Mar 31st, 11:15 AM - 12:00 PM

Using OpenRefine to Standardize and Augment Your Data

Blake Galbreath
Washington State University, blake.galbreath@wsu.edu

Using OpenRefine to Standardize and Augment Your Data

Blake L. Galbreath
Core Services Librarian
Washington State University

Presentation Overview

1. What is it?
2. Installing
3. Quick orientation
4. GREL
5. Transformations
6. Facets and Clusters
7. Augmenting data

What is OpenRefine?

formerly GoogleRefine

Definitions

Awesome!

Program that runs locally, displays in browser

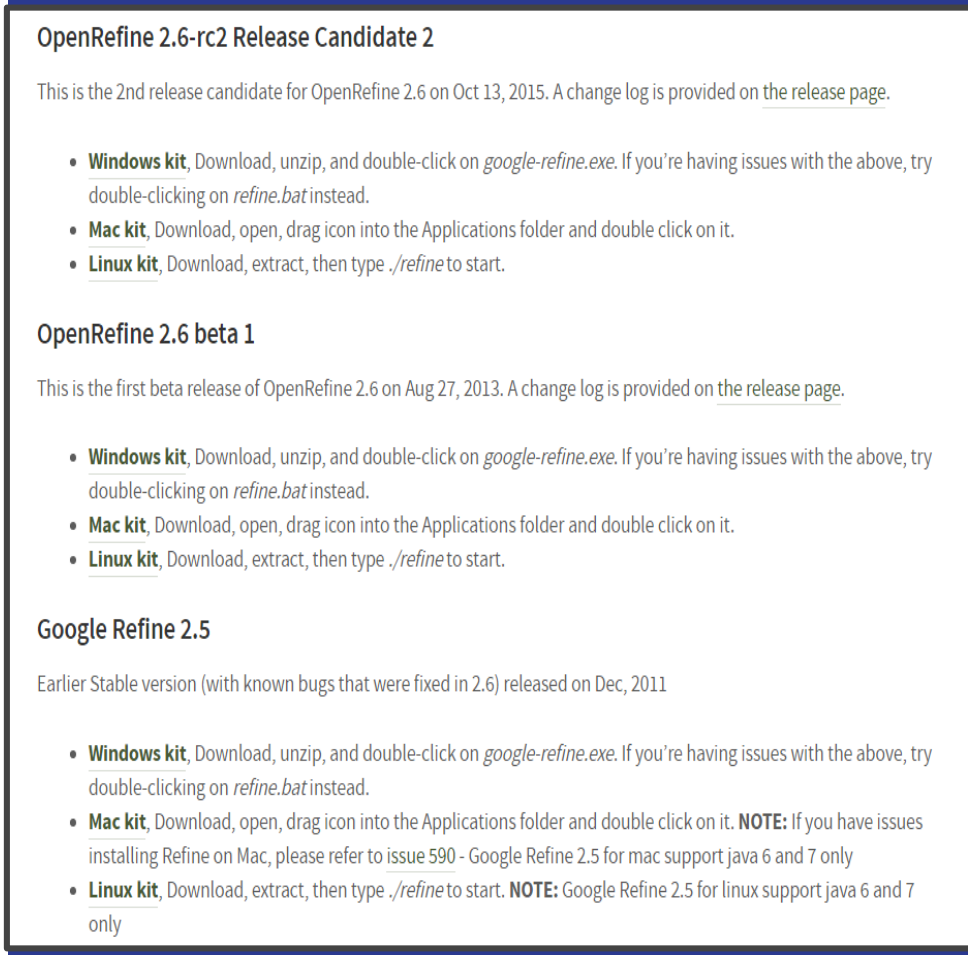
Allows you to deal with data in various formats

Can fetch information from outside sources

Installing OpenRefine

http://openrefine.org/download.html

Read the notes



OpenRefine 2.6-rc2 Release Candidate 2

This is the 2nd release candidate for OpenRefine 2.6 on Oct 13, 2015. A change log is provided on the release page.

- **Windows kit**, Download, unzip, and double-click on *google-refine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it.
- **Linux kit**, Download, extract, then type *./refine* to start.

OpenRefine 2.6 beta 1

This is the first beta release of OpenRefine 2.6 on Aug 27, 2013. A change log is provided on the release page.

