

A Comparative Study of Data Reuse Among Quantitative Social Scientists and Archaeologists

Ixchel M. Faniel
OCLC Research
faniel@oclc.org

Julianna Barrera-Gomez
OCLC Research
barreraj@oclc.org

Adam Kriesberg
University of Michigan
School of Information
akriesbe@umich.edu

Elizabeth Yakel
University of Michigan
School of Information
yakel@umich.edu

Abstract

This paper presents preliminary findings from a comparative study of data reuse in the quantitative social science and archaeology disciplines. Quantitative social scientists have been engaged in large-scale data sharing and reuse from centralized repositories for over 50 years. In contrast, archaeologists are transitioning from sharing and reusing data on a small-scale with colleagues and museums to large scale sharing and reuse via centralized repositories. In this study, we consider whether approaches to supporting data reuse in quantitative social science can be applied to the archaeological community. Currently we are examining data reuse practices in both disciplines via three points of comparison and will discuss preliminary findings regarding: 1) the nature of context needed during reuse, 2) the use of a bibliography of data related literature, and 3) the role of intermediaries.

Keywords: data reuse, data repositories

Introduction

A common data infrastructure, albeit necessary to support the central goals of eScience, may not arise given minimal research to discover common ground among different disciplines (Borgman, 2007). Comparative studies among eScience initiatives have been informative, but attention has tended to focus on degrees of disciplinary difference. We have learned how different levels of interdependence among scientists and task uncertainty have implications for design and use of eScience infrastructures and how the different histories and configurations of disciplines might impact how researchers contextualize and document research data and processes (Birnholtz & Bietz 2003; Carlson & Anderson 2007; Fry, 2006). However, we know less about the similarities that exist among disciplines and how they can be drawn upon to develop common approaches to support data reuse.

In this study, we explore data reuse in two disciplinary communities – quantitative social science and archaeology. Quantitative social scientists have been engaged in large-scale data sharing and reuse from centralized repositories, some of which have implemented procedures and standards and built expertise and reputations over 50 years (e.g. Inter-university Consortium of Political and Social Research). In contrast, the archaeological community is transitioning from sharing and reusing data on a small scale with colleagues and museums to large scale sharing and reuse via centralized repositories. They too have implemented procedures and standards, but also look to the well-established data repositories to inform their work (e.g. Open Context). A major objective of this study is to examine the data reuse practices in both disciplinary communities to determine whether approaches that support data reuse among the quantitative social scientists might be applied to the archaeological community. In the following paragraphs, we draw from the data reuse literature to discuss the three ways we are comparing data reuse practices. We then describe our research methods and end with a discussion of the preliminary findings and implications.

Acknowledgement: This research was made possible by a National Leadership Grant from the Institute for Museum and Library Services, LG-06-10-0140-10, "Dissemination Information Packages for Information Reuse".

Faniel, I., Barrera-Gomez, J., Kriesberg, A., & Yakel, E. (2013). A comparative study of data reuse among quantitative social scientists and archaeologists. *iConference 2013 Proceedings* (pp. 797-800). doi:10.9776/13391

Copyright is held by the authors.

Data Reuse Literature

This current study examines three points of comparison found in data reuse studies that have implications for the design, development, and use of data repositories. The first is what information about the context of data production is needed during reuse. Although studies agree that context is critical for reuse, there is disagreement as to whether enough can be captured to discover and evaluate reusable data (e.g. Faniel & Jacobsen, 2010; Zimmerman, 2008). The latter hinges on the nature of the context, particularly how it is generated and whether it is captured during data production, rather than on a specific discipline or data type (e.g. Carlson & Anderson, 2007; Jirotko et al., 2005). In separate studies, both field anthropologists and HIV/AIDS laboratory researchers were found to rely on “craft like methods” and tacit knowledge that were difficult to explicate through systematic documentation of the research process (Birnholtz & Bietz, 2003; Carlson & Anderson, 2007). The second point of comparison is the use of bibliographies of data related literature. As an additional means of documenting data, they serve multiple purposes during the data reuse process, including discovering relevant data, constructing systematic data samples, understanding how data have been conceptualized and measured, and keeping informed about community debates around the data (Faniel & Jacobsen, 2010; Faniel, Kriesberg, & Yakel, 2012; Zimmerman, 2007). Lastly, we compare the role intermediaries play between data producers and reusers. While a repository itself and the people who staff it serve as intermediaries, research has also found data reusers rely on other third parties, such as advisors and colleagues, to facilitate data reuse (Faniel et al., 2012; Jirotko et al., 2005).

Methods

Our findings are drawn from data collected during three rounds of interviews conducted between June 2011 and April 2012. In total, we spoke with 66 participants: 44 quantitative social scientists (22 novices and 22 experts) and 22 archaeologists. In our series of semi-structured hour-long interviews, we asked respondents to discuss their experiences reusing data in their particular field of research. Topics of inquiry included how respondents discovered and evaluated data for reuse and their experiences and thoughts about digital data repositories. All interviews were audio recorded and transcribed. We then coded the transcripts using NVivo, a qualitative data analysis software tool. To achieve inter-rater reliability, two members of our project team coded each group of transcripts.

