# OpenRefine

## Tips and tricks for cleaning up archival metadata

Emily Sommers, Digital Records Archivist, U of T Archives

# Getting started

Download Open Refine: https://openrefine.org/download.html

- Installation instructions: https://docs.openrefine.org/manual/installing/

Documentation manual: https://docs.openrefine.org/

# Things you can do with OpenRefine

- Find typos in controlled vocabulary fields
- Find duplicates
- Create new column with pre-populated value
- Add leading zeros to numbers
- Merge columns
  - Merge columns (with different separators)
- Split column into multiple columns
- Add slug to all files from one series

- Clean up dates
  - Split dates into two columns (start & end)
  - When you have **start** and **endDate** columns and you want to populate the **endDate** column with the same value as **start** (if there was nothing to split!)
  - Remove various characters, punctuation, and circa notations
  - Converting free-text dates to ISO-8601 machine readable

# Find typos

**Why this is useful?**

- To avoid adding additional terms to a controlled taxonomy / vocabulary list
- Sometimes controlled vocabulary columns are turned into facets or drop-down lists, and you wouldn’t want these to include typos

**Select column > Facet > Text Facet**

- Sort by count to quickly identify typos

# Find duplicates

**Why this is useful?**

- You may have columns that require unique values, e.g. identifier in a file list

**Select column > Customized facets > Duplicates facet**

- True = duplicates
- If you want to facet the identified duplicates, then

**Select column > Facet > Text Facet**

- Can sort by name or count

# Create new column with pre-populated value

**Why this is useful?**

- Quickly populate a column with the same value, e.g. Level of Description

**Select column > Edit column > Add column based on this column...**

- Give column a name
- Replace ‘value’ with the term you want to fill down in quotation marks
    - “File”

# Add leading zeros

**Why this is useful?**

- Quickly add leading zeros to boxes or file numbers so that they are consistent lengths

**Select column > Edit cells > Transform**

- "000"[0,3-length(value)] + value
  - 3 is the length, so if you want more or less leading zeros, adjust accordingly

# Merge columns (with the same separator)

**Why this is useful?**

- Sometimes you may want to merge data from multiple columns into one, e.g. when creating a ‘Citation’ field for a digital collection.

Go to one of the columns you would like to join, then
**Edit column > Join columns**

- You can add separator between the contents of each column.
- You can overwrite combining information into the original column or create a new column for the combining contents.

# Merge columns (with different separators)

**Why this is useful?**

- Merge columns to create accession/box(file)
- Similar to CONCATENATE function in Excel

Go to one of the columns you would like to merge (i.e box no.), then **Edit cells > Transform**

- ‘B1991-0013’ + ‘/’ + value + ‘(‘ + cells[‘columnName’].value + ‘)’

Value of this column (box no.)

Value of other column

# Split column into multiple columns

**Why this is useful?**

- Sometimes you want to split a column into more useful pieces of data, i.e Surname and First Name

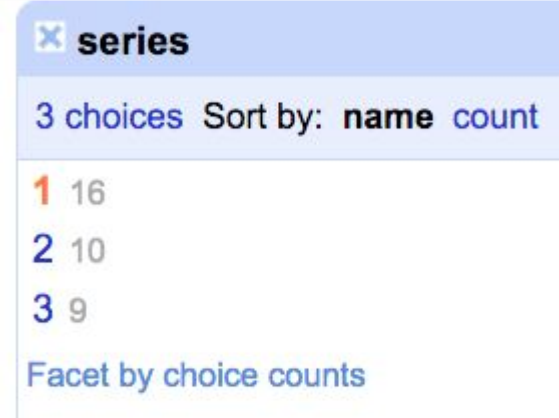
**Select column > Edit column > Split into several columns...**

- Can split by separator or by field lengths

# Add slug to all files from one series

**Select series column > Facet > Text facet**

Select a series



series

3 choices Sort by: name count

1 16

2 10

3 9

Facet by choice counts

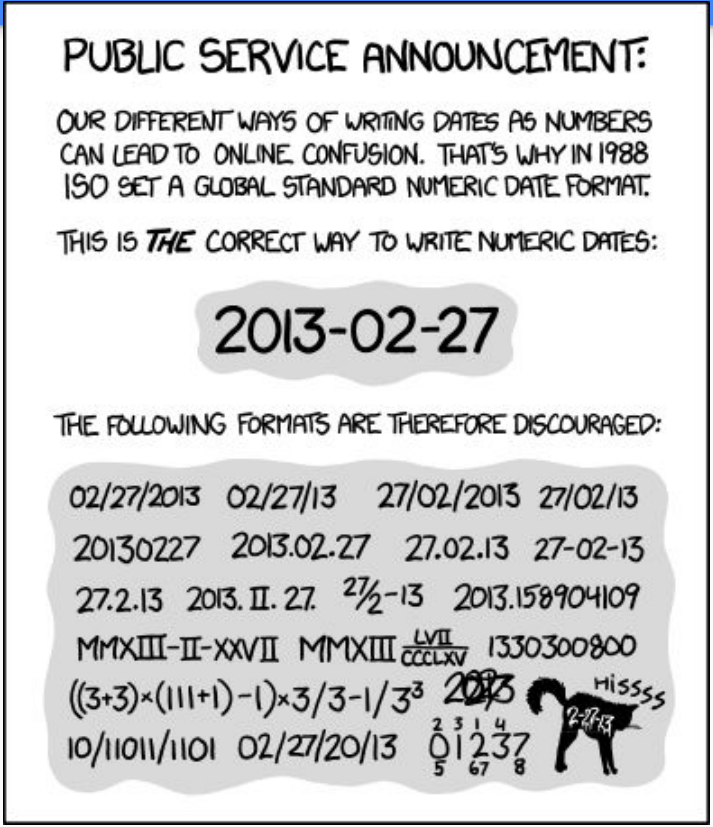
If there is already a slug column:

- Slug column > Edit cells > Transform
- Expression: change “value” to “series-slug”

If there is no slug column:

- Series column > Edit column > Add column based on this column
- New column name: qubitParentSlug
- Expression: change “value” to “series-slug”

# Clean up dates


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
Hissss
2-27-13
10/11011/1101 02/27/20/13 01237

“ISO 8601”, XKCD, https://xkcd.com/1179/

# Clean up dates

**Why this is useful?**

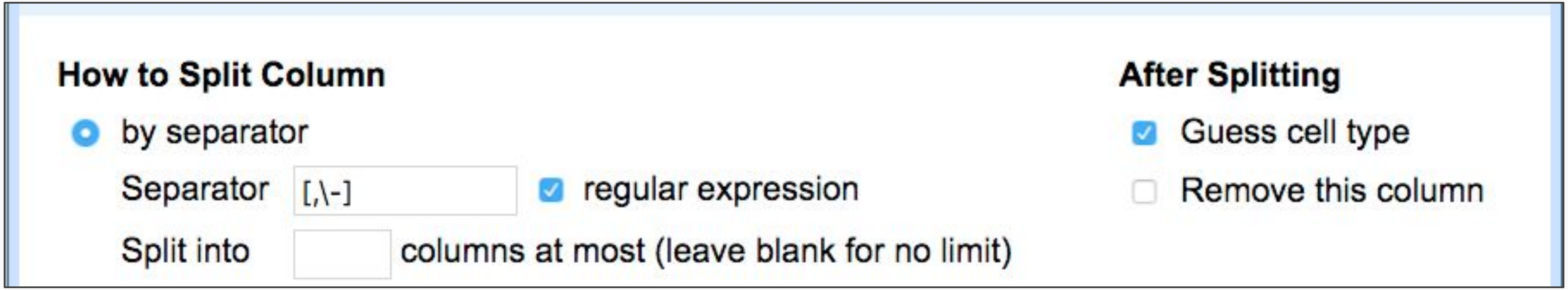
- Sometimes you need to turn free text field with approximate dates, into machine-readable dates to support date range searching and sorting

Cassie Schmitt, “Date Formats”,
https://icantiemyownshoes.wordpress.com/2014/04/24/clean-up-dates-and-openrefine/

# Clean up dates

## Split dates into two columns (e.g. start and end)

1. Look and see what the separators are, most likely - and ,
   - **Facet > Text facet**
2. Select column to split
   - **Edit colum > Split into several columns..**
   - By separator [,\-]
     - Make sure regular expression is checked
     - Remove this column is unchecked



**How to Split Column**

● by separator

Separator [,\-] ☑ regular expression

Split into [ ] columns at most (leave blank for no limit)

**After Splitting**

☑ Guess cell type

□ Remove this column

# Clean up dates

**When you have start and endDate columns and you want to populate the endDate column with the same value as start (if there was nothing to split!)**

1. Select all blank cells in endDate column
    - Facet > Customized facets > Facet by blank > true
2. Fill blank cells with the values from the startDate column
    - Edit cells > Transform
    - Expression: cells[‘startDate’].value where ‘startDate’ is the column header

# Clean up dates

## Remove various characters, punctuation, and circa notations

Isolate rows that may have these things

- Select column > Text filter > [

then...

- Edit cells > Transform
    - value.replace('[ca. ','').replace(']','').replace('[','')
    - ...and whatever else might be in the date column

# Clean up dates

## Converting free-text dates to ISO-8601 machine readable

- Will depend on how the dates are written, but here are basic steps:
    - Use Facets or Text Filter to isolate rows with dates that are more than just year
    - Split into 2 or 3 columns - day, month, year or month, year
    - Replace month with numeric month (e.g. Jan to 01)

```
value.replace(‘Jan. ‘, ’01’).replace(‘Feb. ‘, ’02’).replace(‘Mar. ‘, ’03’).replace(‘Apr. ‘, ’04’).replace(‘May ‘,
’05’).replace(‘Jun. ’, ’06’).replace(‘Jul. ‘, ’07’).replace(‘Aug. ‘, ’08’).replace(‘Sep. ‘, ’09’).replace(‘Sept. ‘,
’09’).replace(‘Oct. ‘, ’10’).replace(‘Nov. ‘, ’11’).replace(‘Dec. ‘, ’12’)
```

- 
    - In the eventStartDate column, transform the freetext dates (4 Jan. 1972) with data from the day, month and year columns
        - Edit cells > Transform
        - cells[‘year’].value + ‘-’ + cells[‘month’].value + ‘-’ + cells[‘day’].value

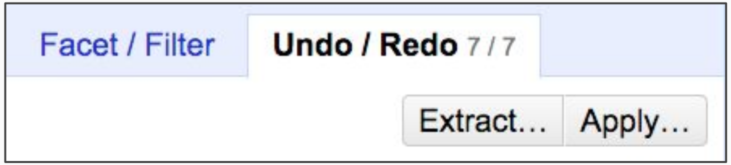
# Other common transformations

- Delete blanks
- Remove whitespace
- Unescape HTML entities
- To titlecase
- To uppercase
- To lowercase
- To number
- To date
- To text

**Select column > Edit cells > Common transforms**

See "Common Transformations"
https://guides.library.illinois.edu/openrefine/commontransform

You can also extract your steps if you think you'll be repeating them again on another dataset



# Some resources besides Google

University of Illinois Library OpenRefine LibGuide

Library Carpentry: OpenRefine

Chaos → Order blog

Katrina Cohen-Palacios, “Wikidata and Archivists”

You got this!