Title
Data, data use, and inquiry: A new point of view on data curation

Permalink
https://escholarship.org/uc/item/24x7c6pq

Authors
Wallis, Jillian C.
Wynholds, Laura A.
Borgman, Christine L.
et al.

Publication Date
2012

License
CC BY-NC-ND 4.0

Peer reviewed
Data, Data Use, & Inquiry: A new point of view on data curation

Jillian C. Wallis, Laura A. Wynholds, Christine L. Borgman*, Ashley Sands, Sharon Traweek**
Department of Information Studies, University of California, Los Angeles, +1(310)206-0029
*Department of Information Studies, University of California, Los Angeles, +1(310)825-6164
**History and Gender Studies Departments, University of California, Los Angeles, +1(310)825-4601
jwallisi@ucla.edu, wynholds@ucla.edu, borgman@gseis.ucla.edu,

ABSTRACT
Data are proliferating far faster than they can be captured, managed, or stored. What types of data are most likely to be used and reused, by whom, and for what purposes? Answers to these questions will inform information policy and the design of digital libraries. We report findings from semi-structured interviews and field observations to investigate characteristics of data use and reuse and how those characteristics vary within and between scientific communities. The two communities studied are the researchers at the Center for Embedded Network Sensing (CENS) and users of the Sloan Digital Sky Survey (SDSS) data. We found that the interactions between inquiry, data, and use fall into three categories: foreground vs. background, use of the same data for different actions, and sources of data for reuse. The data practices of CENS and SDSS researchers have implications for data curation, system evaluation, and policy. Some data that are important to the conduct of research are not viewed as sufficiently valuable to keep. Other data of great value may not be mentioned or cited, because those data serve only as background to a given investigation. Metrics to assess the value of documents do not map well to data.

Categories and Subject Descriptors
H.3.7 [Digital Libraries]: User Issues.

General Terms
Documentation, Design, Human Factors, Standardization.

Keywords
Scientific data, data practices, data sharing, data citation.

1. INTRODUCTION
Data curation is an immediate concern of the digital libraries community and a theme of this conference. Data are proliferating far faster than they can be captured, managed, or stored – all of which are prerequisites to actual curation, which means adding value to content. Data management plans, as required by funding agencies, are concerned with keeping data for future uses. Data management plan requirements are predicated on the expectation that data will be reused. How are data used and how are they reused? What types of data are most likely to be reused, by whom, and for what purposes? Answers to these questions will inform information policy and the design of digital libraries.

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL’12, June 10-14, 2012, Washington, DC, USA.
Copyright 2010 ACM 1-58113-000-0/00/0010...$10.00.

Notions of the “use” of information range from counts of downloads to complex analyses of behavioral processes. Frequency counts of database use offer comparisons of popularity, but do not offer insights to how those downloaded objects are used once retrieved. Information needs and uses studies, which have a long tradition in the information science literature [5][17], are more concerned with how people seek documents than with the uses they make of the content therein. The science and technology studies literature is more concerned with the processes by which publications and data are created than with subsequent curation activities [12][13][15]. Research on the seeking and use of documents does not transfer directly to the seeking and use of research data. Little of the research in the social studies of science on data origins has been applied to problems of data management and curation. Much needed are contemporary studies of scientific data use that can inform information policy and the design of digital libraries for data.

Over the last decade we have been conducting interview and field studies to study scientific data practices in environmental sciences, marine biology, ecology, seismology, computer science, engineering, and astronomy [3][4][21][22][23][24][25]. Three phases of our interview studies include questions about researchers’ use of data collected by themselves, by their teams, and by other parties. We identify the characteristics of the “use” of various types of data from the perspectives of individual researchers and of teams. The same data may be used in multiple ways, depending on the research activity. “Uses” may be understood much differently by researchers than by digital library designers, librarians, and archivists. Explanations of “data use” will may inform information policy and the design of digital libraries.

This paper addresses two complementary research questions, at two research sites:

1. What are the characteristics of data use and reuse within each research community?
2. How do characteristics of data use and reuse vary within and between research communities?

2. BACKGROUND
The ubiquity of “use” as a term makes it particularly difficult to study. The dictionary definitions most applicable to the use of information, documents, or data are the verb forms “to put something into action or service,” “to consume or to take regularly,” “to carry out a purpose of action by means of,” and the noun forms “the act or practice of employing something,” or “the fact or state of being used” [27]. Documents are put into action by reading them, in lighted rooms with eyes focused on a page. That
page may be on paper or on a screen. Data also may be put into action through reading, but more often they require computers to act upon them. Data may be used to carry out purposes such as evidence of phenomena, calibration of instruments, or contextual information for the study. Actions on data tend to involve analyses and mining, often in combination from multiple sources.

Data are more complex and varied than documents, making “use” even more difficult to study. Data are amorphous, taking forms that range from physical specimens to bit streams. To the extent that data are stable objects that can be described in familiar terms, practices developed for printed documents – from personal letters to published books – may apply, at least for description. Once retrieved, the differences in “use” become even more apparent. Comparisons between uses of documents and of data offer a starting point for analyses.

2.1 Models of Information Use

Even in the study of documents, information use is difficult to capture and has been called ‘theoretically underdeveloped’, as Savolainen [19] notes: “the processes of information use largely tend to remain ‘black-boxed’ in information studies.” Information-seeking studies tend to avoid the question of how documents are used, and instead focus on search processes of users.

