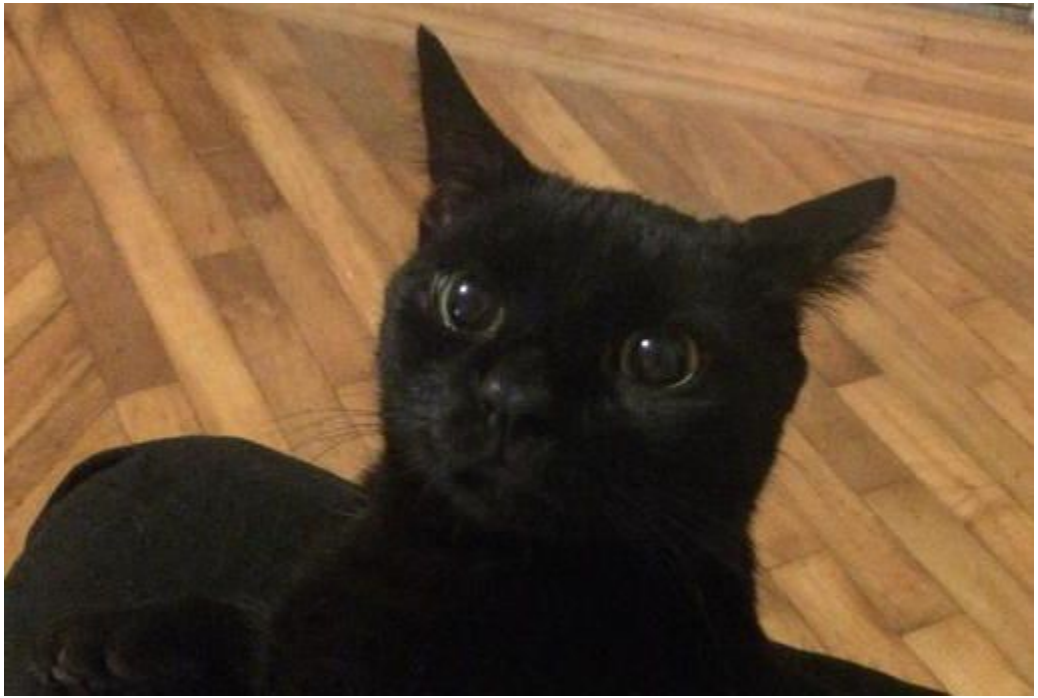
Cleaning & Publishing SHARE Metadata

A Hands on Tutorial with OpenRefine

Christina Harlow, @cm_harlow

Slides, Examples, + Install

github.com/cmh2166/SHAREOpenRefineWkshop



Don't Have OpenRefine Installed?

You can try to install during OpenRefine Intro, or watch & test later on your own.

Agenda

- Introduction
- Importing Data
- Data Munging
- Reconciliation
- Publishing & Exporting
- Wrap-up

Agenda

- Introduction 
- Importing Data
- Data Munging
- Reconciliation
- Publishing & Exporting
- Wrap-up

Goals for & Getting Help in this Workshop

Goal:

- Learn How to Assess & Remediate SHARE Metadata with OpenRefine

Help:

- Speak Up, Check Instructions / Repository Documents, Check Online

"Hacker School Rules" Please

- No feigning surprise
- No well-actually's
- No back-seat driving
- No subtle -isms

https://www.recurse.com/manual

What is OpenRefine?

- OpenRefine = power data tool
- Since 2012, community-sourced
- OpenRefine.org
- github.com/OpenRefine/Openrefine
- Java (& Jetty) tool that runs locally
- GUI runs in your chosen browser (NOT INTERNET EXPLORER)

OpenRefine: Making Data Work

- Start in Data Journalism
- Transforms everything to tabular format
  - Works best with less-nested data formats
- Facets & Subset Selection in a GUI
- Data Transforms according to GREL & Regular Expressions
- String Matching Algorithms Available
- OpenRefine Extensions Available
- External Data Matching via a "Reconciliation Service API"

OpenRefine & RDF / Linked Data

- Native support for importing RDF/XML, RDF Ntriples
- Original Freebase Extension
- DERI RDF Extension / LODRefine
  - RDF Document Reconciliation
  - RDF Skeleton, Mapping
  - RDF Export: RDF/XML, RDF Turtle
  - Ask me if you want the extension

OpenRefine Help & Resources

- OpenRefine.org
- github.com/OpenRefine/Openrefine
  - Especially the Wiki on this GitHub Repository
- www.meanboyfriend.com/overdue_ideas/tag/openrefine/
  - Posts & Help by Owen Stephens, library technologist in the UK
- http://data-lessons.github.io/library-openrefine/

Agenda

- Introduction
- Importing Data 
- Data Munging
- Reconciliation
- Publishing & Exporting
- Wrap-up

First, start OpenRefine.
Second, create a Project that we will import data into.

