Organization: OCLC Research

Primary mentor: Devon Smith, Lead Engineer, OCLC Research

Project Title	**Metadata Record Similarity: Identification and Clustering**
Description	Clustering within the WorldCat bibliographic utility en masse is often computationally infeasible since the data is so large and complex. This project will focus on finding ways to "fuzzy cluster" MARC (MAchine readable cataloging) metadata records so that computationally expensive processes can be run on the resultant, smaller fuzzy clusters. In a very similar vein, when given an exemplar, quickly finding records that are similar to the exemplar will also be explored.
Problems/ Research Questions	How to quickly process large volumes of data, exclude "obviously" non-matching records, and include the most similar records.
Techniques	Any/all of string similarity metrics, match scoring, fingerprinting, locality sensitive hashing, Kolmogorov complexity. Most any approach that can be justified by some literature would be acceptable.
Tools/ Languages used	R or Python (libraries to be identified)
Data	Description: WorldCat Data Type: MARC (MAchine readable cataloging) Data Size: Sample
Outcome	An algorithm and an implementation thereof that can be used to quickly identify likely match candidates or small fuzzy clusters for further, more computationally expensive processing. This outcome will feed into several ongoing research

	initiatives at OCLC, furthering our goals of increasing the quality and utility of our data.
Milestone Timeline	1) Week 1: • Establish environment • Preliminary data analysis and exploration 2) Month 1: • Literature review • Experiment with similarity functions and clustering algorithms • Develop evaluation plan 3) Months 2-5: • Iteratively develop, experiment, and evaluate a similarity function and a clustering algorithm 4) Month 6: • Prepare and present results.
References	• MAchine readable cataloging, see: https://www.loc.gov/marc/, and explore the "Bibliographic" link, first link on left-hand, ToC (table of contents). • Worldcat, explore: https://www.worldcat.org/