RESEARCH ARTICLE

How do properties of data, their curation, and their funding relate to reuse?

Libby Hemphill[1,2] | Amy Pienta[1] | Sara Lafia[1] | Dharma Akmon[1] |
David A. Bleckley[1]

[1]Inter-university Consortium for Political
and Social Research (ICPSR), University
of Michigan, Ann Arbor, Michigan

[2]School of Information (UMSI),
University of Michigan, Ann Arbor,
Michigan

Correspondence
Libby Hemphill, Inter-university
Consortium for Political and Social
Research, University of Michigan,
P.O. Box 1248, Ann Arbor, MI
48106-1248, USA.
Email: libbyh@umich.edu

Abstract

Despite large public investments in facilitating the secondary use of data, there is little information about the specific factors that predict data's reuse. Using data download logs from the Inter-university Consortium for Political and Social Research (ICPSR), this study examines how data properties, curation decisions, and repository funding models relate to data reuse. We find that datasets deposited by institutions, subject to many curatorial tasks, and whose access and preservation is funded externally, are used more often. Our findings confirm that investments in data collection, curation, and preservation are associated with more data reuse.

1 | INTRODUCTION

Data archives are receiving more data than they have capacity to curate and preserve and need to make decisions about which curation actions to take on which datasets. We know that curation matters (Goodman et al., 2014; McLure et al., 2014) but not which curation decisions or metadata enhancements are associated with increased use. Knowing how often data are reused is key to making good collection development decisions. Data reuse refers to tracking the return on investment of curation that increases all kinds of reuses such as obtaining information about a study through reviewing its data and documentation, interacting with the data through data management and data analysis, and using the data to produce new knowledge or to collect new data. Archives need ways of prioritizing which data are likely to be most worthy of curation effort and what curation practices result in the highest use.

Funders, publishers, and researchers themselves have increased demand for data sharing. For instance, the Sorbonne Declaration of Research Data Rights (Sorbonne Declaration on Research Data Rights, 2020), signed by networks of research universities from around the world, encourages universities to share data and governments to establish appropriate regulations. In the United States, the National Science Foundation (2020) and National Institutes of Health (2003) have both adopted data sharing policies that apply to all grantees. NSF's policy indicates that data should be shared with other researchers

"at no more than incremental cost and within a reasonable time" (p. XI-17). Managing and ensuring access to the large and increasing stream of data is a significant challenge as repositories seek a workable model for maximizing the impact of their work (Kitchin et al., 2015). These requirements put tremendous pressure on data repositories to process data quickly and efficiently. Prior work indicates that archives obscure much of the work that goes into preparing data for reuse (Plantin, 2019). The invisibility of the curators' labor may lead outsiders to underestimate the costs and time required to prepare data for sharing and reuse (Thomer et al., 2022).

Understanding relationships between reuse and its predictors requires being able to measure both reuse and the factors that impact it. There are potential problems with some of the past data reuse measures in the literature such as using data citation, which is likely to underestimate reuse (Park et al., 2018; Robinson-García et al., 2016; Silvello, 2018), and data downloads, which may overestimate reuse (Borgman et al., 2015). Downloads capture, but cannot disambiguate, a breadth of uses of interest to archives (e.g., in teaching, for new research projects), and we therefore adopt downloads as an informative measure of data use. We attempt to control for some of the overestimation effect with downloads by measuring unique users who downloaded data and not just raw download numbers.

What about the data and its curation impacts how often it is downloaded? When users look for data to use, they search by keyword or phrase (and not study name or data producer) more than two-thirds of the time (Pienta et al., 2017); this pattern suggests that attaching subject terms to data will make them more discoverable. Data users also often turn to data that are produced by researchers or institutions they know and who have provided information about the context of the data's collection and production (Birnholtz & Bietz, 2003; Faniel et al., 2019). Funding for data archiving services often includes additional resources for promotion. Funders also set specific collection development policies that can be more selective and focused on particular audiences; for instance, the National Institutes of Health's BRAIN Initiative: Data Archives for the BRAIN Initiative specifically supports the creation and management of a data archive for BRAIN Initiative data. ICPSR's general archive, which is membership-funded, has broad and varied audiences. We expect that the additional resources and audience-targeting that accompanies external funding will lead to more data downloads. We generate variables related to properties of the data (e.g., who produced it), the curation actions the archive took (e.g., attaching subject terms), and the funding model for the data to understand how those features of a dataset influence its reuse.

1.1 | Study setting

The Inter-university Consortium for Political and Social Research (ICPSR) maintains the world's largest archive of digital social science data and has been growing its collection for over 55 years. ICPSR is a member-funded consortium that responds to the needs of its membership by identifying high-value data collections for archiving. It also receives funding from federal agencies, private foundations, and institutions to archive particular datasets or collections; in these externally funded collections, many of the selection decisions are made by funders rather than the consortium. ICPSR generates and captures metadata about studies in its collections including the number of variables in datasets, the datasets' primary investigators and depositors, question text, and other documentation for variables, among other metadata records.

ICPSR is widely known for archiving survey and interview data produced by government agencies and collected with federal funding. ICPSR also serves as an all-purpose data repository for the social sciences domain offering data archiving services for small to large research projects. Because of its broad collection development policy (ICPSR, 2021), the ICPSR archive also includes videos, image collections, administrative records, clinical research, and more. The ICPSR data holdings grow through the work of ICPSR's acquisitions staff that conducts outreach to the research community to add data to the archive. There are also unsolicited deposits of data from the research community who know ICPSR as well as returning depositors.

