

OpenRefine Loves Data

Meredith L. Hale, Metadata Librarian

mhale16@utk.edu

February 12, 2026

Slides: <https://tiny.utk.edu/ORLovesData>



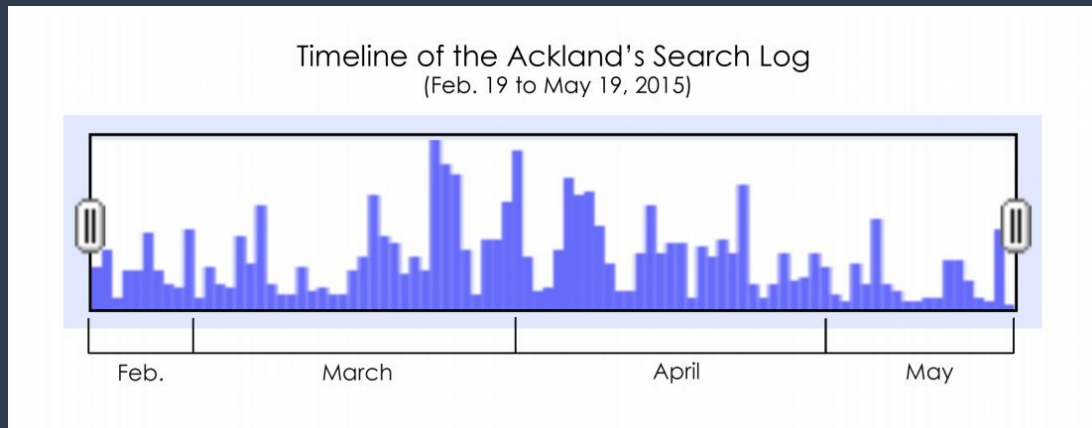
What is OpenRefine Good For?



- Analyzing Data
- Cleaning Data
- Enhancing Data (with URIs, coordinates, etc.)
- Changing the Format of Data
- Automating Routine Data Processes

OpenRefine for Analysis

- My first foray into OpenRefine focused on analyzing log data from the Ackland Art Museum



Finding Your Data

- Pull from an endpoint using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
 - <https://github.com/vphill/pyoaiharvester>
 - UTK Libraries endpoint - <http://dloai.lib.utk.edu/cgi-bin/XMLFile/dlmodsoai/oai.pl>
- Application Programming Interfaces (APIs)
 - [Digital Public Library of America \(DPLA\)](#)
- Other Open Data Sites
 - <https://www.data.gov/>
 - <https://data.worldbank.org/>
 - <https://www.census.gov/>
- Create your own data set

Finding Your Data

Department of Data & Digital Scholarship (3DS)

- Contact: dataservices@utk.edu
- https://libguides.utk.edu/find_data

Finding Social Science Data for Research

Selected data resources categorized by academic discipline

What is Secondary Data?	National Center for Health Statistics (NCHS)
Getting Started	NCHS is one of the leading providers of health care statistics and data. NCHS provides public access to both administrative data (birth and death records) and to a wide range of health survey data (National Health Interview Survey, National Survey of Family Growth, National Nursing Home Survey, Longitudinal Study on Aging, etc.)
Frequently Used Data	
Data Resources by Topic	Agency for Healthcare Research and Quality (AHRQ)
Crime & Justice	Provides access to two main data sources: Medical Expenditure Panel Survey - MEPS is designed to provide data on the cost and utilization of healthcare, as well as data on health insurance. Healthcare Cost and Utilization Project - HCUP provides data on hospital usage and outcomes.
Economic & Finance	
Education	
Health	Centers for Disease Control
International	Search for Data & Statistics within the 'More CDC Topics' category (located in the top right portion of the main CDC page).
Labor/Employment	
Multi-Topic	
Tennessee	
Citing Data	
Get Help	

Consider OpenRefine's Limits

- Question whether your data will be effectively manipulated using OpenRefine
- Memory allocation can be increased from the standard 1024MB to enhance performance speeds
 - <https://docs.openrefine.org/manual/installing/#increasing-memory-allocation>
- Limits to OpenRefine
 - No more than one million total cells
 - Input file size maximum of 50 megabytes (MB)
- My personal experience has shown that spreadsheets with more than 15,000 rows or 256 columns cannot be easily accommodated

Not all Data are Diamonds . . .

Katie Rawson and Trevor Muñoz, "Against Cleaning", <http://curatingmenus.org/articles/against-cleaning/>.



