Research Data:
Who will share what, with whom, when, and why?

Christine L. Borgman
Professor & Presidential Chair in Information Studies
University of California, Los Angeles

Fifth China – North America Library Conference 2010
8-12 September 2010
Beijing

Research Data Sharing Theme Speakers (North America):
Christine L. Borgman, UCLA Information Studies
Lee Dirks, Microsoft Scholarly Communication Program

Table of Contents

ABSTRACT

The deluge of scientific research data has excited the general public, as well as the scientific community, with the possibilities for better understanding of scientific problems, from climate to culture. For data to be available, researchers must be willing and able to share them. The policies of governments, funding agencies, journals, and university tenure and promotion committees also influence how, when, and whether research data are shared. Data are complex objects. Their purposes and the methods

by which they are produced vary widely across scientific fields, as do the criteria for sharing them. To address these challenges, it is necessary to examine the arguments for sharing data and how those arguments match the motivations and interests of the scientific community and the public. Four arguments are examined: to make the results of publicly funded data available to the public, to enable others to ask new questions of extant data, to advance the state of science, and to reproduce research. Libraries need to consider their role in the face of each of these arguments, and what expertise and systems they require for data curation.

INTRODUCTION

The data deluge has arrived. Long predicted by the science community [1], the popular press is now heralding the wide availability of data for use by anyone, anywhere. Not only has *Nature*, a premier science journal, published feature sections on "big data" [2; 3], so have *WIRED* magazine [4], and the *Economist* [5]. Libraries are responding with reports of their own, assessing what actions they can, should, and should not pursue [6; 7].

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 astroinformatics and computational biology [8]. Domain scientists and computer scientists work closely together in many fields, with varying degrees of tension over their mutual research interests.

Along with the proliferation of data has come the concern for how those data are to be captured, curated, and maintained for future use. Libraries have become the most likely institution to host data, a responsibility they are approaching with some apprehension. Data are different objects than books and journals, in ways both obvious and subtle. Appropriate economic and institutional models for data curation are far from apparent [9].

Unstated assumptions about what are data and why they should be shared underlie the library's concerns about data curation, researchers' concerns about data release and access, and public policies for data sharing. This short paper explores some of those assumptions, focusing on the implications for library roles in data curation and for the design of curatorial systems. In considering what roles libraries might play in data curation – a theme of this conference – it is useful to explore the perspectives of the scientists conducting research that results in the data that libraries may curate.

WHAT ARE DATA?

All too rarely do those promoting the sharing and curation of data define "data" explicitly or acknowledge the diversity of forms that data may take. The definition established in a National Research Council report suggests the complexity of the

concept: "*Data* are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors." [10, p. 15]. The notion of "data" can vary considerably among collaborators [11], and even more so between disciplines.

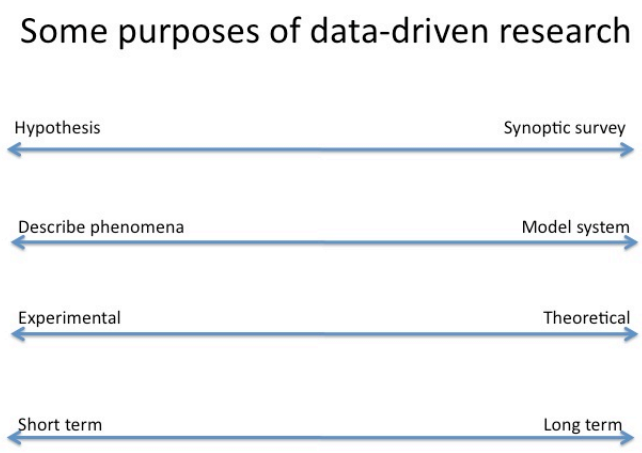
Some types of data have both immediate and enduring value, some gain value over time, some have transient value, and yet others are easier to recreate than to curate [12; 13; 14]. Many of these distinctions depend upon the category of data, as identified in an influential National Science Foundation report [15]: 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 sciences, which are the subject of the long-lived data report, exemplify all of these categories. While useful as a general framework, these four categories tend to obscure the many kinds of data that may be collected in any given scholarly endeavor [16].

