UCLA
Publications

Title
When use cases are not useful: Data practices, astronomy, and digital libraries

Permalink
https://escholarship.org/uc/item/4tk5d7hx

Authors
Wynholds, Laura
Fearon, David
Borgman, Christine L
et al.

Publication Date
2011

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

When Use Cases Are Not Useful: Data Practices, Astronomy, and Digital Libraries

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

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

ABSTRACT
As science becomes more dependent upon digital data, the need for data curation and for data digital libraries becomes more urgent. Questions remain about what researchers consider to be their data, their criteria for selecting and trusting data, and their orientation to data challenges. This paper reports findings from the first 18 months of research on astronomy data practices from the Data Conservancy. Initial findings suggest that issues for data production, use, preservation, and sharing revolve around factors that rarely are accommodated in use cases for digital library system design including trust in data, funding structures, communication channels, and perceptions of scientific value.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, standards, user issues.*

General Terms
Management, Design, Human Factors, Standardization.

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

1. INTRODUCTION
As science becomes more dependent upon digital data, the need for data curation and for data digital libraries becomes more urgent [8]. We lack institutions and practices for access to research data that are comparable to the roles served by libraries and publishers for access to publications [2]. The endeavor to build infrastructure for observational scientific data has been fraught with unanswered questions regarding data granularity, data structures, definitions of data, and functions of data. On the other hand, digital libraries "hold the potential to move beyond merely disseminating resources toward creating environments that support the analysis required to understand them" [1].

In 2007, the National Science Foundation (NSF) established the DataNet Initiative, which in 2009 funded two major grants, (i) the Data Conservancy (DC), based at Johns Hopkins University, and (ii) DataONE, based at the University of New Mexico. The Data Conservancy, comprised of interdisciplinary teams at a dozen institutions, approaches data practices via a shared vision of science data curation, as "not an end, but rather a means to collect, organize, validate, and preserve data to address the grand research challenges that face society" [7]. The communities under study are astronomy (led by the UCLA team), life sciences (led by the University of Illinois team), and earth sciences (led by the team at the National Center for Atmospheric Research). The three teams inform the design decisions for a fourth team building data repository tools and applications.

The UCLA team approaches astronomy with three questions, (i) what are the data management, curation, and sharing practices of this community, (ii) who uses what data when, with whom, and why, and (iii) what data are most important to curate, how, and for whom? While focusing on communities around major sky surveys, we are investigating what these communities consider to be their data, their criteria for selection, their criteria for trust, the data problems they currently face, and the data problems they have resolved. We consider these elements to be crucial for efficient system design and for the creation of services for sustainable digital collections. Our initial findings suggest that issues for data production, use, preservation, and sharing revolve around the sources of data and the integration of data sources into knowledge discovery tools. We identified a broad range of concerns for data curation, access, and management. These vary widely by factors that rarely are considered in digital library system design: trust in data, funding sources, communication channels, and perception of scientific contributions.

2. LITERATURE REVIEW

2.1 Astronomy Data Practices
Over the last three decades, the volume of data produced by astronomical observatories and telescopes has grown by several orders of magnitude, from gigabyte scale in the 1990s to terabyte scale in the 2000s to petabyte scales today. In this context, the astronomy research community has developed networked databases, remote access capabilities to guide instruments – both ground- and space-based – and funding structures – both public and private to support terabyte and petabyte scale digital infrastructures [6]. Many of these data-intensive projects have pushed the technical limits of computing and of database design [8]. The investment in terabyte and petabyte scale projects has resulted in fewer small research projects and more collaborative efforts, many of which are international. However, while the large and technically challenging projects tend to dominate the limelight, astronomers also conduct smaller-scale investigations with more limited data collection and reuse of data from other projects, such as sky surveys.

NASA provides access to current and historical project data via a handful of dedicated data centers, which include the High Energy Astrophysics Science Archive Research Center (HEASARC), the Space Telescope Science Institute (STScI), and the Infrared Processing and Analysis Center (IPAC). Many ground-based observatories, especially those with public funding, also provide online access to datasets, such as the National Radio Astronomy

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL'11, June 13-17, 2011, Ottowa, Canada.
Copyright 2011 ACM 978-1-4503-0744-4/11/06...$10.00.