Use metrics such as citations and download counts are popular indicators of the use of scholarly works. Citations are at best imprecise indicators of the use of information-bearing objects [2][14]. Smith [20] was among the first to recognize that despite the assumption that citations imply use, it is difficult – if not impossible – to establish the type of use an individual citation may represent. Efforts to classify the types of use embodied in citations have a long history. Garfield [9] summarized uses of citations as: paying homage to pioneers; giving credit for related work (homage to peers); identifying methodology, equipment, etc.; providing background reading; correcting one’s own work; correcting the work of others; criticizing previous work; substantiating claims; alerting to forthcoming work; providing leads to poorly disseminated, poorly indexed, or uncited work; authenticating data and classes of fact-physical constants, etc.; identifying original publications in which an idea or concept was discussed; identifying original publications or other work describing an eponymic concept or term; disclaiming work or ideas of others (negative claims); and disputing priority claims of others (negative homage).

While the use of publications has a different structural basis than the use of data, publications are central to scientific practice, and the means by which most data are reported. The relationship between citations and data is further complicated by the difficulty of differentiating references to publications versus references publications that contain or describe data. The activities and reward mechanisms are structured by a larger ecosystem of values and rewards that goes well beyond the immediate uses of publications and data [11].

2.2 Models of Scientific Inquiry and Data

Identifying what has differentiated science from other modes of inquiry has occupied researchers for over a century. Many studies have attended to daily research practices in all kinds of sciences in order to identify what characterizes these activities as science. Some of the significant findings include the recognition of many kinds of research methods, the many ways of generating numerical measurements, and the ever-present process of consensus building and revision regarding research design, data, and analytic strategies. Each scientific community differentiates equipment, data, analytic tools, and findings that currently are regarded as stable (widely used and no longer under active debate) and the search for new kinds of questions, methods, equipment, data, and analytic strategies in each field.

More specifically: (i) In basic research state-of-the-art project design, equipment, data collection, databases, and analytic tools are always under revision. (ii) Stable equipment, data, and analytic tools are used to calibrate stable backgrounds against which new kinds of data can begin to be recognized and the long process of data evaluation can proceed. (iii) Basic research communities know how to build knowledge in the context of this mixture of stability and instability in design, equipment, data, and analysis. They develop different approaches to answering shared questions. Data curation for basic research must accommodate this on-going daily work with stability and instability, including in large-scale data sets. To do this we need to understand basic research practices with data.

Scientists pursue research questions, develop hypotheses and theories, and gather data as evidence to address those questions. The forms and types of data will vary by many factors, including the stage of inquiry, characteristics of the research domain, and how much is known about the research problem. Social and organizational structures also vary considerably between “big science” and “little science” fields, a distinction made by Derek de Solla Price [18] and much studied since [7][8]. Established forms of empirical, theoretical, and computational science are supplemented, rather than supplanted by data-driven science. Many argue that great scientific promise lies in the reuse of data.

2.3 Domain-Specific Data Practices

Little science and big science are distinguished by such factors as the maturity of the field, the consistency of research methods, the degree of shared instrumentation, and the volumes of data produced [18]. Fields also are distinguished by the types of data they collect and use, such as observational, experimental, computational, and records [26]. None of these factors are static. Choices of data, metadata, analytic tools, and specializations are constantly being revised. Many, if not most scientific fields are becoming more data-intensive with advances in instrumentation such as sensor networks. As new instruments and forms of data become available and as communities respond to new requirements for data management plans, scientific practices are in flux. The flux creates opportunities to study data production, use, and reuse.

In ecology, a field in which we are studying the evolution of practices as they deploy embedded sensor networks, science takes the form of “naturalistic realism” in which models of reality are established and then compared to the real world [10][16]. The practice of ecological science is largely inductive, beginning with accumulation of observations in the field, with the intent to discover patterns. In the search for repeatable patterns, parameter estimation often is more useful than formal hypothesis testing. When patterns are well established, deductive methods may be applied in which data are gathered to test hypotheses [16]. Ecology researchers are studying complex systems that do not lend themselves as well to consistent methods and measures as do the physical sciences [1]. However, long-term observations are essential to provide baseline comparisons for current research.

In astronomy, a field we began to study in 2009, science relies on observational and theoretical studies of celestial objects and phenomena. Much of astronomy can be described as falling with the norms of Big Science as described above. The scale of
instruments, both in terms of cost and skillsets, favor large collaborations. They use increasingly large datasets and complex models [26]. At present, dozens of airborne, ground, and underground-based observatories are generating astronomical data. Since 1972, about 200 space and solar system telescopes, observatories, and probes in this international field have yielded observational data at unprecedented scale. Observational data are produced on petabyte scales. Even after data reduction, digital libraries may hold hundreds of terabytes of astronomical observations. Astronomers share instruments and share data through a complex network of collaborations and funding arrangements. New instruments and new analytic tools are continually under development.

Astronomy data are grouped by the kinds of equipment being used (ground and spaced-based), parts of the cosmos under study (solar system, exoplanets, galaxy evolution, star formation, dark energy, etc), and the means of collecting data: radio waves, microwave & infrared, ultraviolet, X-rays, and gamma rays. Specialists cluster around these topics, but many astronomers use data from various frequencies. Astronomers tend to be familiar with many kinds of international digital libraries, how to access them, and techniques for investigating them. Many astronomers find historical data to be of value, as a result some century-old glass plates are being digitized for studies of how phenomena have changed over time [6].

