SECTION 2.2.2

MANAGING, SHARING, AND PUBLISHING DATA

Susan Ivey, Sophia Lafferty-Hess, Peace Ossom-Williamson, and Katie Barrick

INTRODUCTION

Research data management (RDM) is a set of foundational practices and decisions regarding the maintenance and care of data produced during a research project. RDM can help ensure the stability, accessibility, and transparency of research materials and increases impact of research through data citation and reuse. The growing interest in RDM was partly driven by public interest in greater research transparency. As noted in section 2.2.1, "Introduction to Open Data," this interest is reflected by an uptick in action taken by policymakers and federal grant-funding agencies, a driver for information professionals' continued involvement in RDM. The first part of this section covers best practices for managing data so they can be understood and used, and the second part addresses sharing research data.

So why is data management important?

RESEARCH DATA MANAGEMENT

DATA SET STABILITY AND LOSS

At any point in the research cycle, there is always a chance of accidental data loss or destruction. Data loss or destruction occurs in a variety of ways ranging from software or computer failure to theft of hard drives and computers, lab members leaving and taking data with them, and even natural disasters. If steps are not taken to store, back up, and secure data, the consequences may be severe. Not only are data lost, but labor and time investments as well, which can set a project back. Not all data can be replaced.

RESEARCH ACCESSIBILITY, TRANSPARENCY, AND IMPACT

While levels of accessibility differ between disciplines, data management and data sharing help enable greater trust in research (see section 2.2.1, "Introduction to Open Data"). Other researchers can use those data to replicate or build on the original research. They can help expand a layperson's or disciplinary newcomer's understanding of a particular project beyond

the published article and results. Sharing the details of a study's methodology and analyses allows greater transparency and replicability or the achievement of consistent results within studies that seek to answer the same scientific question.[1] Good data management can support sharing these research outputs in a manner that the data can be reused, which reduces some of the costs of data collection.

BEST PRACTICES FOR RDM

Methods and data types vary across disciplines, and individual research projects can involve multiple data types as well. This can make planning and managing research data a challenge. The following general best practices focus on making robust directions and documentation available that will enable researchers to effectively work with, reproduce, and reuse theirs and others' data. Some steps are universal, while others are specific to the particular type of research being performed. Steps prior to and during research are described here.

Planning Ahead

Before beginning research, it is best practice to write a *protocol*—a study plan that details the purpose of the research, the data that will be collected or compiled, the variables that will be measured and tested, the tests that will be run, how data will be prepared, and how outliers will be dealt with. Protocols should be shared as transparently as possible prior to beginning research by writing and sharing protocols (through systems like Protocols.io or PROSPERO[2]) and creating a *data management plan* (DMP). In a DMP, the plans for collecting, storing, organizing, documenting, and sharing research data are described. A DMP lays the groundwork for beginning to document research methods thoroughly for future reporting when writing articles, preparing presentations, or sharing other research outputs. Using a DMP will also provide a vehicle for starting to talk and make decisions about areas to plan for, including where and how to store data and methods, consistency in file naming and versioning, and establishment of a schedule for updating documentation.*

Consistency

RDM is an active and ongoing activity throughout research. Once steps have been put into place, it is important to regularly engage in quality control and documentation. Researchers should provide accurate and detailed description of data and how they were collected from the start. Some tools, such as electronic lab notebooks (ELNs), have built-in ways to effectively document research methods, including data collection and description, which avoids manual steps. Other tools, such as Stata, R, and OpenRefine, that are used in data cleaning and analysis allow you to build "research pipelines" that define the exact actions that were performed. Research pipelines are discussed in more detail in section 2.2.3, "Supporting Reproducible Research." Researchers can also use manual documentation, including data dictionaries and README files, to consistently record relevant metadata.

Researchers should also make a copy of the original raw data prior to taking any steps, leaving the original file or files untouched. It is also useful to engage in versioning to take snapshots of the process of change that occurred. This can be done by saving a new file with a descriptive file name and version number each time a new step is taken in order to preserve and revisit what came before. These data files should be described as to what changes were

* For an example protocol, see https://doi.org/10.7910/DVN/23835/KA20IB. For an example DMP, see https://doi.org/10.5281/zenodo.1240420.

made and the software used to process them, and these files can be made available through sharing during and after research.

Transparency

When writing about the research that was conducted, researchers should refer to standard guidelines for reporting what was done, like those listed on Equator Network[3] These guidelines include things like the inclusion and exclusion criteria that were used, what data types were analyzed, the tests that were run, how outliers were handled, and what other interpretations and decisions were made, among many other methodological details that are not strictly about data. You can use repositories for sharing more detailed methodology and data. The "Sharing and Publishing Data" section below provides more information about data sharing.

Verifying Reproducibility

When beginning to engage in these practices, the best way to determine if they were done effectively is to see if a researcher can reproduce the research methods as reported and end up with the same results. Having an individual who was not involved in the original research try to replicate the research methods using the files and documentation that would be made available upon completion is a good way to confirm that the descriptions provided are clearly usable to someone who was not originally part of the study. If this can be done, the documentation is likely sufficient. Section 2.2.3 will discuss reproducibility in more detail.

RDM FOR QUALITATIVE AND HUMANITIES RESEARCH†

While scholars doing qualitative and humanist work do not always think of their objects of analysis as data, the objects of study and evidence available for analysis has expanded in these fields just as it has in quantitative fields. In other words, qualitative researchers and humanist scholars, too, are facing a data deluge.[4] And with that data deluge comes the concomitant need for data management. Most strategies for data management will be similar across fields, as described throughout this section. However, librarians supporting data management in qualitative and humanist fields should bear in mind the following additional challenges:

- Humanists may not consider their work *data*, preferring instead terms such as *evidence*, *citations*, or *primary sources*.[5] Reaching out actively to these research communities using the academic jargon specific to the discipline you are interacting with will increase your credibility and help researchers understand the value you bring to their projects.
- Researchers in qualitative and humanist fields may not only lack specific technical skills useful to managing research data, but they also may be anxious at being asked to develop or expand this skill set. As with reaching out to researchers in any field, developing relationships and searching for the simplest solutions to a researcher's problem will be important.
- Arts and humanities researchers often face barriers to managing and sharing data due to intellectual property issues, while qualitative researchers need to consider the privacy of their human subjects; for more on these issues, see section 2.2.1, "Introduction to Open Data."

† Thank you to Gabriele Hayden for contributing the "RDM for Qualitative and Humanities Research" content in this section.

SHARING AND PUBLISHING DATA

Sharing data can be defined most simply as making data available to others. However, sharing data ranges from providing a copy of as-is files to a researcher who requests them to publishing a well-described, curated data set in an open source format through a reputable data repository. In this chapter, the focus is primarily on sharing by *publishing data*. Sharing data involves extensive preplanning along with a commitment toward good data preparation, description, and quality control; these actions are resource- and time-intensive, but they have many benefits (see figure 2.1).

