Data Transformation and Enrichment with OpenRefine: Part 1

Ryan Mendenhall

Columbia University Libraries

Oct. 3, 2018

Plan for today

- Background on OpenRefine and data transformation
- Create a project in OpenRefine
- Use faceting, clustering, and editing tools to visualize and transform “messy” data

Created by corpus delicti
from Noun Project

But first. . .
Introductions

Who has some experience with OpenRefine?

What do we mean by data clean-up?

Created by Nhor
from Noun Project

What is OpenRefine?

- History: Freebase Gridworks → GoogleRefine → OpenRefine
- Most recent version 3.0 (we will use v. 2.8)
- Runs in the browser at http://127.0.0.1:3333
- GREL: Google Refine Expression Language
- Formats: delimited text (CSV, TSV), Excel, XML, JSON, RDF, web services / APIs / endpoints
- Documentation: GitHub, Website, email list, books (see resource list)

Image by Google Inc. [BSD (http://opensource.org/licenses/bsd-license.php)], via Wikimedia Commons

Why use OpenRefine?

- Easily identify common errors in datasets
- Basic visualization of datasets (faceting and clustering)
- Built-in algorithms assist with identifying patterns
- Easily transform large datasets for more consistent output
- Match your data with external datasets (covered in part 2)
- Import external data into your dataset (covered in part 2)
- →Linked Data / Semantic Web

Scope of exercises

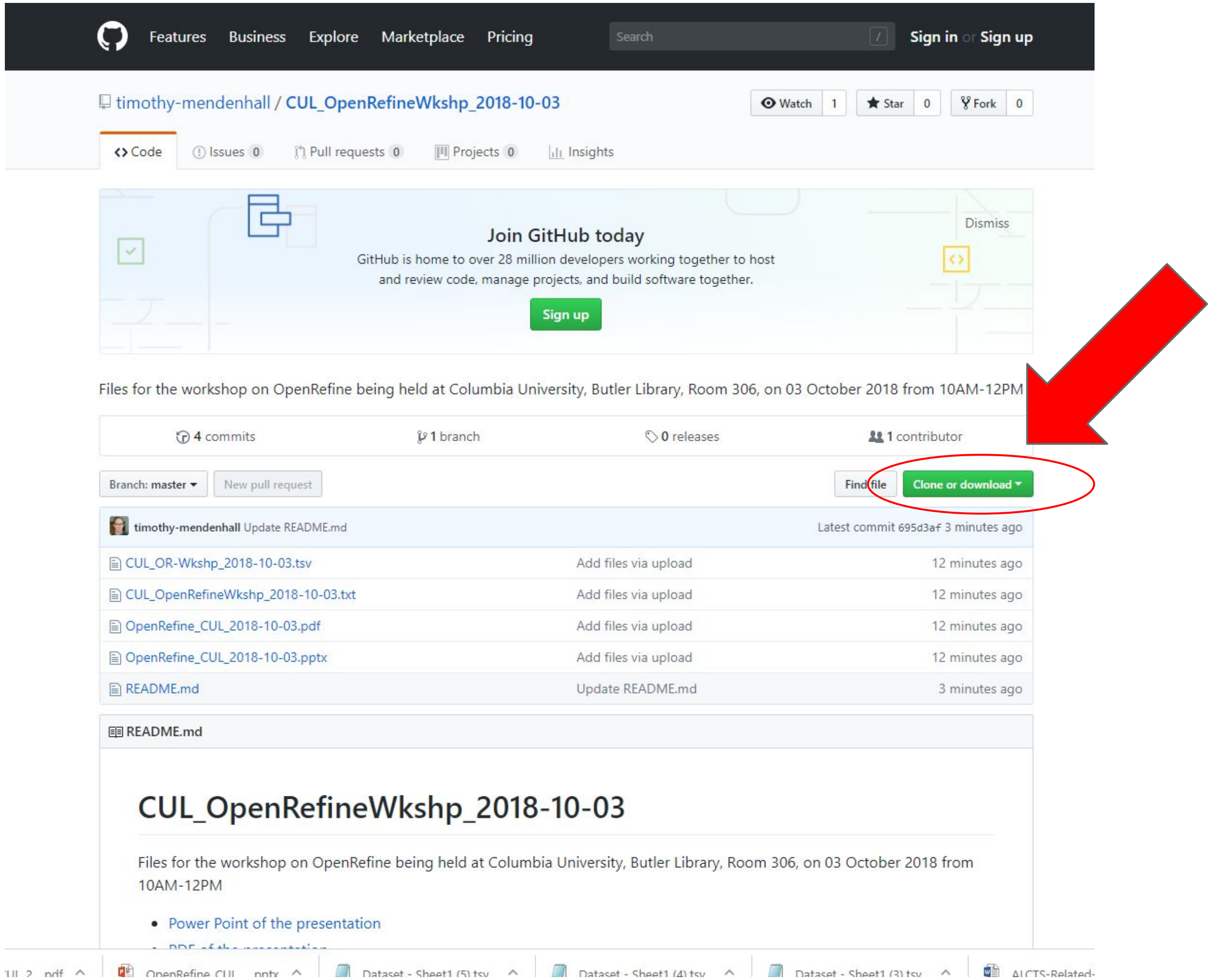
- Our data is adapted from an import spreadsheet created for RBML's Hubert H. Harrison Papers to upload content into CUL's Digital Library Collections (DLC)
- Import data and creating your first project
- Basic features for exploring and editing data: faceting and clustering
- Advanced clean-up using some lite GREL and regular expressions
- Exporting your project
- What else would you like to learn today?

