Data in the Sciences

By: Karen Stanley Grigg

Grigg, K. (2015). Data in the Sciences. In Lynda Kellam & Kristi Thompson (Eds.), *Databrarianship: The Academic Data Librarian in Theory and Practice*.

Abstract:

This chapter presents an introduction to scientific data and its relevance to librarians and libraries. The characteristics of science data in general, and in relation to a number of scientific disciplines, are identified. The disciplines discussed have been chosen because they demonstrate notable aspects of data management, either because of the type of data used or because of the requirements external agencies place on researchers. Federal funding agencies' requirements for data management and sharing are discussed, along with initiatives to promote sharing of data, and notable large datasets and repositories are identified. Though the chapter mainly focuses on United States (U.S.) funding agencies, it also lists some international archives. Broad discussion of data management in the sciences, and how libraries and librarians can embed themselves in the data lifecycle, are presented, along with specific examples of how libraries have become involved with research data services.

Keywords: scientific data | libraries | funding agencies | data management

Article:

*****Note: Full text of article below**

CHAPTER 12

Data in the Sciences

Karen Stanley Grigg

THIS CHAPTER PRESENTS an introduction to scientific data and its relevance to librarians and libraries. The characteristics of science data in general, and in relation to a number of scientific disciplines, are identified. The disciplines discussed have been chosen because they demonstrate notable aspects of data management, either because of the type of data used or because of the requirements external agencies place on researchers. Federal funding agencies' requirements for data management and sharing are discussed, along with initiatives to promote sharing of data, and notable large datasets and repositories are identified. Though the chapter mainly focuses on United States (U.S.) funding agencies, it also lists some international archives. Broad discussion of data management in the sciences, and how libraries and librarians can embed themselves in the data lifecycle, are presented, along with specific examples of how libraries have become involved with research data services.

Science Data Management—Early Adopters

In the United States, federal funding agencies were among the first to issue data management and open access policies for their grant recipients. In 2005, the United States National Institute of Health (NIH) created a public access policy that encouraged researchers to provide open access to its funded research. In 2007, a new NIH Public Access Policy bill was signed by President George W. Bush that required the NIH to provide open access to all the research it funded. Investigators would now be required to submit their final manuscripts to PubMedCentral, the NIH's open access repository. In 2011, the United States National Science Foundation (NSF) required that all proposals submitted to the NSF must include a data management plan of two pages or less. In contrast, other agencies have implemented their data management and sharing policies more recently. For example, the Office of Digital Humanities released its data requirements in 2014, and the Institute of Museum and Library Services implemented its data management plan requirements in 2015.

Scientific research and researchers have unique characteristics and needs that have pushed data management in the sciences to the forefront. One such characteristic of scientific research is that the size of the data presents specific challenges for storage and curation. Another unique aspect of scientific research is that some disciplines, such as medicine, have special privacy and security concerns. Scientific datasets can be extremely large, often taking up terabytes, even petabytes, of storage space. In 2011, *Science* magazine polled its peer reviewers about the size of the datasets used and found that "about 20% of the respondents regularly use or analyze datasets exceeding 100 gigabytes, and 7% use datasets exceeding 1 terabyte. About half of those polled store their data only in their laboratories—not an ideal long-term solution. Many bemoaned the lack of common metadata and archives as a main impediment to using and storing data."[1] Additionally, 80 percent said that their institution did not have sufficient funding for data curation.[2] While more recent data was not available, datasets are unlikely to have become any smaller.

Genomics and astronomy generate some of the largest datasets, often reaching one or more petabytes (PB). The Sloan Digital Sky Survey, for example, produces about five PB a day, and the proposed Large Synoptic Survey Telescope (LSST) will record 30 trillion bytes of image data daily from the top of a mountain in Chile, a data volume equal to two daily Sloan Digital Sky Surveys.[3] Not only can storage be a limiting factor, but due to their size, these datasets can be difficult to move to remote locations, download, or process. Scientific data are expensive and complex to analyze and manipulate, as the data require high performance computer servers and storage solutions that strain the budgets of individual institutions. These servers must manage and transfer petabytes and terabytes of data in distributed computing environments, generate simulations, and provide the ability to share and transfer large sets of data to remote or local sites.