Investigators collect data for many purposes, using many methods. Those purposes and methods influence both what they consider to be their "data," and the conditions under which they are willing to share those data with others. The criteria for identifying data and for sharing are not well understood yet. Understanding practices, problems, and policies for data is an expanding area of research in the fields of information studies and social studies of science. Research to date is largely characterized by case studies in individual disciplines [17; 18; 19; 16; 20; 21; 22; 23; 24; 25].

Purposes of Data-Driven Research

Among the many purposes for which data are collected, a few dimensions emerge. None of these are distinct scales, but the contrast suggests the range of possibilities (Figure 1).

Figure 1: Purposes of Data-Driven Research



The first dimension illustrated is specificity of purpose, ranging from hypothesis-driven inquiry to synoptic survey research. An investigator might be pursuing a specific question, often at a specific site, perhaps about a specific phenomenon, to test one or more hypotheses. This type of research might take place in a laboratory, in a field setting, or some combination. At the other end of this dimension are synoptic surveys. These are surveys that attempt to provide a comprehensive view of some whole entity or system, such as the earth or sky. Global climate modeling depends upon consistent data collection of climate phenomena around the world at agreed upon times, locations, and variables [26]. Synoptic sky surveys such as the Sloan Digital Sky Survey [27], Panoramic Survey Telescope and Rapid Response System (PAN-STARRS) [28], and Large Synoptic Sky Telescope (LSST) [29] are producing – or will soon produce – a deluge of astronomy data. Multiple synoptic surveys of ecological and geophysical data also exist [17]. Synoptic surveys are massive efforts that usually require substantial amounts of public or private funding. They are conducted to serve a large community and thus usually the data are made publicly available. Investigators and others can mine the data to ask their own questions or to identify bases for comparison with data from other sources.

A second dimension of purposes is the range of specificity from studies that describe particular kinds of events or phenomena to studies that model entire systems. Climate research provides examples along this dimension. Weather data, in the short term, can be used to describe or predict rain, snow, wind or other events. Those data can be combined with other kinds of data that have long-term value, and with principles of physics, to model the climate of the earth [26]. Ecology data provide a contrasting example. Researchers may be studying a specific phenomenon such as harmful algal blooms (HAB). They may collect data for months or even years to capture events before, during, and after an HAB event [30; 31]. Their goal is to understand the processes that trigger an event and how those processes evolve. It has proven difficult to aggregate the study of such biological events into comprehensive systems models of the type used in climate research [17], due largely to differences in data characteristics. Whereas the physical sciences have established constants and standard measures, biological organisms are individually distinct, requiring specialized methods and measures [13].

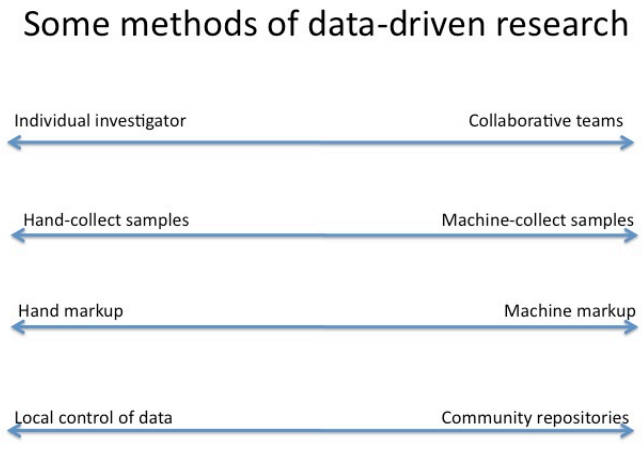
A third dimension of purposes is from experimental to theoretical research. Sometimes an investigator sets up an experiment to control some variables and to test others. While feasible in a laboratory, few opportunities exist to control the atmosphere or the universe. In the latter case, experiments can be conducted on theoretical models, which is commonly done with elaborate models such as those for climate. In theoretical research, whether climate or astronomy, data may be simulated, rather than collected from "the real world." Experiments and models can be recreated, and thus the data may not be curated for the long term. Observations of the physical universe, in contrast, occur at a unique place and time and can never be reconstructed. Thus more emphasis usually is placed on curating observations [12].

