Metadata Cleanup with Linked Data and OpenRefine

Greer Martin

Metadata Technologies Librarian

Loyola University Chicago

https://id.loc.gov

LIBRARY OF CONGRESS

ASK A LIBRARIAN | DIGITAL COLLECTIONS | LIBRARY CATALOGS

Search | Search Loc.gov | GO

The Library of Congress > Linked Data Service

ID.LOC.GOV

- Linked Data Service
- About
- Search
- Download
- Technical Center
- Contact Us
- Privacy Policy

ID.LOC.GOV – Linked Data Service

ID.LOC.GOV provides both interactive and machine access to commonly used ontologies, controlled vocabularies, and other lists for bibliographic description. To search everything, use the search box below. Or, select from the list below to learn more and search only in that vocabulary.

Search id.loc.gov | GO

Subjects, Thesauri, Classification

- LC Subject Headings (LCSH)
- LC Classification (LCC)
- LC Children's Subject Headings
- LC Medium of Performance Thesaurus for Music (LCMPT)
- Thesaurus for Graphic Materials (TGM)
- AFS Ethnographic Thesaurus
- Subject Schemes
- Classification Schemes

Agents

- LC Name Authority File (LCNAF)
- Cultural Heritage Organizations

Ontologies

- BIBFRAME Ontology

Cataloging

- Aspect Ratio
- Authentication Action
- Broadcast Standard
- Carriers
- Color Content
- Content Types
- Description Conventions
- Encoding Level
- File Type
- Generation
- Groove Width/Pitch/Cutting
- Identifier Status
- Illustrative Content
- Intended Audience
- Issuance
- Layout
- LC Demographic Group Terms (LCDGT)

Languages

- MARC Languages
- ISO639-1 Languages
- ISO639-2 Languages
- ISO639-5 Languages

Preservation Vocabularies

- Preservation Vocabs (all)
- Actions Granted
- Agent Type
- Event Outcome
- Preservation Level
- Content Location Type
- Copyright Status
- Cryptographic Hash Functions
- Environment Characteristic
- Environment Function Type

From Library of Congress Subject Headings

Details | Suggest Terminology

College students

URI(s)
- http://id.loc.gov/authorities/subjects/sh85028356
- info:lc/authorities/sh85028356
- http://id.loc.gov/authorities/sh85028356#concept

Instance Of
- MADS/RDF Topic
- MADS/RDF Authority
- SKOS Concept

Scheme Membership(s)
- Library of Congress Subject Headings

Collection Membership(s)
- LCSH Collection - Authorized Headings
- LCSH Collection - General Collection
- LCSH Collection - May Subdivide Geographically

Variants
- College life
- College students--Education
- Universities and colleges--Students
- University students

Broader Terms
- Students

https://www.wikidata.org/wiki/Wikidata:Main_Page

English | Not logged in | Talk | Contributions | Create account | Log in

Main Page | Discussion | Read | View source | View history | Search Wikidata

WIKIDATA

- Main page
- Community portal
- Project chat
- Create a new Item
- Create a new Lexeme
- Recent changes
- Random item
- Query Service
- Nearby
- Help
- Donate

Print/export

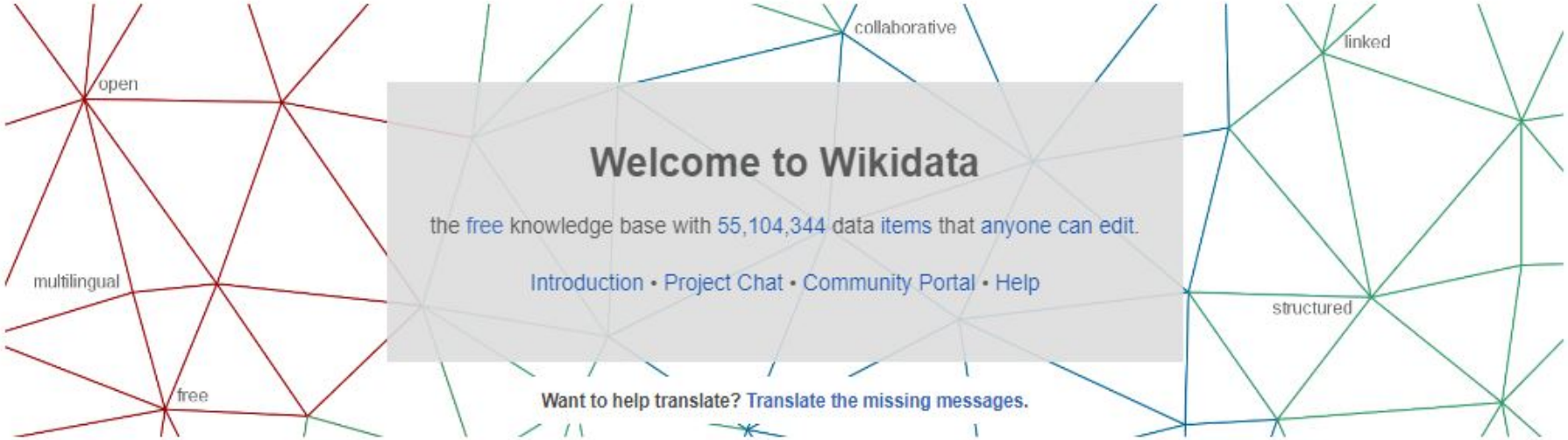
- Create a book
- Download as PDF
- Printable version

In other projects

- Wikimedia Commons
- MediaWiki
- Meta-Wiki
- Wikispecies
- Wikibooks
- Wikinews
- Wikipedia
- Wikiquote
- Wikisource
- Wikiversity
- Wikivoyage
- Wiktionary

Tools

- What links here



Welcome to Wikidata

the free knowledge base with 55,104,344 data items that anyone can edit.

Introduction • Project Chat • Community Portal • Help

Want to help translate? Translate the missing messages.

Welcome!

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.

Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others.

Wikidata also provides support to many other sites and services beyond just Wikimedia projects! The content of Wikidata is available under a free license, exported using standard formats, and can be interlinked to other open data sets on the linked data web.

Get involved

Learn about Wikidata

- What is Wikidata? Read the **Wikidata introduction**.
- Explore Wikidata by looking at a featured showcase item for author Douglas Adams.

