Data, Data Use, and Scientific Inquiry: Two Case Studies of Data Practices

Laura A. Wynholds, Jillian C. Wallis, Christine L. Borgman, Ashley Sands, Sharon Traweek*

Department of Information Studies, University of California, Los Angeles

*History and Gender Studies Departments, University of California, Los Angeles

wynholds@ucla.edu, jwallisi@ucla.edu, borgman@gseis.ucla.edu, ashleysa@ucla.edu, traweek@history.ucla.edu

ABSTRACT

Data are proliferating far faster than they can be captured, managed, or stored. What types of data are most likely to be used and reused, by whom, and for what purposes? Answers to these questions will inform information policy and the design of digital libraries. We report findings from semi-structured interviews and field observations to investigate characteristics of data use and reuse and how those characteristics vary within and between scientific communities. The two communities studied are researchers at the Center for Embedded Network Sensing (CENS) and users of the Sloan Digital Sky Survey (SDSS) data. The data practices of CENS and SDSS researchers have implications for data curation, system evaluation, and policy. Some data that are important to the conduct of research are not viewed as sufficiently valuable to keep. Other data of great value may not be mentioned or cited, because those data serve only as background to a given investigation. Metrics to assess the value of documents do not map well to data.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: User Issues.

General Terms

Documentation, Design, Human Factors, Standardization.

Keywords

Scientific data, data practices, data sharing, data citation.

1. INTRODUCTION

Data curation is an immediate concern of the digital libraries community and a theme of this conference. Data are proliferating far faster than they can be captured, managed, or stored – all of which are prerequisites to actual curation. A major motivation for data curation is keeping data for future use and reuse. Data curation faces the challenge that the data must be housed, managed, and made accessible prior to use, but actual uses may not be known until after sizeable investments are made. Data curation research has been engaged with outstanding questions of use: How are data used and how are they reused? What types of data are most likely to be reused, by whom, and for what purposes? Answers to these questions are needed to inform information policy and the design of digital libraries to support the capture, management, use, and reuse of data.

Prior work in information studies has not focused on these questions of use and reuse. Information needs and uses studies, which have a long tradition in the information science literature [4, 10], are more concerned with how people seek documents than with the uses they make of the content therein. Research on the seeking and use of documents does not transfer directly to the seeking and use of research data. The science and technology studies literature is more concerned with the processes by which publications and data are created than with subsequent curation activities [6–8]. Little of the research in the social studies of science on data origins has been applied to problems of data management and curation. The present study assesses scientific data use with the intent to inform information policy and the design of digital libraries for data.

Over the last decade we have been conducting interview and field studies to investigate scientific data practices in environmental sciences, marine biology, ecology, seismology, computer science, engineering, and astronomy[1, 3, 14–19]. Portions of the studies included questions about how data were collected by individuals, teams, other parties. This paper addresses two research questions, at two research sites:

1. What are the characteristics of data use and reuse within each research community?
2. How do characteristics of data use and reuse vary within and between research communities?

We identify the characteristics of the "use" of various types of data from the perspectives of individual researchers and of teams. The same data may be used in multiple ways, depending on the research activity. "Uses" may be understood much differently by researchers than by digital library designers, librarians, and archivists. Explanations of "data use" will inform information policy and the design of digital libraries.

2. BACKGROUND

Each scientific community differentiates equipment, data, analytic tools, and findings that currently are regarded as stable (widely used and no longer under active debate) and the search for new kinds of questions, methods, equipment, data, and analytic strategies in each field [9, 11]. Data curation must accommodate stability and instability of daily work and of data sources, large and small.

Scientists pursue research questions, develop hypotheses and theories, and gather data as evidence to support their inquiry. The forms and types of data will vary by many factors, including the stage of inquiry, characteristics of the research domain, and how much is known about the research problem [5, 7, 8]. Choices of data, metadata, analytic tools, and specializations are constantly being revised [19]. Many, if not most, scientific fields are becoming more data-intensive with advances in instrumentation such as sensor networks. As new instruments and forms of data become available and as communities respond to new requirements for data management plans, scientific practices are

JCDL'12, June 10-14, 2012, Washington, DC, USA.

in flux. The flux creates opportunities to study data production, use, and reuse.

