UCLA

Publications

Title

Follow the Data: How astronomers use and reuse data

Permalink

<https://escholarship.org/uc/item/6p92h3b3>

Journal

Proceedings of the American Society for Information Science and Technology, 49(1)

ISSN

1550-8390

Authors

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

Publication Date

2012-10-29

DOI

10.1002/meet.14504901341

Peer reviewed

Follow the Data: How astronomers use and reuse data

Ashley Sands

Information Studies
University of California
Los Angeles
ashleysa@ucla.edu

Christine L. Borgman

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

Laura Wynholds

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

Sharon Traweek

Gender Studies & History
University of California
Los Angeles
trawee@history.ucla.edu

ABSTRACT

We analyze the people and infrastructure involved in the building, sustaining, and curation of large astronomy sky surveys. Our research assesses what new infrastructures, divisions of labor, knowledge, and expertise are necessary for the proper care of data. Between May 2011- February 2012, we conducted fourteen interviews employing Sloan Digital Sky Survey (SDSS) data use as the focus. SDSS is a multi-faceted, multi-phased data-driven telescope project with hundreds of collaborators and thousands of users of the open data. The *Follow the Data* interview 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.

The interviews revealed the ways these astronomers *discover, locate, retrieve, and store* external data for their research. Any given astronomy research project may employ multiple methods to discover, locate, retrieve, and store multiple datasets. Our research finds that informal and formal methods are used to discover and locate data, including person-to-person contact. Data retrieval and storage methods are often determined by the size of the dataset and the amount of infrastructure available to the researcher. Astronomy research practices are evolving rapidly with access to more data and better tools. The poster presentation will report further on how those data are used and reused in astronomy.

Keywords

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

INTRODUCTION

Since 2009 the Knowledge Infrastructures team at UCLA has studied the data practices of astronomers using semi-structured interviews and participant observation fieldwork. Astronomers draw ever more heavily on repositories of astronomical observations. These resources may supplement or supplant direct observation from telescopes (Wynholds, Wallis, Borgman, Sands, & Traweek, 2012). We analyze the people and infrastructure involved in the building, sustaining, and curation of large sky surveys; these generate massive amounts of data that serve multiple scientific purposes (Fearon, Borgman, Traweek, & Wynholds, 2010).

Our research assesses what new infrastructures, divisions of labor, knowledge, and expertise are necessary for the proper care of 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.

We report here on interviews with astronomers about their use of the Sloan Digital Sky Survey (SDSS) ("Sloan Digital Sky Survey"). SDSS is a multi-faceted, multi-phased, data-driven telescope project with around 400 participants involved in the SDSS-I and SDSS-II data collection activities. 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 it began taking data in 1998, SDSS has been a pioneer in open data projects. Following a short proprietary period, SDSS provides public data releases. They were the first to ensure prompt public release of data; many other collaborative telescope projects now emulate the SDSS data practices. The nature of SDSS data access, use, and curation are important focus points for understanding future big science projects.

METHODS

Our study population includes astronomers (students, post-docs, research scientists, and faculty) working with SDSS data. Interviewees represent a range of career stages, and include both builders and users of the SDSS. Interviewees were identified through bibliographic searches of papers citing SDSS. This corpus includes 14 interviews of 13 individuals, conducted between May 2011 and February 2012, in total about 19 interview hours. Interviews lasted from 50 minutes to 2 hours 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.

The *Follow the Data* interview protocol proved an effective way to identify 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.

PRELIMINARY RESULTS

The interviews revealed how these astronomers *discover*, *locate*, *retrieve*, and *store* external data for their research. We report preliminary results here. Full results will be available by the time of the conference.

Discover and Locate

Researchers first must *discover* the existence of relevant resources and then *locate* those resources. Any given research project may draw upon multiple datasets. These datasets can include catalogs, source lists, data releases, value-added catalogs, cross-match catalogs, simulation outputs, data papers, technical papers, and data collected by the authors. These resources can be discovered via searching data release papers and the open literature, by communication with colleagues, and other means. Researchers seek not only traditional datasets. They also search for existing algorithms, codes, queries, and other tools.

Often, our interviewees locate needed datasets directly from archives, such as those run by NASA, or via public releases, such as the SDSS official website. Distinct from the SDSS and other large examples, smaller datasets can be difficult to locate even once their existence is discovered.

Informal data requests and transfers are commonly used to locate desired data. Our interviewees explained their methods for locating these data, including browsing personal websites and contacting colleagues directly. Despite the age of big data and ease of Internet access, astronomers may engage in a large amount of informal

networks of communication prior to discovering and locating data. Our interviewees employ both formal and informal ways of discovering and locating multiple kinds of data—all of which may be used to inform a single scientific research article.

Retrieve and Store

Once located, researchers must *retrieve* the relevant data and then choose a location to *store* it for use. Interviewees use a number of methods for retrieving and storing data. Once discovered and located, the user interface and size of the required dataset determine the way the scientist retrieves the data. While SDSS has effective data retrieval tools, smaller datasets often have limited user interfaces.

While some datasets may be small enough to download to a local laptop over a wireless connection, the SDSS data is about 130 TB in total size. Researchers must determine what query to use to identify the desired dataset, how much data they will be able to retrieve, as well as how and where the data will be stored. These storage considerations need to be weighed against the desire to work with the data quickly. Some university departments have chosen to download large sets of the SDSS data and to keep them locally on the university server cluster. Others are unable to provide that level of server space and CPU power to their researchers. Even once discovered and located, astronomers have choices to make in how to retrieve and store data.

The steps of data discovery, location, retrieval, and storage all take place before data use or reuse takes place (Wynholds, Wallis, Borgman, Sands, & Traweek, 2012). 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.

CONCLUSIONS

Large sky surveys, including the SDSS, have significantly shaped research practices in the field of astronomy. However, these large data sources have not served to homogenize information retrieval in the field. There is no single, standardized method for discovering, locating, retrieving, and storing astronomy data.

Discover and Locate

Informal methods are often used to discover and locate data. Astronomers may use generic Internet search engines to discover necessary datasets in addition to formal literature searches. While such techniques are helpful to scholars in the short term, the signal to noise ratio is problematic. Of greater concern long-term is that scattered individual datasets are not likely to remain available indefinitely. Curated repositories are a more reliable solution.

Retrieve and Store

The choices astronomers make while retrieving and storing their data have implications to their science. The size of available bandwidth and CPU space limit the astronomer's

decision of how much data to retrieve. These restricting factors also impact the number and kind of backups that can be made. Even with publically available data, the have-and-have-nots may now be determined by available infrastructure, which can affect the astronomer's ability to perform science.

Astronomy research practices are evolving rapidly with access to more data and better tools to discover, locate, retrieve, and store data. The poster presentation will report further on how those data are used and reused in astronomy.

ACKNOWLEDGMENTS

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).

REFERENCES

- Fearon, D. S., Borgman, C. L., Traweek, S., & Wynholds, L. A. (2010). Curators to the Stars (Poster). *Proceedings of the American Society for Information Science and Technology* (Vol. 47). Presented at the ASIST Annual Meeting, Pittsburgh, PA. doi:10.1002/meet.14504701425
- SDSS SkyServer Web Site Traffic. Retrieved June 10, 2012, from <http://skyserver.sdss.org/log/en/traffic/>
- Sloan Digital Sky Survey. Retrieved August 26, 2010, from <http://www.sdss.org/>
- Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., & Traweek, S. (2012). Data, data use, and scientific inquiry: Two case studies of data practices. *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, JCDL '12* (pp. 19–22). New York, NY, USA: ACM. doi:10.1145/2232817.2232822