ICPSR uses a "curation level" framework for standardizing common curation actions (ICPSR, 2020). All datasets undergo thorough disclosure risk review and remediation. Level 1 is considered baseline curation. In Level 1, curators create a study website with descriptive metadata, a PDF codebook that explains what each variable represents, and data files for all major statistical software packages (Stata, SPSS, and plain text). Level 2 includes the actions taken in Level 1 and seeks to increase usability through reformatting data as necessary (e.g., converting numbers stored as strings into numeric variables), standardizing missing values, correcting spelling, and making labels more understandable to secondary data users. Level 3 builds on the two previous levels and adds customized documentation and indexes survey question text in the Social Sciences Variable Database (SSVD) to make them searchable. Curating non-tabular data such as qualitative or spatial datasets falls under Level 3.

ICPSR provides access to its public and membership-viewable data through its website. ICPSR maintains download logs about its holdings that we analyze to

evaluate the impact of curation decisions and data attributes on data use. Because it disseminates a wide variety of data and applies a broad set of curation actions, ICPSR can provide a great deal of insight into the characteristics that predict data's use.

1.2 | Related literature

Data reuse fills many user needs, not just the ability to explore new research questions with data. Our expectation of data use in a domain repository, such as ICPSR, is necessarily broad and encompasses wide-ranging purposes such as performing secondary data analysis, informing research design, teaching/training students, study replication, verification of published results, determining compliance with data sharing mandates, meeting broad public accountability/access, and likely other less well-known purposes. ICPSR accommodates a wide range of data types used by diverse fields to respond to data users' wide-ranging needs (ICPSR, 2021, 2022). Accordingly, one can imagine several ways to measure data's reuse, including through page views, downloads, and citations. Data citations are an increasingly common metric for capturing the impact of data reuse (Silvello, 2018), but inconsistent citation practices limit utility of that measure (Kratz & Strasser, 2015; Pasquetto et al., 2017). Furthermore, reliance on formal citation as the sole measure of data reuse fails to capture the full range of activities that signal the data's value and impact, especially to repository managers. In fact, a 2013 dissertation that examined ICPSR's data usage revealed discrepancies between bibliometric measures of impact and study download counts (Fear, 2013): some datasets that ranked in the top 10 most downloaded studies ranked much lower using bibliometrics, indicating download counts account for uses outside publications.

Given the clear limitations of bibliometrics for adequately capturing data's impact, researchers, repository managers, and funders have increasingly focused on download activity to measure data's use and impact. A 2015 study investigated how management transaction logs (including download counts) could be leveraged to describe users (Borgman et al., 2015; Borgman et al., 2018). As the authors noted, transaction logs capture the traces users leave as they interact with the archive; however, they reveal very little on their own about why they are using the data. Also, downloads are also subject to inflation because users may download the same data more than once, users may not actually use data that they have downloaded, or downloads may be triggered by scripts rather than human users. Still, they conclude that logs are some of the best resources repositories have for knowing how the repository is being used. Some studies have focused on data reuse patterns tied to a particular repository, seeking to understand the value of alternative measures of reuse that are not bibliometric-focused, finding evidence that data downloads are a useful indicator of data's impact (Fear, 2013; He & Han, 2017).

Precedent exists for using downloads counts in the scholarly publication realm. To serve journal database providers and librarians that need to measure return on investment, Counter, an international non-profit organization, oversees a standard that enables publishers to report use of their electronic resources in a consistent way; and libraries to compare data across a number of publishers and vendors. Recognizing the special needs of data (e.g., versioning, defining what constitutes the item to count, etc.), several teams of researchers, working primarily through the Research Data Alliance and the Make Data Count project, have proposed a standard for the generation and distribution of usage metrics for research data (Fenner et al., 2018). The resulting Code of Practice for Research Data Usage specifies metric types for reporting that include the "total number of times a dataset was retrieved (the content was accessed or downloaded in full or a section of it)."

As researchers and practitioners grapple with developing widely accepted, non-bibliographic metrics for data's impact, they are leveraging a variety of approaches to examine data reuse. Data reuse studies have largely focused on citation practices (Park et al., 2018), citation patterns (Belter, 2014; Fear, 2013), and patterns of who is using the data and for what purposes (Bishop & Kuula-Luumi, 2017). Several studies examine patterns of data reuse in specific scientific domains, including qualitative social sciences (Bishop & Kuula-Luumi, 2017), genetics and heredity (Park et al., 2018), and oceanography (Belter, 2014). Yet scant research ties reuse patterns captured in metrics to data's traits or the curation that aims at making them more reusable.

Instead, many studies of data reuse examine researchers' satisfaction with reuse (Faniel et al., 2016), researcher's attitudes toward data reuse (Yoon & Kim, 2017), data reusers' trust in data (Yoon, 2017), how researchers decide whether to reuse data (Faniel et al., 2019), and the factors that influence data's reusability (Akmon et al., 2011; Niu, 2009; Zimmerman, 2008). These studies are based primarily on surveys of and semi-structured interviews with data reusers and reveal important considerations for data reusers.

Data reusers are most satisfied with their reuse of social science data when data are "comprehensive, easy to obtain, easy to manipulate, and believable" and when the documentation is high-quality (Faniel et al., 2016,

p. 1412). As researchers evaluate data for reuse, they base their trust in the data on the reputation of the data producer and high-quality data preparation and documentation (Yoon, 2017). Furthermore, they look at important contextual clues when deciding whether or not to use data, including data production information, repository information, and data reuse information (Birnholtz & Bietz, 2003; Faniel et al., 2019). Data reusability depends on understanding the context of the data's production. In scientific research, tacit and craft knowledge is abundant, which makes communicating information—through comprehensive documentation about data—particularly challenging but also critically necessary (Akmon et al., 2011; Carlson & Anderson, 2007).