Policies such as requirements for data management plans are predicated on the expectation that data will have future value for reuse. The question of use in relation to scientific data is particularly important for digital libraries. Conceptual frameworks of what constitutes use have a long history in relation to documents, bibliographic sources, and texts [13], but less so for data. These frameworks have been formalized into metrics that have become enormously important in assessing scholarly work. It remains unclear, however, how these concepts and metrics of use will apply to data and whether the same constructs can be applied without modification. Efforts to standardize data citation try to map extant publishing frameworks to data, revealing the considerable differences in granularity and stability.

The ubiquity and generality of "use" as a term makes it particularly difficult to study. Even in the context of documents, information use is difficult to capture and has been called 'theoretically underdeveloped' [12]. Use metrics such as citations and download counts are popular indicators of the use of scholarly works. Citations are at best imprecise indicators of the use of information-bearing objects [2]. Comparisons between uses of documents and of data offer a starting point for analyses. Data are more complex and varied than documents. Data are amorphous, taking forms that range from physical specimens to bit streams. While the use of publications has a different basis than the use of data, publications are central to scientific practice and are the means by which most data are reported. The relationship between citations and data is further complicated by the difficulty of differentiating references to publications versus references publications that contain or describe data.

3. RESEARCH METHODS

The research reported here compares findings from parallel studies in sensor networks and astronomy. Since 2002 we have studied data practices, management, and curation in the Center for Embedded Network Sensing (CENS), a multi-disciplinary, science and technology research center. Research protocols and interview questions used in the CENS data practices studies were later adapted to research on astronomers. Our astronomy focus in this paper is builders and users of the Sloan Digital Sky Survey (SDSS), a large, long-term, and well-known data-driven project.

CENS research is "small science," with small teams and emergent data collection methods [3], whereas the SDSS research is "big science," with large teams and elaborate data collection methods. However, the two sites have much in common, allowing us to make a series of comparisons. CENS and SDSS share these attributes: decade-long projects with multiple sources of funding, collaborations of multiple institutions, and multiple teams that have evolved over the course of their research cycles. They are obligated by their funders to share data and findings. Participants collaborate within and across disciplines at each site. Scientists and technology researchers also collaborate to develop instruments and software. Both types of collaborations require substantial infrastructure investments for data collection.

Our CENS and SDSS studies employed semi-structured interviews and field observations to investigate data use and reuse within these communities. Interviews ranged from 45 minutes to 2 hours, with an average of 60 minutes per interview. Participants were asked about their data and data practices. In most cases, their data were observations or output from models; in some cases data may be code, software, or computer systems.

All interviews were recorded and transcribed. Interviewees and interviews were assigned unique identifiers and names redacted upon request per Institutional Review Board procedures. Interviews were coded using NVivo software. The initial codebook was developed for the first round of CENS interviews in 2006 and significantly revised for the second round of CENS interviews in 2009. The SDSS codebook was developed from the CENS 2006 codebook and adapted to the specifics of the astronomy research questions. Inter-coder reliability tests were performed by our team on both projects. Findings presented below are based on selected questions and relevant responses to questions about data types, data sources, and uses of data.

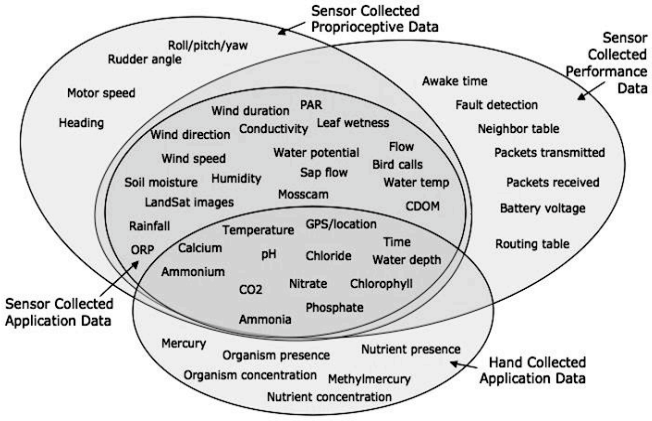
4. RESULTS

The following results are organized by the two study sites, CENS and SDSS. Within each of these two sections, results first are described for data sources and types, then by type of inquiry, and lastly by uses and reuses of data within each community. Comparisons between these sites are presented in the discussion section.

