UCLA
Publications

## Title
The durability and fragility of knowledge infrastructures: Lessons learned from astronomy

## Permalink
https://escholarship.org/uc/item/3x46256r

## Journal
Proceedings of the Association for Information Science and Technology, 53(1)

## ISSN
2373-9231

## Authors
Borgman, Christine L.
Sands, Ashley E.
Darch, Peter T.
et al.

## Publication Date
2016-12-27

## Copyright Information
This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, availalbe at https://creativecommons.org/licenses/by-nc-nd/4.0/

Peer reviewed

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

# The Durability and Fragility of Knowledge Infrastructures: Lessons Learned from Astronomy[1]

**Christine L. Borgman**
Department of Information Studies, University of California, Los Angeles, USA
christine.borgman@ucla.edu

**Peter T. Darch**
School of Information Sciences, University of Illinois at Urbana-Champaign, USA
ptdarch@illinois.edu

**Ashley E. Sands**
Department of Information Studies, University of California, Los Angeles, USA
ashleysa@ucla.edu

**Milena S. Golshan**
Department of Information Studies, University of California, Los Angeles, USA
milenagolshan@ucla.edu

## ABSTRACT

Infrastructures are not inherently durable or fragile, yet all are fragile over the long term. Durability requires care and maintenance of individual components and the links between them. Astronomy is an ideal domain in which to study knowledge infrastructures, due to its long history, transparency, and accumulation of observational data over a period of centuries. Research reported here draws upon a long-term study of scientific data practices to ask questions about the durability and fragility of infrastructures for data in astronomy. Methods include interviews, ethnography, and document analysis. As astronomy has become a digital science, the community has invested in shared instruments, data standards, digital archives, metadata and discovery services, and other relatively durable infrastructure components. Several features of data practices in astronomy contribute to the fragility of that infrastructure. These include different archiving practices between ground- and space-based missions, between sky surveys and investigator-led projects, and between observational and simulated data. Infrastructure components are tightly coupled, based on international agreements. However, the durability of these infrastructures relies on much invisible work – cataloging, metadata, and other labor conducted by information professionals. Continual investments in care and maintenance of the human and technical components of these infrastructures are necessary for sustainability.

**Keywords**

Knowledge Infrastructures, scientific data, astronomy, stewardship.

[1] This paper is dedicated to A.J. (Jack) Meadows (1934-2016), astronomer, information scientist, and pioneer in scientific communication.

*ASIST 2016, October 14-18*, 2016, Copenhagen, Denmark.

© 2016 Christine L. Borgman, Peter T. Darch, Ashley E. Sands, Milena S. Golshan. This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). To view a copy of this license, visit: https://creativecommons.org/licenses/by/4.0/

## INTRODUCTION

Infrastructures, whether for transportation, telecommunications, or scholarly work, are much more fragile than they appear. While some parts have proved durable, such as the Roman aqueducts, most of the roads connecting them have long since crumbled away. Knowledge infrastructures, which are "robust networks of people, artifacts, and institutions" for producing, exchanging, and sustaining knowledge (Edwards, 2010, p. 17), similarly have some durable components such as printed books and the libraries that have stewarded them for centuries. Yet these components are fragile, as funding for research libraries declines and as digital materials fade away (Rumsey, 2016). Infrastructures – transportation and scholarship alike – build on an installed base and are linked with conventions of practice (Star & Ruhleder, 1996). New components emerge to fill gaps, to extend capabilities, or to replace existing infrastructures altogether. Infrastructures develop and evolve, converge or diverge, or fade away when they no longer are needed, funded, or maintained.

For the purposes of this article, "durability" is persistence over time. A particular component tends to be durable when it continues to serve its intended purposes adequately, provided that sufficient resources (financial, material, and human) have been invested in its care and maintenance. "Fragility" similarly describes a component of infrastructure that is subject to failure or degradation, usually due to uncertain availability of the resources necessary to sustain it. A component that has hitherto proved durable nevertheless can become extremely fragile. Infrastructures – and the links between them – require continuous care and maintenance (Borgman, 2000; Star & Strauss, 1999). To regard a component as intrinsically durable would obscure this necessary work.

Knowledge infrastructures for science are evolving in concert with computational and storage capabilities, growth in data production, and scientific policies that require more open access to publications and data. Astronomy is an ideal domain for studying the durability and fragility of

knowledge infrastructures, due to its longevity, scale, and transparency. Knowledge of the skies has accumulated over millennia. Today's astronomy databases incorporate star catalogs constructed over the course of centuries. While astronomy is far from a unified field, its boundaries are more apparent than those of most research domains. Astronomers share telescopes, data archives, common software tools, metadata services, and other large investments in infrastructure. Community investments are coordinated internationally, as major telescope missions have partners from multiple countries. The American Astronomical Society provides the unifying function of publishing the primary English-language journals of the field.

This paper explores the durability and fragility of knowledge infrastructures for astronomy, drawing upon a large and long-term study of infrastructures in multiple scientific domains. We frame the problem and provide initial findings, then point to further explications that are under way.

## LITERATURE REVIEW AND BACKGROUND

A brief survey of knowledge infrastructures in science and astronomy sets the context for this study.

### Knowledge Infrastructures in Science

All infrastructures are fragile in the long term, as institutions, technologies, social arrangements, and individual stakeholders change over time. Infrastructures may survive due to parts that are durable, to commitments to sustain the capabilities through periods of change, or in some cases due to benign neglect. Particularly useful in thinking about fragility and durability are the eight dimensions of infrastructure identified by Star and Ruhleder (1996, p. 113): Embeddedness, transparency, reach or scope, learned as part of membership, linked with conventions of practice, embodiment of standards, built on an installed base, and becomes visible upon breakdown. Their model originated in a large study of new scientific technologies being introduced to biology. Many scholars subsequently employed these dimensions to study infrastructures (Borgman, 2000, 2015; Bowker, 2005; Edwards et al., 2013).