Scientific data are often decentralized, generated in a laboratory by many researchers. This dispersion often leads to data being spread across multiple storage devices and without a standardized method of naming conventions or metadata. Additionally, many of the people in the laboratory producing the data are graduate students and postdoctoral researchers, who may take their data with them when they leave or leave the institution without a schema to allow those who come after them to interpret data results. Those researchers who direct the workflow are not necessarily those who produce and manipulate the research data. Because of this decentralization, it is crucially important that the researchers create detailed plans for data preservation, curation, storage, and metadata conventions in their initial data management plan. Creating standardized conventions for file naming and formats is important for long-term curation of research data. It is useful for entire disciplines to discuss standards, rather than leaving the decisions to be carried out institution by institution, in order to enable useful sharing of data between researchers.

The willingness to openly share scientific data varies by discipline. MacMillan states, "Data sharing among the sciences does not have an all-inclusive, uniform or over-arching data culture, a function of both values and of sheer quantities and types of data produced."[4]

Tenopir et al. surveyed a total of 1,329 scientists in a variety of disciplines regarding their data sharing practices and perceptions. Survey data revealed that those in basic sciences, such as atmospheric science, environmental science and ecology, and biology are cultures that most favor data sharing, while health sciences, computer science, and engineering are disciplines less likely to share data.[5] Disciplines in which data are costly to obtain and store, such as astronomy or meteorology, tend to have a culture that favors and encourages the open sharing of data.[6] Confidential data are a significant concern with medical research in the United States and in other countries, and highly competitive disciplines that are financially lucrative are likely to be more reluctant to share data.[7] While science data share some commonalities, large size and production in laboratory environments among them, individual scientific disciplines have varying characteristics and attitudes towards sharing data. This chapter focuses on some of the major scientific disciplines that have unique datasets, funding agency requirements, or challenges with analyzing, storing, and sharing data. As science librarians should become familiar with the major datasets and repositories in their subject areas, this chapter lists some of the major data repositories for each discipline.

Genomics/Biomedical Sciences

Genomics datasets require a massive amount of storage space and are expensive to sequence and store. An early genomic project that garnered much attention is the Human Genome Project, which mapped the "human genome" and was declared complete in 2003. The field of Bioinformatics has emerged in order to deal with the complexity of organizing and analyzing genomic data.

Though the cost of DNA sequencing is high, it has decreased sharply due to the increase over time in computing power. In 2009, the average raw genome cost $154,714 to sequence. In 2014, the cost had dropped to $4,905.[8] As cloud storage solutions become increasingly common, the cost of sequencing raw genomic data will continue to decrease, though cloud solutions must guarantee data security for wide scale adoption to be possible. Some non-commercial cloud data solutions include Bionimbus Protected Data Cloud (PDC), an open source cloud-computing platform developed at the University of Chicago, and NCI Cancer Genomics Cloud Pilots, an NIH trusted partner. Commercial cloud computing solutions include the Amazon Web Services Cloud and the newly released Google Genomics Cloud Platform.

Biomedical human research presents many barriers to creating a culture of data sharing. Due to privacy concerns, researchers are often reluctant or unable to

make their data publicly available. A study on gene sequence data articles published between 2006 and 2009 showed that cancer researchers who use human subjects tended to not share their datasets, and only a fourth of the studies had stored their raw data in repositories. Data sharing was more prevalent when studies were funded by NIH, and thus were covered by the NIH Data Sharing Policy.[9] A more recent survey of biomedical researchers in the U.S. Southwest confirmed that about 88% of respondents agreed with the NIH Resource Sharing Plan, and over half would be likely to participate in a virtual biorepository of human cancer biospecimens.[10]

Genomics and Biological Sciences Datasets/Repositories

- 1000 Genomes (http://www.ncbi.nlm.nih.gov/geo/): Launched in 2008, this international project is a catalog of human genetic variation. Scientists have sequenced genomes of over a thousand participants in a variety of ethnic groups.
- Entrez Databases (http://www.ncbi.nlm.nih.gov/genbank/): Entrez is a web portal that consists of a number of health sciences databases at the National Center for Biotechnology Information (NCBI) web site. Entrez provides a federated search engine that allows users to search across all databases.
- National Center for Biotechnology Information (NCBI)'s Gene Expression Omnibus (NCBI GEO) (http://www.ncbi.nlm.nih.gov/geo/): One of the NCBI databases and an international public repository for functional genomic datasets.
- GenBank (http://www.ncbi.nlm.nih.gov/genbank/): GenBank, another NCBI database, is an annotated collection of all publicly available DNA sequences.
- The cBio Cancer Genomics Portal (http://cbioportal.org): This portal is an open access repository that provides access to large-scale cancer genomics datasets.
- Dryad Digital Repository (http://www.datadryad.org/): Dryad is a scientific data repository that makes the data corresponding to scientific journal articles available and re-usable. Authors, journals, and publishers can facilitate data archiving at the time of article publication and receive a permanent Digital Object Identifier (DOI), which can be included in the published article.