Fear's (2013) dissertation study of ICPSR investigated the factors that influence data reuse, where reuse was measured using both bibliographic and download metrics. Specifically, she examined the impact of curation status (curated vs. uncurated), data producer information, connection with data producer, data prominence, dataset size, and discipline of the dataset on reuse impact. She found that curation status was the most significant predictor of the number of downloaders a dataset received, followed by the h-index of the data producer. She also found dataset size—as indicated by the number of variables in the study—had a significant association with the rate at which the data were downloaded. Interestingly, the study's interviews revealed that researchers prefer data from government sources or other highly reputable institutions. Fear excluded studies with institutional authors from her analysis to use h-index as a proxy for author reputation (a measure that does not apply to institutions), and therefore cannot tell us whether data produced by institutions receive more downloads. Furthermore, Fear's analysis—conducted long before ICPSR implemented standardized levels of curation—treated curation activity as a binary (curated vs. uncurated) and hence was unable to identify the impacts of different kinds of curation activity.

Other research on dataset search and reuse among social scientists found that researchers look for data from investigators and institutions that they trust, that contain individual questions they are interested in, and that match keywords they use to search (Gregory et al., 2019). Pasquetto et al. (2019) found that researchers reusing data preferred to collaborator with the original data collectors so they could ask questions about and understand the data's context and purposes. Metadata such as the individual questions asked in surveys (Gregory et al., 2019), what individual variables mean and measure (Jones et al., 2006), and details about how data were processed (Pasquetto et al., 2019) can help potential users decide whether a dataset is right for them and how best to use it.

Archived data have varying levels of usability. Large, uncurated data collections that rely solely on the contributor to prepare the data and documentation may be only minimally accessible. ICPSR invests significant resources curating the data in its archives, and overall ICPSR observes high use of its collections: for instance, 36,190 unique users downloaded 660,946 data files in 2020. However, even ICPSR applies curation in varying intensity across studies, guided by the state of the data deposited, the expected interest in the dataset, and the resources available for a particular study.

1.3 | Our contributions

In this paper, we asked: How do *data attributes*, *curatorial decisions*, and *archive funding models* impact research *data usage*? Based on prior literature about the impacts of curation on data reuse, we predicted that several data attributes—specifically being part of a series, having more variables, deposited by institutions, and having more metadata terms—would be associated with higher data usage. We also predicted more downloads for data that were subject to more curatorial actions and where external funding was available to support ingest, curation, and access. We found that data attributes, curation level and number of subject terms, and external funding were associated with more data usage.

2 | MATERIAL AND METHODS

2.1 | Data overview

We analyzed data usage for 380 studies released by ICPSR from January 1, 2017, to April 30, 2021. We limited our analysis to those studies that had data files available for download to any ICPSR member or the public (i.e., no studies with only restricted use data). We computed the number of "data users" for each study in our sample as follows: extract all unique download users, defined as a unique user downloading one or more data files associated with a study between January 1, 2017, and April 30, 2021, from ICPSR's administrative web statistics. Table 1 presents the number of studies released and data users by year, and Figure 1 shows the frequency distribution of total data users by study. Uniqueness was based on IP address. Users must login to ICPSR's website to download data, which allows us to exclude ICPSR staff downloads from our analysis.

ICPSR provided use data from its administrative database, which contains information on study characteristics related to data that are stored as study and/or variable

metadata. The data include properties of the data, descriptions of work ICPSR performed, how the work was funded, and how many users accessed the data through ICPSR's website. Table 2 provides definitions of the variables used in our analysis. We do not include data about studies that are housed in other archives, ICPSR faculty or staff use, restricted-access datasets, or self-published datasets in openICPSR. We selected January 1, 2017, as a start date for the sample because it reflects the beginning of ICPSR's transition to centralized curation. In the new organizational structure, a

centralized group of curatorial staff record details about curatorial actions taken on data being prepared for dissemination; their standardized set of records make this an ideal dataset for our analysis. Prior to this change, curation decisions were not recorded centrally, and curation staff worked independently of one another. They reported to different supervisors and used their own processes; some tools and standards were still shared across curation staff.

Over the period of analysis, ICPSR instituted several changes to its curation policies. In 2018, ICPSR implemented standardized curation levels and terminology (ICPSR, 2020); we have harmonized curation level information from 2017 to the 2018 levels. We understand that higher levels of data curation at ICPSR are more extensive, demanding more effort and staff time spent on curation activities (Lafia et al., 2021). Level 1 studies receive ICPSR's base level of curation and can generally be disseminated more quickly, while Level 3 is ICPSR's most extensive level of curation. In 2018, ICPSR limited the number of subject terms that the data curators can apply to a study (15 subject

