Research Data, Reproducibility, and Curation

Digital Social Research:

A Forum for Policy and Practice

Oxford Internet Institute Invitational Symposium

http://www.oii.ox.ac.uk/events/?id=487

13 March 2012

Christine L. Borgman

Professor & Presidential Chair in Information Studies

University of California, Los Angeles

borgman@gseis.ucla.edu

The "data deluge" that was the eScience call to arms (Hey & Trefethen, 2003), quickly passed the "tsunami" stage (Southan & Cameron, 2009), and shows few signs of slowing down. Much of that data is runoff, leaving science policy makers concerned with how to capture, manage, and curate data in ways that might make them useful to others.

While a laudable goal, enforced by requirements for data management plans and data sharing, the curation infrastructure to support those requirements is almost nonexistent. A few fields, such as genomics, seismology, and astronomy, have established standards, practices, and repositories to curate data. Even in these fields, curation is uneven. Genomics data are deposited at varying rates, with varying degrees of quality. Seismology data are more likely to be deposited if the research was conducted with funds from public agencies than if done with private foundation funds. Similarly, space-based astronomy missions tend to have more comprehensive data management than ground-based missions. These fields are exceptions: in the "long tail" of science, repositories are rare, standards for data structures and metadata may not exist, and data handling practices are based on local craft knowledge.

Rationales for sharing data are many, and rarely are made explicit. They tend to fall in four categories: (1) to reproduce or to verify research, (2) to make results of publicly funded research available to the public, (3) to enable others to ask new questions of extant data, and (4) to advance the state of research and innovation. These rationales differ by the arguments for sharing, by beneficiaries, and by the motivations and incentives of the many stakeholders involved (Borgman, 2012, forthcoming).

Of particular interest for this forum is the first rationale, that of reproducibility, and how it may apply to data curation. Pressure is mounting to share data for the purposes

of reproducing research findings. A recent special issue of *Science* on replication and reproducibility examines the approaches, benefits, and challenges across multiple fields (Ioannidis & Khoury, 2011; Jasny, Chin, Chong & Vignieri, 2011; Peng, 2011; Ryan, 2011; Santer, Wigley & Taylor, 2011; Tomasello & Call, 2011). The authors encourage data sharing to increase the likelihood of replication, while acknowledging the very different methods and standards for reproducibility in each field discussed.

Reproducibility or replication of research is viewed as "the gold standard" for science (Jasny et al., 2011), yet it is the most problematic rationale for sharing research data. Reproducing a study confirms the science, and in doing so confirms that public monies were well spent. However, the argument can be applied only to certain kinds of data and types of research, and rests upon several questionable assumptions about relationships among data, epistemology, research methods, sources of evidence, and scientific practice.

The relationship between rationales for data sharing and data curation appears to be an unexplored topic, and one appropriate for discussion by a forum of scholars and policy makers. Among the questions are these: Should data policy be driven by requirements for reproducibility or replication? How do reproducibility and replication vary between fields? Should data curation practices strive to maintain the products associated with each research paper, to promote reproducibility? Should curation practices strive to standardize data structures and practices within a field, to promote data mining and the ability to ask new questions? To what extent are these goals synonymous and to what extent are they in conflict? What are the costs and benefits associated with data curation, and to whom? What data are most worthy of curation, and who decides? What workforce skills are required for data curation within each scientific community? What workforce skills for data curation should be embodied in libraries, archives, and university infrastructure? Who should fund data curation, by what means, and where should research data be maintained, by whom, and for how long?

REFERENCES

Borgman, C. L. (2012, forthcoming). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology.

Hey, A. J. G. & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective. In Berman, F., Fox, G. & Hey, A. J. G. (Eds.). Grid Computing: Making the Global Infrastructure a Reality. Chichester, Wiley. Retrieved from http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf on 20 January 2005.

Ioannidis, J. P. A. & Khoury, M. J. (2011). Improving validation practices in "omics" research. Science, 334(6060): 1230-1232.

Jasny, B. R., Chin, G., Chong, L. & Vignieri, S. (2011). Again, and again, and again. Science, 334(6060): 1225.

Peng, R. D. (2011). Reproducible Research in Computational Science. Science, 334(6060): 1226-1227.

Ryan, M. J. (2011). Replication in Field Biology: The Case of the Frog-Eating Bat. Science, 334(6060): 1229-1230.

Santer, B. D., Wigley, T. M. L. & Taylor, K. E. (2011). The Reproducibility of Observational Estimates of Surface and Atmospheric Temperature Change. Science, 334(6060): 1232-1233.

Southan, C. & Cameron, G. (2009). Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences Data. In Hey, T., Tansley, S. & Tolle, K. (Eds.). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA, Microsoft: 117-123. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/ on 5 March 2012.

Tomasello, M. & Call, J. (2011). Methodological challenges in the study of primate cognition. Science, 334(6060): 1227-1228.