The conundrum of sharing research data

Christine L. Borgman, Professor & Presidential Chair
Department of Information Studies, University of California, Los Angeles
Borgman@gseis.ucla.edu

Requested citation to this version:
Borgman, Christine L. (2011, submitted). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology.*

> *We must all accept that science is data and that data are science, and thus provide for, and justify the need for the support of, much-improved data curation (Hanson, Sugden & Alberts, 2011).*

ABSTRACT

The deluge of research data has excited researchers, policy makers, and the general public. Not only might research be reproducible, but new questions can be asked, with great benefit to research, innovation, education, and the citizenry. However, very little

data is being shared, despite the best efforts of funding agencies and journals. This article explores the complexities of data, research practices, innovation, incentives, economics, intellectual property, and public policy associated with the data sharing conundrum – "an intricate and difficult problem." Research data take many forms, are collected for many purposes, via many approaches, and often are difficult to interpret once removed from their initial context. Rationales for sharing data vary along two dimensions: whether motivated by research concerns or by leveraging public investments, and whether intended to serve the interests of researchers who produce data or the interests of potential re-users of data. Four rationales for sharing research data are identified and positioned on these dimensions. Researchers' incentives to share their data depend not only on these rationales, but on characteristics of their data and research practices, funding agency policies, and resources for data management. Much more is understood about why researchers do *not* share data than about when, why, and how researchers *do* share data, or about when, how, and why researchers or the public reuse data. The model and research agenda are illustrated with examples from the sciences, social sciences, and humanities.

INTRODUCTION

The data deluge has arrived. Long predicted by the science community (Hey, A. J. G. & Trefethen, 2003), the popular press is now heralding the wide availability of data for use by anyone, anywhere. Not only have *Nature* and *Science*, the premier science journals, published feature sections on "big data" (Community cleverness required, 2008; Data's shameful neglect, 2009; Dealing with data, 2011), so have *WIRED* magazine (Anderson, 2008), and the *Economist* (Data, Data Everywhere, 2010). Universities are assessing their rights, roles, and responsibilities for managing and for exploiting data from their researchers (The University's Role in the Dissemination of Research and Scholarship, 2009; Lyon, 2007).

Grand expectations for the data-rich world include discoveries of new drugs, a better understanding of the earth's climate, and improved ability to examine history and culture. The growth of data in the "big sciences" such as astronomy, physics, and biology has led not only to new models of science – collectively known as the "Fourth Paradigm" – but also to the emergence of new fields of study such as astro-informatics, computational biology, and digital humanities (Borgman, 2009; Kowalczyk & Shankar, 2011).

If the rewards of big data are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others. Underlying this simple statement are thick layers of complexity about the nature of data, research, innovation, and scholarship, incentives and rewards, economics and intellectual property, and public policy. Sharing research data is thus a

conundrum – "an intricate and difficult problem" (Merriam-Webster's Collegiate Dictionary, 1993).

The "dirty little secret" behind the promotion of data sharing is that not much sharing may be taking place. Despite pressure from funding agencies and findings that sharing research data increases citation rates (Piwowar, Becich, Bilofsky & Crowley, 2008; Piwowar & Chapman, 2010; Piwowar, Day & Fridsma, 2007), relatively few studies document consistent data release. Data sharing activities appear to be concentrated in a few fields, and practices even within these fields are inconsistent. In nine years of studying data practices in a National Science Foundation Science and Technology Center, we have found that little research data is circulated beyond the research teams that produce them, and few requests are made for these data (Mayernik, 2011; Wallis, Mayernik, Borgman & Pepe, 2010). Data often do not exist in transferrable forms. Some data are not sharable for ethical or epistemological reasons. In many cases, it is not clear what are "the data" associated with a research project.

This article explores the complexity of data sharing, examining the roots of current discourse, the problematic notion of "data" per se, current policy arguments in favor of data sharing, differing perspectives of stakeholders, and associated ethical, professional, and epistemological aspects of research data. It is a foray into a labyrinth worthy of book-length examination.

WHY IS DATA SHARING URGENT?

Sharing research data is not a new topic of discussion in research and policy circles. Thoughtful reports on the reasons to improve data sharing and curation date at least from the 1980s (Fienberg, Martin & Straf, 1985), with many more reports to follow (Preserving Scientific Data on Our Physical Universe, 1995; Bits of Power: Issues in Global Access to Scientific Data, 1997; Long-Lived Digital Data Collections, 2005; Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age, 2009; Harnessing the power of digital data for science and society, 2009; Berman, F. et al., 2010; Dalrymple, 2003; Esanu & Uhlir, 2003; 2004; Hanson et al., 2011).

"Data sharing" has many meanings in these reports, which rarely are made explicit. For the purposes of this article, data sharing is the release of research data for use by others. Release may take many forms, from private exchange upon request to deposit in a public data collection. Posting datasets on a public website or providing them to a journal as supplementary materials also qualifies as sharing. The degree of usefulness, trustworthiness, and value of shared data varies widely, however. Some may be richly structured and curated. Others may be raw files with minimal documentation. Similarly, the intended users may vary from researchers within a narrow specialty to the general public.

Funding agencies have begun requiring data release to varying degrees, and with varying degrees of enforcement. The National Institutes of Health added a data management plan requirement in 2003 for grants over $500,000 (Long-Lived Digital Data Collections, 2005). The National Science Foundation long has had this statement requiring data sharing in its grant contracts, but has not enforced the requirement consistently (Grant Policy Manual, 2001; NSF Data Sharing Policy, 2010):

> Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.

In 2010, the NSF made the long-anticipated announcement that all future grant proposals would require a two-page Data Management Plan that addresses the above requirement and that the Plan would be subject to peer review (NSF Data Management Plans, 2010; NSF Proposal Preparation Instructions, 2011). The NSF requirement is thus more comprehensive than that of NIH, which applies only to larger grants and is negotiated between investigators and program officers rather than being subject to peer review.

U.K. funding agencies began to formulate data release policies in the latter 1990s (Wellcome Trust statement on genome data release, 1997; Wellcome Trust Policy on Access to Bioinformatics Resources by Trust-Funded Researchers, 2001; Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility, 2003; ESRC Research Data Policy, 2010; Lyon, 2007). In response, the Digital Curation Centre (DCC), founded as part of the U.K. eScience initiatives, created a series of templates for data management plans corresponding to the requirements of individual U.K. funding agencies (DCC Data Management Plans, 2011). The DCC is adding links to U.S. agency requirements, and the DCC templates are the basis for many of the planning documents now being developed by U.S. universities (Abrams, Cruse & Kunze, 2009; Witt, Carlson, Brandt & Cragin, 2009).

Similarly, selected journals long have required the deposit of data and other research documentation associated with published articles. Requirements to deposit genome sequences are best known (Summary of principles, 1996; Wellcome Trust statement on genome data release, 1997; Genome Canada Data Release and Sharing Policy, 2005; Berman, H. M. et al., 2000; Hilgartner, 1998), but journals in economics and many other fields also require access to data. The mechanisms of enforcement may be formal, for example by requiring deposit in specific collections such as the *Protein Data Bank*, with the structure entry number included in the article (Protein Data Bank, 2011); or less formal, such as links to sources.

Journal policies also have become more rigorous about data access as of late. *Science*, in an editorial accompanying a special issue on data, announced more extensive requirements, such as sharing computer code "involved in the creation or analysis of data" and including a "specific statement regarding the availability and curation of data" in the article's acknowledgements (Hanson et al., 2011). Also recently announced are new data archiving policies by "key journals in evolution and ecology" including *The American Naturalist, Evolution, the Journal of Evolutionary Biology, Molecular Ecology,* and *Heredity* (Whitlock, McPeek, Rausher, Rieseberg & Moore, 2010: 145). These journals are requiring or encouraging the deposit of data in public archives.

While none of these actions alone caused the sense of urgency for data sharing, the NSF requirement for data management plans appears to be the tipping point, at least in the U.S. The National Science Foundation has an encyclopedic scope across the sciences and social sciences, excluding only the arts, humanities, and medicine – and even funds grants in these areas for projects that address scientific problems. The NSF requirement applies to all proposals, of any size, in any directorate. While these are data *management* plans and not data *sharing* plans, they do strongly encourage sharing and they are subject to peer review. Thus an investigator's ability to articulate what her or his data are, how they will be managed, how they will be shared, and if not shared why, will influence whether or not a project is funded. In making these plans part of the peer review process, the NSF has provoked a broad conversation about data sharing among stakeholders in publicly funded research.

What is often not explicit in the discussions of data management plans and data sharing requirements are the competing interests and differing motivations of the many stakeholders involved. These motivations and interests, as well as the incentives and disincentives for sharing of those who produce research data, need to be brought to the foreground.

WHAT ARE DATA?

A starting point to discuss the conundrum of sharing research data is to examine the complex notion of "data." An artifact or observation may be, at best, "alleged evidence," to use Michael Buckland's pithy phrase (Buckland, 1991). Data may exist only in the eye of the beholder: the recognition that an observation, artifact, or record constitutes data is itself a scholarly act. Data curators, librarians, archivists, and others involved in data management may be offered a collection that is deemed data by the collector, but not perceived as such by the recipients. Conversely, an investigator may be holding collections of materials without realizing how valuable they may be as data.

Data is a difficult concept to define, as data may take many forms, both physical and digital. Among the most widely cited definitions is this one, from a National Academy of Sciences report: "*Data* are facts, numbers, letters, and symbols that describe an object,

idea, condition, situation, or other factors." (A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases, 1999, p. 15). A more current working definition, from an internal Academy document, is particularly useful for discussions of sharing:

> The term "data" as used in this document is meant to be broadly inclusive. In addition to digital manifestations of literature (including text, sound, still images, moving images, models, games, or simulations), it refers as well to forms of data and databases that generally require the assistance of computational machinery and software in order to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socioeconomic data; and other forms of data either generated or compiled, by humans or machines. (Uhlir & Cohen, 2011).

The above notion of "data" transcends the sciences and other domains of scholarship, acknowledging the many forms that data can take. Data sources also vary widely. In the physical and life sciences, most data are gathered or produced by researchers, such as by observations, experiments, or models. In the social sciences, researchers may gather or produce their own data, or they may obtain data from other sources such as public records of economic activity. The notion of data is least well developed in the humanities, although the growth of digital humanities research has led to more common usage of the term. Humanities data most often are drawn from records of human culture, whether archival materials, published documents, or artifacts (Borgman, 2007; 2009).

The term "dataset" is sometimes conflated with the notion of "data." However, definitions of "dataset" in the scientific literature have at least four common themes – grouping, content, relatedness, and purpose – each of which has multiple categories (Renear, Sacchi & Wickett, 2010). While "dataset" may be useful to refer to a collection of data for the purposes of citation, the term does little to clarify what is meant by data.

Communities and Data

In the data management plan requirement, NSF sidesteps the definition of data with the first of its Frequently Asked Questions (NSF Data Management Plans, 2010):

> **1. What constitutes "data" covered by a Data Management Plan?** What constitutes such data will be determined by the community of interest through the process of peer review and program management. This may include, but is not limited to: data, publications, samples, physical collections, software and models.

