ISSN 2161-3974

JeSLIB 2021; 10(3): e1200
https://doi.org/10.7191/jeslib.2021.1200

Journal of eScience Librarianship
putting the pieces together: theory and practice

Commentary

An Insider's Take on Data Curation: Context, Quality, and Efficiency

Skylar Hawthorne

University of Michigan, Ann Arbor, MI, USA

Abstract

This commentary describes how context, quality, and efficiency guide data curation at the University of Michigan's Inter-university Consortium for Political and Social Research (ICPSR). These three principals manifest from necessity. A primary purpose of this work is to facilitate secondary data analysis but in order to so, the context of data must be documented. Since a mistake in this work would render any results published from the data inaccurate, quality is paramount. However, optimizing data quality can be time consuming, so automative curation practices are necessary for efficiency. The implementation of these principles (context, quality, and efficiency) is demonstrated by a recent case study with a high-profile dataset. As the nature of data work changes, these principles will continue to guide the practice of curation and establish valuable skills for future curators to cultivate.

Correspondence: Skylar Hawthorne: skyd@umich.edu

Received: March 25, 2021 **Accepted**: June 11, 2021 **Published**: August 11, 2021

Disclosures: The author reports no conflict of interest.

Introduction

I curate data for the world's largest social science data archive. Here at ICPSR, researchers and institutions affiliated with hundreds of universities entrust us with their data because our fastidious curation process increases the data's value, posterity, and potential for discovery and citation. As a curator, I've worked on high-impact studies across social science disciplines using psychology data from Yale, education data from Columbia, economic data from the University of Chicago, and much more. This experience has given me an insider's take on data curation at ICPSR and has elucidated three guiding principles: context, quality, and efficiency.

Documenting Data Context

Many social science researchers follow a linear workflow; they collect data, run statistical analyses, then publish their results. Curators, however, must consider what happens to data, how others can use it, and what they might use it for. The initial insights gleaned from data can be groundbreaking, but no single research team can maximize the full potential of their data. This is why secondary data analysis, in which researchers use existing data to discover new results, is so valuable (Gregory 2020).

For example, since the release of the U.S. Transgender Survey in 2015, an additional 54 data-related publications have built upon it. These include the first report on the lives of transgender people in rural America (Movement Advancement Project 2019), state policies and healthcare use among transgender people (Goldenberg et al. 2020), and even an insight into the association of transphobic discrimination and alcohol misuse among transgender adults (Kcomt et al. 2020). If the researchers for the U.S. Transgender Survey did not release their data, then other researchers would not have been able to use their data to discover novel insights.

However, data produced by other researchers can be difficult to use. In some cases, there are discrepancies between the data and its documentation. In other cases, variables are entirely undocumented. There may also be truncated variables, suspicious characters, open-ended strings that exceed character limits, and other quality concerns that inhibit potential users' abilities to understand the data (Sun and Khoo 2017). Furthermore, ethical concerns regarding confidentiality and privacy limit researchers' willingness to share and publish data in the social sciences (Akers and Doty 2013).

That's where data curators come in. We serve as mediators, translators, editors, publishers, librarians, and—of course—curators. We review the data for confidentiality concerns and redact disclosive information accordingly. We document undocumented variables so that researchers can interpret them. We identify discrepancies between the data and its documentation so the two can be reconciled. Curators at ICPSR also create a comprehensive codebook with

JeSLIB 2021; 10(3): e1200
https://doi.org/10.7191/jeslib.2021.1200

summary statistics for every variable. In short, we do everything we can to ensure the data is easy for other researchers to interpret and re-use.

That said, when the data comes to us, it isn't exactly raw. It has been "cooked" (in the sense that the Principal Investigators or organizations have, in many cases, collected and used it themselves), but the data rarely comes with a good recipe. Curators create recipes by documenting ingredients and transformation steps, which are collected in workflows (Plantin 2019). This enables other researchers to safely "consume" or recombine cooked data into new fusion dishes (secondary data publications).

