Automating Controlled Vocabulary Reconciliation

Anna Neatrour

Metadata Librarian

Jeremy Myntti

Interim Head, Digital Library Services

Project Overview

Summary

- Metadata inconsistency
- Overview of vendor authority process
- Further work with Open Refine
- Next steps

http://www.utahindians.org

Inconsistency

Gosiute Indians
Goshute Indians

Navajo Indians
Navaho Indians

Salt Lake
Salt Lake City
Salt Lake City (Utah)

Bishop, Dail Stapley
Bishop, Dale Stapely
Bishop, Dale Stapley

Beckwith, Frank A. (1876-1951)
Beckwith, Frank Asahel (1876-1951)
Beckwith, Frank A.
Beckwith, Frank A. (1876-1951)
Beckwith, Frank Asahel (1876-1951)
Beckwith, Frank Asahel, 1876-1951

Woven basket or jug;
http://content.lib.utah.edu/cdm/ref/collection/UU_Photo_Archives/id/13887

Project Timeline

June-Sept. 2012 – Define project

Oct. 2012 – May 2013 – Testing

June 2013 – Contracted with Backstage Library Works

June 2013-Feb. 2014 – Continued testing

Feb.-May 2014 – 17 collections processed

June-Aug. 2014 – Manual review (intern)

April 2015-today – Explore OpenRefine

Methodology

http://content.lib.utah.edu/cdm/ref/collection/uaida/id/14697

<title>A group of St. George (Sibwit) Paiutes and Wickiups (cedar)</title>

<subjec>Paiute Indians; Ute Indians--History; Wickiups; Indians of North America--Dwellings;</subject>

<covspa>Utah;</covspa>

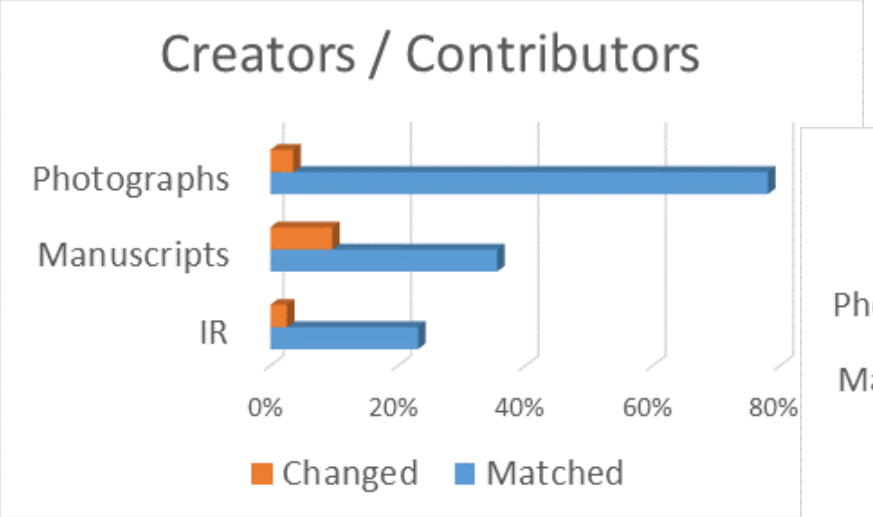
<descri>A group of people sitting and standing in front of a brush shelter;<descri>

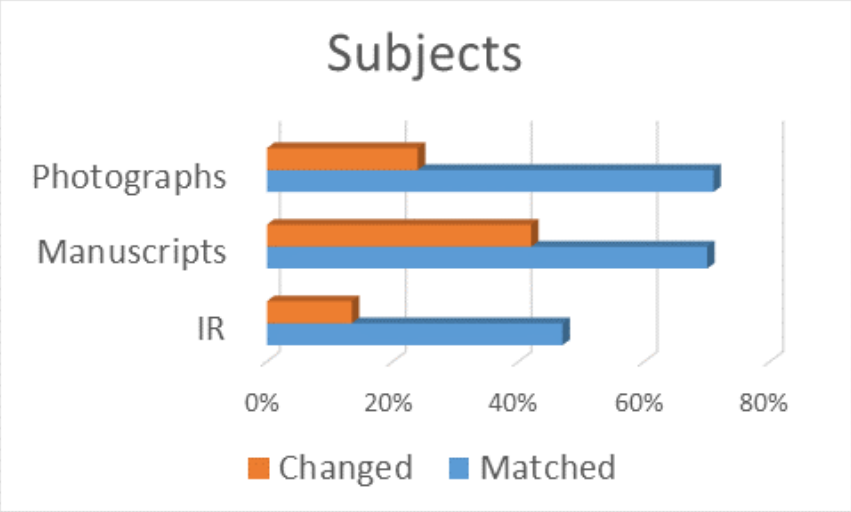
<publis>Digitized by: J. Willard Marriott Library, University of Utah;</publis>

<type>Image;StillImage;</type>

<format>image/jpeg;</format>

Backstage: statistics and reports





Unmatched report

7 records	creato	Albin, William M.
1 record	creato	Allen, Rufus C.
1 record	creato	Allred, William J.
8 records	creato	Alter, Cecil J.
2 records	creato	American West Center
15 records	creato	Arentz, Bob
1 record	creato	Arnold Stone, Elizabeth
1 record	creato	Arrowpeen

Change report

2 records	Old	creato	Leupp, Francis Ellington, 1849-1918
	New	creato	Leupp, Francis E. (Francis Ellington), 1849-1918
321 records	Old	creato	Liebler, Harold Baxter, 1889-1982
	New	creato	Liebler, H. Baxter (Harold Baxter), 1889-1982
21 records	Old	creato	Matheson, Alva
	New	creato	Matheson, Alva L.