NSF's choice of the term "community of interest" echoes the practice of the digital archiving world, where policies are framed in terms of the "designated community" (Reference Model for an Open Archival Information System, 2002). It is left to the investigator – or to the data archive – to designate the appropriate community of interest.

Therein lies the rub. An investigator may be part of multiple, overlapping communities of interest, each of which may have different notions of what are data and different data practices. The boundaries of communities of interest are neither clear nor stable. In the case of data management plans, an investigator is asked to identify the appropriate community for the purposes of a specific grant proposal and for the proposed duration of that award.

Communities of interest are narrower than disciplines or research specialties. Communities of practice (Lave & Wenger, 1991; Wenger, 1998) and epistemic cultures (Knorr-Cetina, 1999) are groupings commonly used in social studies of science. *Communities of practice* is a concept originated by Lave and Wenger to describe how knowledge is learned and shared in groups, a concept subsequently much studied and extended (Osterlund & Carlile, 2005). *Epistemic cultures*, in contrast, are neither disciplines nor communities. They are more a set of "arrangements and mechanisms" associated with the processes of constructing knowledge, and include individuals, groups, artifacts, and technologies (Knorr-Cetina, 1999; Van House, 2004). Common to both communities of practice and epistemic cultures is the idea that knowledge is situated and local. Nancy Van House (2004: 40) summarizes this perspective succinctly: "There is no 'view from nowhere' – knowledge is always situated in a place, time, conditions, practices, and understandings. There is no single knowledge, but multiple knowledges."

It is the difficulty of bounding either "community of interest" or "data," not to mention bounding the intersection of these two concepts, that makes data sharing requirements so challenging to articulate. Thus the next sections are devoted to explicating the many forms of research data that might be created and shared, or at least made sharable.

Categories of Data

Some types of data have both immediate and enduring value, some gain value over time, some have transient value, and yet others are easier to recreate than to curate (Preserving Scientific Data on Our Physical Universe, 1995; Bits of Power: Issues in Global Access to Scientific Data, 1997; Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age, 2009). Many of these distinctions depend on the category of data, as identified in an influential National Science Board

report (Long-Lived Digital Data Collections, 2005): observational, computational, experimental, and records.

Observational data include weather measurements and attitude surveys, either of which may be associated with specific places and times or may involve multiple places and times (e.g., cross-sectional, longitudinal studies). *Computational* data result from executing a computer model or simulation, whether for physics or cultural virtual reality. Replicating the model or simulation in the future may require extensive documentation of the hardware, software, and input data. In some cases, only the output of the model might be preserved. *Experimental* data include results from laboratory studies such as measurements of chemical reactions or from field experiments such as controlled behavioral studies. Whether sufficient data and documentation to reproduce the experiment are kept varies by the cost and reproducibility of the experiment. *Records* of government, business, and public and private life also yield useful data for scientific, social scientific, and humanistic research.

The physical and life sciences, which are the focus of the long-lived data collections report, exemplify all of these categories. As noted above, the social sciences encompass all of these categories, especially given the growth in modeling of social systems. Humanities scholars also model places and spaces, but are the least likely to perform experiments. While useful as a general framework, the National Science Board's four categories tend to obscure the diversity of data that may be collected in any given scholarly endeavor.

Investigators collect data for many purposes, using many methods. Research purposes, methods, and approaches all influence what investigators consider to be their "data," the degree to which those data might be sharable, and the conditions under which researchers are willing to share those data with others. The criteria for identifying data and for sharing are not yet well understood. Understanding practices, problems, and policies for data is an expanding area of research in the fields of information studies and social studies of science (Borgman, 2007; Bowker, 2000; 2005; Edwards, Mayernik, Batcheller, Bowker & Borgman, 2011, forthcoming; Karasti, Baker & Halkola, 2006; Mayernik, 2011; Mayernik, Batcheller & Borgman, 2011; Palmer, 2005; Renear & Palmer, 2009; Ribes, Baker, Millerand & Bowker, 2005; Ribes & Finholt, 2007; Wynholds, Fearon Jr, Borgman & Traweek, 2011; Zimmerman, 2007).

Purposes for Collecting Data

A brief survey of the purposes for which research data are collected will illustrate some of the complexities that arise in making them available to other potential users. Figure 1 presents three dimensions along which data collection may vary. These dimensions are neither exhaustive nor mutually exclusive. For each dimension, the first pole is the more local and flexible type of purpose, whereas the second pole is more global and

systematized. Scenarios drawn from our research on data practices are used to illustrate these dimensions.

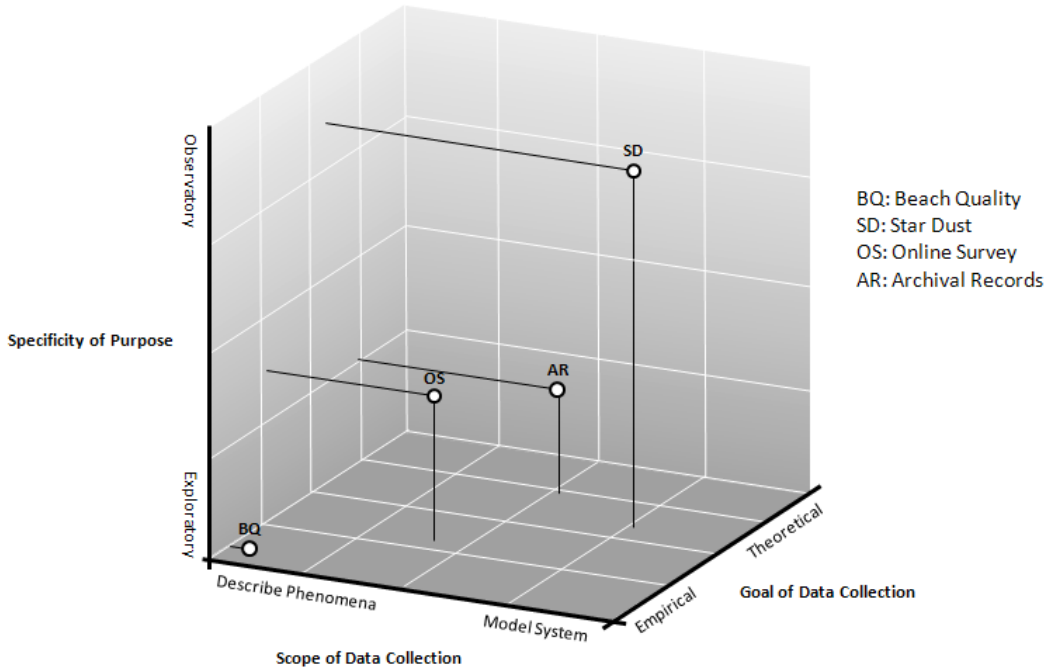

Figure 1: Purposes for collecting data. *Figure by Jillian C. Wallis.*

Specificity of Purpose

The first dimension illustrated is specificity of purpose, ranging from exploratory research to building observatories. Exploratory investigations pursue specific questions, often at a specific site, usually about a specific phenomenon, and may take place in a laboratory, a field setting, or some combination.

Studies to identify sources of bacteria and other beach contaminants offer examples of exploratory research. In this case, one or two students collect water samples, selected for time of day, location, weather conditions (e.g., dry or rainy), and other factors. Using a small portable wet lab, they dilute the samples to standard pH levels. The dilution varies by the expected concentration of bacteria, a judgment that requires scientific expertise specific to this type of research.

Once samples are collected, diluted, and brought to the campus lab, one of two processing methods is applied. The simplest and least expensive method is to culture the samples for 24 hours, then to count the bacteria. This method is slow and too insensitive to distinguish between human and animal sources of bacteria. The more sophisticated method is Quantitative Polymerase Chain Reaction (QPCR), adapted from medical applications, which requires greater expertise and is much more expensive. This method is faster and more sensitive, but results will vary between laboratories due to

choices of local protocols, filter material, machine type and model, and handling methods. Protocols and results are shared between partner laboratories seeking to perfect the method, but little other than the methods of data collection, protocols, and final curves might be reported in the journal articles. Biological samples are fragile; they degrade quickly or are destroyed in the analysis process.

At the other end of the specificity dimension are observatories, which are institutions for the observation and interpretation of natural phenomena. Examples include NEON and LTER in ecology (National Ecological Observatory Network, 2010; U.S. Long Term Ecological Research Network, 2010; Porter, 2010), GEON in the earth sciences (GEON, 2011; Ribes & Bowker, 2008), and synoptic sky surveys in astronomy (PAN-STARRS, 2009; Large Synoptic Sky Telescope, 2010; Sloan Digital Sky Survey, 2010). Observatories attempt to provide a comprehensive view of some whole entity or system, such as the earth or sky. Global climate modeling, for example, depends upon consistent data collection of climate phenomena around the world at agreed upon times, locations, and variables (Edwards, 2010).

The value of observatories lies in systematically capturing the same set of observations over long periods of time. Astronomical observatories are massive investments, intended to serve a large community. Investigators and others can mine the data to ask their own questions or to identify bases for comparison with data from other sources. Studies of the role of dust emission in star formation make use of observatory data. In one case, a team of astrophysics researchers queries several data collections that hold observations at different wavelengths, extracting many years of observations taken in a specific star-forming region of interest. They apply several new methods of data analysis to model physical processes in star formation. By combining data from multiple observatories, they produce empirical results that enable them to propose a new theory. They release the combined dataset when they publish the journal article describing their results.

Scope of Data Collection

The second dimension of Figure 1 is the scope of data collection. At one pole are studies that describe particular events or phenomena; at the other are studies that model entire systems. The beach quality research above is descriptive in nature, while the star dust example is closer to the model end of the dimension, as the purpose of their research is to model physical processes. Climate research spans this spectrum. Weather data, in the short term, can be used to describe or predict rain, snow, wind or other events. Systematic climate observations are used as inputs to models of physical process to study the earth's climate. The same observations may be input to multiple models, each developed by different research teams, and each with its own set of parameters and theories (Edwards, 2010).

Survey research tends to fall in the center regions of these dimensions, although individual studies may be more or less specific and may vary considerably in scope. An online survey of student attitudes can be conducted by a single investigator and deployed to a large number of universities. Large-scale surveys are suitable for hypothesis-testing and description of populations. Interviews are better for exploratory studies, and also can be used to develop theories. Conversely, because large online surveys are highly structured and usually anonymized, the resulting data are easier to share. Interviews are more open-ended, personalized, harder to anonymize, and difficult to code in consistent formats.

Among the great promises of data sharing is the ability to aggregate and compare multiple local studies. However, such integration of data is not always possible or desirable. Many research domains are concerned with rich descriptions of complex phenomena that are associated with specific times and places. Marine biologists, for example, study local phenomena such as harmful algal blooms. They may collect data for months or even years to capture conditions before, during, and after an event. Their goal is to understand the processes that trigger an event and how those processes evolve (Borgman, Wallis, Mayernik & Pepe, 2007; Gobler, Boneillo, Debenham & Caron, 2004). Small studies may cumulate into larger endeavors; in that case, the data from each individual study may become more valuable as the data cumulate, enabling comparisons across time periods and locations. In other cases, small studies may be one-off investigations of individual phenomena at a particular time and place (Borgman, Wallis & Enyedy, 2006; Bowker, 2000; Karasti et al., 2006; Karasti, Baker & Millerand, 2010).