Observatory and the National Optical Astronomy Observatory Science Archive. With a variety of significant archives and the integration efforts of Virtual Observatory (VO) projects, astronomy is ahead of many fields by having hundreds of terabytes of data in standardized, queryable form [5].

2.2 Design Approaches for Data Infrastructure

Historically, software design processes, such as the waterfall model tend to assume that tasks are concrete and workflows are well understood. Use cases are a well-established method for generating system design requirements [4]. However, use cases are most effective in situations where the practices are relatively homogeneous and where the contexts of use can be identified and documented clearly [10]. The risks in the use case approach lie in the complexity of the task of establishing goals for system design. An aphorism of the design process is that most of the mistakes are made on the first day, by choosing the wrong problem to address. The further into the design process those problems are identified, the more expensive and difficult they are to address.

The design of repositories to support discovery with scientific data requires the expertise of computer science, information studies, and science and technology studies. The NSF DataNet solicitation envisions the broadening of infrastructure for data access and integration making possible new kinds of science. Achieving this goal, however, raises formidable questions and presents significant design challenges for data infrastructure, including the heterogeneity of data as a concept [11], the heterogeneity of data use [9], and the heterogeneity of approaches to data in scholarly publishing systems [3].

If we are to understand who uses what data, for what purposes and why, then understanding the dimensions that contribute to decisions of use and purpose is essential. Factors identified in the literature that influence use include tacit knowledge, source, intended audience, trust, reliability, validity, technological dependencies, description and documentation, as well as how they change [9,12]. Each of these aspects presents significant complexity for system design.

3. METHODS

Results reported here are based on two rounds of data practices interviews conducted between February and August 2010, plus follow-up interviews with key informants. The recorded interviews total 32 sessions ranging from 50 minutes to 2 hours and average an hour in length. Several additional half-day sessions provided feedback on our approach and preliminary findings. The 32 interviews represent 27 individuals across 9 institutions, predominantly in the western US. Questions with our interviewees included their type of research, participation in sky survey projects, data challenges, conceptions of data, data sources, data analysis tools, walk-throughs, end of project curation, and funding structures for data. We built the initial contact list from a bibliographic search of researchers in the western US who had worked with the SDSS dataset, and expanded our sample based on recommendations from key informants to establish a mix of subjects at different career stages and types of astronomy. Interview subjects were selected from major research projects and from small, independent projects.

While the initial set of interviews is limited in scope, we believe the richness of the subject pool yielded sufficient breadth for contextualizing the research domain. The participants represented a spectrum of astronomy domains from radio astronomy, star formation studies of microwaves, planet detection in infrared, visible spectrum sky surveys of supernovae, galaxy evolution through x-rays, gamma ray bursts, and dark matter detection. Several of the astronomers work primarily with space-based instruments, and some primarily with simulation models. Overall, our sample exhibited a mix of astronomy data types, practices, and research goals.

4. FINDINGS

We encountered broad differences in data use, curation practices, and requirements between projects, data centers, academic collaborations, and domains of research. We have begun to identify ways that our studies of the data practices of astronomy can inform technology design for the Data Conservancy. The reflexive and iterative processes of research and development are expected to inform each other throughout this five-year project. Among the activities of Data Conservancy teams is to develop a collection policy that will aid in the selection of data to archive. Based on contexts described above, we identified five major factors that impact the use and usability of astronomy data.

4.1 Trust In Sources

Astronomers make significant distinctions between data products in terms of trust. Underlying the notion of trust are concerns about the adequacy of documentation, how the data were processed, and the expertise and the reputation of the researchers who produced the data. Established data projects, such as the Sloan Digital Sky Survey (SDSS) or the Hubble Space Telescope (HST), represent extensively tested, vetted, and publicly reviewed data products. These sources are seen as valid, accurate, well documented, and trustworthy. Even astronomers who did not use Sloan data spoke highly of its quality and reliability. Another astronomer commented about a NASA mission: “The other great thing about [these data] is that it was so well calibrated. They really worked on it…you could trust the calibration.”