Assuring Data Quality

There is no room for error in data curation. If a curator were to accidentally alter the original data in any way, then any results published from that research would be inaccurate. In this way, data curation is like filing taxes: it must be perfect. Such work requires a high level of attention to detail. ICPSR has developed an extensive quality check process that ensures deposited data meet the utmost standards for quality.

Quality checks are a deliberately iterative process that involve both human and computational effort. Redundancy is built into the curation workflow so that no errors slip through. First, curators here must pass proprietary scripts (in a process called Self-QC) that analyze the data for potential problems, like suspicious characters or duplicate variable labels. Then, the data gets submitted for a human review (called 1QC and/or 2QC). Here's the iteration: the reviewer compiles a log of revisions, the curator fixes the study to what they hope is perfection, and then the reviewer sends it back (because it is invariably short of perfection). Figure 1 shows how quality checks (Self-QC, 1QC, 2QC) are iterated in the curation workflow.

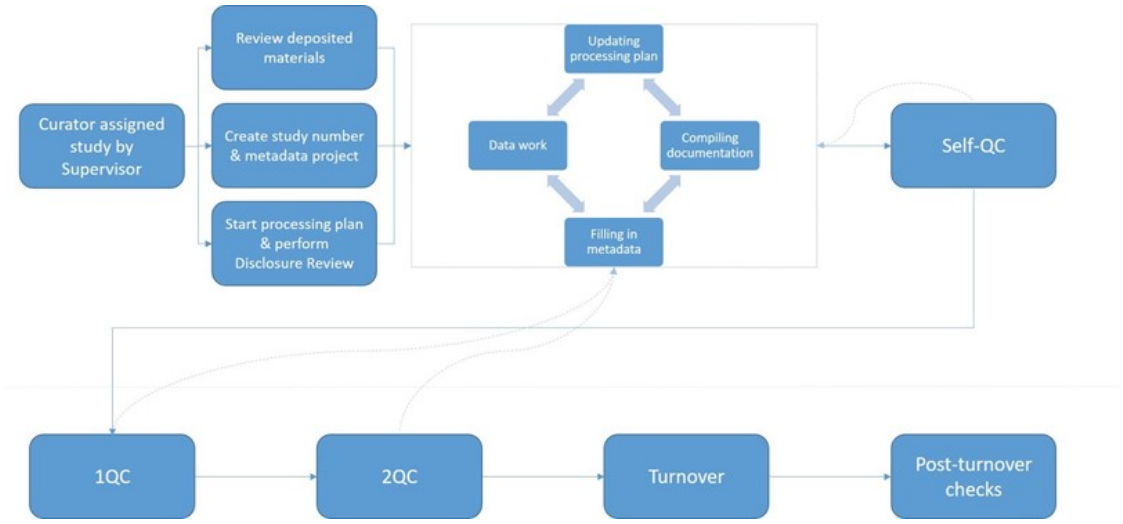

Figure 1: Overview of quality checks in a curation workflow.

Curators often joke about the fact that no matter how hard we try, no study is ever perfect after its first pass. Files must often be deleted, only to be created again. Because some studies have thousands of variable labels, we are at a statistical disadvantage—it's all too easy to miss just one small imperfection. We all know it's helpful to have another person (or script) double-check your work; in data curation, it is necessary.

Curators must also take great care to ensure the safety and privacy of respondents. For every study, we conduct a Disclosure Risk Review (DRR). This process involves checking for data on vulnerable populations, geographic variables, cases with low frequencies, and disclosive open-ended responses. In some cases, we'll run cross-tabulation analysis to determine whether the combination of any two variables in a study could be disclosive. When we do identify potentially disclosive data, we implement one of several DRR remediations. We might top-code variables with low upper frequencies (like income variables with wealthy outliers). We might mask a variable entirely (like specific geographic variables). If there's a time-constraint, open-ended string variables will receive this treatment (blanket masking) as a precaution. If time permits, curators will peruse every single open-ended string (no matter how many there might be) and selectively spot-mask the disclosive data. One such case resulted in this particularly amusing excerpt: "Privacy is not important to me and my name is [redacted]."

