6-2017

Advancing Open Online Community Science Practice: The Open Community Data Exchange

Georg J.P. Link
University of Nebraska at Omaha, glink@unomaha.edu

Kevin Lumbard
University of Nebraska at Omaha, klumbard@unomaha.edu

Matt Germonprez
University of Nebraska at Omaha, germonprez@gmail.com

Sean Goggins
University of Missouri, gogginss@missouri.edu

Advancing Open Online Community Science Practice: The Open Community Data Exchange[1]

Research-in-Progress

Georg J.P. Link
University of Nebraska at Omaha
glink@unomaha.edu

Kevin Lumbard
University of Nebraska at Omaha
klumbard@unomaha.edu

Matt Germonprez
University of Nebraska at Omaha
germonprez@gmail.com

Sean Goggins
University of Missouri
gogginss@missouri.edu

ABSTRACT

Open online communities produce an enormous amount of digital data. To date, open online community science has been conducted piecemeal, one internet address at a time, often without social or scholarly impact beyond one's own research. The Open Community Data Exchange (OCDX) addresses this issue and develops a metadata specification and community infrastructure for describing, discovering, and sharing open online community data between researchers. The goal of this paper is to outline strategies and considerations for advancing how the development of a metadata specification and community infrastructure can reveal new scientific practice for open online community researchers.

Keywords

Open Community Data Exchange; Open Online Community Science; Science of Science; Open Data

INTRODUCTION

Online behavioral data has been used for research in diverse online contexts, such as online learning outcomes (Bishop and Verleger 2013), political use of social media (Nahon and Hemsley 2014), and valued health benefits (Moorhead et al. 2013). The large volume of online behavioral data, combined with its poor description, creates a number of persistent research challenges that (1) limit the discovery and reuse of data; (2) hinder researchers in combining or comparing data; and (3) make the study of how researchers are creating and using data in scientific inquiry difficult.

The Open Community Data Exchange (OCDX) is a metadata specification and community infrastructure aimed at alleviating these shortcomings (Link et al. 2016). The OCDX is rooted in the current practice of science, carrying the potential for changing the way open online community science is conducted. We explore the core activities of the OCDX, a metadata specification and community infrastructure to advance open online community science by advancing the ways that scientists share, discover, and use data, leading to our research question:

> *How do a metadata specification and community infrastructure aimed at the discovery and sharing of open online community data sets impact the practice of science?*

BACKGROUND

Science of Science

Science of science is the study of scientific practice (Turchin 1977). New tools with improved precision for collecting and analyzing data challenge existing practice and lead to new scientific practice (Kuhn 1970). As the scientific enterprise has grown and evolved, it has required organization and categorization of scientific practice and the formalization of scientific

[1] This project is funded through the National Science Foundation (SAVI project: NSF #1449188; #1449209).

language (Turchin 1977). A formal language allows for aggregating knowledge and for making it accessible to an array of users. In this regard, our research explores how a formalized language (as metadata) is understood within scientific practice to uniformly name, label, and represent data or datasets (Borgman 2015).

Metadata

Metadata is "data about data" (Gilliland 2016). Metadata can be generated automatically when data is collected through technology that records time stamps, dates, location, and context. Researchers that manually collect data create metadata by assigning variable names to survey items, assigning labels to rows or columns in spreadsheets, and record contextual information. Example metadata is provenance information, which describes the origin of a dataset, the transformations made on a dataset, and documents other information necessary to replicate or interpret findings (Borgman 2015).

The Open Community Data Exchange

The OCDX is a metadata specification and community infrastructure to assist solving the challenge of open online community data sharing (Link et al. 2016). Online behavioral data must be described consistently in order to be discoverable and reused by others. Similar to what GenBank (Benson et al. 2015) is doing for biological researchers, the OCDX specification enables researchers to describe and annotate data. The OCDX is specifically built on open online community datasets, accompanying analysis scripts, a bill of materials for datasets (OCDX metadata manifest), and the supporting OCDX community infrastructure (Figure 1).

