Discovering Implicit Data References

Faculty supervisor: Libby Hemphill

The main goal of this project is to find ways to use natural language processing and machine learning to automatically connect research outputs (e.g., research papers) and datasets. We will use these approaches to improve the ICPSR Bibliography of Data-related Literature, a continuously-updated, searchable database that provides curated two-way links between data archived at ICPSR and nearly 75,000 published and unpublished works resulting from analyses of data distributed by ICPSR.

Currently, ICPSR staff populate the Bibliography by detecting data use in the literature and linking it back to ICPSR data. If all scholarly publications included formal data citations with persistent identifiers, ICPSR could take advantage of indexing scripts to find a near-complete list of primary and secondary-use publications. But most authors do not cite data formally. Limiting our collection of related publications to only those in which ICPSR study DOIs were used would mean all data use that was not cited properly would go unlinked and uncounted. Two staff members and one part-time research assistant currently conduct manual searches for explicit and informal mentions of data use in the full text of resources provided by electronic social science literature databases, e.g., ProQuest, PubMed, ScienceDirect, and Web of Science, Scopus, and more.

The REMS student's main responsibilities will be
1. Construct a dataset connecting source publications and target datasets
2. Draft a paper describing the challenges of implicit data reference and characterizes common types of implicit reference
3. (if possible) Develop a supervised machine learning algorithm to identify data-related publications

Students who are interested in data citation, metadata, and archives but who do not have the computational skills to build the machine learning model are welcome.

I will meet weekly with the student. I expect students to work well independently but will happily help with all aspects of the project. Elisabeth Moss, who directs ICPSR's Bibliography efforts, will also meet with and support the student throughout the project.