Backstage: standardization

Capitalization, Punctuation, and Updated Authorized Access Points

Forests and Forestry – Utah

forests and forestry -- Utah

Forest lands - Utah

Forests and forestry--Utah

A group of Navajos at Navajo Mountain government school; http://content.lib.utah.edu/cdm/ref/collection/uaida/id/43551

Backstage: problems encountered

Missing MARC tags

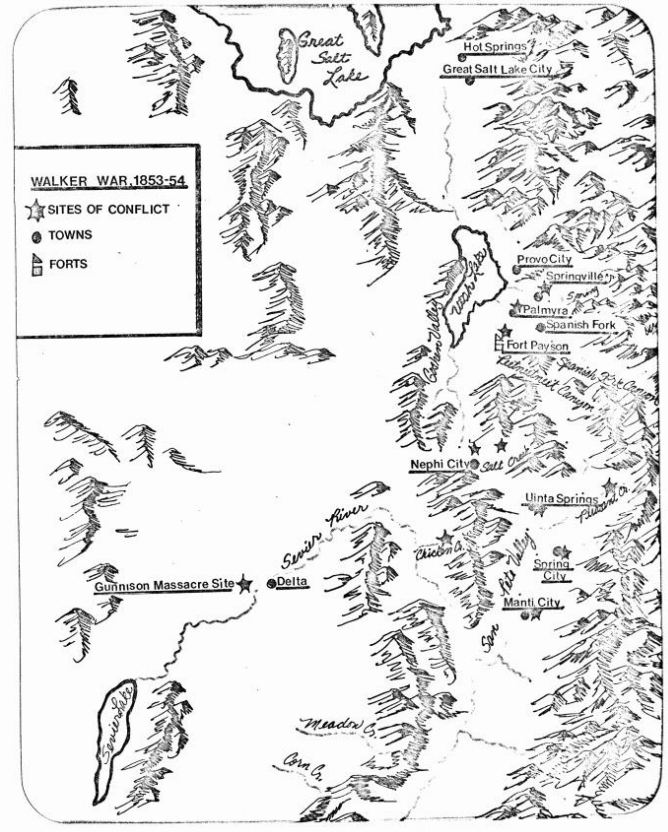
- Names treated as topical headings and vice versa
  - Provo => Provisional IRA

Data in wrong fields

- Date: Price Hiram, 1814-1901

Incorrect match

- Local names matching wrong records
  - Johnson, Abe is not Johnson, F.T.



Walker War Map 1853-1854; http://content.lib.utah.edu/cdm/ref/collection/uaida/id/15474

Intern review and clean-up

OpenRefine project

- Used UAIDA as a pilot, since it had the greatest number of unmatched names due to the size of the collection (over 8,000 items)
- 529 unmatched names after Backstage process

Navajo woman weaving,
http://content.lib.utah.edu/cdm/ref/collection/uaida/id/45379

OpenRefine: two approaches

- Reconciliation process developed by Jenn Wright and Matt Carruthers, University of Michigan Library, https://github.com/mcarruthers/LCNAF-Named-Entity-Reconciliation
- Reconciliation process developed by Roderic Page, http://iphylo.blogspot.com/2013/04/reconciling-author-names-using-open.html