TABLE 1 Number of studies released and data users by year

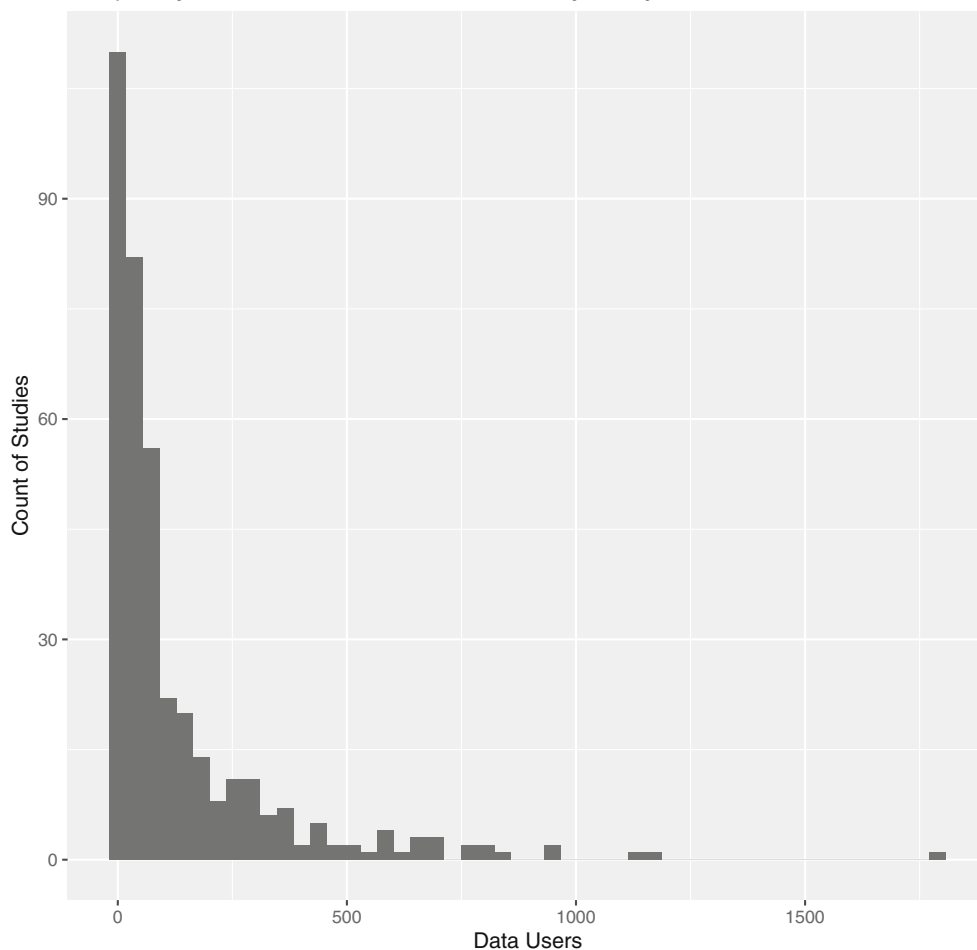| Release year	Studies	Data users
2017	73	15,493
2018	120	19,389
2019	58	7526
2020	97	7463
2021	32	354
Total	380	50,225

FIGURE 1 Frequency distribution of total data users by study

TABLE 2 Variables and their definitions

Variable type	Variable	Definition
Data attributes	*Series*	1 = Study is part of a recurring serial collection with new data archived over time (e.g., repeated cross-sectional studies, longitudinal studies); 0 = Study is not part of a series
	Institutional PI	1 = At least one of the study's principal investigators or depositors is an institution (e.g., U.S. Bureau of the Census); 0 = All of the study's principal investigators are individuals
	Number of variables	Number of variables in the study indicating the size of the study (note: Qualitative studies have zero variables; our sample includes 35 qualitative studies)
Curatorial decisions	*Number of subject terms*	Number of metadata subject terms assigned by staff (including terms supplied by data contributor) to the study, indicating scope
	Curation Level	Level of curation for the study indicating the set of curation activities performed in preparing the study where 3 indicates the most activities and 1 the fewest. Rarely, data and documentation are released in the format provided by the data producer, and these studies are called "fast release" (FR). Level 3, the highest level of curation, serves as the reference group in our regression models.
	SSVD	1 = Variable-level metadata, including variable name, label, and value labels, are indexed for search in ICPSR's social science variable database; 0 = Variables are not indexed for search
	Question text	1 = Question text from data collection instruments or other source documentation manually generated for all variables; 0 = No question text available for search
	SDA	1 = Study data has been processed, compiled, and made available for online analysis; 0 = Not available for online analysis
Archive funding model	*External funder*	1 = Study was released by an externally-sponsored, topical archive (e.g., National Archive of Criminal Justice Data) rather than the member-sponsored archive (i.e., General Archive or Resource Center for Minority Data); 0 = Study was deposited in the ICPSR membership archive
Control variable	*Days*	Number of days the study has been available (from study release to data pull date)
Dependent variable	*Total data users*	Number of unique users that downloaded quantitative data files, specifically, from the study between January 2017 and April 2021

terms); data depositors are able to add their own subject terms as well.

Descriptive information about study attributes is presented in Table 3. The studies we analyzed were distributed across release years; data for 2021 including only studies released on or before April 30. Nearly two-third of studies are part of a series and do not have an institutional PI. The studies are also distributed across levels of curation (1–3). Nearly all studies have variables indexed for search in a public database (the Social Science Variable Database, SSVD); less than half are available for online analysis (Survey Documentation Analysis, SDA). Just over half the studies have complete question text.

About three-fifths of studies are housed in an externally sponsored, topical archive at ICPSR; about 40% are in member-funded archives.

2.2 | Statistical analysis

We used negative binomial regression to analyze the relationships between *data attributes*, *curatorial decisions*, *archive funding models*, and *data reuse*. We present four models[1] of reuse; in each model, the dependent variable is the number of users who downloaded data files. Model 1 included attributes of the data; Model 2 included

TABLE 3 Descriptive statistics for data attributes, curatorial decisions, funding models, and data use

	Overall (*N* = 380)
Series	
No	138 (36.3%)
Yes	242 (63.7%)
Number of variables	
Mean (SD)	1328.158 (3395.758)
Range	0.000–34094.000
Institutional PI	
No	212 (55.8%)
Yes	168 (44.2%)
Curation level	
Level 1	82 (21.6%)
Fast release	11 (2.9%)
Level 2	133 (35.0%)
Level 3	154 (40.5%)
Number of subject terms	
Mean (SD)	12.053 (7.654)
Range	2.000–48.000
SSVD	
No	21 (5.5%)
Yes	359 (94.5%)
Question text	
No	185 (48.7%)
Yes	195 (51.3%)
SDA	
No	211 (55.5%)
Yes	169 (44.5%)
External funder	
No	150 (39.5%)
Yes	230 (60.5%)
Total data users	
Mean (SD)	132.171 (207.820)
Range	0.000–1790.000

