How I Programmatically Reconciled Archive Data with Library of Congress

CLAIRE MARSHALL

Abstract: This project sought to gain greater control over our subject data within ArchivesSpace by reconciling it with Library of Congress Subject Headings. Legacy data migrated from a previous CMS was inconsistently or improperly formatted. Furthermore, few subjects had authority ids mapped to LCSH, and for those that did have them, we were unsure of their accuracy. A programmatic solution was sought that would bulk reconcile and update our subjects with accurate data. I used a combination of Python scripting, SQL queries, web scraping, and the ArchivesSpace API to pull LCSH data and match it against our own. The biggest hurdle I faced over the course of this project was having access to an accurate and easily parsable data set of subject headings. Ultimately, I had to "create" my own in the form of a local database, which presented a convenient means of storing and reusing data as needed. In total, this project reconciled and updated several thousand subjects and took roughly three months to complete.

About the author:
Claire Marshall is the Metadata Migrations Archivist for the Access to Collections project at Smith College, a role she has held since 2018. She did her undergraduate education in History and Italian at Sarah Lawrence College, before graduating with a Master of Arts in History from Marquette University and with a Master of Science in Information from the University of Michigan. She is interested in how programming and automation can be utilized within archival contexts to relieve the burden of tedious manual labor and to facilitate greater access to collections information online.