Summary: “Measuring and Improving the Efficacy of Curation Activities in Data Archives” is a three-year (Fall 2019-Summer 2022), $605,066 ($124,429 cost share) 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. The goal is to investigate how curatorial actions impact the reuse of digital collections. This project will assess stakeholders needs, priorities, and values for data reuse and create curatorial metrics for measuring the impact of curation activities.

Statement of National Need. There is ample evidence that 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 their usability or accessibility. 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? Without being able to answer these questions, digital curators, repository managers and administrators cannot efficiently prioritize, plan, or fund digital curation work.

Empirical studies are needed to evaluate the impact of digital data curation, as are tools to that might help data curators analyze the impact of their curation work. We propose to develop *curatorial metrics* to evaluate the impact and efficacy of specific data curation processes. Curatorial metrics are statistical measures 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 measures from the last five years at the Inter-university Consortium for Political and Social Research (ICPSR),** a highly impactful social science data repository. With more than 50 years of service to the social sciences, ICPSR is the world’s largest archive of digital social and behavioral science data, and is consequently an ideal testbed for this work.

Project Design. This project will be conducted through three phases of work.
Phase 1: Understanding values, priorities, and curatorial actions (Leads: Thomer and Yakel). We will evaluate existing ICPSR curation logs and use records to assess curatorial processes and interview stakeholders (data producers, reusers, repository staff) to understand their values and priorities in supporting data reuse.

Phase 2: Measuring Reuse and Impact (Leads: Akmon and Hemphill). In addition to the traditional measures the literature measuring data reuse impact (e.g., citation and download), we will utilize at least two additional measures of reuse and impact: *secondary impact* and *diversity*. We will use multivariate regression analysis and structural equation modeling (SEM) to determine the relationships among curatorial actions, metadata, the dataset itself, ICPSR services, and reuse and impact.

Phase 3: Generating Curatorial Metrics (Leads: Hemphill and Yakel). The results of the SEM analysis will produce estimations of the strength and direction of relationships between variables. We will develop a set of curatorial metrics that repositories can use to track the impact of their work on the reusability of datasets.

National Impact

The metrics developed will be tested and disseminated to repositories throughout the US and the world. We will promote and describe their use in such a way that any data repository will be able to: 1) Understand ROI of specific curation activities and the places where curators most add value to data and 2) Create mechanisms for long-term tracking of impacts of curatorial activities, Metrics will enable repositories to fine-tune their curatorial processes to support their designated communities as well as to engage in data driven decision making using metrics best suited for their environment.

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 the 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.

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 (Fall 2019-Summer 2022), $605,064 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: $124,429). The goal of this project is to understand how curatorial actions impact the use of digital collections. This project will assess stakeholders needs, priorities, and values for data reuse and create curatorial metrics for measuring the impact of curation activities.

Statement of National Need

Ample evidence suggests that digital curation is critically important to ensuring the preservation, accessibility, and usability of digital collections (Borgman, Scharnhorst, & Golshan, 2018; Faniel & Zimmerman, 2011). 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 furthest 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 efficiently 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. We propose to **develop curatorial metrics to evaluate the impact and efficacy of specific data curation processes**. Curatorial metrics are statistical measures similar to bibliometrics but designed to assess the impact of curatorial work over time on the use of collections. Using curation logs and other records, 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.

Previous indices to of data reuse, frequency and impact have not considered the impact of curatorial actions. For example, Ingwersen and Chaven (2011) created the Data Usage Index, which used searches and downloads as proxy measures for data reuse of data. Weber et al. (Weber et al., 2013) extended and applied Ingwersen and Chavan’s DUI to a climate science repository; while they argue that use metrics could help guide data curation efforts, they do not directly consider the impact of curation work in their work. Fear’s (2013) dissertation, based on ICPSR data, introduced two new measures of data reuse impact: secondary citation (citations of the article citing the data) and diversity (disciplinary breadth of reuse). In her study, no datasets had consistently high impact across both metrics. In the practice realm, EZID and DataCite services track data reuse through traditional bibliometric methods such as citation count (Brase, 2009; California Digital Library, n.d.; Michener et al., 2011; Wilkinson, 2010). Though these metrics are helpful in showing a broad view of data reuse, they also “have missed broad swaths of important activities, including the benefits associated with the collection, management, and preservation of digital resources” (Mayernik, Hart, Maull, & Weber, 2017). Our project will seeks to build a stronger evidence base for curatorial decision-making and actions.

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 data collections are used in a broad spectrum of disciplines, including political science, sociology, demography, economics, education, psychology, criminology, gerontology, public health, public policy, and more. ICPSR’s Data Curation Team consists of 30 experienced data curators who provide 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 on 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.