TABLE 4 Regression results

	Dependent variable: total_data_users
Series (yes)	0.891
Number of variables	1.000**
Institutional PI (yes)	1.322**
Curation Level (fast release)	0.345**
Curation level (Level 1)	1.154
Curation level (Level 2)	0.617**
Number of subject terms	1.031***
SSVD (yes)	0.777
Question text (yes)	1.342*
SDA (yes)	0.750*
External funder (yes)	4.273***
Curation level (fast release):External funder (yes)	0.744
Curation level (Level 1): External funder (yes)	0.606*
Curation level (Level 2): External funder (yes)	0.967
Constant	0.060***
Observations	380
Log likelihood	−2,063.611
Theta	0.959*** (0.064)
Akaike inf. crit.	4,157.222

*$p < .1$; **$p < .05$; ***$p < .01$.

curatorial actions; and Model 3 included a measure of the archive funding model. Model 4 included all three sets of measures and is the model of best fit (using AIC). In all models, we controlled for the number of days a study had been available by using an offset of *ln*(days).

3 | RESULTS

We found that *data attributes*, *curatorial decisions*, and *archive funding models* correlated with *data reuse*.

Table 4 shows the results of the best-fit regression model; results for other models are available in the Appendix. Data that contain more variables and/or are collected by an institutional PI are correlated with greater data reuse.

More curation actions (Level 3, the reference group in Table 4), adding question text, and attaching subject terms also correlated with more data reuse. Having online analysis available is significantly negatively correlated with downloads; studies with SDA are downloaded 25% less often.

External funding for archives is also positively correlated with data reuse. Studies in archives that are funded externally are downloaded over twice as often as member-funded studies.

The interaction between curation level and external funder negatively correlates with fewer downloads when we hold other variables constant—external funding and level 3 curation are associated with more downloads. This interaction may be easier to understand visually, and we provide the marginal effects plot in Figure 2. The figure makes clear that there is more variation in the number of downloads of externally funded studies than

FIGURE 2 Marginal effects of curation level and external funder on data download numbers

Predicted counts of total_data_users
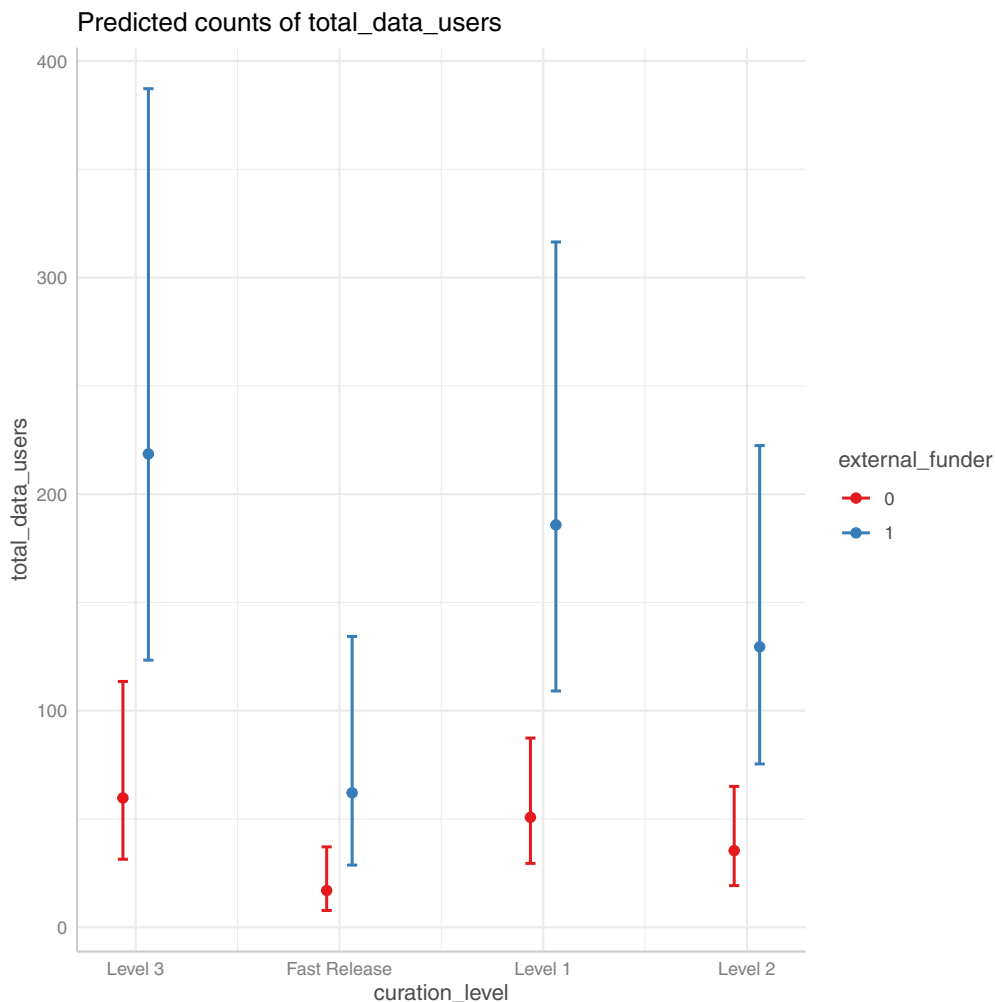
in ICPSR-funded studies. External funding is not associated with more downloads among studies with limited curation (i.e., Fast Release).

Overall, having an institutional PI, receiving Level 3 curation, and having an external funder mean that institutions invested in a dataset's collection and deposit, ICPSR invested time in its curation, and external funders supported ICPSR's efforts. These efforts correlate with more downloads. The effects of additional curation activity are stronger when coupled with external funding.