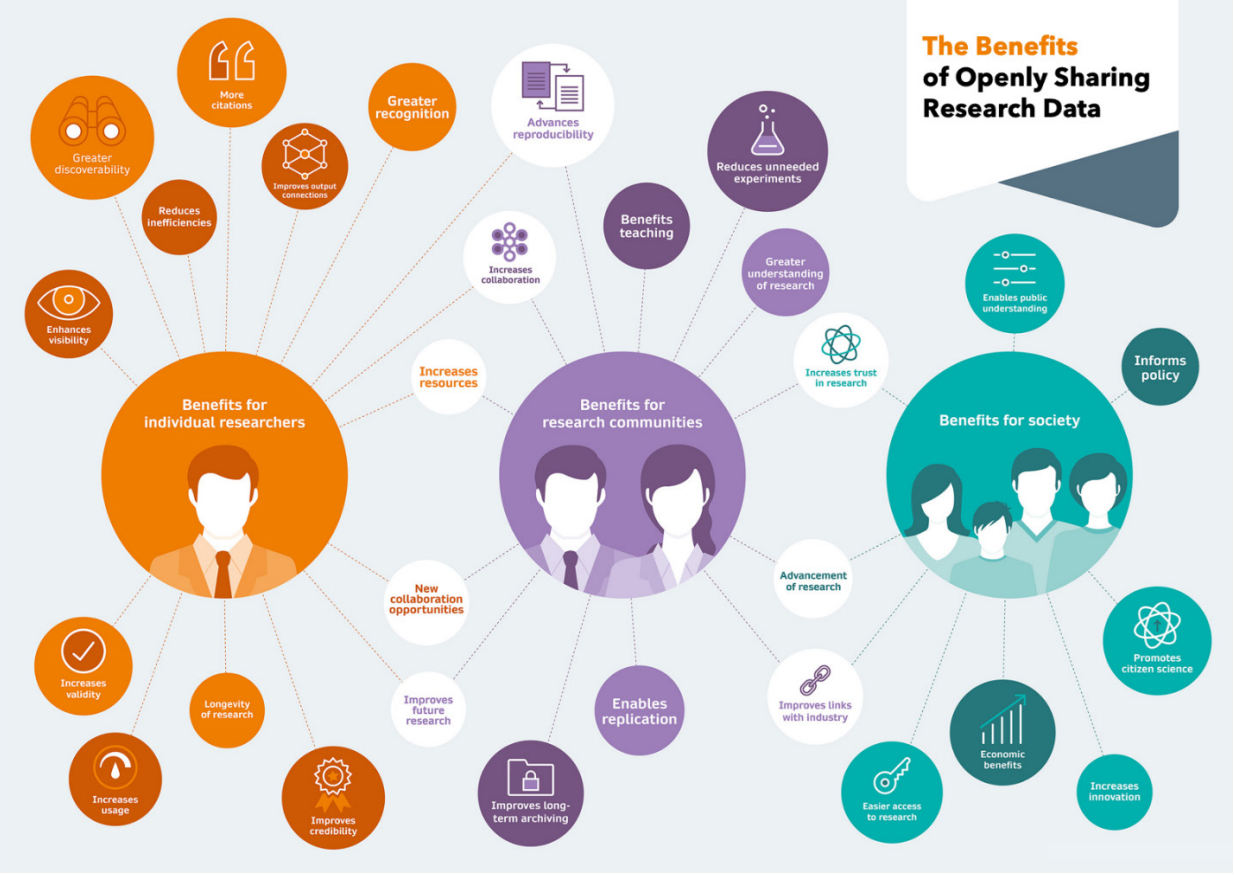

Figure 2.1
The benefits of openly sharing research data. Image adapted from Scientific Data and Mathias Astell, "Benefits of Open Research Data Infographic," Figshare, July 11, 2017, https://doi.org/10.6084/m9.figshare.5179006.v3. CC-BY 4.0, https://creativecommons.org/licenses/by/4.0/.

BENEFITS TO THE RESEARCHER

Complies with mandates—As mentioned in part I, federal granting agencies like the National Science Foundation have decided that the data that are gathered under grants from these agencies are public data because they come from public investment. Therefore, incorporating the tasks necessary for sharing data can help researchers comply with these and other external

mandates. Many scholarly associations and research journals are requiring data be shared publicly. Some examples of this include the joint statement for transparency in research coauthored by the editors-in-chief of several leading science journals[6] and the following statement from the American Political Science Association's *Guide to Professional Ethics*:

> Data access: Researchers making evidence-based knowledge claims should reference the data they used to make those claims. If these are data they themselves generated or collected, researchers should provide access to those data or explain why they cannot.
>
> Production transparency: Researchers providing access to data they themselves generated or collected, should offer a full account of the procedures used to collect or generate the data.[7]

Increases data reuse and impact—Sharing data immediately confirms ownership and provides researchers with a persistent link and time stamp of when the data set has been made available, which can be used if others make a claim to be authors or creators of a data set. Therefore, researchers can make data available as early as it is ready, switching from being "'scooped' to being 'credited,'"[8] and those data that have been shared can be used or reused in ways unimagined by the original researcher. Reuse and citation can enhance the reputation and recognition of the originating researcher because most data-sharing repositories publish data in a way that allows for measuring impact (e.g., number of views and downloads, persistent unique identifiers, and metadata). It has also been found that research studies published with accompanying data, particularly links to data deposited in repositories, have greater instances of citations than those published without the underlying data.[9] In some cases, published and well-cited data sets that have undergone quality control review can be seen as a research output in their own right, separate from any publication. Therefore, some researchers have chosen to cultivate, write grants for, and publish data sets for that particular purpose, rather than as underlying a research study. At present, there is limited recognition of this sort of publication from most promotion and tenure committees at US institutions of higher education, but this situation may change as publishing data is increasingly recognized as part of the research process due to aforementioned mandates and publisher requirements.

Allows for quality control—Somewhat less discussed is the impact of the published data on the researcher and the implications on their conclusions in their published articles. The aim of science is to progress toward innovation and solutions. Therefore, sharing data and making it available as immediately as possible allows for earlier recognition of errors in data entry, analysis, or other areas of data use and reduces the risk of harmful conclusions being implemented. This may protect the researcher from reputation damage later on.

BENEFITS TO OTHERS

One multifaceted benefit of sharing data is its implications for future use. Data that can be reused ensure maximum use of expended resources, improve the progression of science, and prevent data loss.

Ensures future access—Many researchers have experienced the phenomenon of requesting another researcher's data, for either reproducibility purposes or for building upon their research, and being told that the data cannot be found because they were saved somewhere forgotten years prior or that the data cannot be accessed because they are stored on files from proprietary software that is no longer available or on devices that cannot be read (e.g., floppy disks, CDs, and punched cards). Many *data repositories* include curation services, preventing data loss and ensuring future access.

Improves reproducibility and progression of research—Researchers without the resources to collect data can contribute to the progression of research by accessing and analyzing open data. Increased data availability also furthers the progression toward open science and research reproducibility because it allows for research methods and the underlying data to be verified by others, and it also allows others to replicate studies that have been performed. Likewise, researchers can more quickly build upon what has come before due to more immediate access to the data, and analysts at government agencies and institutions can make more informed decisions using available data. Since many publishers do not consider published data to be a previous publication of research, data can even be published during the research process, rather than waiting to accompany the research article, making access occur even sooner. Furthermore, *linked data*, or data that can be connected or pooled, allow for greater statistical research power and generalizability.

Limits the expenditure of resources and human investment—Sharing and preserving data along with sufficient documentation means that researchers—including the original researcher who produced the data—can access them in the future. This practice minimizes the time investment and potential risk for research participants (human or animal) and other resources expended each time data are collected, rather than repeating the data collection process.

The following sections provide information that can help with the selection of data-sharing platforms, which data should be shared (including human participant protections and copyright and ownership), how to prepare data for sharing, when data will be shared (e.g., during research, after publication, or after an embargo), and with whom (e.g., limiting access through system functionality and data use agreements).

DECIDING WHICH DATA TO SHARE

Data—Researchers must determine how many and which data to share. Do they need to share all of the raw data generated in a project? Only the analysis data sets? Only the data underlying a particular publication? Requirements from journals or funders may impact these decisions, and standard practice can vary by discipline. Researchers will also need to make assessments about what data are most useful or required for reuse and reproducibility and weigh other practical considerations, such as size and cost.

Code—In addition to sharing data, sharing software and other types of code is an important step for reproducibility. Code repositories, such as GitHub, and code-sharing platforms and tools, such as Jupyter Notebook, Code Ocean, or ReproZip, are popular choices.[10] Some general repositories, including Zenodo and Figshare, integrate with GitHub to allow for code sharing and citation, as they assign persistent identifiers to a specific release or version of the code.[11]

Lab notebooks and workflows—Open lab notebooks allow researchers to share and publish their research, including data, in real time on the open web. Researchers share their data and methods, whether successful or not, for anyone to see. The goal is to save time for other researchers, to uncover potential mistakes, and to advance science more quickly. Openlabnotebooks.org is a free resource for researchers around the world to create and share an open lab notebook.[12] Other platforms, such as the Open Science Framework (OSF), also allow for the open sharing of work processes and workflows where files may be made immediately available for public view and access.[13]

PLATFORMS FOR DATA SHARING

Researchers utilize a variety of systems, platforms, and methods in order to share their data. This section will provide context about the current environment of data sharing. The first

three methods—*data repositories*, *data papers and data journals*, and *open data portals*—are considered best practice for data sharing, as these methods tend to enforce overall curation including data organization, description, and licensing, which ensure that data are in their most reusable form. These systems also promote discoverability on the open web. Other methods described below that are no longer considered best practices for data sharing include web pages, supplemental material alongside publications, and available by request.