A fourth dimension is where the intended purpose of a study falls on a continuum from short-term to long-term research. A scientist may conduct a small number of experiments or field studies to explore a problem. That set of studies may or may not be part of a larger, long-term body of research. 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. Longitudinal data with consistent measurements are more valuable than are descriptions of isolated events over time.

Methods of Data-Driven Research

The methods by which data-driven research are conducted are at least as diverse as the purposes they serve. As with the dimensions of purposes outlined above, this selection of research methods is intended to be illustrative rather than exhaustive (Figure 2).

Figure 2: Methods of Data-Driven Research



The first dimension of methods ranges from the individual investigator working alone to large collaborative teams. 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 methods, and who has the rights and responsibilities to analyze, publish, and release those data [32; 21; 33].

A second dimension of methods is the degree of handling necessary to acquire data. Investigators in ecology, for example, may spend days, weeks, or months in the field collecting 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. In highly instrumented disciplines such as astronomy and high-energy physics, instruments that observe phenomena in the sky or in accelerators generate vast amounts of data. Those instruments require many years to build, and are based on long-term collaborations among scientists and technologists. Generally speaking, the more hand-crafted the data collection, the less likely that researchers will share their data [34], but practices vary so widely across fields and research teams that any such generalizations are difficult to make.

The third dimension of methods concerns the markup, or documentation of data. Rarely are data self-describing. When data are collected by hand, such as gathering 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 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. In the most highly instrumented research, such as sky surveys, instruments capture contextual information about the data. Even so, 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.

A fourth dimension of methods is the range of control of the data. The smaller and more geographically constrained the research team, the more likely they are to control all aspects of their data management. Spreadsheets often suffice for data management and analysis, especially if the number of observations and data elements is small. Spreadsheets may be the lowest common denominator among small research groups, and provide the means for data exchange within or between teams [35]. At the other end of this dimension are the large repositories necessary to manage the flood of data from telescopes, particle detectors, and other research instruments. Polices for access to data repositories in astronomy, physics, seismology, genomics and other fields vary widely. In some cases, only the teams that contribute data have access to the pool, and in others, data are open for use by anyone, immediately. Embargo periods, in which investigators have first access to the data, also vary widely.

WHY SHARE RESEARCH DATA?

As is evident from the above discussion of the purposes and methods of data-driven research, 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. They weigh a number of considerations in

determining when, how, why, and whether to share their data with others. Researchers also have disparate views about which classes of information actually are their "data."

The following discussion is largely concerned with the incentives and disincentives of scientific investigators and their research partners to share data. These considerations are somewhat different for synoptic surveys. Generally speaking, the latter are multi-year efforts to gather a critical mass of data that are intended for use by the scientific community, although some survey projects are intended for use only by the collaborators funded on that study. Even in large synoptic surveys, investigators may have embargo periods before data are released to other scientists.

Researchers' Incentives and Disincentives to Share Data

Incentives for researchers to share their data include the ethos of open science and peer review; the value of collaborating with others, for which data may be the "glue;" benefits to reputation; and reciprocity. Depositing one's data may be a condition of gaining access to the data of others, and of access to useful tools for analysis and management. Coercion may also play a role: some funding agencies or individual grant contracts may require data contribution as a condition for funding.

Researchers also have multiple incentives *not* to share their data. In most fields, the rewards come from publication, not from data management. Scholars are hired and promoted based on their publication record rather than on the quality of their metadata. Secondly, documenting data is a labor-intensive process even for local use. Documenting methods, instrumentation, and software, and producing metadata at a level that the data are interpretable by others, can require much more labor than documentation for use by oneself or one's team. Thirdly, researchers are concerned about establishing the priority of their claims on research findings in the face of competition. Embargo periods, where they exist, protect the investigator by providing a period of time to analyze data and publish results prior to the public release of their data. Lastly is the set of concerns for intellectual property, both the ability to control one's own resources and the ability to gain access to resources controlled by others [16].