It has proven difficult to aggregate studies of biological events into comprehensive systems models of the type used in climate research, due largely to differences in data characteristics (Aronova, Baker & Oreskes, 2010). The ecological sciences community is promoting data sharing by standardizing data collection practices, as their research tends to focus on local phenomena (Moore, McPeek, Rausher, Rieseberg & Whitlock, 2010; Whitlock, 2011). Data in most of the examples above would be considered observations, per the National Science Board categorization. Some might be considered natural experiments, such as comparisons of phenomena that occur in similar environments. Some of these observations could be aggregated into models of phenomena, others not. The beach quality example above reflects the local characteristics of much data collection. No matter how well they document their field practices, gathering an identical set of water samples is impossible due to changing environmental conditions.

Goal of Research

The third dimension in Figure 1 is the goal of the research, ranging from empirical to theoretical. Most research in the sciences and social sciences relies on empirical data of some sort. Scholars in the humanities often collect data, such as assembling records and

notes from archives. Theoreticians, who may or may not be the same people as those who collect empirical data, draw upon various forms of evidence to propose and to test theories.

Empirical investigations usually control some variables and test others. Those studying beach quality, for example, can control variables by diluting samples to standard pH values and by following consistent protocols. They can compare samples from different sites under different conditions. Their experiments can be replicated by gathering new samples, but they cannot reproduce results, as any given sample can be analyzed only once.

Those studying the atmosphere or the universe have little control over their variables. However, they can conduct experiments on theoretical models. This is commonly done with elaborate models for climate or economies. In theoretical research – whether climate, astronomy, or the economy – data may be simulated, rather than collected from "the real world." Observations of the physical universe occur at a unique place and time and can never be reconstructed, whereas experiments and models can be recreated (Preserving Scientific Data on Our Physical Universe, 1995). Important differences in terminology arise between experimentalists and theorists, however. Observations of the "real world" are data to experimentalists, while theoreticians often consider the simulated observations that are output from models to be their data (Edwards, 2010; Wynholds et al., 2011).

Approaches to Collecting Data

The process of data collection can be approached in many ways. Individuals and communities apply many combinations of techniques to identify, capture, describe, analyze, derive, and manage data. As with the dimensions of purposes outlined above, this selection of approaches is intended to be illustrative rather than exhaustive or mutually exclusive (Figure 2). They also range from the more local and flexible at the first pole to the more global and systematized on the second pole of each dimension.

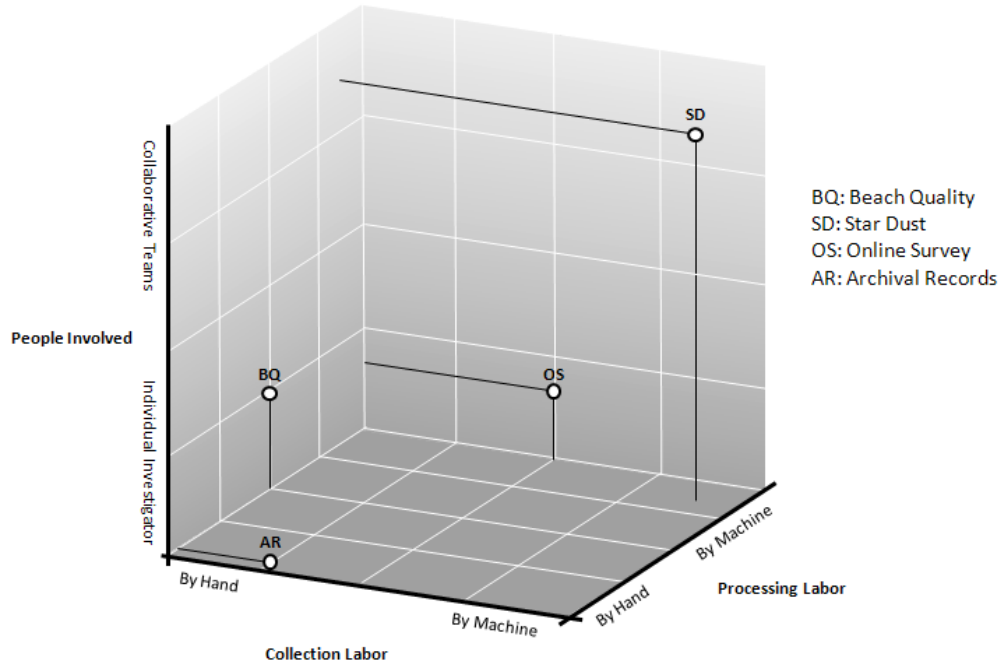

Figure 2: Approaches to collecting data. *Figure by Jillian C. Wallis.*

People Involved

The first dimension of approaches to data collection is the number of people involved. Individuals working alone have complete control over their methods and their data. Teams, who may be widely distributed, have to agree upon what data will be collected, by what techniques and instruments, and who has the rights and responsibilities to analyze, publish, and release those data (Borgman, Bowker, Finholt & Wallis, 2009; David, 2004; Olson, Zimmerman & Bos, 2008; Ribes & Finholt, 2007). The historian works alone with archival records. The sociologist conducting the online survey of student attitudes may work alone or with a small team of students and statisticians. The beach quality research is conducted by five to ten graduate and undergraduate students, and led by a single investigator. In contrast, dozens – if not hundreds or even thousands of people around the world – may be involved in collecting and curating data in observatories. Those who draw upon the data from those observatories may be individuals or teams of any size.

Labor to Collect Data

Approaches to data collection also differ in the amount of human labor required – the second dimension in Figure 2. Investigators in beach quality, marine biology, or other field research may spend days, weeks, or months hand-gathering physical samples of soil, water, or plants, which then must be processed in a laboratory to extract data – a process that also may require days, weeks, or months. Similarly, the historian may spend months or years in historical archives, taking notes on laptops – or only with pencil on paper, by the rules of some archives. Extracting useful data from those notes

may also require months or years. These labor-intensive approaches have the advantage of flexibility and local control by the investigators. They have the disadvantages, from a data sharing perspective, of being difficult to replicate and producing data that are not consistent in form or structure.

Machine-collected observations, whether by telescopes, sensor networks, online survey software, or social network logs, may be labor-intensive to design and develop, but once deployed can produce massive amounts of data that can be used by many people. Major telescopes, both on land and in space, for example, require long-term collaborations among scientists and technologists. Data structures, management, and curation plans are developed in parallel with the design of studies and instruments, a process of a decade or more. Machine-collected data tend to be consistent, structured, and to scale well, but considerable expertise is required to interpret them. Conversely, these forms of data collection are less flexible and adaptable to an individual investigator's research questions. Observation parameters may be hard-wired or coded into the technologies, thus determining the data that can be obtained.

Labor to Process Data

A third dimension of approaches to data collection is the amount and type of processing required for interpretation. At one pole of this dimension in Figure 2 are human processing techniques. These techniques may be simple, such as measuring the dimensions of leaves, assigning geo-spatial and temporal parameters (e.g., precise location and time where gathered), and assigning experimental parameters (e.g., amount of sunlight and shade, proximity to the ground or water). In other cases of collecting physical samples, the actual "data" may be instrument readings (e.g., a type of nitrate as indicated by a voltage measurement on a sensor, or concentration of a bacterium in parts per million of water). Whether the numbers are hand-written in a field notebook or machine generated, they must be associated with a specific sample. Other information such as the type of machine, its calibration, the time, date, and place of data collection, and the method by which the sample was captured are necessary to interpret any given data point (Borgman, Wallis & Enyedy, 2007). Similarly, the historian in an archive is documenting characteristics of the records examined so that those notes can be interpreted later.

In the most highly instrumented research, such as astronomical sky surveys, instruments capture contextual information about the data. While minimal human labor may be required for processing the data, considerable expertise is required to assess the accuracy of data and metadata in these research environments, as minute errors in calibration can influence analysis and interpretation significantly (Mayernik et al., 2011; Wynholds et al., 2011). Statistical analysis can be applied both to online surveys and to the star dust data drawn from observatories. Human processing labor may be minimal, but the domain expertise required for analysis is high, of course. The beach quality

studies fall in the middle of the dimension, as samples gathered by hand may be processed by sophisticated technologies such as QPCR to yield numerical results.

Generally speaking, the more hand-crafted the data collection and the more labor-intensive the post-processing for interpretation, the less likely that researchers will share their data. However, practices vary so widely across fields and research teams that any such generalizations are difficult to make (Hilgartner & Brandt-Rauf, 1994; Pritchard, Carver & Anand, 2004).

WHY SHARE RESEARCH DATA?

As is evident from the above discussion of the purposes and approaches to collecting data, investigators (and their collaborators, students, and staff) devote massive amounts of physical and intellectual labor to collecting, managing, and analyzing their data and to publishing their results. Data are the lifeblood of research in any field, but just what are those "data" varies by purpose, approach, community, and many other local and global considerations. Some of those data may be in sharable forms, others not. Some data are of obvious value to the community, others not. Some researchers wish to share all of their data all of the time, some wish to share none of their data none of the time, and most are willing to share some of their data some of the time. These competing perspectives, the array of data types and origins, and the variety of local circumstances all contribute to the intricacy and difficulty of sharing data.

The pressure to share data comes from many quarters: funding agencies – both public and private, policy bodies such as national academies and research councils, journal publishers, educators, the public at large, and from researchers themselves. These stakeholders each have their own reasons for requiring or encouraging data sharing. In examining the public statements of these entities, some identify explicit benefits of data sharing to specific parties, while others are vague about why data should be shared and who will benefit. A review of policy documents, recent studies of data sharing, and participation in public discourse on the topic yields two dimensions along which rationales for data sharing vary: motivations for sharing research data and the interests that are served by sharing data. These are presented in Figure 3.

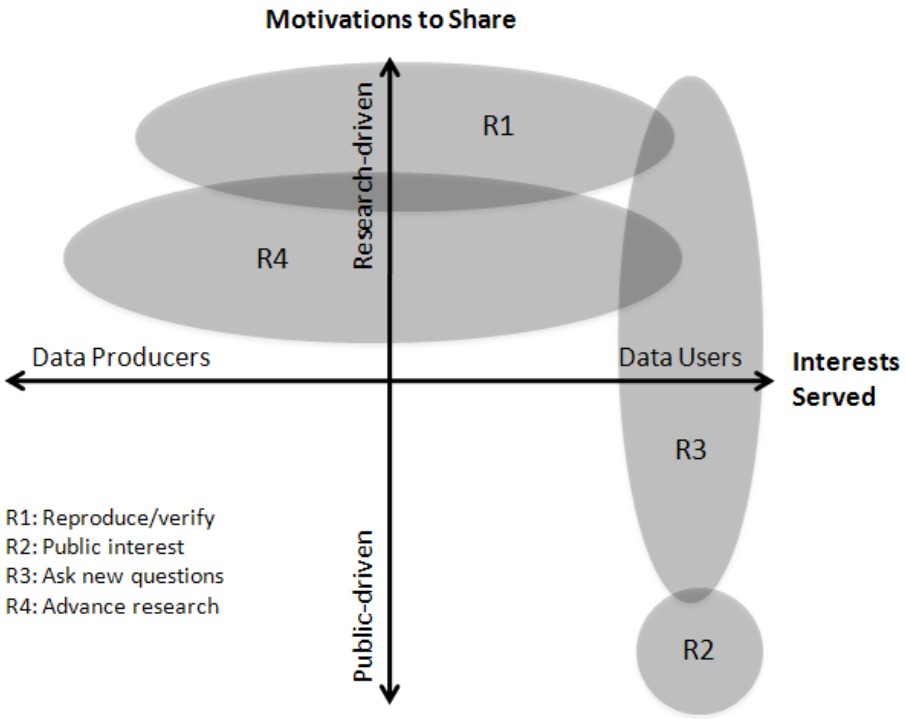

Figure 3: Rationales for Sharing Research Data. Figure by Jillian C. Wallis

Policies to promote data sharing may be motivated by research needs or by service to the public at large. Researchers who produce data may (or may not) feel motivated to share them with others, either within their own communities or with larger audiences. Funding agencies and journals usually promote data sharing to serve their respective communities. Motivations of the many stakeholders in research data may be aligned or may be in conflict.