4.1 Center for Embedded Networked Sensing

CENS researchers collect observational, experimental, and simulation data. We have classified CENS data into four categories of data collected during sensor network deployments: sensor-collected application science data, sensor-collected network health data, sensor-collected proprioceptive data, and hand-collected application science data [16], see Figure 1. Researchers supplemented their data with external data including national observatories such as NOAA, where they find tidal estimates, or NASA's MODIS satellite, which provides remote sensing images, and the USGS for gravitational information [1].

Figure 1: CENS data organized by collection method and use.



Inquiry at CENS falls into three, sequential categories: (i) proof-of-concept development of new equipment, algorithms, and systems, (ii) refining equipment, algorithms, and systems for use by the application scientists, and (iii) scientific discovery on the part of the application scientists using CENS equipment, algorithms, and systems. The first and third types of inquiry are performed by the technology and scientific researchers, respectively, and counts as scholarship to each community. The second type is characterized by technology and application science researchers working closely together in the field to transform proof-of-concept technologies into research-grade or commercial-grade technologies that the application scientists employ in field or laboratory environments.

Despite the diversity of the CENS community with respect to the phenomena they study, the data they collect, and the actions they apply to data, a number of commonalities exist. Almost all the participants mentioned the importance of their data in identifying patterns or trends, which includes identifying interesting locations and phenomena for further study. From these patterns and trends, most of the researchers would then construct hypotheses or models. When out in the field, about half the participants mentioned relying on data to provide real-time feedback to ensure that they are capturing phenomena of interest. Electrical engineering researchers feed scientific data to algorithms that drive data-collection robots or that redistribute networked sensors to improve data capture. Raw data from the sensors must be processed through scientific models that represent "the real world" prior to analysis. Transformations include adjusting data with calibration and "ground-truthing" measures collected about each sensor, feeding data into algorithms or simulations to generate new data, and translating indirect measures to data about phenomena that can measured only by proxy. Data analyses may include inductive correlations, deductive hypotheses, or model testing. Other possible analyses are performance evaluation of equipment, algorithms, methods, systems, and meta-analysis. We identified six dimensions of CENS data: (a) background and foreground, (b) observation, experimental, and simulation data, (c) old and new, (d) collection in lab or field , (e) raw versus processed, and (f) collection by the team or obtained from external sources.

4.2 Sloan Digital Sky Survey

The researchers we interviewed provided examples of observational, experimental, and simulation data that they relied upon in the course of their research. Most participants reported drawing data from multiple sources; notably, many drew from sources across multiple wavelengths. Researchers distinguished between the physical instrument as a data source and data obtained from venues such as repositories or catalogs. For example, a researcher might acquire Hubble Space Telescope (instrument) data that had been processed and published in a journal article (venue). Figure 2 summarizes the types of data products mentioned in the interviews.

Figure 2: Types of data products used.

Data Type	Example
Instrument Outputs	Ground Based Telescopes Space Based Telescopes
Model Outputs	Simulation Outputs
Structured Data Products	Sky Surveys Value Added Catalogs
Federated Queries	Virtual Observatory Services

Scientists' interactions with their data depend upon the specific activity and stage of research. Inquiry types identified in our interviews included hypothesis testing, open-ended inquiry, known object inquiry, and multi-faceted inquiries. Open-ended inquiry refers to cases where the scientists were looking for previously unknown or undescribed phenomena. One astronomer articulated this mode as, *"let's populate it with my data and I'll see what's in there."* Other scholars described more precise, known-object queries which included looking for physical properties such as white dwarf spin periods or calculating the intrinsic flux of specific galaxies. Researchers are more concerned with demonstrating innovative inquiry than with creating good, long-range, well-calibrated data products.