- **Windows kit**, Download, unzip, and double-click on *google-refine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it.
- **Linux kit**, Download, extract, then type *./refine* to start.

Google Refine 2.5

Earlier Stable version (with known bugs that were fixed in 2.6) released on Dec, 2011

- **Windows kit**, Download, unzip, and double-click on *google-refine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it. **NOTE:** If you have issues installing Refine on Mac, please refer to issue 590 - Google Refine 2.5 for mac support java 6 and 7 only
- **Linux kit**, Download, extract, then type *./refine* to start. **NOTE:** Google Refine 2.5 for linux support java 6 and 7 only

How Does it Work?

Super-quick orientation

Launch OpenRefine

Open [Existing] Project or

Create New Project

Import File

Get to work!

GREL: General Refine Expression Language

GREL

Functions

Array

Boolean

Date

Math

Other

String

GREL

Functions: Boolean Example

value = “dog, cat, mouse”

value.contains(“cat”)

→ TRUE

GREL

More please!

https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language

Also, Regex!

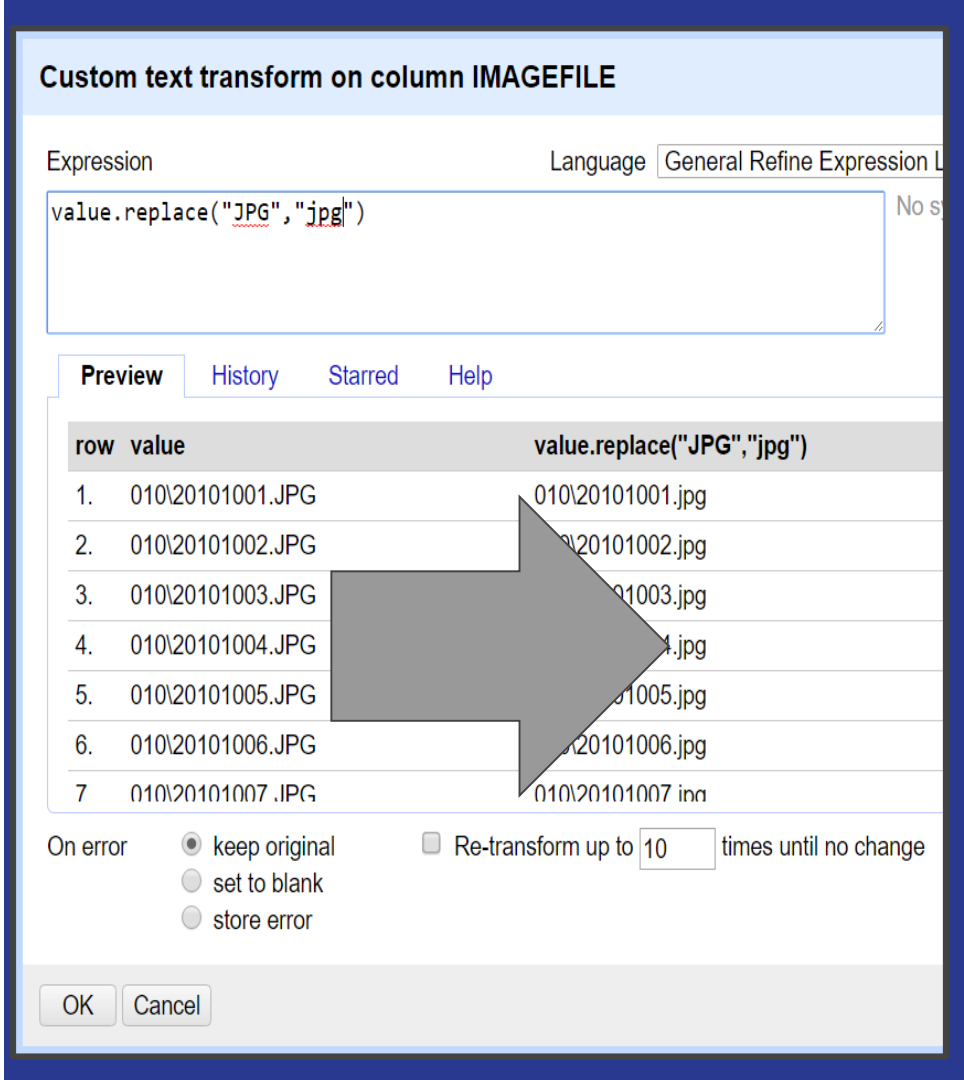
Powerful stuff

```
value.replace(/[A-Z]/,";")

value.replace(/([A-Z])/,";$1")

value.replace(/( [A-Z])/,";$1")
```