Infrastructures are necessary to collect, record, and use data, as these activities are embedded in scholarly practice (Bowker, 2005). Similarly, the practices necessary to interpret data and to maintain their scientific value depend on infrastructures (Ribes & Jackson, 2013). Knowledge infrastructures for science are fragile because they have many points of potential failure. Long-term investments in durable parts of infrastructure are necessary, such as in journals, scholarly societies, data archives, and shared technologies. A single point of failure, such as a network router or a central data archive, can disrupt an entire infrastructure. If well designed, however, some infrastructures can be self-healing, whether by re-routing traffic or by finding alternate paths to a solution (Borgman, 2007; Edwards et al., 2013; Van de Sompel, 2013).

Knowledge infrastructures are expensive to construct and maintain because they must support data collection, analysis, use, and access to information over the long term. Infrastructures also exist at multiple levels of scale, which can create tensions between stakeholders with short-, medium-, and long-term goals (Ribes & Finholt, 2009). The value proposition and burden of costs for scientific infrastructures are much debated (Berman & Cerf, 2013; European Commission High Level Expert Group on Scientific Data, 2010).

### Knowledge Infrastructures in Astronomy

Astronomy, in the most general sense, is the study of the universe beyond the earth's atmosphere. Most astronomers are concerned with celestial objects or physical phenomena; others are concerned with chemical or biological phenomena; and yet others with the history of the universe. These are but a few of the dimensions along which the science varies (Meadows, 1974). While astronomy is often viewed as a quintessential "big science" (Smith, 1992), it also has many "little science" features (Darch & Sands, 2015). Building new instruments, whether on the ground or in space, can take a decade or more from the proposal to "first light." Some kinds of science can be done in a few months or years, if conducted with observing time on extant instruments or with data taken from archives. Infrastructure must adapt continuously, as each new generation of telescopes may produce orders of magnitude more observations than its predecessors (Strauss, 2014).

Telescopes are among the most durable features of the knowledge infrastructures of astronomy, as their scientific usefulness may last for decades. By replacing older cameras and other observing instruments with newer technologies, the scientific life of some telescopes can be extended for many years. Yet even telescopes can be a fragile infrastructure as labor is necessary to ensure that instruments continue to function over time. Funding has to be assembled from multiple public and private sources, each of which can fluctuate over the lifetime of a telescope, instrument, or individual research project. Missions are funded in stages, thus the research and development of a telescope project might be accomplished, but not its construction or subsequent data collection stages. Astronomy relies more heavily on private philanthropy than most scientific endeavors. Until the mid-twentieth century, much of astronomy, at least in the U.S., was privately funded. Astronomers at elite universities had access to large and modern instruments supported by Rockefeller, Carnegie, and other benefactors (McCray, 2004; Williams, 2014). As NASA, the U.S. National Science Foundation, and public agencies in Europe, Japan, China, India, Australia, and elsewhere made greater investments in astronomy, access to telescope time and to data became more equitable (Munns, 2012; McCray, 2004). The number

of professional astronomers in positions at universities and other research centers around the world has grown in parallel (DeVorkin & Routly, 1999).

The astronomy community sets its overall priorities for funding through a negotiated process that is published as the decadal survey (Committee on Survey of Surveys: Lessons Learned from the Decadal Survey Process, 2015). This long negotiation process contributes to the durability of astronomy infrastructure by establishing community commitments, at least for decade-long periods. Public funding for astronomy, like most scientific fields, can be a zero-sum game. Commitments made to new missions, telescopes, instruments, data archives, and individual projects are monies not available for other science. When projects ranked highly in one decadal cycle are ranked much lower in the next, their funding may decline precipitously in favor of new endeavors. As new telescopes come online, others may be decommissioned, disrupting the research of those whose science depends on the older instruments. "Big science" projects that depend on new instruments often are in tension with the "little science" projects that can be accomplished with smaller amounts of funding and data (McCray, 2000). Private philanthropy is easier to find for instruments, buildings, or other durable parts of infrastructure than for the essential maintenance of that infrastructure.

The transition from analog to digital astronomy, which occurred from the 1960s through the 1990s, facilitated fundamental changes in scientific practice (McCray, 2004, 2014). Until the advent of modern digital photography, astronomers spent nights on the mountain, exposing glass plates one at a time. With digital capture, astronomers can specify precise timing, exposure, and data rates. A technician on site can confirm settings and adjust calibration to climate conditions, capturing a data stream for the requesting scientists to analyze later. However, some astronomers still prefer to take their own data on the mountain.

Digital capture results in discrete images that can be copied, transferred, and manipulated far more easily than analog data. Astronomy observations, whether taken on glass plates or digital devices, are measurements of the intensity of electromagnetic radiation (e.g., X-rays, visible light) as a function of position on the sky, wavelength, and time. Glass plates, like books, can survive by benign neglect if kept in adequate environmental conditions. Even that is no guarantee of durability – a flood recently threatened the extensive glass plate collection of the Harvard-Smithsonian Center for Astrophysics (Carlisle, 2016). Astronomers, and astronomy librarians, have maintained the durability of observational data by migrating them to new technologies as they appear (Grindlay et al., 2009). As astronomy data collections have grown in size and in number, continual migration has become far more complex and expensive.

## RESEARCH QUESTIONS AND METHODS

The research reported here explores knowledge infrastructures in astronomy, drawing on our studies of data practices conducted under a series of grants from the National Science Foundation and the Alfred P. Sloan Foundation since 2009. Most of our astronomy research has focused on the Sloan Digital Sky Survey (SDSS) and the Large Synoptic Survey Telescope (LSST), as explained further in Findings. We have asked questions about infrastructure within the SDSS and LSST communities, and also conducted interviews and observations in complementary areas of astronomy. The questions addressed in this paper are these:

- How has astronomy developed, deployed, and managed knowledge infrastructures for their data?
- What factors contribute to the durability and fragility of knowledge infrastructures in astronomy?

