Single-Investigator Culture, Accidental Data Contamination, and Geopolitics of Data Ownership: Three Challenges of Data Sharing in Big Data Science, Cyberinfrastructure, and Cooperative Scientific Work

Kerk F. Kee, Ph.D.
Assistant Professor
Department of Communication Studies
Chapman University
One University Drive, Orange, CA 92866
kerk.kee@gmail.com
www.ekerk.com

ABSTRACT

While the vision of data sharing for big data science, cyberinfrastructure, and cooperative scientific work is exciting, the practice is full of challenges. Based on qualitative methods guided by grounded theory and 70 interviews, three such challenges are presented in this paper. They are single-investigator culture in traditional academic, accidental data contamination, and geopolitics of data ownership. The analysis suggests that in order for data sharing to become an enduring practice, the urgency of the scientific problems needs to become more pressing, in order to put pressure on academic systems, researchers, and governments to shift their current practices towards the vision of community data sharing.

Author Keywords

Data sharing, cyberinfrastructure; big data science.

ACM Classification Keywords

K.4.3 [Organizational Impacts]—Computer-supported collaborative work, H.5.3 [Group and Organization Interfaces]—Computer-supported cooperative work.

INTRODUCTION

Data sharing, reuse, and circulation of resources is one of the most exciting promises of cyberinfrastructure (CI), as it enables large-scale research and data-intensive collaboration using aggregated computational resources and combined datasets through the Internet, high-performance networks, and local machines and to be able to mine publicly-funded datasets accumulated over time [1]. Data intensiveness is a key defining characteristic of CI and e-science projects [2], and the sharing and reuse of big datasets can revolutionize how science can be performed [3]. However, the vision of data sharing, re-use, and circulation faces several practical challenges. This position paper illuminates three such challenges based on thematic analysis of qualitative interviews. As a preview, the challenges are the single-investigator culture of traditional academia, reluctance to show accidental data contamination, and geopolitics of shared ownership of data.

Citation: Kee, K. F. (2014, February). Single-investigator culture, accidental data contamination, and geopolitics of data ownership: Three challenges of data sharing in big data science, cyberinfrastructure, and cooperative scientific work.Paper presented at "Sharing, Re-use and Circulation of Resources in Scientific Cooperative Work" workshop, Computer Supported Cooperative Work (CSCW) Conference, Baltimore, MD.

DATA COLLECTION AND METHOD

The interview data analyzed and presented in this paper was conducted over a period of 32 months, from November 2007 to June 2010. The data set includes 70 interviews with 66 participants from across 17 U.S. states and three other countries [The UK, Germany, and Australia]. Because most of the interviews were conducted in 2008, the analysis primarily reflects the practice of data sharing during this period, although much can be argued to be the same today.

This paper employs grounded theory in qualitative data collection and analysis. According to Corbin and Strauss [4], one of the main goals of grounded theory is to seek to uncover relevant conditions under which social phenomena manifest. I argue in this paper that the relevant conditions for data sharing are the academic culture, inevitable human errors, and geopolitics. The next section unpacks these challenges in details.

SINGLE-INVESTIGATOR CULTURE OF TRADITIONAL ACADEMIA

The first challenge is the tension between the single-investigator culture in many scientific fields and the vision of advancing a scientific community by data sharing. In traditional academic, scientists advance their careers, get tenured and promoted in ranks, by illustrating that they can do good science on their own, including collecting their own data. In the early 21st century, many fields in traditional academic still have the culture of rewarding single investigators. The excerpt below illustrates this tension:

> The data piece… In our communities, we have a culture that's been very much single investigator… [Y]ou talk to them about - we could create a tool that would make it easier for you to work with your data and to get data from all different locations and integrate it and plug it into models and use those models and publish them and share them with other people and be able to plug together different models and visualize the results… The benefits we would get from being more of a community, it's not immediately obvious to people why they should spend a lot of time on that. (Faculty PI, Water Resources Engineering Researcher, Illinois)

Academic cultures are often enduring, although they do evolve over time. Until a field has substantially transitioned into a community-oriented approach to doing science, data sharing will likely persist in the time being. For this to happen, time is a critical factor.