Sharing research data may serve the interests of researchers who produce data, the interests of those who might use research data, or both. Funding agencies promote data sharing to serve the research community and perhaps also to serve the general public. Journals wish to serve their readers, their authors, and their publishers. Researchers' interests in releasing their own data may or may not align with their interests in gaining access to the data of others. Similarly, funding agencies' and journals' purposes for data release may not align with the interests of researchers who create those data.

Neither dimension is absolute; the poles represent relative positions of people or situations. For example, a researcher or policy maker may make one argument on behalf of the producer of data and another on behalf of the users. Similarly, an argument made in the name of scholarship may also benefit the public good. These motivations and interests are not mutually exclusive; rather, they provide a two-dimensional space in which to place the various rationales in favor of sharing research data.

Subtle distinctions in the premises for data sharing may lead to markedly different policies, economic models, research practices, curation practices, and degrees of compliance. Of particular concern is how those premises align with the interests of those whose work produces the data. Accordingly, the following four rationales for sharing research data focus on the concerns of data producers, and on their abilities and incentives to share their data.

The model proposed here is intended to provoke discussion among the many stakeholders in research data. Most of the examples are drawn from the sciences and social sciences, as these are the areas most studied and are on the front lines of current policy debates. This analysis can be extrapolated to the humanities, where similar policies and deposit requirements are in discussion (Kansa, Kansa, Burton & Stankowski, 2010; Unsworth et al., 2006). The ability to implement any data sharing policy will depend on many factors, including local data practices, differences in the intellectual property rights intrinsic to data sources, and the need to maintain confidentiality of human subjects (Borgman, 2007).

1. To reproduce or to verify research

The strongest, yet most problematic, rationale for sharing research data is the ability to reproduce research results. The motivation is fundamentally research-driven but can also be viewed as serving the public good. Reproducing a study confirms the science, and in doing so confirms that public monies were well spent. However, the argument can be applied only to certain kinds of data and types of research, and rests upon several questionable assumptions.

Peer review depends upon the ability of reviewers or referees to judge the reliability and validity of a research report based on the information provided. In only a few fields do reviewers attempt to reanalyze or verify data or to reconstruct all the steps in a mathematical proof or other procedure. Even when data are included with a journal article or conference paper, rarely is enough information provided to reproduce the results. Instrument details and calibration may be omitted, or lab-specific practices may not be documented in sufficient detail. This is normal practice, both because journal space constraints discourage elaborate methods sections and because research expertise relies upon tacit knowledge that is not easily documented (Bowker, 2005; Kanfer et al., 2000; Latour, 1987). Yet whenever published articles are withdrawn from major journals, questions are raised about what the reviewers knew – or should have known – about the data and procedures (Brumfiel, 2002; Couzin & Unger, 2006; Couzin-Frankel, 2010; Normile, Vogel & Couzin, 2006).

Reproducibility is a high bar, and even it has several levels, such as the precise duplication of observations or experiments, exact replication of a software workflow, degree of effort necessary, and whether proprietary tools are required (Reproducible

research: Addressing the need for data and code sharing in computational science, 2010; Stodden, 2009a; b; Vandewalle, Kovacevic & Vetterli, 2009). Observations of the "real world" – such as water samples, algae, air and soil temperature, or comets – are associated with times and places, thus rarely can they be reproduced. It is for this reason that observations are worth curating. Experimental observations may be reproduced in laboratory conditions, given sufficient documentation and access to the same materials, equipment, software, and technical expertise. Given the same set of observations, whether natural or experimental, other researchers may be able to replicate the results, if identical circumstances can be achieved. However, experienced researchers know that minute differences in procedures, machine calibrations, temperature and humidity near the instruments, and other factors can introduce undetected variation that influences replicability and interpretation.

Even the research activities are difficult to replicate, since processing and analysis tools such as statistical, mathematical, and workflow software rarely maintain a precise record of the systems or the data at each transaction (Claerbout, 2010; Goble & De Roure, 2009). Questions of the provenance – both in the archival sense of "chain of custody" and in the computing sense of "transformations from original state" – arise in interpreting data that may have evolved through multiple hands and processes (Buneman, Khanna & Tan, 2000; Gil, 2010; Hunter & Cheung, 2007),

Other impediments to reproducibility include difficulties in gaining access to data and to tools used to create and analyze those data, lack of a licensing regime to provide access to proprietary software and to data (Stodden, 2009b), and conflicts in copyright law, such as differences between the U.S. and Europe in the ability to make proprietary claims on factual matters (Reichman & Uhlir, 2003).

Reproducibility is elusive due to the vagaries of the research process, the varying notions of data, disparate community practices for documenting evidence, the transitory nature of observations, and the combinations of expertise necessary for interpretation of evidence. As a motivation for sharing research data, reproducibility is problematic not only because it applies to so few types of research, but because it risks reducing the research process to a set of mechanistic procedures. Even chemists will acknowledge that their work is as much art as it is science (Lagoze & Velden, 2009a; b). True reproducibility requires deep engagement with the epistemological questions of a given research specialty, and the very different ways in which investigators obtain and value evidence. Re-users of data may not know, or be able to know, what prior actors did to the data. Each step in cleaning or processing data requires judgments, few of which may be fully documented. Later interpretations thus may depend upon multi-level inferences that are statistically problematic (Meng, 2010).

Although reproducibility is a popular term in promoting data release, the concept is often conflated with verification of results. While a less precise notion, verification is easier to accomplish. Peer reviewers and readers can judge whether the research

methods meet community standards, and whether the evidence is appropriate for the claim, and whether the arguments are reasonable, even if they are unable to reconstruct all of the procedures. Potential re-users of data also can apply their own judgments about the veracity and usefulness of the data; their criteria may be highly specific to their research topic (Faniel & Jacobsen, 2010; Wynholds et al., 2011).

Incentives to share research data for the purposes of reproducibility or verification vary widely by type and condition of the data and expectations of the community. Where data deposit is required as a condition of publication, as in the case of genomes and the Protein Data Bank discussed earlier, researchers will comply. If reproducibility requires materials that are very difficult to share, such as specialized animals or cell lines, then community practices will dictate the conditions of data release. Considerable human expertise, labor, and licensing of intellectual property may be involved. Researchers give these reasons and many others for refusing to release data (Campbell et al., 2002; Hilgartner, 1997; 1998; 2002; Hilgartner & Brandt-Rauf, 1994). If materials and documentation are highly automated, if no licensing restrictions apply, and if the researcher has completed his or her publication of the results, then data sharing is more likely to occur.

However, researchers may never be "done" with their data. In cases where a research career is based on long-term study of a specific species, locale, or set of artifacts, data become more valuable as they cumulate. Researchers in these situations may be particularly reluctant to release data associated with a specific publication, as it might mean releasing many years of data. Similarly, reproducing the data associated with any given publication is problematic, as the set of observations reported may depend heavily on prior studies and on interpretation of much earlier data.

Scholars may wish to verify findings either to build upon or to refute them. The generalized rationale of sharing data for reproducibility or for verification of results lies near the research-driven pole of the motivation dimension in Figure 3. It spans the interests of data producers, who can benefit by having their findings verified by others to reinforce their veracity, and the interests of researchers who would use others' data. The greater possibility of replication or verification occurs with data produced for the purposes of observatories, modeling systems, or for theory building – the far end of the dimensions in Figure 1 – or those collected by large teams, by technologies, and subject to machine processing – the far poles of Figure 2. These categories of data are more likely to be captured and described consistently than are data collected for exploratory investigations, to describe phenomena, for empirical studies, by individual investigators, or processed by hand. Innovation in research requires new methods of research design, analysis, and technology.

From an epistemological perspective, reproducibility and verification is the most problematic of the four arguments for sharing data. Often the research creativity lies in identifying a new method required to approach an old problem. Research outcomes

often depend much more on interpretation than on the data per se. Separating data from context is a risky matter that must be balanced carefully against demands for reproducibility.

2. To make results of publicly funded research available to the public

Public sentiment for sharing research data is based on the rationale that tax monies should be leveraged to serve the public good. Data produced with public funds should be available for use and should not be hoarded by researchers. These notions are implicit in the OECD principles, to which several of the funding agency policies refer. Open access to research data is a means to leverage public investment in research (OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007). The OECD document also builds upon an earlier U.S. study, explicitly quoting this passage: "The value of data lies in their use. Full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from publicly funded research" (Bits of Power: Issues in Global Access to Scientific Data, 1997).

U.S. public policy tends toward openness of research information. Federal law waives copyright protection on data and information directly produced by government agencies, putting those materials into the public domain in the U.S. (Reichman & Uhlir, 2003). Data and information resulting from research grants to universities and other agencies do not fall under the same law.

The argument for access to the products of public research funding applies both to data and to publications, but is playing out differently between the two. The public monies-public goods rationale has succeeded in the biomedical research community for the open deposit of publications, but not without resistance, especially on the part of publishers. Biomedical information has a substantial audience, including biomedical researchers, clinicians, the pharmaceutical industry, and patients, so it is not surprising that this was the first frontier for open access. Publications resulting from NIH funding must be deposited into PubMed Central within 12 months of publication, an embargo period that protects the journals (NIH Public Access Policy, 2005). However, with regard to data, the NIH requires the release only for grants over a certain size, and allows these data to be embargoed for a certain period of time to enable the investigators to publish their findings. NIH does not require that the data be deposited in any particular resource, only that it be released.

The Wellcome Trust, the largest funder of biomedical research in the United Kingdom, requires both publications and data from its grants to be made available. They support multiple types of open access publication, but do not follow the NIH model of requiring

deposit in a specific repository (Wellcome Trust position statement in support of open and unrestricted access to published research, 2005; Fazackerley, 2004).

The National Science Foundation policy is ambiguous with respect to what must be released. Publications are defined as a type of data within the scope of the Data Management Plan, but NSF does not specifically require that publications resulting from their grants be released openly (NSF Data Management Plans, 2010). In light of growing pressure for open access to publications, the situation may change (Directory of Open Access Journals, 2009; Open Content Alliance, 2009; Beaudouin-Lafon, 2010; Crow, 2009; Kaiser, 2008; Ware, 2010; Young, 2009). One result of the popularity of the NIH publication deposit requirements is a bill introduced into Congress in 2010 (H.R. 5037) to require federal research granting agencies to make resulting publications available to the public (Federal Research Public Access Act of 2009).