Our astronomy work builds upon comparative and longitudinal studies of data practices in scientific and engineering domains, including embedded sensor networks, biology, undersea science, medicine, and physical sciences (Borgman et al., 2015; Borgman, Darch, Sands, Wallis, & Traweek, 2014; Darch et al., 2015; Darch & Sands, 2015; Pasquetto, Sands, Darch, & Borgman, 2016).

We draw on our studies of data practices in astronomy to explore the knowledge infrastructures on which this community depends. Interviews and observations are used to gather information on how astronomers conduct their research, how they generate or acquire data, and how they manage and exploit those data in the short and long term. We also ask specific questions about infrastructure components, relationships among them, and the origin and evolution of those components. Our questions about knowledge infrastructures cut across dimensions such as scale, central or distributed data collection, and characteristics of data management.

| Sites | Interviews | People | Period |
|---|---|---|---|
| SDSS | 136 | 118 | 2009-2016 |
| LSST | 58 | 50 | 2013-2016 |
| Astronomy Infrastructure | 37 | 26 | 2009-2016 |
| Total | 231 | 194 | 2009-2016 |

**Table 1. Data sources used for research reported in this paper**

As shown in Table 1, we draw on interviews, ethnographic participant-observation, and analysis of webpages and other documents. Ethnographic work has been conducted intermittently, from one day to several weeks at a time, over a period of seven years. Interviews are recorded and professionally transcribed. For analytical coding of interview transcripts, field notes, and documents, we used

NVivo 9, a qualitative analysis software package, and analyzed for emergent themes using grounded theory (Glaser & Strauss, 1967).

## FINDINGS

The findings are organized as follows, and include literature references due to the array of public documents on which we draw. First we present short summaries of the SDSS and LSST projects as a means to explain the role of sky surveys in astronomy knowledge infrastructures. Second, we describe the components of knowledge infrastructures in astronomy that have proved the most durable, with a focus on data management. Third, we identify some of the features of astronomy research that contribute to the fragility of these infrastructures.

### Sky Surveys: Case Studies

Astronomy sky surveys are research projects to capture uniform data about a region of the sky. They long predate modern telescopes and digital data archives. Early civilizations tracked the night sky throughout the year, creating star catalogs that could be used for purposes such as navigation. Today's knowledge infrastructures in astronomy incorporate historical star catalogs (Genova, 2013).

#### *Sloan Digital Sky Survey (SDSS)*

The Sloan Digital Sky Survey, named after its largest funder, the Alfred P. Sloan Foundation, is notable for its commitment to timely data releases via a public data archive. SDSS planning began in the 1990s and survey data collection began in 2000, mapping about one-quarter of the night sky with a focus on galaxies, quasars, and stars. A 2.5-meter optical telescope at Apache Point Observatory in New Mexico was designed, built, and deployed for the collection of the SDSS survey data. Multiple instruments on the telescope collected optical and spectroscopic data. The first phase of the SDSS project (SDSS-I) ran from 2000 to 2005; SDSS-II covered 2005 to 2008. Each was funded as an independent project. SDSS-II expanded the scientific goals and broadened the participation. SDSS-III continued with largely new leadership, collaborating institutions, and scientific goals. SDSS-III collected data through summer of 2014, when SDSS-IV began (Ahn et al., 2012; Finkbeiner, 2010; Gray et al., 2005; "Sloan Digital Sky Survey: Home," 2016).

The SDSS data remain heavily used; a July 2016 search of the astronomy section of the SAO/NASA Astrophysics Data System (ADS) yields more than 10,000 papers mentioning "SDSS" in the title or abstract (ADS, 2016). The actual number of papers using SDSS data is probably much higher, given the common practice of reusing data without citing them in publications (Goodman et al., 2014; Pepe, Goodman, Muench, Crosas, & Erdmann, 2014).

SDSS data are held by the investigators for a short proprietary period to process them into a useful scientific form, and then the data are released openly to the world. Individual investigators, small projects, and educators thus have access to high quality data on which to conduct their own research, with or without external funding. The SDSS dataset serves a wider array of users and uses than anticipated by the architects of this influential sky survey. SDSS data have been reused in multiple scientific communities and have become the basis for citizen science projects such as Galaxy Zoo, which led to Zooniverse (Darch, 2011; Zooniverse, 2014). Astronomers tend to view these public engagement efforts as important investments because they help sustain taxpayer support for continued funding.

#### *Large Synoptic Survey Telescope (LSST)*

The Large Synoptic Survey Telescope (LSST) is a sky survey based on a telescope currently being built in Chile ("LSST project schedule," 2015). LSST is due to launch a decade-long phase of data collection in 2022, generating up to 15 terabytes of data nightly (LSST Science Collaboration et al., 2009). It will provide data for small teams of scientists to answer fundamental questions about multiple topics, including the solar system, near Earth objects, the Milky Way, and the evolution of the universe. LSST is also of significant interest in particle physics, as one scientific goal is to study dark energy (Ivezić et al., 2014).

The National Science Foundation is the primary funder of the LSST, with contributions for particular components from other sources. For instance, the camera is supported by the U.S. Department of Energy, while the telescope's mirrors are funded primarily by private sources. Initial discussions about LSST began in the 1990s, and by 2001 LSST was ranked as the most important ground-based facility in the decadal survey (Committee for a Decadal Survey of Astronomy and Astrophysics; National Research Council, 2010). Research and development began in 2004, and in 2014, NSF approved funding for LSST to transition to the official Construction phase. This transition is accompanied by ramping up the infrastructures for data collection, management, and accessibility. Significant aspects of this data management work are distributed across five sites in the U.S.