IMLS announced on March 11, 2019 that ICPSR is among the 30 finalists for the 2019 National Medal for Museum and Library Service. National Medal winners will be announced later this spring. The National Medal is the nation's highest honor given to museums and libraries for service to their communities. For 25 years, the award has celebrated institutions that demonstrate extraordinary and innovative approaches to public service and are making a difference for individuals, families, and communities. ICPSR is the only data organization among the nominees for 2019, and this nomination recognizes its status as a leader among archives. We expect that processes and outcomes from ICPSR will be of use and interest to other libraries and archives.

Project Design

Our primary research questions are

1. What impacts do specific curatorial actions have on the impact or reuse of research data?
2. How should curatorial actions be prioritized to balance impact and return on investment?

To address these questions, we propose a mixed-methods study combining qualitative research (interviews, content analysis) and statistical modelling to identify important factors that impact and predict data reuse. We will identify factors of significance for reuse through interviews with data producers/sharers and reusers and map those onto curatorial actions. By combining these variables, we will create a model that predicts the impact of curatorial metrics and thus identify those which create the highest return on investment. Generating the statistical models requires that we define measures such as *curatorial action* and *reuse*, and the interviews with stakeholders will inform how we define those measures.

To identify curatorial actions and other features of datasets and ICPSR services that influence reuse, we will evaluate existing ICPSR curation logs and use records (such as downloads and citations). Curation logs contain data about specific data transformations or preservation steps, and connecting them to use records enables us to identify which actions are associated with higher rates of reuse or access. Table 1 summarizes our research activities, the data we will collect, the types of analysis we will conduct, and the artifacts will we generate.

We will begin by analyzing studies included in the Resource Center for Minority Data (RCMD) collection. A *study* at ICPSR is a collection of data produced by a single research study along with codebooks and supplemental documentation and may contain multiple data files. RCMD focuses on data about underrepresented populations, and our project team includes two RCMD staff members (Hemphill and Thomas). We've chosen to start with RCMD because it has limited resources but a strong interest in increasing reuse and visibility of the data in its collection. RCMD staff are committed to working with ICPSR's curation team to implement changes to our processes based on our project's findings.

We will additionally conduct interviews with stakeholders throughout the project to better understand their values and priorities in supporting reuse. The metrics we develop in this project will answer these questions and will enable archivists to make informed decisions about their curatorial activities.

Our project takes place via three phases of work:

1) Understanding values, priorities, and curatorial actions (method: semi-structured interviews, content analysis)
2) Measuring reuse and impact (method: multivariate regression, structural equation modeling)
3) Generating curatorial metrics (method: structural equation modeling, path analysis)

Phase 1: Understanding values, priorities, and curatorial actions (Leads: Thomer and Yakel)

Defining Curatorial Activities

Curatorial activities are the specific transformations, alterations, or improvements made to data products to improve their fitness-for-use or preservation readiness. These include tasks such as data cleaning and normalization; creating and improving metadata and other documentation; the application of controlled vocabularies or (meta)data standards; and so on. These tasks vary and depend on the type of data being curated; the scope and focus of the organization doing the curation; and the designated community the curators seek to serve. There has been some research examining curatorial actions (Daniels, Faniel, Fear, & Yakel, 2012; Pienta, Alter, & Lyle, 2010; Yakel, Faniel, & Maiorana, n.d.). Chao, Cragin and Palmer (2015) derive a typology of data curation concepts, activities, and terms through a qualitative study of earth science researchers and show how the typology can support cost analysis of different curatorial activities. However, further work is needed to examine efficacy of these tasks with different data types and then to empirically demonstrate the costs and benefits to a repository and a user community.

In our first phase of work, we extend these prior efforts by further **identifying curatorial actions through content analysis and annotation of ICPSR curation logs.** Curation logs are available for 2,828 studies over a five year period. We will annotate curation logs following the RDA/TDWG Curatorial Metadata and Attribution Model (Thessen, Woodburn, & Koureas, 2018), which proposes a common framework for describing and citing curatorial work. This will allow us to take a grounded approach to identifying curatorial actions, and the dataset will be used later as a foundation for later work, in which we generate new curatorial metrics.

Curation logs (an example log is provided in the supporting documents) capture a variety of curatorial actions taken by ICPSR including standardizing variable names, defining missing values, attaching question text, completing metadata fields such as study design and sampling approach, among others. Analyzing curation logs enables us to study the impact of data transformations. Curation and archive management staff also use JIRA tickets and processing plans (an example ticket and processing plan are provided in the supporting documents) to discuss curation activities, and these tickets contain decisions about modifications curation will make, estimates of the time required, and actual time worked. Analyzing JIRA tickets enables us to measure the impact of aspects of curation work that are not captured in the data and its transformations. At the time of this writing, we have 673 JIRA tickets to include in our analysis. JIRA ticket data, such as the length of discussion about a dataset or mismatches between the time estimated and the actual time required to complete curation activities may indicate complications in the data or its structure that our other measures may miss.

