OpenRefine Workshop

Meredith L. Hale, Metadata Librarian
mhale16@utk.edu

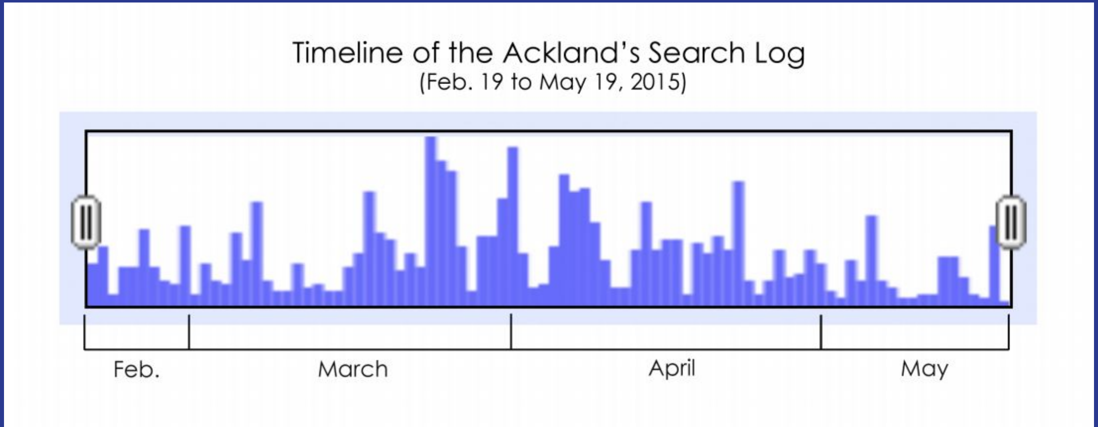
Open Sandbox Series event on February 16, 2020

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://www.pewresearch.org/internet/datasets/
  - https://data.worldbank.org/
- Create your own data set

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
- OpenRefine Lessons for Digital Humanities

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, Heroku-hosted service

Getty Vocabularies

- Getty Reconciliation Instructions

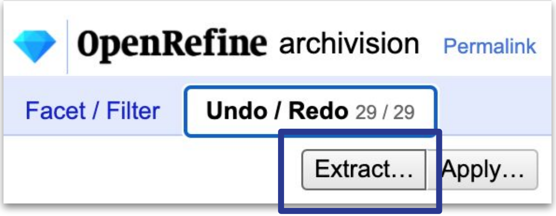
Virtual International Authority File (VIAF)

