Reusing scientific data: A research framework

Ixchel M. Faniel
School of Information
University of Michigan
1085 S. University Avenue
Ann Arbor, MI 48109-1107
ifaniel@umich.edu

Trond E. Jacobsen
School of Information-North
University of Michigan
1075 Beal Avenue
Ann Arbor, MI 48109-2112
trond@umich.edu

ABSTRACT

Increasing the supply and dissemination of scientific data is no guarantee it will be reused. To achieve greater data reuse, especially over the long term and on a large scale, we contend that a more systematic study of data reuse practices is needed. In this position paper we introduce a data reusability assessment framework, compare and contrast findings from its application to the earthquake engineering community with prior data reuse studies, and outline future research. Understanding data reuse is a critical dimension of designing systems and policies that support and accelerate collaborative science using cyberinfrastructure.

Author Keywords

Data reuse, data quality, data sharing, trust

INTRODUCTION

To date the issues and challenges related to scientific data reuse have received less attention than those related to data management and sharing [1-4]. These issues are important, but CSCW researchers should not assume data supply - creation and dissemination - is the fundamental obstacle [5, 6]. Regardless of the scale of technology investments to manage data or the elegance of incentives to share data, understanding how scientists decide whether or not to reuse data is critical if greater scientific collaboration and innovation are to be achieved.

Scholarly communication is changing where the actual data, not just claims about the data, are being made available for reuse. The changes are driven in part by nature of emerging scientific challenges and in part because of requirements from funding agencies to share data. Despite significant progress in several notable cases, in most disciplines there is a "scandalous shortfall in the sharing of data by researchers" [7].

We contend that this is due in part to the need to provide detailed information about the context of data production. Context information describes the set of interrelated environmental conditions where data are produced. It may include the names of the lead researchers, descriptions of test facilities equipment, details about test setups, including procedures, materials, and specimens, descriptions of data limitations, and descriptions of data conversions (i.e. from raw to converted, corrected, and derived).

Some context information is being captured as metadata in centralized data repositories. However, the types of context information needed to promote data reuse are unlikely to be provided by metadata alone [8]. A common obstacle has been capturing the subtle aspects of context information that are tacit, hard to articulate, or beyond the original research objectives. It has also been difficult to predict what kind and how much context information should be captured to ensure that data are reusable for a potentially broad and heterogeneous array of scientists and non-scientists.

This position paper introduces a data reusability assessment framework to address these difficulties. The framework directs us to ask: 1) what evaluative questions do people ask themselves when assessing the reusability of others' data; 2) what context information do people need to answer the evaluative questions; and 3) how do they get the context information? We derive and describe the types of context information important (2) in one community of scientists primarily by examining considerations relevant to questions (1) and (3). We believe the data reusability framework can inform data sharing efforts and system design, focus research on data reuse, and highlight differences and commonalities in reuse practices within and across disciplines. The workshop will produce valuable feedback about the framework and stimulate discussions about data reuse practices and research.

THE REUSABILITY ASSESSMENT FRAMEWORK

Drawing from prior research, we find that scientists generally ask themselves a series of evaluative questions when assessing the reusability of someone else's data:

- Are the data relevant? It is an assessment of the degree to which the data meet the problem at hand, including the strategies scientists use to locate data [9]

This position paper was accepted to the CSCW 2010 Workshop on "The Changing Dynamics of Scientific Collaboration".
CSCW 2010, Savannah, GA, February 6-10, 2010

- Can the data be understood? Here scientists want to determine whether they can comprehend the meaning and intent behind the data [8].
- Are the data trustworthy? Scientists ask whether data are credible, reliable, and valid. Credibility assessments are typically based on the attributes of data producers [10], while reliability and validity assessments are direct evaluations of the data [11].

