

Clean Your Data!

getting started with OpenRefine

University of Idaho Library

Evan Peter Williamson
ewilliamson@uidaho.edu

2015-10-14

Objectives:

- Introduce OpenRefine
- Install OpenRefine
- Orientation to the interface
- Walkthrough some examples
- Free time for questions

What is OpenRefine?



<http://openrefine.org/>

Exciting trailers from Google!

Introduction: https://youtu.be/B70J_H_zAWM

Data Transformation: https://youtu.be/cO8NVCs_Ba0

Data Augmentation: <https://youtu.be/5tsyz3ibYzk>



Refiner
@RefineDude



Follow

Look what I'm learning about
@OpenRefine! #Awesome
#DoubleAwesome https://youtu.be/B70J_H_zAWM

← Reply ↻ Retweet ★ Favorite ⋮ More

12:41 PM - 14 Oct 15 · Embed this Tweet

Free,
Open source,
Extensible,
Java app,
runs in web
browser
(offline)

BL Flickr Images Book sub: x

127.0.0.1:3333/project?project=1461778412627

Google refine BL Flickr Images Book subset csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 0

8287 rows Extensions: Named-entity recognition Freebase RDF

Refresh Reset All Remove All

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

All	Identifier	Edition Statement	Place of Publication	Date of Publication	Publisher	Title	Author	Contributors	Corporate Author
1.	206		London	1879 [1878]	S. Tinsley & Co.	Walter Forbes. [A novel.] By A. A.	A. A.	FORBES, Walter.	
2.	216		London; Virtue & Yorston	1868	Virtue & Co.	All for Greed. [A novel. The dedication signed: A. A. A., i.e. Marie Pauline Rose, Baroness Blaze de Bury.]	A., A. A.	BLAZE DE BURY, Marie Pauline Rose - Baroness	
3.	edit 218		London	1869	Bradbury, Evans & Co.	Love the Avenger. By the author of "All for Greed." [The dedication signed: A. A. A., i.e. Marie Pauline Rose, Baroness Blaze de Bury.]	A., A. A.	BLAZE DE BURY, Marie Pauline Rose - Baroness	
4.	472		London	1851	James Darling	Welsh Sketches, chiefly ecclesiastical, to the close of the twelfth century. By the author of "Proposals for Christian Union" (E. S. A. [i.e. Ernest Appleyard])	A., E. S.	Appleyard, Ernest Silvanus.	
5.	480	A new edition, revised, etc.	London	1857	Wertheim & Macintosh	[The World in which I live, and my place in it. By E. S. A. [i.e. Letitia Willgoss Stone.] Edited by ... J. H. Broome.]	A., E. S.	BROOME, John Henry.	
6.	481	Fourth edition, revised, etc.	London	1875	William Macintosh	[The World in which I live, and my place	A., E. S.	BROOME, John Henry.	

Facet / Filter Issuance type change

2 choices Sort by: name count Cluster

continuing 19
monographic 8268
Facet by choice counts

Facet / Filter Publisher change

1989 choices Sort by: name count Cluster

- 152. D. H. Edwards 1
- 82. 1-298. E. P. for H. Seyle 1
- . 42. Nichols & Co. 1
- "Albion" Press 1
- "Art Journal" Office 1
- "Brighton Gazette" Printing Co. 1
- "Craven Herald" 1
- "Daily Herald" Office 1
- "Engineering" Office 1
- "Home Words" Publishing Office 1
- "Judy" Office 3
- "Mirror" 1

Facet / Filter Date of Publication change reset

1,540.00 — 1,920.00

Numeric 6528 Non-numeric 1578 Blank 181 Error 0

Tabular Data

Column

Cell

Row

	A	B	C	D	
1	key	attribute1	attribute2	attribute3	
2	record1	23	no	idaho	
3	record2	42	no	ID	
4	record4	42	yes	Idaho	
5					
6					

Import: TSV, CSV, custom separator txt, Excel, XML, JSON, Google Spreadsheets, RDF
From: file, archive (zip), URL, clipboard, or Google

“a power tool for working with messy data”

- more powerful than a spreadsheet
- more interactive and visual than scripting
- more provisional / exploratory / experimental / playful than a database

David Huynh, <http://www.davidhuynh.net/spaces/nicar2011/tutorial.pdf>

What is Messy Data?

2015-10-14	\$1,000	ID
10/14/2015	1000	I.D.
10/14/15	1,000	US-ID
Oct 14, 2015	1000 dollars	idaho
Wed, Oct 14th	US\$1000	Idaho,
42291	\$1k	Ihaho
“Using OpenRefine by Ruben Verborgh and Max De Wilde, September 2013”		

Refine Use Cases:

Clean - discover and fix inconsistency with faceting, clustering, cell transforms, GREL expressions...

Transform - change formats or reshape with split/join multi valued cells, split columns, transpose columns/rows...

Extend - enrich data by combining files, merging projects, fetching URLs, reconciliation with online databases...


Automate - reuse your processing routine by exporting operation history in JSON!