Identifying Values and Priorities for Reuse

In order to build meaningful metrics of impact, we need to understand what curators and researchers consider meaningful impact (Piwowar, 2013; Punzalan, Marsh, & Cools, 2017). We will **conduct two rounds of qualitative, semi-structured interviews with data providers and re-users, repository staff and leadership to better understand their values and priorities in supporting reuse**. The first round of interviews will focus on how understanding how stakeholders currently measure—or wish they could measure—the impact of their collections. What kinds of reuse do they value? What kinds of impact do they want their collections and data to make? Interview transcripts will be coded in NVIVO using a constant comparative methodology (Glaser, 1965). We will draft summaries of the coding results to inform Phases 2 and 3 activities.

	Data Required	**Analysis Approach**	**Outcomes and Artifacts**
RQ1: What impacts do specific curatorial actions have on the impact or reuse of research data?			
What classes of data transformations occur at ICPSR? What, if any, curation actions occur on nearly all data sets?	● Curation logs ● JIRA tickets ● Interviews with Curators	● Annotation of curation logs w/ RDA/TDWG standard ● Constant-comparative coding of interviews in NVivo	● Annotated curation logs ● Annotated JIRA tickets ● Annotated interview transcripts ● Coding results
How should we measure reuse and impact?	● ICPSR's current measures ● Interviews with data sharers /reusers ● Interviews with ORs and other archives	● Constant-comparative coding of interviews in NVivo	● Annotated interview transcripts ● Coding results (e.g., how other archives measure reuse and impact)
How should we establish baselines for expected reuse and/or impact?	● ICPSR access logs ● Interviews with ORs and other archives ● Interviews with data sharers /reusers	● Regression analysis ● Constant-comparative coding of interviews in NVivo	● Table of regression results ● Summary of regression results ● Coding results (e.g., what kinds of reuse do PIs and archives anticipate for different datasets)
How are curation actions related to reuse? Are there notable differences in use and access between the uncurated data in OpenICPSR and the curated data in the general / topical archives?	● ICPSR's existing measures ● Curation logs ● JIRA tickets ● New measures (e.g., altmetrics, diversity, secondary citation)	● Regression analysis ● Structural equation modeling (SEM)	● Table of regression results ● Summary of regression results ● Complete, valued SEM diagram ● Summary of SEM results
RQ2: How should curatorial actions be prioritized to balance impact and return on investment?			
What do various curation actions cost? What characteristics of the data impact the cost of curation actions?	● Curation logs ● JIRA tickets ● Interviews with Curators	● Annotation of curation logs w/ RDA/TDWG standard ● Constant-comparative coding of interviews in NVivo	● Annotated curation logs ● Annotated JIRA tickets ● Annotated interview transcripts ● Cost estimates
What kinds of reuse are most important to data producers, curators, and to ORs? How much effort are data producers / reusers willing to expend to prepare data for sharing / reuse?	● Interviews with PIs who share data ● Interviews with data reusers	● Constant-comparative coding in NVivo	● Summaries/memos of coding results

Table 1. Overview of Research Activities

Phase 2: Measuring Reuse and Impact (Leads: Akmon and Hemphill)

The main goal of Phase 2 is to generate measures (i.e. variables) of reuse, impact, and actions so that we can understand the impacts of each specific measure by modeling their relationships. We use the outcomes from Phase 1 to inform the construction of these variables. Then in Phase 3, we will use the models produced in this phase to develops metrics, or systems of measuring these relationships and deciding which measures to include.

ICPSR's Current Measures of Use and Impact

ICPSR captures numerous measures of study impact, many of which are presented publicly on each study's home page under *usage report* and *data-related publications*. Some studies are part of a larger *series*, which are repeated studies, with new waves of data added periodically, and each study within the series receives its own usage report. To populate *usage reports*, ICPSR captures a number of usage metrics for each study, including *total downloads*, *total sessions*, and *total users*. Each study in ICPSR's collection has its own homepage, and its usage report is publicly visible there. Downloads refers to the number of times a study's files have been downloaded. Sessions measures the number of unique user visits who downloaded files. Users captures the number of users who downloaded files.