Learn about data

New to the wonderful world of data? Develop and improve your data literacy through content designed to get you up to speed and feeling comfortable with the fundamentals in no time.

Item: *Earth (Q2)* | Property: *highest point (P610)* | custom value: *Mount Everest (Q513)*

Other Data Sources for Reconciliation

- FAST (Faceted Application of Subject Terminology)
- VIVO Scientific Collaboration Platform
- VIAF
- Sharedshelf Built Work Registry
- OpenCorporates
- dbpedia
- GeoNames

OpenRefine

- Powerful data cleanup tool
- Formerly known as Google Refine
- Java tool
- GUI in browser
- Functionality: Export/import data, facets, clusters, clean, GREL, reconciliation

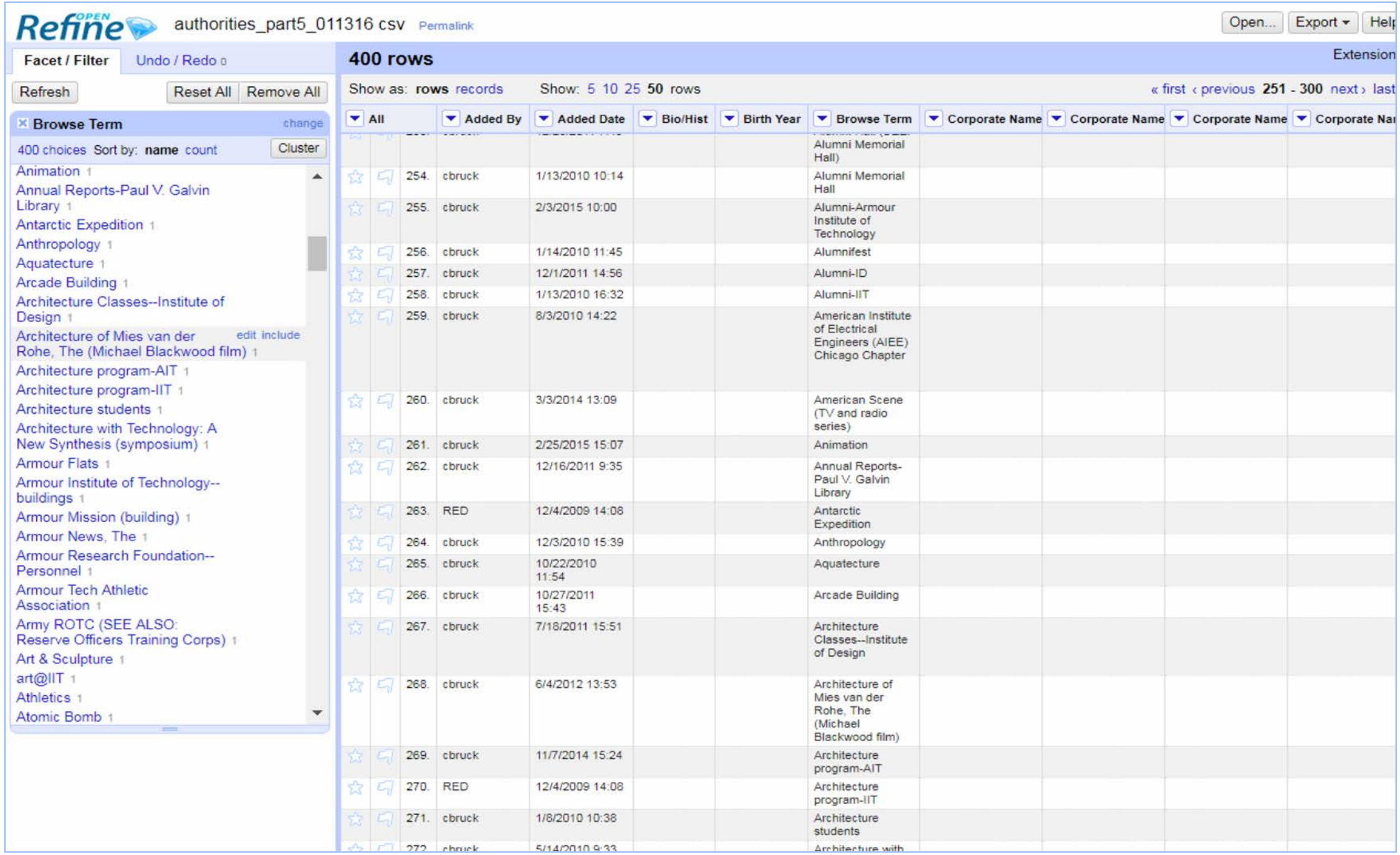
ArchivesSpace Migration

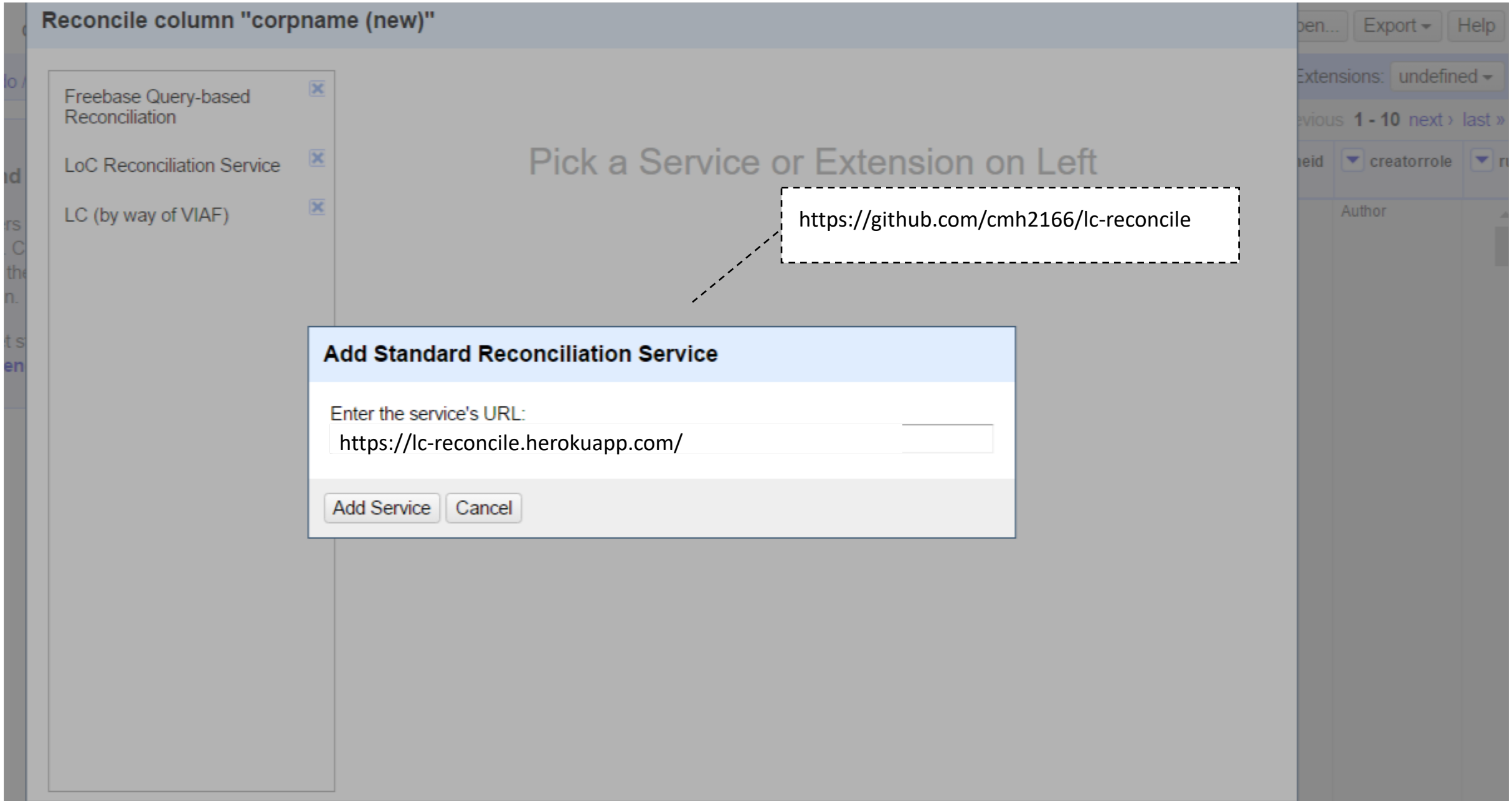
- 1,500 collections, mostly university records
- Use Re:Discovery Proficio for archival records management system
- Inconsistent use of authorities and controlled vocabularies

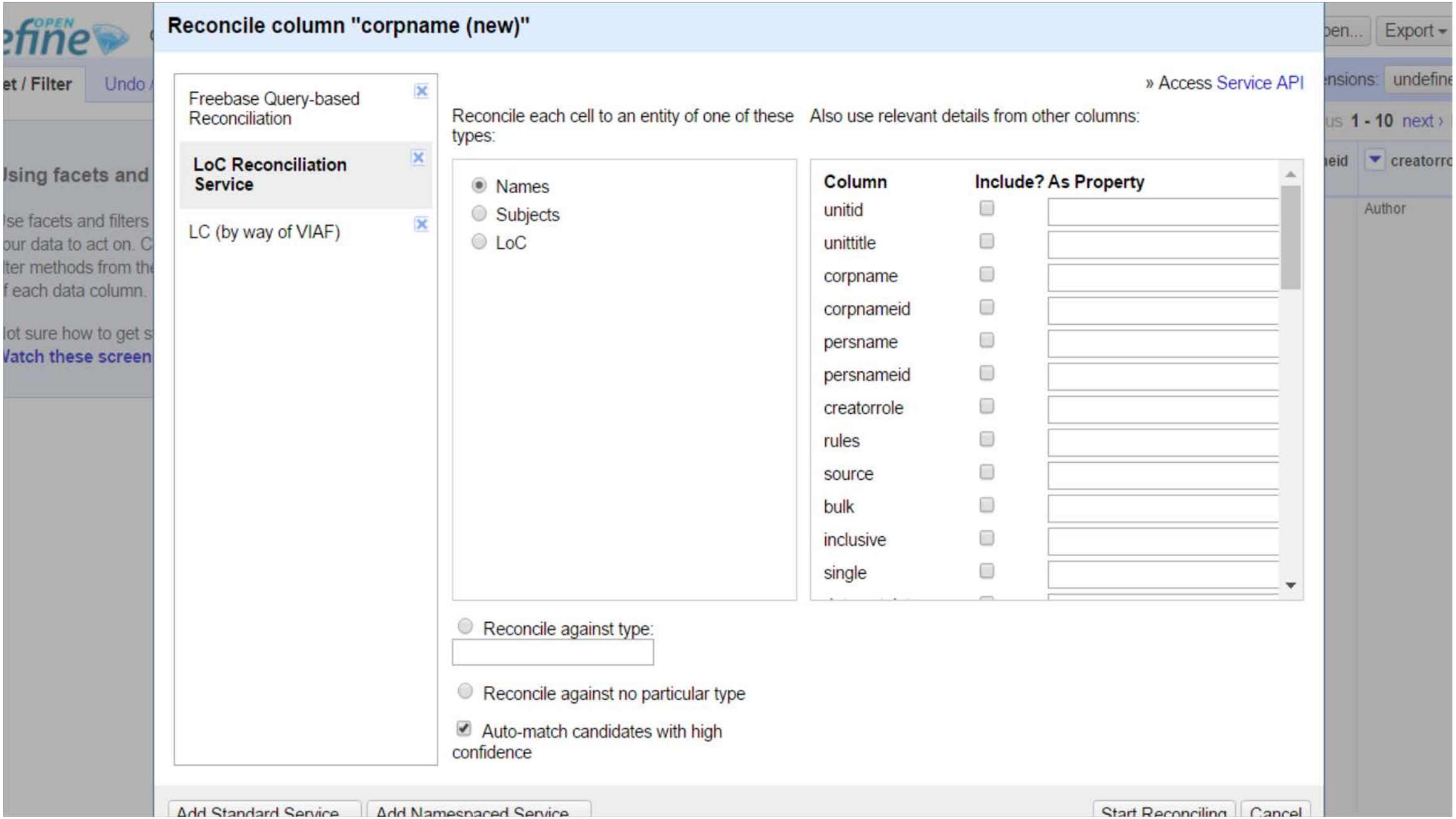
University Archives & Special Collections, Paul V. Galvin Library, Illinois Institute of Technology