Transformations

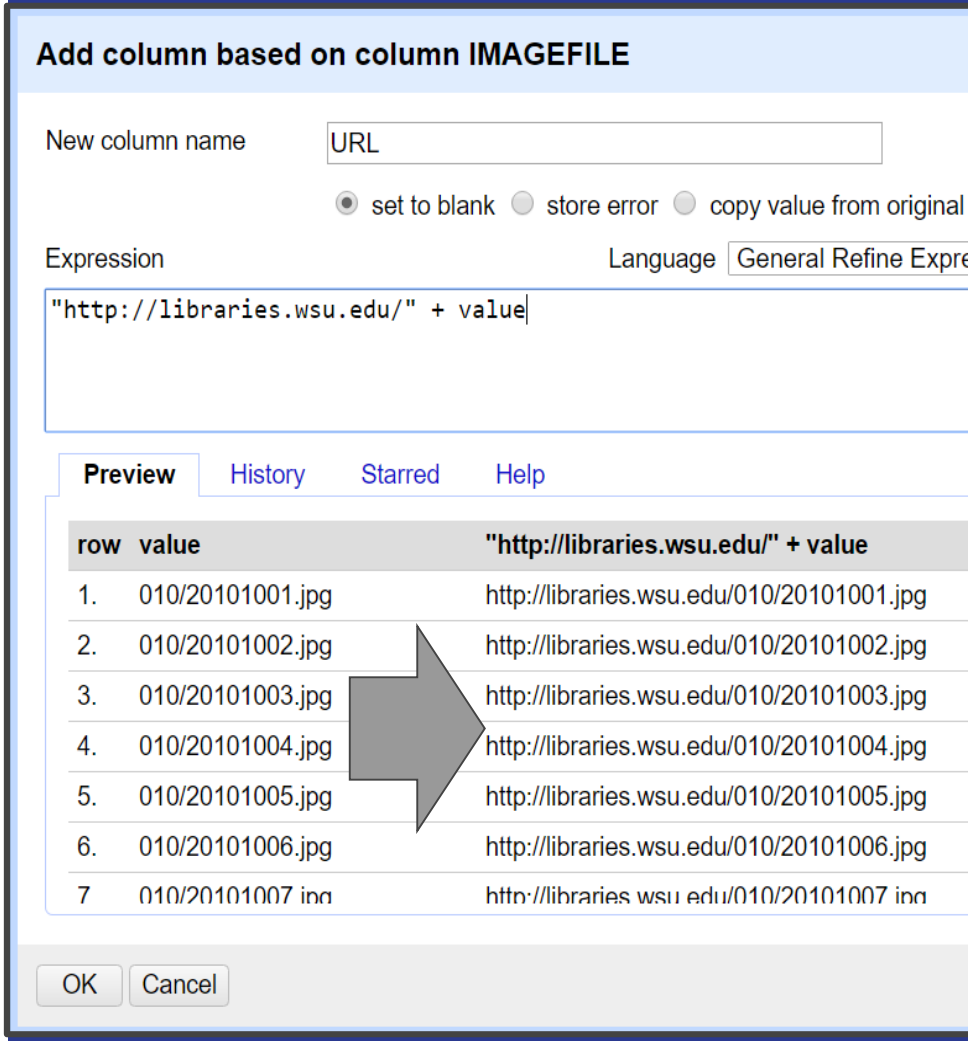
Manual Transforms

Replace



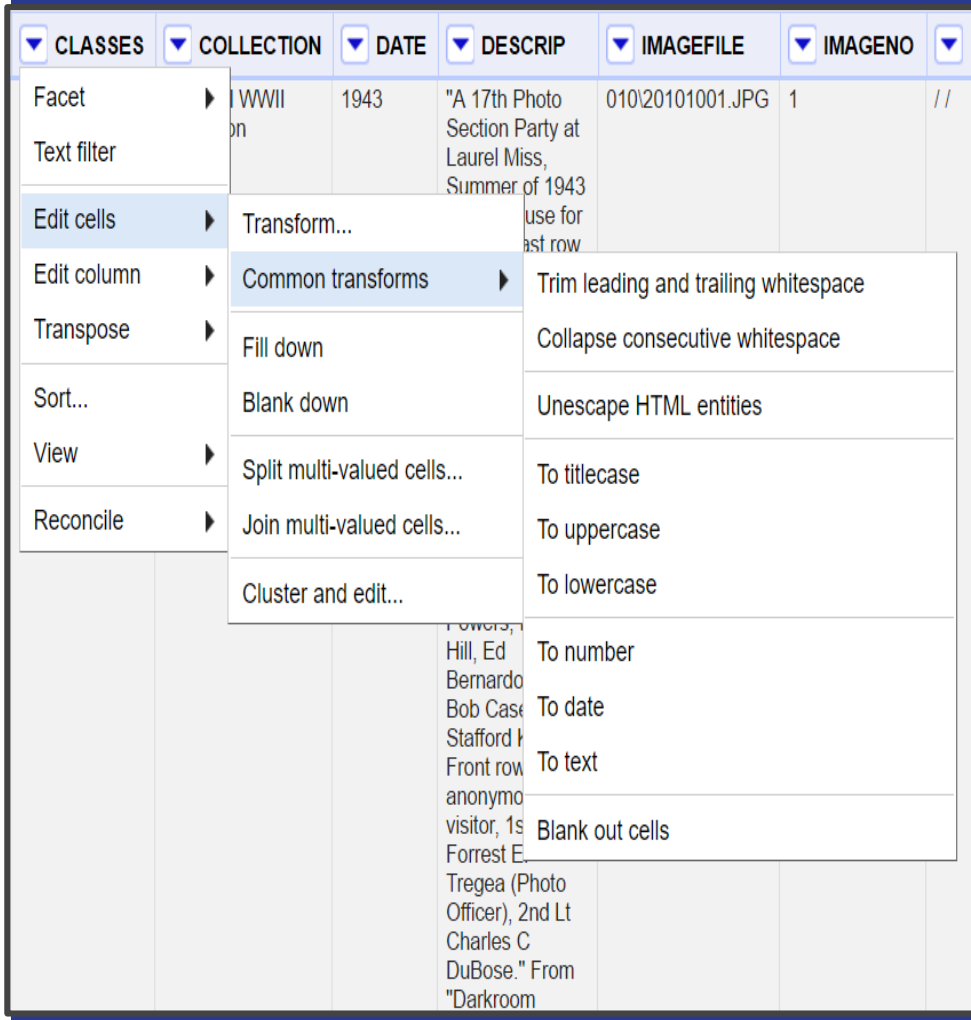
Manual Transforms

Concatenate



Common Transforms

Trim, etc.



Facets and Clusters

Prepare data <u>before</u> facet/cluster

Filter or Cluster

1. Edit cells → Transform

   (into delimited strings)
2. Edit cells → Split

   multi-valued cells
3. Clean up with

   a. Facet → Text facet

      OR

   b. Edit cells → Cluster

Place ; before capital letter, if leading space

Step 1

value.replace

(/(space[A-Z])/,";$1")

