Reconciling Archival Data with Library of Congress Subject Headings

Claire Marshall | Metadata Migration Archivist | cmarshall@smith.edu

Smith College Libraries

Goal

The goal of this project was to reconcile the subjects in ArchivesSpace against Library of Congress Subject Headings (LCSH) in order to improve their accuracy and research value. The project aimed to add LCSH authority ids and authorized phrasing to ArchivesSpace subject records wherever possible, as well as to standardize punctuation and subject construction.

Background

Having been migrated to ArchivesSpace from a previous content management system, our subject data inherited problems resulting from long-standing inconsistencies in manual entry, e.g., non-standard punctuation, complex subjects that existed as a single string instead of being broken into individual terms, few or uncertain authority ids, etc. In short, we did not know how many subjects were actually valid according to Library of Congress.

Solution

Automation to the rescue! In order to reach our goal, I used a combination of Python scripts and SQL queries, in addition to web scraping, to programmatically reconcile our subject data. The most important of these tools are located in our GitHub repository: https://github.com/smith-special-collections/a2c-tools.

Process

Because some subjects would have direct matches in LCSH and others would only have partial matches (i.e., the heading and one or more subdivisions might be matched with some part(s) remaining unmatched), two slightly different approaches were needed.

Challenges

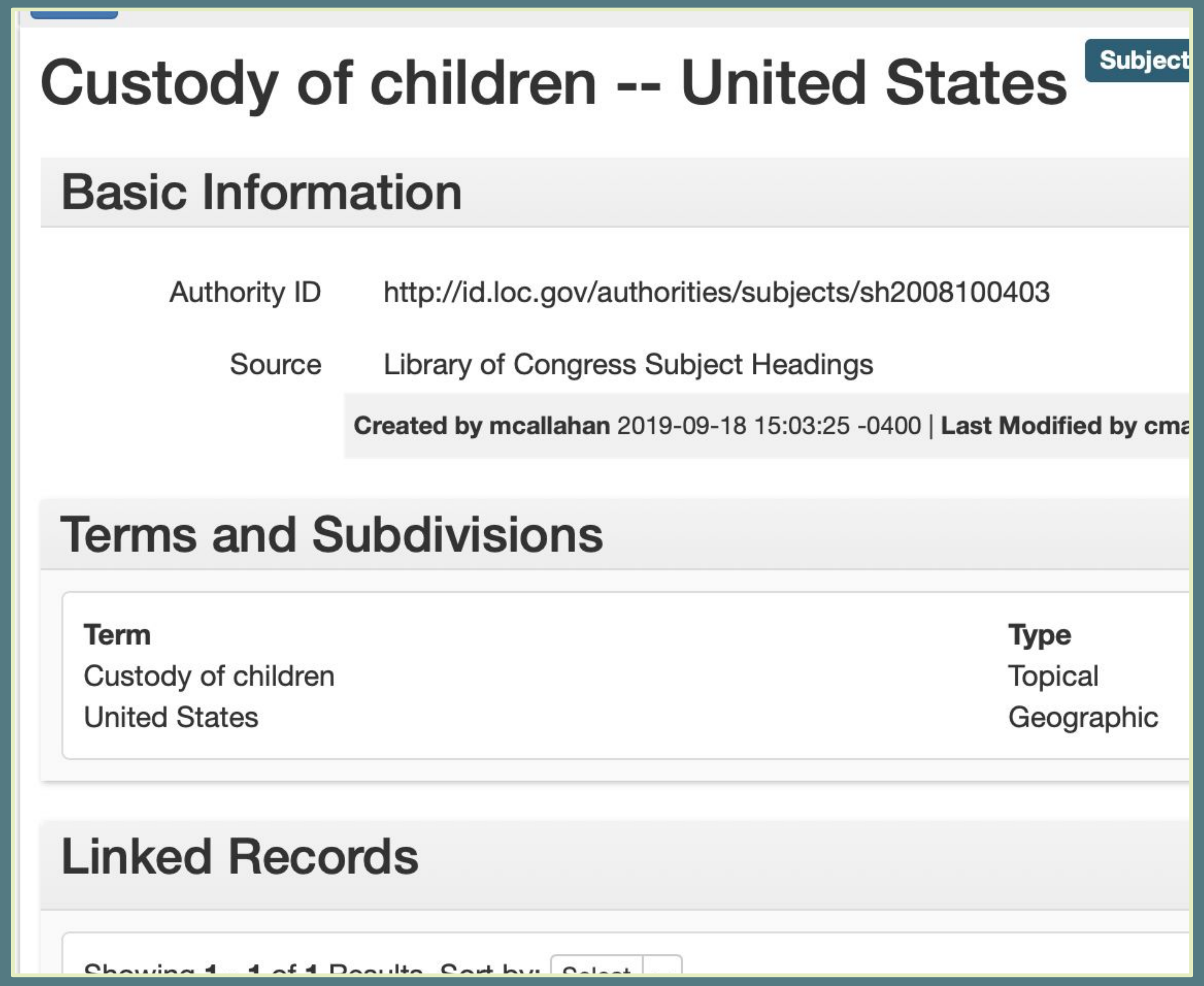
- Library of Congress subject heading data exists in extremely large files that have a complicated nested structure
- There is no separate list for Library of Congress subdivisions
- ArchivesSpace subject record schema requires controlled type values for each individual term in a complex subject. These types do not exist within the Library of Congress data file
- Traditional Python fuzzy matching is extremely slow when working with large data sets

Numbers

Total subjects before project	7,244
Subjects with terms broken down and term types matched to LCSH	3,255
Subjects matched directly to LCSH	1,745
Subjects with External Document subrecords added	2,625

