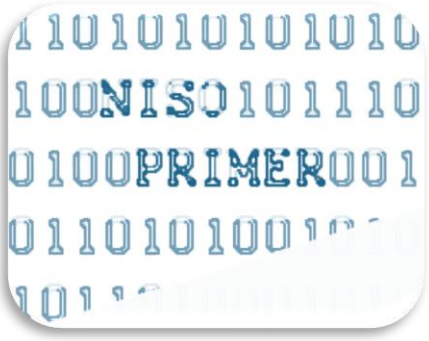
Research Data Management

By Carly Strasser

A Primer Publication of the
National Information Standards Organization



About the NISO Primer Series

This NISO Primer Series is a three-part series of documents that provide introductory guidance to users of research data. Meant to provide insight and instruction to researchers collecting data, these primers discuss the latest developments in research data and the new tools, best practices, and resources now available

For current information on the status of this publication contact the NISO office or visit the NISO website (www.niso.org).

Published by
National Information Standards Organization (NISO)
3600 Clipper Mill Road
Suite 302
Baltimore, MD 21211
www.niso.org

ISBN: 978-1-937522-65-0

CONTENTS

INTRODUCTION

The way research is conducted has changed dramatically in the last two decades. New methods and tools (software, hardware, instruments, and equipment), new sources of data, and the increasing connectivity of global research via the internet mean that researchers across the globe are making progress at an unprecedented pace. But with this paradigm shift comes significant challenges, most notably reproducibility of research and transparency of methods and workflows.

Meeting the challenges of 21st century research require sound research data management. By carefully planning, documenting, and preserving data, the goals of having reproducible and transparent research data are far easier to meet. Further, well-managed data are easier to use and reuse, which translates to more collaboration for researchers and maximum return-on-investment for funders. This primer will cover the basics of research data management, with the goal of helping researchers and those that support them become better data stewards.

PLANNING FOR DATA MANAGEMENT

The many adages about planning certainly apply to research data management: it is the most critical step for ensuring longevity and utility of datasets. Before the first data point is collected, researchers should invest time in carefully considering:

- How will the data be documented? What metadata will be used? What naming schemes will be used for data, files, samples, etc.?
- What personnel, software, and hardware will be required to manage the data effectively?
- Who will be responsible for ensuring that data management remains a priority throughout the project?
- Where will the data ultimately be stored? Who will have access? What policies will accompany data use and reuse? Whose resources will be used?
- When will the data no longer be useful and can be discarded?

Data management plans (DMPs)

Many funders are beginning to recognize that planning before beginning a research data project is critical, and data management plans (DMPs) are increasingly required as part of the proposal submission process. Most DMPs have five basic components intended to address the major sections of project data management:

1. A description of the types of data that will be collected or generated during the project
2. The standards that will be used for those data and their associated metadata
3. A description of the policies pertaining to the data that will be collected or generated
4. Plans for archiving and preservation of the data generated
5. A description of the resources that will be needed to accomplish data management, including personnel, hardware, software, and budgetary requirements

Data description

Describing the data that will be collected is potentially more nuanced than it might appear. Much of this potential complication stems from the fact that the very definition of "data" is fraught with debate. Some funders explicitly list potential research outputs that should be considered data for the purposes of a DMP. Outputs might include code or scripts, images, software, video, or models. These are not traditionally defined as "data," but by considering all research outputs as

data, disciplines that have not historically used the word data, such as the humanities, can more easily understand how best to complete a DMP for funders. In addition, groups such as theoreticians that produce methods, as opposed to traditional datasets, can better understand what products to consider sharing upon completion of a project.

Generally, the description of the data should include what the researcher expects to collect, how they will collect it, and their plans for processing and analyzing it. This should not be considered a supplement to the "Methods" section of a proposal; instead it should focus on what activities will be conducted around the data to ensure it is collected properly, well-documented, and considered carefully before the project begins. See *Documenting Research Data* below for more information.

Standards

Identifying standards (for both data and metadata) at the beginning of a research project can greatly assist in properly managing the data over the course of the work, thereby making data sharing easier after the project ends. There are many metadata standards available and it can be daunting for researchers to select one given their potential limited knowledge of data stewardship. The best way to identify an appropriate metadata standard for a given dataset is to identify the destination repository that will be used for the data at the close of the project. These repositories may have required or requested data and/or metadata standards for submission. Another useful strategy for identifying both data and metadata standards is to consult colleagues with similar datasets and copy their approach. See *Metadata* below for more information.