3. RESEARCH METHODS
The research reported here compares findings from parallel studies in sensor networks and astronomy. Since 2002 we have studied data practices, management, and curation in the Center for Embedded Network Sensing (CENS), a multi-disciplinary, science and technology research center. Research protocols and interview questions used in the CENS data practices studies were later adapted to research on astronomers. Our astronomy focus in this paper is the Sloan Digital Sky Survey (SDSS), a large, long-term, and well-known data-driven project.

CENS research is “small science,” with small teams and emergent data collection methods, whereas the SDSS research is “big science,” with large teams and elaborate data collection methods. However, the two sites have much in common, allowing us to make a series of comparisons. CENS and SDSS are decade-long projects with multiple sources of funding, involve multiple institutions within each of these two communities, and multiple teams that have evolved over the course of their research cycles. They are obligated by their funders to share data and findings. Scientists and technologists collaborate at each site. Both require substantial infrastructure investments in technology for data collection.

The CENS and SDSS studies employed semi-structured interviews and field observations to investigate data use and reuse within these communities. Interviews ranged from 45 minutes to 2 hours, with an average of 60 minutes per interview. Participants were asked about their data and data practices. In most cases, their data were observations or output from models; in some cases data may be code, software, or computer systems.

All interviews were recorded and transcribed. Interviewees and interviews were assigned unique identifiers and names redacted upon request per Institutional Review Board procedures. Interviews were coded using NVivo software. The initial codebook was developed for the first round of CENS interviews in 2006 and significantly revised for the second round of CENS interviews in 2009. The SDSS codebook was developed from the CENS 2006 codebook, adapted to the specifics of the astronomy research questions. Findings herein are based on selected questions and relevant responses, as described in more detail below.

Figure 1: Selected interview questions from both sites

<table>
<thead>
<tr>
<th>Data Types</th>
<th>Within your work, what is typically considered to be “data”?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Sources</td>
<td>How do you distinguish between different levels or states of data? (CENS only)</td>
</tr>
<tr>
<td>Data Use</td>
<td>What are the main sources of data for your research projects? (SDSS only)</td>
</tr>
<tr>
<td>Data Use</td>
<td>Do you routinely or have you ever used data that you did not generate yourself, or from beyond the immediate project team?</td>
</tr>
<tr>
<td>Data Use</td>
<td>When you look at data, what are you hoping to find in it?</td>
</tr>
<tr>
<td>Data Use</td>
<td>When, if ever, do you reuse your datasets?</td>
</tr>
</tbody>
</table>

3.1 Research Sites

3.1.1 Center for Embedded Network Sensing
We have documented and facilitated the data practices of a distributed, collaborative, an interdisciplinary research center, the Center for Embedded Networked Sensing (CENS), from its inception in August, 2002 through its closing in July, 2012 [3][4][21][22][23][24]. CENS is a National Science Foundation (NSF) Science and Technology Center based at UCLA, with four other partner institutions in California. The mission of CENS is to develop sensing systems for scientific and social applications through collaborations between scientists, computer scientists, engineers, and experts in other domain areas. Over 300 faculty members, students, and research staff are now associated with CENS. Technology research partners in CENS include computer scientists, electrical engineers, and statisticians, and application scientists include seismologists, habitat ecologists, environmental engineers, and marine biologists. Other members of the Center come from urban planning, design and media arts, and information studies.

Interview data for this paper were collected from two rounds of data practices interviews, 43 in total, with participants from the CENS community. During the first round of interviews, in 2005-2006, 22 participants were selected using stratified random sampling, which stratified participants based on whether their research fell within the realm of technology or application science. During the second round, in 2009-2010, 21 participants were selected using stratified random sampling, but this time, they were stratified by the magnitude of their coefficient of betweenness centrality (i.e., the degree of connectedness they had with the rest of a co-authorship network constructed from CENS publications). Five people were selected by chance for both samples.

At the time of the first round of interviews in 2006, CENS was comprised of approximately 70 faculty and other researchers, about 140 student researchers, and some full-time research staff who were affiliated with the five participating universities. While the second round of interviews were conducted in 2009, the composition of the Center was very similar to that of 2006, with approximately 70 faculty, 140 students, and several full-time research staff. Interview passages provided in the results note the year in which the interview took place, the anonymized
participant identification number, and whether the interviewee is an application scientist (AS) or technology researcher (T).

3.1.2 Sloan Digital Sky Survey
Our astronomy research centers around how astronomers use the Sloan Digital Sky Survey (SDSS). 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 user base of the data is much larger. The data service (SkyServer) has over 2500 registered users and hundreds of millions of anonymous queries logged. Examining how astronomers use SDSS provides a rich view of how SDSS advanced database technology for astronomy, and advanced the sharing of data by releasing the data to the public.

Results reported here are based on two rounds of interviews with users of SDSS conducted between February 2010 and August 2011, plus open ended follow-up interviews with key informants [25]. The 32 interviews represent 27 individuals at 9 institutions. The interviews included questions about the participant’s 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.

The initial contact list was based on a bibliographic search of researchers in the western US who had worked with the SDSS data. We sought a mix of subjects at different career stages and types of astronomy, information difficult to find without personal contacts. Interview subjects were selected to span both major research projects and small, independent projects. Subjects represented multiple roles in sky surveys, including both central collaborators in large data-driven projects and users of multiple data products. These interviewees include software developers, university faculty, postdocs, senior graduate students, and other researchers using data from instrument collaborations. SDSS principals in other parts of the U.S. were key informants.