Browse Term	Term Type
College Bowl	Topic term
College of Architecture--IIT	Corporate name
College of Psychology	Topic term
College of Psychology-IIT	Corporate name
College of Science and Letters--IIT	Corporate name
Collens Scholarship Program	Topic term
Collens, Lewis (Lew)	Personal name
Collens, Marge	Personal name
Collins, John	Personal name
Colonade Park (aka Branch Brook), Newark, New Jersey	Geographic term
Commencements	Topic term
Committee on the Future of IIT (COMFIIT)	Topic term
Commons (building)	Topic term
Communication Engineering School	Corporate name
Community Relations	Topic term
Comprehensive Plan for Chicago (1965)	Topic term
Computer Engineering	Topic term
Computer Science	Topic term
Computers	Topic term
Computers & Computing, History of	Topic term
Concrete Canoe	Topic term

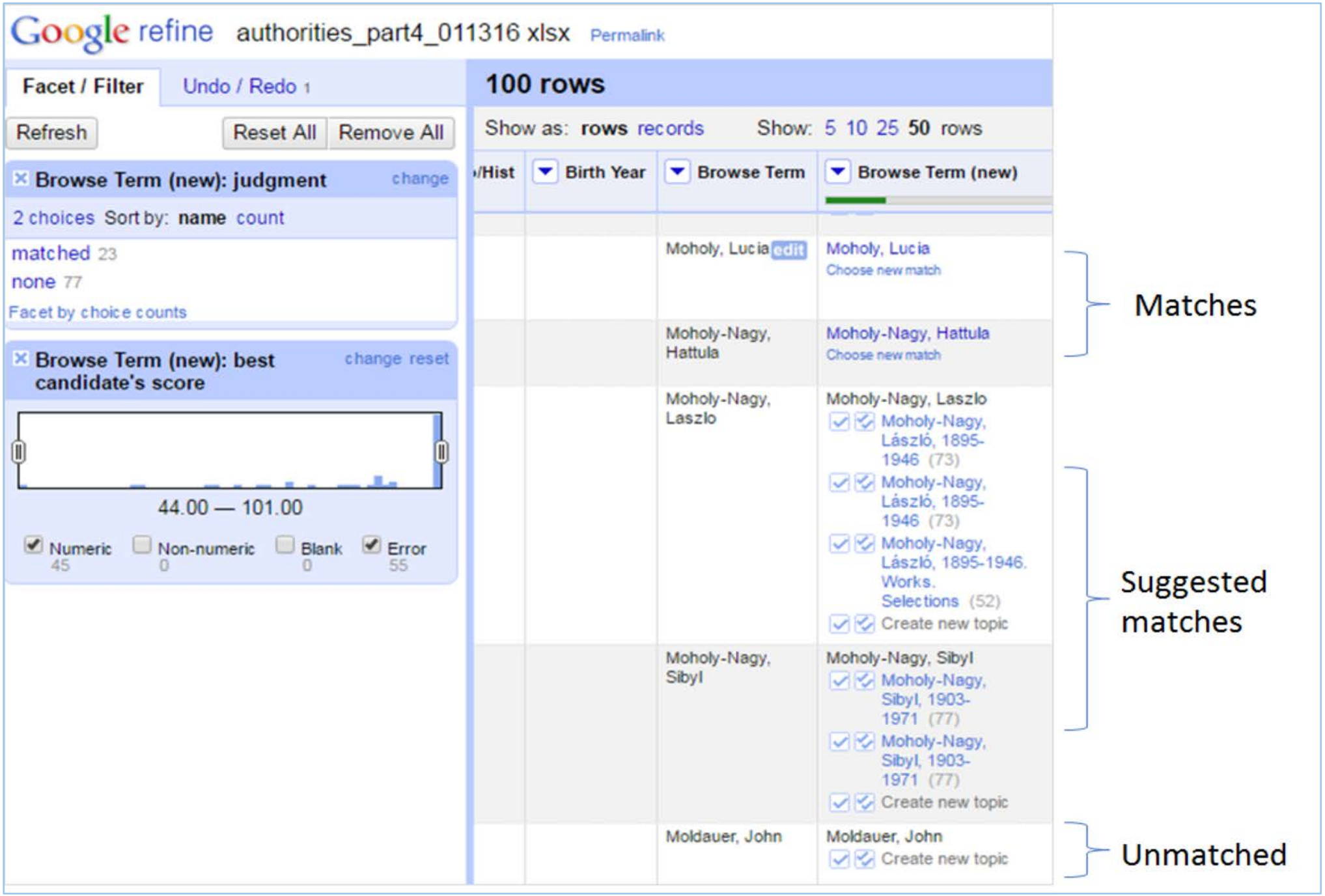
Browse Term	Term Type
Hermann Union Building (AKA: HUB) (SEE: Hermann Hall)	Topic term
Hern, Matthew	Personal name
Heron, Reginald	Personal name
Heskett, John	Personal name
Heuermann, Magda	Personal name
Hewitt, Louise	Personal name
Heyd, Kurt	Personal name
Higgins, ???	Personal name
Higgins, Geoffrey T.	Personal name
Higgins, Michael	Personal name
High Schools-Chicago	Topic term
Hilberseimer, Ludwig	Personal name
Hilberseimer, Ludwig K.	Personal name
Hill, Doug	Personal name
Hilvers, Robert Joseph	Personal name
Himmelreich, Dean	Personal name
Hirano, Marjorie Yasuko	Personal name
Hiroshima, Japan	Geographic term
Hispanics	Topic term
Hock, Daniel	Personal name
Hocker, Thomas	Personal name
Hoffman, Ellis	Personal name
Holabird and Roche	Corporate name