Then... What Data Can I Import?

- TSV (tab-separated values)
- CSV (comma-separated values)
- Excel
- JSON (javascript object notation)
- XML
- Google Spreadsheet
- …

And... How Can I Import that Data?

- Files from your local machine
- Specify a URL with structured data
- Copy & Paste data
- Retrieve data from Google Drive

Follow along with

Instructions/Importing.md

In our Workshop's GitHub Repository

Create a Project & Import Data from the Option Most Relevant to You:

OAI-PMH URL

XML File

CSV File

SHARE API URL

Agenda

- Introduction
- Importing Data
- Data Munging 
- Reconciliation
- Publishing & Exporting
- Wrap-up

First, let’s acquaint ourselves with our OpenRefine Project’s interface...

Instructions/Interface.md

Metadata Munging in OpenRefine
Ways to Normalize, Remediate Data:

- Join, Split Rows
- Splitting, Renaming Columns
- Faceting, Clustering & Filtering
- Google Refine Expression Language (GREL)

https://github.com/OpenRefine/OpenRefine/wiki

Follow along with

Instructions/Munging.md

In our Workshop's GitHub Repository

For Your Project, Move Rows & Columns so your data table best represents your records. Be aware of the Records / Rows split.

Faceting, Clustering & Filtering

- `Facets` group all values in a column
- `Text Filters` facets according to provided string
- `Clustering` groups facet values by various string-matching algorithms
- While a facet/filter is in place, only that data is changed
- You can batch edit facet values

Follow along with

Instructions/Faceting.md

In our Workshop's GitHub Repository

For Your Project, Select 1-2 columns that could be normalized. Apply a facet, a text filter, a custom filter, & clustering. Feel free to edit as well.

GREL Transformations

- 'GREL' or the Google Refine Expression Language
- Applies Transformations to cells in a column
- Relies on the Data Type of the cell
- https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language

Follow along with

Instructions/GREL.md

In our Workshop's GitHub Repository

For Your Project, try to apply some of the GREL transformations provided in GREL.md on an appropriate column.

Agenda

- Introduction
- Importing Data
- Data Munging
- Reconciliation 
- Publishing & Exporting
- Wrap-up

Reconciliation broadly...

Compare values in my dataset with values in an external dataset, if deemed a match, link and pull in external datapoint information

... aka Matching

Matching: Add column by fetching URL...

- HTTP requests to external data API
- Can take far longer to pull data
- Requires parsing returned data (in a new column) with GREL
- See Instructions/Reconciliation/addcolumnexamples.md for some examples of this

Matching: Standard Recon Service API

- RESTful API between OpenRefine and external data
- Handles JSON reconciliation objects btwn datasource API + Openrefine

Follow along with

Instructions/Matching.md

In our Workshop's GitHub Repository

For Your Project, select an appropriate column (authors, for example) to match to VIAF using the hosted VIAF Reconciliation Service.

Agenda

- Introduction
- Importing Data
- Data Munging
- Reconciliation
- Publishing & Exporting 
- Wrap-up

Exporting Data

- There are some common tabular representations you can export to
- You can also use the Templating feature to export data according to a template
  - This is a great example of templating: http://digitalscholarship.utsc.utoronto.ca//content/blogs/converting-spreadsheets-modsxml-using-open-refine

Follow along with

Instructions/Exporting.md

In our Workshop's GitHub Repository

For Your Project, experiment with the Export Template option.

Agenda

- Introduction
- Importing Data
- Data Munging
- Reconciliation
- Publishing & Exporting
- Wrap-up 

Questions? Requests?

github.com/cmh2166/SHAREOpenRefineWkshop