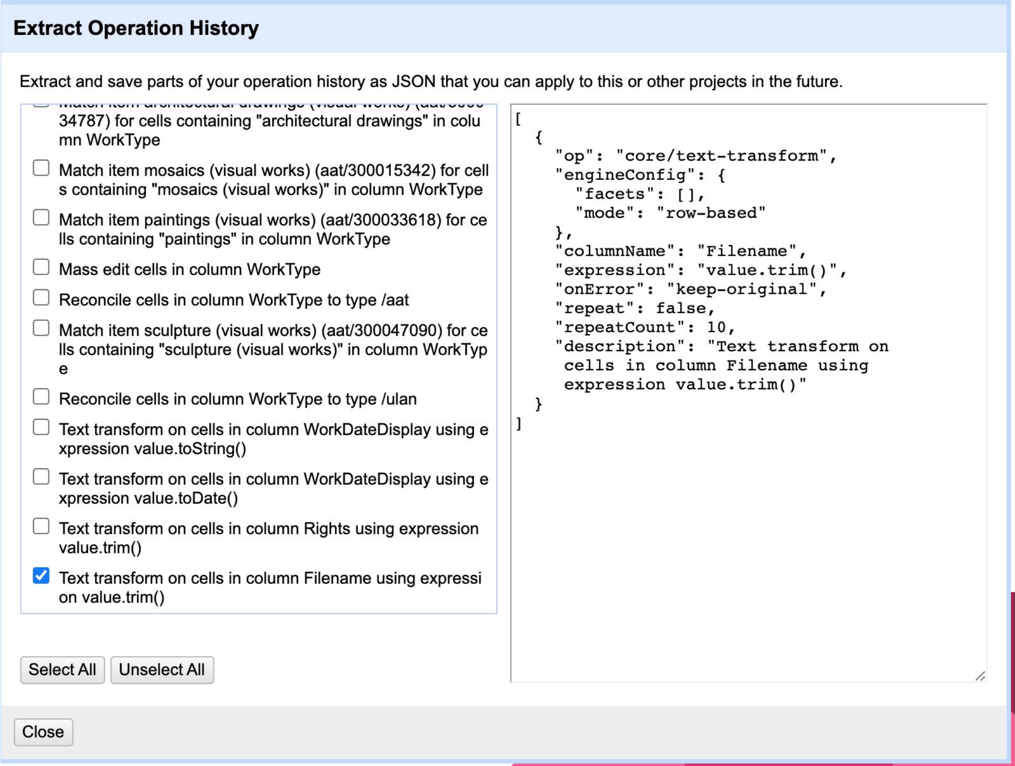
- Jeff Chiu's service

Geonames

- Christina Harlow's service
- University of Tennessee, Knoxville, Heroku-hosted service

Automating Processes





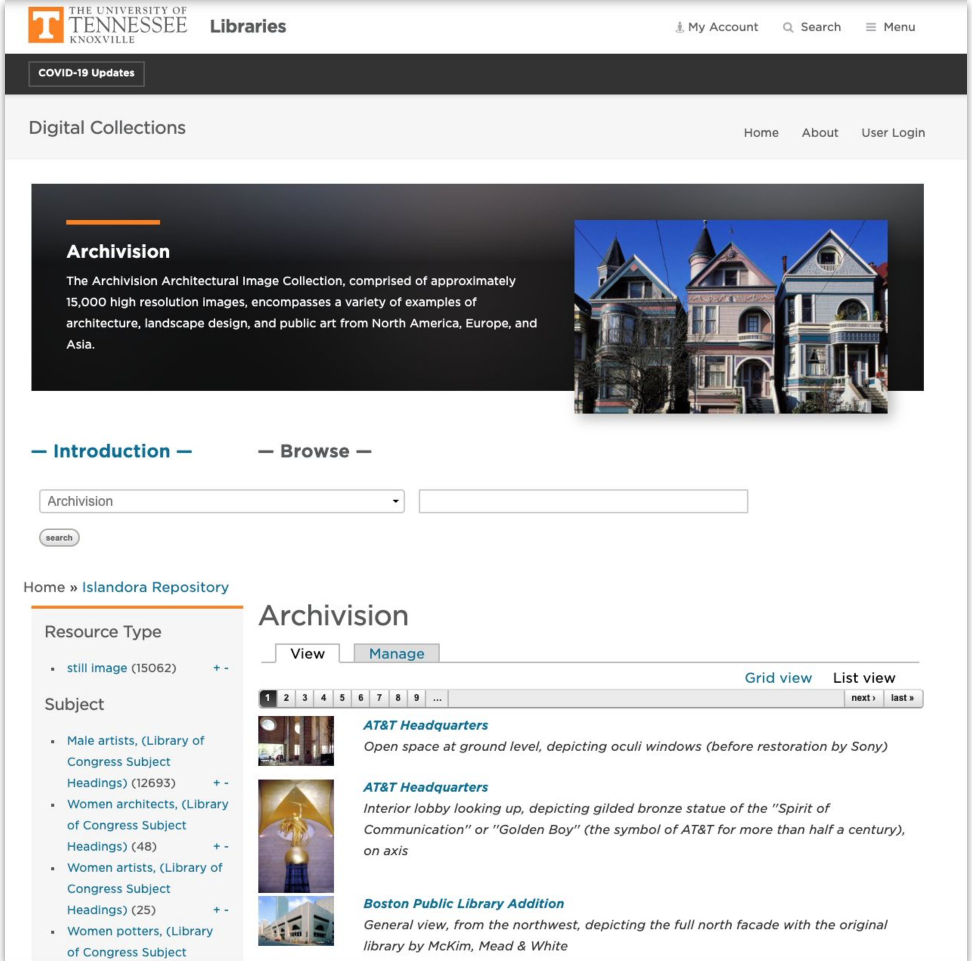
Extract Operation History

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()

```
[
  {
    "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()"
  }
]
```