Policies and procedures

Properly planning for data collection involves identifying whether there are any rules or policies that govern the data that will be collected. These should be noted in the DMP, along with plans for how others will gain access to data, whether there will be embargo periods before others can gain this access, and whether there are existing obligations to share (or not share) the data due to funder or organizational policies. Sensitive data (e.g., those involving human subjects or endangered species) may be subject to federal or state laws and will likely require approval from institutional review boards. There is still discussion around the subject of intellectual property rights related to data, as well as the extent to which U.S. copyright law applies to datasets. But generally, the researcher should describe the type of license or waiver that will apply to the data.

Archiving and preservation

The DMP should include information about where the data will be archived for the long-term. This likely includes one or more repositories that have preservation strategies in place for ensuring long-term usability of and accessibility to the data. By identifying the data's eventual destination, a researcher is better able to plan for other aspects of data management, including file formats, transformations that might be required before deposit, and metadata standards and content. The DMP should also identify which project personnel will be responsible for the preservation of the data in the selected repositories.

Resources needed

Stellar data management is a potentially costly endeavor; the DMP should identify what resources will be required to properly carry out the plan. This will include personnel time, hardware, software, and archival fees for organizing, managing, backing up, and securing data. These costs are not insignificant, and it is recommended that researchers consult their institutional service providers (libraries, IT departments) to ensure they appropriately budget for data stewardship activities.

Best practices for data management planning

Data management planning should not be limited to creating the minimal DMP required by a funder. Instead, it should be treated as an invaluable part of the data life cycle that will ensure the data are usable both for the primary researcher over the course of the project, as well as any future users of the data.

Design naming schemes

Many researchers begin collecting data and accumulating files without having an organizational system in place to maximize efficiency and usability. By spending time designing how samples/data (physical or digital) will be named, problems with duplicate names or identity confusion, as well as future renaming and sorting tasks, can be avoided. Naming schemes should be descriptive, unique, and reflect the content/sample. For example, naming a particular soil collection `sample1` is not helpful if there is also a water sample named `sample1`. Better choices might be `Soil01_SiteB_2014` and `Water01_SiteB_2014`.

When deciding on organizational schemes for digital files, consider future dependencies (will a particular data file be required for a script?), file formats (should all .csv files be grouped together?), and order of analysis. Use folders and tags to help organize files logically.

Design spreadsheets

Many researchers, regardless of discipline, use spreadsheets for organizing and visualizing their data. Sometimes, spreadsheets are used to the detriment of data documentation and dataset provenance: features that make spreadsheets user-friendly and appealing for data manipulation also render retracing the steps taken to results nearly impossible.

Ideally, such manipulations and analyses would be carried out with scripts to ensure provenance, ensuring documentation of the entire workflow. It is unlikely, however, that researchers will stop using spreadsheets in the near future. Given this fact, there are a few best practices associated with spreadsheets that can help enable usability of the data in the long term:

1. Keep the raw data (unprocessed and unanalyzed) on a separate tab. All manipulations should be carried out on other tabs or in other files to ensure that the original data are not lost.

2. "Atomize" information within the spreadsheet. This means breaking up information so that only one type of data is in any given cell. For example, rather than providing a location in a cell as "Austin, TX," this should be broken into two cells: one for city (Austin) and one for state (TX).

3. Consider breaking spreadsheets into more discrete, standalone tables. Rather than having all of the information for location of collection (site name, site number, latitude, longitude, city, county, state, country) directly in the data table, move this information to a separate table specifically designed for describing location. The location number can then suffice within the main data table, and users can refer to the site table for more information. This strategy mimics relational databases.

Create a metadata collection plan

Deciding on a metadata standard is not enough. Before data are collected, researchers should consider exactly how data will be described. Some disciplines use the concept of a "data dictionary," which contains information about the data such as meaning, relationships to other data, origin, usage, and format. Other information might include how null values will be identified, what naming conventions will be used for samples, how many significant digits should be included in an instrument's output, or what units should be used. If data collection sheets (digital or analog) will be used, these should be designed before data collection to comply with the metadata schema chosen for the dataset.

Establish a backup strategy

Researchers should ensure they have a clear, detailed plan for how they will back up their data. Considerations include the hardware and software needed, the frequency of backups, who will be in charge of performing these backups, where and how long the backups are stored, and what steps should be taken in case of loss of the main instance of the data. If backups are performed automatically, a plan should be in place for testing and verifying that these backups are taking place and are capturing the necessary information correctly and consistently.