The ICPSR Bibliography of Data-Related Literature links over 75,000 research publications to the ICPSR data on which they are based and populates the *data-related publications* tab on ICPSR study homepages. These works include journal articles, books, book chapters, government and agency reports, working papers, dissertations, conference papers, meeting presentations, unpublished manuscripts, magazine and newspaper articles, and audiovisual materials. Showcasing the impact of a given dataset on the field, ICPSR bills the ICPSR-generated bibliography as a key indicator of return on the original investment in the creation of the data. ICPSR makes no claims on the exhaustiveness of the Bibliography. Many authors do not include a formal citation of the data used in their work and/or do not inform ICPSR of their published materials (Moss & Lyle, 2018). Therefore, this collection likely underreports utilization of ICPSR data and should only be viewed as a partial statement of its impact on research.

Generating New Measures of Reuse and Impact

Following Fear (2013), we will **construct two additional measures of reuse and impact**: *secondary impact* and *diversity*. Secondary impact is a measure of how many times the reuse publications have been cited. We can construct this measure by gathering citation data for all items in the bibliography that are not the original PI's publications. Diversity is a measure of the breadth of disciplines that use the data and can similarly be constructed from the bibliography. Because both of these measures rely on citations in the bibliography, they likely capture the lower boundaries of reuse and are therefore conservative measures. We can also construct an ensemble measure that includes some or all of the reuse measures, potentially weighting and combining them differently for different purposes. Data from our interviews with depositors and repository staff will help determine the best methods for measuring reuse and impact.

Measuring Relationships between Curatorial Actions, Reuse, and Impact

We will **use multivariate regression analysis and structural equation modeling (SEM)** to determine the relationships among curatorial actions, metadata, the dataset itself, ICPSR services, and reuse and impact. We have chosen these analyses because they are appropriate for measuring the structure of relationships between multiple variables where we have both measured variables (e.g. number of downloads, type of access provided) and latent constructs (e.g., metadata completion, dataset properties). In Figure 1, we provide an SEM model with initial variables and constructs for inclusion, but the complete list of variables will depend on the measures of impact and curatorial action that we develop in the initial stages of the project (i.e., through interviews). The figure includes measurable variables that are related to our five main latent variables: curation (disclosure risk review, curation level, missing variable definitions, question text, value labels, and variable names), the dataset's metadata properties (method, sample, and completeness), the dataset itself (format, age, whether it is part of a series, and its primary investigator or creator, what actions PIs took to prepare data for reuse), and ICPSR services (online analysis, access rules). Each of these constructs are likely related to reuse but not directly measurable and are therefore good candidates for inclusion in an SEM. We include features of the dataset itself, including the size and scope, the time of its release, the reputation of its author, and the number and variety of observations it contains because those measures are known to impact reuse (Fear, 2013; Pienta et al., 2010; Pienta, Hemphill, & Akmon, 2019). We expect that the types of access (e.g., restricted, membership-only, public) and whether or not online analysis is available to also impact reuse.

