Cleaning up legacy metadata for ETDs:
Strategies, tools and a look into the future

Xiping Liu, Albert Duran, Anne Washington
UNIVERSITY of HOUSTON | LIBRARIES

Background

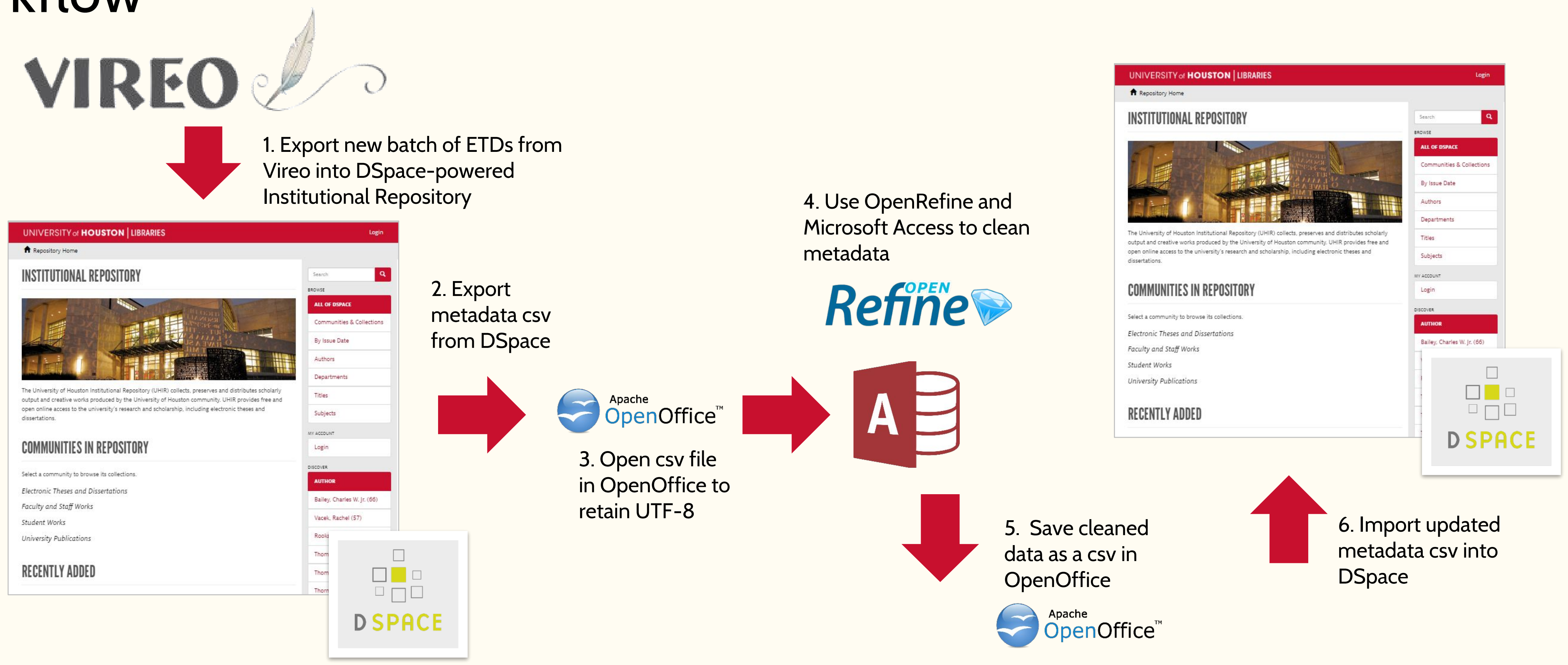
Since July 2015, the University of Houston (UH) Libraries Metadata and Digitization Services (MDS) department has collaborated with the Digital Repository Services (DRS) department to improve the quality of legacy Electronic Thesis and Dissertation (ETD) metadata in the UH Institutional Repository. The team developed workflows and strategies that improved the ETD metadata quality, strengthened the communication and collaboration between the DRS department and the MDS Metadata Unit, and provided insight into opportunities for future development.

