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

Authors

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

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

2. School of Information (UMSI), University of Michigan, Ann Arbor, USA

Author contributions

Conceptualization, A.P., L.H., and D.A.; Methodology, L.H., S.L., A.P., , and D.B.; Resources, D.B. and L.H.; Data Curation, D.B.; Writing - Original Draft, L.H., A.P., D.A., S.L., and D.B.; Supervision, A.P. and L.H.; Project Administration, L.H. and D.B.; Funding Acquisition, L.H., D.A., and A.P.

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.

Keywords
data archives, data curation, data sharing, data metrics, data reuse, value of curation, FAIR principles, administrative records

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. 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.

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). In this paper, we adopt downloads as a measure of data use and 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's 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., 2018); this pattern suggests that attaching subject terms to data will make it more discoverable. Data users also often turn to data that is 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.

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 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.

Study Background

Data reuse encompasses many activities, including exploring new research questions, planning new research projects or data collection efforts, teaching students, verifying results, and providing the curious with more information about the data that underlie published results. Accordingly, one can imagine a number of 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, 12/2015; 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 ten 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 based on the "traces they leave in the system" (Borgman et al., 2015). 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 (Borgman et al., 2015). 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, M., Lowenberg, D., Jones, M., Needham, P., Vieglais, D., Abrams, S., Cruse, P. Chodacki, J., 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)."

Deep Blue DOI: http://dx.doi.org/10.7302/1639

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 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 (Fear, 2013). 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.

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. Here, we test the relationship between data attributes, archival decisions, resources for curation, and data usage.

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.

Results

We found that *data attributes*, *curatorial decisions*, and *archive funding models* correlated with *data reuse*. Table 1 shows the results of the best-fit regression model; results for other models are available in Supplementary materials. Data that contain more variables and/or are collected by an institutional PI are correlated with greater data reuse.

Base level curation (Level 1, the reference group in Table 1), 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.

Table 1. Regression Results

Deep Blue DOI: http://dx.doi.org/10.7302/1639

	Dependent variable:
	total_data_users
series1	0.891
vars	1.000**
inst_pi1	1.322**
curation_levelFast Release	0.299***
curation_levelLevel 2	0.534**
curation_levelLevel 3	0.867
numterms	1.031***
ssvd1	0.777
qtext1	1.342*
sda1	0.750*
external_funder1	2.590***
curation_levelFast Release:external_funder1	1.228
curation_levelLevel 2:external_funder1	1.595
curation_levelLevel 3:external_funder1	1.650*
Constant	0.070***
Observations	380
Log Likelihood	-2,063.611
theta	0.959*** (0.064)
Akaike Inf. Crit.	4,157.222
Note:	*p<0.1; **p<0.05; ***p<0.01

The interaction between curation level and external funder positively correlates with more downloads when we hold other variables constant. This interaction may be easier to understand visually, and we provide the marginal effects plot in Figure 1. 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.

