

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](mailto:blake.galbreath@wsu.edu)

**Let us know how access to this document benefits you.**

Follow this and additional works at: <https://pdxscholar.library.pdx.edu/onlinenorthwest>

---

Galbreath, Blake, "Using OpenRefine to Standardize and Augment Your Data" (2017). *Online Northwest*. 8.  
<https://pdxscholar.library.pdx.edu/onlinenorthwest/2017/schedule/8>

This Presentation is brought to you for free and open access. It has been accepted for inclusion in Online Northwest by an authorized administrator of PDXScholar. For more information, please contact [pdxscholar@pdx.edu](mailto:pdxscholar@pdx.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](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

Custom text transform on column IMAGEFILE

Expression Language

```
value.replace("JPG", "jpg")
```

No s

**Preview** [History](#) [Starred](#) [Help](#)

row	value	value.replace("JPG", "jpg")
1.	010\20101001.JPG	010\20101001.jpg
2.	010\20101002.JPG	010\20101002.jpg
3.	010\20101003.JPG	01003.jpg
4.	010\20101004.JPG	4.jpg
5.	010\20101005.JPG	01005.jpg
6.	010\20101006.JPG	010\20101006.jpg
7.	010\20101007.JPG	010\20101007.jpg

On error  keep original  Re-transform up to  times until no change  
 set to blank  
 store error

OK Cancel



# Manual Transforms

Concatenate

**Add column based on column IMAGEFILE**

New column name

set to blank  store error  copy value from original

Expression  Language

**Preview** History Starred Help

row	value	"http://libraries.wsu.edu/" + value
1.	010/20101001.jpg	http://libraries.wsu.edu/010/20101001.jpg
2.	010/20101002.jpg	http://libraries.wsu.edu/010/20101002.jpg
3.	010/20101003.jpg	http://libraries.wsu.edu/010/20101003.jpg
4.	010/20101004.jpg	http://libraries.wsu.edu/010/20101004.jpg
5.	010/20101005.jpg	http://libraries.wsu.edu/010/20101005.jpg
6.	010/20101006.jpg	http://libraries.wsu.edu/010/20101006.jpg
7.	010/20101007.jpg	http://libraries.wsu.edu/010/20101007.jpg

