Selected Papers of Internet Research 16:
The 16th Annual Meeting of the
Association of Internet Researchers
Phoenix, AZ, USA / 21-24 October 2015

DISCOURSES, A COMMUNITY OF SCIENTISTS, AND LONG-TAIL DATA IN THE CLOUD

Catherine F. Brooks
University of Arizona

P. Bryan Heidorn
University of Arizona

Gretchen Stahlman
University of Arizona

Steven Chong
University of Arizona

Data in the Sciences

Data collection, use, and management are significant activities in the scientific enterprise, and are evolving given the onset of a big data world. Many scientists (e.g., scholars or instructors) do not have access to the computing infrastructure needed to work with large data sets. This is particularly true at many biological research stations – while some stations are part of or loosely affiliated with major universities, others are stand-alone entities. The station researchers are relatively transient, spending seasons at the stations then returning to their home institutions – their data scatter in their differing formats (Estrin et al., 2003). This is particularly true for little data that often do not conform to established standards (Borgman et al., 2007).

Previous research on the long tail of data (Heidorn, 2008; Palmer, Cragin, Heidorn, & Smith, 2007) suggests that a significant portion of data collected in previous eras was actually lost or went unused, data collected may have been used for specific projects then left to die on floppy disks or personal computers. Data sharing on cloud-based platforms provides a way to share costs for needed infrastructure and offers hope that previously-lost 'dark' data will be brought to light. Discussions of day-to-day computing needs and data-sharing possibilities are an important initial step in enabling problem-solving for contemporary scientists – facilitating this type of discussion was the focus of the workshop interrogated in this study.

Suggested Citation (APA): Brooks, C. F., Heidorn, P. B., Stahlman, G., and Chong, S. (2015, October 21-24). *Discourses, a community of scientists, and long-tail data in the cloud*. Paper presented at Internet Research 16: The 16th Annual Meeting of the Association of Internet Researchers. Phoenix, AZ, USA: AoIR. Retrieved from http://spir.aoir.org.

Data sharing and related infrastructure dilemmas are of interest across a wide variety of scientists, salient for those engaging with ‘little’ or big data. Scientists tend to work with technologist and engineers as part of their working teams (smaller projects often rely on graduate students working in laboratories). So, it is easy to imagine the ways in which shifts in scholarly practice, data collection, and the management of information – especially in an age of big data – are of paramount importance across sectors and to those coming from the entire research spectrum. To interrogate contemporary concerns in science about data management, particularly that in the cloud, the following research question was posed:

RQ: How do issues with sharing data in the cloud get discursively situated by an interdisciplinary group of scientists?

Method

This qualitative work (Miles & Huberman, 1994) drew themes from within the data themselves, and followed protocols that are well established in qualitative research traditions (Lindlof, 1995). Data collection and organization was IRB approved and involved the audio recording of whole-workshop talk between a leader and workshop attendees. These recordings were then transcribed and analyzed.

This qualitative work (Miles & Huberman, 1994) drew themes from within the data themselves, and followed protocols that are well established in qualitative research traditions (Lindlof, 1995). Data collection and organization was IRB approved and involved the audio recording of whole-workshop talk between a leader and workshop attendees. These recordings were then transcribed and analyzed.

The central premise of the workshop was that new science could be enabled through the development of shared cloud-based cyberinfrastructure. The work was funded under the National Science Foundation’s Software Infrastructure for Sustained Innovation program[1]. Alongside one primary workshop leader, there were three other workshop facilitators, 29 participants, and five assisting doctoral students.

Findings: Framing Shared Problems

As described previously, these scientists came together to resolve a shared problem, considering tools that can support sharing data in the cloud. If dark data from smaller projects were drawn together, scientists would have a broad set of data available to them. Shawn, the workshop planner began the first meeting by explaining dark data:

…you could pull it together and get to a critical mass to make it easier to share and useful for science.

He continued to explain that if scholars share and work together though the Internet or other web-based platforms, more data will be made available, viewable, and usable –

[1] This workshop was sponsored in part by the National Science Foundation through the following collaborative SI^{2}-S2I2 grants: 1216726, 1216754, 1216872, 1216879, 1216884. http://www.nsf.gov/si2/

data will become increasingly ‘democratized’ and access across communities of scholars and practitioners will be enhanced. As Shawn continued,

…there is a huge growth in the amount of data that we can acquire. Sequenced data, in particular, is outstripping the Moore’s Law, so, in fact, the amount of data we have is growing faster than our computing capacity to process that data.

To give context for the current interrogation, then, Shawn’s talk shows that shifts in computing infrastructures and cloud-based solutions need consideration.

Findings: Capital, Commercialized Threats, and the Economy of Innovative Science

In some ways, sharing data on the cloud implies trust of commercial interests. In such an environment, data become the commodity providing a rich place for commercialized competition.

Moderator: … programs, like [those with] Amazon …that are put in place for science data… they’ll take data, literally for free, and host it for free within certain constraints that they’re still often a little fuzzy about…

Participant: Right.

Moderator: I think this is actually a wingding for everybody…

Indeed, commercial concerns were continually voiced relative to companies like Amazon and Google, as Shawn discussed:

Some of you may remember that Google would send you a suitcase with a hard drive that said, “Just send your data. We’ll take care of the rest.” It took them less than six months to close that program down… [Laughter]

For these scientists, individualized academic culture, institutionalized legal concerns, and a broad capitalist culture were all considered disruptions or barriers to the common scientific enterprise and to ‘figuring out’ how to share and manage data with and through contemporary Internet, web, or cloud-based tools.

Conclusion

Contemporary science has witnessed recent shifts that are powerful, cultural, and sit within and well beyond the confines of academia. The very nature of our knowledge-related capabilities has changed. Computing infrastructure is needed, most practicing scientists cannot manage those needs alone. This study provides an early glimpse of how these issues are situated by scientists and in relation to the broader milieu in contemporary science.

References

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries, 7(1-2), 17-30.

Estrin D, Michener W, Bonito G, and the workshop participants (2003) Environmental Cyberinfrastructure Needs for Distributed Sensor Networks: A Report from a National Science Foundation Sponsored Workshop. Scripps Institution of Oceanography, La Jolla, CA. August 12-14, 2003.

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. Library Trends, 57(2), 280-299.

Lindlof, T. R. (1995). Qualitative communication research methods. Thousand Oaks, CA: Sage.

Mayernik, M. S., Batcheller, A. L., & Borgman, C. L. (2011). How institutional factors influence the creation of scientific metadata. ACM International Conference Proceeding Series, 417–425. doi:10.1145/1940761.1940818.

Miles, M. B., & Huberman, A. M. (1994). An expanded resource: Qualitative data analysis (2nd ed.). Thousand Oaks, CA: Sage.

Palmer, C. L., Cragin, M. H., Heidorn, P. B., & Smith, L. C. (2007). Data curation for the long tail of science: The case of environmental sciences. Paper presented at the Third International Digital Curation Conference, Washington, DC.