Today's Workspace

Please download the workshop files from here:

https://bit.ly/2y6LVrx

Then right-click: Winzip → Extract to here



Join GitHub today

GitHub is home to over 28 million developers working together to host and review code, manage projects, and build software together.

Files for the workshop on OpenRefine being held at Columbia University, Butler Library, Room 306, on 03 October 2018 from 10AM-12PM

timothy-mendenhall Update README.md — Latest commit 695d3af 3 minutes ago

CUL_OR-Wkshp_2018-10-03.tsv	Add files via upload	12 minutes ago
CUL_OpenRefineWkshp_2018-10-03.txt	Add files via upload	12 minutes ago
OpenRefine_CUL_2018-10-03.pdf	Add files via upload	12 minutes ago
OpenRefine_CUL_2018-10-03.pptx	Add files via upload	12 minutes ago
README.md	Update README.md	3 minutes ago

README.md

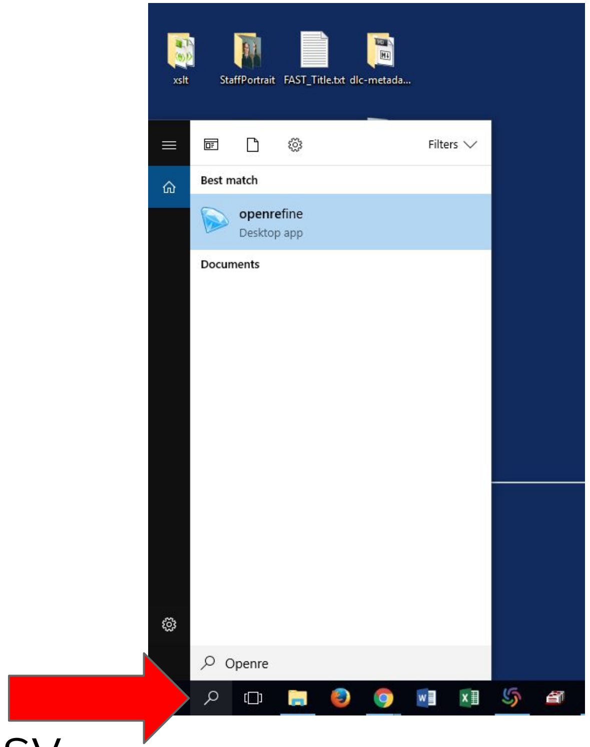
CUL_OpenRefineWkshp_2018-10-03

Files for the workshop on OpenRefine being held at Columbia University, Butler Library, Room 306, on 03 October 2018 from 10AM-12PM

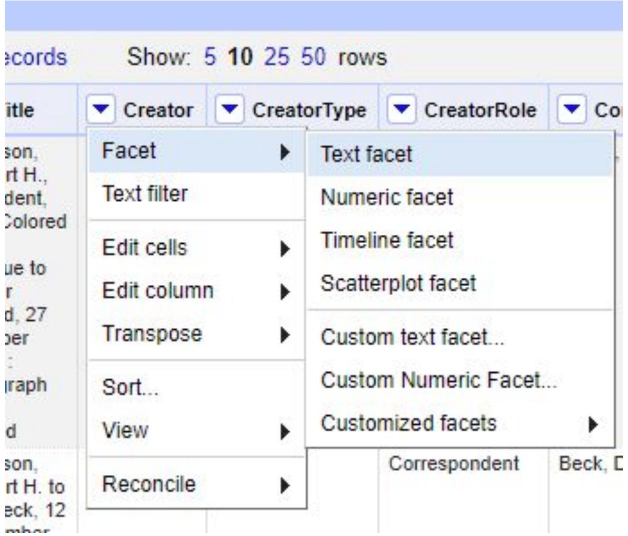
- Power Point of the presentation
- PDF of the presentation

Import data and create project

- Start the OpenRefine program
- If the browser window doesn't open: http://127.0.0.1:3333
- Import the dataset file CUL_OR-Wkshp_2018-10-03.tsv
- Why might we work with a "flat" data format like .tsv (delimited text)?