Policy Arguments for Sharing Research Data

The policy motivations for sharing research data are many, but rarely are they connected explicitly to the incentives of researchers to share their data with others or to the reasons for which libraries might curate data. Policy arguments for sharing data – some explicit and some implicit – vary along two dimensions, as presented in Figure 3.

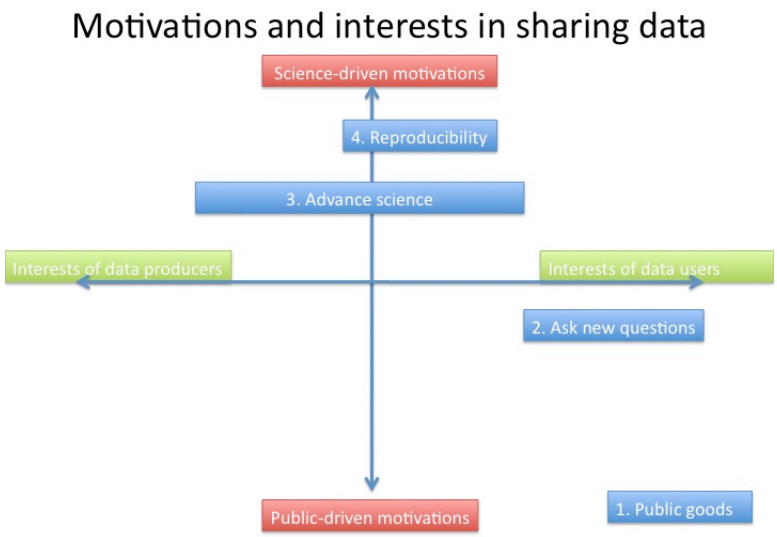


Figure 3: Motivations and interests in sharing data

On the vertical axis is the motivation for sharing, either toward the value of data for the public (bottom) or toward the value of the data for science (top). On the horizontal axis is the locus of the interests served, either toward the producers of data (left) or the prospective users of the data (right). Neither dimension is absolute; the poles represent relative positions of people or situations. For example, a scientist 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 science may also benefit the public good.

While these may seem to be subtle distinctions, the arguments can lead to markedly different levels of data sharing by the scientific community and to very different policy and design models by the library community. Four arguments for sharing research data are introduced, in the inverse order in which they resonate with the scientific community. These are general statements made in the interest of provoking discussion among librarians and policy makers, as a detailed treatment of these issues would require a book-length discourse. National and international policies for sharing data will be addressed in another session at this conference.

1. To make the results of publicly funded data available to the public

The most general argument in favor of sharing research data applies to research that was conducted with public funds. If taxpayer monies produced the research, then the taxpayers should have access to the results. U.S. public policy tends more in this direction than most other countries; the law waives copyright protection on data and information directly produced by government agencies, putting those materials into the public domain [36]. Data and information resulting from research grants to universities and other research agencies do not fall under the same law. However, the public monies for public good argument is the motivation for depositing publications arising from funding by the U.S. National Institutes of Health (NIH) into PubMed Central, for example [37]. The Wellcome Trust, which is the largest funder of biomedical research in the United Kingdom, has similar policies [38; 39]. Both NIH and the Wellcome Trust also have policies about depositing or making available the data from funded research, at least for grants over a certain size [40; 41; 42]. Biomedical data and publications have a

substantial audience, including biomedical researchers, clinicians, the pharmaceutical industry, and patients.

The public monies-public goods argument has succeeded in the biomedical research community for the deposit of publications; but not without resistance, especially on the part of publishers. The U.S. National Science Foundation has discussed similar policies for its grants for many years, but has not yet implemented a general policy. The NSF grant policy manual contains statements that investigators are expected to make their data available upon request [43]. However, these statements are difficult to enforce. They do not contain specific language defining what "data" are to be released, for example. Given the plethora of physical and intellectual objects that might be considered data, and the amount of contextual information required to make sense of them, a uniform policy across an agency that supports research in many disciplines, as NSF does, may be infeasible. Multiple policies, adapted to specific disciplinary units within NSF, are under discussion.

The public monies for public good argument for making research data available is too general to gain strong support in the scientific community. This argument is positioned in the lower right quadrant of Figure 3, as one driven by the interests of the general public and potential data users. Arguments that emphasize benefits to the scientific community are gaining more traction.

