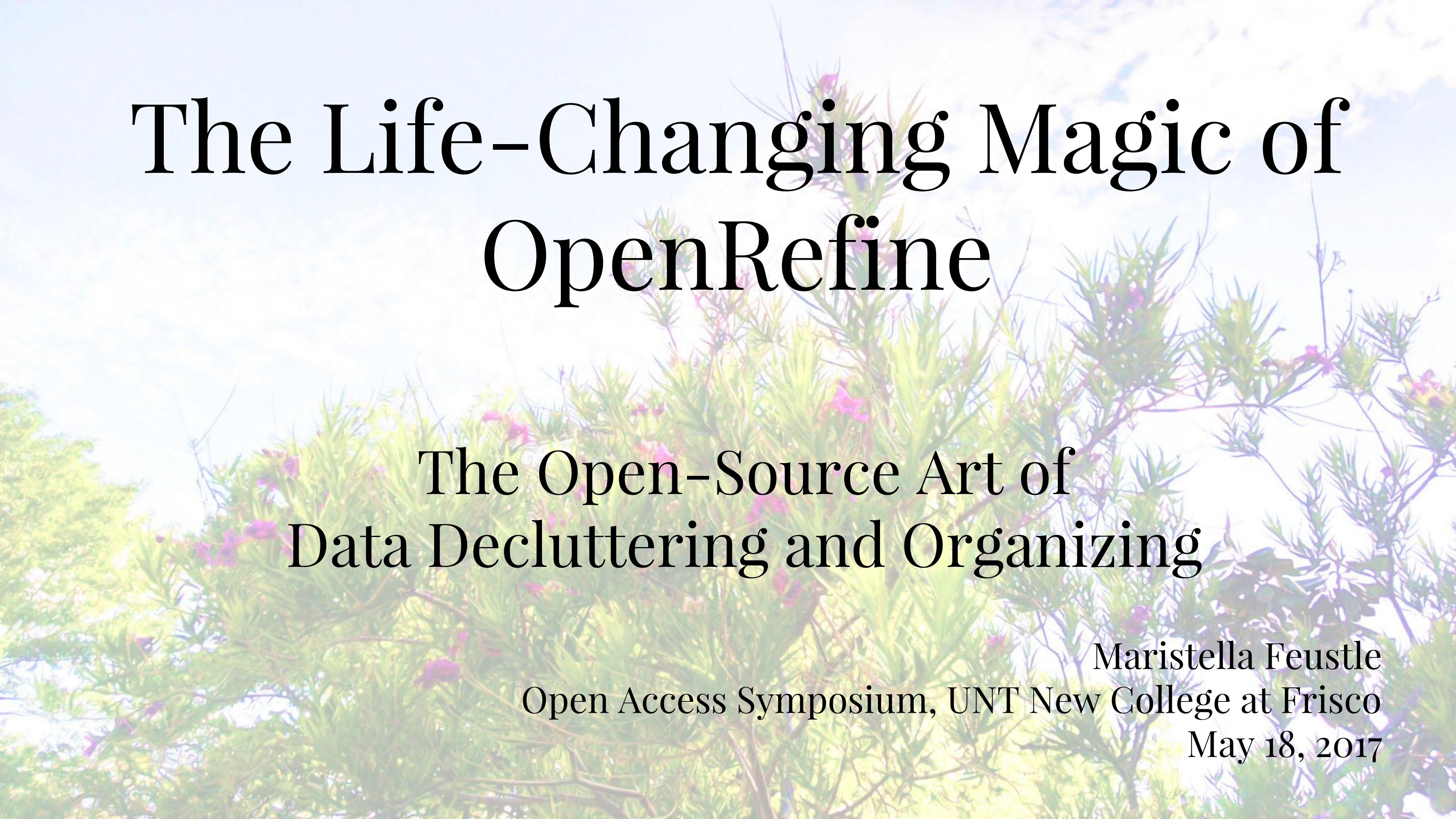


The Life-Changing Magic of OpenRefine

The Open-Source Art of Data Decluttering and Organizing

Maristella Feustle

Open Access Symposium, UNT New College at Frisco
May 18, 2017

Why OpenRefine?

Establishing meaningful relationships within and between datasets requires that all of the data are being read as intended.

