Awash in Stardust: Data Practices in Astronomy

Laura Wynholds
University of California, Los Angeles
wynholds@ucla.edu

David Fearon
University of California, Los Angeles
fearon@ucla.edu

Christine L Borgman
University of California, Los Angeles
borgman@gseis.ucla.edu

Sharon Traweek
University of California, Los Angeles
traweek@history.ucla.edu

ABSTRACT

One of several major research initiatives into the grand challenge of data curation, the Data Conservancy (DC), funded by the National Science Foundation‟s DataNet Initiative, is investigating data use, sharing, and preservation in multiple fields of science. Our group at the University of California, Los Angeles is conducting a deep case study of astronomy and astrophysics. DC partners at Cornell, Illinois, the National Center for Atmospheric Research, and the National Snow and Ice Data Center are studying data practices in several other science domains. The DC is a collaborative multi-sited research project that will offer new insights into data practices in an array of physical and life sciences. The mandate of the project is to "research, design, implement, deploy and sustain data curation infrastructure for cross-disciplinary discovery with an emphasis on observational data." [4].

This poster will summarize findings from the first year of UCLA‟s research on astronomers and astronomy data. Our approach to studying data practices is complementary to that of our DC project partners, most of whom are surveying a broader set of fields less deeply. The UCLA team is part of Data Conservancy information science and computer science (IS/CS) team, which will share methods and findings. Our overall goal is to compare comparative data practices and data curation requirements across a range of physical and life science fields.

Astronomy is considered to be at the forefront of data-driven science. Hanisch and Quinn, in explaining the development of the Virtual Observatory, wrote, "Astronomy faces a data avalanche. Breakthroughs in telescope, detector, and computer technology allow astronomical instruments to produce terabytes of images and catalogs…These technological developments will fundamentally change the way astronomy is done. These changes will have dramatic effects on the sociology of astronomy itself."[7].

Over the course of the last ten years, astronomy data projects have grown from terabyte scales to petabyte scales, and the data deluge has affected many more sciences, large and small. Long predicted by the science community [8], not only has *Nature*, a premier science journal, published feature sections on "big data" [2] so have *Wired* Magazine [1], and the *Economist* [5].

However, significant tensions surrounding big data projects are present in the field, as expressed by two *Nature* editors: "Astronomy is in an era of unprecedented change…more and more astronomy papers are showing evidence that familiarity with the essential „dirtiness‟ of data and models is being lost. …Worries that the centuries-old culture of astronomy is being eroded have been voiced in the community for several years, especially in cosmology where the big-science approach now dominates." [12]

Data curation of these complex digital objects presents a significant challenge facing both scientific research and scholarly record keeping institutions. Bowker and Star [14] argued that of the problems of aggregating data within an information system are reflective of the sociotechnical systems that yielded the data. Following that argument, the quest to build repositories for data becomes largely a quest to fold the practices of an established community into evolving technological solutions. Thus it is essential to study the data practices of communities whose data is to be curated. Astronomy is a rich domain in which to study data practices, and the Data Conservancy offers a diverse environment in which to compare data curation challenges across the sciences.

We approach astronomy data practices with three questions:

1. What are the data management, curation, and sharing practices of astronomers and astronomy data centers, and how have they developed?
2. Who uses what data when, with whom, and why?
3. What data are most important to curate, how, for whom, and for what purposes?

The first question focuses on what people do, how they manage data, and what counts as relevant research data to generate, use, keep, and discard. The second question addresses the social contexts, networks, and communities within which these practices occur. The third question focuses on specific aspects of data curation, such as deciding what data will be of future use to others, assigning responsibilities for organizing and describing datasets for use, identifying incentives and disincentives for individuals or groups to curate their data, and developing tools and services necessary to exploit those data.

At the core of our astronomy case study is an analysis of the large sky surveys, as these generate massive amounts of data that fuel both inquiry and the tensions outlined above. The first year of the project has been concerned with capturing a broad perspective of the empirical and theoretical research that can be accomplished

iConference 2011, February 8-11, 2011, Seattle, WA, USA

with astronomical observations, comparing data activities associated with sky surveys to other types of inquiry.

Our starting point has been the Sloan Digital Sky Survey (SDSS) [13], which began data capture in 2000 and recently completed its final data release of the SDSS-II project. This groundbreaking optical survey telescope and accompanying digital dataset provides distributed access to data for one quarter of the sky. We are studying the development, practices and challenges of data management and curation in the SDSS, as well as the project‟s impact on astronomy. Our study of subsequent sky survey projects, such as PanSTARRS [11] and LSST [10], will offer insights to the role and value of synoptic surveys in physical science research.