ETD Metadata Upgrade Goals

- Align metadata with Texas Digital Library (TDL) standards*
- Correct department and discipline names
- Standardize advisor and committee member names
- Normalize date fields
- Remove broken URLS
- Develop sustainable workflow for bi-annual additions to DSpace

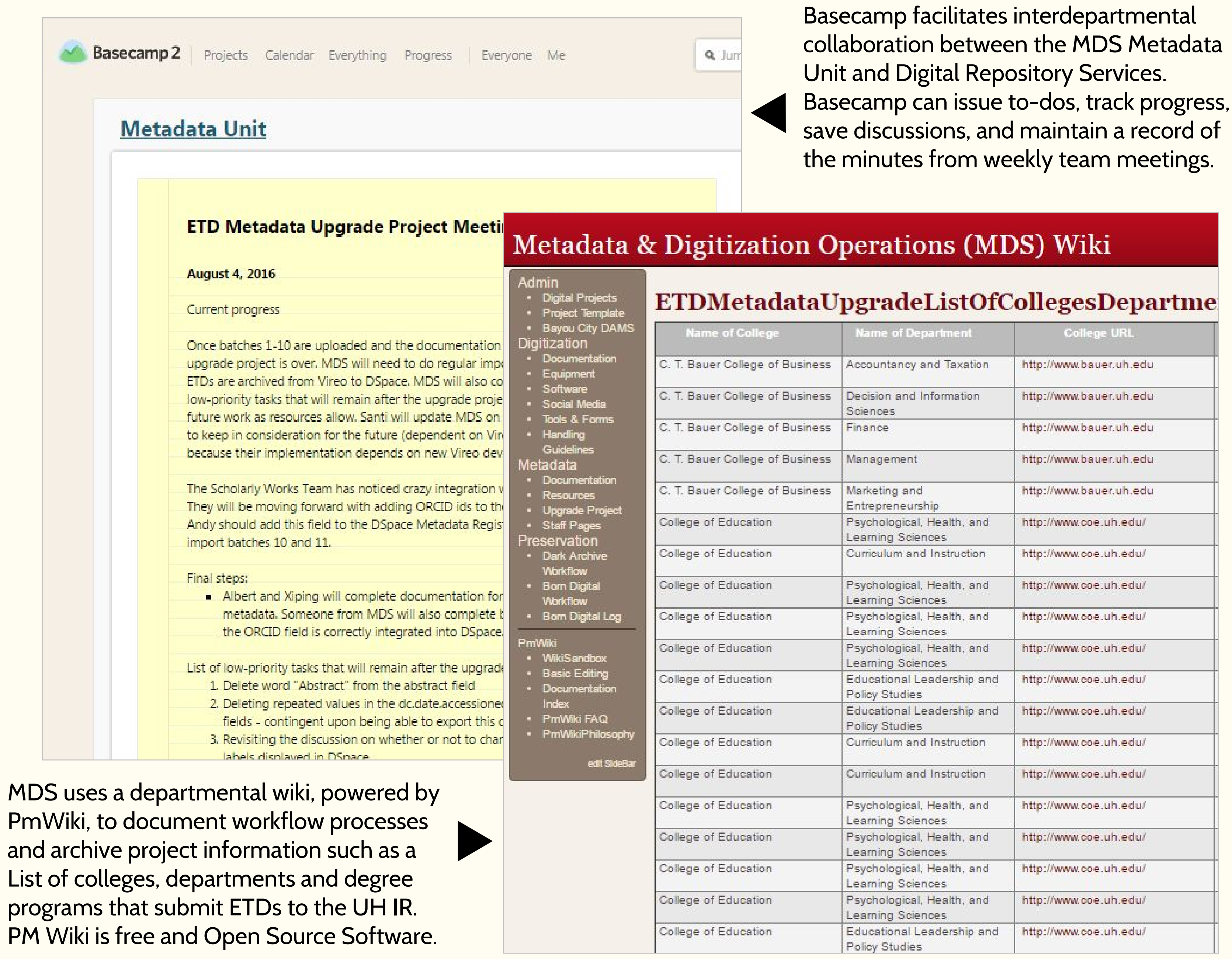
*http://hdl.handle.net/2249.1/68437

Workflow



1. Export new batch of ETDs from Vireo into DSpace-powered Institutional Repository
2. Export metadata csv from DSpace
3. Open csv file in OpenOffice to retain UTF-8
4. Use OpenRefine and Microsoft Access to clean metadata
5. Save cleaned data as a csv in OpenOffice
6. Import updated metadata csv into DSpace

Communication and Documentation



Basecamp 2 | Projects Calendar Everything Progress | Everyone Me

Metadata Unit

ETD Metadata Upgrade Project Meeti

August 4, 2016

Current progress

Once batches 1-10 are uploaded and the documentation
upgrade project is over. MDS will need to do regular imp
ETDs are archived from Vireo to DSpace. MDS will also co
low-priority tasks that will remain after the upgrade proje
future work as resources allow. Santi will update MDS on
to keep in consideration for the future (dependent on Vir
because their implementation depends on new Vireo dev

The Scholarly Works Team has noticed crazy integration v
They will be moving forward with adding ORCID ids to th
Andy should add this field to the DSpace Metadata Regis
import batches 10 and 11.