4. RESULTS
In the following section, we will describe the characteristics of data use and reuse within each research site, and how characteristics of data use and reuse vary within and between research sites. Results are organized by the two study sites, CENS and SDSS. Within each of these two sections, results first are described for data sources and types, then by type of inquiry, and lastly by uses and reuses of data within each community. Comparisons between these sites are presented in the discussion section.

4.1 Center for Embedded Networked Sensing
4.1.1 Data Sources and Types
CENS researchers collect observational, experimental, and simulation data. We have classified CENS data into four categories, which are shown in Figure 2 [4]. During sensing system deployments, networked sensors are placed in the field or in the lab to collect data on the scientific application, on the performance of the sensors themselves, or – for robotic sensor technology – proprioceptive data about the robot to facilitate navigation. The fourth category is hand-collected data for the scientific application, such as water and soil samples. Sensors can measure physical parameters, such as temperature or barometric pressure very reliably, but chemical and biological sensors are less effective. The chemical sensors tend to work only for very short time. Few biological sensors other than cameras are available. Many of the application scientists therefore collect hand-samples for the more chemical and biological variables that cannot be captured by sensors.

Each of the four data types has multiple variables; these are examples from a longer list. Some data serve only one purpose, but most serve multiple purposes as illustrated by the intersecting sets. When we asked our subjects about capturing, using, sharing, and preserving data from deployments, and about capabilities they desired in digital libraries to support their data, the primary (if not sole) interest was in the sensor-collected application science data, at the center of the figure, as these constituted the scientific data. Computer science and engineering researchers were as concerned about the quality and accessibility of these data as were the domain scientists. Conversely, the computer science and engineering researchers took little interest in maintaining access to sensor performance data or to proprioceptive datasets that are essential to their own research (these are the sets at the top right and left of the figure, respectively). These latter forms of data appear to serve transient purposes for these researchers, with minimal archival value.

Data that CENS researchers do not collect or generate come from a variety of external sources. Sources include national-scale observatories like NOAA, where they find tidal estimates, NASA’s MODIS satellite, which provides remote sensing images, and the USGS for gravitational information. Data sources also include repositories, such as Crawdad, which hosts 802.11 measures. Occasionally CENS researchers will use data from a source that is both an observatory and a repository, such as the Southern California Ocean Observing System, which collects ocean variables and also hosts data from researchers.

4.1.2 Types of Inquiry
Inquiry at CENS falls into three, sequential categories: (i) proof-of-concept development of new equipment, algorithms, and systems, (ii) refining equipment, algorithms, and systems for use by the application scientists, and (iii) scientific discovery on the part of the application scientists using CENS equipment, algorithms, and systems. The first type of inquiry is performed by the technology researchers and counts as scholarship to that community. The third type of inquiry is performed by the application science researchers and counts as scholarship to that community. The middle category is the most collaborative and sets CENS apart from other academic technology research. Technology and application science researchers work closely together in the field to take proof-of-concept technologies and transform them into research-grade or commercial-grade technologies that the application scientists can use to perform research. To the application scientists, much of this work looks like user testing, and to the technology researchers, it looks like...
product support. Although crucial to making the technology useful, few researchers are able to publish the results from this category of inquiry.

The products of CENS research are sensing systems that enable scientific researchers to observe the previously unobservable and to ask new types of questions. One of the application science students gave the following reason for why he worked at CENS: “I’m just motivated that I think it’s going to allow biologists to ask a lot of questions and answer a lot of questions that before have just been completely impossible to do. So just kind of in general being able to ask more interesting scientific questions and being able to come up with better answers.” (2006-14, Student, AS)

This participant sees the potential for significant domain change as a result of access to finer grained data provided by networked sensing systems. Not only could researchers ask new types of questions, but the quality of their work overall would benefit. Another application science faculty member (2010-11, Faculty, AS) gave a similar defense for what was new about CENS research, by supporting the integration of observation and sensor data they are increasing the quality of scientific research.

4.1.3 Data Use

The actions performed by CENS participants on data map well to models of normative science. Almost all the participants mentioned using data to identify patterns or trends, which includes identifying interesting locations and phenomena for further study. From these patterns and trends, most of the researchers would then construct hypotheses or models. When out in the field, about half the participants mentioned using data for some kind of realtime feedback that allows them to ensure that they are capturing phenomena of interest. The researchers also use these data to drive data-collection robots or to change the distribution and density of the networked sensors to improve data capture.

Many of the variables of interest to the CENS researchers could not be used in the form output by their sensors, so they would use one set of data to make another set of data usable for analysis. Activities include using calibration and “ground-truthing” metadata to adjust sensor values, feeding data into algorithms or simulations to generate new data, and using sensor values as indirect measures to calculate the data about phenomena that cannot be measured. Ultimately the data are used for analysis, from inductive correlations to deductive hypothesis or model testing. The range of analysis also includes performance evaluation of equipment, algorithms, methods, and systems, as well as meta-analysis.

Despite the diversity of the CENS community with respect to the phenomena they study, the data they collect, and the actions they apply to data, a number of commonalities exist. Types of data use in CENS include distinctions between background and foreground data, observation and experimental data, old and new data, whether the researcher was in the lab or in the field, raw versus processed data, and whether the data were collected or externally sourced. Here we explicate the data types and the actions applied to them.