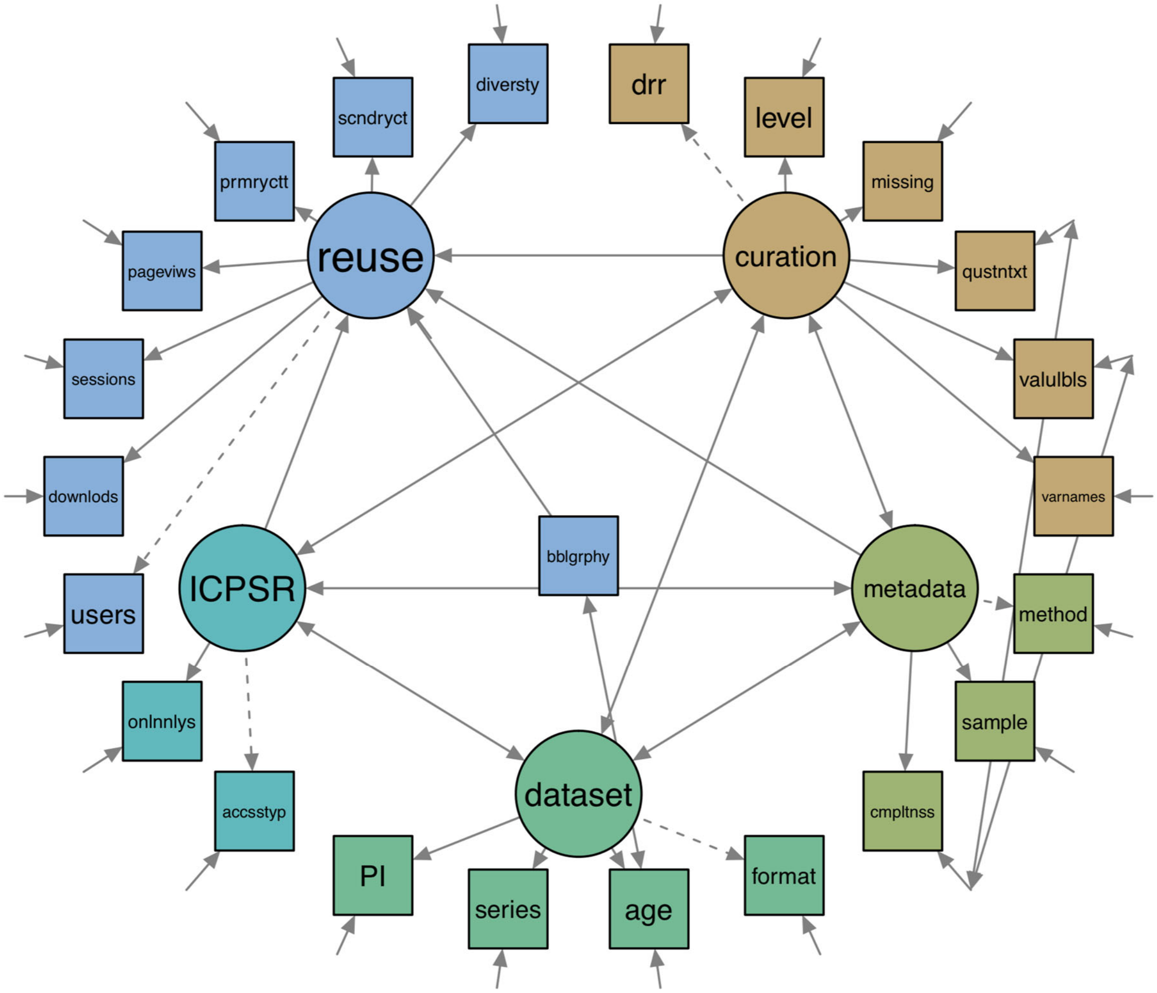

Figure 1. Initial Structural Equation Model (SEM) describing relationships among curation activity, other ICPSR services, metadata, the dataset, and its reuse. Variable names are abbreviated. Squares represent measurable variables. Circles indicate latent variables. Double-header arrows indicate covariance, and single-headed arrows indicate a directional relationship between variables.

In both regression and SEM, ICPSR's existing measures of use will serve as the dependent variables in our initial models. One of the goals for the first year of the project is to develop alternative measures of reuse and impact (*secondary impact* and *diversity*), and these newly-developed measures will serve as variables in later models. We will construct the curation activity and metadata variables from the annotated curation logs and JIRA tickets created in Phase I.

Phase 3: Generating Curatorial Metrics (Leads: Hemphill and Yakel)

Through Phase 1, we will build a deep understanding of curatorial actions through content analysis and annotation of ICPSR curation logs and investigate the different types of impact sought by data producers, curators, and reusers through interviews. The first round of interviews will explain how stakeholders value various types of reuse and how those values may vary among datasets and organizations. In Phase 2, we examine current measures of data reuse at ICPSR and in the literature and combine these variables in a series of multivariate regression analyses and structural equation modeling (SEM) to determine the relationships among curatorial actions, metadata, the dataset itself, ICPSR services, and reuse and impact. We continue this process in Phase 3, to further **isolate the variables that lead to different types of impact**. The results of the SEM analysis will produce estimations of the strength and direction of relationships between variables (i.e., each line in the model above will receive a value). SEM also makes it possible to analyze the paths between variables so that we can explain how measures influence curation which, in turn, influences reuse. For instance, we will be able to explain the amount of variance in reuse that curation accounts for and estimate the impact of each of the individual curation measures.

We will be able to build SEMs that estimate the relationships between different combinations of variables so that we can weigh different types of reuse or different curation activities according to the values of a particular archive and its available resources. If a repository is interested in demonstrating broad-based use of its collections through a high download or citation count, we will be able to analyze the curatorial actions leading to collections with high citation counts and control for the type or data. Likewise, if the SEM analysis reveals that the inclusion of question text has a high impact on diversity/breadth of reuse, then archives that value diversity will want to invest in creating complete question text. However, if question text has no impact on primary citations, then archives that value primary citations will not need to invest in question text preservation. We will further **validate these metrics with staff and leaders in diverse repositories in a second round of interviews**, in which we ask participants how effectively these new metrics reflect their priorities and values**.** These interviews will also be coded in NVIVO using the constant comparative method.

Metrics will be further tested and refined through analysis of ICPSR's internal curation documents. This documentation explains how the organization estimates the costs associated with various curation actions. In developing metrics, we will use the cost estimates to gauge returns correlated with efforts and types of reuse. For instance, one document for guiding curation budgeting process includes file content (quantitative, qualitative, or spatial data), number of files, number of variables, disclosure risk, and other complexity measures. Our models include these and related measures, making it possible to determine the return on investment for curation activities. These models with be reusable by other archives, which could assign different weights or coefficients to these measures depending on their estimates of the actual costs of these activities within their organization.

