IFLA WLIC 2017 WROCŁAW

Submitted on: 06.07.2017

What do data curators care about? Data quality, user trust, and the data reuse plan

Sposito, Frank Andreas
University of Denver
Email: frank.sposito@gmail.com

Abstract:

Data curation is often defined as the practice of maintaining, preserving, and enhancing research data for long-term value and reusability. The role of data reuse in the data curation lifecycle is critical: increased reuse is the core justification for the often sizable expenditures necessary to build data management infrastructures and user services. Yet recent studies have shown that data are being shared and reused through open data repositories at much lower levels than expected. These studies underscore a fundamental and often overlooked challenge in research data management that invites deeper examination of the roles and responsibilities of data curators. This presentation will identify key barriers to data reuse, data quality and user trust, and propose a framework for implementing reuser-centric strategies to increase data reuse. Using the concept of a "data reuse plan", it will highlight repository-based approaches to improve data quality and user trust, and address critical areas for innovation for data curators working in the absence of repository support.

Keywords: Data curation, data quality, data reuse, user trust, data reuse plan

Introduction

What do data curators care about? The answer is not as obvious as it may seem. Though there is fairly strong consensus surrounding what constitutes the scope of research data management as a set of sociotechnical practices, the day-to-day activities of data curators, and the formal titles under which they operate, can vary considerably across institutional settings and international borders (Tenopir et al, 2017). The motivations for the development of data curation programs are likewise diverse, often emerging from:

- the policy environment in which research is conducted, such as funder requirements, and broader regulatory constraints, including open data and data security
- local organizational priorities, including scholarly communication initiatives and the creation of multipurpose digital infrastructures

Data curators, depending on their institutional commitments, can care about anything from:

- promoting data literacy among local user communities (data creators)
- supporting grant compliance among research communities
- increasing the efficiency and effectiveness of repository infrastructures
- easing the frustrations of users with repository infrastructures
- providing direct project support for scholars struggling to manage and communicate the often voluminous data output their research generates

This litany is by no means exhaustive. Generally speaking, data curators care about all of these things and more, across the entire data lifecycle, from inception, through preservation, to potential reuse.

The diversity of practices, motivations, and concerns that define the scope of research data management is evident in many influential definitions of data curation. Yet amid the wide range of abstract roles and responsibilities, two overarching purposes—the two things data curators care about most—emerge: preservation and reuse.

- The Digital Curation Centre effectively equates data curation with digital preservation, and refers to reuse only obliquely (DCC, 2017)
- The University of Illinois explicitly mentions reuse as the end-stage of the data lifecycle, but redefines it as reusability, that is, preservation such that reuse becomes possible (Cragin, Heidorn, Palmer & Smith, 2007).
- The European Commission's FAIR guidelines for open research data elaborate on the meaning of reusability by underscoring the importance of metadata quality (among other things) in long-term preservation and access, but provides scant detail about what constitutes reuse and how, specifically, it should be encouraged.

That data curation should be understood within the context of digital preservation is not in itself problematic. Yet the tendency of most understandings of data curation to focus on preservation for reus*ability* rather than reuse per se raises important questions as to whether the digital preservation model for data curation is sustainable as a practice.

Data are unique as preserved digital objects in that their value consist exclusively in their *extrinsic* properties, that is, their *fitness for use* (Altman, 2012). In this sense they differ from digitized cultural heritage objects, which hold intrinsic value and warrant preservation regardless of whether they are ever utilized for research or any other purpose.

- "Scientific data preservation is pointless unless the data are used now and in the future" (Parsons & Duerr, 2005; p. 31).

From this perspective, to care about data, is tantamount to caring about data reuse. Indeed, the entire revenue model for data curation as a practice lies in the assumption that data curation will ultimately increase data reuse (Brown, Wolski & Richardson, 2015; Treadway et al, 2016; Borgman, 2012). Thus the value-proposition for funders in underwriting data curation programs and infrastructures lies precisely in amplifying their return on investment for data-intensive research outputs—that is, leveraging the long-tail of data through actual, not just possible, reuse.

