Curators to the Stars

David S. Fearon Jr.
Information Studies
University of California
Los Angeles
dfearon@ucla.edu

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

Sharon Traweek
Women's Studies
University of California
Los Angeles
traweek@history.ucla.edu

Laura Wynholds
Information Studies
University of California
Los Angeles
lawynholds@gmail.com

POSTER ABSTRACT

We are introducing to the ASIS&T community what will be, to date, the most extensive study of data practices for astronomy and astrophysics from the Information Science field. We approach astronomy data curation 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 is about 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 tasks 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. The poster will summarize findings from our first year of research.

Our research team, based at the University of California Los Angeles' Center for Embedded Networked Sensing is part of a five-year project, the Data Conservancy (DC), funded by the National Science Foundation's DataNet Initiative (Data Conservancy, 2010). DC's partner institutions are investigating data use, sharing, and preservation in multiple fields of science. UCLA 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 research project that will offer new insights into data practices in an array of physical and life sciences. Research will be translated into practice via the design of a data repository. The poster will summarize findings from the first year of research on astronomers and astronomy data.

What can the information sciences learn from astronomy about data curation? Astronomy is a pioneer in big science projects with large-scale digital datasets. Telescopes on the ground and in orbit can stream data instantly and internationally. Large instruments can stream data directly into institutional data centers; those data may be available immediately or in periodic data releases some months later. Collaborative use of instruments has fostered the ongoing development of standards for data formats and interoperability. New technologies brought changes to the profession and to research practices accompanying data production, analysis, sharing, and preservation. The investments in shared research instruments and information technologies that characterize big science also support smaller-scale projects by astronomers using "virtual observatories" from their offices, making data management a more personal responsibility. Astronomy offers a rich and complex setting to study data curation practices and to identify challenges applicable to many fields of science.

At the core of our astronomy case study is an analysis of the large sky surveys, as these generate massive amounts of data that serve multiple scientific purposes. We are comparing data activities associated with sky surveys to those of point-based observations and to other types of astronomical inquiry. The first year of the project is concerned with capturing a broad view of empirical and theoretical research that can be accomplished with astronomical observations. Our starting point is the Sloan Digital Sky Survey (SDSS) (Sloan Digital Sky Survey, 2010), which recently completed its final data release of the SDSS-II project. This pioneering optical telescope and accompanying digital dataset, online since 2000, provides distributed access to data for one quarter of the sky, surveyed across eight years. The Johns Hopkins University libraries will curate these data as part of the Data Conservancy project. We are studying the development of the SDSS, its practices of data management and curation, hurdles overcome and remaining, and its impact on astronomy. Our study of subsequent sky survey projects (PAN-STARRS, 2009; Large Synoptic Sky Telescope, 2010) will offer insights to the role and value of synoptic surveys in physical science research.

To better understand the social contexts of these projects, we are interviewing people who have worked in multiple roles in sky surveys and who use sky survey data in their own research. These people include software developers, university faculty, postdocs, and other researchers using data from networked astrophysical instruments. We are examining practices and curation issues of data centers that

support virtual astronomy projects, and are investigating the work of the International Virtual Observatory Alliance (IVOA) (Hanisch & Quinn, 2002) to build standards and tools for interoperable data archives and instruments. We are comparing the range of curation requirements for managing large-scale archives and smaller collections of research data.

Our methods follow from our three research questions about data practices, social contexts, and curation requirements in these astronomy settings: 1) We examine data practices through qualitative ethnography, including in-depth interviews and site observations; and 2) we map the social context of projects by analyzing documents about projects and their history, and people's networks of professional affiliations and research activities.

To support this multi-modal research, we are examining the extensive documentation of the SDSS program, including an archived listserv discussion group of its builders and users. We are also building a research database to help integrate our qualitative analysis of interviews and project documentation (Wynholds, Borgman, Traweek, Fearon Jr & Fidler, 2010). 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 in less detail. The UCLA DC subgroup participates within the DC information science and computer science team, collaboratively sharing methods, findings, and enabling a comparative analysis of practices across a range of physical and life science fields.

Our initial fieldwork at astronomy sites has found broad differences in curation practices and requirements between data centers and smaller university faculty groups, and also significant diversity among data centers and among areas of astronomy research. Identifying generalizable comparisons is an ongoing challenge. We see historical and cultural changes at large and small levels, including the professionalization of data management and informatics roles in astronomy.

We see significant diversity in what counts as data among those studying each wavelength, and between observational and theoretical approaches. For example, among the astrophysicists we have interviewed who primarily use computational modeling, some must archive the results of supercomputer runs, while others retain algorithms but discard data generated by simulations. Data archiving practices for sky surveys appear to vary widely by wavelength, due to differences in data volume, format and complexity. Similarly, astronomy data use may be subdivided by practices of ground-based and space-based instruments.

Adoption of computational approaches to knowledge discovery is uneven across the astronomy community. Requirements for rich science-driven investigation may conflict with engineering approaches to data archives, according to some of our respondents. We also are seeing considerable variation in the use of sky surveys, from scientific inquiry to calibration of other instruments.

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 (Cragin, Palmer, Carlson & Witt, 2010) show "no field-wide norms" for sharing data among the researchers they interviewed, and diverse use of data repositories even within a research field. Data practices appear to vary widely among disciplines within the physical and life sciences, and even more so between them. Data integration in astronomy, for example, relies upon established constants and physical laws to calibrate instruments and to determine measurement standards. Such standards are rare for fields with diverse phenomena like biology. Also of interest to ASIS&T will be our initial efforts to translate our research findings into design features for a data repository.

Keywords

Scientific data practices, digital data curation, astronomy, information behavior, user-centered design, ethnography, science & technology studies, data repositories, collaboration.

Acknowledgements

The Data Conservancy is funded by the National Science Foundation award number 830976, Sayeed Choudhury, P.I., Johns Hopkins University.

REFERENCES

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: Mathematical, Physical and Engineering Sciences, 368(1926): 4023-4038.

The Data Conservancy. (2010). Johns Hopkins University. Retrieved from http://www.dataconservancy.org/home on 10 August 2010.

Hanisch, R. J. & Quinn, P. J. (2002). The International Virtual Observatory. Retrieved from http://www.ivoa.net/ pub/info/TheIVOA.pdf on 24 August 2010.

Large Synoptic Sky Telescope. (2010). Retrieved from http://www.lsst.org/lsst on 9 August 2010.

PAN-STARRS. (2009). Panoramic Survey Telescope & Rapid Response System. Retrieved from http://pan-starrs.ifa.hawaii.edu/public/ on 14 September 2009.

Sloan Digital Sky Survey. (2010). Retrieved from http://www.sdss.org/ on 9 August 2010.

Wynholds, L., Borgman, C. L., Traweek, S., Fearon Jr, D. S. & Fidler, B. (2010). Embodying research methods into fields and tables: a process. iConference, University of Ill inois at Urbana-Champaign. Retrieved from http://www.ischools.org iConference10/2010index/ on 24 August 2010.