The framework we present here is unique in several respects. First, existing studies typically address one or two evaluative questions, providing a partial view of data reuse at best. At worst it suggests that the assessment examined in a particular study is the only assessment scientists make. For example, environmental planners and botanists use reputations to assess credibility [10, 12], but we do not know if reputation influences assessments of reliability, validity, relevance, or how data are understood.

Second, existing reuse studies typically examine the resources scientists use to make reuse assessments in isolation. For example, we know prior knowledge, face-to-face communication, and community membership are used to make reusability assessments, but we do not know whether and how these resources are used together.

Third, some resources have not been examined much at all. For example, scholars recognize the importance of documentation [8, 12, 13], but few studies have examined how it is used to assess the reusability of data [11]. As a result we know little about the context information documentation provides and whether or not it is useful. Yet, inadequate documentation is likely to impede science as data collection methods continue to become more innovative, complex, and large scale. Even people producing data for their own use are realizing memories, prior knowledge, and oral traditions are a less useful means to rely on context information and increasingly document their data for their own use [2].

Only when data reuse studies examine all data reusability assessments at the same time in the same study can we begin to draw conclusions about scientists' data reuse practices in full. We contend that scientists do not consider these components in isolation. Nor do they rely on only one resource [14, 15]. To understand how working scientists make reuse decisions we must examine all the assessments they make and the resources they use during evaluation, including data documentation. We must examine data reuse practices as they are if we are to improve them.

PRELIMINARY RESEARCH

We applied the data reusability framework to study data reuse among earthquake engineers (EE) affiliated with the George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES). NEES is an NSF-funded cyber-infrastructure (CI) initiative linking numerous sites and researchers for collaboration, access to advanced laboratories, and a shared data repository.

EE researchers rely heavily on experimental studies and numerical computation modeling. Experimental studies are 2-3 year investigations yielding data from instrument readings (i.e. sensors) attached to the specimen under study (e.g. column, beam, etc.). Numerical computation models simulate events and are usually validated using experimental data.

To date, reuse within the EE research community has been primarily local and one-to-one, with researchers asking colleagues to share experimental data typically produced in university labs. File formats vary greatly and no formal standards or guidelines for data documentation exist but is often obtained from data reports and dissertations. The documentation typically spans hundreds of pages and may include such context information as: descriptions of equipment; structural configurations; attachment of test specimens to the equipment; sensor descriptors; input motion files; descriptions of data acquisition systems, and photographs and video capture of structural damage.

Findings

As part of a 2 ½ year investigation of NEES we examined data reuse in semi-structured interviews with 14 EE researchers and through responses from 117 survey respondents. We examined the purposes of data reuse, willingness to use a shared repository, sources and types of data, perceived benefits of data reuse, and how these scientists ask and answer the evaluative questions in the reusability assessment framework. We find that EE researchers rely on different resources (or rely on those resources differently) and balance reuse considerations differently than reported in earlier studies [8, 9, 13].

One issue we aim to explore at the workshop is the degree to which findings about the EE research community anticipate broader future directions in data production, sharing, and reuse in other contexts with large-scale, multi-site, collaborative science and shared data repositories.

Are the data relevant?

EE researchers prefer to make assessments about data relevance with less rather than more context information. First they generate key criteria, (e.g. types of structures), and other required parameters for validating their model. They try to match the criteria against context information available in journal articles or through conversations with members of their personal networks. Our findings confirm those in other studies [9]. We find EE researchers develop problem-specific criteria and use the affordances of journals for quick relevance assessments [16-19]. We also find that some EE researchers in small sub-communities prefer to talk with colleagues in their personal network instead.

Can the data be understood?

Scientists use the resources of prior knowledge, communication with people offering assistance, and data documentation to establish confidence they can understand the data. For example, access to assistance by face-to-face communication is important to understanding data for HIV/AIDS researchers [8] while other communities use help desks and workshops [11]. Scientists also use prior experience collecting comparable data or using similar procedures as a means to understand and use colleagues' data [13]. We find EE researchers use of a mix of prior knowledge and perceptions of the availability of assistance and data documentation to assess if a colleague's data can be understood. Most EE researchers reuse colleagues' data to validate their model, meaning they need to replicate experimental results. Consequently, EE researchers have a need to understand exactly what happened during experiments in full and they rely on documentation much more than communication with colleagues.