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

A more focused argument is that sharing data enables others to ask new questions, whether from an individual dataset or by combining multiple sources. This is a user-driven argument that often ignores the incentives for the producers of those data to release them. Given the investment in acquiring those data and the additional effort necessary to make them useful to others, researchers often ask why they should release their data without some specific benefit to themselves. At a minimum, most researchers want attribution for any data used by others.

WIRED magazine [4], in its enthusiasm for the many potential uses of publicly available scientific data, went so far as to proclaim "the end of theory" – suggesting that the scientific method can be abandoned in favor of data mining. This is a naïve argument that ignores both the complexity of data and the amount of expertise required to interpret them. Data are of little value without adequate data description and documentation of associated context information. Considerable expertise in the research domain is required to interpret data accurately and reliably. Access to data does not a scientist make. Among the most common reasons that scientists give for not sharing their data are concerns that their data will be misinterpreted, misused, or misappropriated without credit [16].

The argument that data should be made available so that others can use them is more science-driven than the general notion of public monies for public goods, in that science

questions are articulated. Hence the argument is placed higher in the public-goods motivation / user-driven interests quadrant of Figure 3 than the first argument.

3. To advance the state of science

A much stronger argument for sharing data is one that focuses on how science can be advanced by sharing data. This is the "fourth paradigm" argument: computational science constitutes a new set of methods beyond empiricism, theory, and simulation [44; 45; 8]. Data are not a method per se, but are a rich resource for any of the empirical, theoretical, simulation, or computational paradigms [46].

Scientists – rather than policy makers and journalists – are arguing for the benefits of sharing data to achieve critical mass. These arguments appear to resonate more in data-intensive fields that benefit from synoptic surveys, such as astronomy, 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 others' data more readily. For example, some sky-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.

Data-intensive fields where data have high monetary value and much of the research is privately funded, such as chemistry, are far less inclined to share their data [47; 48]. Establishing open data and metadata standards for chemistry has been highly contentious in comparison to other scientific fields [49]. Chemists using public funds to conduct research in academe thus may find themselves at a disadvantage to other scientific fields in terms of access to data.

Data sharing to advance the state of science is placed in the upper half of Figure 3, spanning the interests of research data producers and users. This argument is the cornerstone for the Data Conservancy, which is one of the two consortia initially funded by the National Science Foundation in the DataNet Program [50; 51]:

> 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 Data Conservancy will research, design, implement, deploy and sustain data curation infrastructure for cross-disciplinary discovery with an emphasis on observational data.

Data curation thus is viewed as a *means* to support science rather than an end in itself. Our work on astronomers and astronomy data practices is part of this effort; our partners are studying data practices in an array of bioinformatics and physical science fields [52].

Although we are in the early stages of this research project, we can begin to characterize how data curation and data sharing may serve as a means to advance science. One aspect of this approach is that the design of systems should be driven by scientific needs rather than by data management per se. Scientists want the capability to explore data repositories, as well as to make specific queries on structured data. A related matter is the need for tools and services that support scientific inquiry.

Our interviews to date on this project suggest that most astronomers will contribute their data in return for access to the collective resources, especially if they have the funding to do so. Some astronomers have noted that their data receives heavy use when they become available, which demonstrates their value to their funding agencies. Embargo practices in astronomy vary considerably. In some cases, investigators have exclusive control of their data for a proprietary period of one year or more. In other projects, data are released to the community within a few hours of their capture by a telescope. While those investigators may not have a time advantage, they still gain advantages by being most familiar with the instruments.

The advancing science argument for sharing data is placed in the science-driven half of Figure 3, spanning the interests of science-data producers and science-data users.

4. To reproduce research

A narrower argument for sharing research data, but a very important one, is the need to reproduce research results. 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 not 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 [53; 54]. Yet whenever papers are withdrawn from major journals, questions are raised about what the reviewers knew – or should have known – about the data and procedures [55; 56; 57; 58].