The treatment of a disclosive variable is contingent upon two factors that each have two dimensions (summarized in Figure 2). The first factor is release level: Studies are released for either restricted or public use. Since anybody can use public data, these studies have stricter DRR standards. The second factor is the degree of disclosiveness. This is determined by the risk of re-identification and the risk of harm. Still, the ultimate decision remains somewhat subjective and criteria have changed over time. I have found that for almost every study, my supervisor and I have a riveting conversation over whether or not a variable is disclosive and, if so, how it should be treated.

Efficiency in Curation

Given the staffing constraints that many repositories with curation services face (Johnston et al., 2017), efforts must often be directed toward the adoption of highly efficient workflows. At ICPSR, for instance, curators undergo a two-week-long formal training period to learn these best practices. But the learning process is continuous—even curators who have been at the archive for years continue to ask questions. The skill-set needed to curate data is multi-faceted and requires one to carefully balance operational efficiency with precision and accuracy (Henderson 2020).

A recent analysis of curation work logs at ICPSR found that time spent on data curation is not restricted to data transformation, but also includes a broader set of activities, including planning, communication, metadata, and quality reviews (Lafia

JeSLIB 2021; 10(3): e1200
https://doi.org/10.7191/jeslib.2021.1200

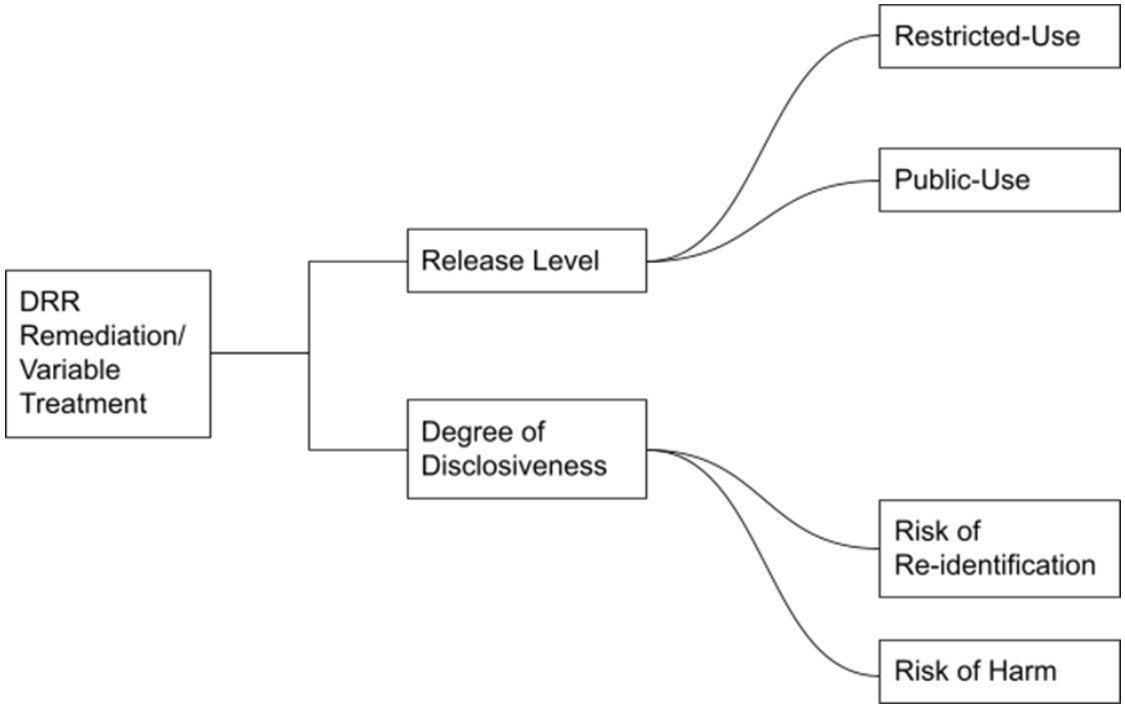

Figure 2: Considerations in disclosure risk reviews.