Many of the researchers label a portion of their data as being “background” or “context” to their work. One participant delineates context data from “his data,” where the context data are “not formally being analyzed or used for this project in any way.” (2006-20, Faculty, AS) Several interviewees indicated that context is not considered “use.” Another participant separates the variables he collects into two buckets, “context” and his “work”:

“We see all of the parameters that I’ve mentioned before, the nutrient structure, the water column structure, all of that is context for our work. And the biological information is what we are trying to understand and predict, based on how those organisms interact with one another, and how they act with their chemical and physical environment.” (2006-02, Faculty, AS)

The participant below also distinguishes between the background and foreground of his research:

“We also have to do things like: what are the territories of the birds, what species are there, what are the territories of the birds, and stuff like that that’s background to doing all this.” (2006-09, Faculty, AS)

All three of the participants quoted above are application scientists, the first is in public health, the second a marine ecologist, and the third an evolutionary biologist. Although not in the same field, they all are studying organisms against a background of physical parameters.

The CENS data types described in the Data Types and Sources section are used in multiple ways by different domains of research at CENS. For instance, battery life data are used by sensing system researchers to assess network health. The same battery life data are used by a habitat ecologist studying leaf cover as to validate photosynthetic active radiation. Once battery life drops below a certain threshold, the data collected by a sensor are no longer considered valid. Robotics researchers who work with marine biologists use the hand-collected water samples to validate data collected by a robotic boat, whereas the marine biologists use hand samples to determine nutrient concentrations at different depths. These are also examples of foreground and background uses of data. Network health and proprioceptive data tend to be in the foreground for technical researchers and application science data are foreground for those conducting science.

Whether the data are observational, experimental, or simulation influences what can be said with them. Application scientists tend to use observational data for inductive research, experimental data for deductive research, and simulation data for predictive modeling. Technology researchers use these data types differently from the application scientists.

“But at the very least if we use the deployment data to convince people, look here is at least one real world situation in which the data looks like this and here’s how our system behaves, would behave if it were using that data or generating that data.” (2010-16, Faculty, T)

In this robotics example, observational data collected via a sensing system are used to demonstrate that the system works, whereas the experimental data are used to evaluate system functionality. Another researcher in environmental engineering reported using his observational data to verify his simulation data.

Environmental observations are irreplaceable and often accrue value with time. They can be incremented on a regular basis to create longitudinal studies, aggregate measures, and trends. In this example, the participant identifies a dichotomy of carbon flux and budget:

“You would be putting together sort of a history with old data. With new data you’re generally looking at a process that’s happening. So like carbon flux now, versus carbon budgets over the long term” (2006-21, Staff, AS)

Over the long-term, the overall carbon budget can be calculated, whereas the current measure can be used to get the carbon flux, demonstrating a difference between the use of old and new data.
The location of the research, in the field or in the lab, influences the actions applied to data. The following participant describes her use of data for real-time feedback and how her uses vary between field and lab:

“In the field we tend to, because I mean, it’s real-time data, it’s coming in rather quickly. And we’re trying to use it to determine where to take samples, where to maybe deploy more buoys and things like that. So in the field we’re looking at more spatial distribution of things, whereas when we get back to the lab we go a little bit more in depth into the temporal side of the patterns.” (2006-10, Staff, AS)

She is talking about the observational data collected in the field, but uses the same type of data for different purposes in the lab vs. the field. The typical in-field/in-lab distinction is between observational and experimental data, but some CENS researchers perform experiments in the field and collect observations in lab.

Raw data are processed to make them usable for analysis, but sometimes researchers return to them for verification:

“There are only one or two people, myself included, that actually get into the data in the database looking at the raw data. For the most part it’s, ‘... I’m noticing something weird. Can you go back and look at the actual data and see what was happening? ... Are there actually no measurements or were the measurements discarded because they were out of range.’” (2006-14, Student, AS)

The CENS community is focused on developing new equipment, systems, and software, and sometimes it difficult to tell whether outliers were sensor artifacts or legitimate phenomena. In the raw data, it may be possible to recognize a sensor fault by a characteristic pattern.

A few of the CENS participants used the same data multiple times. Known data are easy for them to reuse, because they are trusted. Participants would use known data to try new analyses or algorithms, because they know what can be found in them. CENS participants also used known data to verify recently collected data, and to test new hypotheses.

When CENS participants used data from external sources, they were drawn from national-scale observatories like the USGS or NOAA. Observatory data were useful for documenting environmental conditions, predicting the number and distribution of phenomena of interest, and identifying the expected conditions at the planned time of data collection. A couple of participants used externally sourced data for building models; a few others use parameters taken from papers or colleagues to adjust their models. Many participants used external data to verify their data and to generalize to other conditions. A special case of verification is the use of standard datasets to train, test, and benchmark algorithms and systems. Finally, a few researchers mentioned the use of remote sensing data to scale-up their findings, using their own data to parameterize remote sensing data and then viewing the phenomena at the remote sensing scale.

### 4.2 Sloan Digital Sky Survey

4.2.1 **Data Sources and Types**

Contemporary sky surveys represent the work of large collaborations. While our research focuses on SDSS, a number of interviewees used not only SDSS data, but other data products as well. In some cases the interviewees were heavily involved in building other data-driven projects such as the sources listed in Figure 3. Several of the projects were major collaborations, with up to several hundred astronomers, physicists, computer scientists, and other experts involved. In the course of their research, the researchers we interviewed provided examples of observational, experimental, and simulation data. Because the astronomers drew from such a variety of sources we report on a broader spectrum of the use of sky surveys than SDSS alone.