Astronomy

Astronomical research, like genomics, produces large amounts of data and can create great challenges for processing and archiving. Telescopes are becoming

larger and more sophisticated than ever before. In a 2012 interview, Alberto Conti explained:

> Over the past 25 to 30 years, we have been able to build telescopes that are 30 times larger than what we used to be able to build, and at the same time our detectors are 3,000 times more powerful in terms of pixels. The explosion in sensitivity you see in these detectors is a product of Moore's Law—they can collect up to a hundred times more data than was possible even just a few years ago. This exponential increase means that the collective data of astronomy doubles every year or so, and that can be very tough to capture and analyze.[11]

Astronomy has long fostered a culture that is receptive and proactive in terms of data sharing, and, indeed, there are a number of large open access datasets. Norris discusses the culture of sharing in astronomical research: "Because the advance of astronomy frequently depends on the comparison and merging of disparate data, it is important that astronomers have access to all available data on the objects or phenomena that they are studying. Astronomical data have therefore always enjoyed a tradition of open access, best exemplified by the astronomical data centres, which provide access to data for all astronomers at no charge."[12]

a. Managing and archiving these large databases is often difficult due to understaffing, which makes the creation of metadata challenging. Other issues include a lack of standardization of formats, inadequate nomenclature, and the fact that many of the classic, seminal articles in astronomy have not been converted to electronic formats.[13] As so much astronomy discovery depends on the collaborations of astronomers internationally, common data standards and formats have been necessary in order to process and make sense of data from a variety of observatories and laboratories. Community standards have been developed to handle large-scale astronomy projects. For example, the Flexible Image Transport System (FITS), published in 1981, is the standard format for the interchange of astronomical images and arrays.

Astronomy Datasets/Repositories

- The Sloan Digital Sky Survey (http://www.sdss.org): This project has detailed three-dimensional maps of the Universe, containing images of one third of the sky. It also contains spectra for over three million astronomical objects.

- Earth Observing System Data and Information System (EOSDIS) (https://earthdata.nasa.gov/about-eosdis): Processes, archives, and distributes data from a large number of Earth observing satellites and represents a crucial capability for studying the Earth system from space and improving prediction of Earth system change. EOSDIS consists of a set of processing facilities and data centers distributed across the United States that serve hundreds of thousands of users around the world.
- NASA Space Science Data Coordinated Archive (http://nssdc.gsfc.nasa.gov/): NASA's archive for space science mission data,
- NED: NASA/IPAC Extragalactic Database (https://ned.ipac.caltech.edu/): This site collects and distributes astronomical data worldwide, for millions of objects outside the Milky Way galaxy and connects the public to observatories around the world and makes use of data collected from powerful telescopes.

Chemistry

Chemistry research data management practices have lagged behind other disciplines for a variety of reasons. Often, the research work of chemists takes place at the lab group level, and is guided by academics who must focus on managing the project rather than interacting directly with the data. This lack of hands-on contact with the laboratory's data can result in differing opinions among chemists on standards for data management. Chemists often store data in formats that are subject to technological obsolescence, and labs rarely create metadata for research. As graduate students and junior researchers leave for other institutions, data and knowledge become lost.[14]

While researchers in fields such as astronomy and physics are organizing and storing their data on a cross-institutional basis, partially due to the expense of equipment, chemistry is still largely campus-based, and faculty often believe their field is slow to innovate.[15]

Given that the culture of chemistry is more campus-based and less collaborative, chemists are less likely to value sharing their data than researchers in disciplines such as astronomy and geology. In a 2011–2013 study of their research support needs, 67% of interviewed chemists surveyed responded that they had never utilized an online repository, though many did state that they would be more likely to do so if mandated by their institution.[16] While chemists tend to not openly share their data, they do frequently make their data available to colleagues when requested.[17]

Chemistry Datasets/Repositories

- Cambridge Structural Database (http://www.ccdc.cam.ac.uk/about_ccdc/): The Cambridge Crystallographic Data Center (CCDC) compiles

this database, which is an international repository of organic and metal-organic crystal structures obtained via experimental data.