OK Cancel



# Facets and Clusters

# Prepare data before 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 facetOR
  - b. Edit cells → Cluster

Place ; before  
capital letter, if  
leading space

Step 1

```
value.replace
```

```
(/(\b[A-Z])/,";$1")
```

and

```
.replace(";", ",;")
```

---

# Split data on newly-inserted semicolon

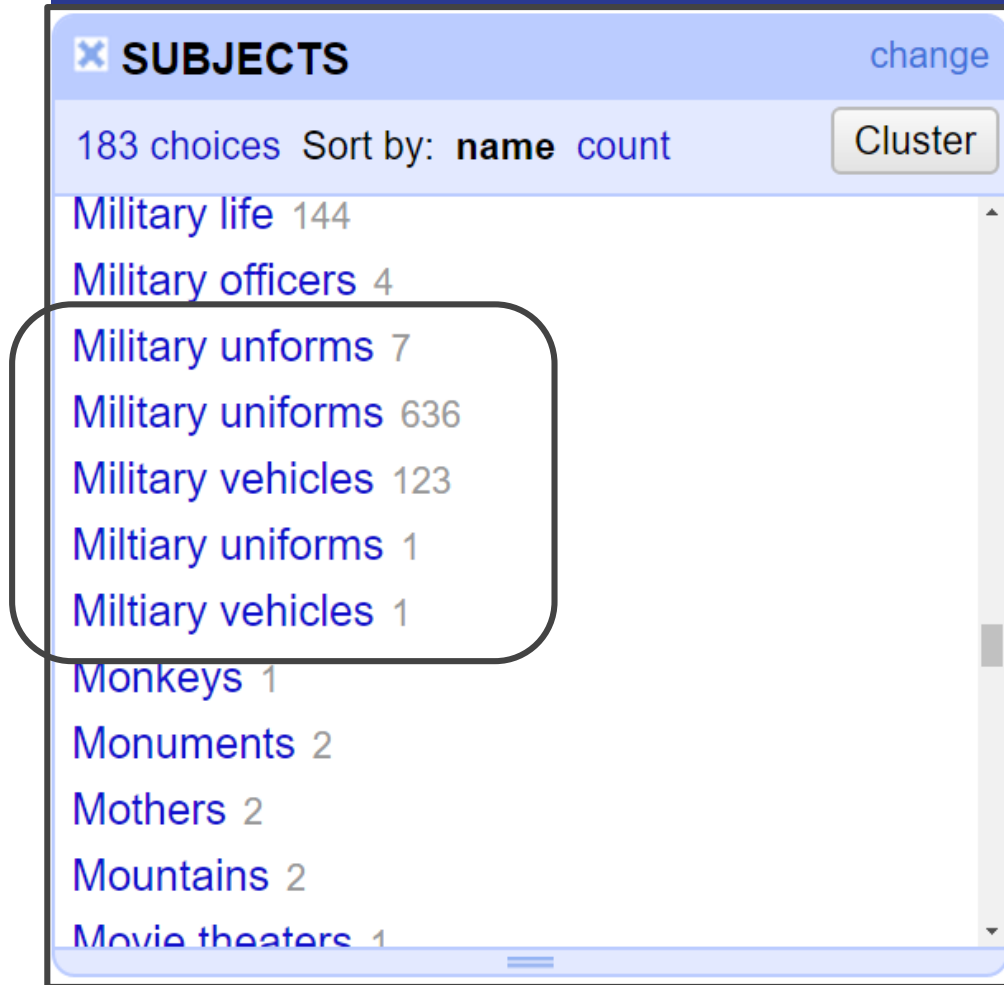
Step 2

The screenshot shows a data table with three columns: 'All', 'CLASSES', and 'SUBJECTS'. The 'CLASSES' column contains a list of categories: 'People;WorldWarII', 'Military life', 'Military uniforms', and 'Parties'. A context menu is open over the 'CLASSES' column, with the 'Split multi-valued cells...' option highlighted. The menu also includes options like 'Facet', 'Text filter', 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', 'Transform...', 'Common transforms', 'Fill down', 'Blank down', 'Join multi-valued cells...', and 'Cluster and edit...'. The table has two rows, with the first row containing a star icon, a speech bubble icon, and the number '1.', and the second row containing a star icon, a speech bubble icon, and the number '2.'.

All	CLASSES	SUBJECTS
★	1.	People;WorldWarII Military life Military uniforms Parties
★	2.	

# Facets

Step 3A



**SUBJECTS** change

183 choices Sort by: **name** count Cluster

Military life	144
Military officers	4
Military uniforms	7
Military uniforms	636
Military vehicles	123
Miltiary uniforms	1
Miltiary vehicles	1
Monkeys	1
Monuments	2
Mothers	2
Mountains	2
Movie theaters	1

# Clusters

Step 3B

The image shows a software interface with a menu open. The menu has several options, and the 'Cluster and edit...' option is highlighted with a blue background and a black border. The interface also shows a table with columns 'STERMS' and 'SUBJECTS'.

STERMS	SUBJECTS
Eastern Oregon State	Biology Boulders Cameras

Menu options:

- Facet
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort...
- View
- Reconcile
- Extract named entities...
- Cluster and edit...



# Clusters

Step 3B

Distance Function Radius 1.0  
levenshtein ▾ Block Chars 6

Row Count	Values in Cluster
	<ul style="list-style-type: none"><li>• <a href="#">Competition</a> (1 rows)</li><li>• <a href="#">Competitions</a> (1 rows)</li></ul>
	<ul style="list-style-type: none"><li>• <a href="#">Paintings</a> (3 rows)</li><li>• <a href="#">Painting</a> (2 rows)</li></ul>

Keying Function Ngram 2  
ngram-fingerprint ▾

Row Count	Values in Cluster
	<ul style="list-style-type: none"><li>• <a href="#">Wood work</a> (1 rows)</li><li>• <a href="#">Woodwork</a> (1 rows)</li></ul>
	<ul style="list-style-type: none"><li>• <a href="#">Metal work</a> (1 rows)</li><li>• <a href="#">Metalwork</a> (1 rows)</li></ul>

# Clusters

More please!

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

---



# Augment Data from Web Services

# Get Wikidata

Wikidata Reconciliation for  
OpenRefine (en)

## Add Standard Reconciliation Service

Enter the service's URL:

Add Service

Cancel

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

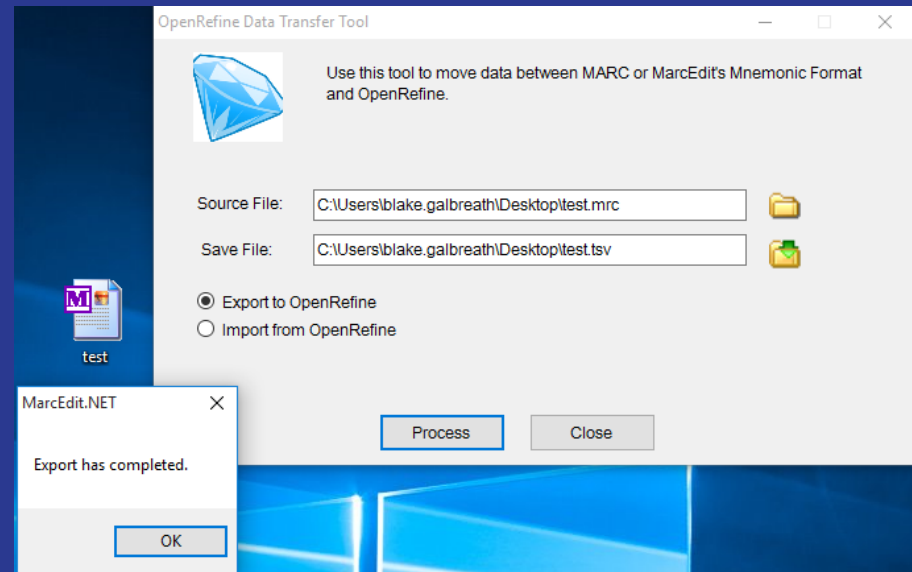
<http://freegeoip.net/xml/onlinenorthwest.org>

<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

---