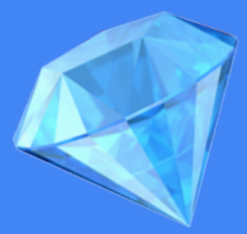
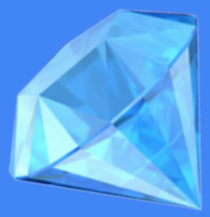
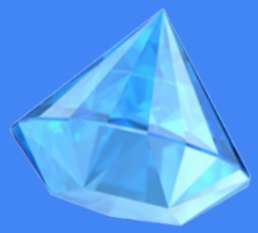
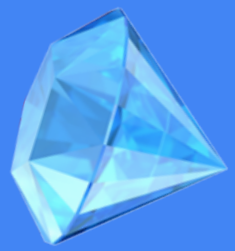




Cleaning Data with OpenRefine

presented by **Alissa McCulloch** (@lissertations)
+ assisted by **Alexis Tindall** (@lexistindall)

VALA Tech Camp 2019, Melbourne VIC

Check your tech

1) Ensure OpenRefine is installed (**openrefine.org**) + install Java if necessary

2) Open the OpenRefine app

3) A terminal window will open! This is normal! Don't touch it

4) OpenRefine should open automatically in your browser. If it doesn't, open a new tab and go to **http://127.0.0.1:3333/**

Why am I here?

- To find out more about OpenRefine
- To get a taste for what OpenRefine can do
- To start thinking about how OpenRefine might be useful for me and my work
- To learn from peers, not experts 🙂

‘If you know something, you can teach something’

What will I learn today?

- Import and export data in and out of OpenRefine
- Facet, filter, cluster and edit data
- Transform data using GREL (General Refine Expression Language)
- Reconcile data against an external source (VIAF)
- Two kinds of datasets: CSV and MARC

Getting data into OpenRefine

- Accepts TSV, CSV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, Google Data
- Can also scrape from web or Google Sheets
- Claims to accept binary MARC files
  - It doesn't, really
  - Converts them to MARCXML instead
  - Use MarcEdit to convert MARC files to TSV files
  - OpenRefine will parse these as TSV, not MARC format

Getting to know your data

- Macro-level editing, not micro-level (not Excel)
- One **record** can have multiple **rows**
- Column-based program 🏛️
  - Menu functions hidden under the triangle
- Complete edit history / undo feature (amazing)
  - Exportable, for use with other projects!
  - Make sure your facets are set first, though

Facets and filters

- These help narrow down + distil your data 
- Also let you perform simple bulk edits
- **Facets** group all the values in a column and let you filter the data by those values
  - Text, numeric, timeline, custom facets
- **Text filters** work like a search box

Clustering and editing

- Harness the power of the algorithm 🤖
- **Clustering** brings together similar but inconsistent values
- You can merge them into a value of your choice
  - Makes your data cleaner and more consistent
- To cluster multi-valued cells, split them first
- Different algorithms give different results
  - Risk of false positives! Clusters are a guide only!

Transformations

- Powerful way of manipulating data in columns ⚡
- Some common transformations are in menus
  - Upper/lowercase changes
  - Trim whitespace
  - Standardise date formats
- Other transformations involve writing code 😮

Transformations… with GREL!

- **G**eneral **R**efine **E**xpression **L**anguage
- Designed to resemble JavaScript
  - `value.function(options)`
  - `function(value, options)`
- e.g. converting dates to a readable format
  - `value.toString("dd MMMM yyyy")`
  - = ‘Convert all values in this column to a string (in the format I specify)’
- Can preview a transformation + save it for later

More transformations

- Can use regular expressions (regex)
  - `value.replace(/\s+/,'')`
  - ‘Replace whitespace with nothing’ = ‘Delete whitespace’
- Can combine GREL scripts 🔀
  - `if(value.contains("test"),"Test data",value)`
  - If (and only if) the cell value contains the string ‘test’ anywhere in it, replace the entire cell with ‘Test data’

Reconciliation + API lookup

- Lookup + reconcile against external values
  - Compare, match or enhance data
  - Import text, numbers, URLs, unique identifiers
  - A bit like creating linked data (ish) = like a hyperlink
- It will only reconcile plain text, not MARC data
  - Super annoying 😒
  - Strip out your MARC subfields before you start
  - There's a recipe in your project guide
- Reconciliation results are a guide only!

Getting data out of OpenRefine

- Export data to TSV, CSV, Excel (.xls and .xlsx), ODS, HTML table, Google Sheets, Wikidata (!)
- Recommend using **Custom tabular exporter**
- Can also export project (.tar.gz) with edit history
  - Exporting ‘data’ and ‘project’ are not the same thing ❎

Congratulations!
You're all wizards 🔮

What have I learned today?

- Import and export data in and out of OpenRefine
- Facet, filter, cluster and edit data
- Transform data using GREL (General Refine Expression Language)
- Reconcile data against an external source (VIAF)
- Two kinds of datasets: CSV and MARC

Acknowledgements

Some text and exercises adapted from Library Carpentry OpenRefine lesson (CC-BY 4.0 licensed)

MARC dataset courtesy Terry Reese

VIAF reconciliation target written by Jeff Chiu

This workshop based on loads of things I learned at work (thanks, work!)

Work some magic!

1) Open your project guide: **lissertations.github.io/openrefine**

2) Choose your dataset: **MARC or CSV Spreadsheet**

3) Work your way through the suggested exercises

4) If you need help, please ask! Alissa's contact details are in the project guide 📩