Organization: OCLC

Mentors: OCLC Membership and Research staff

- Jean Godby, Senior Research Scientist, OCLC Membership and Research
- Devon Smith, Lead Engineer

Project Title: Automatic Identification of Publisher Entities to Support Discovery and Navigation

Description: High-quality bibliographic records typically contain statements of responsibility. The MARC semantics of the relevant fields are broad and redundant, but the contents are often interpreted as names of publishers. This is valuable information for many uses of library data, but publisher names are effectively unavailable because they are represented as strings, not as URIs or other identifiers that are associated with real-world entities. A user searching a collection of library metadata will experience this shortcoming as a list of results that describe the same resource, with perhaps minor variants in spelling and punctuation that are nevertheless too substantial to be caught by duplicate detection algorithms.

The LIS research literature reports pessimistic conclusions about the prospects for disambiguating publisher names and controlling them in authority files. But this project has different aims:

- Apply data-science methods to evaluate evidence that can be used to promote names of publishers to entities; and
- Show how the results can be used to improve discovery and browsing.

Problems: Parse the statements of responsibility in MARC bibliographic records to identify publisher names and distinguish them from the names of other responsible entities. Associate names that are most likely to refer to the same entity. Use this output as evidence that bibliographic records refer to the same resource.

Techniques:

- Extract data from MARC fields, using named entity recognition and similar techniques.
- Apply similarity measures both on extracted names and their local contexts.
    - Apply string similarity metrics on the name strings to identify spelling and punctuation variants.
    - Apply semantic indexing on the bibliographic records from which the names are extracted to identify similar contexts. One possible outcome is evidence for a model that associates names of imprints with a given publisher.
- Associate extracted names with VIAF and similar resources where appropriate.
- Cluster and visualize the results. Mock up a ‘before’ and ‘after’ search result using data obtained from the study.
- Evaluate the clusters with human judges.

Data: OCLC WorldCat records

Outcomes: This project seeks to advance our understanding of the publisher entity in library bibliographic data. Can we make use of the available evidence? What are the impediments to success? Is there a subset of the problem that can be successfully automated? What changes to descriptive practice does this study support?