Final steps:

- Albert and Xiping will complete documentation for metadata. Someone from MDS will also complete t the ORCID field is correctly integrated into DSpace

List of low-priority tasks that will remain after the upgrad
1. Delete word "Abstract" from the abstract field
2. Deleting repeated values in the dc.date.accessione fields - contingent upon being able to export this c
3. Revisiting the discussion on whether or not to char labels displayed in DSpace

Metadata & Digitization Operations (MDS) Wiki

Admin
- Digital Projects
- Project Template
- Bayou City DAMS

Digitization
- Documentation
- Equipment
- Software
- Social Media
- Tools & Forms
- Handling Guidelines

Metadata
- Documentation
- Resources
- Upgrade Project
- Staff Pages

Preservation
- Dark Archive Workflow
- Born Digital Workflow
- Born Digital Log

PmWiki
- WikiSandbox
- Basic Editing
- Documentation Index
- PmWiki FAQ
- PmWikiPhilosophy

edit SideBar

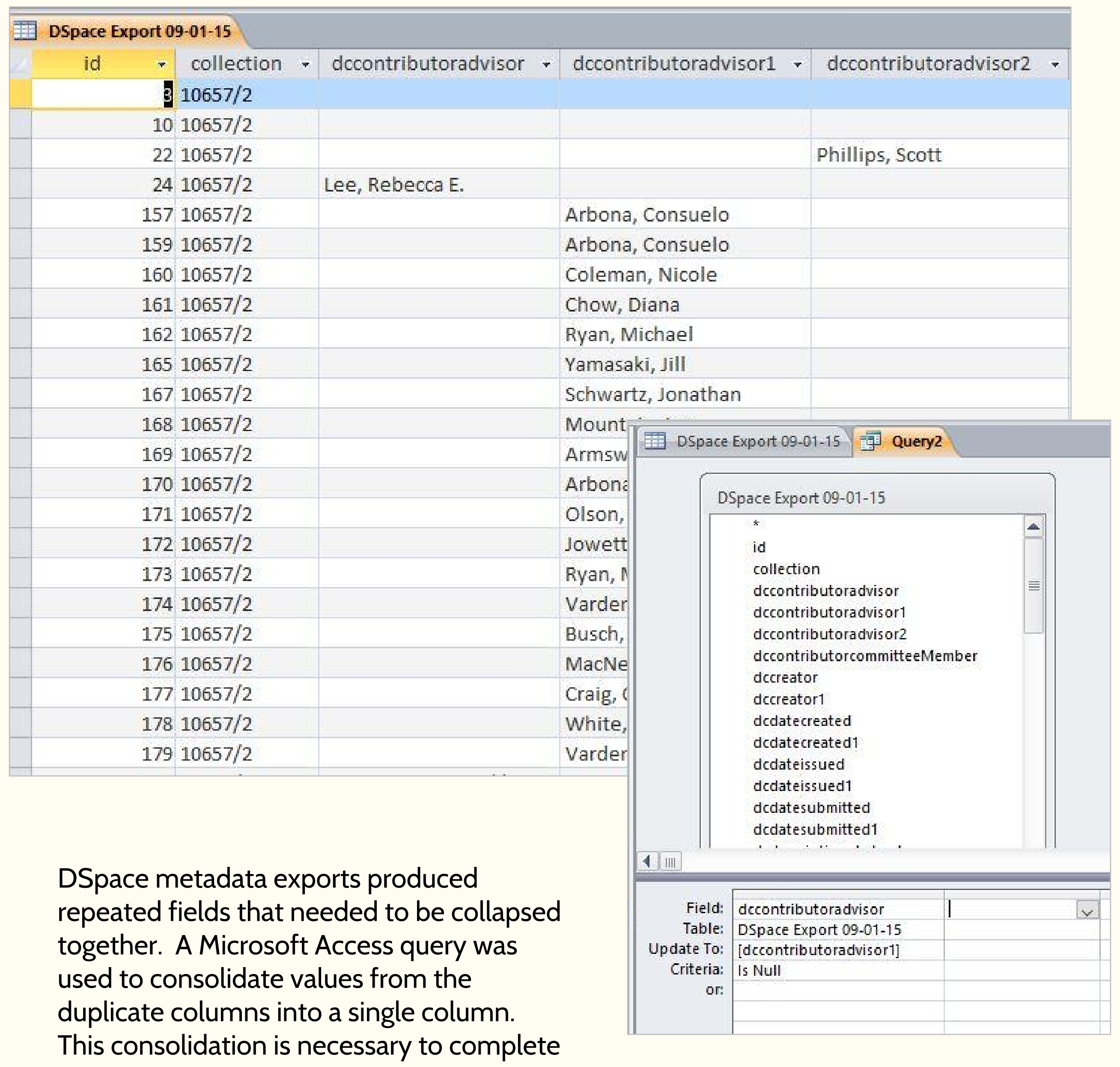
ETDMetadataUpgradeListOfCollegesDepartme