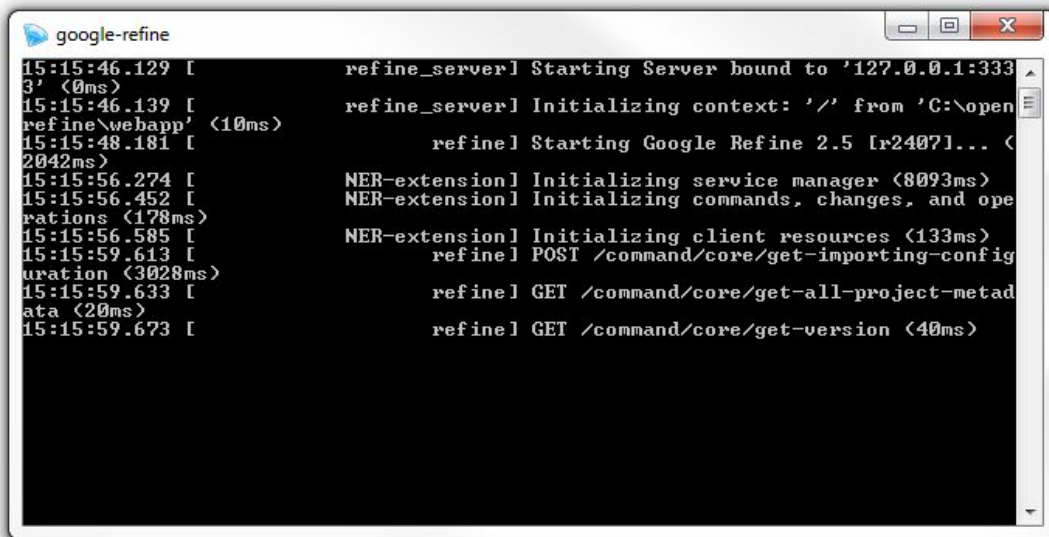
Install Refine:

1. Download package for your OS from <http://openrefine.org/download.html>
2. Windows: unzip the package to desired location (right click > *Extract All*).
Mac: drag dmg to application folder.
3. Do you have Java installed? <http://java.com/en/> (just remember to uncheck option to add Yahoo tools to your browsers when installing...)

<https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions>

Start Refine:

1. Windows: double click 'google-refine.exe' or 'openrefine.exe'. Mac: click Refine icon. 
2. The application should start running in a terminal window.
3. Your default web browser should automatically open <http://127.0.0.1:3333> to host the GUI.



```
google-refine
15:15:46.129 [ refine_server] Starting Server bound to '127.0.0.1:3333' <0ms>
15:15:46.139 [ refine_server] Initializing context: '/' from 'C:\openrefine\weapp' <10ms>
15:15:48.181 [ refine] Starting Google Refine 2.5 [r2407]... <2042ms>
15:15:56.274 [ NER-extension] Initializing service manager <8093ms>
15:15:56.452 [ NER-extension] Initializing commands, changes, and operations <178ms>
15:15:56.585 [ NER-extension] Initializing client resources <133ms>
15:15:59.613 [ refine] POST /command/core/get-importing-config?duration <3028ms>
15:15:59.633 [ refine] GET /command/core/get-all-project-metadata <20ms>
15:15:59.673 [ refine] GET /command/core/get-version <40ms>
```

To shut down: close browser window then close the host window with CTL+C.

Example: University Endowment data

In this demo we are going to play with a data set about University endowments harvested from Wikipedia—so it is very *messy*!

Download “universityData.csv” one of these links:

<http://emudrak.github.io/2015-01-15-cornell/data/biology/universityData.csv>

https://drive.google.com/file/d/0B_lwAVmhWjPGbWRPMDFiUUUzWIE/view?usp=sharing

1. Import data

2. Explore the interface

3. Facet & Cluster

```
value.unescape('url')
```

4. Text Filter & Sort & deduplicate

5. Find & replace Refine style!

```
value.replace("USD", "")
```

```
value.replace("US $", "").replace("US$", "")
```

```
value.replace("US $", "").replace("US$", "").replace("$", "")
```

```
value.contains("million")
```

```
toNumber(value.replace("million", ""))*1000000
```

Example: University Endowment data (2)

6. Dates

```
value.match(/.*(\d{4}).*/)[0]  
value.toString('yyyy')
```

7. Basic Geo lookup

```
"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")  
with(value.parseJson().results[0].geometry.location, pair, pair.lat + ", " + pair.lng)
```

8. Export

9. Undo (Export & Apply)

Based on tutorials from:

Enipedia, http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

DataCarpentry OpenRefine Demo <http://emudrak.github.io/2015-01-15-cornell/lessons/OpenRefine/open-refine-demo.html>

Resources:

OpenRefine Wiki, <https://github.com/OpenRefine/OpenRefine/wiki>

GREL reference, <https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language>

List of OpenRefine tutorials, <https://github.com/OpenRefine/OpenRefine/wiki/External-Resources>

OpenRefine Google Group, <http://groups.google.com/d/forum/openrefine>

Using OpenRefine, by Ruben Verborgh and Max De Wilde, (book available at UI library) <http://ida.lib.uidaho.edu:3500/lib/uidaho/detail.action?docID=10761194>

Tutorials & Examples:

Online course from Big Data University, <http://bigdatauniversity.com/bdu-wp/bdu-course/introduction-to-openrefine/>

Tutorial from Enipedia using university endowment data, http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

Tutorial from School of Data using African hospital data, <http://schoolofdata.org/handbook/recipes/cleaning-data-with-refine/>

Tutorial from Heard Library (Vanderbilt) using campaign donation data, <https://github.com/HeardLibrary/workshops/tree/master/OpenRefine>

Tutorial from ProPublica focusing on data journalism, <https://www.propublica.org/nerds/item/using-google-refine-for-data-cleaning>

Tutorial from Programming Historian focusing on history data, <http://programminghistorian.org/lessons/cleaning-data-with-openrefine>

Blog highlighting Biology data example, <https://practicaldatamanagement.wordpress.com/2014/05/16/help-me-im-covered-in-bees-or-using-openrefine-to-clean-specimen-data/>