Visualizing and editing: faceting



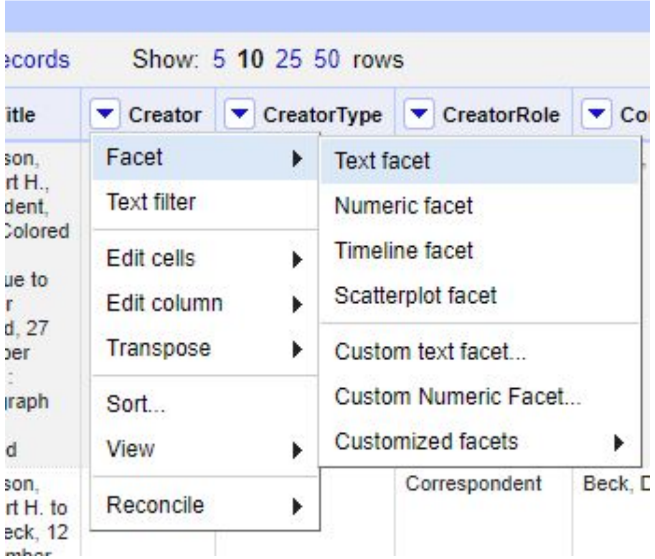
Faceting = OpenRefine's built-in feature to assist with:

- Grouping data
- Sorting data
- Batch-editing data

What happens with faceting?

- An sidebar opens with information on cells with identical values (value, cell count)
- Much like a traditional spreadsheet sort, but more actionable

Visualizing and editing: faceting



Types of facets

- Our focus today: text facet
- Custom facets using GREL and regular expressions (more on those later)
- Other useful facets: word facet, duplicates facet, blank facet, error facet

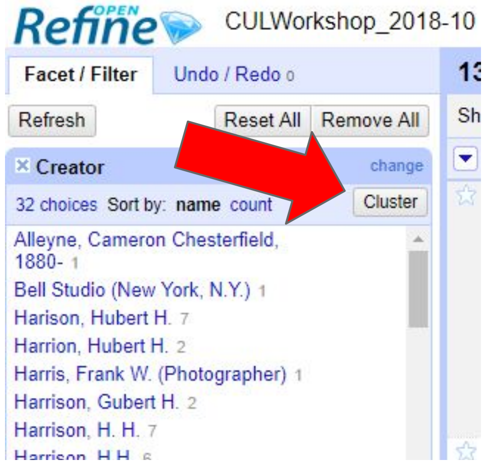
Exercise 1

1) Explore the facets function; try out at least the text facet, and one of the customized facets; explore the sort options in the facet box that appears at left; try some of the other features of the facet box (invert, reset, include, exclude)
2) Facets can also be cumulative--try faceting for all works by Creator “Harrison, Hubert H.” written in French

Visualizing and editing: faceting as filtering

Keep in mind the power of faceting as we progress through the advanced exercises later--Faceting can be a powerful way reconcile or to apply batch changes to subsets of your data, rather than the entire dataset

Visualizing and editing: Clustering



Alleyne, Cameron Chesterfield, 1880- 1
Bell Studio (New York, N.Y.) 1
Harison, Hubert H. 7
Harrion, Hubert H. 2
Harris, Frank W. (Photographer) 1
Harrison, Gubert H. 2
Harrison, H. H. 7
Harrison, H.H. 6

Clustering: takes faceting a step further by applying algorithms to your cell values to suggest matches

- Unlike faceting, can help identify near matches and errors, e.g. “Harrison” will be clustered with “Herrison,” “1877” with “1878” and “1879”
- Matching algorithms are grouped by method; we won’t go into the details here
- Visualizations can help target your editing efforts
- Human review is a key component of the clustering process

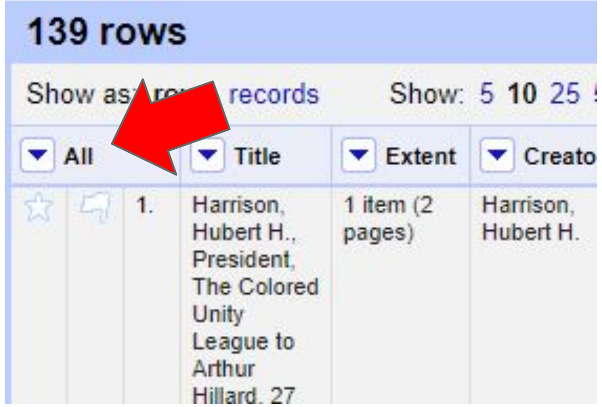
Exercise 2

Using the OriginPlace, play around with some of the faceting and clustering features. Try out different algorithms for clustering. What kinds of matches and clusters are you getting? Which matching algorithms work better?