Many of the astronomers interviewed were reluctant to use data that did not come from thoroughly tested, reliable sources. One astronomer stated she knew personally whom to trust her field, saying “the further you are from that expert, the less trust.” She drew upon her own expertise to assess whether the work “seems right.” Another astronomer explained “Even defining what is the most basic thing, like what's the flux coming from a galaxy?...there are tens or hundreds of ways to do that. And if you don't know what you're looking at, you cannot do precision work.” Several of the astronomers interviewed referred to data that had passed through the hands of other researchers as *secondary data products*: personal datasets created from trusted sources, such as NASA, SDSS, or Hubble. The reticence to trust non-vetted projects raises significant questions about archiving of secondary data products.

Even with trusted data sources, observations from astronomical instruments are handled with a constant awareness of the limitations of their evidential value. The astronomers were wary of the dangers of treating the data as given or as an aggregation of facts rather than as the output of complex observational instruments. To generate an image, for example, a series of interdependent assertions are made about the hardware and physical principles from which raw data are derived, such as the sensor temperature and its responsiveness at that temperature to different wavelengths. Astronomers were intimately familiar with the datasets and complex data processing environments they used, representing a deep investment in understanding the constraints and technical context of the data. Correspondingly, astronomers were reluctant to use datasets for which they lack prerequisite expertise. Staff astronomers at NASA archives respond to

scientific and technical queries about these data. Considerable human expertise, both technical and scientific, is required to support active use of these science data archives.

4.2 Funding and Curation Environments

The methods and degree of data preservation in astronomy vary widely, with significant differences based on funding source, project size, and type of instruments. NASA funds the preservation of space-based missions through its Science Archive Centers. Ground-based observatories preserve data, but the extent of their curation efforts for long-term access varies, with more extensive archiving by well-funded projects such as the National Radio Astronomy Observatory. According to one radio astronomer, a significant portion of historical data in his field is stuck on magnetic tapes, inaccessible to others, and dependent on the supercomputing center where it resides. Research projects of individual astronomers and small teams present a different type of data preservation challenge. Individuals or small groups may draw their data from the archives of large projects or may collect their own observations on various instruments. In either case their resulting analyses and datasets are secondary sources that may be worthy of archiving. Astronomers reported varying practices for preserving these types of data, usually storing only what they expected to use in the future. None reported having preservation resources for their work outside their group of collaborators.

4.3 The Burden of Documentation

Data products in astronomy are the result of complex data transformations requiring significant expertise and judgment on the part of the researcher. Documenting datasets can be especially challenging: “whenever you have not generated the data, then the documentation must be incredibly detailed and incredibly well thought through. Because if I produced the data, then I know exactly what's happening, but if somebody else did, it's so easy to miss essential details or misinterpret things.”

Many astronomers considered it easier to duplicate the original data manipulations than to follow the opaque footsteps of a third party. They were equally hesitant to share their own derivative data products openly, citing the amount of work required to create adequate documentation or expressing concerns that data could be misused or misinterpreted, due to an inadequate understanding of the operating constraints. Rather than consulting documentation, several astronomers reported being contacted by researchers with questions about working with their data. They also reported interacting directly with the researchers whose data they wanted to acquire and understand. An astronomer commented, “I will basically be emailing somebody and saying, ‘Hey you, you made this measurement. Can you send me the numbers from this figure so that I can plot them on my figure of my predictions and see how they match up?’"

Astronomers reported few incentives to document data adequately for sharing. Similarly, the lack of documentation and the amount of work required to retrace steps were significant disincentives to reusing others’ data.

4.4 Appraising Value

Use is a central feature for value; astronomers tend to evaluate and select data for preservation based on anticipated use. Being “useful” expresses a challenge in evaluating scientific value to a community of users. Assessing that value, however, is a non-trivial matter, both for technical and social reasons. For example, the STSci archived the Palomar Sky Survey for its value as a historical and navigational reference. Other groups digitized the Survey’s photographic plate collection. On the other hand, a prominent astronomer questioned the value of digitizing historical plates: "[W]ith modern equipment you can have much higher precision … you're better off just taking new data over short-term baseline and be done with it.” Such disputes point to astronomy’s competing interests in preserving data with immediate utility, not just an historical legacy.