- Chemical Synthesis Database (ChemSynthesis) (http://www.chemsynthesis.com/): ChemSynthesis is an open access chemical database. It contains physical properties of over 40,000 compounds and over 45,000 synthesis references.
- ChemXSeer (http://chemxseer.ist.psu.edu/): Hosted and administered by Pennsylvania State University, ChemxSeer is an integrated digital library and database that hosts published articles and experimental data obtained from chemical kinetics.
- PubChem (http://pubchem.ncbi.nlm.nih.gov/): From the U.S. National Center for Biotechnology Information of the National Institutes of Health (NIH). This database provides chemical information on the biological properties and activities of small molecules. It is organized as three linked databases within NCBI's Entrez Information Retrieval system.

U.S. Federal Agency Requirements on Data Management and Data Sharing†

As previously discussed, given that scientific researchers have varying approaches to sharing data, and that scientific data can be large and complex, the importance of data sharing to further development and expansion of research ensures that data sharing matters to federal agencies. Scientific federal agencies were the first to impose requirements for data management and sharing. The three federal agencies with the largest research and development budgets are the National Institute of Health, The National Science Foundation, and the Department of Defense.[18] This section discusses the requirements from these three agencies, as well as from other important federal granting agencies that support scientific research.

The first federal funding agency to develop and implement a data sharing policy was the National Institute of Health in 2008. In this public access policy, scientists funded by NIH grants are required to submit their final peer-reviewed journal manuscripts to PubMed Central immediately upon acceptance for publication. Increasingly, journals are now automatically submitting these publications to PMC on behalf of authors. However, authors can upload manuscripts themselves via the NIH Manuscript Submission (NIHMS) system, which submits articles to PubMedCentral.

Additionally, researchers who submit an application seeking $500,000 or more in NIH funding for a single year must include their plan for data sharing, or, if data sharing is not possible, their reasons for not sharing. NIH states that "data

† U.S. policies and agencies have been mainly discussed, as the author is most familiar with them. The situation in other countries will vary.

intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects."[19]

In January 2015, the NIH-instituted Genomic Data Sharing (GDS) Policy became effective, and it applies to all NIH-funded research that generates large-scale genomic data. All researchers applying for NIH grants where large-scale genomic data will be generated must create a Genomic Data Sharing plan in their proposal and outline in the budget section of their application the resources they will need to prepare the data for submission to the appropriate repositories. If the sharing of human data is not possible, applicants should provide a justification as to why the data cannot be shared.

Following on the heels of NIH, in 2011 the National Science Foundation (NSF) began requiring that researchers submit a two-page data management plan as part of each funding proposal. This data management plan must include information about the types of data to be gathered, the metadata standards to be used, the policies and provision for reuse, and plans for long-term archiving. Although the NSF has guidelines for archiving, saving, or providing of samples, collections, or research data, it does not mandate how these practices should be done.

The Department of Defense (DoD) has released a draft of its proposed public access plan, which requires that all proposals for research funding include a data management plan, as well as the deposit of metadata for all created datasets to the Defense Technical Information Center (DTIC.) Other federal funding agencies with data management mandates or guidelines include the Environmental Protection Agency, the Department of Energy, National Aeronautic and Space Agency, and the National Oceanic and Atmospheric Administration.

In March 2012, the Obama-Biden administration rolled out the Big Data Research and Development Initiative from the Office of Science and Technology Policy (OSTP), which will invest $200 million to store and provide access to large collections of digital data. This effort involves Defense Advanced Research Projects Agency (DARPA), the NIH, and the NSF. Some of the initiatives include a partnership between the NIH and Amazon Web services to host the 1,000 Genomes Project data on Amazon Web Services, grants for "EarthCube" (a collaborative partnership between the U.S. National Science Foundation's Geosciences Directorate (GEO) and the Division of Advanced Cyberinfrastructure (ACI) and DARPA's XDATA, which develops computational techniques for analyzing large volumes of defense data.