Exercise 3

Using the OriginPlace column, begin making batch edits using the faceting and clustering features

The All column



			All	Title	Extent	Creato
☆		1.		Harrison, Hubert H., President, The Colored Unity League to Arthur Hillard, 27	1 item (2 pages)	Harrison, Hubert H.

At left, the All column hides some useful features:

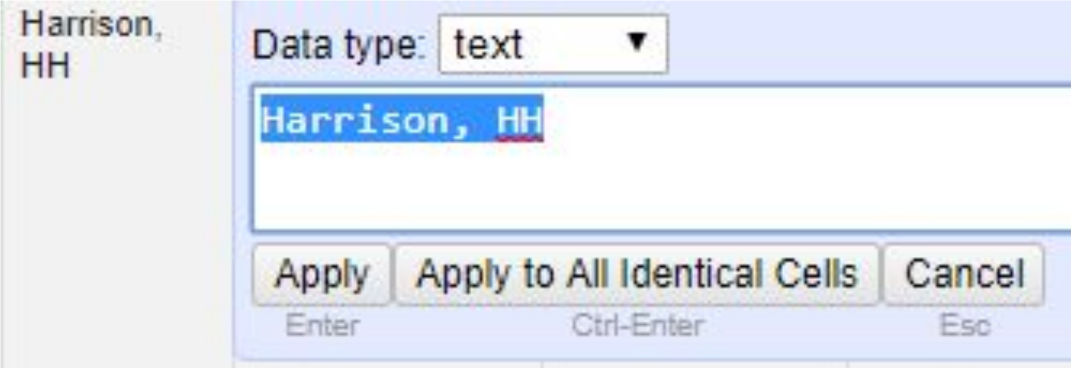
- DELETING via starring and flagging
  - Edit rows→ Star/Flag rows THEN Edit rows→ Remove all matching rows
  - Stars or flags can also be used to make facets
- REORDERING COLUMNS
  - Edit columns→ Re-order/remove columns
  - This can be easier than the “Move left” and “Move right” options available on other columns, especially in large datasets

Exercise 4

Try using the All column to remove all rows with materials of a specific Form, like correspondence, photographs, or ephemera--how could we do this most efficiently?

Try moving around the columns to make your work a bit easier

Editing cells


Harrison, HH

Data type: text

Harrison, HH

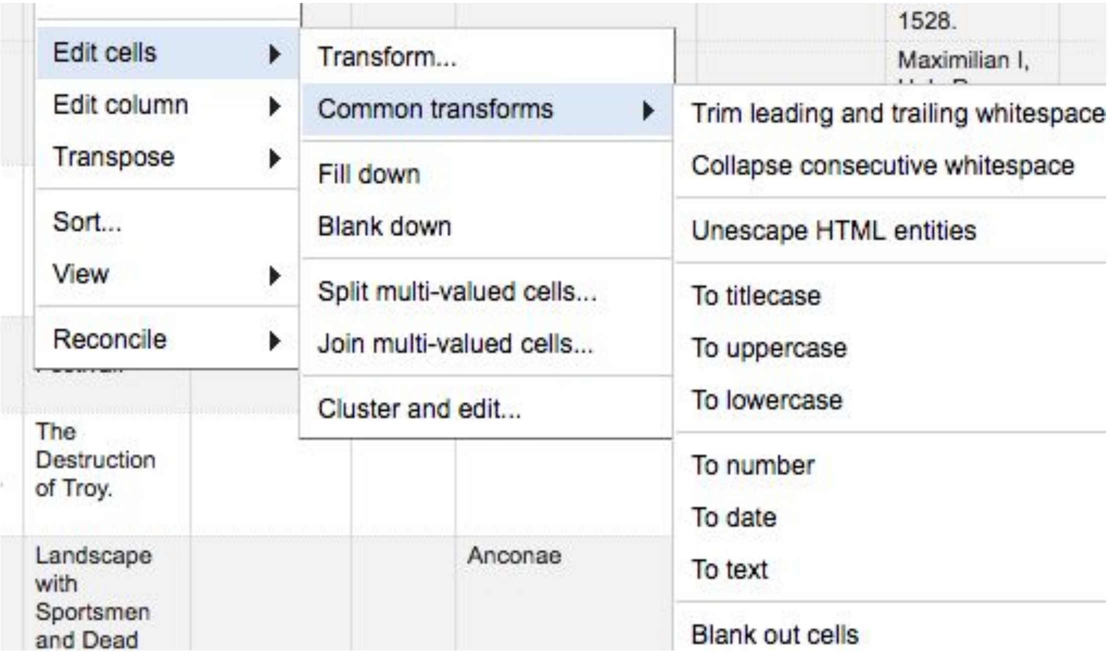
Counterintuitively, let’s now move into more granular changes. . .

