Considerations for Research Data Sharing

Setting the Stage for Greater Openness

ECAR working groups are where EDUCAUSE members come together to create solutions to today’s problems and provide insight into higher education IT’s tomorrow. Individuals at EDUCAUSE member institutions are invited to collaborate on projects that address core technology challenges and advance emerging technologies important to colleges and universities. More information can be found at the ECAR working groups website.

Introduction

In a collaborative, computationally powerful research environment, data sharing holds tremendous promise. Openly available data sets that are well documented with clear rights statements attached allow secondary users to ask new questions of extant data, leading to new discoveries and increased scientific progress. Open research data increases transparency and makes the results of research replicable and verifiable. Funding agencies, wanting to increase the value of their scientific investments, are pressing researchers to make the results of publicly funded research publicly available. IT companies developing research-based products want to promote their discoveries through reproducible experiments using open research data.

Underneath the much-heralded promises of research data sharing, however, lie numerous considerations that impact or hinder research data sharing as a common practice. In this paper we explore some of the topics that institutions and researchers should consider as they encourage and enable research data sharing.

What Is Research Data?

The phrase “research data” has numerous connotations, and interpretation usually depends on the audience. The most prevalent definition of research data is “the recorded factual material commonly accepted in the scientific community as necessary to validate research findings.” Another definition describes research data as that which is “collected, observed, or created, for purposes of analysis to produce original research results.”

Research data has a number of broad categories, including:

- Raw or initially processed data produced at a facility
- Research-ready processed data that has been cleaned, combined, and/or calibrated
- Published output data sets from a detailed analysis of a research-ready data set
- Supplemental to a journal article
- Through exposed metadata (data description) while keeping data private
set; and published catalog-type representation of a published data set. The origins of research data also span a number of categories, including.

- **Observational data** captured in real time (sensor data, survey data, etc.)
- **Experimental data** collected from lab equipment (gene sequences, chromatograms, microCT scans, etc.)
- **Simulation results** generated from test models (astrophysical models of planetary formation and gravity waves, simulations of infectious-disease transmission, etc.)
- **Derived or compiled data** (extracted through text mining, databases, 3D models, etc.)
- **Reference or canonical data** compiled from peer-reviewed data sets and authoritative sources that are published and curated (gene sequence databanks, chemical structures, spatial data portals, the U.S. Census, etc.)
- **Crowdsourced data** collected from amateurs and professionals (fireball tracking, protein folding, gravitational wave detection, SETI@home, Zooniverse, etc.)

In addition to the raw data, research data should comprise all that is necessary to understand the data, such as the surveys, methods, protocols, code, and models used to generate the raw data. This additional material acquired and generated during the process of research may be either digital or analog and could include:

- Text (documents, spreadsheets)
- Laboratory notebooks, field notebooks, diaries
- Questionnaires, transcripts, codebooks
- Media (audio and video, photographs, films)
- Test responses
- Artifacts (slides, specimens, samples)
- Methodologies and workflows
- Operating procedures and protocols
- Data files
- Databases including video, audio, text, images
- Models, algorithms, scripts, code, configuration files
- Applications, input, output, log files, simulation models, schemas
- Research records (correspondence, project files, grant applications, ethics/IRB applications, technical reports, research reports, master lists, signed consent forms)

Research data points are individual facts collected to validate research findings; they are not in themselves information. Information is produced from the processing, organization, structuring, and/or presentation of research data in a given context to make the data useful. To that end, sharing research data enables the additional processing, analysis, and presentation of data to make it useful to more people.
Why Share Data?

Over the past decade, much has been written on the promise and potential of data sharing. Motivations for data sharing are legion, from the prospect of new discoveries to improved research transparency and greater interdisciplinary collaboration, and they span areas including medicine, informatics, and engineering, as well as digital humanities and the like. Releasing data as a product of research increases the value of public funding by allowing scientists to ask new questions of extant datasets and enabling “big data” projects such as addressing climate change, mapping the human genome, and charting the sky. Researchers in some fields of study, such as astronomy and genomics, have more readily embraced data sharing, while those in other fields—particularly researchers with small, disparate data sets—may be more reticent to share data. With funders and publishers mandating data sharing, the scientific endeavor is slowly moving toward openness, but challenges, both real and imagined, continue to impede data sharing. The benefits of data sharing may be extolled at length, but rarely do funding agencies, publishers, and universities define the term data sharing or give particular directives for how data—including experimental context, methodologies, and workflows—must be shared. Researchers often attach conditions to data sharing, such as the first right to publish results from the data, a requirement of proper attribution, and the right to say who may and may not use their data and for what purposes. Researchers will often share data only to the minimum extent required by their grant by putting their data anywhere that generates a digital object identifier (DOI), without considering the best location for their data or curating their data to a level where it might be understandable to future users. To understand the conundrum of data sharing, it’s useful to consider the complementary and competing interests of the various stakeholders embroiled in the data-sharing endeavor. While it’s pleasant to imagine researchers, institutions, libraries and repositories, funding agencies, publishers, and the general public working together in lockstep toward the common goal of scientific achievement and mutually beneficial progress, in reality each stakeholder group has its own priorities and perspectives that influence its respective attitudes. The following describes how each group engages in research data sharing and touches on its specific concerns in this space.

