Organization: Digital Public Library of America (DPLA)

Project title: DPLA Resources and Vocabulary Enrichment for Analytics

Primary mentor: Gretchen Gueguen

Supporting mentors: Emily Gore, Michael Della Bitta, Audrey Altman

Description (2 project options)	DPLA's data has more than 3 million unique subject headings, with only a portion coming from controlled vocabularies. Issues arise when records use slight variations of terms, or synonyms for the same concept. **Project 1:** Develop and test an effective method for analyzing record content and matching content, including keywords, with relevant controlled term from a defined list (LCSH or potentially Wikidata). Methods used will create a consistent vocabulary to aid users, and be replicable as a data source as it is re-ingested into DPLA. **Project 2**: Extends previous work focusing on the format terms in DPLA, to analyze this data and a vocabulary drawing from a select set of 30 terms from the *Art and Architecture Thesaurus* (a linked data ontology supported by the Getty Research Institute). The mechanics of this project are similar to project 1, however, the scope of the target vocabulary has already been identified.
Problems	Address challenges that exist when records use slight variations of terms, or synonyms for the same concept.
Techniques	Term extraction/marching, using NLP and potentially machine learning
Data	DPLA metadata records
Outcome	A method and an approach for identifying term variations and applying terminology more consistent terminology to support analytics