Click in any cell--what options are available to change the cell content?

Exercise 5

Clean up some of the cells in the Type column--for example, click in the top cell, click “Edit,” and change the upper-case T to lower-case: “Text” → “text”. Click “apply to all identical cells”

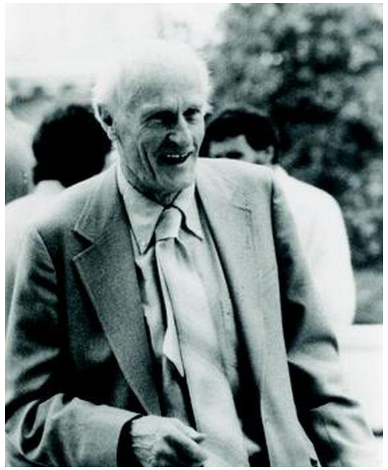
Editing cells



Common editing tasks:
Edit cells → Common transforms

- Explore some of these tasks. . . why are these built-in to OpenRefine?

Editing cells



Stephen Cole Kleene,
inventor of resexes

Regular expressions

- Regular expressions (regexes): a syntax for complex find-and-replace
- Examples:
  - □ \.$ matches all periods at the end of a line
  - □ \d{3}\-\d{4} matches all telephone numbers in format ###-#### e.g. 555-5456
- GREL Syntax: value.replace(/{regex}/, "{replacement}")
- Documentation on basic regexes in OpenRefine

Exercise 6

You notice that all of the values in the Contributor column have a period at the end--this period should be removed from all the cells.

Using Edit cells → Transform -- can you deploy a regular expression (regex) to remove all periods at the end of the Contributor column? (Hint on previous slide)

Exercise 6

Using Edit cells → Transform -- can you deploy a regular expression (regex) to remove all periods at the end of a column?

ANSWER: value.replace(/\.$/, "")

Named groups in regexes

Named groups: a way to refer to patterns in your data and use it in replace statements

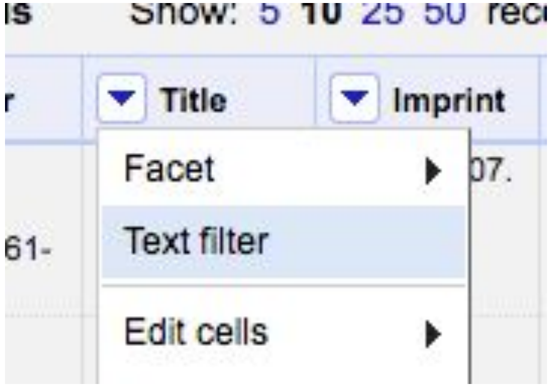
- Example: you need to normalize a date column: all data was entered MM/DD/YYYY, but you need it in the ISO standard format of YYYY-MM-DD
- 10/27/1925 → 1925-10-27
- Find: (\d{2})/(\d{2})/(\d{4})
- Replace: $3-$1-$2

Exercise 7

Using the Edit Cells→ Transform dialogue, clean up the Date1Code column. Convert the date to the standardized YYYY-MM-DD format

Where else might we use named groups to clean our data?