Researchers

As the primary collectors of data, researchers often feel ownership of their data and protectiveness over how it is shared. While most researchers are passionate about sharing information and disseminating their work broadly, there are still legitimate concerns from the research community regarding data sharing. Academic research is an extremely competitive endeavor, and many researchers are slow to share data for fear of being “scooped” or to ensure they “finish” their work before others have an opportunity to analyze their data. Some researchers are afraid their data will be misinterpreted or misused by people who might misrepresent or cherry-pick their data. Infrastructure and best practices are emerging to address researcher concerns about data sharing, and scholarly mechanisms are emerging to reward researchers for sharing data. The NSF has encouraged researchers to list data sets alongside publications in their NSF Biographical Sketch “Research Products” section to elevate data sets to a legitimate research product. Studies are beginning to show that researchers who release data alongside publications have higher citation rates than those who do not release data as a supplement to articles. The nonprofit organization DataCite has created standardized...
formats for data citation and provides persistent identifiers (DOIs) for data sets to aid in locating, identifying, and citing them. As credit mechanisms for data sharing continue to evolve, researchers may respond to the incentives related to making their data available, usable, and citable.

The gold standard of science has always been reproducibility. Sharing the data that underlies publications allows for greater transparency and the ability to validate research findings. In classical science, it was normative to report results in a way that would allow them to be reproduced. It would be ironic and problematic if new data practices and computationally intensive approaches weakened this tradition. Reproducibility is only possible when data, methods, and code are available to secondary users. Data sharing also leverages research and encourages cross-disciplinary and interdisciplinary work by allowing scientists in other fields to ask new questions of existing data.

Libraries and Repositories
Libraries have traditionally collected, preserved, and made accessible the intellectual outputs of their institutions. As data increasingly becomes a research output on par with monographs and journal articles, library faculty and staff consider how they can best acquire and manage data alongside other research resources. A significant amount of research data comes from academia and from collaborations between academia and industry. In the absence of subject-specific or nationally funded repositories, stewardship of the resulting research data can be managed, curated, and shared from within the institution. Libraries, which often leverage existing institutional repositories or dedicated data repositories to accomplish this task, shoulder the costs of data storage and curation, which can be a strain on finite resources.

Implicit in establishing data-curation services is developing a digital infrastructure that supports secure and effective modes of data sharing. Many libraries have worked with researchers to develop requirements for a data repository, including the ability to accept a wide range of file formats, embargo periods, at least 10 years of secure storage, persistent identifiers, suggested data citation, and robust metadata for discoverability.

Libraries have also taken on a pedagogical role in data management, building on their traditional role in bibliographic instruction and expertise in information management. Expanding instruction to best practices in data management, visualization, and analysis will enable more researchers to manage their data effectively, contributing to trustworthy, discoverable, and replicable research.

Higher Education Institutions
An important goal of higher education institutions is to produce new knowledge and to share that knowledge broadly. Much has been said about the benefits for institutions that share data, but the benefits have not been easy to quantify. Thus, while many higher education institutions have taken a positive stance on data sharing, few have funded robust and comprehensive institutional data-sharing services and the attendant technical infrastructure. As funding-agency mandates become more pervasive and researchers more vocal about their data needs, universities will feel increasing pressure to provide and fund data services.

At its core, sharing research data is in line with the university mission to contribute knowledge and information to benefit research and society at large. From a marketing point of view, it can help establish an institution’s reputation as a leader in specific domains and thus can provide a competitive edge for
funding and faculty and student recruitment. Emerging data scientists are drawn to interesting problems and need access to data. Universities, however, may be less inclined to share data with potential commercial value or that gives their researchers a competitive edge over those at other institutions.

**Funding Agencies**

Funding agencies have been the biggest force in moving data sharing into the collective consciousness of researchers through data management and sharing mandates. In 2003, the National Institutes of Health (NIH) became the first funding agency to require a data management plan (DMP) in grant proposals for all funding requests over $500,000. The NSF instituted a DMP requirement in 2011 for grant proposals of any amount but has only recently begun using DMPs in postaward assessment. In 2013, the White House Office of Science and Technology Policy (OSTP) issued a memo directing all federal agencies engaged in more than $100 million in annual research and development to create policies that would make the direct results of research—including publications and data sets—available to and useful for the general public. As a result, all of the largest federal funding agencies require recipients to include a DMP as part of their proposals and make the results of their research publicly available. Each agency has guidelines for data management that should be consulted when developing a research proposal; the University of Minnesota Libraries has created a table that outlines funders’ responses to the OSTP memo.