4 | DISCUSSION

Investments in data—through institutional data collection, curation, and external funding of archive functions—correlate with higher levels of data reuse measured by additional downloads. We analyzed data attributes, curation activities, and archive funding models to determine the impact of each on data reuse. Our results show that the combination of more extensive curation and the use of external funding is strongly

associated with higher data downloads. Datasets that get more ICPSR curation effort, with or without external funding, are downloaded more often; highly curated data that also have external funding are downloaded more often than uncurated data or ICPSR-funded data.

4.1 | Additional curation correlates with more data downloads

Why does curation matter? To understand what specifically about curation explains the correlation between more intense curation and more data downloads, we look specifically at the FAIR principles (Wilkinson et al., 2016); the original principles focus on machine-readable metadata, and here we consider findability, accessibility, interoperability, and reusability more generally. ICPSR's curation activities are geared toward these principles, and our results show that making data findable by attaching subject terms has the biggest impact. Other efforts to make data findable and interoperable, such as indexing in the SSVD and attaching question text, showed mixed results. Indexing

variables was not related to downloads, but attaching question text did correlate with more downloads. Prior work emphasized that social scientists often look for a single question within a survey when deciding to reuse data (Gregory et al., 2019), and attaching question text facilitates this type of search and evaluation. All studies with question text also received Level 3 curation; the regression results indicate that attaching question text leads to roughly 30% more downloads than Level 3 curation alone.

Earlier research emphasized the importance of subject terms in data reusers' searches (Gregory et al., 2019; Pienta et al., 2017). Our findings confirm that subject terms are especially important for connecting reusers with data: each new subject term was related with a 3% increase in downloads. It may be that ICPSR is effective at identifying datasets that would benefit from curation; ICPSR likely invests in datasets that they expect to have more utility. Both overall curation effort (measured by level) and individual actions (i.e., attaching subject terms) correlated with more reuse.

4.2 | Dataset- and collection-specific funding correlates with more data downloads

Why does funding make such a difference? Many funders require that data collected in projects they support be available beyond ICPSR's membership, and we know that more open access is expected to correlate with more use (Turner et al., 2015). Funders may also generate demand by hosting workshops that help researchers discover and use datasets and build a community of users; all externally funded datasets are also publicized by at least two marketing organizations (ICPSR's and the funder's). Prior work underscores the importance of communities of use in facilitating reuse by helping transfer tacit knowledge about the data's context, collection processes, and particular peculiarities (Gregory et al., 2020; Pasquetto et al., 2019). We cannot make a causal claim about funding, but either funders effectively prioritize datasets worth their investment and/or their investments generate demand for the data. Caring for data is an expensive endeavor, and strategic investments may pay off in greater reuse.

Our results do not allow us to make general causal claims about the connections between investments in data before and after deposit. For instance, having an institutional PI may correlate with more downloads because the kind of data institutions collect are already in high demand (e.g., census data, national probability sample surveys). These data may also have established communities of use through institutional affiliates. External funding

often includes additional promotional activities and outreach efforts from the funder. However, we are able to say that when all other variables are held constant, both institutional PIs and external funding are associated with more downloads. The effect of external funding is limited to curated studies, however. The interaction term and margin plots reveal that "fast release" studies that receive external funding are not downloaded more often than their ICPSR-funded counterparts. This suggests that funding alone is not sufficient but must be accompanied by improvements to metadata and findability.

4.3 | Broad audiences and institutional deposits correlate with more data downloads

Beyond funding and curation, datasets that are designed to appeal to broad audiences—those with more variables that were produced by institutions—also attract more users. Data reusers judge whether the original data collectors were competent and trustworthy (Yoon, 2017), and institutional deposits may be seen as more trustworthy than individual PIs'. Our findings are in line with Fear's (2013) earlier study that found study size and curation correlated with additional use.

Making data available for online analysis is correlated with fewer downloads, suggesting that a significant proportion of users meet their data needs through online analysis and do not need to download and work with data locally. Or, users may use online analysis to explore the data and decide they are not right for the project. Gregory et al. (2020) found that nearly 75% of surveyed researchers used exploratory analysis to determine a dataset's fitness for use. Offering online analysis likely facilitates this exploration. It is also helpful to know that offering online analysis reduces downloads because some data—large data or sensitive data, for instance—are safer and more manageable when they stay in one place. Our results indicate that making the data available for analysis rather than for download could be an effective way to make data accessible while ensuring reuse. Online analysis reduces the bandwidth and computing resources that researchers must have locally, making large and sensitive data more accessible. The relationship between downloads and online analysis suggests a broader definition of "reuse" than using existing data to studying new problems (Zimmerman, 2008). Instead, "reuse" may include data exploration and advancing one's thinking about a topic; this breadth is in line with van de Sandt et al.'s (2019) proposal to define reuse as "use of any research resource regardless of when it is used, the purpose, the characteristics of the data and its user" (p. 14).

4.4 | Limitations and future directions

We provide initial evidence for the impact of data curation on data reuse. We found several curation actions, such as the inclusion of question text, which contribute to data reuse. More analysis is needed to explain the ways in which data curation interacts with other factors, such as users' considerations when assessing and selecting data to reuse. Prior work emphasizes the importance of data communities and metadata in researchers' decisions to reuse data (Gregory et al., 2020; Pasquetto et al., 2019). While we confirmed that more intense curation, and the metadata improvements it brings, is generally related to data reuse, future analysis should focus on the impacts of specific metadata additions such as variable descriptions or processing notes. Work in this vein should also examine users' behavior. For instance, a future study could use interviews with data users or the digital traces ICPSR users create in their searches for data to understand what paths users take through the archive, which types of searches (e.g., keywords, variable names) are most successful, and what role online analysis plays in download decisions.