The Data Management Plan (DMP) as a living document

Planning for good research data management does not end when a project begins. Instead, researchers should frequently revisit the plan to ensure follow-through and consistency. A best practice is to set a reminder to revisit the DMP every week and update it as necessary. Analyses, strategies, and data collections might change significantly over the course of a project; these changes should be documented and taken into account via the DMP.

Personnel for data management planning

The responsibility of planning for research data management does not fall solely to the researcher. Institutions and organizations that support researchers, as well as funders, also have a key role in data stewardship. Grants coordinators and sponsors of project offices should pay close attention to DMPs submitted alongside grants and share researchers' plans with institutional groups whose resources may be implicated in the plan. These groups may include libraries, institutional repositories, information technology offices, research support services, data security offices, institutional review boards, and legal counsel.

In the same way that institutions must provide basic infrastructure for research (lab space, internet connection, library access), they should also make provisions for proper stewardship of data throughout the research life cycle. This may include trained staff to assist researchers in data management, access to infrastructure for storage and backup of data during a project, and/or the availability of an institutional repository that ensures preservation and access to research data for the long-term.

DOCUMENTING RESEARCH DATA

Research data must adhere to specific guidelines for identification to be useful in the long-term.

Metadata

Research outputs (data, code, etc.) that are poorly documented are like canned goods with the label removed: the contents may be something desirable, but it is impossible to tell without metadata. High-quality metadata are as critical to effective data sharing as the data itself, since the better the metadata, the more likely a dataset will be able to be reused.

Informal metadata

Researchers not familiar with research data management best practices are usually familiar with informal metadata, although they may not refer to it as such. Generally, this is the information contained in laboratory notebooks, field notebooks, and in "readme" files. As research is conducted, notes are accumulated that help the primary investigator and his or her colleagues to understand the data and workflow, document the decision-making, and annotate the project's progress. Even with its nuanced and personalized structure, this informal metadata can be critical to understanding research outputs. Such information should be "born digital" whenever possible, and digitized when this is not possible. Doing so aids in keeping this critical information to the born-digital dataset. Although informal metadata are not ideal for ensuring interoperability and data discovery, they should be archived alongside corresponding research outputs whenever possible.

Formal (standard) metadata

The second form of metadata for research outputs is standardized. Standard metadata includes information about the research outputs that conforms to a standardized format, has a controlled vocabulary, and is accepted and used by the community. There are numerous metadata standards available for use. The most appropriate metadata standard for a given project will depend primarily on the associated discipline for the data and the intended repository for the dataset's long-term preservation. Examples of discipline-specific metadata schemas include Ecological Metadata Language (used in ecology and environmental science) and FGDC 19115 (used for geospatial data). The primary benefits of standardized metadata are to ensure data are interoperable with similar datasets and to allow for data to be discovered easily. Discoverability is enhanced by the underlying encoding of the metadata standard, which ensures that it is machine-readable.

Standard metadata should be generated during a project but is unfortunately usually only considered when the data repository chosen for preservation requests a particular schema upon data deposition. It is preferable that a standard be decided at the beginning of a project and that the metadata elements needed to conform to the standard are documented at the point of collection. Researchers should contact their chosen repository and ask for guidance and suggestions early in the life of the project. Some repositories offer free software tools and applications that can be used to generate standard metadata, making creating metadata easier for the researcher throughout the course of the project.

Documenting software

A set of data, as customarily defined, is not the only research product. Another common output is software, which is sometimes simply called "code." As research becomes increasingly computationally intensive, code is often generated to clean and organize datasets, analyze them, and produce final outputs such as graphs and tables. These final research outputs represent a summary and/or reworked form of the data; tracing from these final forms back to the original raw dataset is nearly impossible. In the past, researchers and the general public have trusted that the results presented were true and honest representations of the data originally collected. But increasing retractions and skepticism from the public demand that the research community provide better provenance for the results they present.

To enable complete reproducibility of a project, the code used to accomplish these tasks should also be preserved alongside the data. Documenting that code is critical to ensuring that it is understandable to others; this may involve annotating the code with comments and/or providing readme files to explain the relationships among code scripts, datasets, and outputs.

Documenting workflows

Workflows are another type of research output that may help with reuse and reproducibility. Similar to metadata, workflows may be either informal or formal; both can be invaluable for understanding a research project and its outputs.

Informal workflows

Informal workflows can be simple flow charts, describing the path of information (data) from the point of its generation or collection to the final synthesized outputs (graphs, publications). Such workflows may be hand-drawn, documented via code, take the form of readme files, or otherwise be documented by the researcher in a way that will help others better understand the conclusions drawn from the work.