Project outcomes

In summary, the work above will result in a) a empirically-derived typology of curatorial actions that extends prior work in this area; b) qualitative data about stakeholders' values and priorities in facilitating reuse and determining impactful reuse; c) the definition of alternative *measures* of impactful data reuse; and d) a set of curatorial metrics and statistical models that will make it possible to demonstrate the return on investment of curatorial work on a dataset's reuse. These outcomes represent substantial contributions to LIS research and practice; our dissemination plan is outlined further below.

Project Personnel

Libby Hemphill, Associate Professor in the School of Information and Director of the Resource Center for Minority Data at ICPSR, will lead the project. Her expertise is in computational social science, and she will lead the statistical analyses efforts and work with ICPSR staff to implement the metrics we develop. She will also oversee the ICPSR staff and co-advise the graduate student.

Elizabeth Yakel, is a Professor and Associate Dean for Academic Affairs in UMSI. Her research focuses on data reuse and the context needed to ensure meaningful reuse of data by scholars and the public who were not involved in the original research.

Andrea Thomer is an Assistant Professor of digital curation in the School of Information. She conducts research on data curation and information organization, and has prior experience extending data use metrics (Weber et al., 2013). She will lead work identifying and describing curatorial activities, and will contribute to the development of new metrics.

Dharma Akmon is Director of Project Management and User Support and an assistant research scientist at ICPSR. Her research has focused on scientific data practices. She will be involved in all aspects of the research efforts to understand the impact of curation activity on data reuse.

An UMSI PhD student, to be determined, will be responsible for conducting interviews, working with David Thomas to access log data, and working with the faculty to analyze data. The student will also contribute to dissemination efforts by authoring reports and publications.

David Thomas, is Project Manager for the Resource Center for Minority Data and former curation team member at ICPSR. He brings years of experience curating data and supporting researchers in both deposit and reuse. Thomas will assist with project management and provide ICPSR background knowledge on curation activities and processes; he will also provide access and support for the ICPSR log data.

Dory Knight-Ingram is the senior editor in ICPSR's Membership and Communications unit. She is adept in social media management, news releases, video production, and other ways to communicate our mission and highlight our work and our reach. Dory will lead social media outreach efforts.

Jenna Tyson is the Multimedia Designer in ICPSR's Membership and Communications unit. She designs materials that promote, educate, and enhance the mission of ICPSR. She will help create materials for dissemination on social media and in publications.

Our advisory board consists of three experts in data curation and impact metrics. Nicholas Weber, assistant professor in the Information School at the University of Washington, was selected for his expertise in data curation and experience with the Qualitative Data Repository, Dataverse, and Make Data Count (Mayernik et al., 2017; Weber et al., 2013). Anne Thessen, a research scholar at the Ronin Institute and chair of the RDA/TDWG Attribution Metadata Working Group, was selected for her expertise in developing metadata for attributing curatorial and maintenance work (Thessen et al., 2018). Ricardo Punzalan, assistant professor at the University of Maryland College of Information studies, was selected for his expertise in ethnographic data and assessing impact (Punzalan et al., 2017).

Dissemination Plan

Our dissemination plan is designed to make our work and its impacts accessible to both practicing librarians, archivists, and curators managing data in a variety of settings. Researchers on the project will work with ICPSR's Membership and Communications team to ensure timely and wide dissemination of our results. Our communication goals are to (1) create awareness of the the project, and (2) initiate consideration and initial use of our curatorial activities measures and metrics. Our primary audience for communication activities are ICPSR's 784 Official and Designated Representatives (ORs) and the data librarian/curation research community in such organizations as the Research Data Alliance and the Association of College and Research Libraries. We will accomplish our dissemination goals through four primary mechanisms:

1. Webinars,
2. Social media posts,
3. Conference presentations (at both scholarly and professional conferences), and
4. Peer reviewed papers.

ICPSR webinar announcements reach of over 1500 individuals (including 784 in the OR role), and typically 60-80 individuals "attend" a webinar. Webinar recordings and materials are available on the ICPSR YouTube channel. ICPSR's social media team uses Twitter, Facebook, LinkedIn, and Instagram and will broadly promote the project and promote the existence of our metrics through these channels. The ICPSR communication team commonly posts an announcement on its website with links to the research paper or webinar and then links back to that announcement from its social media posts. ICPSR's social media accounts reach over 10,000 users. ICPSR's social media brings digital research data to life and keeps the data community engaged and aware of ICPSR as a leader in data stewardship.