Subjects Matched Directly

- Extract subject data from LCSH JSON file and transform it into a spreadsheet containing authority id and subject term
- Create local Postgresql database
- Perform fast, fuzzy matching against database
- Construct URL from matched LCSH authority id to query LOC website for each subject's individual web page and scrape its metadata for the term type
- Update ArchivesSpace subject records with matched data and break complex subjects into individual terms via the API



Custody of children -- United States

Basic Information

Authority ID http://id.loc.gov/authorities/subjects/sh2008100403

Source Library of Congress Subject Headings

Created by mcallahan 2019-09-18 15:03:25 -0400 | **Last Modified by cma**

Terms and Subdivisions

Term	Type
Custody of children	Topical
United States	Geographic

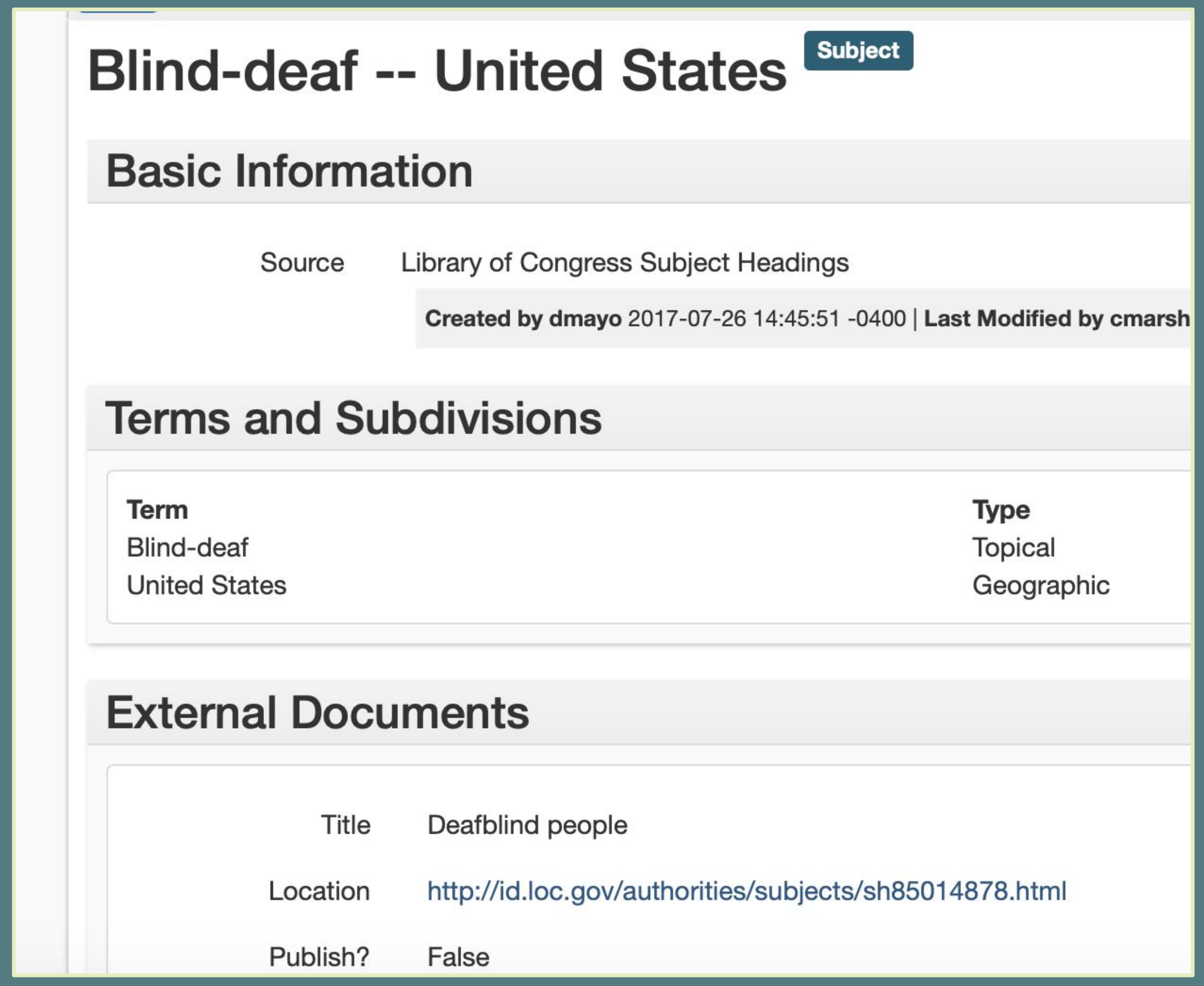
Linked Records

Showing 1 - 1 of 1 Results Sort by: Select

Standardized phrasing, terms broken down with type, and full LOC URI

Subjects Matched by Term

- Break complex subjects into individual terms and match against Postgresql database
- Scrape LOC website for subdivisions and their authority ids and create second local database for subdivision matching
- Update ArchivesSpace records with External Document subrecords containing term and authority id if matched
- Construct URL queries for each subject and extract term type from HTML
- Update subject records in ArchivesSpace with term type data via the API



Blind-deaf -- United States

Basic Information

Source Library of Congress Subject Headings

Created by dmayo 2017-07-26 14:45:51 -0400 | **Last Modified by cmarsh**

Terms and Subdivisions

Term	Type
Blind-deaf	Topical
United States	Geographic

External Documents

Title Deafblind people

Location http://id.loc.gov/authorities/subjects/sh85014878.html

Publish? False

External Document subrecord with preferred phrasing and full LOC URI