Researchers had a number of different ways to approach original and cutting-edge science based on their data, including identifying phenomena not previously observable, obtaining observations at a scale not previously possible, comparisons of data from multiple studies, and asking scientific questions that could not have been asked previously. The approaches were strongly linked with concepts of use, the type of inquiry, and their definition of data. Researchers foregrounded their analysis in discussions of use, listing only data associated with certain types of use. Background uses included obtaining baseline data from established sources, using data to calibrate instruments or analyses, and creating simulations and models. Often the background uses were not mentioned in discussions of data sources, but rather mentioned during discussions of instrument design, data processing, or analysis. Another mode of interaction with the data is to reproduce results to extend or to verify the validity of prior research. Foreground uses, in contrast, are those that drive the scientific inquiry.

5. DISCUSSION

We have reported on our findings for two research questions, based on interviews in two complementary communities, the Center for Embedded Networked Sensing (CENS) and users of Sloan Digital Sky Survey (SDSS) data. Our goal was to identify uses and reuses from the perspective of the individual researchers and their communities, and to apply those findings to information policy and to the design of digital libraries.

We found that CENS researchers and astronomers alike describe their data with respect to the purposes for which they are used. Their data exist only in relation to their research question, hypothesis, model, instrument, or study. Importantly, these researchers also act on data in ways that would be considered "use" by librarians, archivists, digital library developers, information policy makers, and other researchers, and yet are not viewed as "use" by the actors themselves. Such uses include seeking data from external sources and employing them as comparisons or calibration metrics.

Reuse suggests using data repeatedly, whether for the same or different purposes. We identified cases in CENS, particularly, where researchers would maintain some laboratory or field data for later analysis, either alone or in combination with new data. Rarely were these data deposited publicly; rather, they were kept for reuse within the research team. Additionally, both CENS and SDSS users drew upon public repositories of data for background information. Some SDSS users also used data from external sources for foreground purposes to drive their inquiry. Thus taking data from public repositories and sky surveys are reuses in the sense of "using again," but not in the sense of maintaining a team's empirical data in ways that those data might be exploited by others in the future.

Some data that are essential to the research process are not kept at all, and thus are not available for future use by the research team or others. In CENS, these include data produced during the long processes of iterative testing and evaluation of sensor networks, which is the most collaborative part of this type of research. Among SDSS users, these may include similar iterative testing and evaluation of instruments, calibration, and merging of data from multiple sources.

6. CONCLUSIONS

"Use" is a term with little value unless qualified. "Data" adds another level of complexity to the understanding of use. Data are extremely varied and rather than being fixed, they shimmer. Data

uses vary by type of inquiry, type of data, and interactions between them.

Scientific inquiry varies widely in methods and sources of evidence. Data and data practices vary accordingly. Practices vary by domain, by lab, and by individual. CENS and SDSS research sites are exemplars of small science and big science. Both sites are long-term projects, with individuals participating in the science, technology, or both types of research. The individuals we studied included astronomers, computer scientists, electrical engineers, environmental engineers, habitat ecologists, marine biologists, seismologists, and researchers in several related fields.

The complexity of these interactions has multiple implications for data sharing and for the design of digital libraries. One implication is that data that are important to the conduct of research often are not viewed by the researchers themselves as sufficiently valuable to keep. Thus those data are invisible. They may never be available for capture in digital libraries, much less for sharing and reuse.

Another implication is that data in the foreground of a research problem are most likely to be captured, described, and cited. The same data, when viewed as background to a research problem, may not be mentioned in a research report, nor cited explicitly. Yet these data, such as the density and distribution of phenomena, are essential to the research process and expensive to create and maintain. Since scientists vary between disciplines and even between projects on their understanding of "use," measuring the value of a data repository based on citations may grossly underestimate the collection. For example, our findings suggest that researchers may cite foreground but not background uses, even though both uses are essential to their research process.