Science Librarian Support for Data Management

Given the increasing emphasis of data management practices in the sciences, the Association of Research Libraries (ARL) began to discuss the roles that libraries might take as partners in research in the mid-2000s. In 2006, ARL hosted a work-

shop funded by the National Science Foundation to discuss and explore the roles academic research libraries could serve in order to partner with organizations in the science and engineering research lifecycle. A subsequent report (*To Stand the Test of Time: Long Term Stewardship of Digital Data Sets in Science and Engineering*) presented recommendations for data stewardship.[20]

Furthermore, a survey of science librarians at ARL member institutions in 2012 asked respondents about participation in data management and data repositories. The survey showed that 8% of respondents' institutions accepted and stored data, and 11% indicated that the institution was working on implementing a repository. Thirty-two percent indicated assistance with data management was available, and 24% responded that librarians offered data management plan consultations. Forty-four percent of respondents indicated that their job duties included working with institutional/data repositories or data management, with 17% indicating that these duties were forthcoming.[21]

Librarians knew that data services were a newly emerging role and opportunity but were grappling with the scope of what could and should be done. The gradual emergence of federal agency requirements provided a framework of current and future services libraries could provide. As science funding agencies, such as the NLM and NIH, are increasingly mandating data management plans and data sharing, science librarians have a unique opportunity to participate in library efforts to develop and define data management and support services at academic research institutions.

Research Data Management Lifecycle

It is helpful for librarians, in planning support services, to consider the research data management lifecycle. This basic model† shows the research lifecycle as an iterative process with many points of opportunity for libraries to provide services and guidance to researchers. Most scientists begin by formulating and refining a research question while consulting previous studies via journal articles and data analysis. Librarians have long been involved in this stage of the research project, assisting researchers with locating appropriate background information, scholarly journal articles, and datasets, which is the data search/reuse stage of the research data management lifecycle. Assisting users with literature reviews and finding information has been the specialty of the liaison, and liaisons should become familiar and knowledgeable about locating data in their subject areas. The library may have a data services librarian who specializes in locating data in a variety of

† This model has been formulated in various ways. In brief, the idea is that a data management plan needs to take into account initial research, data collection, documentation, analysis, storage and archiving, discovery, and reuse (which starts the cycle over again). See a version of the lifecycle at http://guides.library.ucsc.edu/datamanagement/.

disciplines, but the liaison is often the first point of contact and should have some background knowledge. Liaisons often focus on serving faculty individually but may also embed themselves in a laboratory at the beginning of a research project.

When the research question has been formulated, and the initial data has been gathered, researchers are often required by the funding agency to create a data management plan. Even when such a plan is not required, librarians should emphasize the importance of going through the exercise. Many libraries are offering assistance with data management plans by offering information on data management tools, locating standards for different agencies, and even providing assistance in the writing of these plans.

During the data storage phase, as well as the archival phase, libraries can provide services such as advising on best practices in organizing, naming, and storing research data. Science librarians should take leadership in advising researchers on current standards for metadata and file naming in relevant subject areas, such as the Darwin Core and Integrated Taxonomic Information System (IT IS) for Natural History, the Standard for the Exchange of Earthquake Data (SEED) for Earth Sciences, and the Geospatial Interoperability Framework (GIF) for Geographic and Geospatial Data. Librarians already possess skills in organization and description of materials, and development and assistance with metadata are an opportunity to offer a valuable service to researchers and stay relevant in the university setting.

Institutional and Other Repositories

Given the role libraries already serve in hosting and offering research materials to the university, they are well positioned to offer services for storing, hosting, and archiving scientific research data. Many libraries are already offering institutional repositories to host manuscripts, so hosting, storing, and sharing researcher data are natural extensions. However, libraries with limited budgets and staff are frequently unable to take on this major commitment, and certainly should not try to if a long-term commitment is not possible. When libraries cannot store data internally, librarians should advise and guide researchers to the appropriate repositories for storing their data, and liaisons should become familiar with repositories in their subject areas.

Dedicated Professional Staff

Some libraries have dedicated data services librarians who work with all the colleges and departments who produce data. These librarians work closely with Information Technology staff, faculty, and other relevant library staff to provide a central source for data management and support initiatives. Other libraries do not

have a centralized research and data support librarian, and, hence, have distributed data support work among liaisons, technical services, and library IT staff. Though libraries without dedicated research and data support librarians and with a small staff may have to scale down the services that can be provided to institutional researchers, it is important for the library to define the needs of the researchers, to define the capabilities of library staff, and to begin to make strategic decisions as to what the library can offer. Administrators considering library-wide initiatives should assess the feasibility of having dedicated positions for data support.