Researchers will present findings at appropriate professional and scholarly conferences such as IASSIST, IDCC, RDA, RDAP, ASIST, and ACRL. We have chosen these conferences are likely outlets because they reach both practicing librarians and archivists as well as those researching data curation. ICPSR staff have a strong tradition of attendance and engagement in these conferences. Our goal is to provide periodic updates about the project -- the success of our metrics depends in part on the community's engagement. Therefore, we will plan a conference presentation at least once per year of the project.

We will also produce publications appropriate for peer review in publications such as the *Journal of the Association for Information Science and Technology (JASIST), International Journal of Digital Curation (IJDC), Data Science Journal,* and *PLoS One.* Our team has experience publishing in these venues, and they are the highest-impact journals in the appropriate fields.

Diversity Plan

Our project includes specific plans for engaging diverse and underserved communities by focusing on data about underrepresented populations and recruiting for diversity on our project team.

ICPSR holds a variety of datasets that document underserved communities, and Hemphill directs RCMD, which manages those collections. RCMD's mission is to provide educators, researchers, and students with data resources so that they can produce analysis of issues affecting racial and ethnic minority populations in the United States. Because RCMD is a membership-funded archive, the data and products of this project are of special interest to ICPSR and its curators, and lessons learned in RCMD are likely to be useful to and adopted by other topic archives at ICPSR. 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.

We will also actively recruit students from underrepresented groups to join our research team. The PIs' strong track records of mentoring women, students with disabilities, and students from underrepresented racial and socioeconomic groups will continue, and our institution is committed to supporting our efforts around diversity, equity, and inclusion. Our budget includes funding for one graduate student, and we will apply for additional student support through university programs such as the Undergraduate Research Opportunities Program (UROP), UMSI's Research Experience for Master's Students program, the Michigan Community College Summer Research Fellowship, and the Women and Gender Summer Fellowship Program. These programs all are externally funded and seek to give diverse students research opportunities. We will use a number of existing recruitment resources at U-M, UMSI, and ICPSR to ensure diverse participation in all the proposed activities. Recruitment activities will draw on the expertise of campus units such as the Center for Educational Outreach, the Women in Science & Engineering program, and the National Center for Institutional Diversity.

National Impact

Transformative research. There is an urgent need to support data repositories in their curation efforts. In 2013, Office of Science and Technology Policy director John Holdren directed federal agencies with budgets over $100 million to "develop plans to make the results of federally funded research freely available to the public;" (Holdren, 2013) and funding agencies and journal publishers increasingly require that researchers share and archive their data. This has placed enormous pressure on data repositories to develop effective methods of data curation. By developing a set of curatorial impact metrics, we will create guidelines that will have immediate and broad benefit to a wide range of repositories, both in the US and abroad.

Impact beyond our case. By rooting our metrics development in prior work, we will help ensure that any metrics we develop are adaptable by repositories beyond ICPSR. Specifically, this project will result in four transferable deliverables. First, this project will establish a key set of *measures* for curatorial activities that can be adopted by other repositories. Many curatorial actions can be measured; this project will identify the actions that have the highest impacts on reuse. Second, the project will identify the curatorial actions that yield the highest return on investment for different types of data or different types of reuse. Thus, the project will suggest *metrics,* which we view as a higher-level more subjective assessment based on the values of the institution. For example, if a repository values the interdisciplinary reach of data, we would use metrics that assess cross-disciplinary data use. Third, the project will demonstrate how a variety of existing tools (e.g., JIRA) and standards (e.g., PROV) can be used for the long-term tracking of curatorial actions on datasets. This is essential since impact metrics are often not identifiable in the short term and other repositories will need to be able to collect, retain, and mine their data over time. The result will be an *end-to-end process* for analyzing curatorial action and comparing it to metrics. The process will be characterized and described generally so that it is transferable to other repositories regardless of their size, disciplinary specificity or generality, or data format.

Sustaining beyond funding period: ICPSR is committed to efficient, effective use of curation resources and will use the metrics of this project to inform their standard levels of curation and their curation workflows. ICPSR's levels of curation define the set of actions that are to be performed to curate a study's data before they are disseminated. Armed with a better understanding of how particular curatorial actions impact the use of data, ICPSR will adjust the definitions of the levels to ensure that they include the right set of curatorial actions to support the type of reuse most valued given the data. Furthermore, ICPSR will share information about the ROI of specific activities with the ICPSR Membership and topical archives to inform their decisions about how to invest in curation. Information about the specific curatorial activities that influence the data's secondary impact will also be used to ensure that ICPSR captures and tracks the most meaningful information about curatorial work.