OpenRefine in My Work Today



THE UNIVERSITY OF TENNESSEE KNOXVILLE Libraries

My Account Search Menu

COVID-19 Updates

Digital Collections

Home About User Login

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

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)

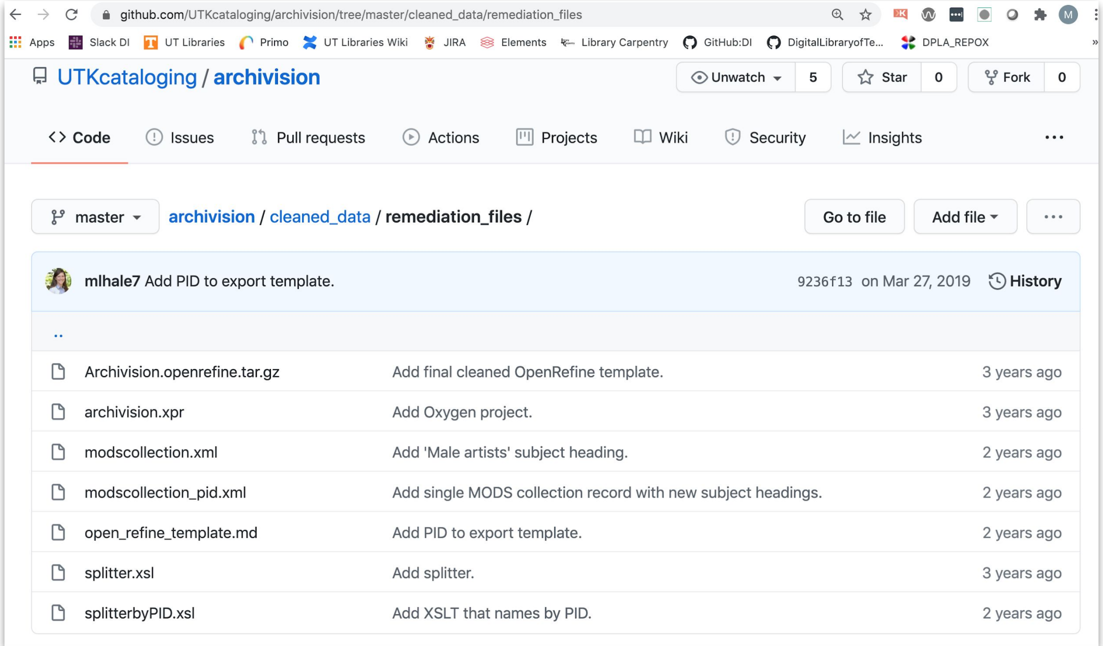
AT&T Headquarters

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 Today



UTKcataloging / archivision

archivision / cleaned_data / remediation_files /

mlhale7 Add PID to export template. 9236f13 on Mar 27, 2019

..

Archivision.openrefine.tar.gz	Add final cleaned OpenRefine template.	3 years ago
archivision.xpr	Add Oxygen project.	3 years ago
modscollection.xml	Add 'Male artists' subject heading.	2 years ago
modscollection_pid.xml	Add single MODS collection record with new subject headings.	2 years ago
open_refine_template.md	Add PID to export template.	2 years ago
splitter.xsl	Add splitter.	3 years ago
splitterbyPID.xsl	Add XSLT that names by PID.	2 years ago

- 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.org/
- Library Carpentry’s “Introduction to Working with Data - Regular Expressions”, https://librarycarpentry.org/lc-data-intro/01-regular-expressions/index.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!