One of the most surprising and disquieting findings in recent empirical studies of open data repositories and their users is that data sharing and actual reuse occur at far lower rates than most scholars and practicing data curators expected (and at least implicitly promised). For example:

- A recent survey of American biomedical researchers found that while 71% of respondents reported sharing data directly with colleagues, only 39% reported ever having used a repository service for either sharing or reuse (Federer et al, 2015).
- A survey of Austrian scholars in 2015 found that only 14% of respondents reported using open data repositories for data sharing, while 54% reported using traditional informal channels of exchanges, such as email or cloud drives (Bauer et al, 2015).

Data curators of all stripes should be deeply alarmed by these anemic patterns of data sharing and reuse. Something seems to be amiss at the heart of data curation. What is it? I will argue that it is a loss of focus on data quality as a critical factor in data reuse.

A brief history of data curation from a data quality and reuse point of view

Ostensibly what all data curators care about are data—that is, collections of digital objects designated for use as evidence in the course of primary research (Borgman, 2015a). Yet in reality it is how they care about data, and why they care, that matters most.

- The concept of care is in fact embedded in the etymology of the word "curation," which comes from the Latin root *cura*, meaning "take care of" or "worry about", but also "treat", "administer" and "supervise".
- Underlying each of these shades of meaning is the idea of "guardianship" or "superintendence," and sometimes "governance."
- The closest semantic equivalents to curators in non-Latinized English are "stewards" and "trustees."
- In many languages, "curation" has taken on the additional meaning of selecting objects for quality and relevance to some antecedent standard of value, normative practice that many data curators, notably those that identify as data librarians and digital preservationists, find beyond the scope of their responsibilities.

From this etymological perspective, data curators care *about* data, but more importantly they care *for* data. And data curators, as guardians and trustees of data, are not only concerned that the data continue to exist (that they are merely preserved, though of course that matters too) but more importantly that they flourish—that is, that they persist in a condition of high functional quality. It is implicit commitment to data quality that has been lost in many recent understandings of the roles and responsibilities of data curators.

Managing research data for quality, in one form or another, has in fact been the core responsibility of data curation since its inception as a distinct sub-discipline within library and information sciences:

- In 1993, a U.S. Department of Energy (DOE) report introduced data curation as a "new professional job category" by explicitly equating it with data quality control: "*Data curation and data quality control.* As [research] databases take on a role similar to [research publications], curation will become increasingly important. Tools are needed to allow and encourage data submitters to take responsibility for the continuing quality of their submissions. Curators must be appointed to oversee long-term quality and consistency of data subsets in community databases. A new professional job category, not unlike museum curators, may develop for these databases. Professional database curators and tools for direct author curation should be supported" (Kingsbury & Snoddy, 1994, p. 16).
- The application of "museum curators" as a metaphor for data curators was meant to underscore the critical importance of ensuring *data quality* in enabling data reuse.
- In 1994, Diane Zorich, citing the DOE report, argued that "'data curators' should emerge as a new professional job category" to oversee the growing body of digital assets generated by museums and archives in the course of their routine preservation and conservation work (Zorich, 1994, p. 431).
- Unlike the DOE report, however, Zorich was not concerned with the curation of primary data per se (which in the museum context is the responsibility of museum curators with disciplinary specializations), but rather secondary data that enable long-term preservation of primary data—what she calls "collections documentation," and what we would now call *metadata*.

- Zorich's focus on secondary metadata rather than primary data allowed her to appropriate a key concept from record-keeping literature into her understanding of data curation: the notion that metadata exist in something akin to a "lifecycle" that parallels the administrative lifecycle of the primary objects those metadata represent.

From this moment forward, data curation become a content-agnostic approach to lifecycle data management geared towards long-term preservation. The development of high-quality (or "rich") metadata became the hallmark responsibility of data curators, while the obligation to ensure the quality of primary data was subsequently relegated to subject matter experts, namely, the original research data creators themselves.

The data reuse plan: tactics for augmenting data quality

While the reasons for low levels of data reuse are not perfectly understood, one common factor is correspondingly low levels of *user trust* in information resources of every kind (Yoon, 2016; Yoon, 2017; Carlson, 2007). User trust and data reuse are connected through the perceived quality of data resources: data quality is, in effect, a bridge of trust between two distinct (but sometimes overlapping) user roles involved in research data exchange, data creators and data reusers (Peer, Green, & Stephenson, 2014).