We acknowledge limitations with our current study that restrict our analyses. First, we used data downloads to measure data use. While data downloads are a widely accepted data usage metric (Cousijn et al., 2019; Fenner et al., 2018), they imply access rather than analysis of the data. To complement this analysis of downloads, we are also analyzing data citations, work that requires a comprehensive and representative bibliography of literature citing the data. Analyzing data citations offers richer insights into the ways that data are used (e.g., to reproduce an existing analysis versus extending a method) and disciplinary differences in data use practices.

We also excluded self-published data (i.e., data deposited in openICPSR), which is not curated, and restricted-use data from our analyses. Self-published data are often shared to satisfy publisher requirements and are not subject to similar selection criteria that curated data deposits must meet. Restricted-use data access is tracked through a separate system at ICPSR, and collecting usage information requires substantial additional data collection. Future work should incorporate additional usage metrics and deposit types to account for potential differences in how self-published and restricted-use data are used.

Data archives must make decisions about how to best allocate their limited resources. Some readers may be tempted to interpret our results as suggesting that archives should focus on large, institutional datasets. However, such a read is potentially dangerous. Science benefits from transparency and open access, and those benefits include small, single-author, low-resource datasets. We suggest that rather than limiting their focus to large, institutional datasets, archives must work to improve the efficiency of curation to ensure that even small, single-author datasets are FAIR. Better tools and practices during data collection and management, such as documenting data processing, will reduce the burden on archives to apply curatorial actions.

5 | CONCLUSION

Archives have responsibilities to use resources efficiently, and understanding the impacts of different investments in data can inform their decision-making. Our analysis suggests that investments in data curation pay off. Level 3 curation, the most extensive level of curation, is most closely associated with more data downloads; external funding is also associated with more downloads, but only when data also undergo curation. Providing online analysis is an effective way to provide access without requiring data downloads. Datasets from trusted sources, like institutions, are in greater demand than those produced by individuals. In conclusion, our data suggest that (a) actively curating data, especially by attaching subject terms, (b) partnering with external funders, and (c) recruiting deposits from institutional data producers are steps archives can take to increase data downloads.

AUTHOR CONTRIBUTIONS
Conceptualization: Amy Pienta, Libby Hemphill, and Dharma Akmon; methodology: Libby Hemphill, Sara Lafia, Amy Pienta, and David A. Bleckley; resources: David A. Bleckley and Libby Hemphill; data curation: David A. Bleckley; writing—original draft: Libby Hemphill, Amy Pienta, Dharma Akmon, Sara Lafia, and David A. Bleckley; supervision: Amy Pienta and Libby Hemphill; project administration: Libby Hemphill and David A. Bleckley; funding acquisition: Libby Hemphill, Dharma Akmon, and Amy Pienta

ORCID

Libby Hemphill https://orcid.org/0000-0002-3793-7281
Amy Pienta https://orcid.org/0000-0003-1174-6118
Sara Lafia https://orcid.org/0000-0002-5896-7295
Dharma Akmon https://orcid.org/0000-0002-1359-0586
David A. Bleckley https://orcid.org/0000-0001-7715-4348

ENDNOTE

[1] The Appendix shows the results of all models. We include only the model of best fit here.

REFERENCES

Akmon, D., Zimmerman, A., Daniels, M., & Hedstrom, M. (2011). The application of archival concepts to a data-intensive environment: Working with scientists to understand data management and preservation needs. *Archives Des Sciences/Editees Par La Societe de Physique et D'histoire Naturelle de Geneve*, 11(3–4), 329–348.

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS One*, 9(3), e92590.

Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: Supporting sharing in science and engineering. *GROUP '03: Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, Sanibel Island, Florida, USA*, 339–348.

Bishop, L., & Kuula-Luumi, A. (2017). Revisiting qualitative data reuse: A decade on. *SAGE Open*, 7(1), 2158244016685136.

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2018). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, 70(8), 888–904.

Borgman, C. L., Van de Sompel, H., Scharnhorst, A., van den Berg, H., & Treloar, A. (2015). Who uses the digital data archive? An exploratory study of DANS. *Proceedings of the Association for Information Science and Technology*, 52(1), 1–4. https://doi.org/10.1002/pra2.2015.145052010096

Carlson, S., & Anderson, B. (2007). What are data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication: JCMC*, 12(2), 635–651.

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing citations and usage metrics together to make data count. *Data Science Journal*, 18, 9. https://doi.org/10.5334/dsj-2019-009

Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of Documentation*, 75(6), 1274–1297.

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, 67(6), 1404–1416.

Fear, K. M. (2013). *Measuring and anticipating the impact of data reuse.* (Unpublished doctoral dissertation). University of Michigan. https://deepblue.lib.umich.edu/handle/2027.42/102481

Fenner, M., Lowenberg, D., Jones, M., Needham, P., Vieglais, D., Abrams, S., Cruse, P., & Chodacki, J. (2018). Code of practice for research data usage metrics release 1. *PeerJ.*, 6, e26505v1. https://doi.org/10.7287/peerj.preprints.26505v1

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P.,

Hedstrom, M., Hogg, D. W., Kashyap, V., Mahabal, A., Siemiginowska, A., & Slavkovic, A. (2014). Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology*, 10(4), e1003542.

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching data: A review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology*, 70(5), 419–432. https://doi.org/10.1002/asi.24165

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or found? Discovering data needed for research. *Harvard Data Science Review*, 2(2). https://doi.org/10.1162/99608f92.e38165eb

He, L., & Han, Z. (2017). Do usage counts of scientific data make sense? An investigation of the dryad repository. *Library Hi Tech*, 35(2), 332–342.

