Title
Follow the Data: How astronomers use and reuse data (poster)

Authors
Borgman, Christine L.
Sands, Ashley
Wynholds, Laura
et al.

Publication Date
2012-10-29

Follow the Data: How Astronomers Use and Reuse Data

Ashley Sands[1], Christine L. Borgman[1], Laura A. Wynholds[1], Sharon Traweek[2]
University of California, Los Angeles, 1. Department of Information Studies, 2. Gender Studies and History

Our research assesses what new infrastructures, divisions of labor, knowledge, and expertise are necessary for the proper care of research data. How do data management, curation, sharing, and re-use practices vary among research areas? Who uses what data when, with whom, and why? These analyses will influence decisions about scientific practice and infrastructure by researchers, curators, funders, and policy makers.

Research Methods

Site: The Sloan Digital Sky Survey (SDSS)

We report here on interviews with astronomers about their use of the Sloan Digital Sky Survey (SDSS). SDSS is a multi-faceted, multi-phased, data-driven telescope project. Our research focuses on the SDSS-I and SDSS-II data collection activities which includes both photometric and spectroscopic observations. While over 400 people participated in the data SDSS-I/II data collection, the number of users who access the data is much larger: the data service (SkyServer) has over 2500 registered users and has logged hundreds of millions of anonymous queries; in May 2012 the SDSS database had 15,194,389 hits ("SDSS SkyServer Web Site Traffic"). Since the 2.5-meter telescope at Apache Point Observatory began taking data in 1998, SDSS has been a pioneer in open data projects. Following a relatively short proprietary period for data calibration, SDSS provides public data releases. The project was the first to ensure prompt public release of data; many other collaborative telescope projects now emulate the SDSS data practices including Pan-STARRS and LSST. The nature of SDSS data access, use, and curation have implications for understanding future Big Science projects.

Study Population

Our study population comprises astronomers working with SDSS data. Interviewees represent a range of career stages (students, post-docs, research scientists, and faculty), and include both builders and users of the SDSS. Interviewees were identified through bibliographic searches of papers citing SDSS. This corpus includes 14 interviews, conducted between May 2011 and February 2012, in total about 18 interview hours. Interviews lasted from 55 minutes to 1 hour 50 minutes each. Interviews were audio-recorded, transcribed, and uploaded into the data analysis software NVivo 9. The UCLA team members coded each interview. Inter-coder reliability tests ensured consistent coding practices between team members.

Interview Protocol

The *Follow the Data* interview protocol proved an effective way to identify research data sources, types of data, and uses of data. The protocol identifies a single publication authored by each interviewee and uses it as a lens, looking backward and forward, to identify data uses leading into and out of the publication. Prior to the interview session, the interviewer performs a close reading of the text and identifies authors, data sources, links to data, and other relevant aspects to discuss during the interview. This background research enables a rich interview, addressing questions identified during the close reading of the text.

Data Practices of Astronomers

Discover and Locate

Astronomy research practices are divided by many factors including the phenomena under study and the methods and tools used including wavelength, ground v. space-based, and theoretical v. observational research. Every astronomer then makes choices about their data practices beginning with the approaches used in their subset of the field. Our population used a number of both informal and formal methods to obtain data for a single published article. These included methods from formal academic literature searches to search engine searches to contacting other scholars directly. Aside from what is traditionally considered 'data', astronomers also seek other existing algorithms, codes, queries, and other tools to perform their research. These findings will be discussed in another presentation.

Retrieve and Store

Astronomers must make choices about their scientific methods based on factors outside their control. In order to use data, choices must be made on what data can be logistically retrieved and/or stored. Factors determining the feasibility include the size of the dataset and the infrastructure and resources available to the astronomer. The decisions required in these initial stages can impact the scale, method and other aspects of the scientific project, in turn affecting the outcome of the research. Unfortunately, a lack of infrastructure to handle large datasets can constrain science. Beyond being constrained by physical infrastructure, astronomers may not store data because they do not consider it important. Astronomers tend to only store what they plan to use again in the future (Wynholds 2010).

Use and Reuse

Astronomers use and reuse data for two main reasons: foreground or background research. While information professionals tend to refer to background uses of data as 'use', they are not necessarily viewed as 'use' by the astronomer. SDSS users often used public repositories for background information, while they used data from external sources for foreground purposes to drive their scientific goals (Wynholds 2012). The astronomers we spoke to explained that it is not a simple task to reuse someone else's data. A large of amount of tacit knowledge (in terms of large datasets) or area expertise (in terms of personal datasets) is necessary in order to reuse data efficiently and validly (Wynholds 2010). Astronomers sometimes choose to re-gather or re-process data themselves instead of trusting the work of others or gaining enough background knowledge to use the data.

Curate and Preserve

The level of astronomy data curation and preservation varies amongst the different subsets of fields, size of the research project, and nature of the funding agency. Curation of data is not a simple process and can be further complicated through data that is distinct in volume, format and complexity (Wynholds 2011). True curation and funding software is expensive and human expertise is lacking. Additionally, there are few incentives for faculty-track astronomers to spend their time curating and preserving data instead of producing more 'science'. Our research continues to look specifically at the way the SDSS data are being preserved and these conclusions will be presented in a future paper.

This research is funded by the U.S. National Science Foundation ("Data Conservancy" OCI0830976, S. Choudhury, PI, Johns Hopkins University, and "Knowledge & Data Transfer: the Formation of a New Workforce" # 1145888. C.L. Borgman, PI; S. Traweek, Co-PI) and the Alfred P. Sloan Foundation ("The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective" # 20113194. C.L. Borgman, PI; S. Traweek, Co-PI). UCLA Knowledge Infrastructures http://knowledgeinfrastructures.gseis.ucla.edu/

Image: SDSS. "This is the globular cluster Palomar 5, which is a cluster of stars orbiting the Milky Way at a distance of 210 thousand light years. Most of the fainter stars in the picture belong the to cluster; the brighter stars are foreground stars elsewhere in the Milky Way", http://www.sdss.org/gallery/gal_data.html