Data Wrangling with Google Refine

Ronald Metoyer
Associate Professor
Oregon State University
What is Google Refine

Formerly know as Freedbase Gridworks
Tool for working with “messy” data
where “messy” = inconsistent
Also useful for restructuring the data
A scripting language targeted to the data table format
Can also link to web services & databases
Getting Started

https://code.google.com/p/google-refine/

Download & Install

Watch tutorial videos

Launch the executable – it will run in the default browser
Import Data

Supports several formats:

*tsv, csv, xls, xlsx, xml, json, google data docs*
Facets

Facets give you an overview of the column
Look for things out of the ordinary in columns

Can create custom facets for custom filter groups
e.g. facet by first letter in the word

Several Types:
text, numeric, custom, etc.
GREL & Expressions

Expressions give you power

You can create custom facets with expressions

You can filter with expressions

You can transform cells with expressions
Columns & Rows

Column Specific Operations

- Add columns
- Split columns

Add columns based on current column

- Useful for deriving new values from a column
- Should keep old column around until no longer needed!
Structural Edits

Sometimes you want to change columns to rows

To change an entire row to a column, Refine has transpose operations (see documentation!)

To change a variable number of rows to columns

Select those parts of the rows and create columns from them
Useful Features

Undo/Redo History
  Can see your actions and save the history
  (I’ll save the history from these demos!)

Export