Our findings also reinforce concerns from the social studies of science that making data "mobile" may remove so much of their context as to make them useless [8][10][19]. Data sharing efforts, data management plans, and data citation practices all require a more nuanced understanding of data "use" to be effective.

7. ACKNOWLEDGEMENTS

Research reported here is supported in part by grants from the National Science Foundation (NSF) and the Alfred P. Sloan Foundation (Sloan): (1) The Center for Embedded Networked Sensing (CENS) is funded by NSF Cooperative Agreement #CCR-0120778; (2) Towards a Virtual Organization for Data Cyberinfrastructure, NSF #OCI-0750529; (3) Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures: NSF #0827322; (4) The Data Conservancy, NSF Cooperative Agreement (DataNet) award OCI0830976; and (5)The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective, Sloan Award # 20113194.

8. REFERENCES

[1] Borgman, C., Wallis, J.C. and Enyedy, N. 2006. Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology. *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries* (Alicante, Spain, Sep. 2006), 170–183.

[2] Borgman, C.L. ed. 1990. *Scholarly Communication and Bibliometrics*. Sage.

[3] Borgman, C.L., Wallis, J.C. and Enyedy, N. 2007. Little Science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*. 7, 1-2 (2007), 17–30.

[4] Case, D.O. 2006. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*. Academic Press.

[5] Furner, J. 2003. Little book, big book: Before and after Little Science, Big Science: A review article, Part I. *Journal of Librarianship and Information Science*. 35, 2 (2003), 115–125.

[6] Latour, B. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press.

[7] Latour, B. and Woolgar, S. 1979. *Laboratory life: The Social Construction of Scientific Facts*. Sage Publications.

[8] Lynch, M. and Woolgar, S. eds. 1988. *Representation in scientific practice*. MIT Press.

[9] Maurer, B.A. 2004. Models of Scientific Inquiry and Statistical Practice: Implications for the structure of scientific knowledge. *The Nature of Scientific Evidence: Statistical, philosophical, and empirical considerations*. The University of Chicago Press. 17–50.

[10] Paisley, W.J. 1968. Information needs and uses. *Annual Review of Information Science and Technology*. 3, (1968), 1–30.

[11] Reid, R. and Traweek, S. eds. 2000. *Cultural Studies of Science, Technology, and Medicine*. Routledge.

[12] Savolainen, R. 2009. Epistemic work and knowing in practice as conceptualizations of information use. *Information Research: An International Electronic Journal*. 14, 1 (Mar. 2009).

[13] Smith, L.C. 1981. Citation analysis. *Library Trends*. 30, 1 (1981), 83–106.

[14] Wallis, J.C. and Borgman, C.L. 2011. Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. *Proceedings of the Annual Meeting of the American Society for Information Science and Technology* (New Orleans, LA, 2011).

[15] Wallis, J.C., Borgman, C.L., Mayernik, M.S. and Pepe, A. 2008. Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation*. 3, 1 (2008), 114–126.

[16] Wallis, J.C., Borgman, C.L., Mayernik, M.S., Pepe, A., Ramanathan, N. and Hansen, M. 2007. Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries* (Budapest, Hungary, Sep. 2007), 380–391.

[17] Wallis, J.C., Mayernik, M.S., Borgman, C.L. and Pepe, A. 2010. Digital libraries for scientific data discovery and reuse: from vision to practical reality. *Proceedings of the 10th Annual Joint Conference on Digital Libraries* (Gold Coast, Queensland, Australia, 2010), 333–340.

[18] Wynholds, L. 2011. Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects. *International Journal of Digital Curation*. 6, 1 (Mar. 2011), 214–225.

[19] Wynholds, L., Fearon, D.S., Jr., Borgman, C.L. and Traweek, S. 2011. When use cases are not useful: Data practices, astronomy, and digital libraries. *Proceedings of the 11th Annual Joint Conference on Digital Libraries* (Ottawa, Canada, Jun. 2011), 383–386.