How does data curation impact data reuse?

Faculty supervisors: Libby Hemphill, Andrea Thomer

Despite the potential for innovation and advancement that data sharing holds, we don’t yet know how to prioritize datasets for professional preparation and archiving. It’s likely that some datasets hold more reuse potential than others, and data sharing policies should prioritize high-value data over others instead of being one-size-fits-all. This project will help us understand the relative impacts of features of datasets (e.g., questions asked, populations included, topics covered) and curatorial actions (e.g., variable standardization, documentation improvements) on data reuse (e.g., citations, downloads). It will produce metrics to explain the return on various types of resource investment—e.g., what kinds of curatorial action increase data reuse and by what margin?

ICPSR’s standard levels of curation and its open-access archive, openICPSR, provide useful natural experiments for studying the impact of investing in curation activities on the reuse of social science data sets. The levels of curation available at ICPSR provide different types of curatorial activities (e.g., standardizing variable names, recoding categorical variables, providing question text for surveys), and the curation tracking system in use enables us to compare the relative influence and cost of various actions on the data’s reuse. As an open archive in which datasets receive no professional curation activities, openICPSR serves as a useful control. Datasets in openICPSR are prepared by the depositing researchers and are provided “as-is,” meaning the data files and supplementary materials include only unmodified files the depositor provided. Depositors may choose openICPSR or the curated archives for different reasons, making it an imperfect comparison for the curated archives, but the 600+ studies archived there do provide useful data about how “uncurated” data is structured, documented, and reused.

ICPSR Curation work is documented through SPSS syntax files, JIRA tickets, and processing plans. The syntax files enable us to analyze specific actions such as variable renaming and data type conversions. JIRA tickets allow us to measure the impact of time investments and activities that happen outside SPSS. Processing plans provide context for these other data sources and will support analysis of other data.

The REMS student will be responsible for

1. Constructing a dataset that includes study- and file-level variables explaining the processing history and reuse of datasets available from ICPSR
2. Conducting statistical analysis to determine the relative impact of processing variables on reuse variables in that dataset.
3. Identifying statistical measures that would be useful as long-term metrics of curatorial impact.
Successful candidates for this REMS position will have

- Commitment to responsible stewardship of data resources and curiosity about digital curation processes and decision-making
- Experience with SPSS syntax files or automating data preparation through R, Python, or Stata
- Knowledge of logistic regression and structural equation modeling
- Familiarity with bibliometrics or other scholarly/scientific metrics.

We will meet weekly with the student. We expect students to work well independently but will happily help with all aspects of the project. We will also connect students with ICPSR staff who can help access and decipher syntax files.