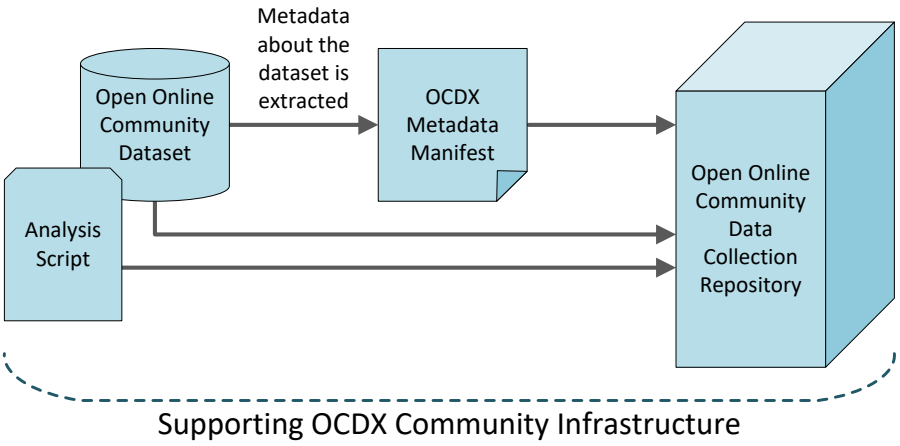

Figure 1: The OCDX-related relationships between open online community datasets, analysis scripts, metadata manifests, data repositories, and supporting infrastructure.

THEORETICAL FRAMEWORK

When data is separated from its context, meaning is lost which poses difficulties. An increasing distance between data creators and data users magnifies the problem that "precise context is never fully transferable" (Borgman 2015, p. 219). Metadata can bridge the distance by capturing meaning in a defined set of information (Borgman 2015). In this, the data creator is tasked with capturing or creating the metadata and making it available to the data user who must understand the metadata to judge the quality, applicability, and usefulness of the data. Likewise, a data user relies on metadata for understanding the nature of the data, the data context, the collection method, and transformations applied to it which is important to judge the authenticity, trustworthiness, and usefulness (Figure 2).

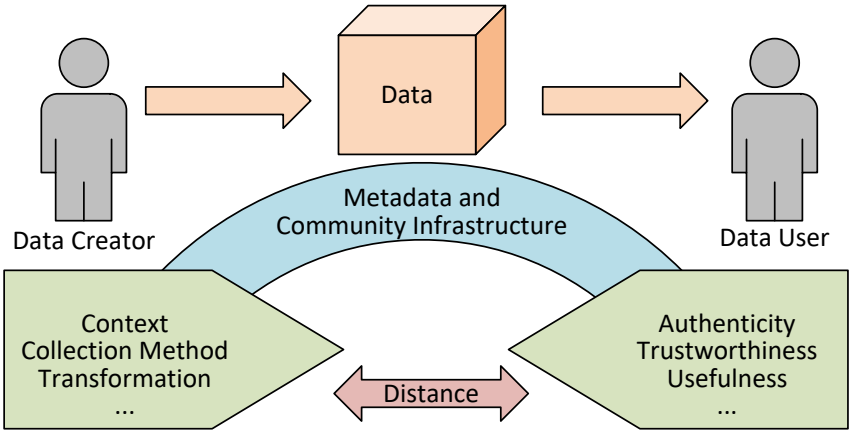

Figure 2: The metadata and community infrastructure bridge the distance between data creator and data user.

PROPOSED METHODS

We address the research question aimed at how the ODCX is understood in the practice of science by conducting field study research and interviews with early adopters of the OCDX standard. We focus at first on small focus groups and a select number of users before inviting more participants. Engaged field study will allow us to understand the cognitive, social, and technological structures of the field by building from our own experiences, thus constructing a "system of meaning within which our experience is embedded" (Dourish 2014, p. 7). The interviews and field notes will be examined and analyzed to expose the practices employed by researchers utilizing open data. We accomplish this through exploratory data analysis (Tukey 1962, 1977, 1980), confirmatory statistical analysis, and modeling to describe how researchers apply metadata specifications in open online community science. Additionally, we will examine electronic trace data in the form of OCDX metadata documents. We will employ Netnographic (Kozinets 2015) techniques to aggregate and study how metadata documents are produced by open online community researchers. Our focus will be to understand the impact of the metadata specification in scientific practice, as well as what data sharing can tell us about the scientific practice for those engaged. The qualitative phenomenological perspective provides a lens to construct how the data and its meaning in the research process are understood differently before and after introducing the metadata and community infrastructure (Creswell 2013). We will verify our data by presenting findings and conclusions to participants of the field study and eliciting their feedback (Miles and Huberman 1994).

