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

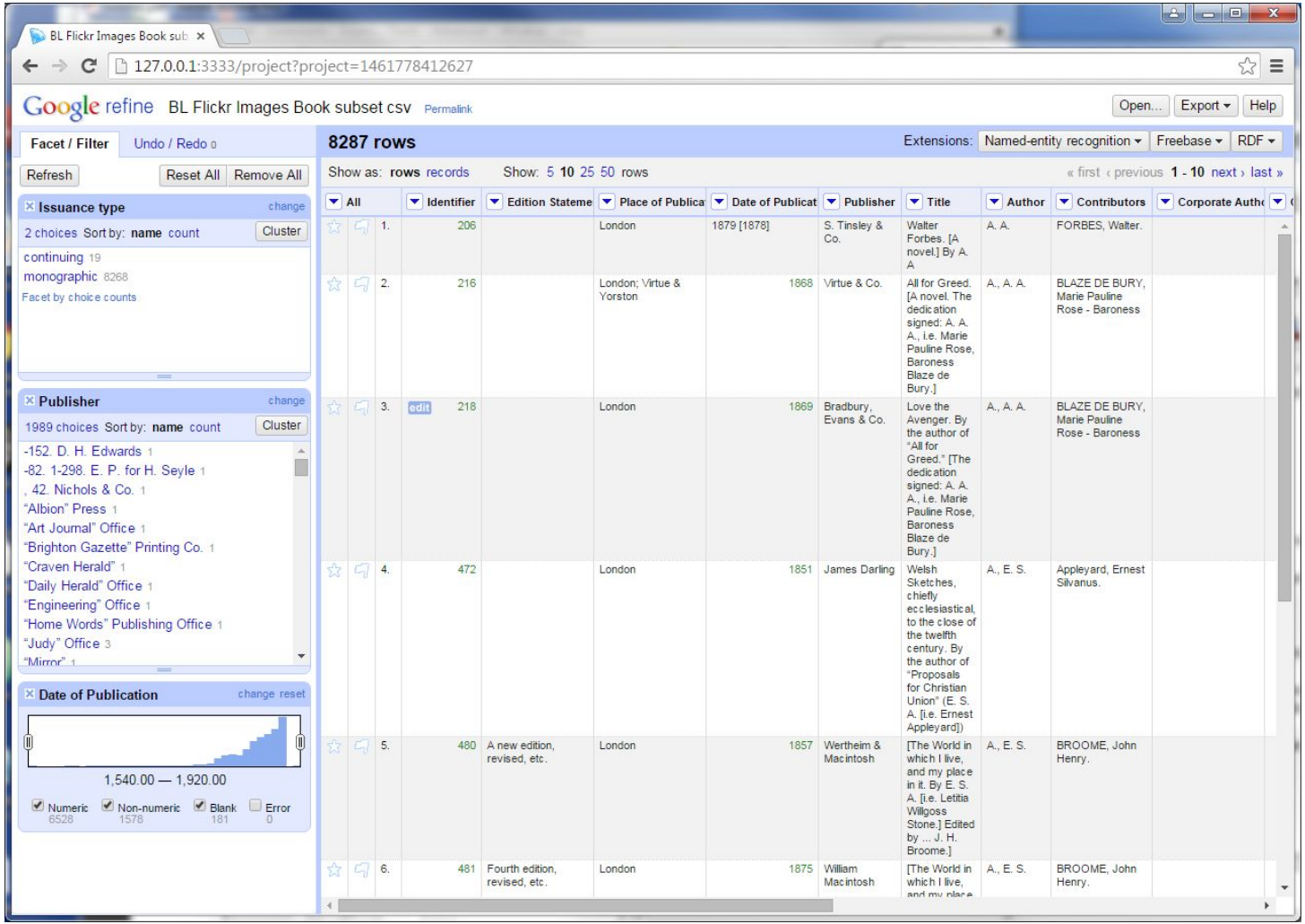
Free,

Open source,

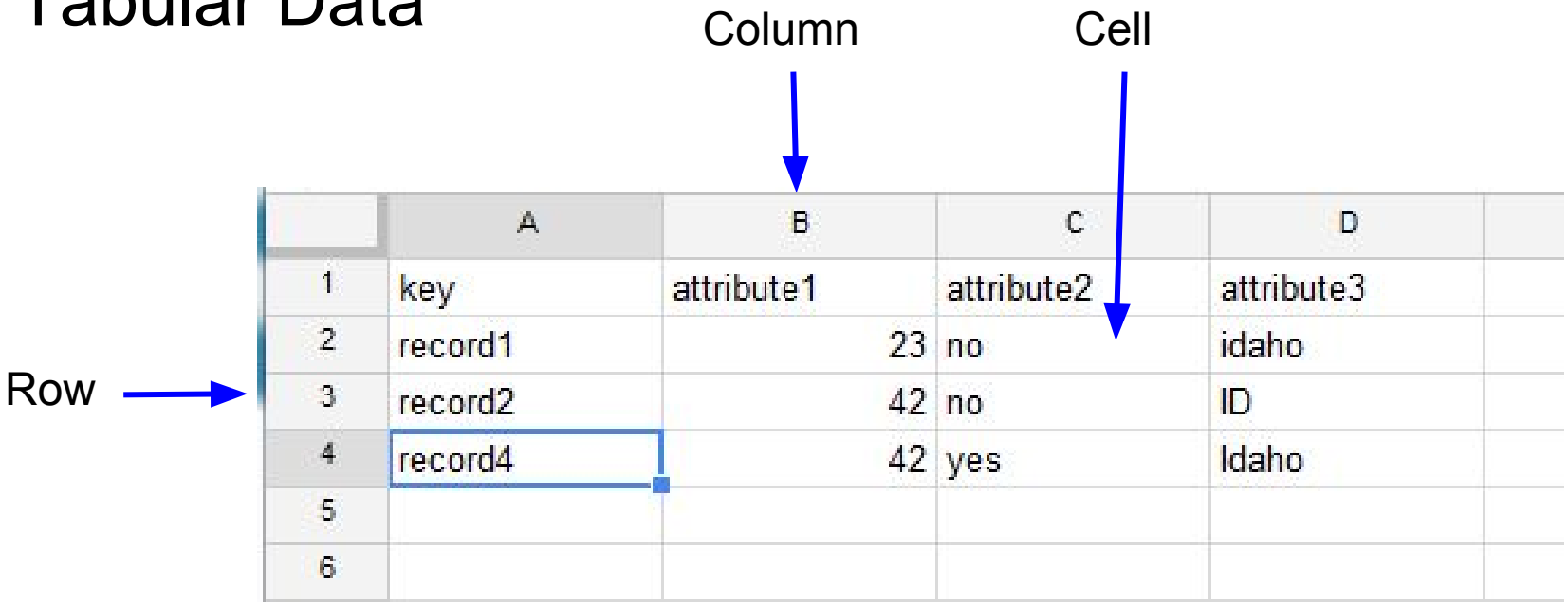
Extensible,

Java app,

runs in web browser (offline)



Tabular Data



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!

