Heather Coates

Ensuring research integrity

The role of data management in current crises

Modern scientific and academic research is facing several crises, including a lack of credibility and risk of data loss. Credibility pertains to whether people trust that the scholarly record represents the world in an objective and accurate way. Credibility is damaged when the public learns of retractions, misconduct investigations, and controversial findings or methods. Researchers risk losing credibility when they violate the expectations of colleagues, institutions, funding agencies, and the public. These expectations include competence, honesty, integrity, and benefit.[1]

The issue of credibility is compounded by a lack of shared understanding about what is typical or acceptable. Public expectations differ from those of researchers. The public is generally not aware that acceptable practices are highly context-specific and differ across research communities. Gaps also exist between the expectations expressed to trainees and the daily practices of more experienced researchers trying to maintain research programs in the face of shrinking funding resources. Public cases of misconduct illuminate gaps between the expressed values and the reality of current practices. Thus, discussions about ethical behavior in research are informed by shared values, expectations, and acceptable practices.

The scientific method assumes that the scholarly record will self-correct as the cumulative weight of evidence favors one explanation of events over others. Patterns revealed by high-profile cases of misconduct, such as that of Diedrick Stapel,[2] highlight many weaknesses in our current system for ensuring research integrity. Investigations into similar cases further highlight the inadequacy of existing processes for self-regulation and timely self-correction.[3]

Though many research processes have transitioned from print to digital, the standards and training used to ensure integrity have not. Continued progress depends on robust peer review processes and available resources for reproducing and validating prior findings. Many scientific findings will eventually be proven inaccurate or incorrect by new technologies, methodologies, or interpretations. Therefore, the integrity of the scholarly record depends on the reliability of verification and self-correction mechanisms rather than the accuracy of a single dataset.

We need more critical appraisal of the processes used to generate, process, and analyze data, which requires greater transparency through easy discoverability and access to the data behind published findings. However, the data alone are not sufficient. Comprehensive documentation describing the context in which the data were generated is essential. While the importance of transparency in resolving credibility and

Heather Coates is digital scholarship and data management librarian at IUPUI University Library Center for Digital Scholarship, e-mail: hcoates@iupui.edu

Contact series editors Zach Coble, digital scholarship specialist at New York University, and Adrian Ho, director of digital scholarship at the University of Kentucky Libraries, at crlnscholcomm@gmail.com with article ideas

integrity issues is recognized,[4, 5] it is not yet a standard practice.

Academic research practices are influenced by other factors, notably criteria for achieving promotion and tenure. These criteria can overemphasize novel and sensationalized conclusions published in a few high-impact journals. Such rewards directly conflict with the tenet that scientific knowledge is advanced through the accumulation of evidence. Kevin Smith, director of copyright and scholarly communication at Duke University Libraries, has remarked that review practices of big scientific journals cut against scientific progress.[6] In practice, many disciplines have outsourced evaluation of research quality to journal editors and publishers by inappropriate use of citation metrics. This happens despite recognition that journal-level impact factors are not effective indicators of the quality of individual articles,[7, 8, 9] much less a research project. Moreover, stakeholders in the current system have competing interests, each with limited individual power to investigate and hold researchers accountable for misconduct.[10] This is not an environment conducive for conducting high-quality ethical research.

A change in culture is long overdue. The process for shifting toward a culture that prioritizes integrity over visibility must be informed by a better understanding of how the current system operates. We understand very little about how expectations, incentives, and conflicting interests influence the way research is conducted. In particular, it may be that differences between norms of practice at the professional, institutional, and departmental levels are significant contributors to research misconduct. Culture change is complex and slow, so we first need to understand which research practices are effective in promoting integrity and then determine how to encourage and reward those practices.

Data loss

Our ability to generate complex, massive data has outstripped our ability to store, manage, and use it. The digital environment is straining practices refined on print materials. While the computing revolution has given us enormous capacity to generate, store, analyze, and visualize data, these systems and media are fragile. They do not have the stability or durability of paper. Communities of researchers, institutions, and funding agencies need to begin having conversations about prioritizing what data to preserve, curate, and share. The ultimate question here is: Will the data be accessible and usable when it is needed?