Data repositories—Often considered the gold standard for research data sharing, data repositories can provide storage, preservation, description, and access to data sets and metadata that describes the data set and its data files, thus aiding researchers in meeting the increasingly cited and required FAIR data principles.[14] There are many types of data repositories, including discipline-specific, general repositories that accept data regardless of the discipline, institutional repositories, and membership-based repositories that require a membership to deposit or access data. Researchers are generally advised to utilize discipline data repositories if one exists in their field for highest impact and visibility of their data, but general data repositories, institutional data repositories, and membership-based repositories can be valuable options as well.

It is important to understand that not all repositories are created equal, though, and a researcher must do their due diligence to determine if the repository they choose is the most appropriate, offers required features, and is trustworthy. One should consider things such as long-term sustainability plans, preservation policies, license options, embargo ability, file size and data set size limits, and robust metadata for increased findability and reusability. Datacite's Registry of Research Data Repositories (https://www.re3data.org) and FAIRsharing (https://fairsharing.org) are global registries of research data repositories that cover research data repositories from different academic disciplines and provide researchers, funders, journals, and academic institutions with a tool to search, compare, and evaluate thousands of repositories.

It is worth noting that there is a lot of current activity around assessment, evaluation, and certification of data repositories. US federal agencies, journal publishers, and other international entities are grappling with how to evaluate, and thus require or recommend, data repositories that are suitable for publicly funded and open data, and this work is expected to continue to impact the data-sharing landscape for the foreseeable future.

Data papers and data journals—Data papers, which are a fairly new type of publication, describe a data set by providing information about the context and content of the data package. Rather than analyzing the data, the purpose of a data paper is to promote data reuse. The actual data themselves are generally deposited into a data repository, and the two are linked via citations with a unique identifier (e.g., DOI). Data papers describe the data much more robustly than does the metadata that describes the data set in a repository, as they can provide more description and context. Data papers can be found in *data journals*, which are publications that contain only data papers. Methodology or data sections of standard publications may also refer to the underlying data and provide contextual information about the data set, although normally not as comprehensively as information found within a data paper.

Open data portals—Functioning like a catalog, open data portals contain metadata records describing and linking to open data, thus facilitating discovery and reuse of open data. Some examples include the National Institute of Mental Health Data Archive (NDA), the Indiana Spatial Data Portal (ISDP), the US government's Data.gov, and the European Union Open Data Portal (EU ODP).[15] DataPortals.org provides a comprehensive list of open data portals, which is curated by government agencies, nongovernmental agencies, and international organizations and is run by the Open Knowledge Foundation.[16]

Websites—Some researchers choose to share their data on a personal, project, or institutional web page. While sharing data on a web page or via a social or professional networking profile can help increase visibility of a researcher's work, it does not meet best practices for long-term preservation or sustainability of data.[17] It also would not fulfill funder or journal mandates for data sharing due to the lack of long-term preservation of the data themselves and of the web page. If researchers wish to share data on a web page, it is best if they do so *in addition* to depositing the data set into a data repository, linking to the persistent identifier of the data set.

Supplementary material alongside an article—The practice of sharing data as supplementary data to a journal article has been occurring for a number of years. As funding agencies and journal publishers increase their focus on reusability and reproducibility of data, though, sharing data as supplementary material no longer meets the data-sharing requirements of most funding agencies and journal publishers. Supplementary materials are not persistently available and have limited standardization in the file formats used and in their organization.[18]

Available by request—Another method of sharing data is when the researcher or a member of the research or project team grants access to data upon request, thus giving the researcher control of when those data are shared and with whom. This method, however, is one of the least open methods of sharing data and is not viewed as sufficient by funding agencies or journal publishers, as these requests are rarely fulfilled, particularly as an article ages.[19]

DATA CURATION

At the foundation of making data open are the RDM best practices explored in the first part of this section, which researchers will ideally implement during the active phase of a project. These set the stage for effective sharing; however, in most cases researchers in collaboration with information professionals will need to further prepare or *curate* data, documentation, and other associated materials to make them openly available in a structure and format that facilitates reproducibility and reuse. Below are a selection of key curation activities. For an even more comprehensive list, see the Data Curation Network's "Definitions of Data Curation Activities."[20]

Quality assurance—Data are complex digital objects that often go through multiple iterations as they are collected, processed, and analyzed. This research process can result in errors, missing information, or structural issues. Researchers and data curators have an opportunity to identify and address any quality assurance issues prior to publishing the data to ensure data accuracy and integrity. While these quality assurance checks will vary across data types and disciplines, some common issues include the coding of missing values, out-of-range values, structures that do not support portability or harmonization with other data sets, incorrect embedded metadata, and the potential inadvertent deductive disclosure of sensitive information.

Documentation and metadata—When preparing data for sharing, researchers should include documentation to explain the content of the data themselves and the context of the study. In some fields they may use structured metadata standards that support interoperability; other researchers might use unstructured files such as README files. Documentation practices vary but the presence of some form of documentation or metadata is essential for others to understand and effectively use the data.

File formats—During the course of a research project, data may be generated and analyzed using proprietary systems or equipment. While it is understandable to work within a format that is ideal for data gathering and analysis, in some cases these formats may limit access by

others in the future. When possible, researchers should consider using or converting files to open, standard file formats to allow for broader and longer-term access and to support interoperability.

Associated code, programs, and other materials—To reach the goal of more reproducible research requires sharing not only data but also associated code files, software programs, and any other files necessary to verify and reproduce the outputs from a study. What is needed to prepare a complete reproducible package again varies, but at a minimum one should include code files used to process and analyze the files, documentation on software and dependencies, and detailed instructions for reproducing the results. More computationally advanced methods also exist that package the computing environment within containers. See section 2.2.3, "Supporting Reproducible Research," to learn more.

TECHNICAL INFRASTRUCTURE FOR DATA SHARING

Publishing data in a repository allows researchers to take advantage of system functionalities that support making data easier to find, discover, access, cite, and preserve for the long term. In conjunction with curation actions described above and coupled with the legal and ethical best practices, these systems provide the technical infrastructure needed for realizing the FAIR principles and making data open.

Persistent unique identifiers—Repositories assign persistent identifiers to data sets to enable ongoing access and discovery. There are various types of identifiers, such as digital object identifiers (DOIs), handles, and Archival Resource Keys (ARKs). A persistent identifier ensures a link will continue to resolve to the digital object landing page even as infrastructure and URLs change or data sets are removed.

Metadata—Repositories enable discovery through metadata that describes a data set. Different repositories will use different metadata standards for description depending upon their community standards and disciplinary norms. General or institutional repositories often use domain-agnostic metadata standards such as Dublin Core or DataCite, which support the minimum fields needed for discovery, citation, and interoperability among systems.[21]

Citations—Through the use of standardized metadata and persistent identifiers, repositories provide the necessary information for the creation of data citations. Repositories may also provide structured citations that can be exported for ease of use. It is best practice for researchers to include data citations for any data set they reuse and include them in the reference list of their article—this facilitates tracking impact and reuse of data sets and helps elevate data sets as scholarly objects. The *Joint Declaration of Data Citation Principles* outlines the primary purpose, function, and attributes of data citations.[22]

Licensing and terms of use—Repositories support the use of standardized licenses or terms of use to communicate reuse requirements to others. Increasingly, researchers are using Creative Commons licenses for this purpose.[23] However, an underlying assumption of Creative Commons is that the person assigning the license maintains copyright for those materials, which for data is not always clear. For this reason and in the spirit of openness, some repositories and open data advocates encourage assigning a *CC0 Public Domain Waiver* or the Open Data Commons Public Domain Dedication and License, which releases data into the public domain without restrictions on reuse.