Because our study considers these resources together, we can demonstrate that for EE researchers all three are important, but data documentation is more important in data reuse decisions. A series of three paired sample t-tests showed a significant mean difference among the three resources – documentation, assistance, and prior experience. This finding may hold great significance if other communities using large-scale CI for basic research confront needs for context information in ways that are comparable to those among EE researchers.

Are the data trustworthy?

Of the three components of trustworthiness – credibility, reliability, and validity – EE researchers place greater emphasis on data reliability and validity rather than credibility. In contrast to scientists examined in other studies [e.g. 10, 12, 13], EE researchers do not assess data credibility directly nor do they use colleagues' reputations as a proxy. Instead they use colleagues' documentation to assess data reliability and validity. EE researchers use different kinds of context information to assess whether colleagues' data are reliable and valid. They use the context information to retrace the experiment, including how data were collected, processed, and analyzed. They also use context information to identify the errors during experiments and how they were addressed. Interestingly, the major concern in trusting colleagues' data for EE researchers is not whether problems occurred during experiments. Rather it is whether colleagues' documentation contains sufficient context information describing errors and their resolution.

FUTURE RESEARCH

Our preliminary results support propositions that data reuse practices vary across scientific communities, user types, data types, and reuse purposes [20-22]. Unfortunately, we know little about such differences, how they influence data reuse, or what they imply for CI environments to support long-term, large-scale data management, sharing, and reuse. Our goals are to: 1) verify potential differences, 2) use the differences to compare within and across scientific communities, and 3) determine their impact on the kinds of context information needed to assess reusability. We focus here on two future studies. The first compares how data reuse practices vary across user types within a community (e.g. scientists, citizens, policy makers, etc.). The second examines how reuse practices vary across data types and scientific communities.

Study 1: User Types within a Scientific Community

A National Science Board (NSB) report describes three categories of digital data collections: research, resource, and reference collections, based on user type [22]. Our focus would be on how different types of specific users (e.g. scientists vs. practitioners) assess the reusability of data. Particular attention would be paid to the amount and kinds of context information different user types need when assessing data reusability.

Study 2: Comparing Three Scientific Communities

The NSB report also notes that observational, computational, and experimental data lead to different archival and preservation choices affecting how data are collected, arranged, and maintained [22]. Such choices also affect how data are shared and reused. Research suggests the history and configuration of a scientific discipline influences scientists' ability to contextualize and document their data [11]. We will conduct a case study comparison of scientists' reuse of colleagues' data within three scientific communities. The objective is to understand how the amount and kind of context information accompanying data varies across data types and scientific communities and how these differences influence reusability assessments.

Our Goals and Contributions

Our goal is to understand the varying ways people combine the data reusability assessment framework with different degrees and kinds of context information when making reuse decisions. The proposed studies should help us understand not only differences in data reuse practices among the different user types and why they occur, but also the differing reusability assessments made and an inventory of the kinds of context information needed. We also believe the findings would be helpful to those designing and developing data intensive CI environments that support the needs of scientists and non-scientists and deal with different data types and scientific communities.

CONCLUSION

This position paper describes ongoing work we believe makes several valuable contributions to future research on data reuse. We believe our framework can give shape to a shared research agenda for the study of data reuse and the kinds of context information needed to support it. Our findings suggest that future research examine the range of evaluations made and resources used to assess reusability. Although we have outlined the multiple assessments and resources at play within the EE community, future research may identify even more. We also urge increased attention by researchers to data documentation given broad scientific and policy goals to conduct more innovative, complex scientific studies producing data intended for long-term, large-scale sharing and reuse.