CONCLUSION

While advancing science through the OCDX for sharing open community science data is not trivial, a metadata specification and community infrastructure can bridge distances between researchers. Consequently, research will no longer be confined to single locations as research spanning internet addresses and times is made possible. The OCDX bridges the needs of researchers by providing a standard for collecting and sharing research data. As such, researchers may be able to ask "bigger questions" by combining datasets, performing new analyses on secondary data, and bridging disciplines. We intend to uncover these and potentially other outcomes of introducing a metadata and community infrastructure.

ORCID

Georg J.P. Link https://orcid.org/0000-0001-6769-7867

Kevin Lumbard https://orcid.org/0000-0001-9306-3040

REFERENCES

1. Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. 2015. "GenBank," *Nucleic Acids Research*, (43:D1), pp. D30-35 (doi: 10.1093/nar/gku1216).
2. Bishop, J. L., and Verleger, M. A. 2013. "The flipped classroom: A survey of the research," in *ASEE National Conference Proceedings, Atlanta, GA*, (Vol. 30).
3. Borgman, C. L. 2015. *Big data, little data, no data: Scholarship in the networked world*, MIT Press.

4. Creswell, J. W. 2013. *Qualitative inquiry and research design: Choosing among five approaches*, (3rd ed.), Los Angeles: SAGE Publications.
5. Dourish, P. 2014. "Reading and Interpreting Ethnography," in *Ways of knowing in HCI*, J. S. Olson and W. Kellogg (eds.), New York: Springer, pp. 1–24.
6. Gilliland, A. J. 2016. "Setting the stage," in *Introduction to Metadata*, M. Baca (ed.), (3rd ed.), Los Angeles: Getty Publications (available at http://www.getty.edu/publications/intrometadata/setting-the-stage/).
7. Kozinets, R. V. 2015. *Netnography: redefined*, SAGE Publications.
8. Kuhn, T. S. 1970. *The structure of scientific revolutions*, (O. Neurath, ed.), (Second Edition, Enlarged., Vol. II), University of Chicago press, Second Edition.
9. Link, G. J. P., Germonprez, M., Goggins, S., Hemsley, J., Rand, B., and Squire, M. 2016. "The Open Community Data Exchange: Advancing data sharing and discovery in open online community science," in *OpenSym '16 Companion*, Presented at the OpenSym 2016, Berlin, Germany: ACM, August 17 (doi: 10.1145/2962132.2962142).
10. Miles, M. B., and Huberman, A. M. 1994. *Qualitative data analysis: An expanded sourcebook*, (2nd ed.), Thousand Oaks: Sage Publications.
11. Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A., and Hoving, C. 2013. "A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication," *Journal of Medical Internet Research*, (15:4:e85), p. 1-16.
12. Nahon, K., and Hemsley, J. 2014. "Homophily in the guise of cross-linking: Political blogs and content," *American Behavioral Scientist*, (58:10), pp. 1294–1313 (doi: 10.1177/0002764214527090).
13. Tukey, J. W. 1962. "The future of data analysis," *The Annals of Mathematical Statistics*, (33:1), pp. 1–67.
14. Tukey, J. W. 1977. *Exploratory data analysis*, Addison-Wesley series in behavioral science, Reading, Mass: Addison-Wesley Pub. Co.
15. Tukey, J. W. 1980. "We need both exploratory and confirmatory," *The American Statistician*, (34:1), pp. 23–25 (doi: 10.2307/2682991).
16. Turchin, V. F. 1977. *The phenomenon of science*, New York: Columbia University Press.