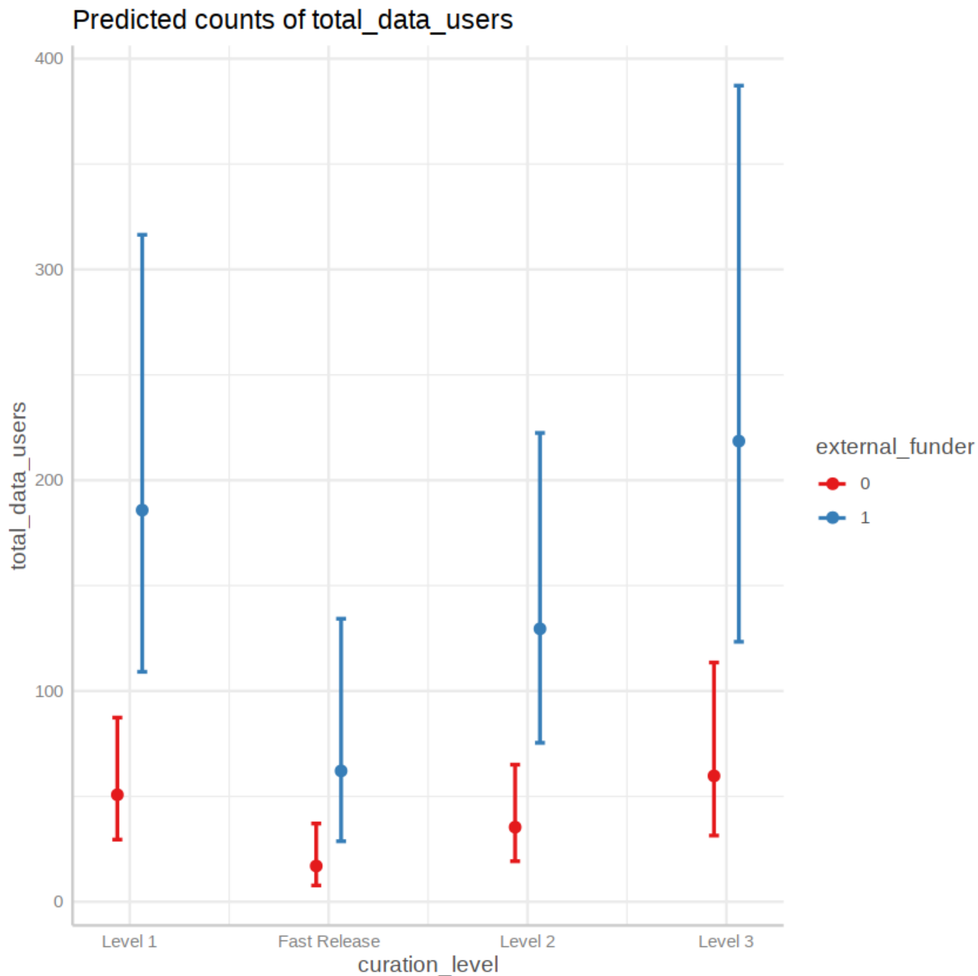

Figure 1. Marginal effects of curation level and external funder on data download numbers

Discussion

We analyzed data attributes, curation activities, archive funding models to determine their impacts on data reuse as measured by downloads. Overall, we found that investments in data–through institutional data collection, ICPSR curation, and external funding of archive functions–correlated with additional use. Archives have responsibilities to use resources efficiently, and understanding the impacts of different investments in data can inform their decision-making. Specifically, our findings suggest that partnering with external institutions and completing more curation tasks (especially attaching subject terms) are actions digital archives can take to ensure high levels of data reuse.

Datasets that are designed to appeal to broad audiences–those with more variables, that were produced by institutions, and that receive dedicated funding–attract more users. We cannot say definitively whether the larger investments in these data cause more reuse, but our data do suggest that intentional investments pay off. Data reusers judge whether the original data collectors were competent and trustworthy (Yoon, 2017), and institutional deposits may be seen

Deep Blue DOI: http://dx.doi.org/10.7302/1639

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.

Investments in curating data correlate with more data reuse. When ICPSR invests curation effort, with or without external funding, datasets are downloaded more often. It may be that ICPSR is effective in 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.

Having external funding support for archival functions was the strongest predictor of reuse; studies in archives with external funding were more than twice as likely to be downloaded as other datasets. Many funders also require that data they support be publicly available, and open access does correlate with more use. Funders may also generate demand by hosting workshops that help researchers discover and use datasets; all externally funded datasets are also publicized by at least two marketing organizations (ICPSR’s and the funder’s). Again we cannot make a causal claim here, 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.

External funding of repository services is clearly related to more data downloads. How does curation help ensure data reuse? We look specifically at its impacts on the FAIR principles (Wilkinson et al., 12/2016); the original principles focus on machine-readable metadata, and here we consider findability, accessibility, interoperability, and reusability more generally.