ACKNOWLEDGEMENTS

This research is based on work supported by the National Science Foundation, Award # CMMI-0714116 to the University of Michigan. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Baker, K. S. and Yarmey, L. (2008), "Data Stewardship: Environmental Data Curation and a Web-of-Repositories", 4th International Digital Curation Conference, Edinburgh, Scotland, December, 2008.
2. Borgman, C. L., Wallis, J. C., Mayernik, M. S. and Pepe, A. (2007) In *ACM/IEEE Joint Conference on Digital Libraries* Vancouver, BC.
3. Karasti, H. and Baker, K. S. (2008), "Digital Data Practices and Long Term Ecological Research Program Growing Global" *The International Journal of Digital Curation,* **2,** 42-58.
4. Karasti, H., Baker, K. S. and Halkola, E. (2006), "Enriching the Notion of Data Curation In E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network " *Computer Supported Cooperative Work,* **2006,** 321-358.
5. Schofield, P. N., Bubela, T., Weaver, T. and Portilla, L. (2009), "Post-publication sharing of data and tools" *Nature,* **461,** 171-173.
6. Toronto International Data Release Workshop (2009), "Prepublication data sharing" *Nature,* **461,** 168-170.
7. Editors (2009), "Data's shameful neglect" *Nature,* **461,** 145.
8. Birnholtz, J. P. and Bietz, M. (2003), "Data at Work: Supporting Sharing in Science and Engineering", *ACM Conference on Supporting Group Work, Sanibel Island, FL, 2003.*
9. 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,** 5-16.
10. Van House, N. A. (2002), "Trust and Epistemic Communities in Biodiversity Data Sharing", *ACM Joint Conference on Digital Libraries, Portland, OR, 2002.*
11. Carlson, S. and Anderson, B. (2007), "What are Data? The Many Kinds of Data and Their Implications for Data Re-Use" *Journal of Computer-Mediated Communication,* **12**.
12. Van House, N. A., Butler, M. H. and Schiff, L. R. (1998), "Cooperative Knowledge Work and Practices of Trust: Sharing Environmental Planning Data Sets", *ACM Conference on Computer Supported Cooperative Work, Seattle, Washington, 1998.*
13. Zimmerman, A. (2008), "New knowledge from old data: The role of standards in the sharing and reuse of ecological data" *Science, Technology, & Human Values,* **33,** 631-652.
14. Bourne, P. (2005), "Will a Biological Database be Different from a Biological Journal" *PLoS Computational Biology,* **1,** 179-181.
15. De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Goderis, A., Michaelides, D. and Newman, D. (2008), "myExperiment: Defining the Social Virtual Research Enviornment", *4th IEEE International Conference on e-Science, Indianapolis, Indiana, December, 2008.*
16. Bishop, A. P. (1999), "Document structure and digital libraries: how researchers mobilize information in journal articles" *Information Processing and Management,* **35,** 255-279.
17. Ellis, D., Cox, D. and Hall, K. (1993), "A Comparison of the Information Seeking Patterns of Researchers in the Physical and Social Sciences" *Journal of Documentation,* **49,** 356-369.
18. Sandusky, R. J. and Tenopir, C. (2007), "Finding and using journal article components: Impacts of disaggregation on teaching and research practice" *Journal of the American Society of Information Science and Technology,* **59,** 970-982.
19. Stewart, L. (1996), "User acceptance of electronic journals: Interviews with chemists at Cornell University" *College & Research Libraries,* **57,** 339-349.
20. Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G. C., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. and Wouters, P. (2004), "Promoting Access to Public Research Data for Scientific, Economic, and Social Development" *Data Science Journal,* **3,** 135-152.
21. Borgman, C. L. (2007) *Scholarship in the Digital Age: Information, Infrastructure, and the Internet,* MIT Press, Cambridge.
22. National Science Board (2005) *Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century*, National Science Foundation, Washington, D.C.