Preservation—The technical infrastructure behind repositories should ideally provide depositors with some assurances that data will be safely stored, backed up, and monitored to ensure ongoing access. When selecting a repository, considerations surrounding preservation plans and policies will help researchers make data openly available not only today but also

into the future. Repositories can demonstrate their trustworthiness through certifications, audits, and transparency in policies and procedures.

Embargoes—Embargoes are a request or requirement that data continue to be restricted for a specified period of time, and these can be placed for various reasons. It is currently common practice for a researcher to limit or completely restrict access to their data in order to allow time for completing and publishing their own work; however, there are benefits to making data immediately open. The most notable benefit is the ability of the data creator to claim prior ownership if there is a dispute or plagiarism concern. In the case of secondary research or commissioned research, restrictions may have been placed by the data owner or the company from which the request for research was commissioned. It is recommended that embargoes be avoided or limited to the length of time necessary to protect privacy and confidentiality and comply with research requirements.

Versioning—The majority of data-, code-, and file-sharing platforms allow for *versioning*, which provides data creators with the ability to make updates to the shared data. Using versioning, researchers can publish their data sooner and make updates by re-uploading the newer version of the data. Versioning may also be used to correct errors identified after publishing a final data set. The data citation in a data repository is changed to reflect the new version number. Therefore, previous files and data sets are still available, and each version can be tracked for provenance purposes.

ETHICAL AND LEGAL CONSIDERATIONS OF DATA SHARING

One of the first steps in ethical data sharing is understanding any potential limitations. There can be various reasons why data either cannot be shared or would need to have certain restrictions on access and use placed on them. While RDM greatly enables sharing data and open data, it is important to think critically about potential privacy and access concerns.

Sensitive and Restricted Data

Researchers have an ethical obligation to ensure they are not disclosing information that should be protected, including sensitive and restricted data. In many cases, these data come from human subjects. When dealing with data from human subjects, there is a potential for harm if data were disclosed or permission for data sharing was not granted during informed consent. Other examples of sensitive data include export-controlled data and geographic information concerning the location of endangered species or archeological sites. These data may be made available in other ways, including by request (as mentioned above) or via a protected environment—as is available virtually through the ICPSR virtual data enclave and physically in the US Census Bureau's Federal Statistical Research Data Centers.[24] (See additional information in section 2.2.4, "Ethics of Open Data.") Two specific types of sensitive or restricted data are

- *Personally identifiable information (PII) or protected health information (PHI)*—PII is defined as any information that can be used to trace or reasonably infer an individual's identity by direct or indirect identifiers. Types of PII data include, but are not limited to, names, social security numbers, street or e-mail addresses, biometric identifiers, place and date of birth, race, and religion. The National Institute of Standards and Technology (NIST) published a guide to protecting PII in 2010 that details additional types.[25] Protections for PII are detailed below under the regulation and policy section. Protected health information (PHI) is information in a medical record that can be used to identify individuals and typically generated or disclosed during health care service.

- *At-risk data* refers to data that if exposed, could potentially put subjects in harm's way. Biodiversity data is an example of at-risk data. Although imperiled animal and plant species are protected by the Endangered Species Act of 1973 in the United States and other laws elsewhere, publishing data on them openly may leave them vulnerable to poaching and environmental degradation. Consider the case of the succulent stealers: In 2015, a couple in Spain pieced together various sources of public data, including data from scientific journals, to track down locations of endangered succulents to poach and to resell.[26] Sharing biodiversity data carries risk and the solutions are not clear cut. Some members of the scientific community believe location data should be restricted, while others believe there is benefit in sharing this information. Tulloch and colleagues developed a decision tree aimed at assessing the risks and benefits of sharing biodiversity data; it may help librarians and researchers discuss data-sharing options.[27]

Most protected and at-risk data are regulated by regulations, policies, and rules designed to protect said data from theft, fraud, or exploitation. Table 2.4, while not an exhaustive list, highlights a few major pieces of legislation.

TABLE 2.4
Major laws intended to shield protected and at-risk data.

Policy	Year	Protections	Resources
United States: Health Insurance Portability and Accountability Act (HIPAA)	1996	Protects protected health information (PHI), information in a medical record that can be used to identify individuals and typically generated or disclosed during health care service	"Research," US Department of Health and Human Services, https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html
United States: Genetic Information Nondiscrimination Act (GINA)	2008	Protects individuals from discrimination in both employment and health insurance on the basis of their genetic information	122 Stat. 881 Genetic Information Nondiscrimination Act of 2008, https://www.govinfo.gov/app/details/STATUTE-122/STATUTE-122-Pg881 "Genetic Discrimination," National Human Genome Research Institute, https://www.genome.gov/about-genomics/policy-issues/Genetic-Discrimination
United States: Family Educational Rights and Privacy Act (FERPA)	1974	Protects the privacy of and governs access to student education records	"Privacy and Data Sharing," US Department of Education, https://studentprivacy.ed.gov/privacy-and-data-sharing
European Union: General Data Protection Regulation (GDPR)	2016	Aims to provide control to individuals over their own personal data and defines six personal data principles	"Principles Relating to Processing of Personal Data," https://gdpr.eu/article-5-how-to-process-personal-data/ European Commission: EU data protection rules, https://ec.europa.eu/info/law/law-topic/data-protection/eu-data-protection-rules_en GDPR.eu
United Kingdom: Data Protection Act	2018	The UK's implementation of GDPR	"Data Protection Act 2018," https://www.legislation.gov.uk/ukpga/2018/12/section/1

SECONDARY AND PROPRIETARY DATA

When researchers are using secondary data, they also are responsible for doing their due diligence to ensure they have the rights to share. Data may be proprietary, may have restrictions on redistribution, or terms of use may be unclear. *Proprietary data* refers to collected or generated information that is controlled by a company or other organization and restricted in how it can be accessed or used. The restrictions are often documented in contracts or other legal documents. Proprietary data privacy concerns typically come into play when a research project includes private sector organizations as partners or funders, which is common in pharmaceutical and biotechnology research. Proprietary data may be accessible through terms of use or data use agreements (more on these below) wherein restricted or nonpublic data can be shared with registered users. Researchers should share secondary data only when there are no legal barriers in place.

COPYRIGHT AND OWNERSHIP

Within the United States, there are questions surrounding whether certain types of data can be protected by copyright. While the specifics of this legal discussion are beyond the scope of this section, current precedent states that facts cannot be copyrighted and that there must be some aspect of creativity to copyright a work. Therefore, while applying copyright to data is ambiguous, even with primary data, researchers need to understand who *owns* the data. Does an institution claim ownership? A funder? What are the agreements with collaborators regarding ownership? And given these agreements, are there any expectations regarding sharing? These are all important questions to answer prior to ethically sharing data. Researchers should share primary data only when there are no barriers in place due to copyright or ownership.

STRATEGIES FOR ADDRESSING PRIVACY CONCERNS

There is no one-size-fits-all approach for addressing data privacy. Nevertheless, here we discuss a few strategies for mitigating these concerns. It is important to acknowledge that a combination of several strategies should be considered as you work with researchers on how they should handle sensitive data and that this list is not exhaustive.

Data collection instruments and documentation are often among the first materials created for a research project. When creating these materials, we need to keep what we've learned about sensitive data in mind. There are several topics to consider:

Consent Form Language

Researchers may inadvertently limit their ability to share and publish research data from human subjects before they've even started data collection. Overly restrictive language, such as, "All records will be kept private," or, "The study data will be shared only with institutional researchers" can prompt sharing and publishing issues. How will records be kept private? Which records? Which institution and what researchers? Striking a balance between participant confidentiality and data-sharing best practices is critical. Fortunately, there are several resources for librarians and researchers to discuss when drafting or reviewing consent language:

- ICPSR's "Recommended Informed Consent Language for Data Sharing," https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/conf-language.html
- TalkBank's "IRB Approval," https://talkbank.org/share/irb/

- Qualitative Data Repository's "Informed Consent," https://qdr.syr.edu/guidance/human-participants/informed-consent
- UK Data Service's "Consent for Data Sharing," https://www.ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing/consent-forms

Identifying Information

As previously discussed, exposing human subjects' data such as PII may endanger research participants by opening their information to theft, fraud, or exploitation. There are many steps a researcher can take to mitigate risk and protect participants.