Name of College	Name of Department	College URL
C. T. Bauer College of Business	Accountancy and Taxation	http://www.bauer.uh.edu
C. T. Bauer College of Business	Decision and Information Sciences	http://www.bauer.uh.edu
C. T. Bauer College of Business	Finance	http://www.bauer.uh.edu
C. T. Bauer College of Business	Management	http://www.bauer.uh.edu
C. T. Bauer College of Business	Marketing and Entrepreneurship	http://www.bauer.uh.edu
College of Education	Psychological, Health, and Learning Sciences	http://www.coe.uh.edu/
College of Education	Curriculum and Instruction	http://www.coe.uh.edu/
College of Education	Psychological, Health, and Learning Sciences	http://www.coe.uh.edu/
College of Education	Psychological, Health, and Learning Sciences	http://www.coe.uh.edu/
College of Education	Psychological, Health, and Learning Sciences	http://www.coe.uh.edu/
College of Education	Educational Leadership and Policy Studies	http://www.coe.uh.edu/
College of Education	Educational Leadership and Policy Studies	http://www.coe.uh.edu/
College of Education	Curriculum and Instruction	http://www.coe.uh.edu/
College of Education	Curriculum and Instruction	http://www.coe.uh.edu/
College of Education	Psychological, Health, and Learning Sciences	http://www.coe.uh.edu/
College of Education	Psychological, Health, and Learning Sciences	http://www.coe.uh.edu/
College of Education	Psychological, Health, and Learning Sciences	http://www.coe.uh.edu/
College of Education	Psychological, Health, and Learning Sciences	http://www.coe.uh.edu/
College of Education	Educational Leadership and Policy Studies	http://www.coe.uh.edu/

Basecamp facilitates interdepartmental collaboration between the MDS Metadata Unit and Digital Repository Services. Basecamp can issue to-dos, track progress, save discussions, and maintain a record of the minutes from weekly team meetings.

MDS uses a departmental wiki, powered by PmWiki, to document workflow processes and archive project information such as a List of colleges, departments and degree programs that submit ETDs to the UH IR. PM Wiki is free and Open Source Software.