The motivation toward data sharing for funding agencies is clear. Funding agencies spend billions of taxpayer dollars for the purpose of advancing scientific progress and benefiting society. The data collected as a result of this funding has tremendous value and should be shared as a product of research alongside publications. Despite their commitment to data sharing, most funding agencies have not provided clear instructions or mechanisms for data sharing. As a result, some researchers see this mandate as a drain on the funds and resources against their primary goal for a project. However, many funding agencies consider data sharing and management as the “new normal” for research in the 21st century. Researchers and institutions need to find ways to accommodate this new standard expectation.

**Publishers**

Publishers, both commercial and open source, benefit from research data sharing in a number of ways. When data sets used in publications are shared, the value of the articles increases by providing the data for verification and citation. This verification and citation can help enhance the transparency of how analyses are performed and the conclusions from such analyses; providing articles based on verifiable and reproducible experiments can in turn improve the reputation of a given publication. However, like funding agencies, most publishers do not provide data curation and repository resources. As with new expectations from funding agencies, this expectation from publishers and readers is pushing researchers and institutions to provide these resources.

**General Public**

A final stakeholder is the general public. When researchers share data, the general public has access to the results of research funded with taxpayer dollars. Exposing data to all who are interested enables the results to be checked, and additional results and implications can be discovered. Public access to data is increasing and of acute interest to data scientists. The inherent transparency of sharing data also enhances the public trust by allowing independent verification. Since a significant majority of research
data is funded by public money, one could argue that the data is partially owned by the public. Another benefit of open research data is the ability to leverage it to teach students how to process and analyze data. Lastly, well-curated data sets are a tremendous shared resource for K–12 and undergraduate education because they can help teach students how to reproduce science and replicate known results.

**Barriers and Considerations**

In addition to the competing interests of numerous stakeholders, a variety of barriers and considerations have hindered the sharing of research data. Even when an individual or group is motivated to share data, social, technical, and legal realities complicate the free and easy flow of data. The following section unpacks a selection of obstacles researchers must consider when sharing or reusing data.

**Social and Cultural Influences**

Data sharing is significantly hindered by social and cultural influences. At present, scientists are rewarded for discovery and publication, not for making data publicly available. Furthermore, research data is not typically documented to the extent necessary for unknown future use. It is often not required as a condition of a research award, and even if it were, few researchers are trained to do it well. Fortunately, the library community is stepping up to assist researchers with tools and formal training, yet library expertise with research data sharing is still relatively nascent and unevenly distributed.\(^9\)

In early 2016, an editorial in the *New England Journal of Medicine* entitled “Data Sharing” created something of a firestorm.\(^10\) The authors used the phrase “research parasites” to describe researchers who increasingly demand access to data for secondary use. Some researchers express fear that others will misuse or misinterpret original data; this concern may be valid, given the lack of documentation. For some, it is a fear of being “scooped” or a concern that they have further research to perform on the data set. Sometimes it is the natural concern of being scrutinized or proven wrong. Some researchers argue that data sharing should happen through collaboration or authorized follow-on studies, not through independent use of research data—in other words, only in the ways data sharing has traditionally occurred. But do traditional methods work in a world where issues such as big data, data from unique or restricted instruments, and questions of scientific fraud or error have become critical? Even if they work in a limited sense, are we optimizing limited resources by requiring researchers to unnecessarily and artificially start from scratch when similar data already exists?

---

**An Ideal Use Case**

In 2003, funding agencies joined with researchers from universities, nonprofit groups, and medical imaging/pharmaceutical companies for a unique public-private partnership and collaborative research effort to find the biological markers that show the progression of Alzheimer’s. The project—the Alzheimer’s Disease Neuroimaging Initiative, or ADNI—has expanded worldwide into WW-ADNI. A key feature of ADNI has been its focus on collaboration and rapid data sharing. Multiple repositories store the data redundantly at institutions worldwide. Uniquely, no single researcher or institution owns the data. Both data and ideas are shared. The approach allowed the study to be conducted at a scale not possible by individual researchers and research organizations. The NIH recently awarded another $40 million in funding to extend this project. With the goal of accelerating the pace of analysis and findings, ADNI investigators make their collected data widely available. The study has been used in over 1,200 research papers. MRI- and PET-scanned brain images, as well as clinical, genetic, and fluid biomarker data, are available to qualified researchers worldwide through a web-based interface. To date, more than 8,500 researchers have sought access.
A series of articles published in *Nature* examines another dimension—the issue of cognitive bias in research.\(^1\) The articles suggest that the scientific community should invite scrutiny, essentially across the entire research life cycle. Crowdsourcing the design, conduct, and analysis of research and research data will result in better science and thus should be pursued. To do this requires a culture change in academic research that begins at the funding level.