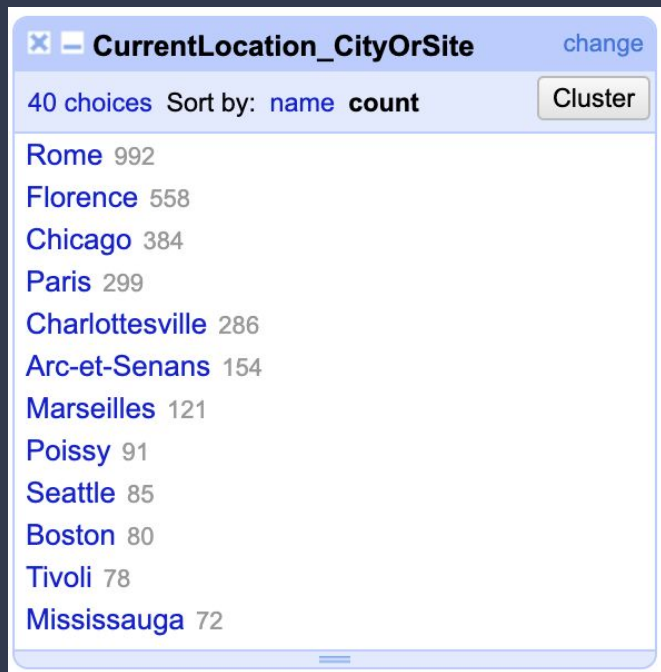
VS



"Crystal Diamonds" by [Kim Alaniz](#) is licensed under [CC BY-ND 2.0](#)

"Amethyst geode (Serra Geral Formation, Lower Cretaceous; southeastern Brazil) 3" by James St. John is licensed under [CC BY 2.0](#)

Cleaning your Data



Built-in functions

- Facet (by text, number, timeline, word)
- Change data format (text, date, number)
- Normalize spacing
- Normalize case
- Split values into different columns
- Clustering to normalize values

Documentation

- [OpenRefine User Manual](#)

Videos

- [Introduction to OpenRefine](#), Sarah Hatoum (Pratt)
- [Using Open Refine to Update, Clean up, and Link Your Metadata to the Wider World](#) (ALCTS)
- [Clean your Data: Getting Started with OpenRefine](#), Evan Williamson (Un. of Idaho)

Enhancing Your Data – Reconciliation

- Reconciliation means standardizing value variations in accordance with an established controlled form of the value

Waterhouse, John
Waterhouse, John W.
Waterhouse, J. W.



Waterhouse, John William
<http://vocab.getty.edu/ulan/500027032>



John William Waterhouse,
Destiny, 1900.

Enhancing Your Data – Reconciliation

Library of Congress Authorities

- [Christina Harlow's service](#)
- [Michael Phillip's service](#)
- [Jeff Chiu's service for LoC Name Authority Files](#)
- [University of Tennessee, Knoxville \(built on Harlow's code\)](#)
 - [Instructions](#)

Getty Vocabularies

- [Getty Reconciliation Instructions](#)

Virtual International Authority File (VIAF) & ORCID

- [Jeff Chiu's service](#)

Geonames

- [Christina Harlow's service](#)
- [University of Tennessee, Knoxville \(built on Harlow's code\)](#)
- [FornPunkt service](#)

Automating Processes



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.

- ☐ Match item architectural drawings (visual works) (aat/300034787) for cells containing "architectural drawings" in column WorkType
- ☐ Match item mosaics (visual works) (aat/300015342) for cells containing "mosaics (visual works)" in column WorkType
- ☐ Match item paintings (visual works) (aat/300033618) for cells containing "paintings" in column WorkType
- ☐ Mass edit cells in column WorkType
- ☐ Reconcile cells in column WorkType to type /aat
- ☐ Match item sculpture (visual works) (aat/300047090) for cells containing "sculpture (visual works)" in column WorkType
- ☐ Reconcile cells in column WorkType to type /ulan
- ☐ Text transform on cells in column WorkDateDisplay using expression value.toString()
- ☐ Text transform on cells in column WorkDateDisplay using expression value.toDate()
- ☐ Text transform on cells in column Rights using expression value.trim()
- ☒ Text transform on cells in column Filename using expression value.trim()

Select All Unselect All

Close

```
[
  {
    "op": "core/text-transform",
    "engineConfig": {
      "facets": [],
      "mode": "row-based"
    },
    "columnName": "Filename",
    "expression": "value.trim()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10,
    "description": "Text transform on cells in column Filename using expression value.trim()"
  }
]
```

Archivision

The Archivision Architectural Image Collection, comprised of approximately 15,000 high resolution images, encompasses a variety of examples of architecture, landscape design, and public art from North America, Europe, and Asia.