The ethos of openness is fundamental to LSST data management principles (Borgman et al., 2014), although subject to negotiation and restrictions. Code used to build LSST data management infrastructure will be open source and globally available; LSST datasets will be openly accessible within the U.S. and Chile. However, external access to the LSST data will be based on agreements negotiated with individual countries.

### Durability of Knowledge Infrastructures in Astronomy

Astronomy, as an international and distributed scientific endeavor, has made massive investments in knowledge infrastructures over the last several decades. Most obvious are the large telescopes and sky surveys, funded by multiple sources and countries. Here we focus on the less obvious

durable components that contribute to data production, management, and stewardship. These include investments in standards, data archives, metadata and discovery systems, and an overall infrastructure fabric. Most of these investments were made in the last few decades, since astronomy became a digital science.

*Data Standards*

Agreements on data standards, developed in the 1970s and widely adopted by the latter 1980s as part of the transition from analog to digital astronomy, underpin many of the later infrastructure developments in astronomy. The Flexible Image Transport System (FITS) is a file format that encodes essential information about the instrument, conditions of observation, wavelength, time, and sky coordinates in a standard data format. FITS adoption enabled astronomers to combine digital records of observations from multiple instruments (Hanisch et al., 2001).

*Data Archives*

Data archives in astronomy are many, varied, and scattered around the world (Committee on NASA Astronomy Science Centers, & National Research Council, 2007). More data appear to originate from space-based than ground-based missions, as discussed further below. Legacy data, such as scans of glass plates, also are becoming more widely available (Grindlay et al., 2009). FITS remains the most common format for observations in these archives. However, our interviewees explain that data standards are a necessary, but far from sufficient, condition for interpreting archived data or for merging data from multiple sources. To use archived data effectively, these scientists require information about the research questions, methods, and observational conditions under which those data were collected. In turn, the people who create and maintain data archives, including the metadata and documentation about them, are essential parts of the knowledge infrastructures for the community. Astronomers often staff help desks, usually on a rotating basis, as both technical and domain knowledge are necessary to exploit data archives for scientific purposes.

*Metadata and Discovery Systems*

Over the last several decades the astronomy community has constructed extensive infrastructures to integrate data archives, publications, and other information artifacts necessary for their science. Three such systems, two in the U.S. and one in France, were initiated between 1970 and 1995.

Databases to catalog celestial objects and other astronomical phenomena mentioned in publications were first established in the early 1970s, building upon historical practice of creating star catalogs. Objects in our galaxy are cataloged in SIMBAD (the Set of Identifications, Measurements, and Bibliography for Astronomical Data), which is based at The Centre de Données Astronomiques de Strasbourg (CDS) in France. Catalogers read new astronomy publications as they appear, creating metadata records for each mentioned celestial object that can be identified (Centre National de la Recherche Scientifique, 2012; Genova, 2013; Perret et al., 2015; "SIMBAD Astronomical Database," 2016). Now that publications are available as digital text, many objects can be identified algorithmically, but some degree of manual cataloging and verification remains essential to the integrity of each of these databases. As of this writing (July 2016), SIMBAD contains identifiers for 8.3 million unique objects that were mentioned in more than 320,000 papers. Objects outside our galaxy are cataloged in the NASA Extragalactic Database (NED), which was founded in the late 1980s (Helou, Madore, Bicay, Schmitz, & Liang, 1990; "NASA/IPAC Extragalactic Database (NED)," 2016). Solar system and planetary data are cataloged in the NASA Planetary Data System (National Aeronautics and Space Administration, Jet Propulsion Laboratory, 2014).

NASA established the Astrophysics Data System (ADS) in 1993 as a means to coordinate access to its many data systems (ADS, 2016). This was a period of rapid technological change, with the World-Wide Web launching about the same time. For some interesting reasons that we will pursue in a later paper, the Smithsonian Astrophysical Observatory / NASA Astrophysics Data System, as it is now known, became a sophisticated bibliographic system, despite its name. ADS contains records of core astronomy publications back to the 19th century, plus has extensive coverage of grey literature (Kurtz et al., 2000, 2005). ADS curates bibliographic records and links between publications, records of celestial objects, and data archives in CDS, NED, and elsewhere (Accomazzi & Dave, 2011; Borgman, 2013; Kurtz et al., 2005). Astronomers use ADS daily to find information, due to its sophisticated searching features, comprehensive coverage, and analytical tools.

Through a series of partnership agreements, ADS, SIMBAD, NED, and CDS are heavily interlinked, offering an array of tools and services for searching, visualizing, and manipulating observational data. Taken together, these four systems establish relatively clear boundaries of astronomy as a science. However, these boundaries are always in flux. For example, LSST expands the boundaries of astronomy by focusing on dark energy through its collaboration with high energy physics. Broader partnerships may disturb the ability of astronomy to maintain these infrastructures – or they may enhance them.

*Infrastructure Fabric*

Astronomers are well aware of their knowledge infrastructures and can describe articulately their strengths, weaknesses, and gaps. In the 2000 decadal survey, the National Virtual Observatory (NVO) rose to the top priority for funding in its category (Astronomy and Astrophysics Survey Committee, 2001; NVO Interim Steering Committee, 2001). The virtual observatory has several

names and incarnations. Some refer to the international collaboration known as the International Virtual Observatory Alliance that coordinates national initiatives (IVOA, 2016). Based at the Space Science Telescope Institute in Baltimore, the NVO developed a series of technologies and standards. The project was intended to provide long-term public funding for building shared infrastructure in astronomy. The U.S. NVO later became the Virtual Astronomical Observatory. In 2014, the assets of the VAO were transferred to NASA (US Virtual Astronomical Observatory, 2014).