**Data Ownership**

A major goal of research data sharing is the suite of new discoveries that can be made from reused data. Clear data ownership rights are important because intellectual property, data licensing agreements, and patents provide a framework under which data sharing can occur. Without clear rights attached to data sets, future researchers won’t know whether or how they can reuse data.

Until recently, the question of data ownership in academic institutions was rarely discussed except in cases of exceptionally large or valuable data sets such as the [Sloan Digital Sky Survey](https://www.sdss.org/) or the [Utah Population Database](https://uupsis.utah.edu/). For the average researcher or research team collecting small to medium-sized data sets in the hopes of answering specific research questions, the data rarely left the researchers, and therefore ownership appeared to lie with them. However, as research has become increasingly data-driven and data sets are treated as valuable research products, funding agencies, institutions, and researchers are identifying data ownership and IP restrictions in data management plans before data collection even begins. Studies show that researchers are still unclear on data ownership. A 2012 study at the University of North Carolina asked researchers about their views on data ownership; 2,743 researchers responded (see table 1).\(^1^2\)

**Table 1. Who owns research data?**

<table>
<thead>
<tr>
<th>Response</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher</td>
<td>46%</td>
</tr>
<tr>
<td>University</td>
<td>15%</td>
</tr>
<tr>
<td>Public</td>
<td>9%</td>
</tr>
<tr>
<td>Funding agency</td>
<td>8%</td>
</tr>
<tr>
<td>Publisher</td>
<td>1%</td>
</tr>
<tr>
<td>Other</td>
<td>15%</td>
</tr>
<tr>
<td>No opinion</td>
<td>7%</td>
</tr>
</tbody>
</table>

The results of the survey show two interesting facts. First, as recently as four years ago there was no clear consensus on data ownership and, second, there are multiple stakeholders that may be in a position to assert some claim over research data. Another study conducted at the University of Oxford in 2012 showed that out of 11 data management tasks, researchers had the lowest confidence dealing with copyright, licensing, or other intellectual property issues related to data sets.\(^1^3\)
One reason researchers are unclear on data ownership is because academic research data falls in a gray area between proprietary data (such as the Google search algorithm or the formula for Coca-Cola), which is clearly closed and subject to trade-secret law, and government data, which is open by default unless it’s confidential or sensitive in some way.14 Academic research data is often funded with public monies, either through the research institution or an outside funding agency, yet until the 2013 OSTP memo, most researchers were not required to make their data openly available except upon request.

Another complication is that it isn’t clear whether data are even eligible for copyright protection. Under the law, facts lack the requisite creativity to be copyrightable.15 As the preeminent expert on copyright law, Melville Nimmer said, “The discoverer of a scientific fact as to the nature of the world, an historical fact, a contemporary news event, or any other ‘fact’ may not claim to be the author of that fact. If anyone may claim authorship of facts, it must be the supreme author of us all. The discoverer merely finds and records.”16 However, if data are selected and arranged in specific ways, as, for example, "collection and assembling of preexisting materials or of data that are selected in such a way that the resulting work as a whole constitutes an original work of authorship," they might be eligible for copyright.17 Lastly, data laws are not harmonized worldwide, which is particularly troublesome because so much scientific discovery crosses national boundaries.

A final complication is that there seems to be widespread agreement that, owing to their intellectual capital contributions, researchers should have ownership (or at least some control over the management and release) of data they generate, in apparent opposition to the growing number of institutions that have created policies claiming institutional ownership of data. In the past 10 years, Duke University, Johns Hopkins University, Stanford University, the University of Kentucky, the University of Utah, Virginia Commonwealth University, and scores of other institutions have established policies that explicitly claim institutional ownership of the data. In a recent court case in California, a judge issued an injunction for a researcher to return a database to the University of California, San Diego, claiming that the database rightfully belongs to the institution, not the researcher.18

As a practical matter, however, most institutions leave the stewardship or “care and feeding” of data to researchers.19 Researchers collect, describe, manage, and even share data at their own discretion the vast majority of the time. When researchers do share their data—either because of funding mandate or goodwill—they generally select a repository for their data and even assign rights to the data or dedicate data to the public domain without consulting their institution. This raises the obvious question as to whether researchers should have the legal right to assign, for instance, a Creative Commons license to their data without the legal rights holder’s permission.