Remediation



DSpace Export 09-01-15

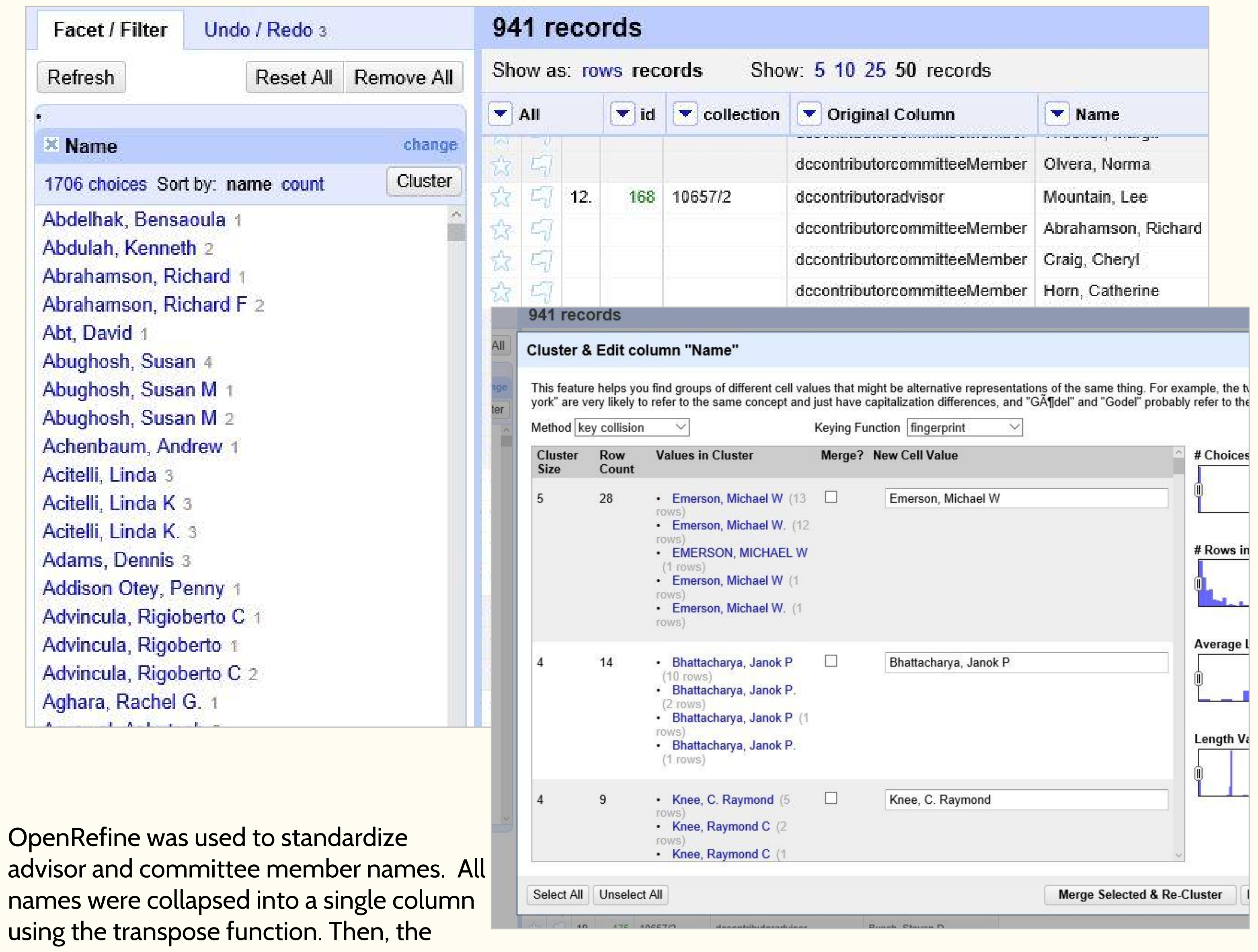
id	collection	dccontributoradvisor	dccontributoradvisor1	dccontributoradvisor2
3	10657/2			
10	10657/2			
22	10657/2			Phillips, Scott
24	10657/2	Lee, Rebecca E.		
157	10657/2		Arbona, Consuelo	
159	10657/2		Arbona, Consuelo	
160	10657/2		Coleman, Nicole	
161	10657/2		Chow, Diana	
162	10657/2		Ryan, Michael	
165	10657/2		Yamasaki, Jill	
167	10657/2		Schwartz, Jonathan	
168	10657/2		Mount	
169	10657/2		Armsw	
170	10657/2		Arbona	
171	10657/2		Olson,	
172	10657/2		Jowett	
173	10657/2		Ryan, M	
174	10657/2		Varder	
175	10657/2		Busch,	
176	10657/2		MacNe	
177	10657/2		Craig, (	
178	10657/2		White,	
179	10657/2		Varder	

DSpace Export 09-01-15 | Query2

DSpace Export 09-01-15
*
id
collection
dccontributoradvisor
dccontributoradvisor1
dccontributoradvisor2
dccontributorcommitteeMember
dccreator
dccreator1
dcdatecreated
dcdatecreated1
dcdateissued
dcdateissued1
dcdatesubmitted
dcdatesubmitted1

Field: dccontributoradvisor
Table: DSpace Export 09-01-15
Update To: [dccontributoradvisor1]
Criteria: Is Null
or:

DSpace metadata exports produced repeated fields that needed to be collapsed together. A Microsoft Access query was used to consolidate values from the duplicate columns into a single column. This consolidation is necessary to complete additional remediation tasks and import the new metadata successfully into DSpace.