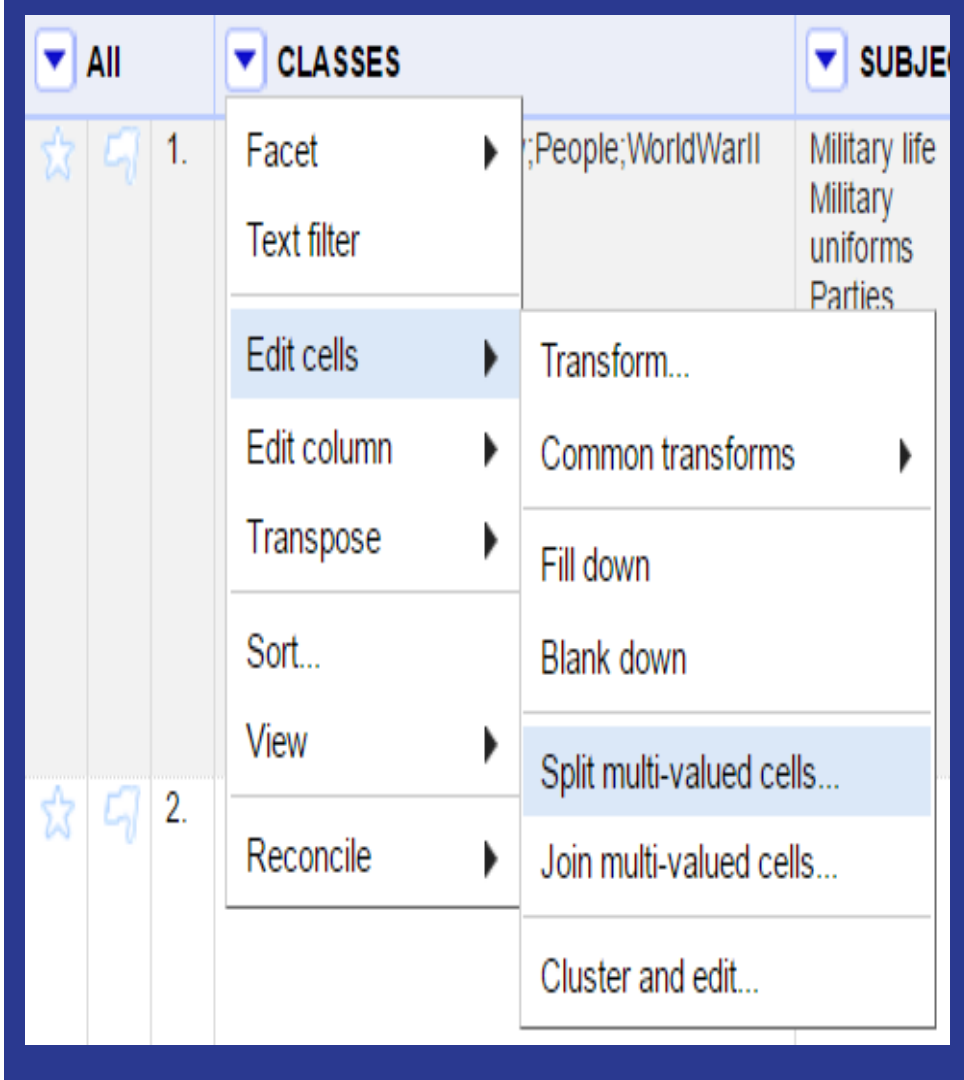
and

.replace("; ",";")