Use OpenRefine to:

- Clean up data - standardize, correct, rearrange
- Automate tedious editing
- Match with outside controlled vocabularies
- Prep and export data for use in programs like MarcEdit, Tableau, Gephi, Raw, CartoDB, Google FusionTables, and more

Before and after

When you let your mom cut your hair and she tells you what a handsome young man you are

A brief history

Freebase Gridworks

Google Refine

OpenRefine

GREL - General Refine Expression Language

Many versions to choose from. Even “release candidate” versions are useful.

Creating a project

Grid-type data is most user-friendly, but OpenRefine can import numerous file types, including .zip files, and existing OpenRefine projects.

Data “lives” on your computer

Security pros and cons

Can be as simple as an Excel spreadsheet inventory, or even a text file

Memory issues with manipulations of very large files

Then what?

How you proceed depends on:

1. The final form you want your data to have
2. How much intervention your data requires to get there

Today’s data sets

https://goo.gl/8AQ31E

Project 1: Tabular data

Columns are the basis of organization in OpenRefine.

If your data is already in rows and columns (Excel, CSV, TSV, etc.), there is little translation to do in importing a project: “What you see is what you get.”

Check encoding for special characters.

Specify a first row as header, if applicable.

Basic transformations

Moving things around

Editing data

Demonstration: Column order, sorting, facets, clustering.

Project 2: Data as text

OpenRefine can do admirable work even within a single column, but can also render text into columns if there is a consistent pattern of rows

Achieving a consistent pattern may require some prep work in adding or editing line breaks, or your columns may “jog” as rows wrap around.

Demonstration:

1. Un-prepped data.
2. Prepped data.

Project 3: Data as XML

EAD: Encoded Archival Description - an XML framework of structured data for archival collections, including hierarchical levels of description.

We are interested in the hierarchical levels, so we select that point at which to import data.

Demonstration: Sort, join multi-valued cells, split and rename columns, reconcile to external controlled vocabulary

Fun with GREL

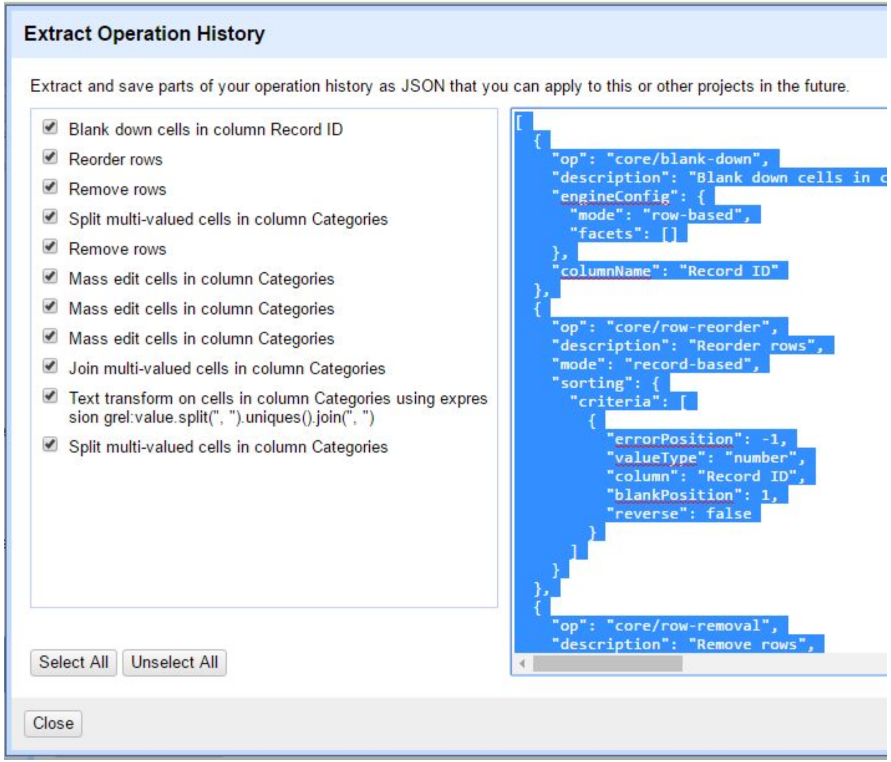
Many alternatives exist to writing code from scratch for common transformations:

https://data-lessons.github.io/library-openrefine/04-basic-functions-II/

https://github.com/OpenRefine/OpenRefine/wiki/Recipes

http://arcadiafalcone.net/GoogleRefineCheatSheets.pdf

Lather, rinse, repeat



Extract Operation History

Extract and save parts of your operation history as JSON that you can apply to this or other projects in the future.