Efforts to make data more 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. 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 (Pienta et al., 2018). 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.

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. It’s helpful to know that offering online analysis reduces downloads because some data, large data or sensitive data for instance, is safer and more manageable when it stays 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.

Our analysis suggests that investments in data curation pay off whether they are from external sources or through consortium efforts like ICPSR. Level 1 curation, the base level of curation, is most closely associated with data downloads; additional curation does not seem to impact downloads unless accompanied by resources from an external funder. 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 (1) partnering with external funders, (2) performing a base level of curation and attaching subject terms, and (3) recruiting deposits from institutional data producers are steps archives can take to increase data downloads.

Material and Methods

Data Overview

We analyzed data usage for 380 studies released by ICPSR from January 1, 2017 - 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 - April 30, 2021, from ICPSR’s administrative web statistics. 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.

Table 2. Number of studies released and downloads 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

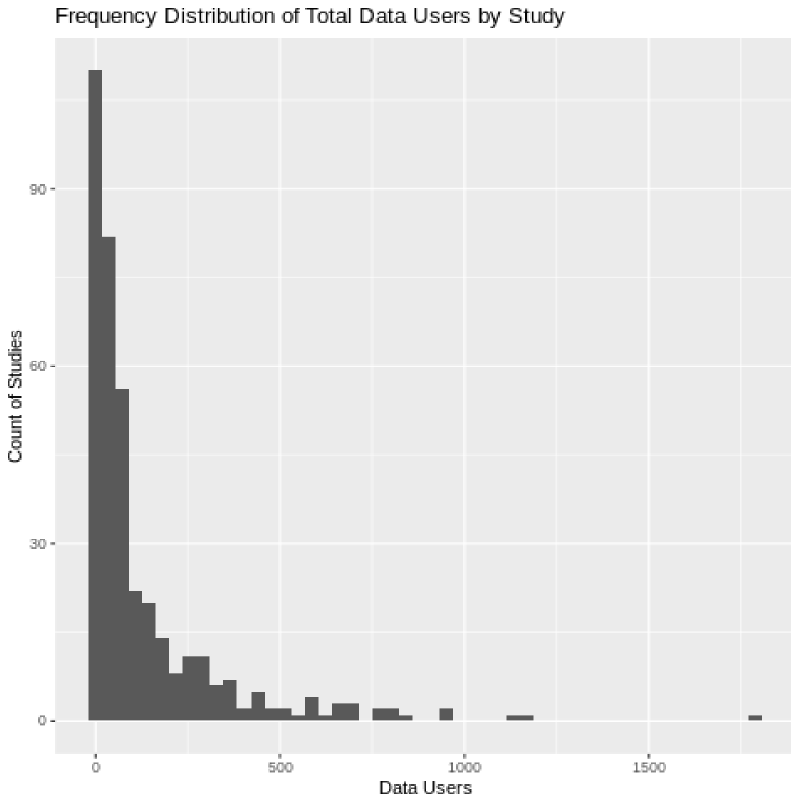

Figure 2. Frequency distribution of total data users by study

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 includes 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. 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 structure, a centralized group of curatorial staff record details about curatorial actions taken on data being prepared for dissemination; their records make this an ideal dataset for our analysis.

Table 3. 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., United States Bureau of the Census); 0 = All of the study's principal investigators are individuals
	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	*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). FR 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 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.

Over the period of analysis, ICPSR instituted several changes to its curation policies. In 2018, ICPSR implemented standardized curation levels and terminology; 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 terms); data depositors are able to add their own subject terms as well.

Descriptive information about study attributes are presented in Table 4. 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.