Name
1706 choices Sort by: name count

- Abdelhak, Bensaoula 1
- Abdulah, Kenneth 2
- Abrahamson, Richard 1
- Abrahamson, Richard F 2
- Abt, David 1
- Abughosh, Susan 4
- Abughosh, Susan M 1
- Abughosh, Susan M 2
- Achenbaum, Andrew 1
- Acitelli, Linda 3
- Acitelli, Linda K 3
- Acitelli, Linda K. 3
- Adams, Dennis 3
- Addison Otey, Penny 1
- Advincula, Rigioberto C 1
- Advincula, Rigoberto 1
- Advincula, Rigoberto C 2
- Aghara, Rachel G. 1

941 records

Show as: rows records Show: 5 10 25 50 records

All	id	collection	Original Column	Name
			dccontributorcommitteeMember	Olvera, Norma
12.	168	10657/2	dccontributoradvisor	Mountain, Lee
			dccontributorcommitteeMember	Abrahamson, Richard
			dccontributorcommitteeMember	Craig, Cheryl
			dccontributorcommitteeMember	Horn, Catherine

Cluster & Edit column "Name"

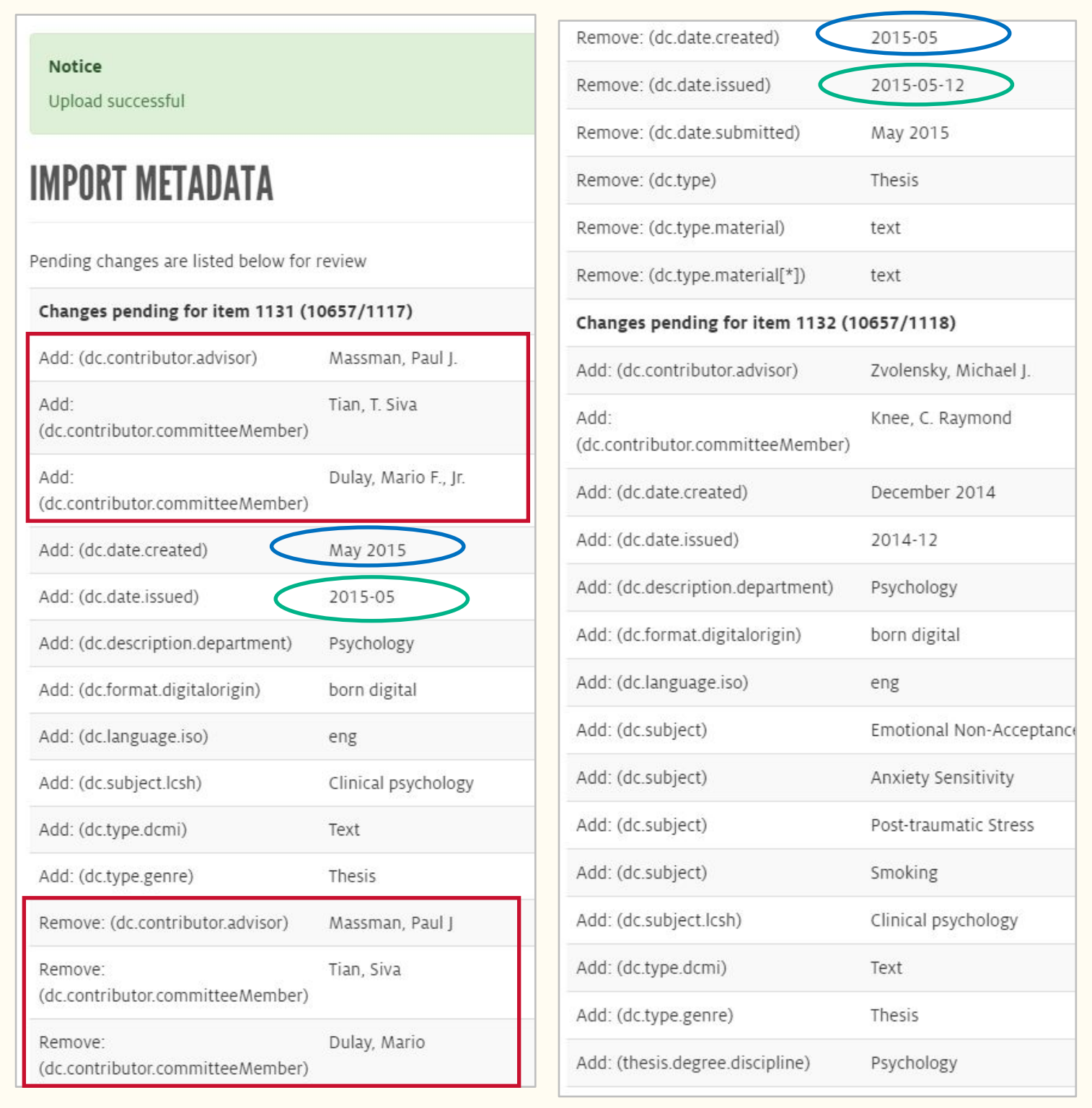
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the t york" are very likely to refer to the same concept and just have capitalization differences, and "GÃ¶del" and "Godel" probably refer to the