Developing plans for storage, backup, and sharing that aim to protect PII and ensure confidentiality should occur in the early stages of a research project. Data collection should be monitored and finalized, or cleaned data need to be reviewed for direct or indirect identifiers. Taking it one step further, consider how a research data set could be used when combined with public information such as voting records or recreational genealogical databases. De-identification and anonymization are not technically perfect, as recent studies have noted how research participants can be traced through a variety of methods: Sweeney (1997),[28] Homer and colleagues (2008),[29] Narayana and Shmatikov (2008),[30] Sweeney, Abu, and Winn (2013),[31] and Rocher, Hendrickx, and de Montjoye (2019).[32] The best way to protect these data is to not collect them in the first place if possible, or to keep identifying information to a minimum.

Access

As discussed earlier, sharing sensitive data may be difficult, but it is not impossible. There are several options for researchers harboring concerns about data privacy and sensitive data:

- *Repositories with limited access options*—Some data repositories, commonly serving the health sciences, offer restricted access for data sets. These options may require users to first register for access to microdata, which can be used to send automated access requests to authors or to track usage of the data.
- *Embargos*—Some open data repositories allow authors to select embargo periods for their data sets. A data set that is under embargo means, that while it has been submitted to a repository, it is not available for download and use right away.
- *Data use agreements (DUAs)*—DUAs established permitted users and uses of a data set. DUAs should be specific to individual data sets to prevent misuse and unauthorized access. DUAs typically contain language that prohibit users from attempting reidentification or contact with subjects, as well as language that prohibits data use outside of an explicitly described project. Terms may vary across data sets, repositories, and institutions.

RDM SERVICES

The landscape of RDM services has evolved with the expanding needs of researchers, publisher and funder data mandates, and institutional policies. Many institutions have developed robust RDM services such as the University of Minnesota Research Data Services team, the University of California, Berkeley, Research Data Management Program, and the Cornell University Libraries Research Data Management Service Group.[33] To support institutional efforts to formalize and grow RDM services, a number of institutional collaborations like The Data Curation Network, professional organizations, and national conferences have emerged as resources for service support and professional development.[34]

As previously discussed in section 2.2.1, "Introduction to Open Data," RDM service models in research institutions vary in terms of support, maturity, offerings, and capacity. Here's a practitioner spotlight on Lisa Johnston, research data management/curation lead and codirector of the University Digital Conservancy at the University of Minnesota Libraries, discussing how she got started in data management work and built services and a network to inform them:

> My career started trending toward data librarianship in 2008 when I was asked to serve on a new library group for "E-Science and Data Services" that was charged to address how the collaborative and digital nature of data management, sharing and publishing was influencing research needs at our large university. My focus was on user-needs assessment and education and by performing surveys and holding focus groups with hundreds of faculty and students, the library was able to take an early front seat in understanding and addressing researcher data management needs on our campus. In the absence of a formal campus-wide group looking at data issues, in 2013 I started the university's first "informal Community of Practice for Research Data Management" and to my surprise nearly 50 people wanted to attend our kick-off event! Many years later, our University has a much more unified approach and now I spend a lot of my time working across campus, collaborating with individuals outside the library from units such as information technology (IT), office of research, and supercomputing, as well as cross-institutionally, addressing data curation challenges collaboratively with my colleagues from other universities in the Data Curation Network. The data problem space is vast and there are enough challenges for us all—so collaboration at every level is key. The best way to get started is to attend campus events (or hold your own!) and seek out individual conversations to ask people about their challenges and how they are addressing them. If your experiences are anything like my own—those same people may be ones who you partner with for years to come!

The following services are increasingly common offerings in academic libraries and may serve as inspiration for the services that you can offer to researchers.

CONSULTATIONS

Research consultations are familiar to information professionals, especially subject or liaison librarians working with faculty and students within specific disciplines. RDM consultations can be viewed as an extension of traditional library research consultations. RDM consultations may range from helping research teams choose an electronic lab notebook or advising on sustainable data formats for storage and preservation to reviewing publisher data-sharing policies when a manuscript is ready for submission. Consultations are a growing area of service, especially in assisting researchers with grant applications by helping draft DMPs.[35]

SUPPORT FOR DMPS

Support for DMPs can encompass several types of services and tools. For instance, several research support teams offer help in the review of DMPs. Such review may assist with ensuring language regarding data sharing is not too restrictive or if adequate storage or backup plans are documented. Similarly, some research libraries promote DMPTool, an online tool that assists users with the development of DMPs by utilizing plan templates for many major funding agencies such as the National Endowment for the Humanities, the National Science Foundation, and the National Institutes of Health. Additional support may take the form of keeping

abreast of funding agencies' changing requirements and expectations for grant recipients and communicating updates to relevant users. Integrating examples of funder requirements into communications or educational outreach can be useful for not only supporting researchers, but also building working relationships between researchers and support services.

Workshops and Instruction

Research data services may also include educational outreach such as workshop facilitation and instruction. Instructional sessions and workshops are highly customizable types of outreach, and so the logistics, content, and delivery of RDM workshops and instruction can vary across institutions depending on the size or type of institution, the audience, and researcher needs. However, as DMPs are required by all major federal funding agencies, research data service providers often find it most beneficial to target graduate students and early-career researchers new to data management foundations and concepts.

Educational outreach also benefits information professionals by establishing relationships with faculty, graduate students, and other service providers on campus. Fostering these relationships can lead to further collaboration, partnerships, and expansion of services. As mentioned in section 2.2.1, "Introduction to Open Data," potential campus partners that may be interested in collaboration could include campus IT, research computing, instructional support centers, and individual departments or colleges.

Scoping educational outreach for the audience in mind is key. For example, a health sciences librarian may choose to develop a workshop series heavily focused on working with institutional review boards and data privacy. Their colleagues in other disciplines may choose to focus content on the basics—file naming, organization, and backup, for example—for a class of undergraduate students completely new to RDM. Consider what is most appropriate for the target population.

Being Embedded on a Research Team

Another research data service involves a data librarian being embedded as a long-term member of a research team. These roles can incorporate traditional librarian concepts of information retrieval, management, and organization, albeit with a data management focus. Duties of an embedded data librarian may include standardizing data collection language and creating data dictionaries, assisting with search strategy development and execution, and organization of relevant research materials.[36] However, partnerships can extend beyond traditional skills. Information professionals can leverage their skill sets to better organize generated or collected data files, integrate metadata schemata to facilitate data discovery, and assist with writing.

A case study by Wang and Fong found that being successfully embedded with a research team required an awareness of "emerging professional practices" and gaining a "deeper understanding of your users' evolving data needs."[37] The study suggests involvement with appropriate professional organizations such as the International Association for Social Science Information Service and Technology (IASSIST) or the Research Data Access and Preservation Association (RDAP), and interest groups within the Association of College and Research Libraries (ACRL) is important in the familiarization process as such organizations provide spaces for peer-to-peer engagement and sharing of experiences. Subject librarians may have an advantage with regard to awareness of professional practices due to knowledge and expertise with a given discipline.[38] While the Wang and Fong case study examines the dual embedment of a data librarian and earth sciences librarian, a data librarian operating as the only

information professional on a research team should consider consulting with the appropriate subject librarian for more information on a given discipline's research and data trends.

Curation

With recent funder mandates and publisher requirements encouraging data sharing, researchers must make decisions and take action to prepare their data adequately for sharing. Some of these services and programs may be hosted by individual departments like university libraries or information technology, or they may be launched as a collaborative or joint responsibility. As an emerging service in academic libraries, data curation is another opportunity for information professionals to exercise their knowledge and experience in data management.

The involvement of information professionals in data curation varies like any other service within an institution. Generally, professionals can expect curatorial duties to include data validation, file transformation, the creation of documentation, the assignment of a DOI to a data set, and so forth (see "Data Curation" above). Similarly, the timing of curatorial work may differ between programs and services. Information professionals may be asked to curate data as they are collected (see "Being Embedded on a Research Team" above), but they increasingly are assigned to curate data sets submitted to institutional repositories after data sets are considered complete.