Despite the community commitment to the National Virtual Observatory in the decadal survey of 2000, and a scientific board to oversee the initiative, controversy arose quickly. Some welcomed the critical mass of scientific and software expertise at one locale to build shared infrastructure. Others viewed the effort as overly concentrated at one site, and too far removed from the daily practices of scientific end users. The NVO efforts did not rise to a high funding priority in the 2010 decadal survey, nor did other investments in data archives desired by some of our research subjects. Funding for the core development activities of IVOA, NVO/VAO, and other national initiatives has largely disappeared. However, the U.S. VAO legacy and main infrastructure components are currently sustained by the NASA archives at the Infrared Processing and Analysis Center, the High Energy Astrophysics Science Archive Research Center, and the Space Telescope Science Institute (US Virtual Astronomical Observatory, 2014). The VAO still exists as an entity on a voluntary basis, and member institutions continue to participate in the International Virtual Observatory Alliance. However, the virtual observatory has proved to be far less durable than its proponents expected. A much fuller history and analysis of the virtual observatory as knowledge infrastructure is needed, which we defer to a later paper.

### Fragility of Astronomy Knowledge Infrastructure

While observational data is the common substrate of astronomy, the sustainability of these data varies widely by research specialty, funding sources, uses, and many other factors. Astronomy, like any academic discipline, employs a diverse array of research methods, instruments, technologies, and theories. Research specialties are inherently unstable due to changes in scientific practice, to local differences in naming, and to the ways in which individuals cross boundaries. Large projects draw their teams from disparate specialties, which can exacerbate collaborative frictions. Our explorations to date reveal three dimensions of astronomy research that contribute to the fragility of their knowledge infrastructures. This is not an exhaustive list, but rather a starting point to examine the durability and fragility of astronomy infrastructures.

#### *Ground vs. Space-Based Missions*

The most fundamental distinction in data practices encountered in our studies of astronomy infrastructure is between projects to build telescopes on the ground and those to launch them into space. Space-based missions, largely funded by NASA in the U.S., invest in long-term data archiving and access as part of the overall project. Examples include the Hubble, Chandra, and James Webb telescopes. Data from space-based missions are archived by NASA science centers for indefinite periods of time, long after the mission itself may have concluded (Committee on NASA Astronomy Science Centers, & National Research Council, 2007). Ground-based missions, largely funded by NSF and private philanthropy in the U.S., may invest substantial project resources in the design and deployment of data archives, but funding for the archive usually ends with funding for the mission. Examples include sky surveys and major instruments such as the observatories at Mauna Kea, Mt. Wilson, and Palomar, to name a few. While funding for astronomy research varies by country and region, the differences between approaches to archiving ground vs. space-based data appear to originate in scientific practice rather than in funding, per se.

Given that astronomy data remain scientifically valuable long after a mission ends, we have asked why data originating in space are privileged over those originating on the ground, especially given that ground-based telescopes long predate space missions. The usual answer is that space-based missions are more expensive (perhaps 50 times more) and thus the cost of curation is a relatively small addition to the total budget. The proportional cost difference does not explain the lack of long-term investment in valuable data from ground-based telescopes, a fact often lamented by astronomers whose work relies on those data sources.

Another factor is that space-based data may be easier to archive, as the instruments need fewer calibration adjustments. The primary calibration occurs before launch. The background sky in space is static, whereas sky and cloud cover on the ground are different each night. Ground-based telescopes require continual cleaning and calibration adjustments in response to observing conditions. New instruments can be added to telescopes on the ground to extend their scientific life. Conversely, most space-based instruments continue to collect data with the hardware resources they had on launch day. However, software can be used to modify calibration, whether on the ground or in space.

Many astronomers mentioned the different organizational cultures of NASA and NSF in response to our questions about investments in data archiving. NASA takes a long view of the data as part of the scientific mission. These are observational data, taken from one instrument, at one time, in one place, and thus cannot be reconstructed. While the same can be said about observational data from ground-based instruments, NSF generally makes smaller and shorter-term grants than does NASA and its international counterparts. With the exception of investments in large-scale facilities such as supercomputers, supercolliders, and

similar shared instruments, NSF tends to fund individual projects, and often very small projects (Heidorn, 2008), whereas NASA maintains most of its data within its own data centers. LSST may prove to be an exception among NSF-funded projects, given the emphasis on data, but data collection is still some years in the future.

*Empirical vs. Theoretical Scientific Inquiry*

The distinction between empiricists, also self-referenced as observers, and theorists is far less sharp than that between ground- and space-based missions. Observers use theory or alerts to determine what data to collect, and may generate new theories from empirical investigations. We use the term empiricists to include those who collect their own data from telescopic instruments – some of whom build their own instruments to do so – and those who use data from the archives of ground- or space-based missions. Theorists are astronomers who construct models of phenomena. They may use data from archives to launch their models. Some theorists in our studies claim not to use any data; others consider their simulated data, their input data, the models themselves, or the output of their models to be their data. Simulated data may be structured in the same ways as observational data, enabling them to be analysed by the same sets of tools. Any one individual can be both an observer and a theorist, although most astronomers tend to concentrate in one or the other areas.

Empiricists need to manage the data they use in their own research, which may or may not include data acquired from astronomy databases. Theorists often acquire data from astronomy archives, and they need ways to manage both these data and outputs of their simulations, some of which may be very large. Astronomers who build simulations appear to have infrastructure needs similar to those of modelers in fields such as climate science or turbulence. Models and their outputs are maintained for varying lengths of time, depending on how difficult they are to reproduce.