___

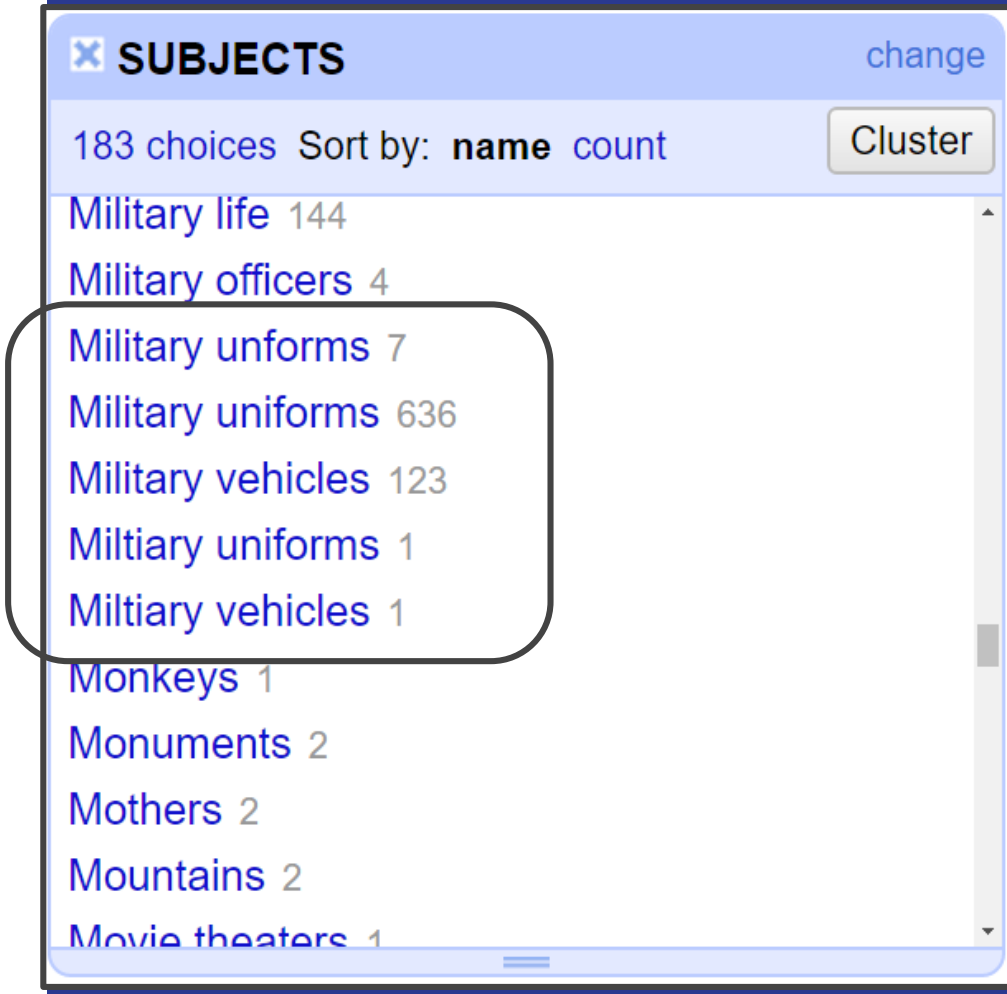
Split data on newly-inserted semicolon

Step 2



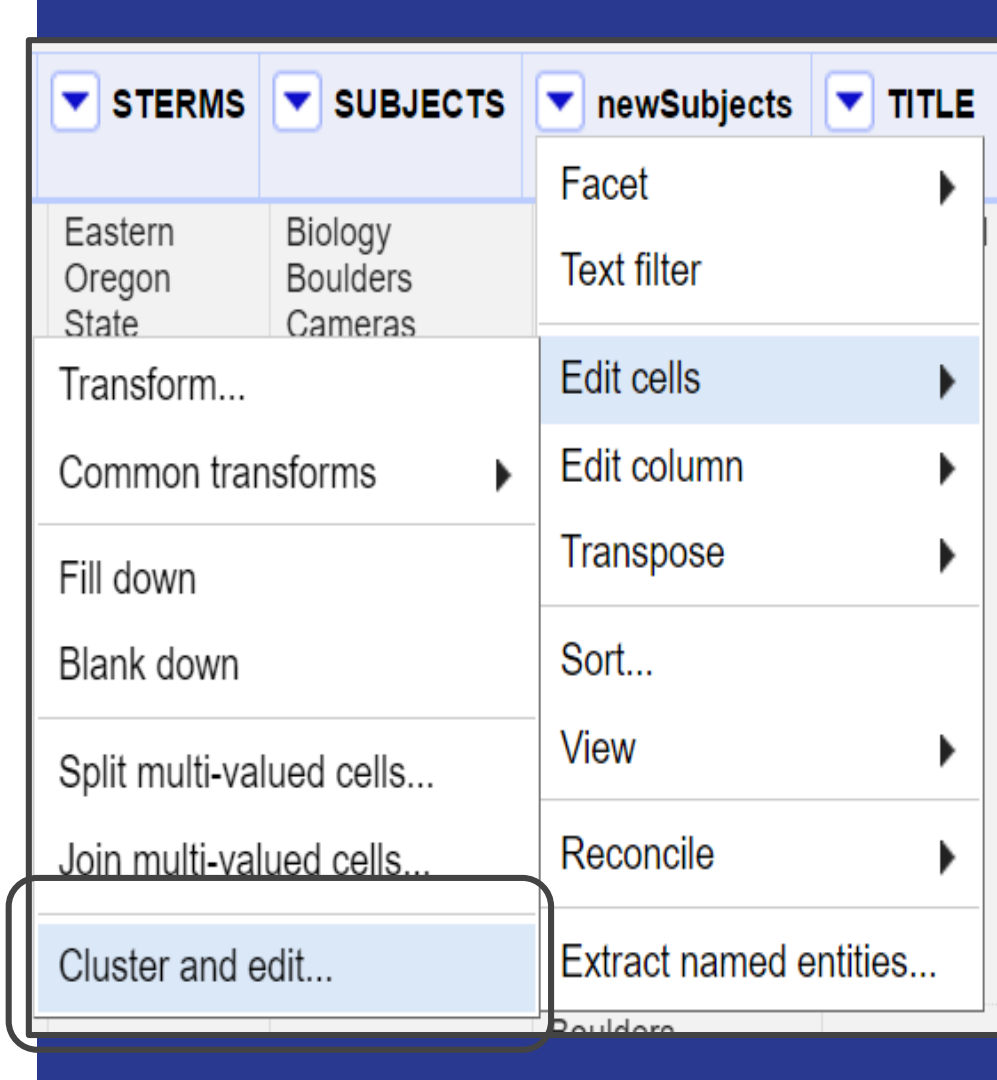
Facets

Step 3A



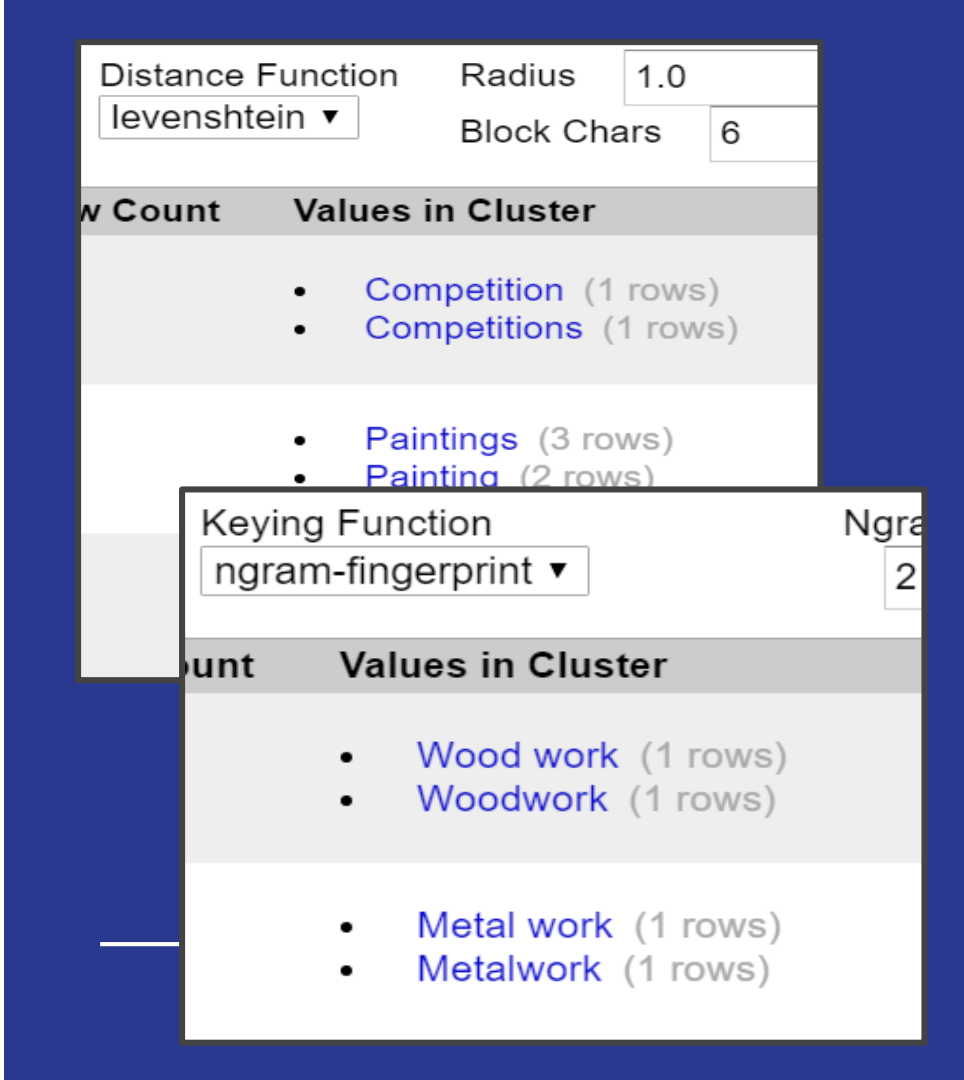
Clusters

Step 3B



Clusters

Step 3B



Clusters

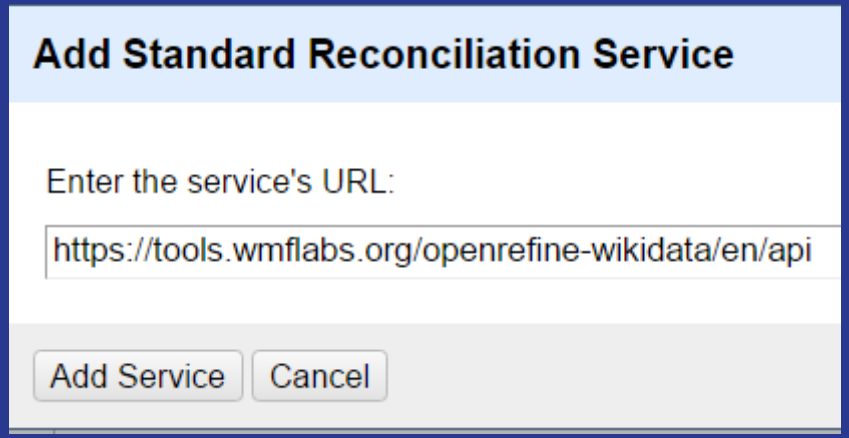
More please!

https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

Augment Data from Web Services

Get Wikidata