Best Practices versus Reality

In a perfect world, data collection does not begin until after a protocol or DMP is finalized and shared, infrastructure is agreed upon and set up, and the research team is trained in data management policies and procedures. The DMP would include a file-naming system and folder structure, a chosen metadata schema, and a clear budget detailing expected costs for storage, security, and plans for sharing. The data's documentation would be robust: data collection is described sufficiently and includes pristine samples of any collection instruments. Data analysis and transformation are detailed, down to the version number of any software used. All files and their interdependencies are listed, and all variables are defined in a data dictionary. Files are stored securely, backed up in several places, and then shared publicly if possible in an appropriate repository.

In reality, the key to RDM is understanding the fundamentals and best practices, then adapting them to one's circumstances and the needs of the researchers. Some investments are required when setting up a data management workflow. Time, labor, and funder requirements will primarily decide what best practices are practical, applicable, and most efficient.

CONCLUSION

Clearly, data management and sharing involve complex and context-specific decision-making, keeping in mind current and future use as well as funding agency and journal mandates. Researchers must decide how to manage data they are generating and using, as well as where they will share the data and how they will prepare them for sharing. Ethical dimensions must also be considered—sensitive and restricted data may never be shared openly or may be made available only decades after risk of harm has subsided. These complexities may seem daunting to researchers and impact their willingness to share. As openness becomes more of an expectation in academia, norms and incentives for data sharing will also need to shift accordingly to build a broader culture that supports data sharing. While the benefits are numerous, barriers both technological and cultural still exist. Researchers should consult with

others, such as librarians or other information professionals, when more expertise is needed for overcoming barriers and engaging in good data-sharing practices.

ADDITIONAL RESOURCES

Arnold, Becky, Louise Bowler, Sarah Gibson, Patricia Herterich, Rosie Higman, Anna Krystalli, Alexander Morley, Martin O'Reilly, and Kirstie Whitaker. *The Turing Way: A Handbook for Reproducible Data Science*, version 0.0.4. Zenodo, March 25, 2019. https://doi.org/10.5281/zenodo.3233986.

Briney, Kristin A., Heather Coates, and Abigail Goben. "Foundational Practices of Research Data Management." *Research Ideas and Outcomes* 6 (2020): e56508. https://doi.org/10.3897/rio.6.e56508.

Cornell University Research Data Management Service Group. "Guide to Writing 'Readme' Style Metadata." Accessed April 30, 2020. https://data.research.cornell.edu/content/readme.

Data Curation Network. "Data Primers." GitHub. Accessed April 30, 2020. https://github.com/DataCurationNetwork/data-primers.

DMPTool. Home page. https://dmptool.org/.

FAIRsharing.org. Home page. Accessed April 30, 2020. https://fairsharing.org/.

Information Commissioner's Office. "Data Protection Impact Assessments." Last modified 2020. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/.

Library of Congress. "Recommended Formats Statement." Accessed April 30, 2020. https://www.loc.gov/preservation/resources/rfs/.

Markowetz, Florian. "Five Selfish Reasons to Work Reproducibly." *Genome Biology* 16 (2015): article 274. https://doi.org/10.1186/s13059-015-0850-7.

McGeever, Mags, Angus Whyte, and Laura Molloy. "Five Things You Need to Know about RDM and the Law: DCC Checklist on Legal Aspects of RDM." Digital Curation Centre, September 2015. https://www.dcc.ac.uk/guidance/how-guides/rdm-law.

McKiernan, Erin C., Philip E. Bourne, C. Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, et al. "Point of View: How Open Science Helps Researchers Succeed." *eLife* 5 (2016): e16800. https://doi.org/10.7554/eLife.16800.

National Digital Stewardship Alliance. "Levels of Digital Preservation," version 2.0. Last modified 2019. https://ndsa.org//publications/levels-of-digital-preservation/.

Research Data Alliance. "Metadata Standards Directory." Accessed April 30, 2020. http://rd-alliance.github.io/metadata-directory/standards/.

Re3data.org: Registry of Research Data Repositories. Home page. Accessed April 30, 2020. https://doi.org/10.17616/R3D.

Riley, Jenn. *Understanding Metadata: What Is Metadata, and What Is It For? A Primer*. Baltimore, MD: NISO, 2017. https://www.niso.org/publications/understanding-metadata-2017.

DISCUSSION QUESTIONS

1. Have you ever used open data? If so, were there restrictions on use, and how easy or difficult was it to access and use the data?

2. When should data be shared in the research and publication life cycle?
3. What do you see as some of the barriers for sharing data? How can these barriers be overcome?

NOTES

1. National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science* (Washington, DC: National Academies Press, 2019), https://doi.org/10.17226/25303.
2. Protocols.io, home page, https://www.protocols.io/; National Institute for Health Research, PROSPERO, https://www.crd.york.ac.uk/PROSPERO/.
3. Equator Network, home page, https://www.equator-network.org/.
4. Christine L. Borgman, *Big Data, Little Data, No Data* (Cambridge, MA: MIT Press, 2015), https://mitpress.mit.edu/9780262529914/big-data-little-data-no-data/.
5. Miriam Posner, "Humanities Data: A Necessary Contradiction," *Miriam Posner's Blog*, June 25, 2015, https://miriamposner.com/blog/humanities-data-a-necessary-contradiction/.
6. Jeremy Berg et al., "Joint Statement on EPA Proposed Rule and Public Availability of Data," *Science* 360, no. 6388 (2018): eaau0116, https://doi.org/10.1126/science.aau0116.
7. American Political Science Association, *A Guide to Professional Ethics in Political Science*, 2nd ed. (Washington DC: American Political Science Association, 2012), 9–10.
8. Mathias Astell, "Benefits of Data Sharing for You," *Research Data Community* (blog), SpringerNature, December 1, 2017, https://researchdata.springernature.com/posts/28549-benefits-of-data-sharing-for-you.
9. Heather A. Piwowar and Todd J. Vision, "Data Reuse and the Open Data Citation Advantage." *PeerJ*, no. 1 (2013): e175, https://doi.org/10.7717/peerj.175; Giovanni Colavizza et al., "The Citation Advantage of Linking Publications to Research Data," *PLOS ONE* 15, no. 4 (2020): e0230416, https://doi.org/10.1371/journal.pone.0230416; Garret Christensen et al., "A Study of the Impact of Data Sharing on Article Citations Using Journal Policies as a Natural Experiment," *PLOS ONE* 14, no. 12 (2019); e0225883, https://doi.org/10.1371/journal.pone.0225883.
10. GitHub, home page, accessed July 21, 2020, https://github.com/; Jupyter, home page, accessed July 21, 2020, https://jupyter.org/; Code Ocean, home page, accessed July 21, 2020, https://codeocean.com/; ReproZip, home page, accessed July 21, 2020, https://www.reprozip.org/.
11. Zenodo, home page, accessed July 21, 2020, https://zenodo.org/; Figshare, home page, accessed July 21, 2020, https://figshare.com/.
12. Openlabnotebooks.org, home page, accessed July 21, 2020, https://openlabnotebooks.org/.
13. OSF, home page, accessed July 21, 2020, https://osf.io/.
14. Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (March 2016): article 160018, https://doi.org/10.1038/sdata.2016.18
15. National Institute of Mental Health, "NIMH Data Archive," accessed July 21, 2020, https://nda.nih.gov/; Indiana University, "Indiana Spatial Data Portal," University Information Technology Services, accessed July 21, 2020, https://gis.iu.edu/; Data.gov, home page, accessed July 21, 2020, https://www.data.gov/; European Commission, "EU Open Data Portal," accessed July 21, 2020, https://data.europa.eu/euodp/en/home.
16. DataPortals.org, home page, accessed July 21, 2020, https://dataportals.org/; Open Knowledge Foundation, home page, accessed July 21, 2020, https://okfn.org/.
17. Erzsébet Tóth-Czifra, "One More Word about ResearchGate/Academia.edu and Why Using These Platforms Will Never Be Equal to Proper Self-Archiving," *DARIAH Open* (blog), May 31, 2020, upd. June 5, 2020, https://dariahopen.hypotheses.org/878.
18. Diana Kwon, "The Push to Replace Journal Supplements with Repositories," *Scientist*, August 19, 2019, website archive, https://web.archive.org/web/20190829200838/https://www.the-scientist.com/news-opinion/the-push-to-replace-journal-supplements-with-repositories--66296; Carlos Santos, Judith Blake, and David J. States, "Supplementary Data Need to Be Kept in Public Repositories," *Nature* 438, no. 7069 (2005): 738–738, https://doi.org/10.1038/438738a.
19. Timothy H. Vines et al., "The Availability of Research Data Declines Rapidly with Article Age," *Current Biology* 24, no. 1 (2014): 94–97, https://doi.org/10.1016/j.cub.2013.11.014.
20. Lisa R. Johnston et al., "Data Curation Activities," Data Curation Network, University of Minnesota Digital Conservancy, 2016, https://datacurationnetwork.org/data-curation-activities/.
21. Dublin Core Metadata Initiative, "Dublin Core," accessed April 27, 2020, https://dublincore.org/specifications/dublin-core/; DataCite, "DataCite Metadata Schema," accessed April 27, 2020, https://schema.datacite.org/.
22. Data Citation Synthesis Group, *Joint Declaration of Data Citation Principles*, ed. M. Martone (San Diego: FORCE11, 2014), https://doi.org/10.25490/a97f-egyk.