Method: key collision Keying Function: fingerprint

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
5	28	Emerson, Michael W (13 rows) Emerson, Michael W. (12 rows) EMERSON, MICHAEL W (1 rows) Emerson, Michael W (1 rows) Emerson, Michael W. (1 rows)	□	Emerson, Michael W
4	14	Bhattacharya, Janok P (10 rows) Bhattacharya, Janok P. (2 rows) Bhattacharya, Janok P (1 rows) Bhattacharya, Janok P. (1 rows)	□	Bhattacharya, Janok P
4	9	Knee, C. Raymond (5 rows) Knee, Raymond C (2 rows) Knee, Raymond C (1	□	Knee, C. Raymond

OpenRefine was used to standardize advisor and committee member names. All names were collapsed into a single column using the transpose function. Then, the facet and cluster functions were used to choose and populate the preferred name throughout the dataset.

Re-import



Notice
Upload successful

IMPORT METADATA

Pending changes are listed below for review

Changes pending for item 1131 (10657/1117)

Add: (dc.contributor.advisor)	Massman, Paul J.
Add: (dc.contributor.committeeMember)	Tian, T. Siva
Add: (dc.contributor.committeeMember)	Dulay, Mario F., Jr.
Add: (dc.date.created)	May 2015
Add: (dc.date.issued)	2015-05
Add: (dc.description.department)	Psychology
Add: (dc.format.digitalorigin)	born digital
Add: (dc.language.iso)	eng
Add: (dc.subject.lcsh)	Clinical psychology
Add: (dc.type.dcmi)	Text
Add: (dc.type.genre)	Thesis
Remove: (dc.contributor.advisor)	Massman, Paul J
Remove: (dc.contributor.committeeMember)	Tian, Siva
Remove: (dc.contributor.committeeMember)	Dulay, Mario
Remove: (dc.date.created)	2015-05
Remove: (dc.date.issued)	2015-05-12
Remove: (dc.date.submitted)	May 2015
Remove: (dc.type)	Thesis
Remove: (dc.type.material)	text
Remove: (dc.type.material[*])	text

Changes pending for item 1132 (10657/1118)

Add: (dc.contributor.advisor)	Zvolensky, Michael J.
Add: (dc.contributor.committeeMember)	Knee, C. Raymond
Add: (dc.date.created)	December 2014
Add: (dc.date.issued)	2014-12
Add: (dc.description.department)	Psychology
Add: (dc.format.digitalorigin)	born digital
Add: (dc.language.iso)	eng
Add: (dc.subject)	Emotional Non-Acceptanc
Add: (dc.subject)	Anxiety Sensitivity
Add: (dc.subject)	Post-traumatic Stress
Add: (dc.subject)	Smoking
Add: (dc.subject.lcsh)	Clinical psychology
Add: (dc.type.dcmi)	Text
Add: (dc.type.genre)	Thesis
Add: (thesis.degree.discipline)	Psychology

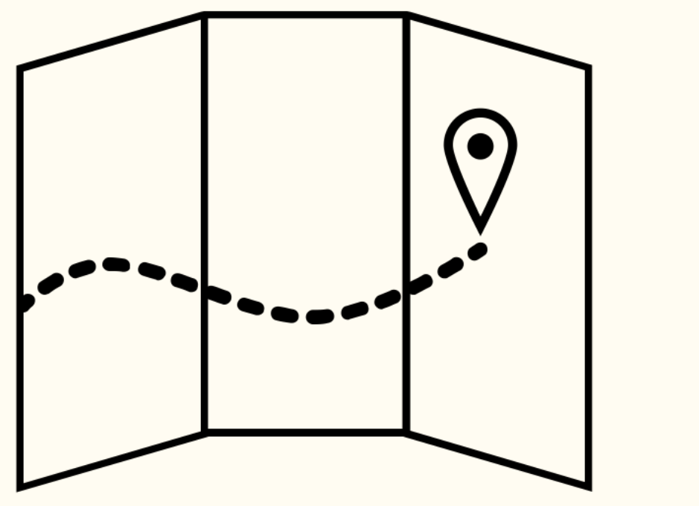
Once all of the metadata clean-up tasks were complete, the updated metadata csv files were imported into DSpace and records were updated.