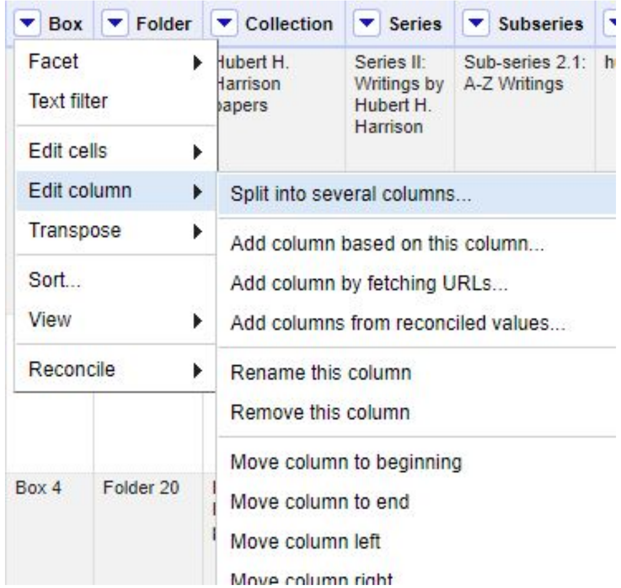
Custom filters and facets



Now that we've used some GREL and regular expressions, you have the basic tools to build your own text filters and custom facets

- Facet→ Custom text facet OR Custom numeric facet
- Text filter

Splitting and joining



Columns and cells can be split or joined based on:

- Delimiters (such as , ; |)
- Fixed lengths (split or join at the nth position)
- GREL and regexes

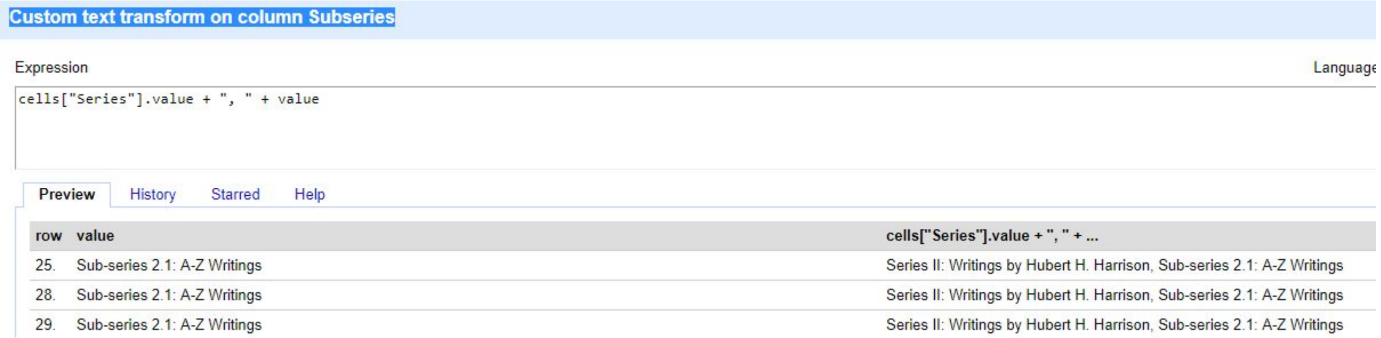
While splitting columns based on a delimiter is pretty self-explanatory, splitting and joining cells, and joining columns is a bit trickier. . . Let's start with splitting columns. . .

Exercise 8

Edit column→ Split into several columns

1. What data might we want to split into multiple columns or cells/rows?
2. See if you can split BoxFolder into two columns (“Box” “Folder”) based on the delimiter “, “

Joining column data



Custom text transform on column Subseries

Expression

Language

```
cells["Series"].value + ", " + value
```

row	value	cells["Series"].value + ", " + ...
25.	Sub-series 2.1: A-Z Writings	Series II: Writings by Hubert H. Harrison, Sub-series 2.1: A-Z Writings
28.	Sub-series 2.1: A-Z Writings	Series II: Writings by Hubert H. Harrison, Sub-series 2.1: A-Z Writings
29.	Sub-series 2.1: A-Z Writings	Series II: Writings by Hubert H. Harrison, Sub-series 2.1: A-Z Writings

Joining column data:

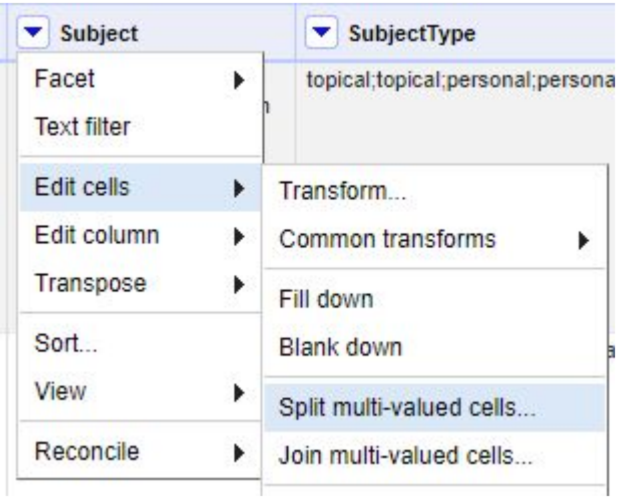
1. Edit Cells→ Transform
2. Insert the column name and delimiter (like a space or semicolon) into this GREL script: **cells[“COLUMN_NAME”].value + “DELIMITER” + value**
3. GREL will generate an error if your columns are not well-named, e.g. if the column name has a space or non-alphanumeric character

Exercise 9

Join some column data!