Wikidata Reconciliation for OpenRefine (en)



1. Reconcile > Start Reconciling

Get Wikidata

Wikidata Reconciliation for OpenRefine (en)

2. Select entity to reconcile against

3. Create URLs where titles match

Get CrossRef Data

Fetching URLs

1. Fetch URLs from ISSNs

"http://api.crossref.org/journals/" + value

Get CrossRef Data

Parse JSON

2. Add column based on fetched column

value.parseJson().message.title

value.parseJson().message.publisher

Questions

Extras! If we have time...

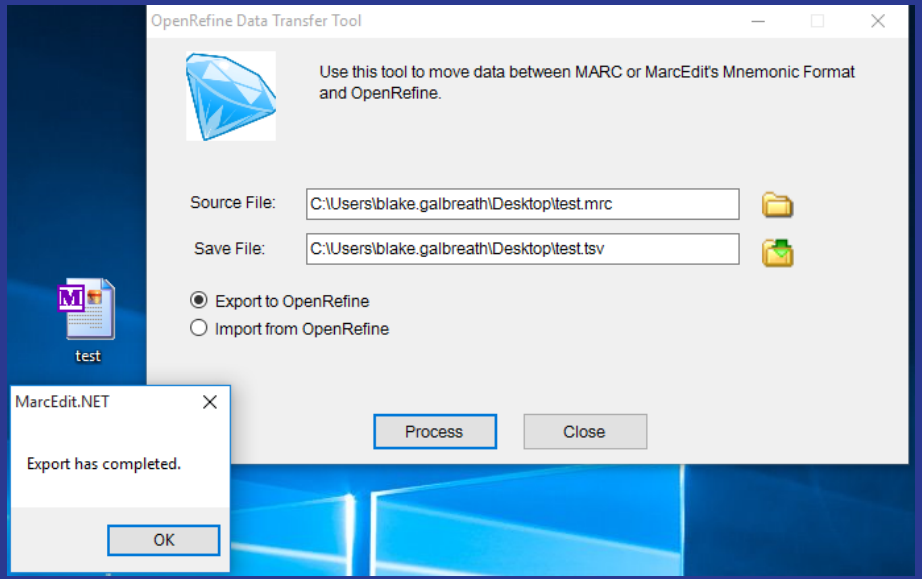
Project via Web Address

http://freegeoip.net/xml/onlinenorthwest.org

http://freegeoip.net/json/onlinenorthwest.org

MARC Files

1. Use MarcEdit to convert .mrc → .tsv



MARC Files

2. Import file into OpenRefine

3. Maintain column order:

   Record# | Tags | Indicators | Content

4. Export files as .tsv