OpenRefine
authorities_part5_011316 csv
Permalink
Open...
Export
Help
Facet / Filter
Undo / Redo 0
Refresh
Reset All
Remove All
Browse Term
change
400 choices Sort by: name count
Cluster
Animation 1
Annual Reports-Paul V. Galvin Library 1
Antarctic Expedition 1
Anthropology 1
Aquatecture 1
Arcade Building 1
Architecture Classes--Institute of Design 1
Architecture of Mies van der Rohe, The (Michael Blackwood film) 1
edit include
Architecture program-AIT 1
Architecture program-IIT 1
Architecture students 1
Architecture with Technology: A New Synthesis (symposium) 1
Armour Flats 1
Armour Institute of Technology--buildings 1
Armour Mission (building) 1
Armour News, The 1
Armour Research Foundation--Personnel 1
Armour Tech Athletic Association 1
Army ROTC (SEE ALSO: Reserve Officers Training Corps) 1
Art & Sculpture 1
art@IIT 1
Athletics 1
Atomic Bomb 1
400 rows
Extension
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 251 - 300 next › last
All
Added By
Added Date
Bio/Hist
Birth Year
Browse Term
Corporate Name
Corporate Name
Corporate Name
Corporate Nar
Alumni Memorial Hall)
254. cbruck 1/13/2010 10:14 Alumni Memorial Hall
255. cbruck 2/3/2015 10:00 Alumni-Armour Institute of Technology
256. cbruck 1/14/2010 11:45 Alumnifest
257. cbruck 12/1/2011 14:56 Alumni-ID
258. cbruck 1/13/2010 16:32 Alumni-IIT
259. cbruck 8/3/2010 14:22 American Institute of Electrical Engineers (AIEE) Chicago Chapter
260. cbruck 3/3/2014 13:09 American Scene (TV and radio series)
261. cbruck 2/25/2015 15:07 Animation
262. cbruck 12/16/2011 9:35 Annual Reports-Paul V. Galvin Library
263. RED 12/4/2009 14:08 Antarctic Expedition
264. cbruck 12/3/2010 15:39 Anthropology
265. cbruck 10/22/2010 11:54 Aquatecture
266. cbruck 10/27/2011 15:43 Arcade Building
267. cbruck 7/18/2011 15:51 Architecture Classes--Institute of Design
268. cbruck 6/4/2012 13:53 Architecture of Mies van der Rohe, The (Michael Blackwood film)
269. cbruck 11/7/2014 15:24 Architecture program-AIT
270. RED 12/4/2009 14:08 Architecture program-IIT
271. cbruck 1/8/2010 10:38 Architecture students

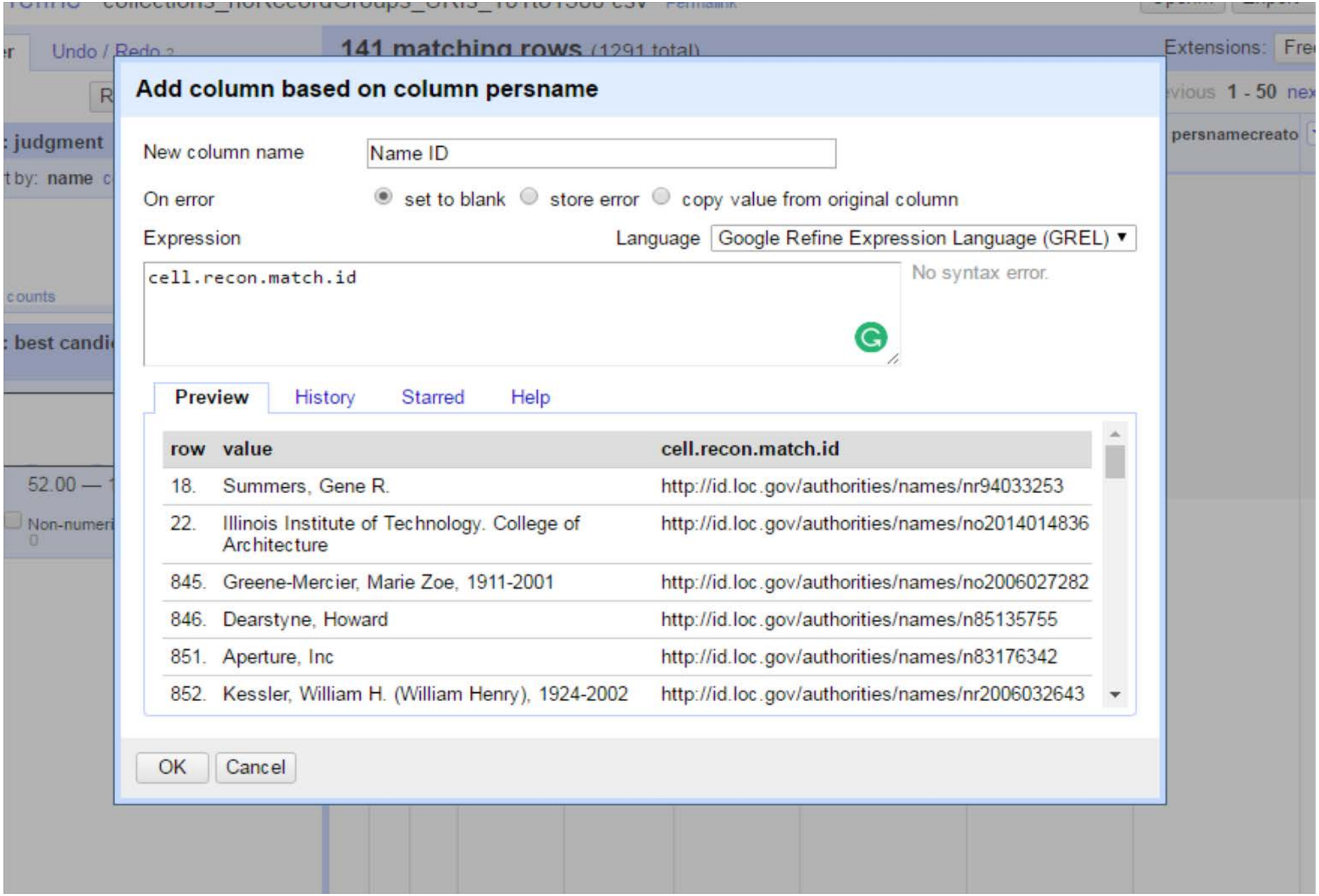

Reconcile column "corpname (new)"
Freebase Query-based Reconciliation
LoC Reconciliation Service
LC (by way of VIAF)
Pick a Service or Extension on Left
https://github.com/cmh2166/lc-reconcile
Add Standard Reconciliation Service
Enter the service's URL:
https://lc-reconcile.herokuapp.com/
Add Service
Cancel
Export
Help
Extensions: undefined
1 - 10 next › last »
creatorrole
Author