Subject Guides

Many academic libraries have created subject guides specifically on data management to assist campus researchers in all areas of data, from discovering existing datasets, to working with data management plans and tools, to finding subject-specific repositories. Researchers, however, may not discover these guides unless they are heavily marketed and easy to locate from library web sites. Science librarians should make sure that these subject guides are presented to their affiliated departments.

Assessment

Before undertaking wide-scale data management services, it is important to assess the needs and practices of researchers across campus. Assessment tools (such as survey instruments, focus groups, and interviews) can be used to determine how faculty in different departments are approaching creating data management plans and processing and storing research data, and to identify their attitudes and practices when it comes to sharing data. Researchers in departments can indicate which services both in the library and also throughout campus would be valuable.

Partnering across the Organization

Librarians involved with data services and management should also identify other campus entities that are already providing data management assistance and reach out to these offices and offer to partner.† Many times, research faculty are not aware of all the services available to them on campus, and the library can serve a valuable role in marketing and communicating on behalf of these entities. A few examples demonstrate how university libraries are participating in providing data management services and resources to researchers. These efforts are not specifi-

† For more information and examples on collaborative support for research data management initiatives, see the chapter in this volume by Hofelich Mohr, Johnston, and Lindsay called "The Data Management Village."

cally limited to science data, but science librarians can use these models as starting points in their own efforts to provide data support.

Purdue University took an early leadership role in offering data services to its researchers, focusing initially on science and engineering research. Through an initiative that evaluated interest in collaborating with researchers, the results led to the creation of a full-time data research scientist position.[22] Purdue built a data curation profiles toolkit that provides a snapshot of the researcher's data at any given time. Additionally, in 2011, the university launched the Purdue University Research Repository (PURR), which provides a collaborative, virtual research environment and working space enabling researchers to create a data management plan, upload research data, and publish datasets.

In 2009, Georgia Tech decided to assess data management practices on campus in advance of the upcoming NSF requirement for data management plans. At that time, 40% of researchers surveyed believed that creating data management plans was unnecessary, and 47% indicated that they did not know much about data management. Seventy-three percent were interested in data storage and preservation assistance, and 40% were interested in information about developing data management plans.[23] As a result, Georgia Tech now offers assistance with data management plans, allows data to be included in SMARTech (Georgia Tech's institutional repository that is comprised of many different communities and collections), and offers a "Data Management Planning" course to help satisfy Georgia Tech's "Responsible Conduct of Research Compliance Policy."

The Johns Hopkins Libraries, in response to big data challenges, developed and implemented Data Management Services in 2010, which provides data management planning support, data consulting, and archival services across disciplines. Researchers can use the JDH Data Archive to store their data if no other repositories are available. The consultants work with researchers to determine their data management needs (how they are planning to manage, store, and share their data) and advise on metadata standards.

The University of Virginia Library's Research Data Services provides extensive and robust data management services to campus researchers. One of the original contributing institutions in the development of the DMPTool, which provides templates for major funding agencies' required formats for DMPs. In addition, librarians offer data management workshops to researchers and are available to help with data management plans for all University of Virginia researchers. The StatLab provides consulting and workshops for researchers on using statistical methods and software.

Health Science librarians are keenly poised to assist researchers with NIH compliance issues. Duke Medical Center Library started monitoring the NIH Public Access policy in 2006 and created a web-based guide to inform researchers on how to comply. Librarians reached out to key research offices around campus and were given the ability to take the lead on addressing researcher compliance.

NIH has created the Public Access Compliance Monitor (PACM), an online system that lists citations found in PubMed citing grants that should fall under the NIH Public Access Policy. Librarians can download reports of non-compliant PIs; researchers who are out of compliance are individually contacted by the library about their compliance issues, given instructions on how to comply, and given contact information so that they can work with librarians to get assistance on uploading manuscripts and other required functions. All library staff receive training in order to handle basic compliance issues, and a core team is available for more complex problems.

Science librarians who want to add data support to their job descriptions should begin by assessing the needs of their departments and researchers. Meeting department heads and head researchers, talking with graduate students, and finding out what services and dedicated staff already exist in each department is important in order to ascertain what service gaps may exist. Important types of information to gather include types and formats of data generated, storage and backup practices, methods used to organize data, whether data management plan help is needed, sharing habits, and identification of units across campus that assist researchers in their departments. Social science data librarians wanting to expand their services to include science data can partner with science librarians, who can provide subject expertise, knowledge of research practices, and connections to key contact people, even if these science librarians lack expertise in data management and support.