— Introduction —

— Browse —

Archivision

search

Home » Islandora Repository

Resource Type

- still image (15062) + -

Subject

- Male artists, (Library of Congress Subject Headings) (12693) + -
- Women architects, (Library of Congress Subject Headings) (48) + -
- Women artists, (Library of Congress Subject Headings) (25) + -
- Women potters, (Library of Congress Subject Headings) (25) + -

Archivision

View

Manage

Grid view

List view

1 2 3 4 5 6 7 8 9 ... next last »



AT&T Headquarters

Open space at ground level, depicting oculi windows (before restoration by Sony)



AT&T Headquarters

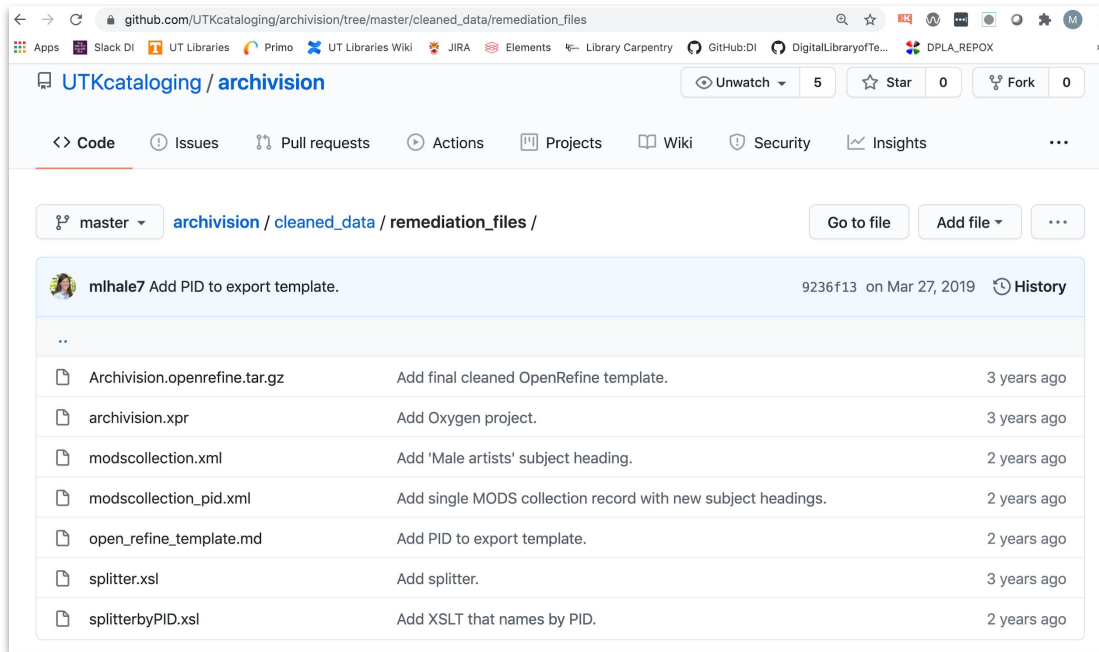
Interior lobby looking up, depicting gilded bronze statue of the "Spirit of Communication" or "Golden Boy" (the symbol of AT&T for more than half a century), on axis



Boston Public Library Addition

General view, from the northwest, depicting the full north facade with the original library by McKim, Mead & White

OpenRefine in My Work



- GitHub - <https://github.com/UTKcataloging>
- Example export template - https://github.com/UTKcataloging/archivision/blob/master/cleaned_data/remediation_files/open_refine_template.md

Resources

- “Free Your Metadata”, <https://freeyourmetadata.knows.idlab.ugent.be/>
- Library Carpentry’s “Introduction to Working with Data - Regular Expressions”, <https://librarycarpentry.github.io/lc-data-intro/01-regular-expressions.html>
- “Open Refine User Manual”, <https://docs.openrefine.org/>
- Verborgh, Ruben and Max De Wilde. *Using OpenRefine*. Olton: Packt Publishing, Limited, 2013. ([Library Ebook](#))
- Workshop GitHub Repository, <https://github.com/mlhale7/OpenRefineWorkshop>

Demo!