Figure 4 summarizes the responses from interviewees to question, “What are the main sources of data for your research projects?” Most participants reported drawing data from multiple sources; notably, many reported drawing from sources across multiple wavelengths. Of particular interest are the levels of granularity represented in descriptions of data sources. Researchers described their data sources as originating both from the physical instrument as well as the venue where the data were subsequently acquired. For example, a researcher might have acquired Hubble Space Telescope (instrument) data that had been highly processed and published in a journal article (venue). Figure 4 summarizes the types of data products mentioned in the interviews. Figure 5 summarizes the venues from which the data were acquired.

**Figure 3: Types of Data Products in Figure 4.**

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Instrument Outputs</strong></td>
<td>Ground Based Telescopes</td>
</tr>
<tr>
<td></td>
<td>Space Based Telescopes</td>
</tr>
<tr>
<td><strong>Model Outputs</strong></td>
<td>Simulation Outputs</td>
</tr>
<tr>
<td><strong>Structured Data Products</strong></td>
<td>Sky Surveys</td>
</tr>
<tr>
<td></td>
<td>Value Added Catalogs</td>
</tr>
<tr>
<td><strong>Federated Queries</strong></td>
<td>Virtual Observatory Services</td>
</tr>
</tbody>
</table>
**Figure 4: Sources of Data Named by Participants**

<table>
<thead>
<tr>
<th>Type</th>
<th>Source Named</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalog (Data) index</td>
<td>SIMBAD, VizieR</td>
<td>Observational</td>
</tr>
<tr>
<td>Curated Data Collection</td>
<td>NASA Exoplanet Database</td>
<td>Observational</td>
</tr>
<tr>
<td>Data Archive</td>
<td>Multi-mission Archive at STScI (MAST), Infrared Science Archive (IRSA)</td>
<td>Observational</td>
</tr>
<tr>
<td>Federated Data Query Services</td>
<td>Virtual Observatory Services (NVO, IVOA)</td>
<td>Observational</td>
</tr>
<tr>
<td>Ground Based Instruments</td>
<td>DEep Imaging Multi-Object Spectrograph (DEIMOS), Keck Observatories, Laser Interferometer Gravitational-Wave Observatory (LIGO)</td>
<td>Observational</td>
</tr>
<tr>
<td>Mixed</td>
<td>SDSS Value Added Catalogs</td>
<td>Observational</td>
</tr>
<tr>
<td>Physical Constants</td>
<td>NIST Atomic Spectra Database</td>
<td>Experimental</td>
</tr>
<tr>
<td>Publications Index</td>
<td>SAO/NASA Astrophysics Data System</td>
<td>Mixed</td>
</tr>
<tr>
<td>Simulation</td>
<td>Millennium Simulation Database</td>
<td>Simulation</td>
</tr>
<tr>
<td>Space Based Instruments</td>
<td>Chandra X-Ray Observatory, Fermi Large Area Telescope, Far Ultraviolet Spectroscopic Explorer (FUSE), Galaxy Evolution Explorer (GALEX), Hubble Space Telescope, Spitzer Space Telescope, Wide-field Infrared Survey Explorer (WISE), XMM X-ray Telescope</td>
<td>Observational</td>
</tr>
<tr>
<td>Space Based Sky Surveys</td>
<td>Two Micron All Sky Survey (2MASS), Infrared Astronomical Satellite Survey (IRAS)</td>
<td>Observational</td>
</tr>
</tbody>
</table>

**Figure 5: Venues from which data are acquired.**

- Publications: Journals, Conference Proceedings
- Indexes (e.g. ADS, SIMBAD, VizieR)
- Other Astronomers (unmediated, one-on-one)
- Instrument (directly, time)
- Collaborations
- Archives

As shown in 4, researchers who use SDSS data also use many other data products. The venue from which they acquire data is related to their expertise and to the role of the data in their research. For example, researchers who were expert with SDSS data would acquire data from multiple venues, but were most concerned that the upstream data processing activities were valid scientifically and documented well enough to be transparent. Researchers who were less familiar with SDSS data tended to consult with experts or to draw data from published journal articles.

### 4.2.2 Types of Inquiry

Scientists’ interactions with their data depend upon the specific activity and stage of research. Data use in astronomy, like in many disciplines, is tightly coupled with inquiry and analyses. Inquiry types identified in our interviews included hypothesis testing, open-ended inquiry, known object inquiry, and multi-faceted inquiries.

We use open-ended inquiry to refer to cases where the scientists were looking for previously unknown or undescribed phenomena.

One astronomer articulated this mode as, “let’s populate it with my data and I’ll see what’s in there” (Interview 012-02). Other scholars described more precise, known-object queries which included looking for physical properties such as white dwarf spin periods or calculating the intrinsic flux of specific galaxies.