Responsible data management to improve research integrity

Data is a key piece of the scholarly record. As such, the way in which data is managed has an impact on the integrity of the scholarly record as well as the potential for data curation, sharing, and reuse or secondary analysis. This is recognized by the Office of Research Integrity,[11] the National Academies of Science,[12] federal funding agencies requiring data management plans,[13] and initiatives like FORCE11.[14] Kenneth Pimple describes data management as "the neglected, but essential, twin to the 'scientific method.'"[15] While careless data management can lead to gaps and errors in the scholarly record, effective data management strategies do exist for preventing inaccuracies. Documentation is one such strategy. Despite its importance in validating published results and preventing the types of errors that lead to retractions, researchers commonly acknowledge that research data are typically poorly documented. One cause may be the heavy administrative burden of conducting academic research.[16]

Variability in the research methods and processes used in different disciplines limits adoption of universal recommendations for managing data. However, some common focus areas are emerging from data management training programs:[17, 18, 19] identifying and addressing ethical and legal obligations; providing detailed documentation that follows standard practices for the field; planning and execution of a data storage and backup plan; and archiving data for reuse and secondary analysis. Sharing data requires careful consideration and planning to protect sensitive

information, while maximizing the availability of the data for verification, secondary analysis, and reuse. Responsible management of the data includes balancing ethical and legal obligations to the funding agency, institutions, researchers, and participants (if applicable). At minimum, the data underlying published findings should be available for verification purposes. There are many options beyond dark and open data, but many researchers are not aware of them.

Acknowledging responsible data management as foundational for research integrity is not sufficient. We need to value the processes and products of research equally by: 1) creating incentives for responsible management of data, 2) developing standards and practices for peer review that balance evaluation of methodological quality and research integrity with potential impact, and 3) carefully considering the resources necessary to responsibly manage and preserve newly created data for five-to-ten years after publication. The last is vital, given our current lack of capacity to store, manage, and use existing data. Efforts to develop better data management technologies and infrastructure have begun, but are not keeping pace with data creation.

Our role in supporting responsible data management

The academic community must improve the integrity of the scholarly record to regain credibility. The library has an important role to play in this by providing ongoing access to the scholarly record and teaching effective practices for ensuring its integrity. Despite our long-time role as stewards of the scholarly record, our involvement with research ethics education is often limited to teaching about plagiarism. Yet, our knowledge extends beyond this to include many of the principles for the responsible conduct of research,[20] excepting mentor and trainee responsibilities. As a profession, our deep engagement with scholarly communication and information management brings practical knowledge to ethics training that is often theoretical in nature. This is an opportunity to leverage our instructional skills to improve the data management practices of researchers and take a more active role in preventing misconduct throughout the research process. Providing education and support further upstream should result in more informed and carefully documented projects, leading to greater integrity and reduced data loss. Such instruction should focus on strategies such as data management planning, file organization and naming, storage and backup, metadata and documentation, and data registration and citation.

As practitioners, we can lead by example. Encouraging the editors of library and information science journals to implement review and data availability policies similar to those proposed by Frances S. Collins and Lawrence A. Tabak[21] would increase the transparency and integrity of our own scholarly records. We can provide greater support for faculty in implementing tools for data deposit, registration, and citation into their process. We can support open science initiatives like study registration and replication studies by introducing these approaches into our outreach and instruction and by collaborating with researchers to implement strategies for transparency. Maintaining long-term access to data and contextual information can be aided by operating an institutional repository or participating in consortia like the Digital Preservation Network. As scholars, we can examine and measure the impact of these changes in practices and incentives on various research communities.

Although none of us will be involved with all of these activities, we should be cognizant of the importance of data management for research integrity and prepared to initiate these discussions with our faculty, students, and administration.

Notes

1. Stephanie. J. Bird, "Responsible research: what is expected? Commentary on: 'Statistical power, the Belmont Report, and the ethics of clinical trials,'" *Science and Engineering Ethics* 16 (2010): 693-6, doi: 10.1007/s11948-010-9248-9.

2. John M. Budd, "The Stapel case: An object lesson in research integrity and its lapses,"

Synesis 4 (2013): G47-53, www.synesisjournal.com/vol4_g/Budd_2013_G47-53.pdf.