For many astronomy data collections the future use and value are difficult to anticipate. As an example, one astronomer reported on his research group’s unanticipated use of the Hubble Space Telescope archive for evidence of cosmic strings. Their group searched the entire archive, building a sophisticated algorithm for searching empty spaces in images taken for other purposes to find faint images of the phenomena, as they could occur in any random direction. They did not find the strings, but the search suggested an upper limit for cosmic string mass, and contributed a new data mining method to the community.

The NASA data center managers occasionally incorporate catalogs from individual researchers, primarily those regarded by the field to be of general reference use. One of the data center managers explained: “So not only do we serve what’s using the [space telescope] archive itself, we serve these higher level science products from these particular teams” choosing “secondary datasets that will be useful to people studying those primary things.”

4.5 Interoperability and Integrating Multiple Sources

Among the astronomers we interviewed, nearly all drew observations about target objects from multiple sources of data. Their purposes included (i) finding corroborative information about target objects or phenomena from other sources, such as positions or light intensities, or modelers matching simulations to observational benchmarks for modeling, (ii) making new discoveries within existing archived collections, and (iii) making discoveries by compiling and mining data from multiple instruments and inter-connected archives. This practice adds a layer of complexity, both to the evaluation for use, but also for tracing the provenance of sources for secondary data. An astronomer studying the orbits of stars around black holes, for example, used “both ground-based optical telescopes and radio telescopes, as well as space-based telescopes.…So a lot of my research does indeed involve digesting the data that comes from all those telescopes and working with it closely.”

One of the NASA data archive managers explained integration among their archives: “this is like combining data from two very different observatories…to do that properly you have to understand a lot about the properties of each of the two telescopes and that information is within [the archive]…think of it as a virtual observatory within just [one wavelength] and then you encapsulate that and you go to the next level by combining it with the x-rays and the their optimums.” Interconnection is not merely linking database queries, but facilitating science and knowledge discovery within multiple data types and formats that correspond to multiple wavelengths and features of astronomical phenomena and to varying conditions of the instruments that capture them.

Astronomers reported that poor interoperability among archives and a steep learning curve for integrating sources presented both a challenge and a liability. As one data manager explained, “you risk having people grab data from these distributed archives and not really understand the data.” Researchers may draw incorrect inferences from merged data. “I think there actually are papers that are having that problem now.” The NASA data centers

approach this challenge by providing comprehensive documentation, as well as domain specialists who are tasked with being a resource to the community.

A major aspect of interoperability and integration of data sources relies on domain specific tools. Most of the astronomers interviewed expressed willingness to learn new analysis tools, or programming and database languages, if they see a benefit for their research; however, limited capacities, poor documentation and ease of use of such tools present disincentives. One of the astronomers involved in VO infrastructure projects commented, "there's some off the shelf tools, they're not terribly scalable, most people don't even know how to use them, not even such as they are. The learning curves are very steep, their penetration in communities is very low."

One factor influencing the astronomy community's adoption rate is a perception of a lack of significant discoveries. One astronomer explains the problem as a "vicious cycle," with too few interoperable discovery tools to facilitate those discoveries. Another astronomer commented: "Show some major new results that came about as a result of data mining, and then explain those tools in the language of astronomers, and then we will start to see a sea change." Some projects build their own tools for integrating data sources, which frequently requires collaborating with applied computer scientists and spending "at least 6 months to a year learning each other's language and culture, and eventually, maybe something will succeed. And you still have two chairs pulled up next to the same computer and you will do stuff together." This astronomer expressed the grand goal for discovery tools that are sustained over time is to work "forever and for everybody… that you don't have to redo from scratch again."

5. CONCLUSIONS

While astronomy may appear to be a "solved problem" with its many repositories and relatively high rate of funding, we have learned that their data are heterogeneous, their collections highly distributed, and many important datasets have no home at all. The challenge for the Data Conservancy is to incorporate research on data practices into the development of prototypes and operational systems with the goal of assisting scientists in their quest "to address the grand research challenges that face society." Significant concerns about what data need to be captured, maintained, and made useful in what form for whom should determine specific operational features of a data digital library for the astronomy community. Data curation has become one of the grand challenges for information science research.