23. Creative Commons, "Share Your Work," accessed July 21, 2020, https://creativecommons.org/share-your-work/.
24. ICPSR, "Data Enclaves," accessed July 21, 2020, https://www.icpsr.umich.edu/web/pages/ICPSR/access/restricted/enclave.html; US Census Bureau, "Federal Statistical Research Data Centers," accessed July 21, 2020, https://www.census.gov/fsrdc.
25. Erika McCallister, Tim Grance, and Karen Scarfone, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, NIST Special Publication 800-122 (Gaithersburg, MD: National Institute of Standards and Technology, April 2010), https://doi.org/10.6028/NIST.SP.800-122.
26. Adam Welz, "Unnatural Surveillance: How Online Data Is Putting Species at Risk," *YaleEnvironment360* (blog), September 6, 2017, https://e360.yale.edu/features/unnatural-surveillance-how-online-data-is-putting-species-at-risk.
27. Ayesha I. T. Tulloch et al., "A Decision Tree for Assessing the Risks and Benefits of Publishing Biodiversity Data," *Nature Ecology and Evolution* 2, no. 8 (August 2018): 1209–17, https://doi.org/10.1038/s41559-018-0608-1.
28. Latanya Sweeney, "Weaving Technology and Policy Together to Maintain Confidentiality," *The Journal of Law, Medicine & Ethics*, 25(2–3 (1997): 98–110, https://doi.org/10.1111/j.1748-720X.1997.tb01885.x.
29. Nils Homer et al, "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays," *PLOS Genetics 4*, no. 8(August 2008): e1000167, https://doi.org/10.1371/journal.pgen.1000167.
30. Arvind Narayanan and Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets," *2008 IEEE Symposium on Security and Privacy,* Spring 2008, Oakland, CA, USA, 2008, pp. 111-125, doi: 10.1109/SP.2008.33.
31. Latanya Sweeney, Akua Abu, and Julia Winn, "Identifying Participants in the Personal Genome Project by Name," April 29, 2013, http://dx.doi.org/10.2139/ssrn.2257732.
32. Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Communications 10*, 3069 (2019), https://doi.org/10.1038/s41467-019-10933-3.
33. University of Minnesota Libraries, "Research Data Services," https://www.lib.umn.edu/services/data; University of California, Berkeley, "Research Data Management Program," https://researchdata.berkeley.edu/; Cornell University Libraries, "Research Data Management Service Group, https://data.research.cornell.edu/.
34. Data Curation Network, "Partner Institutions," accessed September 24, 2020, https://datacurationnetwork.org/about/partners/.
35. Carol Tenopir et al., *Research Data Services in Academic Libraries*, white paper (Middletown, CT: Choice, 2019), https://www.choice360.org/content/2-librarianship/5-whitepaper/tenopir-white-paper-2019/tenopir_121019_rds.pdf.
36. Victoria Goode and Blair Anton, "Welch Informationist Collaboration with the Johns Hopkins Medicine Department of Radiology," *Journal of eScience Librarianship* 2, no. 1 (2013): 16–19, https://doi.org/10.7191/jeslib.2013.1033; Sally Gore, "A Librarian by Any Other Name: The Role of the Informationist on a Clinical Research Team," *Journal of eScience Librarianship* 2, no. 1 (2013): 20–24, https://doi.org/10.7191/jeslib.2013.1041.
37. Minglu Wang and Bonnie L. Fong, "Embedded Data Librarianship: A Case Study of Providing Data Management Support for a Science Department," *Science and Technology Libraries* 34, no. 3 (July 3, 2015): 228-240, https://doi.org/10.1080/0194262X.2015.1085348.
38. Jeremy Garritano and Jake Carlson, "A Subject Librarian's Guide to Collaborating on e-Science Projects," *Issues in Science and Technology Librarianship*, no. 57 (Spring 2009), https://docs.lib.purdue.edu/lib_research/140.

BIBLIOGRAPHY

American Political Science Association. *A Guide to Professional Ethics in Political Science*, 2nd ed. Washington DC: American Political Science Association, 2012.

Astell, Mathias, "Benefits of Data Sharing for You." *Research Data Community* (blog), SpringerNature, December 1, 2017. https://researchdata.springernature.com/posts/28549-benefits-of-data-sharing-for-you.

Berg, Jeremy, Philip Campbell, Veronique Kiermer, Natasha Raikhel, and Deborah Sweet. "Joint Statement on EPA Proposed Rule and Public Availability of Data." *Science* 360, no. 6388 (2018): eaau0116. https://doi.org/10.1126/science.aau0116.

Borgman, Christine L. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press, 2015. https://mitpress.mit.edu/9780262529914/big-data-little-data-no-data/.

Christensen, Garret, Allan Dafoe, Edward Miguel, Don A. Moore, and Andrew K. Rose. "A Study of the Impact of Data Sharing on Article Citations Using Journal Policies as a Natural Experiment." *PLOS ONE* 14, no. 12 (2019): e0225883. https://doi.org/10.1371/journal.pone.0225883.

Code Ocean. Home page. Accessed July 21, 2020. https://codeocean.com/.

Colavizza, Giovanni, Iain Hrynaszkiewicz, Isla Staden, Kirstie Whitaker, and Barbara McGillivray. "The Citation Advantage of Linking Publications to Research Data." PLOS ONE 15, no. 4 (2020): e0230416. https://doi.org/10.1371/journal.pone.0230416.

Cornell University Libraries. "Research Data Management Service Group." https://data.research.cornell.edu/.

Creative Commons. "Share Your Work." Accessed July 21, 2020. https://creativecommons.org/share-your-work/.

Data Citation Synthesis Group. *Joint Declaration of Data Citation Principles*. Edited by M. Martone. San Diego: FORCE11, 2014. https://doi.org/10.25490/a97f-egyk.

DataCite. "DataCite Metadata Schema." Accessed April 27, 2020. https://schema.datacite.org/.

Data Curation Network. "Partner Institutions." Accessed September 24, 2020. https://datacurationnetwork.org/about/partners/.

Data.gov. Home page. Accessed July 21, 2020. https://www.data.gov/.

DataPortals.org. Home page. Accessed July 21, 2020. https://dataportals.org/.

Dublin Core Metadata Initiative. "Dublin Core." Accessed April 27, 2020. https://dublincore.org/specifications/dublin-core/.

Equator Network. Home page. https://www.equator-network.org/.

European Commission. "EU Data Protection Rules." Accessed September 24, 2020. https://ec.europa.eu/info/law/law-topic/data-protection/eu-data-protection-rules_en.

———. "EU Open Data Portal." Accessed July 21, 2020. https://data.europa.eu/euodp/en/home.

Family Educational Rights and Privacy Act (FERPA). Title 34 Code of Federal Regulations. Pt. 99. 2020.

Figshare. Home page. Accessed July 21, 2020. https://figshare.com/.

