Research Ethics for Open Online Community Data: A Case Study of Human Subjects Research Online

Katie Shilton
College of Information Studies
University of Maryland College Park
kshilton@umd.edu

Sean Goggins
School of Information Science and Learning Technologies
University of Missouri
GogginsS@missouri.edu

Brian Butler
College of Information Studies
University of Maryland College Park
bsbutler@umd.edu

Susan Winter
College of Information Studies
University of Maryland College Park
sjwinter@umd.edu

Abstract

Understanding research ethics for open online community data is an ongoing challenge. As part of a larger initiative to prototype a “data factory” for open online community data, our team is investigating research ethics challenges researchers experience when collecting, managing, and analyzing this type of data. Our case study will present initial analysis of interviews with researchers engaged with open online community data. These interviews will focus on ethical challenges they have faced, open ethical questions, and what resources might help future researchers meet ethical challenges.

Author Keywords

Research ethics; online community data;

ACM Classification Keywords

K.4.1. Public Policy Issues: Ethics

Introduction

Open online communities (OOC) have emerged as significant drivers of innovation, economic activity, and social well-being. OOCs are networking and collaboration phenomena that facilitate the collective construction of tangible or intangible products using flexible, distributed, and non-hierarchical forms of organization. OOCs play important roles in a wide variety of areas, including software development, knowledge management, education, health, and scientific discovery.

Scholars and practitioners from different disciplines (e.g. computer science, sociology, mathematics, economics, physics, anthropology, organization science, and communications) engage in OOC research. For example, management scholars in free and open source software (FOSS) focus on developing theories of collaboration on these projects drawn from rich, qualitative methods [5], while software engineering scholars address developer coordination tools and specific issues of how to make sense of electronic trace data through software repository mining [3, 12]. HCI scholars in FOSS are particularly focused on how tools might be designed to support different modes of collaboration [4].

Building an Online Data Factory

Building a community of scholars who address differences in research aims, data and method will enable a new, interdisciplinary synthesis of OOC knowledge. To meet this challenge, we are building an *Open Collaboration Data Factory*. The Data Factory will include a common set of easily accessible and replicable research datasets and processes that support research addressing the design and theoretical challenges raised by diverse multi-community OOC systems. The Data Factory will also develop ethical and privacy guidelines for processing OOC data as an integral part of its mission and deliverables.

Research Ethics for OOCs

Open online community data presents a range of challenges to traditional research ethics. Traditional U.S. research ethics as compiled by the Belmont Report focus on respect for persons, beneficence, and justice [8]. OOC data demands that we rethink or reinterpret the traditional ways each of these principles has been represented in U.S-based research.

In the U.S., respect for persons has most widely been interpreted as a mandate for collecting informed consent from participants. Informed consent is challenged by OOC data, however. The data is often nominally public, and collecting informed consent may be difficult or impossible. And in many cases, the public nature of the data makes it unnecessary. But if we interpret respect for persons broadly, we must consider that much of this data documents work processes and practices that *would* have demanded informed consent for data collection in other settings. Contributors to OOC forums may have no idea such data could be harvested by researchers. For example, researchers who investigate sensitive issues such as values or political conflicts have struggled with whether informed consent was necessary [6, 13].

Beneficence is the second Belmont principle challenged by OOC research. Generally understood as assessment of risks and benefits of the research, it is a principle that guides researchers to think through possible negative consequences of their work. While much OOC research in management or HCI may have few risks, research in domains such as health care may have more concern about risks to subjects. And researchers may be challenged to think about the scope and extent of risk in OOC research. Are there unintended consequences of such research, as there have been in other domains (such as Facebook research [14])? One large issue around risk in OOC datasets focuses on anonymity. Preserving anonymity for OOC participants may be a difficult challenge. Reidentification risks abound in big datasets [7, 9], and further investigation of OOC data is needed to determine when and how reidentification risks reside in OOC data.

OOC researchers must also consider whether their research presents a risk not just to individuals, but to the entire community they study. OOC research may also pose a risk to groups and communities. While anonymizing individual-level data may protect individuals from additional scrutiny and exposure, groups and communities are identifying and highlighted. Future attention (or even the current study) may damage the community and potentially deprive involved individuals of an important source of identity and support.