*Sky Surveys vs. Investigator-Led Inquiry*

The goal of sky surveys is to document specific characteristics of the night sky within a certain range of the electro-magnetic spectrum, using one telescope that may have multiple instruments. While conceived and led by scientific investigators, sky surveys are a different kind of science than investigator-led studies of specific phenomena or celestial objects. The latter may be short or long term, be conducted by one or many investigators, and draw on one or many data sources. Sky surveys and investigator-led inquiries are synergistic. Surveys are systematic efforts to document the night sky. They produce rich sets of observations worthy of "follow up." Later investigator-led projects pursue phenomena identified in the surveys, leading to new findings and theories. Surveys produce far more data and more events of potential scientific interest than that team's survey scientists can pursue themselves. The need for more follow-up investigation about observations in sky surveys is among the arguments often given for open access to astronomy data.

While sky surveys provide a degree of durability for observational data, those same data may become more fragile through reuse by other investigators. Astronomers often acquire data from multiple surveys and other sources to study objects or phenomena. As data are integrated to form new datasets, they support new scientific questions. However, those derived datasets often fail to be sustained. Investigators may hold them as long as they remain useful, and may discard them if too large to maintain. Few astronomy archives can accept derived data that have multiple, poorly documented, or unknown provenance.

## DISCUSSION AND CONCLUSION

Astronomy has developed, deployed, and managed knowledge infrastructures over long periods of time. Observations collected millennia ago for documenting the movement of stars and planets throughout the year originally served purposes of navigation, religion, and culture. As the science has become more digital, more data-intensive, and more collaborative, those infrastructures, divisions of labor, knowledge, and expertise have evolved. Now those early star catalogs provide continuity in studies of the universe. Modern sky surveys, such as the Sloan Digital Sky Survey and the Large Synoptic Survey Telescope, contribute durability to these knowledge infrastructures by serving as trusted collections that are heavily used by the community.

As astronomy became digital, the community established standards, tools, metadata, and discovery systems to exploit and sustain access to their data. All of these systems and services must be maintained continuously. Metadata and discovery systems such as CDS, NED, and ADS became durable parts of the field's infrastructure over a period of several decades. Particularly notable is the amount of expert human labor necessary to identify and catalog individual celestial objects and types of phenomena. This invisible work creates the links between components of these infrastructures (Borgman, 2000; Star & Strauss, 1999). The process of developing an infrastructure fabric under the rubric of the virtual observatory also reveals the fragility of the larger knowledge infrastructure of the field. Stakeholders endorsed the need for infrastructure investments, but disagreed on matters of centralization, standards, and other features. The durability of CDS, NED, ADS, and other essential components depends on periodic renewal of funding. Directors of these agencies, and the communities who rely on these resources, must explain and argue for the value of these investments on a regular basis.

Despite the durability of the data produced by sky surveys, the astronomy community does not embrace these investments unanimously. Some scientists view large infrastructure investments as monies not spent on smaller projects that produce results on shorter scientific time frames or that employ graduate students and post-doctoral

fellows on local projects. Other scientists defend investments in sky surveys and space missions on the basis that they produce observational data that can be mined for a generation. Astronomy is the rare science that has a community mechanism to negotiate these tensions, namely the decadal survey. Consensus is not the same as unanimity, and priorities shift in each decadal review cycle.

The durability and fragility of knowledge infrastructures in astronomy appears partly to be a function of how those infrastructures are deployed in different specialties, and perhaps in different countries and scientific policy regimes. Space-based missions incorporate a long-term commitment to maintaining the value of the observational data collected. Even here, the degree of commitment varies considerably, whether measured by the percentage of project funding devoted to data management, by the number of staff, or by the range of data stewardship services that are provided. Ground-based sky surveys produce archives of observational data that others can use to follow up events and phenomena, and yet do not make the same long-term commitment to maintaining the scientific usefulness of those data beyond the life of the project. From a science policy perspective, the difference in commitment to ground-based and space-based data remains curious. The commitment to maintaining observational data is higher than to simulated data or models, however. Even when observational data are maintained in durable repositories, derivations of those data that result from later reuse may not be sustainable. What is most apparent from our studies is that the degree of human labor devoted to data collection, metadata creation, data curation, data integration, and stewardship is massive and underappreciated (Sands, 2016).

Infrastructure is fragile, even for one of the most durable of sciences – astronomy. The invisible work necessary to maintain individual systems, tools, technologies, standards, and other resources – much of it done by information professionals – may only become visible upon breakdown. Thus the fundamental tenets of infrastructure apply here (Star & Ruhleder, 1996). The durability of the knowledge infrastructure for astronomy is not guaranteed. Constant vigilance remains essential.

## ACKNOWLEDGMENTS

This research was supported by grants from the National Science Foundation (#1145888, C.L. Borgman, PI; S. Traweek, Co-PI, and #0830976), and the Alfred P. Sloan Foundation (#20113194, C.L. Borgman, PI; S. Traweek, Co-PI, and #201514001, C.L. Borgman, PI). We are grateful to the many members of the astronomy community who granted us interviews, access to their laboratories and offices, and provided rare or unpublished documents. We also thank Irene V. Pasquetto and Bernadette M. Randles for commenting on earlier drafts of this paper.

## REFERENCES

Accomazzi, A., & Dave, R. (2011). Semantic Interlinking of Resources in the Virtual Observatory Era. Astronomical Society of the Pacific, 442, 415–424. Retrieved from http://arxiv.org/abs/1103.5958

ADS. (2016). The SAO/NASA Astrophysics Data System. Retrieved April 11, 2016, from http://adswww.harvard.edu/

Ahn, C. P., Alexandroff, R., Allende Prieto, C., Anderson, S. F., Anderton, T., Andrews, B. H., … Zinn, J. C. (2012). The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey. The Astrophysical Journal Supplement Series, 203(2), 21. http://doi.org/10.1088/0067-0049/203/2/21

Astronomy and Astrophysics Survey Committee. (2001). Astronomy and Astrophysics in the New Millennium. Washington, DC: National Academy of Sciences. DOI: 10.17226/9839

Berman, F., & Cerf, V. G. (2013). Who will pay for public access to research data? Science, 341(6146), 616–617. http://doi.org/10.1126/science.1241625

Borgman, C. L. (2000). From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World. Cambridge, MA: MIT Press.

