| | |
|---|---|
| **Organization:** University of Pennsylvania<br><br>**Project title:** Semi-automatically assigning keywords to medieval manuscripts on OPenn | |
| **Primary mentor:** Dot Porter, Curator, Digital Research Services, Schoenberg Institute for Manuscript Studies, Kislak Center for Special Collections, Rare Books and Manuscripts, University of Pennsylvania<br><br>**Supporting mentor:** Doug Emery, our Special Collections Digital Content Programmer, University of Pennsylvania--Libraries |  |

| | |
|---|---|
| **Description** | We publish our medieval manuscript descriptions on a platform called OPenn (http://openn.libraries.upenn.edu). The descriptions are TEI/XML files (converted from MARC records), and we would like to extract data from these records, map that data to keywords, and then put those keywords back into the records. This will enhance the searchability of the collection, and will also make advanced data visualization easier. |
| **Problems** | Although the fields used across the files are consistent, formatting of many terms used are inconsistent (especially things like dates and geographic descriptors) and the use of Library of Congress Subject Headings makes federated searching based directly off the TEI data a messy prospect. We have developed a set of keywords for our current digitization project (https://github.com/leoba/bibliophilly-keywords) and we need to assign these keywords to records for manuscripts that have already been digitized. The fellow will help us map our messy data to the keywords, then put those keywords back into the records. |
| **Techniques** | Data extraction from the TEI files, data cleanup (using OpenRefine or purpose-built code), regeneration of TEI using existing systems. Development of repeatable workflow. |
| **Data** | TEI/XML, consistently formatted. List of keywords to map to. |
| **Outcome** | This project will feed into an ongoing project to improve the search providing access to OPenn (currently non-existent), and will also make visualization of the data possible in ways not possible now. |