Major trusted data sources such as the SDSS are central to astronomy research, especially when well calibrated, vetted, and publicly accessible. Our conversations with data center managers, however, gave a sense of the considerable resources required to take responsibility for and to make available these types of datasets, including a great deal of human expertise in stewarding the creation of documentation of the data and their source instruments. Similar expertise may need to be considered in the digital library context for other types of highly specialized scientific data. Incorporating smaller secondary, multiple-source datasets presents different kinds of challenges. A central goal for astronomy curation has been the capacity of data archives to enhance knowledge discovery by facilitating integration of multiple data sources.

In preparing secondary data products for access, repositories need to address issues of trust, documentation, interoperability, and ongoing value. Other factors associated with secondary data use may be central to acquisition and design decisions. Astronomers judge the value of secondary sources case-by-case by whether they trust the quality of the data generated in other projects, the combined sources, the adequacy of the data's metadata, and the documentation of its production. This analysis suggests that digital libraries make decisions on acquiring secondary data collections after a deeper assessment of its community of users.

Curation requires active efforts to capture, maintain, and sustain not only scientific observations, but also the associated information about the instruments that collected those observations and other scientific context. Scientific data are heterogeneous, and both interpretation and management require deep expertise in the domain. Generic repositories for scientific data do not appear to be feasible, short of reducing the commonalities to the lowest common denominator. Use cases have insufficient scope to characterize the richness of this community's data management needs.

6. ACKNOWLEDGEMENTS

The Data Conservancy is funded by NSF award number 830976, G. Sayeed Choudhury, P.I., Johns Hopkins University.

7. REFERENCES

[1] Audenaert, N. and Furuta, R. What humanists want: how scholars use source materials. *Proceedings of the 10th annual joint conference on digital libraries*, ACM (2010), 283–292.

[2] Borgman, C.L. *Scholarship in the digital age*. MIT Press, Cambridge MA, 2007.

[3] Chavan, V. and Ingwersen, P. Towards a data publishing framework for primary biodiversity data. *BMC Bioinformatics 10*, Suppl 14 (2009), S2.

[4] Cotran, L.C. and Taylor, R.N. *Applying Software Design and Requirements Engineering Techniques to System Conception*. Institute for Software Research, University of California, Irvine, 2010.

[5] Djorgovski, S.G. Virtual astronomy, information technology, and the new scientific methodology. *Proceedings of the Computer Architecture for Machine Perception, 2005*, (2005).

[6] Goodman, A. and Wong, C.G. Bringing the night sky closer: Discoveries in the data deluge. In T. Hey, S. Tansley and K. Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft, Redmond, WA, 2009, 39-44.

[7] Hanisch, R. and Choudhury, S. The Data Conservancy: Building a Sustainable System for Interdisciplinary Scientific Data Curation and Preservation. *Proceedings of the PV 2009 conference*, European Space Agency (2009).

[8] Hey, T., Tansley, S., and Tolle, K. Jim Gray on eScience: A transformed scientific method. In T. Hey, S. Tansley and K. Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft, Redmond, WA, 2009, xix-xxxiii.

[9] Hilgartner, S. and Brandt-Rauf, S.I. Data access, ownership and control: Toward empirical studies of access practices. *Knowledge 15*, (1994), 355-372.

[10] Lilly, S. Use case pitfalls: top 10 problems from real projects using use cases. *Technology of Object-Oriented Languages and Systems. TOOLS 30. Proceedings*, (2002), 174–183.

[11] Renear, A.H., Sacchi, S., and Wickett, K.M. Definitions of Dataset in the Scientific and Technical Literature. *ASIST 2010*, (2010).

[12] Wallis, J.C., Mayernik, M.S., Borgman, C.L., and Pepe, A. Digital libraries for scientific data discovery and reuse: from vision to practical reality. *Proceedings of the 10th annual joint conference on Digital libraries*, ACM (2010), 333-340.