One interviewee explained the transition from known object inquiry to open-ended inquiry in sky surveys:

“At first surveys were done to answer a particular question. And then in the olden days the quality of the data was such that there just wasn't much else to do with it other than the original purpose. But as the quality of the data started going up, suddenly it became clear that they can be used for many, many more things. And properly calibrating and documenting and that will support that then, I think. Sloan survey was probably one of the key turning points in that regard. If you look at the original Sloan survey proposals, documents, plans, whatever, it used to be spectroscopic survey of so many galaxies and its purpose was to measure large-scale structure of the universe. Then they said, "Gosh, how are we going to come up with a catalog of galaxies that we understand?" And they decided, "Oh, we will go through an imaging survey, too, in order to select targets." Now, they could have thought about this a little more and [said] you know these digitized plates now are not so bad. They will fine for our purpose. It would have been fine... But they did it, and a good thing, too. Because... all manner of science that was never envisioned in the original plans came out of it.” (012-02)

The need for researchers to demonstrate that their inquiry is original leads to tensions between creating good, long-range, well-calibrated data products and being seen as conducting cutting edge science. Academic researchers already receive few rewards for creating high quality data products. Being seen as conducting cutting edge science was an important factor for many of the researchers we interviewed, much of which revolved around whether the resulting science would be advanced enough to justify the large expense of the instruments involved. Another of the astronomers has focused on shifting the cutting edge aspects onto the tools and analysis rather than the data themselves:

"In sky surveys, it was understood a long time ago that [their] informational content is very rich and it will support science that [in the] original use of the data you didn't think about. In point of observation, that’s different right? You really want spectra of just one quasar and you just give it your all and whether or not it's good for anything else that’s... You don't care. So it really depends on the purpose of the raw image... or how do you use it is being envisioned. You just have this one question in mind that that's what you really care about. If it happens to be good for something else, wonderful, but that's just the side event.” (012-02)
We observed data being used within ‘normal science’ for the development and testing of hypotheses, theories and models. Researchers also had a number of different ways to approach original and cutting edge science based on their data, including identifying phenomena not previously observable, obtaining observations at a scale not previously possible, comparisons of data from multiple studies, and asking scientific questions that could not have been asked previously. The approaches were highly linked with concepts of use, the type of inquiry, and the definition of data. Of particular importance is the way that researchers foreground some uses in comparison with others that are less prominent. Background uses include obtaining baseline data from established sources, using data to calibrate instruments or analyses, and creating simulations and models. Often the background uses were not included in discussions of data sources used, but rather were mentioned during discussions of instrument design, data processing, or analysis.

One major activity in astronomy is to combine many layers and sources of data about a known object or set of objects. As mentioned in the Data Sources and Types section, many researchers used data from other projects and researchers, but with a careful eye on validity and quality of documentation:

“We measured our own spectrum but then we want to correlate the properties that we measured from our spectra with these data values that other people had calculated... And I’m not so comfortable with that actually but I know that we have to ... to understand some of the larger context of the objects that we’re looking at. ...Not preferable in my training, but I understand that in this context it makes sense.” (070-01)

Unlike the concerns voiced above, SDSS data were held in high regard in all interviews, with some astronomers calling it the “gold standard.” The high quality and homogeneity of the SDSS data also allows researchers to use them to leverage other older and lower quality surveys:

“But the Sloan was so uniform and very deep, you can get samples of rare objects from it. So, if you take the Sloan rare objects and you find those in some other all sky survey, you can find out what characteristics do the rare objects have in the old sky surveys. Then [you can] look for other things like that in the other three quarters... you can sort of leverage the Sloan to create, to find all these rare objects in the other three quarters of the sky.” (127-01)

In the above case, the SDSS data are essential background for the research. Another background use of SDSS data was for astrophysical models and simulations. Theoretical astrophysicists prefer to use well calibrated and vetted observational data for comparisons with their simulation outputs. They have similar concerns regarding validity, but lack the expertise in observational data processing that would allow them to evaluate the data. The theoretical astrophysicists we interviewed also tend to draw observational data from published, vetted sources:

“I’ve mostly used... data products where somebody else has taken data from the SDSS and done something with it, analyzed it, and then produced new results which then I’d take... The most common ways that I’ve used that kind of information is to actually compare data from the SDSS with predictions from my own modeling.” (017-01)

Another mode of interaction with the data is to reproduce results to extend or to verify the validity of prior research. Respondents tended to be more concerned with being transparent in their own work than reviewing others’ data. Another astronomer was concerned about the long-term preservation of data:

“over the years I worry about being able to reproduce observations. Because in the old days, everything was, whether it's an image or a spectrum, it was recorded on a photographic plate. And the plate was kept in an archive forever. But now it's all digital. And just like our digital cameras, you can press the delete button and it's gone. So, we really need to be a lot more careful about what we're doing... And so, I actually do think it is sensible to be setting up archives and taking a lot of effort to make sure that you get all the metadata too so that you could actually, A, re-reduce the data if you have access to the digital, original digital material. But at the very least [be able to] reproduce the measurement that was made so that you know exactly what the setup was.” (073-01)

Scale played several roles in data uses. Past and potential future use was an important consideration to justify the investment of documentation and ongoing support of data products. We saw a spectrum in regard to the conception of the users from small scale to large: Self—small collaboration—community—discipline—multiple disciplines—general public. Large-scale projects such as SDSS represent a much more elaborate infrastructure to accommodate use and users than had been previously attempted. Another aspect of scale is that rather than working individually or hierarchically, large collaborations around sky surveys tend to have an active body of stakeholders who participate in designing larger scale use from the ground up.

5. DISCUSSION
We have reported on our findings for two research questions, based on interviews in two complementary research communities, the Center for Embedded Networked Sensing (CENS) and users of Sloan Digital Sky Survey (SDSS) data: (1) What are the characteristics of data use and reuse within each research community? (2) How do characteristics of data use and reuse vary within and between research communities? Our goal is to identify uses and reuses from the perspective of the individual researchers and their communities, and to apply those findings to information policy and to the design of digital libraries.

Dictionary definitions of “use” focus on actions and purposes. Indeed, we find that CENS researchers and astronomers alike describe their data with respect to the purposes for which they are used. Data exist only in relation to some research question, hypothesis, model, instrument, or study. However, these researchers also act on data in ways that would be considered “use” by librarians, archivists, digital library developers, information policy makers, and other researchers, and yet are not viewed as “use” by the actors themselves.

