Measuring and Improving the Efficacy of Curation Activities in Data Archives

Summary: "Measuring and Improving the Efficacy of Curation Activities in Data Archives" is a three year, $739,066 National Digital Infrastructures and Initiatives project led by investigators at the University of Michigan School of Information in partnership with the Inter-university Consortium for Political and Social Research (cost share: $108,537). The goal of this project is to understand curatorial actions impact the use of digital collections. This project will (a) create curatorial metrics for measuring the impact of curation activities and (b) develop an open source software tool for modifying metadata and tracking curation activity.

Statement of National Need

There is ample evidence that digital curation is critically important to ensuring the preservation, accessibility and usability of digital collections. However, we have relatively less data about the impact of *specific curatorial actions* on the usability or accessibility of digital collections. We similarly have little data analyzing the *return on investment* of different curatorial actions. What curatorial processes have the most immediate or farthest reaching benefits? To what degree do different types of curation improve use and accessibility? What curatorial actions must be prioritized to support different kinds of reuse—and what might be delayed until a later date? Without being able to answer these questions, digital curators, repository managers and administrators cannot clearly prioritize, plan, or fund digital curation work. Furthermore, the developers of curatorial tools cannot assess or prioritize which features and functionalities will best increase curatorial efficacy or data reuse.

Empirical studies are sorely needed to evaluate the impact of specific aspects of digital data curation, as are tools to that might help data curators analyze the impact of their curation work. We propose to use *curatorial metrics* to evaluate the impact and efficacy of specific data curation processes. Curatorial metrics are statistical measures similar to bibliometrics, but designed to show the use of digital collections and the impact of curatorial work over time. We will **develop and analyze a range of curatorial metrics from the last five years of data curation at the Inter-university Consortium for Political and Social Research (ICPSR),** a highly impactful social science data repository. Following our findings from this evaluation, we will **develop a software tool that will support high impact data curation processes while also creating and capturing the data necessary for further metric evaluation in the future.**

Why ICPSR. With more than 50 years of service to the social sciences, ICPSR is the largest archive of digital social and behavioral science data in the world. ICPSR curates, preserves, and disseminates original social science data for research, instruction, and policy evaluation. ICPSR's Data Curation Team consists of 30 experienced data curators who provide data curation and management services for all ICPSR archiving projects. The Data Curation Team has detailed logs of curatorial activities undertaken in the repository over the last 5 years. ICPSR has also invested significant resources to identify the research uses of these datasets in journals, working papers, and other research products.

ICPSR's broad scope, significance, and extensive documentation makes it an ideal testbed in which to evaluate the impact of curatorial actions. openICPSR, ICPSR's uncurated open access repository, provides a natural "control" case that makes it possible for us to compare rates and kinds of use between curated data (ICPSR's general and topical archives) and uncurated data (openICPSR). The existing data curation logs are stored in a range of formats and files; through our work we will be able to develop a more streamlined workflow for the evaluation of curation work, which will be broadly adoptable by other repositories and curation settings.

Project Design

This project will be conducted through two phases of work.

Phase I: Evaluation of curation logs, and refinement of curatorial metric

We will evaluate existing ICPSR curation logs and use records (such as downloads and citations) to identify high impact curatorial processes: that is, specific data transformations or preservation steps that are associated with higher rates of reuse or access. Curation logs are available for 2828 studies over a period of 5 years. We

will annotate curation logs following the RDA/TDWG Curatorial Metadata and Attribution Model (REF), which proposes a common framework for describing and citing curatorial work. This will make it possible to then extend curation metrics such as the the Data Usage Index to analyze these data and thereby identify correlations between curatorial processes and use and access. For instance, are there curation actions that occur on nearly all data sets? Are there curation activities that are associated with higher rates of different kinds of use or access? Are there notable differences in use and access between the uncurated data in openICPSR and the curated data in the general and topical archives? By answering these questions, we will identify high impact curatorial actions, and will furthermore identify important metrics that might

Results of phase I: an annotated set of curation logs; a dataset showing correlations between use and access to dataset and specific curatorial actions; framework of high impact curatorial actions; refined metrics for evaluation of return on curatorial investment; research paper detailing key findings

Phase II: Tool development

Our work in Phase I is enabled by the existence of ICPSR's curation logs and extensive records on the research uses of each dataset; other repositories may not keep such detailed records. In Phase II, we will make it possible for others to build on our work by developing a tool that enables high impact data curation processes and the simultaneous creation of curation logs, which will allow other repositories to conduct similar analyses of the return on investment of their work. The specific features of the tool will follow from the prior analyses—we will prioritize common tasks that have high impact on data discoverability and use. The tool will also capture and display activity to facilitate attribution efforts. We will then test the tool's impact by inviting data librarians and ICPSR Official Representatives—both highly engaged communities already invested in data quality—to use it and then analyze the impact of their modifications on the data. Lastly, we will release an open source version of our tool so that other archives can invite variable-level metadata modification in/on their collections.

Results of phase II: an open-source software tool that supports variable-level metadata editing and curation activity tracking, research paper detailing the impacts of that editing on data reuse

National Impact

- Understand ROI of specific curation activities and the places where curators add the most value to data
- Open source software that allows community contributions to improve metadata
- Enable long-term tracking of impacts of curatorial activities, including provenance and attribution
- Create underlying ways to measure curatorial work so that any data repository can assess which of their curatorial activities have the most positive impact on reuse

Diversity Plan

ICPSR holds a variety of datasets that document underserved communities. By understanding the value of curatorial activities on all collections as well as subsets of digital data, we can leverage this information to better disseminate these unique collections. Furthermore, the analyses may reveal how metadata can be used to erase systematic biases in representing the data.

Budget

The estimated budget is $739,066 from IMLS. This amount of funding includes salaries and wages ($197,920); fringe benefits ($59,376); student support ($207,915); travel ($18,000); advisory board costs ($1,500); computer support ($19,109); indirect cost ($235,246). Although cost share is not required, we included $108,537 in cost-share for this proposal.