Borgman, C. L. (2007). Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, MA: MIT Press.

Borgman, C. L. (2013, May). Keynote Presentation: "ADS, Astronomy and Scholarly Infrastructure." Presented at the Astrophysics Data System 20th Anniversary Symposium, Harvard-Smithsonian Center for Astrophysics, Cambridge, MA. Retrieved from http://conf.adsabs.harvard.edu/ADSXX/

Borgman, C. L. (2015). Big Data, Little Data, No Data: Scholarship in the Networked World. Cambridge, MA: The MIT Press.

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. International Journal on Digital Libraries, 16(3–4), 207–227. http://doi.org/10.1007/s00799-015-0157-z

Borgman, C. L., Darch, P. T., Sands, A. E., Wallis, J. C., & Traweek, S. (2014). The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management. In 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL) (pp. 257–266). London: IEEE Computer Society. http://doi.org/10.1109/JCDL.2014.6970177

Bowker, G. C. (2005). Memory Practices in the Sciences. Cambridge, Mass.: MIT Press.

Carlisle, C. M. (2016, March 8). Flood Threatens Photographic Plates. Sky & Telescope. Retrieved from http://www.skyandtelescope.com/astronomy-news/

Centre National de la Recherche Scientifique. (2012). Aladin Sky Atlas. Retrieved March 21, 2013, from http://aladin.u-strasbg.fr/

Committee for a Decadal Survey of Astronomy and Astrophysics; National Research Council. (2010). New Worlds, New Horizons in Astronomy and Astrophysics. Washington, D.C.: The National Academies Press. DOI: 10.17226/12951

Committee on NASA Astronomy Science Centers, & National Research Council. (2007). Portals to the Universe: The NASA Astronomy Science Centers. Washington, D.C.: National Academies Press. DOI: 10.17226/11909

Committee on Survey of Surveys: Lessons Learned from the Decadal Survey Process. (2015). The Space Science Decadal Surveys: Lessons Learned and Best Practices. Washington, D.C.: National Academies Press. DOI: 10.17226/21788

Darch, P. T. (2011, Michaelmas Term). When Scientists Meet the Public: An Investigation into Citizen Cyberscience (DPhil Dissertation). University of Oxford, Corpus Christi College.

Darch, P. T., Borgman, C. L., Traweek, S., Cummings, R. L., Wallis, J. C., & Sands, A. E. (2015). What lies beneath?: Knowledge infrastructures in the subseafloor biosphere and beyond. International Journal on Digital Libraries, 16(1), 61–77. http://doi.org/10.1007/s00799-015-0137-3

Darch, P. T., & Sands, A. E. (2015). Beyond big or little science: Understanding data lifecycles in astronomy and the deep subseafloor biosphere. In iConference 2015 Proceedings. Newport Beach, CA: iSchools. Retrieved from https://www.ideals.illinois.edu/handle/2142/73655

DeVorkin, D. H., & Routly, P. (1999). The Modern Society: Changes in Demographics. In D. H. DeVorkin (Ed.), The American Astronomical Society's First Century (pp. 122–136). Washington, DC: American Astronomical Society.

Edwards, P. N. (2010). A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Cambridge, MA: The MIT Press.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). Knowledge infrastructures: Intellectual frameworks and research challenges (p. 40). Ann Arbor, MI: University of Michigan. Retrieved from http://knowledgeinfrastructures.org

European Commission High Level Expert Group on Scientific Data. (2010). Riding the wave: How Europe can gain from the rising tide of scientific data. Retrieved from https://www.fosteropenscience.eu/

Finkbeiner, A. K. (2010). A Grand and Bold Thing: the Extraordinary New Map of the Universe Ushering in a New Era of Discovery. New York: Free Press.

Genova, F. (2013). Strasbourg Astronomical Data Center (CDS). Data Science Journal, 12, WDS56-WDS60. http://doi.org/10.2481/dsj.WDS-007

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory: Strategies for qualitative research. Chicago: Aldine Pub. Co.

Goodman, A. A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., … Slavkovic, A. (2014). Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Computational Biology, 10(4), e1003542. http://doi.org/10.1371/journal.pcbi.1003542

Gray, J., Liu, D. T., Nieto-Santisteban, M. A., Szalay, A. S., DeWitt, D. J., & Heber, G. (2005). Scientific Data Management in the Coming Decade. SIGMOD Rec., 34(4), 34–41. http://doi.org/10.1145/1107499.1107503

Grindlay, J., Tang, S., Simcoe, R., Laycock, S., Los, E., Mink, D., … Champine, G. (2009). DASCH to Measure (and preserve) the Harvard Plates: Opening the ˜100-year Time Domain Astronomy Window, Astronomical Society of the Pacific, 410, p. 101. Retrieved from http://adsabs.harvard.edu/abs/2009ASPC..410..101G

Hanisch, R. J., Farris, A., Greisen, E. W., Pence, W. D., Schlesinger, B. M., Teuben, P. J., … Warnock, A. (2001). Definition of the Flexible Image Transport System (FITS). Astronomy and Astrophysics, 376(1), 359–380. http://doi.org/10.1051/0004-6361:20010923

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends, 57(2), 280–299. http://doi.org/10.1353/lib.0.0036

Helou, G., Madore, B. F., Bicay, M. D., Schmitz, M., & Liang, J. (1990). The NASA/IPAC Extragalactic Database. In G. Fabbiano, J. S. Gallagher, & A. Renzini (Eds.), Windows on Galaxies (pp. 109–113). Springer Netherlands. Retrieved from http://doi.org/10.1007/978-94-009-0543-6_15