Associating the release of data and publications is important not only as an economic rationale, but as support for reproducibility, verification, and reuse. The publications are the primary form of documentation for most types of data. They explain the research problem addressed, the methods by which the data were collected, the analyses performed, and the interpretation of the results. Publications add value to data and vice versa.

The Economic and Social Research Council, which is the primary U.K. funding agency for these areas, recently issued a data policy that goes well beyond the scope of the NSF, NIH, and Wellcome Trust requirements. The ESRC not only requires the public release of data, it requires that investigators wishing to create new data must "demonstrate that no suitable data are available for re-use." (ESRC Research Data Policy, 2010: 3). For grants to create new data, ESRC requires a "data management and sharing plan." Grantees also must prepare their data for "re-use and/or archiving" with an ESRC data provider within three months of the end of the award. If the grantees have not offered the data to an appropriate provider in that time period, the agency may withhold the final payment on the award. The ESRC funds repositories that can curate data from their grants. Notably, the ESRC also acknowledges that the repositories can enforce selection policies for the data they collect, lest these repositories become a catch-all for anyone's data, regardless of quality or potential for future re-use.

The "public monies for public good" argument resonates with legislators, taxpayers, and the general public. It also resonates with researchers whose data are readily reusable by those without substantial domain knowledge, such as types of astronomical or earth observations that can support citizen science. Another public-interest-driven rationale is that data release will minimize duplication of research effort, which in turn results in fewer human subjects being required to establish findings (Fischer & Zigmond, 2010).

In most research projects, data are collected by individuals or by small teams. Methods are local and are specific to the research questions at hand. Reusing these types of data requires considerable knowledge of the procedures by which they were collected, which in turn requires considerable expertise in the research specialty. The farther removed from the data collection activity, the harder it is to make use of someone else's data. Thus it is not surprising that concerns for the misinterpretation and misuse of data are common reasons that researchers give for not sharing (Campbell et al., 2002; Hilgartner, 1997; Hilgartner & Brandt-Rauf, 1994).

Researchers are more willing to share their data with those in their immediate area of specialty than with the general public. Those within their community of interest – to use the NSF term – have the expertise to interpret the data and thus are most likely to benefit from access. Making data available to the general public requires much more documentation effort. Researchers are also concerned about possible misuse or misinterpretation, which creates incentives not to share their data broadly. Data from observatories, where investment in documentation and curation is usually included in the research design and funding, are more readily released to the public. Observatory data are only a small subset of extant data resources, and they also can be sensitive. Global and comparative research on climate change depends on open access to data (Overpeck, Meehl, Bony & Easterling, 2011), yet the politicization of climate data (Costello, Maslin, Montgomery, Johnson & Ekins, 2011; Gleick, 2011) makes researchers in these and in other fields wary of releasing their data.

3. To enable others to ask new questions of extant data

A more focused rationale is that sharing data enables others to ask new questions, whether from an individual dataset or by combining multiple sources. This framing also has two strands, one on behalf of researchers and one on behalf of the general public. Researchers have argued that open access to data encourages meta-analysis: the ability to combine data from multiple sources, times, and places to ask new questions (Whitlock, 2011). In this view, access to data is less about inspecting the findings of an individual project and more about the ability to combine data from multiple projects. Indeed, the greatest advantages of data sharing may be in the combination of data from multiple sources, compared or "mashed up" in innovative ways (Butler, 2006). Data are most reliably integrated when collected and processed systematically, in ways that support the standards of large communities. Common data structures, metadata formats, and ontologies help support mining and integration of multiple data sources.

The public good strand of the "ask new questions" rationale was framed most visibly by Chris Anderson, Editor-in-Chief of *WIRED* magazine, in a special issue on data (Anderson, 2008):

> The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding

> the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

Anderson captures the public excitement about the promise of "big data" to explore new questions, and to combine data from multiple sources to identify new relationships. The promise is also reflected in the "fourth paradigm" of data-intensive science, where the sheer scale of data enables analyses never before possible (Hey, T., Tansley & Tolle, 2009).

However, the enthusiasm for data mining masks the expertise required for interpretation. Those in the scientific and technical computing communities who are promoting data-intensive research are well aware of the difficulties involved. Assessing the veracity and integrity of a given dataset requires domain expertise, and that assessment depends upon the extent of the documentation available (Faniel & Jacobsen, 2010). Scientists compute upon large datasets both for exploratory investigations and to generate new theory, which is quite the opposite of Anderson's conclusion that big data means "the end of theory." As Edwards (2010), Rogers (1995), and many others have explained, data and theory are inseparable. Investigations are designed to test or to develop theories, and those theories are used to make sense of the data. Scholarship entails fitting data and theory.

This rationale, to enable others to ask new questions of extant data, serves the interests of users rather than producers of data. The data mashup strand of this argument falls in the upper right quadrant of Figure 3, as the intended others are a peer community of researchers, whereas data mining by anyone, for any reason, falls in the lower right quadrant, primarily serving the general public. Most researchers will share more readily with their peers, given the concerns for labor, interpretation, and likelihood of reuse.

4. To advance the state of research and innovation

The rationale for sharing data which resonates with the widest array of stakeholders is that research and innovation can be advanced more effectively. This is the "fourth paradigm," introduced above: computational science constitutes a new set of methods beyond empiricism, theory, and simulation (Bell, Hey & Szalay, 2009; Gray et al., 2005; Hey, T. et al., 2009). The fourth paradigm claim is appealing, but tends to overreach. Wilbanks (2009) strikes a middle ground, viewing data not as a method per se, but as a rich resource for any of the empirical, theoretical, simulation, or computational paradigms.

One distinction between the "ask new questions" and "advance research" rationales is that the latter is wholly motivated by research interests. It also goes beyond asking new questions of extant data; it addresses the need for more data and for curation of

existing data in ways that ensure their usefulness. Simply put, "Science depends on good data" (Whitlock et al., 2010: 145) and "Data are the main asset of economic and social research" (ESRC Research Data Policy, 2010: 2). Assertions such as these are most common in data-intensive fields that benefit from observatories and synoptic surveys, such as astronomy and social science survey research, and fields in which comparisons across time and space are beneficial, such as some areas of biology and ecology. When data are shared quickly and openly, researchers can draw upon each other's data more readily. For example, some space-based telescope missions alert other astronomy projects when something of interest is spotted, enabling other investigators to turn their instruments toward the specified coordinates. Thus one instrument might identify an object or event and an unrelated project might obtain follow-up observations (Drake et al., 2011),

Fischer and Zigmond (2010), writing in *Science and Engineering Ethics*, identify a number of ways in which data sharing advances the state of science. These include maximizing the use of data, increasing the impact of findings, progressing the state of research faster and farther, laying a broader foundation for knowledge, expanding the scope of research, and diversifying perspectives.

Data from publicly funded research are more likely to be shared than are data resulting from privately funded research, especially in cases where the research is proprietary. Academic researchers in fields where data have high monetary value and much of the research is proprietary, such as chemistry and the biosciences, are at a disadvantage in terms of access to data. Data sharing does occur within academe and between academe and industry, but to a lesser degree than in fields where most research relies on public funds (Haeussler, 2011; Lagoze & Velden, 2009a; b). Establishing open data and metadata standards for chemistry has been highly contentious in comparison to other scientific fields (Murray-Rust & Rzepa, 2004). Open data structures facilitate the massing of large amounts of data that can be mined to ask new questions. In the biosciences, "the likelihood of sharing decreases with the competitive value of the requested information" for both academic and industrial researchers (Haeussler, 2011: 105). The humanities and social sciences have similar problems when their data sources are copyrighted materials owned by others. Scholars may be able to quote small portions of texts, but not reproduce still or moving images or mine digital resources in depth.

The argument that data sharing can advance scholarship goes beyond data release. Once made available, data can be curated in ways that add value for the research community. The notion that data curation is a *means* to advance science is the cornerstone for the Data Conservancy, one of the consortia funded by the NSF DataNet Program (Data Conservancy, 2010; Sustainable Digital Data Preservation and Access Network Partners (DataNet), 2010): "The Data Conservancy (DC) embraces a shared vision: scientific data curation is a means to collect, organize, validate and preserve data so that scientists can find new ways to address the grand research challenges that face society."

The OECD principles have a similar tone: "Sharing and open access to publicly funded research data not only helps to maximise the research potential of new digital technologies and networks, but provides greater returns from the public investment in research." (OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007: 10).

Advancing scholarship is a rationale that spans the interests of data producers and users, thus is positioned in both upper quadrants of Figure 3. If data can be aggregated into a critical mass and curated in ways that make them more accessible and valuable, then those who produce the data can exploit them better, as can other data users. Researchers are more likely to share their data if they know the data will be well managed for future use. Data collected systematically to community standards for structure and content are most easily curated and compared, and researchers are most willing to share them. Many other types of valuable data do not meet these criteria, such as the beach quality and harmful algal blooms examples. While sharing these data also may advance scholarship, the researchers who produce those data are more likely to exchange them within their immediate areas of specialty than to release them to the general public.

DISCUSSION AND CONCLUSIONS

For the last 25 years, the need to share research data has been declared to be an urgent problem. Yet the discussion continues, policies proliferate, and evidence of data sharing is apparent in only a few research fields. Sharing research data is clearly a conundrum: an intricate and difficult problem. Acknowledging that data sharing is difficult does not mean abandoning all hope that some data will be shared with some people some of the time. The challenges are to understand which data might be shared, by whom, with whom, under what conditions, why, and to what effects. Answers to these questions will inform data policy and practice.

The complexity of the "simple statement" in the introduction to this article should by now be apparent:

> If the rewards of big data are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others.

Neither the producers of data nor the agencies that require sharing can agree on what *are* the data. Data take many forms, both physical and digital. They are much more than numbers in a spreadsheet: data can be samples, software, field notes, code books, instrument calibrations, archival records, or a myriad of other information objects, none of which may stand alone. *Sharing* can encompass acts as varied as announcing the existence of data, posting them on a website, or contributing them to a richly curated

repository. *Interpretable* and *reusable* are the most problematic. Interpretable presumes sufficient expertise to assess the integrity of the data and to grasp their meaning. It also presumes adequate documentation of the context of the data creation, processing, and provenance. Reusable is a standard just short of reproducibility. Considerable expertise, effort, restructuring, and proprietary software may be necessary to reuse data. Similarly problematic is the requirement to curate data in such ways that they are "independently understandable" (Reference Model for an Open Archival Information System, 2002). While a laudable goal, it's rarely feasible in any absolute sense, any more than is reproducibility.

Other barriers to sharing research data include lack of reward or credit for sharing, the substantial amount of labor required to document data in reusable forms, concerns for misuse or misinterpretation of data, control over intellectual property, and the need to restrict access or to de-identify data on human subjects or endangered species. Perhaps the largest barrier is the lack of demonstrated demand for research data outside of genomics, climate science, astronomy, social science surveys, and a few other areas. Most funding agencies and review panels evaluate grant proposals on the basis of the new data to be created in support of a research endeavor. Few promote innovation through reuse of data; the ESRC is a notable exception. Until these many barriers are addressed, data sharing is unlikely to increase substantially.