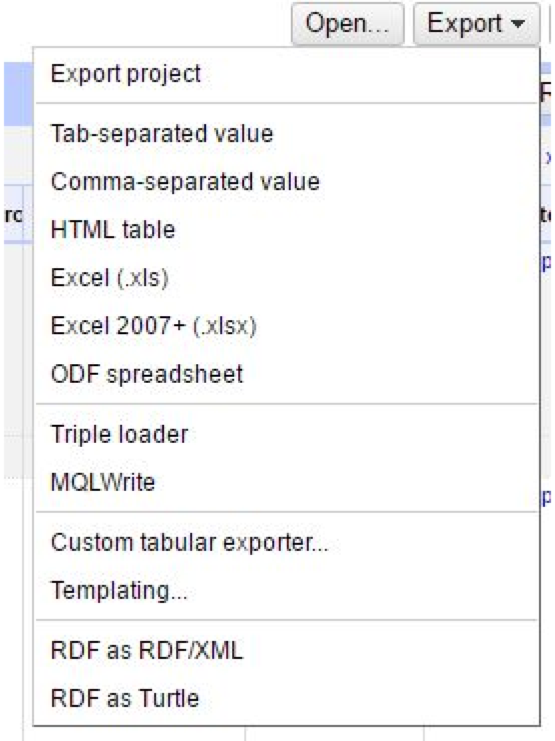
- ☑ Blank down cells in column Record ID
- ☑ Reorder rows
- ☑ Remove rows
- ☑ Split multi-valued cells in column Categories
- ☑ Remove rows
- ☑ Mass edit cells in column Categories
- ☑ Mass edit cells in column Categories
- ☑ Mass edit cells in column Categories
- ☑ Join multi-valued cells in column Categories
- ☑ Text transform on cells in column Categories using expression grel:value.split(", ").uniques().join(", ")
- ☑ Split multi-valued cells in column Categories

```
[
  {
    "op": "core/blank-down",
    "description": "Blank down cells in c
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "Record ID"
  },
  {
    "op": "core/row-reorder",
    "description": "Reorder rows",
    "mode": "record-based",
    "sorting": {
      "criteria": [
        {
          "errorPosition": -1,
          "valueType": "number",
          "column": "Record ID",
          "blankPosition": 1,
          "reverse": false
        }
      ]
    }
  },
  {
    "op": "core/row-removal",
    "description": "Remove rows",
```

Can extract command history to reuse on other datasets.

Export

CSV and other formats.



Other resources:

Terry Reese: MarcEdit and OpenRefine: http://blog.reeset.net/archives/1873

TxDH OpenRefine Webinar: https://www.youtube.com/watch?v=eJHd9n5xcLw

ALCTS, OR, and Metadata (case studies): https://www.youtube.com/watch?v=E-NbMR3_MRw

LSU, Data Wrangling with OR (case studies): https://www.youtube.com/watch?v=xtm6y8yB-Ho

OpenRefine.org, book and videos: http://openrefine.org/

Big Data University: https://bigdatauniversity.com/courses/introduction-to-openrefine/

OpenRefine Wiki: https://github.com/OpenRefine/OpenRefine/wiki

Metadata registry: http://metadataregistry.org/

List of extensions: http://openrefine.org/download.html

Reconcilable data sources: https://github.com/OpenRefine/OpenRefine/wiki/Reconcilable-Data-Sources

Finding datasets

Practice Your Data Mining: https://www.r-bloggers.com/datasets-to-practice-your-data-mining/

Data.gov: https://catalog.data.gov/dataset?tags=museum

Museum of Modern Art: https://github.com/MuseumofModernArt/collection

New York Public Library: What's on the Menu? http://menus.nypl.org/data

New York Public Library: NYC Space/Time Directory http://spacetime.nypl.org/#data

Free Your Metadata: http://freeyourmetadata.org/

DBPedia: http://wiki.dbpedia.org/

Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page

Thank you!

Maristella.Feustle@unt.edu