3. Adam Marcus and Ivan Oransky, *Retraction Watch* (2014), http://retractionwatch.com/ (accessed October 22, 2014).

4. Committee on Science, Engineering, and Public Policy (U.S.). Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, E*nsuring the integrity, accessibility, and stewardship of research data in the digital age* (Washington, D.C.: National Academies Press, 2009), www.nap.edu/catalog.php?record_id=12615.

5. Vedran Katavic, "Retractions of scientific publications: Responsibility and accountability," *Biochemica Medica* 24 (2014): 217-22, doi: 10.11613/BM.2014.024.

6. Kevin Smith, "Are fair use and open access incompatible?" *Scholarly Communications @ Duke,* September 25, 2014, http://blogs.library.duke.edu/scholcomm/2014/09/25/fair-use-open-access-incompatible/.

7. Bjorn Brembs, Katherine Button, and Marcus Munafo, "Deep impact: Unintended consequences of journal rank," *Frontiers in Human Neuroscience* 7 (2013): 291, doi: 10.3389/fnhum.2013.00291.

8. Neal S. Young, John P. Ioannidis, and Omar Al-Ubaydli, "Why current publication practices may distort science," *PLOS Medicine* 5 (2008): e201, doi: 10.1371/journal.pmed.0050201.

9. Jennifer Couzin-Frankel, "Shaking up science," *Science* 339 (2013): 386-9, doi: 10.1126/science.339.6118.386.

10. Anthony L. Zietman, "Falsification, fabrication, and plagiarism: The unholy trinity of scientific writing," *International Journal of Radiation Oncology, Biology, Physics* 87 (2013): 225-7, doi: 10.1016/j.ijrobp.2013.07.004.

11. Nicholas H. Steneck, *Introduction to the responsible conduct of research* (Washington, D. C.: Government Printing Office, 2004), http://ori.hhs.gov/sites/default/files/rcrintro.pdf (accessed October 22, 2014).

12. Committee on Science, Engineering, and Public Policy (U.S.). Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, E*nsuring the integrity, accessibility, and stewardship of research data in the digital age* (Washington, D.C.: National Academies Press, 2009), www.nap.edu/catalog.php?record_id=12615.

13. National Science Foundation, Grant Proposal Guide (Washington, D.C.: National Science Foundation, 2013), www.nsf.gov/pubs/policydocs/pappguide/nsf14001/gpg_index.jsp (accessed October 22, 2014).

14. FORCE11, Joint Declaration of Data Citation Principles (2014), https://www.force11.org/datacitation (accessed October 22, 2014).

15. Kenneth D. Pimple, "Six domains of research ethics," *Science and Engineering Ethics* 8 (2002): 191-205, doi: 10.1007/s11948-002-0018-1.

16. Sandra L. Schneider, Kirsten K. Ness, Sara Rockwell, Kelly Shaver, and Randy Brutkiewicz, 2012 Faculty Workload Survey: Research Report (Federal Demonstration Partnership, 2012), http://smrb.od.nih.gov/documents/reports/8a-FDP-2012-FWS-Research-Report.pdf (accessed October 22, 2014).

17. New England Collaborative Data Management Curriculum (Worcester, MA: University of Massachusetts Medical School, 2013), http://library.umassmed.edu/necdmc/index (accessed October 22, 2014).

18. DataONE, Data Management Education Modules (2012), https://www.dataone.org/education-modules (accessed October 22, 2014).

19. Lisa R. Johnston and Jon Jeffryes, Data Management Workshop Series, Winter 2014, University of Minnesota Libraries (Minneapolis, MN: University of Minnesota Libraries, 2014), http://z.umn.edu/datamgmt14 (accessed October 22, 2014).

20. Nicholas H. Steneck, *Introduction to the responsible conduct of research* (Washington, D. C.: Government Printing Office, 2004), http://ori.hhs.gov/sites/default/files/rcrintro.pdf (accessed October 22, 2014).

21. Frances S. Collins and Lawrence A. Tabak, "Policy: NIH plans to enhance reproducibility," *Nature* 505 (2014): 612-13, doi: 10.1038/505612a.