ICPSR. (2020). *ICPSR curation levels.* https://www.icpsr.umich.edu/files/datamanagement/icpsr-curation-levels.pdf

ICPSR. (2021). *ICPSR collection development policy.* https://www.icpsr.umich.edu/web/pages/datamanagement/policies/colldev.html

Jones, M. B., Schildhauer, M. P., Reichman, O. J., & Bowers, S. (2006). The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics*, 37(1), 519–544. https://doi.org/10.1146/annurev.ecolsys.37.091305.110031

Kitchin, R., Collins, S., & Frost, D. (2015). Funding models for open access digital data repositories. *Online Information Review*, 39(5), 664–681. https://doi.org/10.1108/OIR-01-2015-0031

Kratz, J. E., & Strasser, C. (2015). Making data count. *Scientific Data*, 2(1), 1–5. https://doi.org/10.1038/sdata.2015.39

Lafia, S., Thomer, A., Bleckley, D., Akmon, D., & Hemphill, L. (2021). Leveraging machine learning to detect data curation activities. *Proceedings of 2021 IEEE 17th international conference on E-science. eScience* 2021, virtual. http://arxiv.org/abs/2105.00030

McLure, M., Level, A. V., Cranston, C. L., Oehlerts, B., & Culbertson, M. (2014). Data curation: A study of researcher practices and needs. *Portal: Libraries and the Academy*, 14(2), 139–164.

National Science Foundation. (2020). *Proposal and award policies and procedures guide (NSF publication 20-1).* https://www.nsf.gov/pubs/policydocs/pappg20_1/nsf20_1.pdf

Niu, J. (2009). Overcoming inadequate documentation. *Proceedings of the American Society for Information Science and Technology*, 46(1), 1–14.

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*, 69(11), 1346–1354.

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review*, 1(2), 2–39. https://doi.org/10.1162/99608f92.fc14bf2d

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data. *Data Science Journal*, 16, 8. https://doi.org/10.5334/dsj-2017-008

Pienta, A., Akmon, D., Noble, J., Hoelter, L., & Jekielek, S. (2017). A data-driven approach to appraisal and selection at a domain data repository. *International Journal of Digital Curation*, 12(2), 362–375. https://doi.org/10.2218/ijdc.v12i2.500

Plantin, J.-C. (2019). Data cleaners for pristine datasets: Visibility and invisibility of data processors in social science. *Science, Technology & Human Values*, *44*(1), 52–73. https://doi.org/10.1177/0162243918781268

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, *67*(12), 2964–2975.

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6–20.

Sorbonne Declaration on Research Data Rights (2020). https://sorbonnedatadeclaration.eu/

Thomer, A. K., Akmon, D., York, J., Tyler, A. R. B., Polasak, F., Lafia, S., Hemphill, L., & Yakel, E. (2022). The craft and coordination of data curation: Complicating "workflow" views of data science. In *Deep Blue Documents*. University of Michigan. https://doi.org/10.7302/4017

Turner, W., Rondinini, C., Pettorelli, N., Mora, B., Leidner, A. K., Szantoi, Z., Buchanan, G., Dech, S., Dwyer, J., Herold, M., Koh, L. P., Leimgruber, P., Taubenboeck, H., Wegmann, M., Wikelski, M., & Woodcock, C. (2015). Free and open-access satellite data are key to biodiversity conservation. *Biological Conservation*, *182*, 173–176. https://doi.org/10.1016/j.biocon.2014.11.048

U.S. Department of Health and Human Services. (2003). *Final NIH statement on sharing research data (NIH notice NOT-OD-03-032)*. National Institutes of Health. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

van de Sandt, S., Dallmeier-Tiessen, S., Lavasa, A., & Petras, V. (2019). The definition of reuse. *Data Science Journal*, *18*(1), 22. https://doi.org/10.5334/dsj-2019-022

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

Yoon, A. (2017). Data reusers' trust development. *Journal of the Association for Information Science and Technology*, *68*(4), 946–956.

Yoon, A., & Kim, Y. (2017). Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. *Library & Information Science Research*, *39*(3), 224–233.

Zimmerman, A. S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, *33*(5), 631–652.

APPENDIX: Regression results for other models tested

	Dependent variable:						
	total_data_users						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Series (yes)	1.283*			0.836		0.895	0.891
Number of variables	1.000***			1.000***		1.000**	1.000**
Institutional PI (yes)	1.537***			1.277*		1.331**	1.322**
Curation level (fast release)		0.152***		0.145***	0.301***	0.284***	0.345**
Curation level (Level 1)		0.438***		0.419***	0.846	0.850	1.154
Curation level (Level 2)		0.527***		0.518***	0.606***	0.593***	0.617**
Number of subject terms		1.019**		1.017**	1.032***	1.029***	1.031***
SSVD (yes)		0.607*		0.664	0.685	0.751	0.777
Question text (yes)		1.007		1.031	1.273	1.282	1.342*
SDA (yes)		0.515***		0.538***	0.699**	0.747**	0.750*
External funder (yes)			4.246***		3.783***	3.660***	4.273***
Curation level (fast release): External funder (yes)							0.744
Curation level (Level 1): External funder (yes)							0.606*
Curation level (Level 2): External funder (yes)							0.967
Constant	0.120***	0.472**	0.068***	0.401***	0.087***	0.073***	0.060***
Observations	380	380	380	380	380	380	380
Log likelihood	−2,137.064	−2,115.939	-2,095.893	-2,110.307	-2,069.210	-2,065.205	-2,063.611
Theta	0.703*** (0.045)	0.768*** (0.049)	0.833*** (0.054)	0.786*** (0.051)	0.937*** (0.062)	0.954*** (0.063)	0.959*** (0.064)
Akaike inf. crit.	4,282.127	4,247.877	4,195.787	4,242.614	4,156.420	4,154.410	4,157.222

* $p < .1$; **$p < .05$; ***$p < .01$.