Finally, justice has widely been interpreted as attention to the selection of subjects. We argue that this is an under-investigated area in OOC research. OOC subjects are largely self-selecting. While they may not be likely to include protected populations such as prisoners and children, it might be difficult to tell if participants from such populations (especially children) are included. And from a social justice perspective, OOC participants are generally more affluent and educated than the general population [2, 10]. Reflection is needed about whether this narrow scope generates justice issues.

All of these challenges reside around a fundamental set of questions about whether OOC data is somehow different from existing human subjects datasets, or whether we can draw helpful analogies between this sort of data and more familiar sets of data. Is there something unique about the nature of work [5], the nature of participants [10], or even, as some have suggested, the virtues of participants [1]?

Methods

We will use snowball sampling and citation chaining to find researchers experienced with human subjects issues in OOC research. The interview protocol focuses on how researchers discover and resolve ethical challenges in their work. Interviews will be transcribed and coded using iterative coding by the research team. We expect that areas of focus will include work practices or institutional forces that led to discovery of ethical challenges [11], real-world processes for resolving those challenges, and open questions for the research ethics community.

Results

Findings from this research will be used to support the development of resources for navigating research ethics in OOC investigations. These resources, which might include language for consent forms or IRB applications; flow charts for considering ethical challenges and consequences of data collection, processing, and storage methods; or compilations of best practices from across the community; will be an integral part of the Data Factory for ease of use by the research community.

We will present preliminary results at the *Ethics for Studying Sociotechnical Systems in a Big Data World* workshop. We also hope to use the workshop setting to refine our approaches, ideas, and potential outcomes with the community of research ethics scholars this event will attract.

Acknowledgements

This work is supported by NSF Grants IIS-1449188 and IIS-

References

[1] Benkler, Y. and Nissenbaum, H. 2006. Commons-based Peer Production and Virtue*. *Journal of Political Philosophy*. 14, 4 (2006), 394–419.

[2] Berinsky, A.J., Huber, G.A. and Lenz, G.S. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*. 20, 3 (Jul. 2012), 351–368.

[3] Bird, C., Rigby, P.C., Barr, E.T., Hamilton, D.J., German, D.M. and Devanbu, P. 2009. The promises and perils of mining git. *6th IEEE International Working Conference on Mining Software Repositories, 2009. MSR '09* (May 2009), 1–10.

[4] Dabbish, L., Stuart, C., Tsay, J. and Herbsleb, J. 2012. Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (New York, NY, USA, 2012), 1277–1286.

[5] Howison, J. and Crowston, K. Collaboration through open superposition: a theory of the open source way. *MIS Quarterly*. 38, 1.

[6] Koepfler, J.A., Shilton, K. and Fleischmann, K.R. 2013. A stake in the issue of homelessness: Identifying values of interest for design in online communities. *Proceedings of the 2013 conference on Communities & Technologies* (Munich, Germany, 2013).

[7] Lease, M., Hullman, J., Bigham, J., Bernstein, M., Kim, J., Lasecki, W., Bakhshi, S., Mitra, T. and Miller, R. 2013. *Mechanical Turk is Not Anonymous*. Technical Report #ID 2228728. Social Science Research Network.

[8] Office of the Secretary of The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Department of Health, Education, and Welfare.

[9] Ohm, P. 2010. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*. 57, (2010), 1701.

[10] Ross, J., Irani, L., Silberman, M.S., Zaldivar, A. and Tomlinson, B. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2010), 2863–2872.

[11] Shilton, K. 2013. Values levers: building ethics into design. *Science, Technology & Human Values*. 38, 3 (2013), 374 – 397.

[12] Spinellis, D., Gousios, G., Karakoidas, V., Louridas, P., Adams, P.J., Samoladas, I. and Stamelos, I. 2009. Evaluating the Quality of Open Source Software. *Electronic Notes in Theoretical Computer Science*. 233, (Mar. 2009), 5–28.

[13] Zhou, Y., Fleischmann, K.R. and Wallace, W.A. 2010. Automatic Text Analysis of Values in the Enron Email Dataset: Clustering a Social Network Using the Value Patterns of Actors. *System Sciences (HICSS), 2010 43rd Hawaii International Conference on* (2010), 1 –10.

[14] Zimmer, M. 2010. “But the data is already public”: on the ethics of research in Facebook. *Ethics and Information Technology*. 12, 4 (Dec. 2010), 313–325.

[1]