Where Are We Now?

✓941 corrected records

✓Workflow for new batches

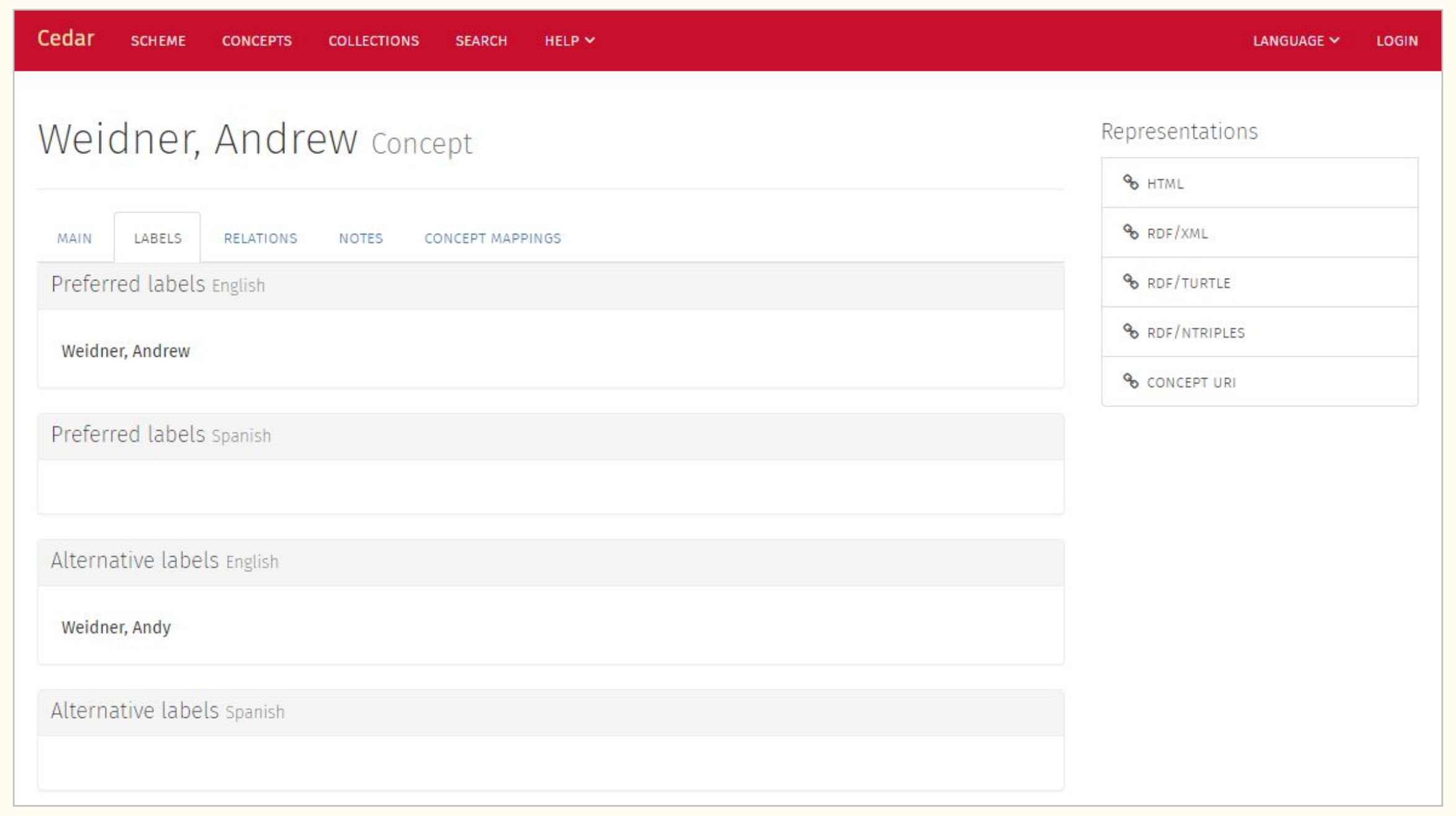
✓Corrections to Vireo



Map by Karsten Barnett from the Noun Project

Next Steps

- Add names to local thesaurus
- Workflow efficiencies
- Data sharing



Weidner, Andrew Concept

Preferred labels English
Weidner, Andrew

Preferred labels Spanish

Alternative labels English
Weidner, Andy

Alternative labels Spanish

Representations