The cyberinfrastructure [59] that supports today's distributed, data-intensive, information-intensive, collaborative research has the potential to support reproducible results. To do so, published articles and papers can be linked directly to the data on which they are based. Reproducibility usually requires access to the software associated with research instruments and data analysis. Technical standards now exist to establish links between related scholarly objects such as journal articles, data, and code [60; 61]. In principle, reproducibility could be achieved by linking all data and documentation related to a specific publication, thus achieving a long-sought scientific goal [62; 63; 64; 65].

However, the goal of reproducibility remains elusive for several reasons. One reason is that the software used to collect or analyze data rarely maintains a precise record of the systems at each transaction [66]. Data analysis often consists of selecting a series of parameters (e.g., using radio buttons on a screen) on many statistical runs. Given the complexity of the data, models, software, and associated information, reconstructing computation-based research precisely is difficult. Workflow and pipeline systems that call various other programs are used for managing data and analysis in some fields, but even these programs provide less than perfect records of data provenance, i.e., each state in which the data existed [67].

Another major hurdle is legal requirements. For research to be fully reproducible, a licensing regime is needed to provide access to proprietary software and to data [64]. Recent changes in copyright law, especially in Europe where proprietary claims can be made on factual matters, make legal access to digital data even more problematic [36]. Yet another barrier is the practices of scholarly journals and conferences, and their ability – or willingness – to host data or to provide links to sources. The permanence of such hosting and linking is also an issue [68].

Reproducibility, while often mentioned as a motivation to share data, is actually a narrow argument. Reproducibility involves the ability to reconstruct the products and processes associated with a specific publication, whereas the argument for advancing science is more concerned with access to large repositories of data. The reproducibility argument is placed higher on the science-driven axis in Figure 3, as it is more strongly focused on scientific matters than on the public good. To the degree that the research in a publication is reproducible, authors' claims would be unassailable. Data users' interests would be served by the verification of claims on which they could build. Given the burden on authors to document their work to this level – which is not always possible – this argument is classified as more user-driven than producer-driven.

LIBRARY INTERESTS IN SHARING RESEARCH DATA

Libraries need to assess the relevance of each of these policy arguments for their own situations. Their choices will influence the expertise required and the services to be provided. Discipline-specific data repositories operated by funding agencies such as NASA or by consortia such as those supporting the Protein Databank face different sets of issues than those of libraries, which are the focus of this conference.

Policy responses

Curating research data is primarily a concern of research libraries. This class of libraries serves a constituency of researchers, thus the latter two arguments are most likely to apply to their situations. National and other governmental libraries may need to respond to this full spectrum of arguments, and more. Their selection, collection, and curation policies may be at least as complex as those of university research libraries.

If data curation is viewed as a means to advance science – the third argument – then libraries need to partner closely with investigators in the sciences and in other disciplines they serve. Because data vary so much by field, and by investigator, generic approaches to data collection are not feasible. Collection development policies may differ between academic departments or even between projects.

Scientists need access to data and to associated tools and services. In distributed cyberinfrastructure environments, they may be indifferent to the location of data, tools, or services. Libraries may play a role in the development of metadata, ontologies, and tools, in methods of tracking provenance, and in establishing policies for data deposit and access. These responsibilities may be separable from those of managing the data per se.

Libraries should be particularly supportive of reproducibility because this notion is central to the scholarly record. Libraries, as memory institutions, have shouldered the burden of maintaining the scholarly record. Sustaining the continuity of scholarship has become ever more difficult as the array of publication types and venues proliferates.

Reproducibility would require a radical extension of cataloging and indexing to include the full network of associated objects. Reproducibility requires elaborate metadata relationship structures, far beyond current practices such as FRBR [69]. Library licensing practices would need to be expanded radically to address the rights issues associated with access to related materials.

Expertise and Services

Libraries as institutions and librarians as information professionals bring a variety of essential capabilities to the problems of research data sharing, curation, and access. These capabilities include expertise in metadata, provenance, licensing, intellectual property, curation, information retrieval, scholarly communication, and publishing. While expertise developed for managing published materials does not translate directly to managing data – a very different type of content – expertise can be adapted, again through partnerships and study. Educating the next generation of librarians and information professionals in data practices, management, and curation also is essential. New courses and curricula on data issues are being taught at several schools of information, including UCLA, Illinois, and North Carolina.