Ivezić, Ž., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., AlSayyad, Y., … Zhan, H. (2014). LSST: From science drivers to reference design and anticipated data products (Version 4.0). Retrieved from https://arxiv.org/abs/0805.2366

IVOA. (2016). International Virtual Observatory Alliance. Retrieved April 12, 2016, from http://www.ivoa.net/

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., & Murray, S. S. (2005). Worldwide use and impact of the NASA Astrophysics Data System digital library. Journal of the American Society for Information Science and Technology, 56(1), 36–45. http://doi.org/10.1002/asi.20095

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Murray, S. S., & Watson, J. M. (2000). The NASA

Astrophysics Data System: Overview. Astronomy and Astrophysics Supplement Series, 143(1), 19. http://doi.org/10.1051/aas:2000170

LSST project schedule. (2015). Retrieved November 8, 2015, from http://www.lsst.org/about/timeline

LSST Science Collaboration, Abell, P. A., Allison, J., Anderson, S. F., Andrew, J. R., Angel, J. R. P., … Zhan, H. (2009). LSST Science Book, Version 2.0. Retrieved from http://arxiv.org/abs/0912.0201

McCray, W. P. (2000). Large Telescopes and the Moral Economy of Recent Astronomy. Social Studies of Science, 30(5), 685–711. http://doi.org/10.1177/030631200030005002

McCray, W. P. (2004). Giant Telescopes: Astronomical Ambition and the Promise of Technology. Cambridge, MA: Harvard University Press.

McCray, W. P. (2014). How Astronomers Digitized the Sky. Technology and Culture, 55(4), 908–944. http://doi.org/10.1353/tech.2014.0102

Meadows, A. J. (1974). Communication in Science. London: Butterworths.

Munns, D. P. D. (2012). A single sky: How an international community forged the science of radio astronomy. The MIT Press.

NASA/IPAC Extragalactic Database (NED). (2016). Retrieved April 11, 2016, from https://ned.ipac.caltech.edu/

National Aeronautics and Space Administration, Jet Propulsion Laboratory. (2014). NASA Planetary Data System. Retrieved March 22, 2014, from http://pds.jpl.nasa.gov/

NVO Interim Steering Committee. (2001). Toward a National Virtual Observatory: Science goals, technical challenges, and implementation plan. In R. J. Brunner, S. G. Djorgovski, & A. S. Szalay (Eds.), Virtual Observatories of the Future. Astronomical Society of the Pacific, 225, p. 353. Retrieved from http://arxiv.org/abs/astro-ph/0108115

Pasquetto, I. V., Sands, A. E., Darch, P. T., & Borgman, C. L. (2016). Open Data in Scientific Settings: From Policy to Practice. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 1585–1596). New York, NY, USA: ACM. http://doi.org/10.1145/2858036.2858543

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. PLoS ONE, 9(8). http://doi.org/10.1371/journal.pone.0104798

Perret, E., Boch, T., Bonnarel, F., Bot, C., Buga, M., Brouty, M., … Woelfel, F. (2015). Working Together at CDS: The Symbiosis Between Astronomers, Documentalists, and IT Specialists. Astronomical Society of the Pacific, 492, p. 13. Retrieved from http://adsabs.harvard.edu/abs/2015ASPC..492...13P

Ribes, D., & Finholt, T. A. (2009). The Long Now of Technology Infrastructure: Articulating Tensions in Development. Journal of the Association for Information Systems, 10(5). Retrieved from http://aisel.aisnet.org/jais/vol10/iss5/5

Ribes, D., & Jackson, S. J. (2013). Data Bite Man: The Work of Sustaining a Long-Term Study. In L. Gitelman (Ed.), "Raw Data" Is an Oxymoron (pp. 147–166). Cambridge, MA: The MIT Press.

Rumsey, A. S. (2016). When We Are No More: How Digital Memory Is Shaping Our Future. London and New York: Bloomsbury Press.

Sands, A. E. (2016). Managing Astronomy Research Data: Data Practices in the Sloan Digital Sky Survey and Large Synoptic Survey Telescope Projects (Dissertation). UCLA, Los Angeles, CA.

SIMBAD Astronomical Database. (2016). Retrieved April 11, 2016, from http://simbad.u-strasbg.fr/simbad/

Sloan Digital Sky Survey: Home. (2016). Retrieved April 11, 2016, from http://www.sdss.org/

Smith, R. W. (1992). The Biggest Kind of Big Science: Astronomers and the Space Telescope. In P. Galison & B. W. Hevly (Eds.), Big science: The growth of large-scale research (pp. 184–211). Stanford, Calif.: Stanford University Press.

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. Information Systems Research, 7(1), 111–134. http://doi.org/10.1287/isre.7.1.111

Star, S. L., & Strauss, A. (1999). Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. Computer Supported Cooperative Work (CSCW), 8(1–2), 9–30. http://doi.org/10.1023/A:1008651105359

Strauss, M. A. (2014). Mapping the Universe: Surveys of the Sky as Discovery Engines in Astronomy. Daedalus, 143(4), 93–102. http://doi.org/10.1162/DAED_a_00309

US Virtual Astronomical Observatory. (2014, September 22). Beyond the VAO. Retrieved April 11, 2016 from http://www.usvao.org/2014/09/22/beyond-the-vao/

Van de Sompel, H. (2013, April). From the Version of Record to a Version of the Record. Presented at the Coalition for Networked Information (CNI) Spring 2013 Membership Meeting, San Antonio, Texas. Retrieved from https://www.youtube.com/watch?v=fhrGS-QbNVA

Williams, T. (2014). The Philanthropy of Stargazing: We're In a New Golden Age of Mega Telescope Projects. Retrieved March 31, 2016, from http://www.insidephilanthropy.com/

Zooniverse. (2014). Galaxy Zoo. Retrieved May 2, 2014, from http://www.galaxyzoo.org/