et al. 2021). To write metadata, curators must learn about the study, read all of its documentation, and extrapolate the relevant information including the study design, sample, and purpose. To write statistical syntax in SPSS, curators must be familiar with the language and the software. Curators must also be competent with PDF editors to create documentation. To do all of this expeditiously, curators must automate parts of the process.

Series of commands, code, or keyboard clicks can be executed simultaneously and repetitively. However, because this causes many changes at once, it can also cause mistakes that are easily missed. As we know, curation requires perfection; to mitigate the vulnerability of automation, ICPSR uses it with caution. Even so, we have several strategies for saving time.

Bash scripts are text files containing a series of commands to be run at once. By writing Linux commands in a text editor (as opposed to the command line), curators can use all the time-saving features of a text editor (like "find and replace"). Furthermore, a text file can be saved. Since curation is deliberately iterative, code must be frequently rerun; saving lists of code as bash scripts saves time.

Since curation involves code, it also involves trouble-shooting. My institute once circulated a meme that read, "Debugging: the classic mystery game where you are the detective, the victim, and the murderer." The process starts with the curator as the murderer, accidentally creating a bug that will cause catastrophe,

then curators become the victim when catastrophe erupts (usually just an error message). Finally, curators become the detective as they try to identify the bug. In other words, the curator is often simultaneously responsible for (unintentionally) creating the bug, identifying the problem, and arriving at the solution. A bug may take hours to find but just seconds to fix (like a missing period). It is up to the curator to (hopefully) avoid such problems in the first place and, when these problems inevitably arise, resolve them efficiently.

Case Study: Context, Quality, and Efficiency in Action

Since the complexities of data are intricate, it is difficult to determine how long it will take to curate a study. Given this, ICPSR rarely assigns due dates. However, I was once given a due date for the National Transgender Discrimination Survey. ICPSR wanted these data released in time for the Transgender Day of Remembrance to show our commitment to affecting real change for transgender rights. By curating these data and documenting their context, ICPSR hopes to help provide researchers with the data needed to advocate for the transgender community. With these data available at ICPSR, future researchers can continue to build upon the groundbreaking results of the initial release.

Though every study iterates through the quality check process, these data felt stuck. They appeared to be caught in a loop, bouncing back and forth between quality checks and corrections. Fortunately, I had written bash scripts for these data, so I didn't have to keep coding each command, but I still had to repetitively re-run the bash scripts.

The week before the due date, these data were still iterating through the quality-check process. Each time, I hoped that my quality reviewer wouldn't find anything that would require an extensive revision; that this iterative process would soon have a linear path to the end. Just one day before the Transgender Day of Remembrance, my reviewer returned the quality check. I opened it up with my fingers crossed, and there was nothing to revise. The study was in perfect condition. I ran the turnover script, released the study, and the next day, the data was made publicly available.

Outlook

Two years ago, there were twenty-one curators at ICPSR. Now, there are thirty-five, and soon there will be even more. The fields of data curation and data science are booming. Without data curation however, there is no data science (Donoho 2017). In fact, the majority of a data scientist's time is consumed with data curation (or data wrangling/munging) (Wickham 2014). This is what ICPSR takes care of, so that secondary researchers don't have to.

When I interviewed at ICPSR, I asked my interviewers about their favorite part of working here. Their answer: the data we get to work with. This has since proven true. Thanks to the constant influx of data from universities around the world,

curators can find their "niche." A curator with a background in criminology works with data from the Bureau for Justice Statistics. A curator with a passion for music curates data for the National Archive of Data on Arts and Culture. Personally, I'm transgender, and I've had the honor of curating the world's largest study on transgender Americans (The U.S. Transgender Survey) and the first national probability sample of transgender Americans (TransPop).