The complex interactions between inquiry, data, and use fall into three categories: foreground vs. background, use of the same data for different actions, and sources of data for reuse.
5.1 Foreground vs. Background Uses of Data

Foreground actions are central to what researchers describe as “doing science,” and are thus privileged over other actions. What is considered foreground and what is considered background depends on the researcher’s point of view, which in turn is structured by the inquiry.

We identified three types of background actions common to CENS and SDSS: (i) Using extant data to increase the precision or accuracy with which new data is taken, (ii) grounding interpretation of results with conditions during data collection, and (iii) comparison of research data with third party sources of data for verification of results. The first type of background action includes the use of prior or externally collected data to identify phenomena or locations of interest and to help plan data collection. The second type includes contextual data that researchers rely on to structure the assertions they can make with their data, such as instrument performance records or known error rates. The third type is performed throughout the research process to corroborate findings and modes of inquiry. Figure 6 lists foreground and background actions applied to data.

Figure 6: Foreground and background actions that can be applied to data.

<table>
<thead>
<tr>
<th>Foreground</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis testing</td>
<td>Verification</td>
</tr>
<tr>
<td>Pattern identification</td>
<td>Generalization</td>
</tr>
<tr>
<td>System evaluation</td>
<td>Model building</td>
</tr>
<tr>
<td>Model testing</td>
<td>System training</td>
</tr>
<tr>
<td>Meta-analysis</td>
<td>Context</td>
</tr>
<tr>
<td>Analysis</td>
<td>Calibration</td>
</tr>
<tr>
<td>Scientific Inquiry</td>
<td>Data Processing</td>
</tr>
<tr>
<td></td>
<td>Software &amp; Code Development</td>
</tr>
<tr>
<td></td>
<td>Instrument Development</td>
</tr>
</tbody>
</table>

Data that are used for planning, context, and validation are not valued even though research could not proceed without them. At CENS background data tend to come from external sources or from collaborators, and are less central than those data they collect for themselves. Astronomers also tend to foreground the use of data from projects with which they work intimately. However, this practice is currently in flux and under active discussion. With the increased popularity of statistical approaches to data aggregation and data mining, and with the increases of well calibrated data available openly, researchers are more free to employ 4th paradigm approaches. Astronomy is also in the position of having the construction of several new instruments postponed due to funding gaps, which places increased pressure on young researchers to make use of extant data.

5.2 Same Data, Different Actions

Among both CENS researchers and SDSS data users, the same data may be put to foreground or background uses by different collaborators. Interviewees privileged foreground actions over background; background actions often were not considered “use.” The battery life example in the CENS data use section shows the battery life data being has a foreground use by the technology researcher and background use by the application scientist. Differences between SDSS use by simulation builders and observational astronomers show similar reversals of foreground and background use across the groups.

5.3 Sources of Data for Reuse

Reuse suggests using data again, whether for the same or different purposes. We identified cases in CENS, particularly, where researchers would maintain some laboratory or field data to use them again. Rarely were these data deposited publicly; rather, they were kept for reuse within in research team. Both CENS and the SDSS users drew upon public repositories of data for background information. To SDSS users, data from external sources might be foreground or background. Taking data from public repositories and sky surveys are uses in the sense of “using again,” but not in the sense of maintaining a team’s empirical data in ways that those data might be used by others in the future.

Some data that are essential to the research process are not kept at all, and thus are not available for future use by the research team or others. In CENS, these include data produced during the long processes of iterative testing and evaluation of sensor networks, which is the most collaborative part of this type of research. Among SDSS users, these may include similar iterative testing and evaluation of instruments, and calibration and merging of data from multiple sources.

6. CONCLUSIONS

“Use” is a term with little value unless qualified. Garfield's typology of the fifteen uses captured in published citations [9] demonstrates the complexity of use with regards to fixed objects. Data adds another level of complexity to the understanding of use. Data are extremely varied and rather than being fixed, they shimmer. Data are used in normative scientific modes, but these uses vary by type of inquiry, type of data, and interactions between them.

Science itself is not uniform and varies widely in approaches, data, and data use. This variety happens not at the domain level, but even between labs and individuals. CENS and SDSS research sites represent small science and big science. Both sites are long-term projects, with individuals falling on the technology and the science side of research. They encompass astronomers, computer scientists, electrical engineers, environmental engineers, habitat ecologists, marine biologists, seismologists, among others.

The complexity of these interactions has multiple implications for data sharing, for the design of digital libraries. One implication is that data that are important to the conduct of research often are not viewed as sufficiently valuable to keep. Thus those data are invisible. They may never be available for capture in digital libraries, much less for sharing and reuse.

Another implication is that data in the foreground of a research problem are most likely to be captured, described, and cited. The same data, when viewed as background to a research problem, may not be mentioned in a research report, much less cited explicitly. Yet these data, such as the density and distribution of phenomena, are essential to the research process and expensive to create and to maintain. If the value of data repository is based on citations, its value may be grossly underestimated. Our findings suggest that researchers may cite foreground but not background uses, for example.

Our findings also reinforce concerns from the social studies of science that making data "mobile," may remove so much of their context as to make them useless [15][17][25]. Data sharing efforts, data management plans, and data citation practices all require a more nuanced understanding of data "use" to be effective.

7. REFERENCES