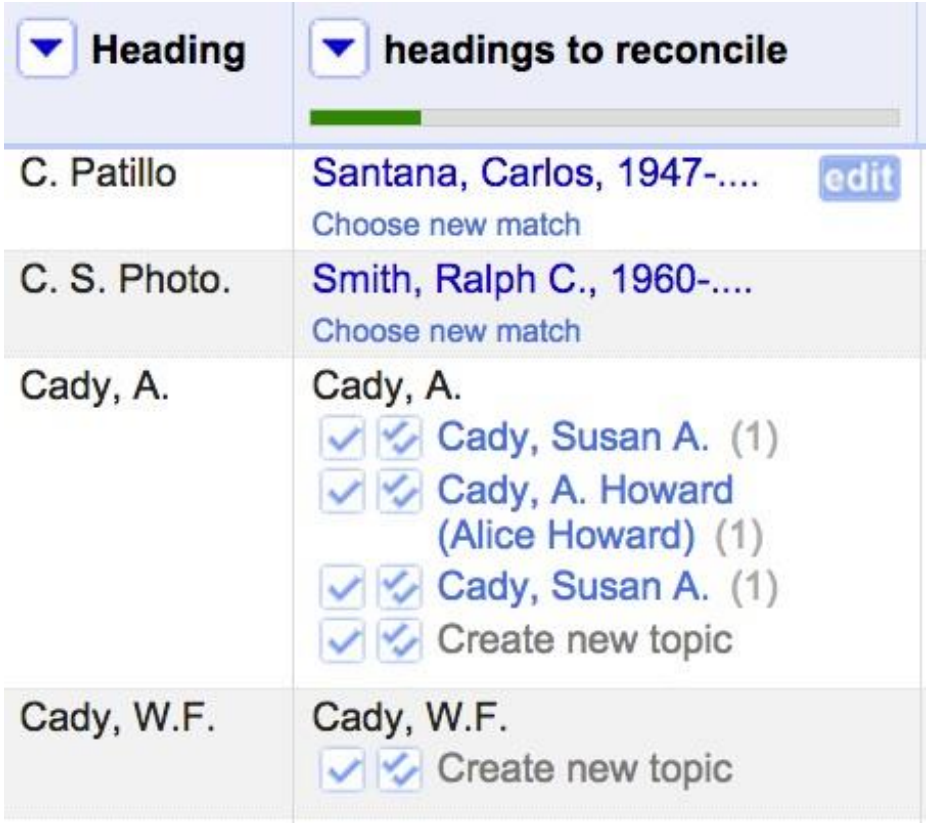
A group of Navajo children and teenagers, http://content.lib.utah.edu/cdm/ref/collection/uaida/id/43285

OpenRefine: differences in results

- Both processes found name matches through searching VIAF.
  - Wright and Carruthers' process looked for a matching LC authority record in the VIAF cluster
    - 81 records were matched, 132 were false matches, and 312 number had no match
  - Page's process matched names to authors in a more general fashion
    - 70 records were matched, 37 were false matches, and 449 had no match.

OpenRefine: manual work

- Check matches against collection and discard false matches



Heading	headings to reconcile
C. Patillo	Santana, Carlos, 1947-.... edit Choose new match
C. S. Photo.	Smith, Ralph C., 1960-.... Choose new match
Cady, A.	Cady, A. Cady, Susan A. (1) Cady, A. Howard (Alice Howard) (1) Cady, Susan A. (1) Create new topic
Cady, W.F.	Cady, W.F. Create new topic

OpenRefine: manual work

- Wright and Carruthers process had 262 undo/redo actions in the open refine project
- Page's reconciliation process resulted in 424 undo/redo actions
  - Could have refined initial results further

Eight Hopi Baskets,
http://content.lib.utah.edu/cdm/ref/collection/uaida/id/45009

Open Refine: student work

- Fall 2015 – student ran additional unmatched items through OpenRefine with Wright & Carruthers process
- Metadata Librarian currently reviewing student work and updating collections

Next Steps

Create local and regional controlled vocabularies

Reconcile across more collections

- CONTENTdm metadata exported in SOLR
- Easier to get list of personal names across all collections
- Explore other reconciliation methods

```
"/uaida","Browning, Daniel M.;"
"/uaida","Holeman, Jacob Harrod, 1793-1857;"
"/uaida","Babbit, A.W.; Lea, Luke;"
"/uaida","Lea, Luke;"
"/uaida","Young, Brigham, 1801-1877;"
"/uaida","Day, H.R.;"
"/uaida","Holeman, Jacob Harrod, 1793-1857;"
"/uaida","Rose, Stephen B. ;"
"/uaida","Young, Brigham, 1801-1877;"
"/uaida","Young, Brigham, 1801-1877;"
"/uaida","Holeman, Jacob Harrod, 1793-1857;"
"/uaida","Young, Brigham, 1801-1877; Lea, Luke;"
"/uaida","Holeman, Jacob Harrod, 1793-1857;"
"/uaida","Holeman, Jacob Harrod, 1793-1857;"
"/uaida","Young, Brigham, 1801-1877;"
"/uaida","Young, Brigham, 1801-1877;"
```

Next Steps

URIs in CONTENTdm, MWDL (Primo), and DPLA

Title	Three people outside a teepee;
Subject	St. Christopher's Mission to the Navajo (Bluff, Utah), http://id.loc.gov/authorities/names/nr91019909; Liebler, Harold Baxter (Harold Baxter), 1889-1982, http://id.loc.gov/authorities/names/n94048511; Tipis, http://id.loc.gov/authorities/subjects/sh94002186;
Creator	Liebler, H. Baxter (Harold Baxter), 1889-1982, http://id.loc.gov/authorities/names/n94048511;
Type	Image;StillImage, http://purl.org/dc/dcmitype/StillImage;
Language	eng, http://id.loc.gov/vocabulary/languages/eng;
Coverage	Bluff, San Juan County, Utah, United States, http://www.geonames.org/5535638/;

http://content.lib.utah.edu/cdm/ref/collection/uaida/id/43183

Questions

Anna Neatrour | anna.neatrour@utah.edu
Metadata Librarian

Jeremy Myntti | jeremy.myntti@utah.edu
Interim Head, Digital Library Services