Formal workflows

In contrast to informal workflows, formal workflows are generated by software and intended for reuse by others. There are several popular workflow documentation software systems, and the number continues to increase as reproducibility becomes more important to researchers. These systems can be complicated and specific to a given discipline's common workflow, and therefore not commonly used. An example of disciplinary use of formal workflows is the use of Taverna by the genetics and genomics communities, who use the workflow software to automate common computational tasks. Although the use of formal workflow systems is not widespread, their use is likely to increase over the next few years as research continues its trajectory towards being born-digital and computationally intensive.

Workflow software

The increasing recognition of data as an important output of scholarly work has caught the attention of those that build applications and services for data, as well as funders interested in supporting data management and sharing. This is evident in the many new tools that have been created with data in mind, both for researchers and for those communities that support them (libraries, data centers, and publishers).

For most disciplines, the process of collecting, analyzing, and presenting data involves multiple computer programs, each requiring different file formats, and all potentially complicating the path that data takes from collection to publishing results (i.e., the provenance of the data). These complicated workflows inhibit reproducibility and reuse, especially since many researchers do not properly document their nuanced workflows.

While researchers continue to increase the number of software tools used in the course of their work, computer scientists are creating software to help capture the many inputs and outputs. Workflow software has been around for more than a decade, but many of the systems created are still too complex for the average researcher to use effectively. This is likely to change as accountability and reproducibility become more important for researchers to demonstrate.

Software for capturing workflows is highly variable, and many applications are created with a specific discipline in mind. Often, they take the form of electronic lab notebooks, intended to help researchers virtually write down their notes as the project moves forward. iPython is one example. Others are execution environments, which means researchers can design and execute a formal workflow. Examples include Taverna and Kepler. Although these systems are promising, researchers without a coding background should proceed with caution since many are focused on integrating with systems like LaTeX, Git, and R.

Computer Emulation

One way to address the complexity of software interaction is to use virtual machines as emulators in order to provide the environment in which the original dataset was processed.

ADMINISTRATION

Dataset governance and use agreements

Sharing research data implies that others may examine, download, and/or use that data in the future. Ensuring that data are available for use and reuse requires proper licensing or waivers that enable these activities. The community norms for rules and regulations around data, i.e., governance, are still being developed, and are complicated by a number of factors unique to data.

Copyright and ownership

Data are facts, which means that they are not subject to restrictions or protections associated with copyright. However, copyright and intellectual property laws do apply to assemblages of data, code, and other research outputs. For example, an individual data point in a lab notebook is not subject to copyright, but the lab notebook itself is protected under copyright laws.

Academic researchers at universities will often find that their university claims ownership of their research outputs, including data. The researchers typically agreed to this arrangement when joining the university. But this is not widely known and understood, and many researchers assume they own their data assemblages, lab notebooks, and other research products. This misconception is exacerbated by the fact that universities rarely exercise their rights to data assemblages, although this may change as new mandates go into place.

Sui generis rights

Another set of rights that may apply to datasets are known as *sui generis*. This is a Latin phrase which means "unique in its characteristics." *Sui generis* rights are intellectual property rights that prohibit extraction or reuse of a database, regardless of creativity or originality of the database. Copyright law emphasizes creativity and originality; *sui generis* rights do not make this distinction. The European Union recognizes *sui generis* rights for fifteen years after database creation, while the United States has not yet enacted such a law. These rights may be important for projects that involve international partners.

Legal mechanisms for addressing data rights

For others to use data without fear of legal recourse, the permissible parameters of use must be understood. Dataset owners can accomplish this using one of three mechanisms.

The first of these is a contract, which might also be called a data access policy, data use policy, or use agreement. Contracts are completely customizable, which makes them appealing for researchers and institutions who are concerned with how and by whom their data might be used. However, contracts make data reuse difficult, and in many cases may reduce the long-term value

of the dataset as a result. The second mechanism is a license, which makes data reuse easier since the terms are universal. The third mechanism is a waiver. If rights to a dataset are waived, the data are considered in the public domain and use is unrestricted. Waivers maximize the utility of datasets but attribution and credit are potentially more difficult.

Sensitive data

Additional complications surround the use of sensitive data, which may include information about human subjects, endangered species, or protected areas. These datasets are potentially subject to laws and regulations intended to prevent widespread sharing of the information and may restrict a researchers' ability to share their datasets, regardless of copyright. As a result, tension exists between privacy and the current move towards open data, open science, and open research. It is important to note that licenses and waivers cannot address privacy and confidentiality; contracts must be used in cases where these types of issues might be important.

