Participant Workbook

Viable Vocabularies:

An Introduction to Controlling, Cleaning, and Linking Your Data

Metadata & Discovery Unit

March 23, 2018

"Street Cleaning Sign" by "HelveticaFanatic"
Licensed under CC BY-SA 2.0. Accessed 2013-09-15. https://www.flickr.com/photos/helveticafanatic/2655427202

ULS/iSchool Digital Scholarship Workshop Series

This workbook: http://bit.ly/vv_workbook; Workshop slides: http://bit.ly/VV_slides

Instructors

Mike Bolam
Head, Metadata & Discovery Unit
Gabi Gulya
Metadata Librarian
Heajin Kim
Authority Control Specialist
Staci Ross
Visiting Cataloging/Metadata Librarian

Learning Objectives

- Participants will be able to define controlled vocabulary and identify its purpose in the research process
- Participants will be able to find appropriate vocabularies for their research needs.
- Participants will be able to select appropriate terms from a controlled vocabulary.
- Participants will be able to perform the basic tasks of cleaning and linking data using OpenRefine.

Additional Resources

Metadata & Controlled Vocabularies

- Metadata & Discovery @ Pitt Guide: https://pitt.libguides.com/metadatadiscovery
- Understanding Metadata: http://www.niso.org/publications/understanding-metadata-2017
- Introduction to Metadata: http://www.getty.edu/publications/intrometadata/

OpenRefine

- Wiki: https://github.com/OpenRefine/OpenRefine/wiki
- User Documentation: https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users
- Using OpenRefine [book – ebook available via PittCat]: https://www.packtpub.com/big-data-and-business-intelligence/using-openrefine
- GREL Resources: http://bit.ly/GREL_wiki

Linked Data

- Free Your Metadata: http://freeyourmetadata.org
- Linked Data for Libraries, Archives, and Museums [book – available at Hillman Library]: http://book.freeyourmetadata.org
- W3C Linked Data: https://www.w3.org/standards/semanticweb/data

Installing OpenRefine:

http://www.openrefine.org

- Direct link to the downloads: http://openrefine.org/download.html

Windows:

- Download the ZIP archive.
- Unzip & extract the contents of the archive to a folder of your choice.
- To launch OpenRefine, double-click on openrefine.exe.

Mac:

- Download the DMG file.
- Open the disk image & drag the OpenRefine icon into the Applications folder.
- Double-click on the icon to start OpenRefine.

Workshop Data

http://bit.ly/vv_data

- Created from the Powerhouse Museum metadata which been released under a CC-BY-SA Creative Commons Attribution Share Alike license.

OpenRefine URL

OpenRefine runs locally on your computer. Your data is not stored online or shared with anyone. It does not require an internet connection, unless you want to reconcile your data with external sources.

If you close your browser, you can get back OpenRefine by pointing it here:

- http://127.0.0.1:3333/ or http://localhost:3333

Reconciliation

The goal of reconciliation is to connect your collection-specific vocabulary to a controlled vocabulary on the Web. There are many options for doing reconciliation within OpenRefine. More details can be found here: https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation

Named Entity Extraction

Reconciliation comes in very handy when your data is already in a structured format. However, many fields contain unstructured text, yet they usually convey a high amount of interesting information. To capture this in machine-processable format,named entity recognition can be used.

Installation

1. Download the latest version of the extension and unzip it. (http://software.freeyourmetadata.org/ner-extension/)
2. Copy the unzipped folder to your extensions folder.

    - To find your extensions folder, choose “Browse workspace directory” from the Refine interface, and navigate to the folder extensions (which you should create if it doesn't exist yet).
3. Start or restart Refine.
4. Open or create a project.
5. Click the Named-entity recognition button, choose Configure API keys...and enter your personal API keys.

Basic GREL Expressions

Remove duplicate comma separated entries in a cell
value.split(', ').uniques().join(', ')
Replace string in cells
value.replace(“+”, ””)
value.replace(“~”, ””).replace(“,”, “”).replace(“-”, “”)
Clean-up character encoding problems
value.unescape(“url”)
Convert number with text to number
toNumber(value.replace(' million', ''))*1000000
Convert day/time to 4-digit year strings
value.replace(/\s+/,'').match(/.*(\d{4}).*/)[0]
value.toString(“yyyy”)
Geocoding
Add column by fetching URLs. ->
"http://maps.google.com/maps/api/geocode/json?sensor=false&address="+ escape(value, "url")
Expression -> with(value.parseJson().results[0].geometry.location pair,pair.lat +"," +pair.lng)