Legal Compliance

In the United States, federal and state regulations constrain or shape certain aspects of data collection, storage, and sharing. Requirements of regulations such as the Health Insurance Portability and Accountability Act (HIPAA),20 the Family Educational Rights and Privacy Act (FERPA),21 and the U.S. Department of State’s International Traffic in Arms Regulation (ITAR)22 must be addressed. Additionally, states and localities may have additional regulations for management of access to research data, especially where human subjects are involved. For example, the Minnesota Department of Health Data Practices Policy delineates rights both to the subjects of research and to the general public who would like to access that research.23 These do not prohibit data sharing but require a high level of data
stewardship. Noncompliance can adversely affect the entire institution, so it behooves universities to develop effective services and training to support researchers working with sensitive data.

**Funding**

Functional data sharing will not happen as an unfunded mandate. Researchers say that there is no room in grant budgets for the data management and curation that are essential for effective sharing. Funders have encouraged researchers to include data curation as a line item in the project budget, but many researchers are still unsure how to calculate those costs upfront. As discussed below, the technical and human resource requirements can be costly. There is also the perennial question about what should be the institutional investment (indirect cost recovery) versus grant support to sustain research data sharing.

Institutions increasingly recognize research data as an asset but—possibly due to culture, complexity, or cost—do not treat it as such. Many institutions do not provide costly infrastructure (or even guidance) regarding data management. When support is available, it is often not mapped to the research data life cycle (taking into account the differing requirements of data collection, analysis, preservation, and sharing). As a result, institutions may have no idea what data they have or how or where it is stored. Researchers pressed to solve their own data-related problems and without institutional support may employ a variety of low-cost solutions, from thumb drives to personal cloud service accounts. For these reasons, many institutions are moving to implement data-governance programs and documenting guidelines for data storage.

**Data Governance**

With data as a primary output and the role of analytics, institutions are recognizing the importance and necessity of managing data as a core institutional asset. As such, the governance of data—how data will be created, stored, used, archived, shared, and deleted—is at the forefront of data discussions. The focus has typically been transactional administrative data, but research data is increasingly a part of these discussions.

Data governance processes and policies define the responsibilities for the protection of institutional data and guide the management of data. Every person who creates or accesses data shares in the responsibility of ensuring that data is used according to institutional policies and guidelines. Data governance has been evolving and is a means to ensure the effective and efficient use of resources and capabilities in the management of data.

Data governance seeks to increase the integrity, transparency, auditability, accountability, and standardization of data through policies and guidelines. Establishing a sustainable data governance program that effectively addresses the way research data specifically will be managed can be difficult. Formal, documented, and repeatable procedures crafted to support regulatory and compliance requirements are recommended for all aspects of research data management. With respect to research data sharing, the governance issues that appear to be most pertinent are access and usage, privacy and security, and classification.

- **Data Access and Usage:** Data governance outlines who can access what data and for what purposes to ensure proper management of the data. For example, research data associated with institutional review boards and human subjects is highly regulated, with elements of complexity that can be difficult to navigate and understand. To start, stewardship of data should be clearly outlined in
the governance process, such as a RACI (responsible, accountable, consulted, informed) matrix, identifying who has the responsibility/authority to grant access to data and the procedure for granting access that should be followed. A process for requesting data should be established, including delineating planned use of the data being requested. Governance rules can ensure that a data-set recipient is authorized to access/view such information, understands the data, understands his/her responsibility as a user, and completes any needed training prior to access/usage. A documented process provides rules and standardization around data access and usage, including federated access (to address the various layers and types of research data). Policies, processes, and guidelines devised by the data governance program should address the usage of data to ensure that users are abiding by the principles of data access, privacy, and management and using the data for only the approved purpose.

- **Data Privacy and Security:** A well-defined and coordinated data governance program in conjunction with IT partners ideally reduces legal, regulatory, and ethical risks related to privacy and security. For example, a governance policy may address how the institution safeguards the privacy and security of research data subjects, both prior and subsequent to publication, while simultaneously ensuring that local, state, and federal privacy and security requirements are addressed.

- **Data Classification:** Policy, processes, and guidelines around the classification of data should be established to ensure proper requirements around various types of data. For example, data can be classified as public, private/confidential, or sensitive and can be de-identified accordingly. As part of an institution’s governance program, data classifications can be identified and defined and access policies and processes devised. The program should outline requirements for the storage, archiving, and deletion of data that adhere to institutional security policies. Rules need to be established for purging and archiving data, taking into account requirements for maintaining, preserving, securing, and accessing data according to regulatory and institutional policy that includes disposing of data based on classifications.