ACCIDENTENTAL DATA CONTAMINATION

Science, like most other human activities, inevitably contains human errors. However, human errors during data collection make researchers reluctant to share their data. Otherwise, the results can be called into questions. Although scientists are humans, and humans understand that errors happen, it is still embarrassing to admit to them. The excerpt below explains this challenge:

> People are sometimes afraid to let people see the warts in their data. So when the guy did the measurements on the third run, he smoked a cigarette and the ash fell in front of the laser and it blanked out for a second and there's a blotch in the middle of the data. He doesn't really want to show that trace, even though he knows it's fine. He'd be embarrassed to explain that he was smoking [in the lab] – that's against the law. (Faculty PI, Combustion Chemistry, Massachusetts)

As suggested in the excerpt above, sometimes the human errors committed are also against established regulation, not simply embarrassing, thus further compounding the challenge to data sharing. However, NSF's recent move toward requiring every proposal to have a data management plan, as well as certain progressive journals making data public along with publications may be changing this culture.

GEOPOLITICS OF SHARED OWNERSHIP OF DATA

While human history and politics draw lines between countries, scientific problems do not stop at national boundaries. In fact, many grand scientific problems exist across borders, making the issue of data-sharing even more pressing. However, national ownership of data presents the third challenge described in this paper. An excerpt is selected to show this argument. As the informant working on this a bi-national project speaks English as a second language, certain linguistic characteristics are maintained in the excerpt below to heighten the readers' awareness of the global dimension of data sharing:

> The main idea in the beginning [was] to create [a] bi-national data set in this case, even though several projects have been achieved in Mexico and the USA in this area… [But] all of the projects usually stop in the river and it's the same case on the Mexican side… [A]ll of the results stop in the border. So the idea at the beginning was to generate a useful database or useful tool to have two agencies on both sides of this basin to improve and to report something, to improve the relationships between Mexico and the United States related with the weather management in this area… So at the beginning of this project we had to contact the people of Mexico and had several meetings along the border to explain to everyone what was the whole idea and what was the proposal for our project just to get the collaboration from everybody. For example, you go to the border and had meetings with stakeholders and they are always afraid, "Hey, you are talking about, you are claiming my water?" (laughs). (Post-Doc, Environmental Engineering, Texas)

As the excerpt above delicately suggests, data can be politicized in the context of scientific research. Datasets are not simply 'owned' by scientists, but the government that has international rights over the region from which data is collected. In other words, data sharing in big data science, cyberinfrastructure, and cooperative scientific work calls for the need to manage the tension between countries over shared problems. Unless the scientific problems being investigated is a pressing issue in the region, severely affecting multiple countries, getting agencies and scientists in different countries to share datasets can be a geopolitical challenge.

CONCLUSION

While the vision of data sharing for big data science using cyberinfrastructure for cooperative scientific work is promising, the practice of data sharing is full of challenges. This paper discusses three such challenges: single-investigator culture of traditional academia, personal reluctance to show accidental data contamination, and geopolitics of data ownership. I argue that in order for data sharing to become an enduring practice, the dialectical tension approach [5] can be used to create a sense of urgency of the scientific problems as pressing, in order to put pressure on academic systems, researchers, and governments to shift their current practices towards the vision of community data sharing. By focusing on challenges and barriers [6], we can gain insights for making progress in adoption and implementation of the practice of data sharing.

ACKNOWLEDGMENTS

I thank all the interview participants for their insights. I also thank Larry Browning, Dawna Ballard, Susan Corbin, and Rion Dooley for their support of the project.

REFERENCES

1. Kee, K. F., Cradduck, L., Blodgett, B., & Olwan, R. Cyberinfrastructure inside out: Definitions and influencing forces shaping its emergence, development, and implementation. In D. Araya, Y. Breindl & T. Houghton (Eds.), *Nexus: New intersections in Internet research* (pp. 157-189). New York: Peter Lang. 2011.
2. Bird, I., Jones, B., & Kee, K. F. The organization and management of grid infrastructures. *Computer, 42*, 1 (2009), 36-46.
3. Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., . . . Wright, M.. Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-ribbon Advisory Panel on Cyberinfrastructure. Washington, DC: National Science Foundation. (2003). Retrieved December 19, 2006 from http://www.communitytechnology.org/nsf_ci_report/.
4. Corbin, J., & Strauss, A. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 13*, 1 (1990), 3-21.
5. Kee, K. F., & Browning, L. D. The dialectical tensions in the funding infrastructure of cyberinfrastructure. *Computer Supported Cooperative Work, 19*, 3-4 (2010), 283-308.
6. Lee, C. P., & Bietz, M. *Barriers to the adoption of new collaboration technologies for scientists*. Paper presented at the ACM Conference on Computer-Human Interaction (CHI), 2009