Reconcile column "corpname (new)"
Freebase Query-based Reconciliation
LoC Reconciliation Service
LC (by way of VIAF)
» Access Service API
Reconcile each cell to an entity of one of these types:
Names
Subjects
LoC
Also use relevant details from other columns:
Column
Include?
As Property
unitid
unittitle
corpname
corpnameid
persname
persnameid
creatorrole
rules
source
bulk
inclusive
single
Reconcile against type:
Reconcile against no particular type
Auto-match candidates with high confidence
Add Standard Service
Add Namespaced Service
Start Reconciling
Cancel

Post-reconciliation

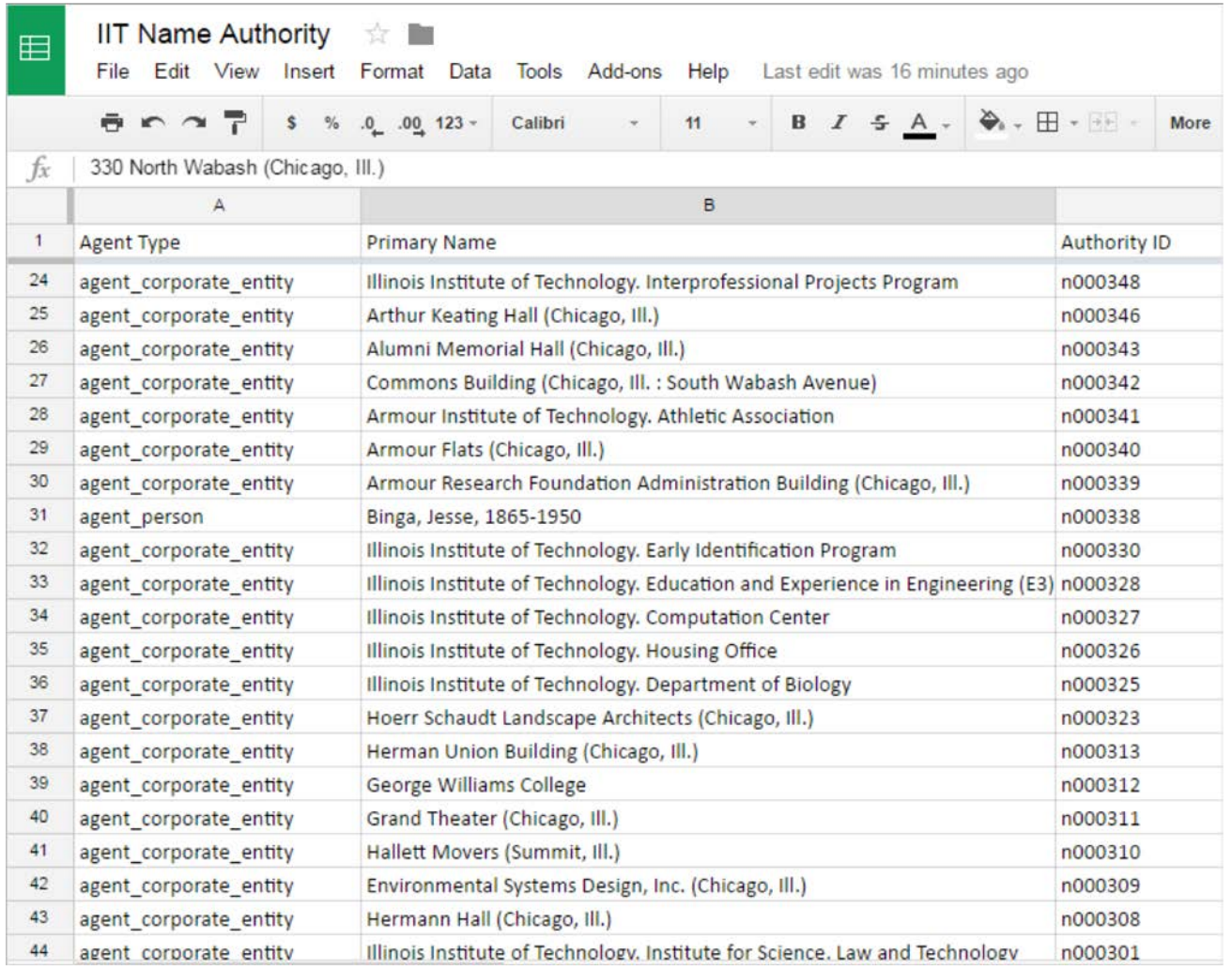


Adding URIs

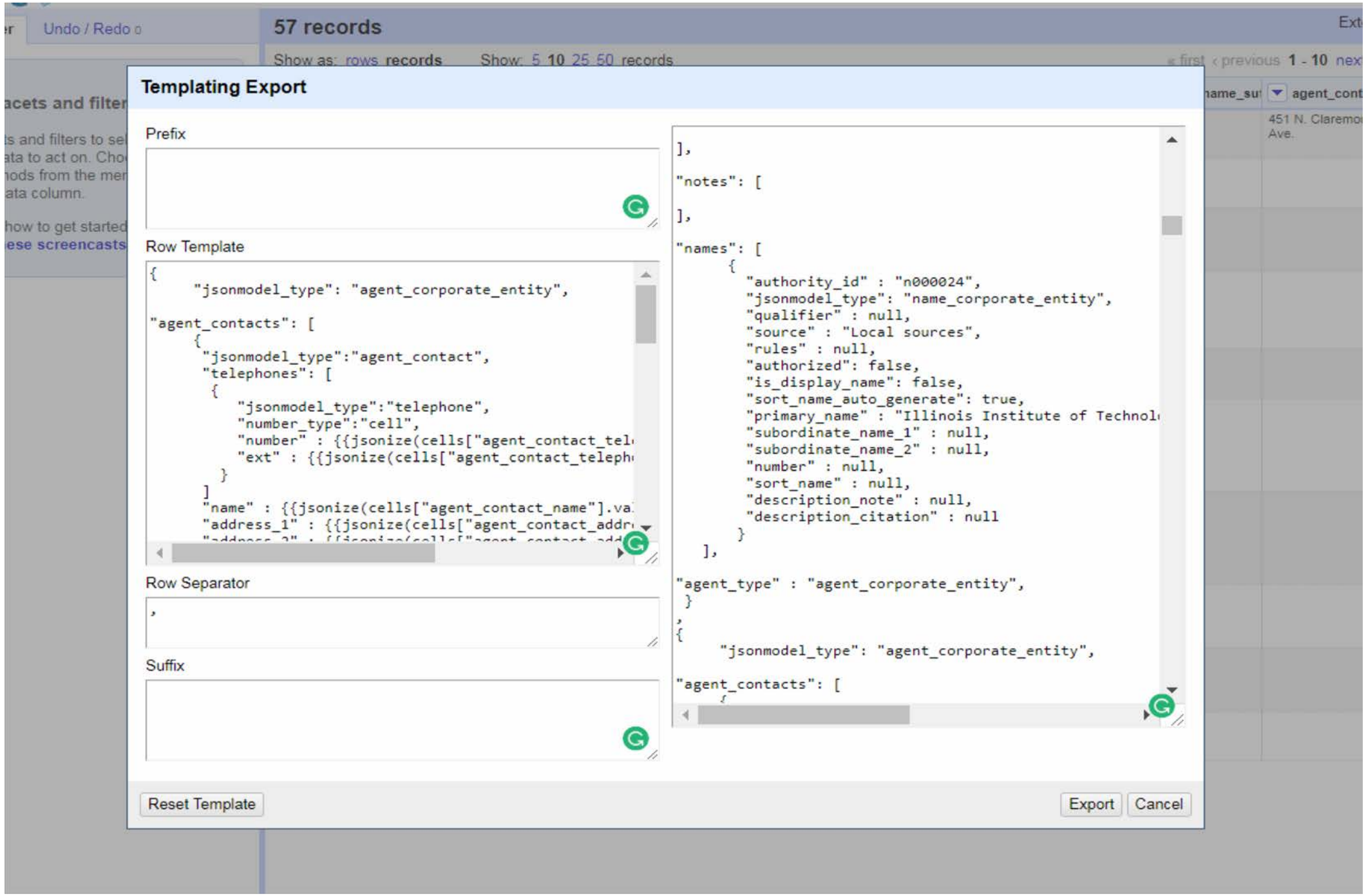

Add column based on column persname
New column name
Name ID
On error
set to blank
store error
copy value from original column
Expression
Language
Google Refine Expression Language (GREL)
cell.recon.match.id
No syntax error.
Preview
History
Starred
Help
row
value
cell.recon.match.id
18. Summers, Gene R. http://id.loc.gov/authorities/names/nr94033253
22. Illinois Institute of Technology. College of Architecture http://id.loc.gov/authorities/names/no2014014836
845. Greene-Mercier, Marie Zoe, 1911-2001 http://id.loc.gov/authorities/names/no2006027282
846. Dearstyne, Howard http://id.loc.gov/authorities/names/n85135755
851. Aperture, Inc http://id.loc.gov/authorities/names/n83176342
852. Kessler, William H. (William Henry), 1924-2002 http://id.loc.gov/authorities/names/nr2006032643
OK
Cancel

Reconcile-csv

- OpenRefine reconciliation service that matches one dataset against another
- Found at: http://okfnlabs.org/reconcile-csv/
- Good for adding local metadata