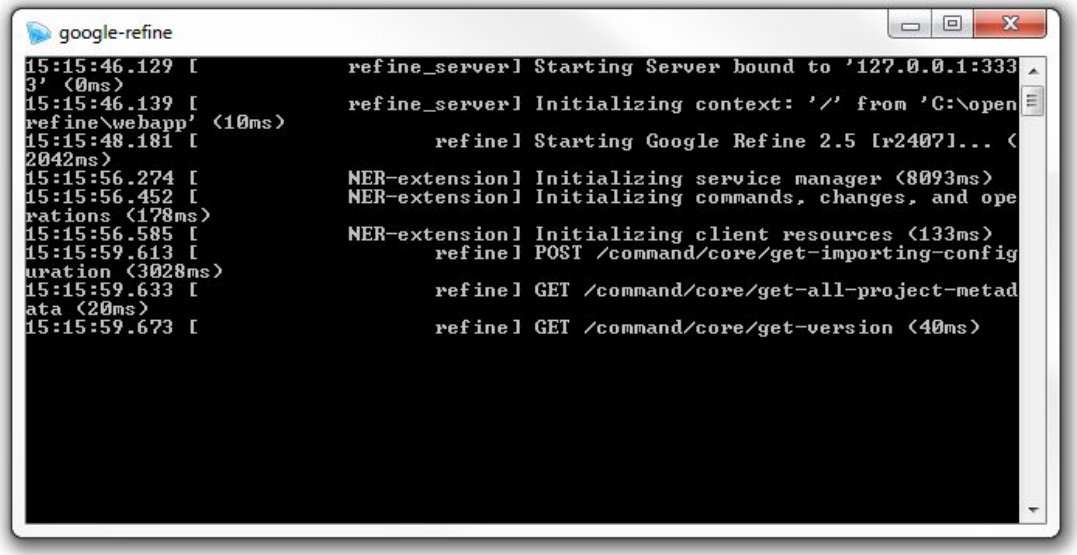
Install Refine:

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.



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_IwAVmhWjPGbWRPMDFiUUszWlE/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/