Import project from original menu
  - Variety of data export options
  - Triple Loader and MQLWrite – must align with pre-existing schema (outside scope)
  - Custom table exporter – Tight control over export. Can select and order columns exported, control date format, reconciliation results
  - Templating – More advanced control using JSON (outside scope)
  - More on RDF exports later

Getting more memory

- Windows
  - Google-refine.l4j.ini
    - # max memory memory heap size
    - -Xmx2048M
- Mac (more complicated)
  - Ctrl-click application, choose Show Folder Contents, Contents, info.plist
  - Find VMOptions – change Xmx1024 to Xmx 2048

Going for more memory

- Consider how much memory you have available.
    - I probably wouldn’t increase above the standard 1gig, if you only have 4 gig RAM on your device.
- Close Refine – ctrl-c (windows)
- Switch to powerpoint.
    - Windows
        - Google-refine.l4j.ini
            - # max memory memory heap size
            - -Xmx2048M
    - Mac (more complicated)
        - Ctrl-click application, choose Show Folder Contents, Contents, info.plist
        - Find VMOptions – change Xmx1024 to Xmx 2048

Advanced Data Operations

- Handling Multi-value Cells
- Alternating between rows & records
- Clustering similar cells
- Transforming cell values
- Splitting data across columns
- Transposing rows & columns

Handling Multi-value cells

- Categories, facet, text facet – “too many to display”
  - Edit cells, split multi-valued – enter “|” and click okay
  - Sort facet by count – show editing option with clothing and dress, but don’t change it.
- Demonstrate rejoining them with comma or other separator, but don’t do it.

Alternating between rows and columns

- Click on numismatics
  - Problem – Just see the rows where numismatics appears, but not the rest of the data.
  - Switch to record view at the top, and will show full record that contain numismatics
  - Point out the way the grey/white goes from rows to columns
- Make sure you’re in row view for the rest of the exercise.

Clustering similar cells

- Categories – Edit Cells, Cluster and edit…
    - Describe the columns – Point out “Biological specimans” – Caps not consistent with others. Fix, select all, and Merge & Recluster
- Other clustering options – ngram fingerprint
    - Looks like good combinations – see a problem? T-shirt vs shirt? Select all, then unselect that one.
- Nearest neighbor w/defaults – probably not that helpful for this dataset, but worth taking a look at.
    - Geological specimens vs geological specimen
    - Hell money vs Shell Money? Looks like a typo to me. Click “browse this cluster” to get a better idea
- Usually best to try out different options and see what works best for your dataset.

Linking Datasets

- Installing Extensions
- Adding reconciliation service
- Reconciling with Linked Data
- Extracting named entities

Installing extensions

- Hit the “open button” in the top left – Look for Browse Workspace Directory - See extensions folder?
- Or…go to installation point, click webapp – see extensions folder?
- Go to http://refine.deri.ie // Downloads.
  - Download latest and unpack the zip file
- Move the rdf-extension folder to the GoogleRefine Extensions folder
- Restart GoogleRefine, and open your project
- Should see an RDF menu on the right side

Adding a reconciliation service

- Click RDF – Add reconciliation service – based on SPARQL endpoint
- You can use any publicly available endpoint, but for the exercise, we’re going to use one set up by the freeyourmetadata.org crew using Library of Congress Subject Headings
  - Name: LCSH
  - Endpoint URL: http://sparql.freeyourmetadata.org/
  - Graph URI: http://sparql.freeyourmetadata.org/authorities-processed/
  - Type: Virtuoso
  - Label Properties – tick only skos:preflabel

Named Entity Extraction

- http://software.freeyourmetadata.org
- Download ner-extension.zip and unpack it.
- Put it in your extensions folder (just like before)
- Restart GoogleRefine
- Create new project, using the same dataset

Adding a reconciliation service

- Endpoint URL – address where the endpoint is located
- Graph URI – identify which dataset within the endpoint to use
- Type – endpoint software – knowing the correct type improves speed
- Label properties – names of fields that can be used to look up the cell value

Linking Data

- This can be slow, depending on internet connection, network traffic, size of data set.
- We’re going to work with an arbitrary subset. Text facet on height, sort by count, include just 215mm results –
    - Text facet on categories (84 categories, 400 rows).
- Categories, Reconcile, Start Reconciling
    - LCSH – might take a minute – Should say skos:Concept & have a path.
- This might take a little while – Going out over the network to pull back these terms
- While running, show NER downloads and talk about other resources.
- Good time to fill out your survey, too!

Linking Data

- Automatically linked those with high confidence
  - Click link – explain LOC page
- Suggests matches for those with lower confidence
  - Can choose or search. Understanding subject headings and structure of vocabulary would help at this point.
- Some it will find no results
- Edit Column – Make new column based on these “Category URLs”
  - cell.recon.match.id with “set to blank” chosen
- Can clear the reconciliation now
  - Reconcile, actions, clear recon data

Named Entity Extraction

- Use a subset again. Same method – Height - Text facet – Pick one around 10 – 885mm?
- Again, will take a while – Finish up survey?
- Check out the results – Linking to DBPedia, the database behind Wikipedia.

Take it to the next level

- Regular Expressions
  - GREL – GoogleRefine/OpenRefine Expression Language
  - JYTHON – Python Written in Java
  - Clojure – A dialect of the LISP programming language
- GREL Resources
  - https://github.com/OpenRefine/OpenRefine/wiki/Google-refine-expression-language

Resources

- OpenRefine Wiki
    - https://github.com/OpenRefine/OpenRefine/wiki
- OpenRefine User Documentation
    - https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users
- Using OpenRefine [book – ebook available via PittCat]
    - https://www.packtpub.com/big-data-and-business-intelligence/using-openrefine
- Free Your Metadata Site
    - http://freeyourmetadata.org
- Linked Data for Libraries, Archives, and Museums [book – available at Hillman Library]
    - http://book.freeyourmetadata.org

Assessment Survey

http://goo.gl/MiDZSm