The rationales for sharing research data vary along two dimensions, as presented in Figure 3. Underlying the data sharing policies of funding agencies, journals, and other stakeholders may be the motivation to advance research or to serve the public good. Rationales also vary by whether they serve primarily the interests of researchers who produce data or the interests of prospective users of research data. Researchers' incentives to share data depend on many factors, including not only the rationale for sharing but the types of data, the purposes for which they were collected, the approaches taken to data collection and processing, concerns for potential misuse or misinterpretation, resources for documenting data, and means to curate and disseminate their data.

Four rationales are presented and are positioned in the model:

1. To reproduce or to verify research
2. To make the results of publicly funded research available to the public
3. To enable others to ask new questions of extant data
4. To advance the state of research and innovation

The first rationale is the strongest from a research perspective, and yet the most problematic. Questions of reproducibility are deeply intertwined with the epistemology of the research specialty. The second and third rationales are the most driven by public interests and are presented from the perspective of those who wish to use data

produced by other parties. The fourth, which also benefits the public, is framed in the interests of data producers, and serves research, innovation, and scholarship.

The analysis presented here focuses on the interests of the researchers who produce data, in an effort to identify types of research in which data sharing is most appropriate, and to identify policies and practices that may encourage data sharing. Incentives to release data depend to a large degree on the labor required, which varies both by the purposes for which data were collected (Figure 1) and the approaches to collecting data (Figure 2). Data collected for observatories or for model building require structure and documentation, which makes them suitable for wider release. Instrumented data may contain automated markup that facilitates release. Hand-collected observations by individual investigators, whether ecological field studies or ethnographies, may require the most labor to document in sharable forms.

Data are more likely to be shared when the policies serve the interests of those who produce the data. This is a simple statement of self-interest. Researchers collaborate, but they also compete for grants, for jobs, for publication venues, for students, and for other resources. They must choose carefully where to spend their time and resources. Time and money spent on documenting data for use by others are resources not spent in data collection, analysis, equipment, publication fees, conference travel, or other research necessities. While it can be argued that good data practices benefit the originating researcher, far less documentation is required to maintain data for one's own use than to release those data to the public. Data release is costly. Even if data sharing is built into the cost of research funding, the requirement may substantially increase the cost of doing research. Data release is more beneficial if those data are curated in ways that make them useful to others over some long period of time. Data curation likewise is very expensive, and unlikely to be justifiable for all forms of data. Issues of selection and appraisal to determine which data are worth curating and for how long are urgent matters that require much more attention. Similarly, more needs to be known about potential uses and users of research data. One reason that researchers do not release data is because they cannot imagine who might use them (Mayernik, 2011); they lack a "recursive public," to use Kelty's term (2008: 3).

However, these concerns beg the question of what *are* the data in any given investigation. Releasing the spreadsheets or statistical files associated with tables in a journal article is a much different requirement than is releasing physical samples, hand-written field notebooks, or raw observations from complex instruments. Numerical data are of little value without the software associated with the data collection, analysis, and processing technologies. That software may be proprietary or it may be crafted locally as part of the research project. In either case, the software necessary to interpret the data may not be sharable. Even if the data were produced with common tools, those data may be unreadable after a few years due to changes in hardware and software,

unless they have been curated well and migrated to new technologies. Yet more problematic is the fact that many new forms of research data are not datasets that exist in bounded forms that can be curated. Rather, they are streams of observations flowing from sensor networks, telescopes, social networks, public cameras, and countless other monitoring technologies. Any such dataset that might be shared is at best a snapshot in time.

Funding agency requirements for sharing data acknowledge that policies must differ by research community. Identifying appropriate policies is challenging because the "research community" does not speak with one voice. Nor does the "astronomy community," the "biology community," the "sociology community," or any other specialty. Collaborative research may cross the boundaries of disciplines, specialties, universities, and countries. Learning the interests of a given community, however narrowly or broadly defined, requires close engagement and study. The social study of science dates to the mid-20th century (Latour & Woolgar, 1979; Merton, 1969; 1973), and the interest in practices associated with data has accelerated in the last decade (Borgman, 2007; Bowker, 2005; Edwards, 2010). Social science and humanities research practices have received far less attention; more studies of these also are needed (Borgman, 2009).

Initiatives such as the NSF DataNet program (Sustainable Digital Data Preservation and Access Network Partners (DataNet), 2010) endeavor to bring researchers, librarians, and systems developers together to understand community-driven design for data curation. Multiple, parallel studies of individual research groups and communities are under way to inform both policy and design. Our research on astronomers, as part of the DataNet program (Data Conservancy, 2010), reveals that community-driven design means selecting and organizing data to reflect specific practices (Wynholds, 2010; Wynholds et al., 2011). At one extreme, very fine details of instrument design and calibration must be associated with data. Multi-dimensional temporal and spatial coordinates also may be essential. At the other extreme, researchers would like to be able to explore massive repositories of data without having to know those fine details. To paraphrase one of the astronomers we interviewed, "only about 10% of all our data has 'eyes on.' We rely on analytical tools to see the rest of it." Several have expressed concern over the design of current data repositories, which may be optimized for database performance rather than for scientific inquiry. Similar insights likely exist for any field whose data may be curated; they await study and partnership.

We are beginning to understand why researchers do not share data readily. We know even less about why researchers do share data and why they reuse data. Thus much more research is needed about practices in fields that *do* share data consistently and practices in fields where data are consistently re-used. Only with this knowledge in

hand, coupled with a richer understanding of the array of physical and digital entities that might be considered "data," can better policies, practices, services, and systems be developed to support the sharing of research data.

ACKNOWLEDGEMENTS

This article is based on an unpublished conference paper (Borgman, 2010). Ideas were developed further through presentations to the National Research Council Board on Research Data and Information and to the Santa Fe Institute. My writing of this article benefited greatly from discussions and comments on earlier drafts by the CENS Data Management team at UCLA – Alberto Pepe, David Fearon, Matthew Mayernik, Katie Shilton, Jillian Wallis, and Laura Wynholds; collaborators Sharon Traweek (UCLA), Catherine van Ingen and Catherine Marshall (Microsoft Research), the Monitoring, Modeling, and Memory research team – Paul Edwards, Steven Jackson, Archer Batcheller, and Ayse Buyuktur (Michigan), Geoffrey Bowker (Pittsburgh), and David Ribes (Georgetown) – and numerous discussions with Paul Uhlir (National Academy of Sciences) and Victoria Stodden (Columbia). Jillian Wallis revised and much improved the figures. Monica Pamela Garcia (UCLA) provided expert bibliographic assistance and fact checking. George Djorgovski (Caltech), Alyssa Goodman (Harvard), Kathryn Mika (UCLA), Alex Szalay (Johns Hopkins), and many anonymous research subjects provided examples of data production and use that are mentioned herein.

Research reported here is supported in part by grants from the National Science Foundation (NSF): (1) The *Center for Embedded Networked Sensing* (CENS) is funded by NSF Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator; (2) *Towards a Virtual Organization for Data Cyberinfrastructure*, #OCI-0750529, C.L. Borgman, UCLA, PI; G. Bowker, Santa Clara University, Co-PI; Thomas Finholt, University of Michigan, Co-PI; (3) *Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures*: #0827322, P.N. Edwards, UM, PI; Co-PIs C.L. Borgman, UCLA; G. Bowker, SCU and Pittsburgh; T. Finholt, UM; S. Jackson, UM; D. Ribes, Georgetown; S.L. Star, SCU and Pittsburgh; and (4) *The Data Conservancy*, NSF Cooperative Agreement (DataNet) award OCI0830976, Sayeed Choudhury, PI, Johns Hopkins University. We also are grateful to Microsoft Technical Computing and External Research for gifts in support of this research program.

REFERENCES

Abrams, S., Cruse, P. & Kunze, J. (2009). Preservation is not a place. International Journal of Digital Curation, 4(1). Retrieved from http://www.ijdc.net/index.php/ijdc/article/viewFile/98/73 on 3 May 2011.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. Wired. Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory on 22 July 2008.

Aronova, E., Baker, K. S. & Oreskes, N. (2010). Big Science and Big Data in Biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957–Present. Historical Studies in the Natural Sciences, 40(2): 183-224.

Beaudouin-Lafon, M. (2010). Open access to scientific publications: The good, the bad, and the ugly. Communications of the Association for Computing Machinery, 53(2): 32-34.

Bell, G., Hey, T. & Szalay, A. (2009). Beyond the data deluge. Science, 323: 1297-1298.

Berman, F., Lavoie, B., Ayris, P., Choudhury, G. S., Cohen, E., Courant, P., Dirks, L., Friedlander, A., Gurbaxani, V., Jones, A., Kerr, A. U., Lynch, C. A., Rubinfeld, D., Rusbridge, C., Schonfeld, R., Smith-Rumsey, A. & Van Camp, A. (2010). Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Retrieved from http://brtf.sdsc.edu/publications.html on 5 May 2010.

Berman, H. M., Westbrook, J., Feng, J., Gilliland, G., Bhat, T. N., Wessig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Research, 28: 235-242.

Bits of Power: Issues in Global Access to Scientific Data. (1997). Washington, DC: National Academy Press. Retrieved from http://www.nap.edu on 28 September 2006.

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

Borgman, C. L. (2009). The digital future is now: A call to action for the humanities. Digital Humanities Quarterly, 3(4). Retrieved from http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html on 14 April 2010.

Borgman, C. L. (2010). Research Data: Who will share what, with whom, when, and why? China-North America Library Conference, Beijing. Retrieved from http://works.bepress.com/borgman/238 and http://www.nlc.gov.cn/yjfw/zm/index_en.html on 17 September 2010.

Borgman, C. L., Bowker, G. C., Finholt, T. A. & Wallis, J. C. (2009). Towards a Virtual Organization for Data Cyberinfrastructure. Joint Conference on Digital Libraries, Austin, TX, ACM.

Borgman, C. L., Wallis, J. C. & Enyedy, N. (2006). Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. 10th European Conference on Digital Libraries, Alicante, Spain, Berlin: Springer. 170-183.

Borgman, C. L., Wallis, J. C. & Enyedy, N. (2007). Little Science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries, 7(1-2): 17-3029 September 2007.

Borgman, C. L., Wallis, J. C., Mayernik, M. S. & Pepe, A. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. Joint Conference on Digital Libraries, Vancouver, British Columbia, Canada, Association for Computing Machinery. 269-277. Retrieved from http://doi.acm.org/10.1145/1255175.1255228 on 1 February 2010.

Bowker, G. C. (2000). Biodiversity datadiversity. Social Studies of Science, 30(5): 643-683.

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

Brumfiel, G. (2002). Misconduct finding at Bell Labs shakes physics community. Nature, 419(Oct 3 News): 419-421.

Buckland, M. K. (1991). Information as thing. Journal of the American Society for Information Science, 42(5): 351-360.

Buneman, P., Khanna, S. & Tan, W.-C. (2000). Data provenance: Some basic issues. Foundations of software technology and theoretical computer science: Proceedings of the 20th Conference, Berlin, Springer. 87-93.

Butler, D. (2006). Mashups mix data into global service: Is this the future for scientific analysis? Nature, 439(7072): 6-7.