The role of libraries in research institutions is evolving from a focus on reader services to a focus on author services (an insight first voiced by Kimberly Douglas of Caltech). The processes associated with gathering or producing research data are a form of authorship, whether or not the researchers accept that view [70]. Faculty, students, post-doctoral fellows, and other research staff are authors of data and of publications. The present and future models of publishing require authors to handle much of the pipeline themselves, not only research and writing, but also formatting, metadata, posting, linking, and submission processes. Libraries are in a position to assist their communities in all of these activities, part of which will be assisting them in sharing their

data in usable forms. Whether or not the libraries hold the data of the researchers they serve, libraries will play essential roles in facilitating the processes of sharing and using research data.

DISCUSSION AND CONCLUSIONS

Research data have become essential scholarly resources to be captured and curated for reuse – but which data, why, and for whose interests? Libraries need to address all of these questions if they are to play a role in selecting, collecting, organizing, curating, and providing access to research data as they have for other scholarly content over the millennia.

For data to be available for selection and curation, they must be shared by those who collected or produced them. Data are complex, messy objects that take many forms. Any individual datapoint or dataset can be transformed multiple times from origin to its use in a publication, and beyond. Data are difficult to define. We know that metadata are needed to describe them and that provenance records are important in interpreting them. But what is the "them"? Lacking a precise definition, data can be viewed along four dimensions of the purposes for which they are collected and four dimensions of research methods. The examples in this paper are drawn from the sciences. The array of purposes and methods is far greater when the full range of data useful to the social sciences and humanities is also considered [16; 71].

The reasons to share data are many and varied. Four arguments are presented for sharing research data:

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

These arguments can be assessed along two axes: from public-driven to science-driven motivations, and from data-producer driven to data-user driven interests. The arguments interact in complex ways with notions of data. Researchers' incentives (and disincentives) to share their data depend on both the reasons for sharing and on their investments in their data. The first two arguments 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 latter two arguments, which also benefit the public, are framed more in the interests of data producers, and serve science more directly. As these latter arguments are more effective with the science community, they also align with the interests of the library community.

Science-driven design starts with the interests of the scientific community. While that may seem an obvious statement, it is a goal more difficult to accomplish than it appears. The "science community" does not speak with one voice. Nor does the "astronomy community," the "biology community," or any other individual field. 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 [72;

73; 74], and the interest in practices associated with data has accelerated in the last decade [16; 53; 26]. Social science and humanities research practices have received far less attention; studies of these also are needed [71].

Initiatives such as the NSF DataNet program [51] endeavor to bring scientists, librarians, and systems developers together to understand data-driven design. Multiple, parallel studies of individual research groups and communities are underway to inform the policy and design of library data curation.

Already we are learning that science-driven design means selecting and organizing data to reflect specific practices. 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, scientists 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 the 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 by libraries; they await study and partnership.

ACKNOWLEDGEMENTS

This paper benefited greatly from comments on earlier drafts by the CENS Data Management team at UCLA – David Fearon, Matthew Mayernik, Katie Shilton, Jillian Wallis, and Laura Wynholds – and by Paul Uhlir of the National Academies.

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; T. Finholt, UM; S. Jackson, UM; D. Ribes, Georgetown; S.L. Star, SCU; 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

[1] 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.

[2] Community cleverness required. (2008). *Nature,* **455**(7209): 1-1.

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

[4] 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.

[5] Data, Data Everywhere: A special report on managing information. (2010). *Economist*: 16-17.

[6] *The University's Role in the Dissemination of Research and Scholarship - A Call to Action* (2009). Association of Research Libraries. Retrieved from www.arl.org/disseminating_research_2009 on 10 March 2009.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Wallis, J. C., Borgman, C. L., Mayernik, M. S. & Pepe, A. (2008). Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation,* **3**(1). Retrieved from http://www.ijdc.net/ijdc/issue/current on 24 November 2008.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] Baker, K. S., Ribes, D., Millerand, F. & Bowker, G. C. (2005). Interoperability Strategies for Scientific Cyberinfrastructure: Research Practice. *Proceedings of the American Society for Information Systems and Technology*.