Preliminary Results

The quantitative social scientists reused a broad range of survey data collected from human subjects and local, state, and federal institutions. They also reused quantitative data that had been transformed from qualitative data, such as measures of the “democratic-ness” of a country produced by analyzing newspaper articles. We found much of the context created for these data was digital (e.g. Microsoft Word documents, online surveys) and static. For instance, a typical research design was developed at the beginning of a project, not much was changed for the duration of the study, and data were collected in standard forms. In contrast, the archaeological data were more varied, consisting of images of objects found during a site survey or excavation, textual descriptions of those objects, geographic location data documenting objects within a site, or GIS shape files documenting a site through time. The context was more dynamic, evolving as the project evolved. During excavation of a physical site, documentation was generated daily and research plans were subject to change based on internal and external factors beyond the excavator’s control (e.g., the nature of excavation discoveries, weather conditions).

Despite these differences in the nature of the context, we found similarities between quantitative social scientists and archaeologists. Both groups expressed a need to understand the data producer’s research methods, especially aspects of context that provided insight into how the data producer carried out the research. Both acknowledged that data producers conducted research in different ways. Knowing more about research procedures helped data reusers understand the nuances and make more informed reuse decisions. For instance, quantitative social scientists wanted to know how data producers defined and measured the variables data were intended to capture. Archaeologists were interested in the type of GPS device data producers used to identify site and object location, since these instruments provide location information at different levels of specificity.

Quantitative social scientists also discussed bibliographies of data-related literature. They used articles written by data producers to get more detailed and clear descriptions about measurement and methods or justifications for research decisions. They also used articles written by community members to see how the data were critiqued and reused. Archaeologists also discussed bibliographies as both a source of data discovery, since raw data are often published in appendices, and as a gateway to locating related project data. They also referenced journal articles as a primary step in discovering both projects and published datasets that might be relevant for reuse. Additionally, for archaeologists bibliographies played a central role in locating associated project reports specialists wrote that included faunal analyses or ceramic typologies.

Quantitative social scientists and archaeologists discussed the role of intermediaries in discovering, evaluating, and understanding data. Social scientists cited the help of colleague networks and mentors in finding relevant data and understanding data limitations. Archaeologists sought advice from colleagues in locating relevant data for reuse, but also asked museum curators questions about the data to get a better understanding of its context when the data producers were not available.

Conclusion

This paper presents preliminary findings from a comparative study of quantitative social scientists and archaeologists. Our immediate plans are to continue data analysis to more fully consider study implications. However, early indications suggest that despite differences in how context was generated and captured during data production there were similarities as well differences in the reuse practices of quantitative social scientists and archaeologists. For instance, data reusers from both disciplines wanted context about the data producer's research methods and were able to get enough detail to be able to reuse the data. Both also used bibliographies of data related literature, but for different purposes. Archaeologists relied on bibliographies to facilitate data discovery, whereas quantitative social scientists used bibliographies to facilitate reuse decisions. Data reusers in both disciplines also relied on intermediaries. However, different types of people were used for different purposes. Quantitative social scientists, particularly novices, relied on faculty advisors who had more data reuse experience to find relevant data and understand data's limitations, whereas archaeologists relied on colleagues and museum curators to locate data and associated context.

References

- Birnholtz, J. P., & Bietz, M. (2003). Data at Work: Supporting Sharing in Science and Engineering. ACM Conference on Supporting Group Work (pp. 339–348). Sanibel Island, FL. doi:10.1145/958160.958215
- Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, Mass: MIT Press.
- Carlson, S., & Anderson, B. (2007). What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. *Journal of Computer-Mediated Communication*, 12(2), 635–651. doi:10.1111/j.1083-6101.2007.00342.x
- Faniel, I. M., & Jacobsen, T. E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Computer Supported Cooperative Work*, 19(3-4), 355–375. doi:10.1007/s10606-010-9117-8
- Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data Reuse and Sensemaking among Novice Social Scientists. In Grove, A. (Ed.), *Proceedings of the 2012 American Society of Information, Science, and Technology*. Silver Spring, MD: American Society of Information, Science, and Technology.
- Fry, J. (2006). Coordination and Control of Research Practice across Scientific Fields: Implications for a Differentiated E-Science. In C. Hine (Ed.), *New Infrastructure for Knowledge Production: Understanding E-Science* (pp. 167–187). Hershey, PA: Information Science Pub.
- Jirotko, M., Procter, R., Hartwood, M., Slack, R., Simpson, A., Coopmans, C., Hinds, C., et al. (2005). Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study. *Computer Supported Cooperative Work*, 14(4), 369–398. doi:10.1007/s10606-005-9001-0

- Zimmerman, A. (2007). Not by Metadata Alone: The Use of Diverse forms of Knowledge to Locate Data for Reuse. *International Journal on Digital Libraries*, 7(1-2), 5–16.
doi:10.1007/s00799-007-0015-8
- Zimmerman, A. S. (2008). New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. *Science, Technology & Human Values*, 33(5), 631–652.
doi:10.1177/0162243907306704