Schedule of Completion

	Fall 2019	**Winter 2020**	**Spring 2020**	**Summer 2020**	**Fall 2020**	**Winter 2021**	**Spring 2021**	**Summer 2021**	**Fall 2021**	**Winter 2022**	**Spring 2022**	**Summer 2022**
Phase 1: Understanding values, priorities, and curatorial actions												
Content analysis of curation logs	▨	▨	▨	▨								
Interviews with stakeholders	Identify values / priorities for reuse	▨	▨						Evaluating metrics w/ stakeholders	▨	▨	
Analyzing interviews			▨	▨	▨						▨	▨
Phase 2: Measuring Reuse and Impact												
Modeling current measures of impact	▨	▨										
Generating new measures of impact					Leveraging interviews	▨	▨					
Measuring Relationships among Curatorial Actions, Reuse, and Impact			Using existing measures	▨				Using new & additional measures	▨			
Phase 3: Generating Curatorial Metrics												
					Existing measures				New measures	▨	▨	
Disseminate Results												
Social Media and Webinars	▨				▨				▨			▨
Conference presentations	RDA		IASSIST				RDAP, ACRL			IDCC		
Peer-reviewed Pubs.						IJDC			JASIST		Data Science.Journal	▨

DIGITAL PRODUCT FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (e.g., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

All applications must include a Digital Product Form.

□ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

We will use the Attribution 4.0 International (CC BY 4.0) Creative Commons license for datasets we intend to create. This is the default license for products provided in openICPSR, and the license ensures we meet federal open-access requirements and enables broad, diverse use of our products.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The data produced will be owned and distributed through ICPSR. The use of these data will fall under ICPSR's standard terms of use. ICPSR adheres to the principles of the Data Seal of Approval, which, in part, require the data consumer to comply with access regulations imposed both by law and by the data repository, and to conform to codes of conduct that are generally accepted in higher education and scientific research for the exchange of knowledge and information. Those wishing to access these data will have to agree not to redistribute these data without ICPSR's consent. The reason for this is to manage version control and to make sure those accessing these data are using the most current versions.

A. 3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

We will review our data for potential disclosure risk and take any necessary remediation steps prior to depositing them in openICPSR. This could include recoding data (e.g. converting dates to time intervals, exact dates of birth to age groups, detailed geographic codes to broader levels of geography, and detailed income to income ranges or categories).If modifications to address identifiers to create a public-use dataset will seriously reduce the analytic utility of the data, ICPSR may release a restricted-use dataset or both public- and restricted-use datasets. Restricted-use datasets retain confidential, identifying information, and are accessible under controlled conditions.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and the format(s) you will use.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

URL:

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

During phase 1 of our study (years 1 and 3), we will conduct interviews and generate transcripts and memos from those activities. We will use data from these interviews to generate new measures of impact and to validate our proposed curation metrics.

During phase 2 (years 1 and 2), we will generate data about curatorial actions and data use by collecting ICPSR curation logs, users, downloads, sessions, and citations. We will generate the data by coding SPSS curation logs and JIRA tickets which require access to SPSS and JIRA. The data we generate will be stored in non-proprietary formats (likely plain text). One likely format of this data is tabular where each row is a study in ICPSR's collection, and each column represents a curatorial action's date, type, and author. The citation data will be stored in RIS and/or bibtex.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

The research activities are exempt from IRB review. Phase 1 activities (interviews) fall under OHRP Exempt Categories #2. Phase 2 falls under category #4 because they use existing ICPSR curation logs, anonymous use records, and public data citations.

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

No.

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

N/A

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

Interviews and manual coding of interviews and texts. We need to access NVivo to code interview data, SPSS to read curation logs, and JIRA to access other curation records; access is provided through university licenses. We will store our outputs in text format.

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

We will produce codebooks and complete ICPSR metadata records for our data. Codebooks will include variable names, labels, values, value labels, summary statistics, missing data definitions, and methodological details. The codebook should provide enough information to make each variable self-explanatory.

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Data produced from this project will be deposited to openICPSR and made available for download and secondary analysis. Our dissemination plan is designed to make our work and its impacts accessible to both practicing librarians, archivists, and curators managing data in a variety of settings. Researchers on the project will work with ICPSR's Membership and Communications team to ensure timely and wide dissemination of our results. Our communication goals are to (1) create awareness of the the project, and (2) initiate consideration and initial use of our curatorial activities measures and metrics.

A.8 Identify where you will deposit the dataset(s):

Name of repository:

ICPSR

URL:

https://www.icpsr.umich.edu/icpsrweb/

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?

The research team will revisit this plan every six months to ensure that we are following appropriate data management during the project and that our plans for managing and sharing data still make sense and meet the project's research, transparency, and dissemination goals.