Table 4. Descriptive statistics for data attributes, curatorial decisions, funding models, and data use

	Overall (N=380)
series	
0	138 (36.3%)
1	242 (63.7%)
vars	
Mean (SD)	1328.158 (3395.758)
Range	0.000 - 34094.000
inst_pi	
0	212 (55.8%)
1	168 (44.2%)
curation_level	
Level 1	82 (21.6%)

Deep Blue DOI: http://dx.doi.org/10.7302/1639

Other	0 (0.0%)
Fast Release	11 (2.9%)
Level 2	133 (35.0%)
Level 3	154 (40.5%)
numterms	
Mean (SD)	12.053 (7.654)
Range	2.000 - 48.000
ssvd	
0	21 (5.5%)
1	359 (94.5%)
qtext	
0	185 (48.7%)
1	195 (51.3%)
sda	
0	211 (55.5%)
1	169 (44.5%)
external_funder	
0	150 (39.5%)
1	230 (60.5%)
total_data_users	
Mean (SD)	132.171 (207.820)
Range	0.000 - 1790.000

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 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 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).

Acknowledgements

We are grateful to Justin Noble at ICPSR for advice and preparation of data in an earlier draft and to Jeremy York for his feedback. Thank you to the team at the University of Michigan's Consulting for Statistics, Computing and Analytics Research (CSCAR) for reviewing our models and interpretations and to the Advanced Research Computing group for providing servers for our data and analysis. This material is based upon work supported by the National Science Foundation under grant 1930645, the Institute of Museum and Library Services grant number LG-37-19-0134-19, and the National Institute of Drug Abuse contract number N01DA-14-5576.

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., Van de Sompel, H., Scharnhorst, A., van den Berg, H., & Treloar, A. (2015). Who uses the digital data archive? An exploratory study of DANS. In *Proceedings of the Association for Information Science and Technology* (Vol. 52, Issue 1, pp. 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.

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*.

Fenner, M., Lowenberg, D., Jones, M., Needham, P., Vieglais, D., Abrams, S., Cruse, P. Chodacki, J. (2018). *Code of Practice for Research Data Usage*. https://www.projectcounter.org/wp-content/uploads/2019/02/Research_Data_20190227.pdf

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.

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.

Kratz, J. E., & Strasser, C. (12/2015). Making data count. *Scientific Data*, *2*(1). https://doi.org/10.1038/sdata.2015.39

Lafia, S., Thomer, A., Bleckley, D., Akmon, D., & Hemphill, L. (2021, April 30). 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.

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., Randles, B. M., & Borgman, C. L. (2017). *On the Reuse of Scientific Data*. https://doi.org/10.5334/dsj-2017-008

Pienta, A. M., Akmon, D., Noble, J., Hoelter, L., & Jekielek, S. (2018). A Data-Driven Approach to Appraisal and Selection at a Domain Data Repository. *International Journal of Digital Curation*, *12*(2). https://doi.org/10.2218/ijdc.v12i2.500

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.

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. (12/2016). 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

	Dependent variable:						
	total_data_users						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
series1	1.283*			0.836		0.895	0.891
vars	1.000***			1.000***		1.000**	1.000**
inst_pi1	1.537***			1.277*		1.331**	1.322**
curation_levelFast Release		0.348***		0.346***	0.356***	0.334***	0.299***
curation_levelLevel 2		1.204		1.235	0.716*	0.697**	0.534**
curation_levelLevel 3		2.284***		2.386***	1.182	1.176	0.867
numterms		1.019**		1.017**	1.032***	1.029***	1.031***
ssvd1		0.607*		0.664	0.685	0.751	0.777
qtext1		1.007		1.031	1.273	1.282	1.342*
sda1		0.515***		0.538***	0.699**	0.747**	0.750*
external_funder1			4.246***		3.783***	3.660***	2.590***
curation_levelFast Release:external_funder1							1.228
curation_levelLevel 2:external_funder1							1.595
curation_levelLevel 3:external_funder1							1.650*
Constant	0.120***	0.207***	0.068***	0.168***	0.073***	0.062***	0.070***
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
Note:							*p<0.1; **p<0.05; ***p<0.01

Deep Blue DOI: http://dx.doi.org/10.7302/1639