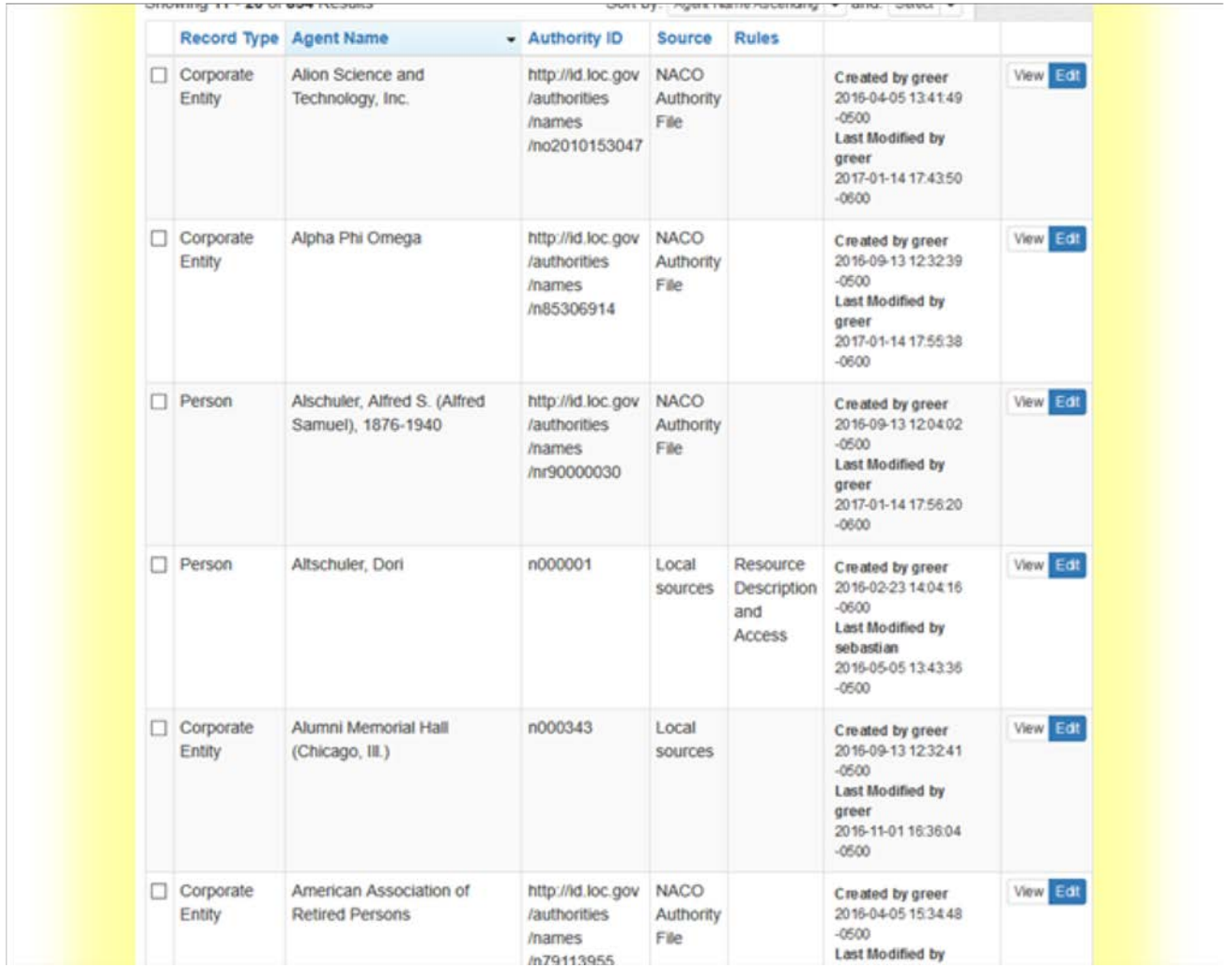


	A	B	
1	Agent Type	Primary Name	Authority ID
24	agent_corporate_entity	Illinois Institute of Technology. Interprofessional Projects Program	n000348
25	agent_corporate_entity	Arthur Keating Hall (Chicago, Ill.)	n000346
26	agent_corporate_entity	Alumni Memorial Hall (Chicago, Ill.)	n000343
27	agent_corporate_entity	Commons Building (Chicago, Ill. : South Wabash Avenue)	n000342
28	agent_corporate_entity	Armour Institute of Technology. Athletic Association	n000341
29	agent_corporate_entity	Armour Flats (Chicago, Ill.)	n000340
30	agent_corporate_entity	Armour Research Foundation Administration Building (Chicago, Ill.)	n000339
31	agent_person	Binga, Jesse, 1865-1950	n000338
32	agent_corporate_entity	Illinois Institute of Technology. Early Identification Program	n000330
33	agent_corporate_entity	Illinois Institute of Technology. Education and Experience in Engineering (E3)	n000328
34	agent_corporate_entity	Illinois Institute of Technology. Computation Center	n000327
35	agent_corporate_entity	Illinois Institute of Technology. Housing Office	n000326
36	agent_corporate_entity	Illinois Institute of Technology. Department of Biology	n000325
37	agent_corporate_entity	Hoerr Schaudt Landscape Architects (Chicago, Ill.)	n000323
38	agent_corporate_entity	Herman Union Building (Chicago, Ill.)	n000313
39	agent_corporate_entity	George Williams College	n000312
40	agent_corporate_entity	Grand Theater (Chicago, Ill.)	n000311
41	agent_corporate_entity	Hallett Movers (Summit, Ill.)	n000310
42	agent_corporate_entity	Environmental Systems Design, Inc. (Chicago, Ill.)	n000309
43	agent_corporate_entity	Hermann Hall (Chicago, Ill.)	n000308
44	agent_corporate_entity	Illinois Institute of Technology. Institute for Science, Law and Technology	n000301


Undo / Redo
57 records
Show as: rows records
Show: 5 10 25 50 records
« first ‹ previous 1 - 10 next
agent_cont
451 N. Claremo
Ave.
Templating Export
Prefix
Row Template
{
"jsonmodel_type": "agent_corporate_entity",
"agent_contacts": [
{
"jsonmodel_type":"agent_contact",
"telephones": [
{
"jsonmodel_type":"telephone",
"number_type":"cell",
"number" : {{jsonize(cells["agent_contact_tel
"ext" : {{jsonize(cells["agent_contact_teleph
}
]
"name" : {{jsonize(cells["agent_contact_name"].va
"address_1" : {{jsonize(cells["agent_contact_addr
],
"notes": [
],
"names": [
{
"authority_id" : "n000024",
"jsonmodel_type": "name_corporate_entity",
"qualifier" : null,
"source" : "Local sources",
"rules" : null,
"authorized": false,
"is_display_name": false,
"sort_name_auto_generate": true,
"primary_name" : "Illinois Institute of Technol
"subordinate_name_1" : null,
"subordinate_name_2" : null,
"number" : null,
"sort_name" : null,
"description_note" : null,
"description_citation" : null
}
],
"agent_type" : "agent_corporate_entity",
}
,
{
"jsonmodel_type": "agent_corporate_entity",
"agent_contacts": [
Row Separator
,
Suffix
Reset Template
Export
Cancel

Conclusion

- Structured data = cleanup and migration is possible!
- OpenRefine for cleaning, reconciliation, JSON export



	Record Type	Agent Name	Authority ID	Source	Rules		
☐	Corporate Entity	Alion Science and Technology, Inc.	http://id.loc.gov/authorities/names/no2010153047	NACO Authority File		Created by greer 2016-04-05 13:41:49 -0500 Last Modified by greer 2017-01-14 17:43:50 -0600	View Edit
☐	Corporate Entity	Alpha Phi Omega	http://id.loc.gov/authorities/names/n85306914	NACO Authority File		Created by greer 2016-09-13 12:32:39 -0500 Last Modified by greer 2017-01-14 17:55:38 -0600	View Edit
☐	Person	Alschuler, Alfred S. (Alfred Samuel), 1876-1940	http://id.loc.gov/authorities/names/nr90000030	NACO Authority File		Created by greer 2016-09-13 12:04:02 -0500 Last Modified by greer 2017-01-14 17:56:20 -0600	View Edit
☐	Person	Altschuler, Dori	n000001	Local sources	Resource Description and Access	Created by greer 2016-02-23 14:04:16 -0600 Last Modified by sebastian 2016-05-05 13:43:36 -0500	View Edit
☐	Corporate Entity	Alumni Memorial Hall (Chicago, Ill.)	n000343	Local sources		Created by greer 2016-09-13 12:32:41 -0500 Last Modified by greer 2016-11-01 16:36:04 -0500	View Edit
☐	Corporate Entity	American Association of Retired Persons	http://id.loc.gov/authorities/names/n79113955	NACO Authority File		Created by greer 2016-04-05 15:34:48 -0500 Last Modified by	View Edit

Resources

- Free Your Metadata: http://freeyourmetadata.org/
- OpenRefine Wiki: https://github.com/OpenRefine/OpenRefine
- Video tutorials: http://openrefine.org

Thank you!

Greer Martin

gmartin5@luc.edu