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A. & Blumenthal, D. (2002). Data withholding in academic genetics: Evidence from a national survey. Journal of the American Medical Association, 287(4): 473-480. Retrieved from http://jama.ama-assn.org/cgi/content/full/287/4/473 on 29 December 2009.

Claerbout, J. (2010). Reproducible computational research: A history of hurdles, mostly overcome. Retrieved from http://sepwww.stanford.edu/sep/jon/reproducible.html on 10 August 2010.

Community cleverness required. (2008). Nature, 455(7209): 1-1.

Costello, A., Maslin, M., Montgomery, H., Johnson, A. M. & Ekins, P. (2011). Global health and climate change: moving from denial and catastrophic fatalism to positive action. Philosophical Transactions of the Royal Society A 369: 1866-188210 June 2011.

Couzin, J. & Unger, C. (2006). Cleaning up the paper trail. Science, 312: 38-43.

Couzin-Frankel, J. (2010). As Questions Grow, Duke Halts Trials, Launches Investigation. Science, 329: 614-615.

Crow, R. (2009). Income models for open access: An overview of current practice. The Scholarly Publishing & Academic Resources Coalition. Retrieved from http://www.arl.org/sparc/publications/papers/imguide.shtml on 17 September 2010.

Dalrymple, D. (2003). Scientific knowledge as a global public good: Contributions to innovation and the economy. In Esanu, J. M. & Uhlir, P. F. (Eds.). The Role of Scientific and Technical Data and Information in the Public Domain. Washington,

DC, The National Academies Press: 35-51. Retrieved from http://books.nap.edu/catalog/10785.html on 6 October 2006.

Data Conservancy. (2010). Johns Hopkins University. Retrieved from http://www.dataconservancy.org/home on 10 August 2010.

Data, Data Everywhere. (2010). Economist: 16.

Data's shameful neglect. (2009). Nature, 461(7261): 145-145.

David, P. A. (2004). Towards a cyberinfrastructure for enhanced scientific collaboration: Providing its 'soft' foundations may be the hardest part. Oxford Internet Institute Research Reports: University of Oxford. 23. Retrieved from http://www.oii.ox.ac.uk on 18 April 2006.

DCC Data Management Plans. (2011). Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/data-management-plans on 12 April 2011.

Dealing with data. (2011). Science, 331(6018): 692-729.

Directory of Open Access Journals. (2009). Open Society Initiative, Scholarly Publishing and Academic Resources Coalition. Retrieved from http://www.doaj.org on 16 August 2009.

Drake, A. J., Djorgovski, S. G., Mahabal, A., Anderson, J., Roy, R., Mohan, V., Ravindranath, S., Frail, D., Gezari, S., Neill, J. D., C. Ho, L., Prieto, J. L., Thompson, D., Thorstensen, J., Wagner, M., Kowalski, R., Chiang, J., Grove, J. E., Schinzel, F. K., Wood, D. L., Carrasco, L., Recillas, E., Kewley, L., Archana, K. N., Basu, A., Wadadekar, Y., Kumar, B., Myers, A. D., Phinney, E. S., Williams, R., Graham, M. J., Catelan, M., Beshore, E., Larson, S. & Christensen, E. (2011). The Discovery and Nature of Optical Transient CSS100217:102913+404220. arXiv, astro-ph.CO. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=1103.5514v1 on 18 April 2011.

Edwards, P. N. (2010). A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Cambridge, MA: MIT Press. Retrieved from http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=12080 on 9 August 2010.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C. & Borgman, C. L. (2011, forthcoming). Science Friction: Data, Metadata, and Collaboration. Social Studies of Science.

Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. (2009). Washington, D.C.: National Academy Press. Retrieved from http://www.nap.edu/ on 4 January 2010.

Esanu, J. M. & Uhlir, P. F. (Eds.). (2003). The Role of Scientific and Technical Data and Information in the Public Domain. Washington, DC: The National Academies Press. Retrieved from http://books.nap.edu/catalog/10785.html on 30 September 2006.

Esanu, J. M. & Uhlir, P. F. (Eds.). (2004). Open Access and the Public Domain in Digital Data and Information for Science: Proceedings of an International Symposium. Washington, DC: The National Academies Press. Retrieved from http://books.nap.edu/catalog/11030.html on 30 September 2006.

ESRC Research Data Policy (2010). Economic and Social Research Council. Retrieved from http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx on.

Faniel, I. M. & Jacobsen, T. E. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. Comput Supported Coop Work, 19(3-4): 355-375.

Fazackerley, A. (2004). Wellcome embraces open access future. Times Higher Education Supplement, 1665(5): 5.

Federal Research Public Access Act of 2009 (2010). Retrieved from http://thomas.loc.gov/cgi-bin/bdquery/z?d111:HR05037:@@@D&summ2=m& on Accessed.

Fienberg, S. E., Martin, M. E. & Straf, M. L. (Eds.). (1985). Sharing Research Data. Washington, DC: National Academy Press. Retrieved from http://books.nap.edu/catalog.php?record_id=2033 on 30 December 2009.

Fischer, B. A. & Zigmond, M. J. (2010). The Essential Nature of Sharing in Science. Science and Engineering Ethics, 16(4): 783-799. Retrieved from <Go to ISI>://000285672100016 on.

Genome Canada Data Release and Sharing Policy. (2005). Retrieved from www.genomecanada.ca/xcorporate/policies/DataReleasePolicy.pdf on 30 September 2006.

GEON. (2011). Retrieved from http://www.geongrid.org/ on 27 April 2011.

Gil, Y. (2010). Provenance XG Final Report. Retrieved from http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/ on 30 March 2011.

Gleick, P. H. e. a. (2011). Climate Change and the Integrity of Science. Science, 328: 689-9010 June 2011.

Goble, C. & De Roure, D. (2009). The impact of workflow tools on data-intensive research. In Hey, T., Tansley, S. & Tolle, K. (Eds.). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA, Microsoft: 137-146. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/ on 16 December 2009.

Gobler, C. J., Boneillo, G. E., Debenham, C. J. & Caron, D. A. (2004). Nutrient limitation, organic matter cycling, and plankton dynamics during an Aureococcus anophagefferens bloom. Aquatic Microbial Ecolology, 35: 31-43.

Grant Policy Manual. (2001). National Science Foundation. Retrieved from http://www.nsf.gov/publications/ on 5 July 2006.

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. & Heber, G. (2005). Scientific data management in the coming decade. CT Watch Quarterly, 1(1). Retrieved from http://www.ctwatch.org/quarterly/articles/2005/02/scientific-data-management/ on 25 August 2006.

Haeussler, C. (2011). Information-sharing in academia and the industry: A comparative study. Research Policy, 40(1): 105-122. Retrieved from <Go to ISI>://000286910500010 on 15 April 2011.

Hanson, B., Sugden, A. & Alberts, B. (2011). Making data maximally available. Science, 331(6018): 649. Retrieved from <Go to ISI>://000287205700050 on.

Harnessing the power of digital data for science and society (2009). Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. Retrieved from http://www.nitrd.gov/about/Harnessing_Power_Web.pdf on 10 March 2009.

Hey, A. J. G. & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective. In Berman, F., Fox, G. & Hey, A. J. G. (Eds.). Grid Computing: Making the Global Infrastructure a Reality. Chichester, Wiley. Retrieved from http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf on 20 January 2005.

Hey, T., Tansley, S. & Tolle, K. (Eds.). (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/ on 16 December 2009.

Hilgartner, S. (1997). Access to Data and Intellectual Property: Scientific Exchange in Genome Research. In. Intellectual Property Rights and the Dissemination of Research Tools in Molecular Biology. Summary of a Workshop Held at the National Academy of Science, February 15-16, 1996. Washington, D.C., National Academy Press: 28-39.

Hilgartner, S. (1998). Data Access Policy in Genome Research. In Thakray, A. (Ed.). Private Science. Oxford, Oxford University Press: 202-218.

Hilgartner, S. (2002). Acceptable intellectual property. Journal of Molecular Biology, 319(4): 943-946. Retrieved from <Go to ISI>://000176505800012 on.

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

Hunter, J. & Cheung, K. (2007). Provenance Explorer-a graphical interface for constructing scientific publication packages from provenance trails. International Journal on Digital Libraries, 7(1): 99-107.

Kaiser, J. (2008). Scientific publishing - Uncle Sam's biomedical archive wants your papers. Science, 319(5861): 266-26610 March 2009.

Kanfer, A. G., Haythornthwaite, C., Bruce, B. C., Bowker, G. C., Burbules, N. C., Porac, J. F. & Wade, J. (2000). Modeling distributed knowledge processes in next generation multidisciplinary alliances. Information Systems Frontiers, 2(3-4): 317-331.

Kansa, E. C., Kansa, S. W., Burton, M. M. & Stankowski, C. (2010). Googling the grey: Open data, web services, and semantics. Archaeologies-Journal of the World Archaeological Congress, 6(2): 301-32615 June 2011.

Karasti, H., Baker, K. S. & Halkola, E. (2006). Enriching the notion of data curation in e-Science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) Network. Journal of Computer Supported Cooperative Work, 15(4): 321-358.

Karasti, H., Baker, K. S. & Millerand, F. (2010). Infrastructure Time: Long-term Matters in Collaborative Development. Comput Supported Coop Work, 19(3-4): 377-415.

Kelty, C. M. (2008). Two bits: the cultural significance of free software. Durham, NC: Duke University Press.

Knorr-Cetina, K. (1999). Epistemic Cultures: How the Sciences Make Knowledge. Cambridge, MA: Harvard University Press.

Kowalczyk, S. & Shankar, K. (2011). Data sharing in the sciences. In Cronin, B. (Ed.). Annual Review of Information Science and Technology. Medford, NJ, Information Today. 45: 247-294.

Lagoze, C. & Velden, T. (2009a). Communicating chemistry. Nature Chemistry, 1: 673 - 678. Retrieved from http://www.nature.com/nchem/journal/v1/n9/full/nchem.448.html on 30 December 2009.

Lagoze, C. & Velden, T. (2009b). The Value of New Scientific Communication Models for Chemistry. 1-71. Retrieved from http://ecommons.cornell.edu/handle/1813/14150 on 17 August 2010.

Large Synoptic Sky Telescope. (2010). Retrieved from http://www.lsst.org/lsst on 9 August 2010.

Latour, B. (1987). Science in Action: How to Follow Scientists and Engineers through Society. Cambridge, MA: Harvard University Press.

Latour, B. & Woolgar, S. (1979). Laboratory life: The Social Construction of Scientific Facts. Beverly Hills: Sage Publications.

Lave, J. & Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge, UK: Cambridge University Press.

Long-Lived Digital Data Collections. (2005). National Science Board. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/ on 18 April 2009.

Lyon, L. (2007). Dealing with data: Roles, rights, responsibilities, and relationships. UKOLN. Retrieved from http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_dealing_with_data.aspx on 23 July 2007.

Mayernik, M. S. (2011). Metadata Realities for Cyberinfrastructure: Data Authors as Metadata CreatorsInformation Studies. UCLA. Los Angeles. Retrieved from http://beta.sensorbase.org/~mayernik/mayernik_dissertation_submitted_08June2011.pdf on 15 June 2011.