**Information Technology Considerations**

Research data sets can be too large to easily be moved or shared across an institution’s normal technical infrastructure. In such cases, researchers need to consult with their institution’s IT and (where available) data management departments or data curation teams (often found within libraries) to create a plan for moving, sharing, and ultimately preserving the data. Going forward, research institutions need to evaluate their current technical infrastructure and develop a long-term strategy for creating one that will support high-performance computing and massive data sets.

As part of an institutional IT infrastructure strategy, it is important to make available storage that is secure and conforms to privacy regulations such as HIPAA and FERPA (see the Legal Compliance section for more information) or other restrictions such as ITAR (under which data must remain on U.S. soil). If state or federal regulations are violated, punitive actions could be levied on both the institution and the individual.

**Metadata**

The National Information Standards Organization (NISO) defines metadata as “structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.” Metadata includes data associated with either an information system or an information object.
for the purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation. Context and metadata are essential; without them, data is not only less useful but also subject to misuse and erroneous conclusions.

In addition to general descriptive metadata, there are discipline-specific standards, some of which may be required from funding agencies to be applied research output. For example:

- **ISO-19115** (geographic information): This standard defines “how to describe geographical information and associated services, including contents, spatial-temporal purchases, data quality, access and rights to use.”
- **MINSEQE** (Minimum Information about a high-throughput nucleotide SEQuencing Experiment): This standard enables the unambiguous interpretation—and facilitates reproduction—of the results of the experiment.
- **FITS** (Flexible Image Transport System): This standard is for data exchange and archiving used by astronomers.

**Interoperability of Sharing Platforms**

Interoperability is the first step in effective data management. Without a functional, secure ability to share data, researchers are unable to effectively collaborate, data curators are not able to evaluate the data for metadata and best practices, and sharing data for reproducibility and reuse becomes problematic. In order to share data effectively, institutions need to invest in secure storage, significant bandwidth, and a set of tools to securely allow data to be shared.

Data is created in a myriad of types and formats, some agnostic, others that require proprietary software to view and manipulate (e.g., MATLAB, Stata). Agnostic formats such as comma separated values (CSV) are recommended because they enable open sharing and because obsolescence is less of a concern for agnostic formats. Proprietary platforms, on the other hand, are updated from time to time, rendering previous versions unusable and unworkable with later data sets or no longer supported at all.

**Data File Formats**

Data sets may need proprietary software to be seen or manipulated, but wherever possible, CSV or some similarly software-agnostic format should be requested from the researcher. When proprietary software is required, this should be noted in the metadata.

Repositories should set a timeline with regular reviews of data file formats to assess the risk of digital obsolescence and have a plan in place for converting data threatened with obsolescence into a useable format with the least amount of loss. If researchers do not intend to put their data into a repository, planning for digital obsolescence should be included in a data management plan drawn up before the data collection has begun.

**Contractual Issues**

Contractual issues may arise with commercial facilities used to house or manage data. As noted earlier, the proliferation of inexpensive or free cloud storage may lead researchers and IT support to consider its use. However, the selection of cloud storage products should be carefully reviewed.
Most cloud storage services outline what they provide in a service level agreement (SLA). Institutions and IT support should consider adopting or developing a process for certifying a cloud service as appropriate for use as a digital repository. The International Organization of Standards (ISO) has developed such a process and documented it in its “Audit and Certification of Trustworthy Digital Repositories” (ISO 16363).  

Ownership of data is one key issue to consider. Some providers consider any data stored on their servers as their property, and in these cases the SLA grants you access to the data. Others concede ownership to the creator but stipulate that they may access and use any data stored on their servers. For example, Apple’s iCloud service has a provision of “No Right of Survivorship” that stipulates that any rights to data terminate when the account owner dies. Google asserts the right to terminate service for inactivity exceeding 180 days.

Another important concern is the permanence of data. Most cloud storage providers do not accept responsibility for damages, corruption, or the loss of data. They all advise that users maintain their own backup of any data stored on their services. In addition, for example, Google may discontinue any service or feature for any reason without liability.

Institutional agreements are generally constructed to address these issues and are becoming more readily available. However, researchers sometimes make use of personal invocations of commercial products that are not bound by these agreements.

**Data Management**

To ensure long-term data preservation, access, and reuse, data collectors must manage their data according to a developing set of best practices. These activities and practices, collectively referred to as data management, start at the conception of a project and extend beyond the life of the research project. Key data management activities include creating a data management plan, de-identifying data that includes personal identifiers, capturing appropriate metadata and documentation, converting data to open formats, and archiving data at the end of a project in an appropriate repository with a persistent identifier. Proper data management increases the value of a data set and gives necessary information to secondary users, including detailed methods, provenance, and information about the quality of a data set.