I became a data curator because this work combines my expertise in coding with my background in psychology and statistics. Since almost every academic field involves data, the backgrounds of other curators are diverse: from library science and creative writing to social work and the humanities. For many of us, this position is the perfect stepping-stone toward a PhD. For all of us, curating data is a manifestation of our passion for the social sciences.

The internet has revolutionized the ways scientists share information (Berners-Lee et al. 2006). In the same way, data curation has revolutionized how scientists share data. With context, quality, and efficiency, this work optimizes data's potential to change the world.

Acknowledgments

I would like to thank the ICPSR Director of Curation, Rujuta Umarji, for providing vital information about the organization. I would also like to thank A.J. Million and Sara Lafia (ICPSR) for their feedback on every draft.

References

Akers, Katherine G., and Jennifer Doty. 2013. "Disciplinary differences in faculty research data management practices and perspectives." *International Journal of Digital Curation* 8(2): 5–26. https://doi.org/10.2218/ijdc.v8i2.263

Berners-Lee, Tim, Wendy Hall, James Hendler, Nigel Shadbolt, and Daniel J. Weitzner. 2006. "Creating a Science of the Web." *Science* 313(5788): 769–771. https://www.jstor.org/stable/3846906

Donoho, David. 2017. "50 years of data science." *Journal of Computational and Graphical Statistics* 26(4): 745–766. https://doi.org/10.1080/10618600.2017.1384734

Gregory, Kathleen. 2020. "A dataset describing data discovery and reuse practices in research." *Scientific Data* 7(1): 1–11. https://doi.org/10.1038/s41597-020-0569-5

Goldenberg, Tamar, Sari L. Reisner, Gary W. Harper, Kristi E. Gamarel, and Rob Stephenson. 2020. "State Policies and Healthcare Use Among Transgender People in the U.S." *American Journal of Preventive Medicine* 59(2): 247–259. https://doi.org/10.1016/j.amepre.2020.01.030

Henderson, Margaret. 2020. "Why You Need Soft and Non-Technical Skills for Successful Data Librarianship." *Journal of eScience Librarianship* 9(1): e1183. https://doi.org/10.7191/jeslib.2020.1183

Johnston, Lisa, Jake Carlson, Patricia Hswe, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart. 2017. "Data curation network: How do we compare? A snapshot of six academic library institutions' data repository and curation services." *Journal of eScience Librarianship* 6(1): e1102. https://doi.org/10.7191/jeslib.2017.1102

Kcomt, Luisa, Rebecca J. Evans-Polce, Carol J. Boyd, and Sean Esteban Mccabe. 2020. "Association of Transphobic Discrimination and Alcohol Misuse among Transgender Adults: Results from the U.S. Transgender Survey." *Drug and Alcohol Dependence* 215(October): 108223. https://doi.org/10.1016/j.drugalcdep.2020.108223

Lafia, Sara, Andrea Thomer, David Bleckley, Dharma Akmon, and Libby Hemphill. 2021. "Leveraging Machine Learning to Detect Data Curation Activities." *arXiv preprint* arXiv:2105.00030 [cs.CL] https://arxiv.org/abs/2105.00030

Movement Advancement Project. 2019. "Where We Call Home: LGBT People in Rural America." https://www.lgbtmap.org/rural-lgbt

Plantin, Jean-Christophe. 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

Sun, Guangyuan, and Christopher Soo Guan Khoo. 2017. "Social science research data curation: issues of reuse." *Libellarium: Journal for the Research of Writing, Books, and Cultural Heritage Institutions* 9(2). https://doi.org/10.15291/libellarium.v9i2.291

Wickham, Hadley. 2014. "Tidy data." *Journal of Statistical Software* 59(10): 1–23. https://doi.org/10.18637/jss.v059.i10