Garritano, Jeremy, and Jake Carlson. "A Subject Librarian's Guide to Collaborating on e-Science Projects." *Issues in Science and Technology Librarianship*, no. 57 (Spring 2009). https://docs.lib.purdue.edu/lib_research/140.

General Data Protection Regulation (GDPR). "Complete Guide to GDPR Compliance." https://gdpr.eu/.

———. "Principles Relating Processing of Personal Data." Article 5 of GDPR. https://gdpr.eu/article-5-how-to-process-personal-data/.

Genetic Information Nondiscrimination Act of 2008, Public Law 110-233, U.S. Statutes at Large 122 (2008): 881–922. https://www.govinfo.gov/app/details/STATUTE-122/STATUTE-122-Pg881.

GitHub. Home page. Accessed July 21, 2020. https://github.com/.

Goode, Victoria, and Blair Anton. "Welch Informationist Collaboration with the Johns Hopkins Medicine Department of Radiology." *Journal of eScience Librarianship* 2, no. 1 (2013): 16–19. https://doi.org/10.7191/jeslib.2013.1033.

Gore, Sally. "A Librarian by Any Other Name: The Role of the Informationist on a Clinical Research Team." *Journal of eScience Librarianship* 2, no. 1 (2013): 20–24. https://doi.org/10.7191/jeslib.2013.1041.

Homer Nils, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays." *PLOS Genetics 4*, no. 8(August 2008): e1000167. https://doi.org/10.1371/journal.pgen.1000167.

ICPSR. "Data Enclaves." Accessed July 21, 2020. https://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/restricted/enclave.html.

———. "Recommended Informed Consent Language for Data Sharing." Accessed September 24, 2020. https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/conf-language.html.

Indiana University. "Indiana Spatial Data Portal." University Information Technology Services. Accessed July 21, 2020. https://gis.iu.edu/.

Johnston, Lisa R., Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart. "Definitions of Data Curation Activities Used by the Data Curation Network." Data Curation Network, University of Minnesota Digital Conservancy, 2016. https://datacurationnetwork.org/data-curation-activities/.

Jupyter. Home page. Accessed July 21, 2020. https://jupyter.org/.

Kwon, Diana. "The Push to Replace Journal Supplements with Repositories." *Scientist*, August 19, 2019. Website archive. https://web.archive.org/web/20190829200838/https://www.the-scientist.com/news-opinion/the-push-to-replace-journal-supplements-with-repositories--66296.

McCallister, Erika, Tim Grance, and Karen Scarfone. *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII): Recommendations of the National Institute of Standards and Technology*. NIST Special Publication 800-122. Gaithersburg, MD: National Institute of Standards and Technology, April 2010. https://doi.org/10.6028/NIST.SP.800-122.

Narayanan, Arvind, and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets." *2008 IEEE Symposium on Security and Privacy (Spring 2008)*, Oakland, CA, USA, 2008, pp. 111-125. doi: 10.1109/SP.2008.33.

National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. Washington, DC: National Academies Press, 2019. https://doi.org/10.17226/25303.

National Human Genome Research Institute. "Genetic Discrimination." Accessed September 24, 2020. https://www.genome.gov/about-genomics/policy-issues/Genetic-Discrimination.

National Institute for Health Research. PROSPERO. https://www.crd.york.ac.uk/PROSPERO/.

National Institute of Mental Health. "NIMH Data Archive." Accessed July 21, 2020. https://nda.nih.gov/.

Open Knowledge Foundation. Home page. Accessed July 21, 2020. https://okfn.org/.

Openlabnotebooks.org. Home page. Accessed July 21, 2020. https://openlabnotebooks.org/.

OSF. Home page. Accessed July 21, 2020. https://osf.io/.

Piwowar, Heather A., and Todd J. Vision. "Data Reuse and the Open Data Citation Advantage." *PeerJ*, no. 1 (2013): e175. https://doi.org/10.7717/peerj.175.

Posner, Miriam. "Humanities Data: A Necessary Contradiction." *Miriam Posner's Blog*, June 25, 2015. https://miriamposner.com/blog/humanities-data-a-necessary-contradiction/.

Protocols.io. Home page. https://www.protocols.io/.

Qualitative Data Repository. "Informed Consent." Accessed September 24, 2020. https://qdr.syr.edu/guidance/human-participants/informed-consent.

ReproZip. Home page. Accessed July 21, 2020. https://www.reprozip.org/.

Rocher, Luc, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. "Estimating the success of re-identifications in incomplete datasets using generative models." *Nature Communications 10*, 3069 (2019). https://doi.org/10.1038/s41467-019-10933-3.

Santos, Carlos, Judith Blake, and David J. States. "Supplementary Data Need to Be Kept in Public Repositories." *Nature* 438, no. 7069 (2005): 738–738. https://doi.org/10.1038/438738a.

Scientific Data and Mathias Astell. "Benefits of Open Research Data Infographic." Figshare, July 11, 2017. https://doi.org/10.6084/m9.figshare.5179006.v3.

Sweeney, Latanya. "Weaving Technology and Policy Together to Maintain Confidentiality." *The Journal of Law, Medicine & Ethics*, 25(2–3 (1997): 98–110. https://doi.org/10.1111/j.1748-720X.1997.tb01885.x.

Sweeney, Latanya, Akua Abu, and Julia Winn. "Identifying Participants in the Personal Genome Project by Name." April 29, 2013. http://dx.doi.org/10.2139/ssrn.2257732.

TalkBank. "IRB Approval." Accessed September 24, 2020. https://talkbank.org/share/irb/.

Tenopir, Carol, Jordan Kaufman, Robert Sandusky, and Danielle Pollock. *Research Data Services in Academic Libraries: Where Are We Today?* White paper. Middletown, CT: Choice, 2019. http://www.choice360.org/content/2-librarianship/5-whitepaper/tenopir-white-paper-2019/tenopir_121019_rds.pdf.

Tóth-Czifra, Erzsébet. "One More Word about ResearchGate/Academia.edu and Why Using These Platforms Will Never Be Equal to Proper Self-Archiving." *DARIAH Open* (blog), May 31, 2020, upd. June 5, 2020. https://dariahopen.hypotheses.org/878.

Tulloch, Ayesha I. T., Nancy Auerbach, Stephanie Avery-Gomm, Elisa Bayraktarov, Nathalie Butt, Chris R. Dickman, Glenn Ehmke, et al. "A Decision Tree for Assessing the Risks and Benefits of Publishing Biodiversity Data." *Nature Ecology and Evolution* 2, no. 8 (August 2018): 1209–17. https://doi.org/10.1038/s41559-018-0608-1.

UK Data Service. "Consent for Data Sharing." Accessed September 24, 2020. https://ukdataservice.ac.uk/learning-hub/research-data-management/ethical-issues/consent-for-data-sharing/.

University of California, Berkeley. "Research Data Management Program." https://researchdata.berkeley.edu/.

University of Minnesota Libraries. "Research Data Services." https://www.lib.umn.edu/services/data.

US Census Bureau. "Federal Statistical Research Data Centers." Accessed July 21, 2020. https://www.census.gov/fsrdc.

US Department of Education. "Privacy and Data Sharing." Accessed September 24, 2020. https://studentprivacy.ed.gov/privacy-and-data-sharing.

US Department of Health and Human Services. "Research." Accessed September 24, 2020. https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html.

Vines, Timothy. H., Arianne Y. K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, et al. "The Availability of Research Data Declines Rapidly with Article Age." *Current Biology* 24, no. 1 (2014): 94–97. https://doi.org/10.1016/j.cub.2013.11.014.

Wang, Minglu, and Bonnie L. Fong. "Embedded Data Librarianship: A Case Study of Providing Data Management Support for a Science Department." *Science and Technology Libraries* 34, no. 3 (July 3, 2015): 228–40. https://doi.org/10.1080/0194262X.2015.1085348.

Welz, Adam. "Unnatural Surveillance: How Online Data Is Putting Species at Risk." *YaleEnvironment360* (blog), September 6, 2017. https://e360.yale.edu/features/unnatural-surveillance-how-online-data-is-putting-species-at-risk.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (March 2016): article 160018, https://doi.org/10.1038/sdata.2016.18.

Zenodo. Home page. Accessed July 21, 2020. https://zenodo.org/.