Mayernik, M. S., Batcheller, A. L. & Borgman, C. L. (2011). How Institutional Factors Influence the Creation of Scientific Metadata. iConference, Seattle, Association for Computing Machinery.

Meng, X.-L. (2010). Multi-party inference and uncongeniality. Berlin, Springer-VerlagAccessed.

Merriam-Webster's Collegiate Dictionary (1993). Springfield, MA, Merriam-WebsterAccessed.

Merton, R. K. (1969). Behavior patterns of scientists. American Scientist, 57(1): 1-23.

Merton, R. K. (1973). The normative structure of science. In Storer, N. W. (Ed.). The Sociology of Science: Theoretical and Empirical Investigations. Chicago, University of Chicago Press: 267-278.

Moore, A. J., McPeek, M. A., Rausher, M. D., Rieseberg, L. & Whitlock, M. C. (2010). The need for archiving data in evolutionary biology. Journal of Evolutionary Biology, 23(4): 659-660. Retrieved from <Go to ISI>://000275761400001 on.

Murray-Rust, P. & Rzepa, H. S. (2004). The next big thing: From hypermedia to datuments. Journal of Digital Information, 5(1): Article No. 248. Retrieved from http://journals.tdl.org/jodi/article/view/130 on 28 December 2009.

National Ecological Observatory Network. (2010). Retrieved from http://www.neoninc.org/ on 20 August 2010.

NIH Public Access Policy. (2005). National Institutes of Health. Retrieved from http://publicaccess.nih.gov/publicaccess_manual.htm on 28 March 2006.

Normile, D., Vogel, G. & Couzin, J. (2006). Cloning - South Korean team's remaining human stem cell claim demolished. Science, 311(5758): 156-157.

NSF Data Management Plans (2010). National Science Foundation. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp on 11 April 2011.

NSF Data Sharing Policy (2010). National Science Foundation. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4 on 18 October 2010.

NSF Proposal Preparation Instructions (2011). Award and Administrative Guide: National Science Foundation. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp on 18 October 2010.

OECD Principles and Guidelines for Access to Research Data from Public Funding (2007). 1-24. Retrieved from www.oecd.org/dataoecd/9/61/38500813.pdf on 4 January 2010.

Olson, G. M., Zimmerman, A. & Bos, N. (Eds.). (2008). Scientific Collaboration on the Internet. Cambridge, MA: MIT Press.

Open Content Alliance. (2009). Retrieved from http://www.opencontentalliance.org/ on 16 August 2009.

Osterlund, C. & Carlile, P. (2005). Relations in practice: Sorting through practice theories on knowledge sharing in complex organizations. The Information Society, 21(2): 91-107.

Overpeck, J. T., Meehl, G. A., Bony, S. & Easterling, D. R. (2011). Climate Data Challenges in the 21st Century. Science, 331(6018): 700-702. Retrieved from <Go to ISI>://WOS:000287205700049 on.

Palmer, C. L. (2005). Scholarly work and the shaping of digital access. Journal of the American Society for Information Science and Technology, 56(11): 1140-1153.

PAN-STARRS. (2009). Panoramic Survey Telescope & Rapid Response System. Retrieved from http://pan-starrs.ifa.hawaii.edu/public/ on 14 September 2009.

Piwowar, H. A., Becich, M. J., Bilofsky, H. & Crowley, R. S. (2008). Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers. Plos Medicine, 5(9): 1315-1319.

Piwowar, H. A. & Chapman, W. W. (2010). Public sharing of research datasets: A pilot study of associations. Journal of Informetrics, 4(2): 148-156. Retrieved from <Go to ISI>://000275515500003 on.

Piwowar, H. A., Day, R. S. & Fridsma, D. B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. Plos One, 2(3). Retrieved from <Go to ISI>://000207445100010 on.

Porter, J. H. (2010). A Brief History of Data Sharing in the U.S. Long Term Ecological Research Network. Bulletin of the Ecological Society of America, 91: 14-20. Retrieved from http://dx.doi.org/10.1890/0012-9623-91.1.14 on.

Preserving Scientific Data on Our Physical Universe. (1995). Washington, D.C.: National Academy Press. Retrieved from http://www.nap.edu/catalog.php?record_id=4871 on 4 January 2010.

Pritchard, S. M., Carver, L. & Anand, S. (2004). Collaboration for knowledge management and campus informatics. University of California, Santa Barbara. 38. Retrieved from http://www.library.ucsb.edu/informatics/informatics/documents/UCSB_Campus_Informatics_Project_Report.pdf on 5 July 2006.

Protein Data Bank. (2011). Retrieved from http://www.rcsb.org/pdb/ on 29 April 2011.

A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases. (1999). Washington, DC: National Academy Press. Retrieved from http://www.nap.edu on 28 September 2006.

Reference Model for an Open Archival Information System (2002). Recommendation for Space Data System Standards: Consultative Committee for Space Data Systems Secretariat, Program Integration Division (Code M-3), National Aeronautics and Space Administration. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf on 4 October 2006.

Reichman, J. H. & Uhlir, P. F. (2003). A contractually reconstructed research commons for scientific data in a highly protectionist intellectual property environment. Law and Contemporary Problems, 66(1&2): 315-462.

Renear, A. H. & Palmer, C. L. (2009). Strategic Reading, Ontologies, and the Future of Scientific Publishing. Science, 325(5942): 828 - 832.

Renear, A. H., Sacchi, S. & Wickett, K. M. (2010). Definitions of Dataset in the Scientific and Technical Literature. American Society for Information Science and Technology, Pittsburgh, Information Today. 1-4. Retrieved from http://portal.acm.org/citation.cfm?id=1920447 on 21 June 2011.

Reproducible research: Addressing the need for data and code sharing in computational science. (2010). Computing in Science & Engineering: 8-12.

Ribes, D., Baker, K. S., Millerand, F. & Bowker, G. C. (2005). Comparative Interoperability Project: Configurations of Community, Technology, Organization. Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries.

Ribes, D. & Bowker, G. C. (2008). Organizing for multidisciplinary collaboration: The case of the Geosciences Network. In Olson, G. M., Zimmerman, A. & Bos, N. (Eds.). Science on the Internet. Cambridge, MIT Press.

Ribes, D. & Finholt, T. A. (2007). Tensions across the scales: Planning infrastructure for the long-term. Proceedings of the 2007 International ACM SIGGROUP Conference on Supporting Group Work, Sanibel Island, Florida, USA, Sanibel Island, Florida, Association for Computing Machinery. 229-238.

Rogers, E. M. (1995). Diffusion of Innovations (4th ed.). New York: The Free Press.

Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. (2003). Meeting organized by the Wellcome Trust, Fort Lauderdale, Florida, Wellcome Trust. Retrieved from www.wellcome.ac.uk/.../groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf on 29 December 2009.

Sloan Digital Sky Survey. (2010). Retrieved from http://www.sdss.org/ on 9 August 2010.

Stodden, V. (2009a). Enabling reproducible research: Open licensing for scientific innovation. 1-55. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1362040 on 17 August 2010.

Stodden, V. (2009b). The Legal Framework for Reproducible Scientific Research: Licensing and Copyright. Computing in Science and Engineering, 11(1): 35-4017 April 2009.

Summary of principles. (1996). International Strategy Meeting on Human Genome Sequencing, Bermuda, The Wellcome Trust. Retrieved from http://www.gene.ucl.ac.uk/hugo/bermuda.htm on 17 December 2005.

Sustainable Digital Data Preservation and Access Network Partners (DataNet). (2010). National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm on 11 August 2010.

U.S. Long Term Ecological Research Network. (2010). Retrieved from http://lternet.edu/ on 20 August 2010.

Uhlir, P. F. & Cohen, D. (2011). Personal communication. Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences. 18 March 2011.

The University's Role in the Dissemination of Research and Scholarship (2009). Association of Research Libraries. 1-8. Retrieved from www.arl.org/disseminating_research_2009 on 10 March 2009.

Unsworth, J., Courant, P., Fraser, S., Goodchild, M., Hedstrom, M., Henry, C., Kaufman, P. B., McGann, J., Rosenzweig, R. & Zuckerman, B. (2006). Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for Humanities and Social Sciences. American Council of Learned Societies. Retrieved from http://www.acls.org/cyberinfrastructure/cyber.htm on 17 July 2007.

Van House, N. A. (2004). Science and technology studies and information studies. In Cronin, B. (Ed.). Annual Review of Information Science and Technology. Medford, NJ, Information Today. 38: 3-86.

Vandewalle, P., Kovacevic, J. & Vetterli, M. (2009). Reproducible Research in Signal Processing. IEEE Signal Processing Magazine, 26(3): 37-47.

Wallis, J. C., Mayernik, M. S., Borgman, C. L. & Pepe, A. (2010). Digital Libraries for Scientific Data Discovery and Reuse: From Vision to Practical Reality. Joint Conference on Digital Libraries, Gold Coast, Queensland, Australia, Association for Computing Machinery.

Ware, M. (2010). Submission fees - a tool in the transition to open access? . 1-13.

Wellcome Trust Policy on Access to Bioinformatics Resources by Trust-Funded Researchers. (2001). Wellcome Trust. Retrieved from http://www.wellcome.ac.uk/doc%5Fwtd002759.html on 5 October 2006.

Wellcome Trust position statement in support of open and unrestricted access to published research. (2005). Wellcome Trust. Retrieved from http://www.wellcome.ac.uk/doc_WTD002766.html on 5 October 2006.

Wellcome Trust statement on genome data release. (1997). Retrieved from http://www.wellcome.ac.uk/doc%5Fwtd002751.html on 5 October 2006.

Wenger, E. (1998). Communities of Practice: Learning, Meaning, and Identity. Cambridge, UK: Cambridge University Press.

Whitlock, M. C. (2011). Data archiving in ecology and evolution: best practices. Trends in Ecology & Evolution, 26(2): 61-65. Retrieved from <Go to ISI>://000287056100006 on.

Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L. & Moore, A. J. (2010). Data Archiving. American Naturalist, 175(2): E45-146. Retrieved from <Go to ISI>://000273650200003 on.

Wilbanks, J. (2009). I have seen the paradigm shift and it is us. In Hey, T., Tansley, S. & Tolle, K. (Eds.). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA, Microsoft: 209-214. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/ on 16 December 2009.

Witt, M., Carlson, J., Brandt, D. S. & Cragin, M. H. (2009). Constructing data curation profiles. International Journal of Digital Curation, 4(3). Retrieved from http://www.ijdc.net/index.php/ijdc/article/viewFile/137/165 on 7 June 2011.

Wynholds, L. (2010). Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects. Digital Curation Conference, Chicago. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/174 on 15 June 2011.

Wynholds, L., Fearon Jr, D. S., Borgman, C. L. & Traweek, S. (2011). When use cases are not useful: Data practices, astronomy, and digital libraries Joint Conference on Digital Libraries, Ottawa, ACM.

Young, J. R. (2009). Physicists set plan in motion to change publishing system. Chronicle of Higher Education, 55(21): A1-. Retrieved from http://chronicle.com/free/v55/i21/21a00104.htm on 10 March 2009.

Zimmerman, A. S. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. International Journal of Digital Libraries, 7(1-2): 5-16.