Our methods follow from our three research questions about data practices, social contexts, and curation requirements in these astronomy settings:

1. Examining data practices through qualitative ethnography, including in-depth interviews and site observations; and
2. Mapping the social context of projects by analyzing documents about projects and their history, and people‟s networks of professional affiliations and research activities.

Within the context of qualitative ethnographies, we are interviewing people who have worked in multiple roles in sky surveys and who use sky survey data in their own research. These interviewees include software developers, university faculty, postdocs, and other researchers using data from networked astrophysical instruments. We are comparing the range of curation requirements for managing large-scale archives and smaller collections of research data.

We are examining the extensive documentation of the SDSS project, including an archived listserv discussion group of its builders and users.

Our initial fieldwork on astronomy sites has found broad differences in curation practices and requirements between projects, data centers, academic collaborations, and domains of research. Identifying generalizable comparisons is a core challenge. We see historical and cultural changes at large and small levels, including the professionalization of data management and the role of informatics in astronomy. Adoption of computational approaches to knowledge discovery appears uneven across the astronomy community. Science-driven research has exhibited tensions with computer engineering approaches to data archives, according to some of our respondents.

We are seeing considerable variation in the use of sky surveys, from scientific inquiry to calibration of other instruments. In conjunction with a considerable variation in use, we see significant diversity in what counts as data among those studying each wavelength, and between observational and theoretical approaches. Among the interviewed theoretical astrophysicists who rely on computational modeling, some archive the results of simulations, while others retain the algorithms but discard the data generated by simulations. Data archiving practices for sky surveys appear to vary widely by wavelength, partially due to differences in data volume, format and complexity. Similarly, astronomy data use may be further divided by practices of ground-based versus space-based instruments. Data practices and data curation requirements within astronomy are far less homogeneous than they may appear from the outside. Similarly, the computation- and data-intensive methods that characterize modern astronomical research are not embraced universally.

Our poster will compare our initial results to those of our Data Conservancy partners' analyses of data practices in other science domains. We may see similar practices of data management and preservation practices among fields; however, early reports by DC partners at Illinois show "no field-wide norms" for sharing data among the researchers they interviewed, and diverse use of data repositories even within a research field. [3] Data practices appear to vary widely within disciplines in the physical and life sciences, and even more so between them.

Keywords

Data curation, data practices, astronomy, sky surveys, collaboration.

REFERENCES

[1] Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, *16*(07).Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

[2] Community cleverness required. (2008). *Nature*, *455*(7209),1. doi:10.1038/455001a

[3] Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A. 368*(1926), 4023-4038. doi:10.1098/rsta.2010.0165

[4] Data Conservancy | Home . Retrieved August 26, 2010, from http://dataconservancy.org/home

[5] Data, data everywhere. (2010). *The Economist*, *Special Report* (February 27, 2010)

[6] Data's shameful neglect. (2009). *Nature*, *461*(7261), 145. doi:10.1038/461145a

[7] Hanisch, R. J., & Quinn, P. J. (2003). The International Virtual Observatory. Retrieved from http://www.ivoa.net/pub/info/TheIVOA.pdf

[8] Hey, A. J. G., & Trefethen, A. E. (2003). The data deluge: an e-science perspective. In *Grid computing-making the global infrastructure a reality* (pp. 809–824). West Sussex, England: Wiley.

[9] Lawrence, A. (2008). Drowning in Data : VO to the rescue. In *Astronomy: Networked Astronomy and the New Media*. Cardiff, UK. Retrieved from http://arxiv.org/abs/0905.2020

[10] Large Synaptic Survey Telescope | Home . Retrieved August 26, 2010, from http://www.lsst.org/lsst

[11] Pan-Starrs - Panoramic Survey Telescope & Rapid Response System. Retrieved August 30, 2010, from http://panstarrs.ifa.hawaii.edu/public/

[12] Sage, L., & Baker, J. (2010). Growing pains. *Nature Physics*, *6*(4), 233. doi:10.1038/nphys1633

[13] Sloan Digital Sky Survey. Retrieved August 20, 2010 from http://www.sdss.org

[14] Star, S. L., & Bowker, G. C. (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge, Mass.: MIT Press.

[15] 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), 5-16. doi:10.1007/s00799-007-0015-8