Most researchers, however, do not have formal training in data management and rarely employ data management practices at a level that facilitates data reuse. While requirements for data management plans have prompted researchers to consider the life cycle of their data in advance of a project, few research teams capture sufficient documentation to enable a secondary user to understand and reuse their data (see figure 1).
Librarians, with their long history of describing, organizing, and archiving information, have taken a leading role in data management. Many university libraries have hired data management librarians to assist researchers, train students and faculty in best practices, and oversee submission of data sets to institutional repositories. Even in the absence of a dedicated data librarian, copious online resources exist to train researchers in data management best practices.\(^{30}\)

Despite the wide variety of available resources, there are reasons that researchers do not incorporate data management best practices in their research. In addition to lacking formal training, researchers are not properly incentivized to spend significant time, energy, and money curating their data for unknown future users. Most researchers would rather spend time on analysis and scholarly writing than creating detailed codebooks or README files, especially when the current system rewards discovery and publication rather than the generation of reusable data sets. It is likely, however, that as funding agencies and journals continue their push for data sharing, as universities build and expand data services and infrastructure, and as young researchers are trained in data management principles, data management will be viewed as a core function of research than an ancillary activity.

**Training and Education**

The research data life cycle begins well before any data is collected. Researcher training should include the creation of effective data management plans and applicable standards (both discipline-specific and general) for data description, technical infrastructure, and sharing/preservation/reuse.

Traditionally, cataloging and now metadata librarians have been trained to create and apply descriptive metadata, usually within the AACR2, RDA, and MARC standards and mostly applied to books and continuing resources such as journals. With more institutions now needing research data to be described,
these standards may no longer be optimal for the description of research outputs such as code, software tools, and data sets. As a result, staff may need to be trained as researchers are in the application of the data life cycle to metadata needs.

Instruction on data handling and cleanup should become standard for all graduate school curriculum. The Software Carpentry Foundation is dedicated to bringing "researchers in science, engineering, medicine, and related disciplines the computing skills they need to get more done in less time and with less pain." Those who complete the courses in skills such as R, Python, UNIX, OpenRefine, and Git are encouraged to participate in Software Carpentry instructor training themselves and to pass these skills on to others in their discipline.

Miller’s concept of the magic number\(^3\) —which states that the human’s working memory can retain an average of seven items, plus or minus two—is often used as a measurement of the effectiveness of search results. This is considered in light of how much metadata is applied to an object and how much of that metadata is unique, which affects meaningful search results. In the world of research data and repositories, this becomes a crucial consideration, especially in light of the self-deposit feature of some repositories. If left to themselves, researchers may name data files in nondescriptive ways (e.g., “data_files.zip”) and may add little descriptive metadata, creating cumbersome results lists for searchers. At a minimum, a research data file should include a descriptive title, the owner’s name, the file size, the file format, and the date of data collection.

Educating researchers on the data life cycle is critical for the capture and application of metadata. Without prior planning, the correct type of metadata may not initially be captured, leading to resource-intensive application of metadata after the fact or the loss of crucial information when graduate students, who are often most knowledgeable about the data and its dependencies, leave the institution.

Data changes over time: Instruments are recalibrated and experiments are rerun. As of now, there is no standard for what constitutes a new “version” of a particular data set. The research process can generate raw data, cleaned data, processed data, and portions of that data used to generate graphs and charts for publication. Data can be processed with proprietary software that may need to be purchased by the reuser. All of these aspects of research data will require metadata that AACR2/MARC does not specifically address (unless in a note). Data often can have relationships with a number of dependencies, including software packages and in-house-created scripts and can even be created out of two separate data sets. All of this information is crucial to reuse and reproducibility, and new standards and the meaning of a version will need to be discussed and decided by the research community.

**Conclusion**

In today’s computationally intensive, big-data-focused, highly collaborative research environment, data sharing has become very important for advancing research knowledge. Data sharing increases the rate of new scientific discoveries, improves research transparency, and creates opportunities for collaboration across disciplines and institutions.

Yet the competitive nature of research, historical and cultural practices, the scarcity of formal support structures and training, and the complexities of the data itself make effective data sharing difficult. Solving the data sharing problem will require collaboration between the researchers (producers and consumers), institutions (mission, IT, legal/IP), libraries and repositories (collection, curation, dissemination), funding
agencies (sponsors, funding), publishers (dissemination), and the general public (sponsor, consumer, beneficiary).

At the top level, we need consistent policies that encourage and require data sharing. Strategies and standards for metadata management and other best practices could be more prescriptive to drive the desired outcomes. Institutions need to do more to support researchers in order to facilitate data sharing. This includes having clear policies, procedures, and infrastructure for data classification, data governance, and data curation. Training for methods in data capture, data curation, and data dissemination should be integrated into the core curriculum alongside data analysis. As discussed earlier, funding is available but guidance is generally not, so data sharing is not executed effectively.