Due to the increasing complexity and expense of working with scientific research data, along with the push towards making publicly funded research available to all, scientists are straining to keep abreast of agency policies, data management plan creation, compliance, metadata standards, and all the other aspects of research data management that compete with their other duties in the laboratory and the classroom. Librarians have the organizational and, often, the technical skills to be valuable partners with our researchers to assist with their data service needs. Even smaller libraries with scant budgetary resources can develop expertise in helping researchers write data management plans, develop helpful guides, and connect their researchers to the appropriate data repositories. Finally, as data are becoming increasingly important as resources for scholarly research, science librarians should become familiar and keep current with the major datasets and repositories in their subject areas.

---

1. Science Staff, "Challenges and Opportunities," *Science* 331, no. 6018 (2011): 692–93, doi:10.1126/science.331.6018.692,693.
2. Ibid.
3. Randal Bryant, Randy H Katz, and Edward D Lazowska, *Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society,* December 22, 2008, http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big_Data.pdf, 2.

4. Don MacMillan, "Data Sharing and Discovery: What Librarians Need to Know," *Journal of Academic Librarianship* 40, no. 5 (September 2014): 541–49, 542.
5. Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame, "Data Sharing by Scientists: Practices and Perceptions.," *PLoS ONE* 6, no. 6 (June 2011): 1–21.
6. Abigail Goben, Dorothea Salo, and Claire Stewart, "Federal Research," *College & Research Libraries News* 74, no. 8 (September 2013): 421–25.
7. Carol Tenopir et al., "Data Sharing by Scientists: Practices and Perceptions."
8. National Human Genome Research Institute, "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)," *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*, October 31, 2014, http://www.genome.gov/sequencingcosts/.
9. Heather A Piwowar, "Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data," *PloS One* 6, no. 7 (2011): 1–13.
10. Mai H. Oushy, Rebecca Palacios, Alan E. C. Holden, Amelie G. Ramirez, Kipling J. Gallion, and Mary A. O'Connell, "To Share or Not to Share? A Survey of Biomedical Researchers in the U.S. Southwest, an Ethnically Diverse Region," ed. Xu-jie Zhou, *PLOS ONE* 10, no. 9 (September 17, 2015): e0138239, doi:10.1371/journal.pone.0138239.
11. Ross Anderson, "How Big Data are Changing Astronomy (Again).," *The Atlantic*, April 19, 2012, accessed March 20, 2015.
12. Ray P Norris, "Can Astronomy Manage Its Data?," *arXiv Preprint Astro-ph/0501089*, 2005. 2–3.
13. Ibid, 3.
14. Matthew P Long and Roger C Schonfeld, "Supporting the Changing Research Practices of Chemists," *New York: Ithaka S+ R,* 2013, http://www.sr.ithaka.org/publications/supporting-the-changing-research-practices-of-chemists/ 11–12.
15. Ibid, 7.
16. Ibid, 35.
17. Ibid, 12.
18. U.S. Office of Science and Technology Policy, "The 2014 Budget: A World-Leading Commitment to Science and Research: Science, Technology, Innovation, and STEM Education in the 2014 Budget," 2008, https://www.whitehouse.gov/sites/default/files/microsites/ostp/2014_R&D-budget_overview.pdf.
19. "NIH Guide: FINAL NIH STATEMENT ON SHARING RESEARCH DATA," February 26, 2003, http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.
20. Amy Friedlander and Prudence Adler, "To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships–The Role of Academic Libraries in the Digital Data Universe," *Association of Research Libraries*, 2006, http://www.arl.org/publications-resources/1075-to-stand-the-test-of-time-long-term-stewardship-of-digital-data-sets-in-science-and-engineering#.VpZvkfkrLRY.
21. Karen Antell, Jody Bales Foote, Jaymie Turner, and Brian Shults, "Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions," *College & Research Libraries* 75, no. 4 (July 2014): 557–74.
22. Catherine Soehner, Catherine Steeves, and Jennifer Ward, "E-Science and Data Support Services: A Study of ARL Member Institutions.," *Association of Research Libraries*, 2010. http://www.arl.org/storage/documents/publications/escience-report-2010.pdf.
23. Susan Wells Parham, Jon Bodnar, and Sara Fuchs, "Supporting Tomorrow's Research," *College & Research Libraries News* 73, no. 1 (January 2012): 10–13.