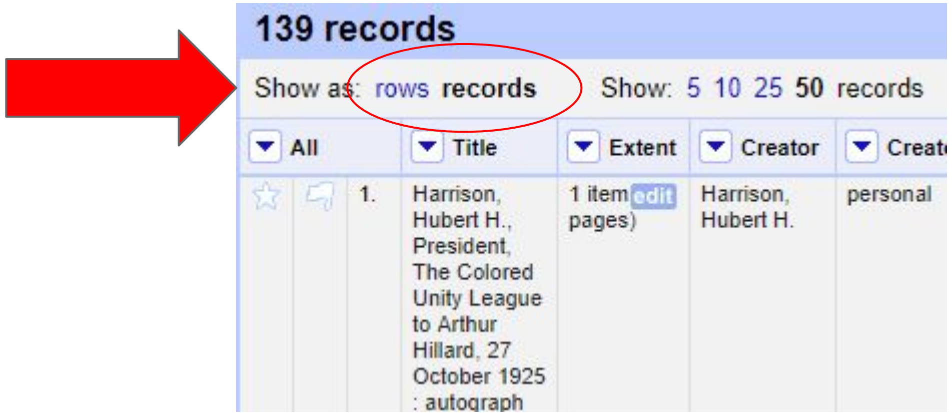
What are some cases where you might want to join data from various columns?

Splitting cells: Rows vs. Records



Look at the Subjects column--there are multiple values in many of the cells

- Why might we want to look at each value separately?
- Splitting cell values brings into relief a powerful feature of OpenRefine: viewing data as Rows versus viewing data as Records



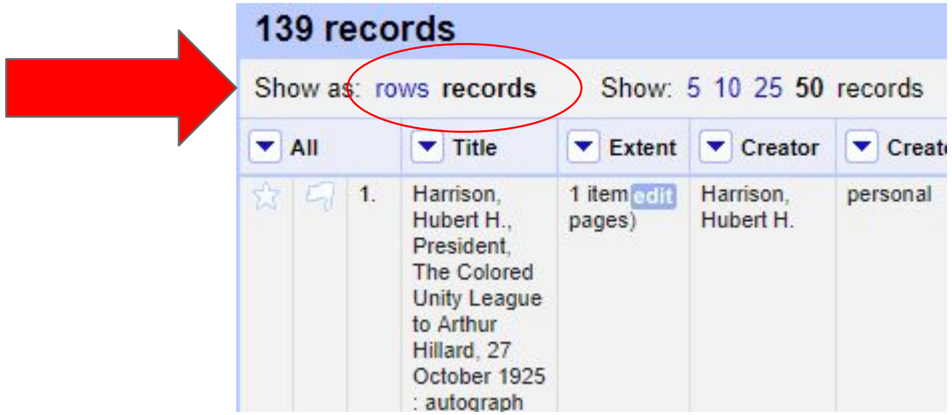
139 records

Show as: rows records   Show: 5 10 25 50 records

All			Title	Extent	Creator	Creat
☆		1.	Harrison, Hubert H., President, The Colored Unity League to Arthur Hillard, 27 October 1925 : autograph	1 item edit pages)	Harrison, Hubert H.	personal

Exercise 10

1) Using the Subject column, click Edit cells→ Split multivalued cells. Split the cells on the delimiter ;
2) After splitting the cell data, toggle between the “Show as” options at the top left of your spreadsheet: Show as rows, Show as records. **How are these views different**?
3) In record view, do some cleanup with faceting and clustering
4) Then, click Edit cells→ Join multivalued cells. Use a semicolon as a delimiter, as before



139 records

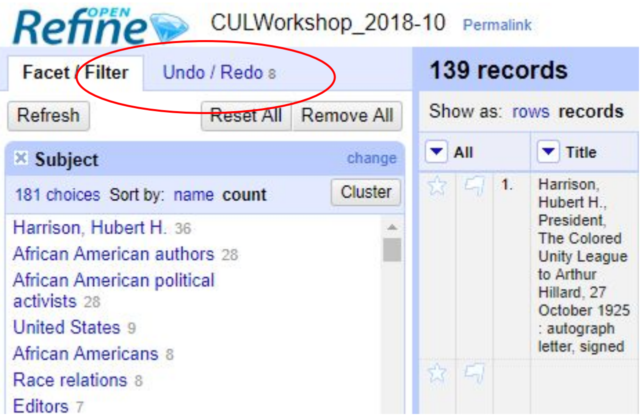
Show as: rows records   Show: 5 10 25 50 records

All		Title	Extent	Creator	Creat
☆	1.	Harrison, Hubert H., President, The Colored Unity League to Arthur Hillard, 27 October 1925 : autograph	1 item edit pages)	Harrison, Hubert H.	personal

Splitting cells

What are some advantages and disadvantages to using the "Split/Join multivalued cells" functions for editing (and reconciling) our data?