- Under conventional data sharing practices, where data is exchanged informally among communities of known creators and reusers, this bridge of trust is strong and well supported by existing and often highly localized professional networks (Carlson & Anderson, 2007; Yoon, 2017).
- With the advent of open data and repository-based sharing, however, the "distance" between data creators and data grows, creating new trust barriers that challenge the entire premise of open data (Borgman, 2015a).
- The idea that the core responsibility of the data curator is to ensure data quality is thus effectively an assertion that data curators are the principal agents for communicating data trustworthiness across the distances that separate relatively anonymous user groups.

Another way of putting this is to say that the channel through which data curators transit trustworthiness to data reusers is data quality.

Yet the fact that data reuse levels are so low is evidence that current models of data curation, the roles and responsibilities of data curators, especially those that emphasize metadata quality over primary data quality, may not be as effective as initially expected. In what follows I will outline recent efforts by some data curation programs to improve primary data quality and increase user trust—all with the aim of increasing downstream data reuse. I will introduce a new conceptual framework , the "data reuse plan", to organize these activities, as well as indicate areas where additional innovation in the roles data curators take and the responsibilities they assume will be required.

The core idea behind a data reuse plan is to elevate data reuse (and subsequently user trust and data quality) as the central rationale for all data curatorial programs and activities:

- The data curator is defined as the agent responsible for ensuring data quality across the data lifecycle with the specific goals of strengthening user trust and increasing the circulation of data resources throughout society.
- The primary user groups for data curators serving in academic libraries, then, are not known data creators, often local faculty or students who produce data as part their

routine research output, but rather unknown data reusers who hope to leverage these outputs for new and often extended insights.

- This approach effectively returns data librarianship to the Ranganathanian view that information resources of all kinds are not for preservation but rather use.
  - ✓ Research data are for reuse
  - ✓ Every reuser their data
  - ✓ Every dataset its reuser

Repositories, in other words, are not archives but rather clearinghouses. Preservation for reusability is not enough. If data are not in fact reused, then data curators must redouble their efforts to connect data with actual reusers, specifically focusing on Ranganathan's third law: find reusers for your data. This is the principal mandate behind the data reuse plan, and it requires deliberate prioritization of data quality and user trust when designing and implementing data curation programs and services.

a. The data reuse plan in repository settings

Repository environments are ideal settings for the development of data reuse plans, since question of reuse potential can be addressed directly at the moment of ingest.

- Identify and designate specific forms of reuse enabled by repository
  - *Reproduction*: affirm veracity of research findings by reanalyzing original data with original methods tools (statistical scripts, code, etc)
  - *Replication*: affirm veracity by reanalyzing new data with original tools and methods
  - *Reanalysis*: explore data for new statistical patterns not identified in original research
  - *Extended integration*: connect data to similarly scoped data sets with common units of analysis but different variables
  - *Open integration*: enable consolidation of separately collected datasets with overlapping variables and identical units of analysis into big data resources
  - *Extended application*: apply data to problems and questions outside its original scope of use with higher order units of analysis
  - *Data journalism*: broadening accessibility of existing data to increase reuse among non-specialist users
- Run automated reproduction analyses on all data submissions prior to ingest
- Create user services/interfaces to facilitate designated reusage
  - Reproduction and replication support tools should be enabled by default for every preserved data set
  - Include data limitations (reliability index) in metadata descriptions
  - Link data sets to published articles
  - Automatically summarize granular data to aggregate units of analysis
  - Track future provenance: how and by whom data is adapted and applied in further reuse
- Create integrated peer review system for ingested datasets
- Leverage repository brand as data publisher

b. The data reuse plan without repository infrastructure

Abstract data librarianship, where specific data usages do not figure into training, presents a far greater challenge for increasing data reuse:

- Above all else: build data literacy programs around a institutional repository service (at least) for educational purposes
  - Open data and data curation are fundamentally repository-based services
  - Commit fully to open data training, including sharing and reuse
- Integrate data curation training into communities of practice
  - Subject-specialist librarians assume greater RDM training and development responsibilities
- Promote reuse (rather than sharing) from the beginning of the research cycle (that is, with researchers)
- Research data models
  - Build research data models with designated reuses (e.g., how can this research be reproduced?)
  - Build research data models with specific reuse communities in mind

References

Altman, M. (2012). Mitigating threats to data quality throughout the curation lifecycle. *Curating For Quality: Ensuring Data Quality to Enable New Science. National Science Foundation, Arlington County, VA*, 1-119.

Bauer, B., Ferus, A., Gorraiz, J., Gründhammer, V., Gumpenberger, C., Maly, N & Steineder, C. (2015). Researchers and Their Data. Results of an Austrian Survey - Report 2015. Zenodo. http://doi.org/10.5281/zenodo.34005

Borgman, C. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059-1078.

Borgman, C. (2015a). *Big data, little data, no data: scholarship in the networked world*. MIT press.

Borgman, C. (2015b). If data sharing is the answer, what is the question? *ERCIM NEWS*, 15.

Brown, R., Wolski, M. & Richardson, J. (2015). Developing new skills for research support librarians. *The Australian library journal*, *64*(3), 224-234.

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*, 12(2), 635-651.

Cragin, M. H., Heidorn, P. B., Palmer, C. L., & Smith, L. C. (2007). An educational program on data curation. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/3493/ALA_STS_poster_2007.pdf

Chignard, S. (2013). A brief history of open data. *Paris Tech Review*, *29*.

Data Seal of Approval. (2017). The core trustworthy data repository requirements. Retrieved from https://www.datasealofapproval.org/en/information/requirements/

Digital Curation Center. (2017). What is digital curation? Retrieved May 30, 2017, from http://www.dcc.ac.uk/digital-curation/what-digital-curation

European Commission. (2016). *Guidelines on FAIR data management in Horizon 2020*. Version 3.0, 26 July 2016. Retrieved December 15, 2016, from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, https://doi.org/10.1016/j.ecoinf.2012.03.004

Federer, L., Lu, Y., Joubert, D., Welsh, J. & Brandys, B. (2015). Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PloS one*, *10*(6), e0129506.

Kingsbury, D., Snoddy, J., & Robbins, R. (1994). Report of the Invitational DOE Workshop on genome informatics, 26-27 April 1993, Baltimore, Maryland. Genome Informatics I: Community Databases. *Journal of Comparative Biology*, *1*, 173-90.

Palmer, C., Weber, N., Renear, A., & Muñoz, T. (2013). Foundations of data curation: The pedagogy and practice of "purposeful work" with research data. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/78099/ArchivesJournal-Weber062015.pdf

Parsons, M. A., & Duerr, R. (2005). Designating user communities for scientific data: challenges and solutions. *Data Science Journal*, *4*, 31-38.

Pasquetto, I., Randles, B., & Borgman, C. (2017). On the Reuse of Scientific Data. *Data Science Journal*, *16*.

Peer, L., Green, A., & Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation*, *9*(1), 263-291.

Tenopir, C., Talja, S., Horstmann, W., Late, E., Hughes, D., Pollock, D. & Allard, S. (2017). Research data services in European academic research libraries. *Liber Quarterly*, *27*(1).

Treadway, J, Hahnel, M, Leonelli, S, Penny, D, Groenewegen, D, Miyairi, N, et al. (2016). *The state of open data report*. Figshare. Retrieved from: https://figshare.com/articles/The_State_of_Open_Data_ Report/4036398.

Wouters, P. & Hack, W. (2017) *Open data: The researcher perspective*, CWTS, Universiteit Leiden, Leiden. Retrieved from https://www.elsevier.com/__data/assets/pdf_file/0004/281920/Open-data-report.pdf

Yoon, A. (2016). Red flags in data: Learning from failed data reuse experiences. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1-6.

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

Zorich, D. M. (1995). Data management: managing electronic information: data curation in museums. *Museum Management and Curatorship*, *14*(4), 430-432.