Working collaboratively, we can accelerate adoption of newer technologies and methods, such as creating shared public ledgers (e.g., blockchains) to record transactions and provide access to particular data. This approach would not store the data itself. It is designed to allow transparent validation of the appropriateness of requests and grant access to data. This approach could alleviate the ownership and security issues. Can we develop or adopt a standard for a digital unique identifier that could help track data provenance? Can we standardize on a code repository such as Github?

Are we ready for greater openness? Research data sharing holds the promise of significantly increased progress in the sciences, arts, and humanities in this age of data-driven discovery. Strong institutional leadership and collaborative interdisciplinary teams are needed to remove the data sharing roadblocks and thereby elevate the impact and efficiency of all research.

Authors

Special thanks go to the following ECAR Research Data Sharing (ECAR-RDS) Working Group authors of this report:

David Ackerman
Chief Digital Officer and Associate Vice President, .edu Services
New York University

Guy T. Almes
Director, Academy for Advanced Telecommunications and Learning Technologies
Texas A&M University

Kirk M. Anne
Assistant Director, Systems & Networking SUNY College at Geneseo

Wendy Bezotte
Data Governance Program Manager
University of Michigan—Ann Arbor

Rebekah Cummings (ECAR-RDS Co-Chair)
Research Data Management Librarian
University of Utah

Curtis W. Hillegas
Associate CIO, Research Computing
Princeton University

Valerie E. Polichar
Director of Research IT Services
University of California, San Diego

Juliane Schneider
Metadata Specialist
University of California, San Diego

Shady Shehata
Principal Data Scientist
D2L Corporation

Brian Stengel
IT Service Owner for Research
University of Pittsburgh

Ralph J. Zottola (ECAR-RDS Co-Chair)
CTO, Research Computing
University of Massachusetts Central Office

“You can have data without information, but you cannot have information without data.”
—Daniel Keys Moran, programmer and science fiction writer
Citation for This Work


Notes


3. The categories of research data are derived and adapted from the Research Data Management Resources website at the Boston University Libraries.


5. NSF Grant Proposal Guide 16-1, “Chapter II: Proposal Preparation Instructions,” January 25, 2016. The guide suggests that this section should include products and notes that “acceptable products must be citable and accessible including but not limited to publications, data sets, software, patents, and copyrights.”


8. We encourage readers to visit the University of Minnesota Libraries web page on Funding Agency Guidelines for more about how the National Endowment for the Humanities, NIH, and NSF guidelines impact how research data is shared.


15. See the U.S. Copyright Office page “What Does Copyright Protect?,” which states, “Copyright does not protect facts, ideas, systems, or methods of operation, although it may protect the way these things are expressed.” However, as noted in U.S. Copyright Office, “Copyright in Derivative Works and Compilation,” Circular 14, October 2013, complications of data may be copyrightable “if the materials are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes a new work.”


19. For more about data stewardship, see Nickolas Backscheider et al., *Establishing Data Stewardship Models*, working group paper (Louisville, CO: ECAR, December 18, 2015).


24. Data governance is discussed at a high level in this paper as a potential solution for safe and secure research data sharing. For more information on institutional data governance (retention, stewardship, ownership of data), see Douglas Blair et al., *The Compelling Case for Data Governance*, working group paper (Louisville, CO: ECAR, March 19, 2015); and Backscheider et al., *Establishing Data Stewardship Models*.


26. See “metadata,” DCMI Glossary. In general, metadata is “data about data”; functionally, it is “structured data about data.”

27. See “geospatial metadata,” Wikipedia.

28. See “Primary Trustworthy Digital Repository Authorisation Body.”

29. To learn more about data management plans (DMPs) and for guidance on developing a DMP service, see Michael Fary and Kim Owen, *Developing an Institutional Research Data Management Plan Service*, working group paper (Louisville, CO: ECAR, January 8, 2013).

30. See, for example:
   - **MANTRA Online Modules**: MANTRA is a free online data management course offered by the University of Edinburgh for those who manage digital data as part of their research.
   - **DataONE Online Modules**: This resource features education modules in PowerPoint that you can download and incorporate into your teaching materials; materials are licensed as CC0, and you may enhance and reuse for your own purposes.
   - **ICPSR Data Management and Curation**: ICPSR is a data repository for social science data. The Data Management and Curation section of its website offers high-quality documentation on how to create DMPs, guides to archiving data, and a guide to creating codebooks.
   - **ICPSR Data Management Tools and Services**: This resource offers a compilation of useful tools and services for managing research data.
   - **Coursera Course in Research Data Management and Sharing**: This is a free online course taught by professors from the University of North Carolina at Chapel Hill and the University of Edinburgh.