Undo/Redo …and reusing your work!



On the top left, there is an undo/redo panel (you may have missed it!)

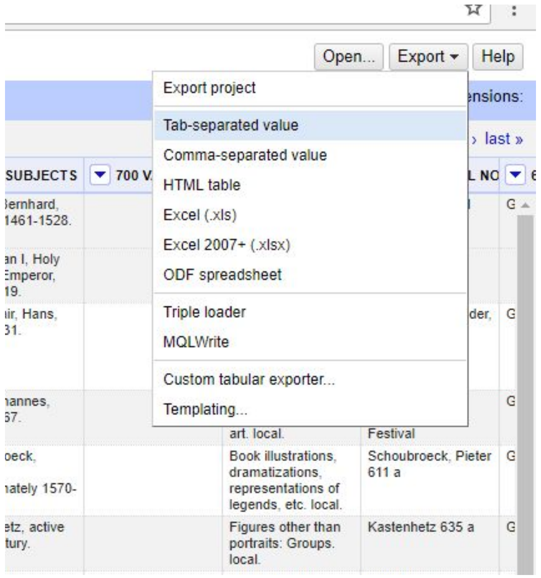
- Essentially, this functions like “Version history” in a Google Sheet or “Track Changes” in Microsoft
- If you make some disastrous batch edits or some scripts go awry, you can always back-up and recover
- Just click on the point in the editing history that you’d like to revert to, and return to editing!

Undo/Redo ...and reusing your work!

Extract/Apply

- Click on the Extract button: a snippet of JSON code appears
- This code can be copied as a kind of batch script that you can re-apply to other datasets (via the Apply button)
- Note that you can also select and de-select steps to include in this code
- How might you use this powerful functionality?

Exporting your data



To get your data out of OpenRefine:

- Click Export (at the top right of screen)
- Pick a data format
- The file will appear in your default downloads folder

Exporting a project

- Export → Export project
- This exports both your data and the entire editing history since you loaded it into OpenRefine
- Can be opened by another instance of OpenRefine

Resources

- http://openrefine.org/
    - *Homepage has basic tutorials, Documentation has links to mailing lists, wikis, and Downloads has links to the GitHub page*
- Verborgh, Ruben and De Wilde, Max. Using OpenRefine. Birmingham : Packt Publishing, 2013. 9781-78328-908-0
    - *A good introduction and reference to using OpenRefine; although based on an older version, most editing features are the same*
    - Especially useful:
        - Chapter 1 Recipe 3
        - Chapter 2 Recipe 2
        - Chapter 3 Recipe 3
- van Hooland, Seth and Verborgh, Ruben. Linked data for libraries, archives and museums : how to clean, link and publish your metadata. London : Facet Publishing, 2014. 9781856049641
    - *A more in-depth exploration of the linked data world, with many examples of how to use OpenRefine*
- Karen Hwang's post for METRO about using OpenRefine

Resources ctd.

- Rawson, Katie and Muñoz, Trevor. “Against cleaning” at *Curating Menus*, referenced 15 May 2018, http://curatingmenus.org/articles/against-cleaning/
  - □ *An exploration of what we’re talking about when we talk about “data cleaning”*

Regular expressions and GREL

- “General Refine Expression Language” at OpenRefine Wiki, referenced 10 May 2018, https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language
  - □ *The primary reference source for GREL. Fairly user-friendly, richly illustrated with examples*
- “Understanding regular expressions” at OpenRefine Wiki, referenced 10 May 2018 https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Regular-Expressions
  - □ *A good introduction to using regular expressions in GREL scripts in OpenRefine*
- https://www.regular-expressions.info/
  - □ *Despite the wonky look, this is an excellent reference for all things Regular Expressions*
- https://regex101.com/
  - □ *Another option for testing your regex*

Questions?

I hope to see you back here next week for Part 2: reconciliation and enrichment!

Ryan Mendenhall

trm2151@columbia.edu