[19] Baker, K. S. & Yarmey, L. (2009). Data Stewardship: Environmental Data Curation and a Web-of-Repositories. *International Journal of Digital Curation,* **4**(2): 1-16. Retrieved from http://www.ijdc.net/index.php/ijdc/issue/view/8 on 30 December 2009.

[20] 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.

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

[22] 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.

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

[24] Zimmerman, A. S. (2003). *Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists* Ph.D Dissertation. School of Information. University of Michigan. Ann Arbor, MI. **280** Pages.

[25] 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.

[26] 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.

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

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

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

[30] 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*.

Vancouver, British Columbia, Canada, Association for Computing Machinery**:** 269-277. Retrieved from http://doi.acm.org/10.1145/1255175.1255228 on June 17-23, 2007 Accessed.

[31] 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.

[32] 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.

[33] Ribes, D. & Finholt, T. A. (2007). *Tensions across the scales: Planning infrastructure for the long-term*. Sanibel Island, Florida, Association for Computing Machinery**:** 229-238Accessed.

[34] 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.

[35] Wallis, J. C., Borgman, C. L., Mayernik, M. S., Pepe, A., Ramanathan, N. & Hansen, M. (2007). Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries. *11th European Conference on Digital Libraries*, Budapest, Hungary, Berlin: Springer. 380-391.

[36] 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.

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

[38] *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.

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

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

[41] *Final NIH Statement on Sharing Research Data, NOT-OD-03-032*. (2003). National Institutes of Health. Retrieved from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html on.

[42] 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.

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

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

[45] 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.

[46] 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.

[47] Lagoze, C. & Velden, T. (2009). 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.

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

[49] 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.

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

[51] *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.

[52] Borgman, C. L. & Palmer, C. L. (2010). The Data Conservancy: Science-driven Information Science. *Berlin Research Colloquium / Berliner Bibliothekswissenschaftliches Kolloquium*. Retrieved from http://works.bepress.com/borgman/236 on 10 August 2010.

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

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

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

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

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

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

[59] Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messina, P., Messerschmitt, D. G., Ostriker, J. P. & Wright, M. H. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon panel on Cyberinfrastructure*. National Science Foundation. Retrieved from http://www.nsf.gov/cise/sci/reports/atkins.pdf on 18 September 2006.

[60] *Object Reuse and Exchange*. (2009). Retrieved from http://www.openarchives.org/ore/ on 15 September 2009.

[61] Pepe, A., Mayernik, M. S., Borgman, C. L. & Van de Sompel, H. (2010). From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *Journal of the American Society for Information Science and Technology,* **61**(3): 567–582. Retrieved from http://www3.interscience.wiley.com/journal/123214737/abstract on 1 February 2010.

[62] Bourne, P. (2005). Will a biological database be different from a biological journal? *PLoS Computational Biology,* **1**(3): e34. Retrieved from http://dx.doi.org/10.1371/journal.pcbi.0010034 on 28 September 2006.

[63] Stodden, V. (2009). 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.

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

[65] Stodden, V. (2010). *The scientific method in practices: Reproducibility in the computational sciences*. MIT Sloan School Working Paper 4773-10. Retrieved from http://ssrn.com/abstract=1550193 on 10 August 2010.

[66] 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.

[67] 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.

[68] Borgman, C. L. (2008). Data, disciplines, and scholarly publishing. *Learned Publishing,* **21**(1): 75-84.

[69] Tillett, B. B. (2004). *What is FRBR?: A Conceptual Model for the Bibliographic Universe*. Library of Congress. Retrieved from http://www.loc.gov/cds/FRBR.html on 10 March 2006.

[70] Wallis, J. C., Borgman, C. L. & Mayernik, M. S. (2010, in review). Who is responsible for data? A case study exploring data authorship, ownership, and responsibility and their implications for data curation. *6th International Digital Curation Conference*, Chicago, Digital Curation Center. Retrieved from http://www.dcc.ac.uk/events/conferences/6th-international-digital-curation-conference on 11 August 2010.

[71] 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.

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

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

[74] 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.