Best practice for sharing

The best practice around sharing (non-sensitive) data to ensure its reuse is to release the dataset into the public domain, without any rules or restrictions inhibiting its reuse. By placing data in the public domain, others are able to download, combine, and refactor data as needed without concern for rules around its use. This is commonly achieved using a Creative Commons Zero (CC0) waiver, which declares that the data is in the public domain.

Waivers (and licenses) should be machine-readable, without any additional requirements for contacting the author or restrictions on how the data can be used. Researchers and institutions sometimes prefer using uniquely crafted data use agreements that are not machine-readable and have restrictions on data use or requirements for contacting the dataset authors. These unique documents significantly impede data reuse and should be avoided. Instead, researchers should opt for license schemas that are commonly used, such as Creative Commons and Open Data Commons.

Note that placing data in the public domain does not preclude cultural norms around attribution. As data citation becomes more prevalent, researchers will receive credit for their datasets in similar ways to how they receive credit for traditional scholarly publications.

Data storage, backups, and security

Storing and backing up data is a necessary concern for any research project involving digital data. However, this seemingly mundane task is too often neglected in the researcher's daily work.

Backups

At a minimum, there should be three copies—original, near, and far—of the full dataset, associated code, and workflows. The first copy, the "original," is the working dataset and associated files. This original copy is usually housed on a researcher's primary computer. The

second copy should live "near" the original, although ideally not in the same physical location. This copy is updated daily via either automatic backup software or manually. Often, the near copy is kept on external hard drives or a shared file server within the researcher's institution. The third copy of the data should be in a physical location "far" from the original and the near copies. It should not be in the same building, and certainly not in the same room. Ideally the far copy is located in an area with different disaster threats. The far copy can be in the form of a cloud-based backup system that undergoes automatic backups and keeps multiple copies of the data stored within the system.

Version control

One of the major challenges of keeping multiple copies of data is version control. Too often researchers use their own idiosyncratic methods for identifying different versions of their files, including datasets. They should be looking to the software community, who has grappled with the version control problem for years, resulting in well-designed version control systems that help document the natural flow of a project. Using these systems, changes to a file are documented with a version number, a timestamp, and an explanation of what has changed. These revisions can be easily compared and restored if necessary. Popular examples of version control systems include Subversion, Git, and Mercurial. Version control systems are not necessarily intuitive to the average researcher without a computer science background, but new web tools like GitHub and Bitbucket are making it easier for newcomers to use these powerful tools.

Cloud storage

An increasingly popular option for storing data is the use of a cloud-based service, such as Dropbox or Google Drive. These are a great solution for many researchers. However, those dealing with sensitive data should be wary of these commercial options. Security of the data is not necessarily guaranteed, and security breaches such as hacked password logs are a real threat. If a dataset contains sensitive material (e.g., personal information about human subjects, endangered species at risk of being hunted, or geographic information about heritage sites), it should be stored on systems with a high level of security rather than free cloud-based options. Researchers with this type of data should work with the information technology group at their institution or organization to ensure that all copies of the data are kept securely and safely.

PRESERVATION

The last step in properly managing research data is preserving it for the long term. Preservation of data is not the same as storage of data. The distinguishing factor between these two is that the goal of data preservation is to ensure that the data can be accurately rendered over time. This involves strategies and policies to allow ongoing usability of the data as well as preservation methods such as maintaining multiple copies and migrating content to new media as needed. Research data preservation can be accomplished by placing the data in a trusted repository for long-term curation.

Preservation best practices

There are simple steps a researcher can take to ensure their data are ready for preservation at the close of a project:

- Select formats for research outputs (data or otherwise) that that are standard and commonly used, open source rather than proprietary, and text-based, rather than binary.
- Ensure that the data receives a unique identifier, which will make it citable and countable. (See *Use and Reuse* section for more information.)
- Create high-quality, machine-readable metadata. (See *Documenting Research Data* above.)
- Ensure that data are licensed properly with permissive terms (see *Dataset Governance* section above).

Repositories

Planning for data management includes selecting the eventual long-term home for the dataset, i.e., the repository. When selecting a repository, researchers should consider the following:

- Where are similar datasets preserved?
- What are the access and use policies for the repository?
- How long will the data be kept? How long should it be kept?
- Who manages the repository? An institution? A commercial provider?
- What are the costs associated with using the repository? How will those costs be paid?
- Are there policies in place for replication, fixity, monitoring, disaster recovery, and business continuity?

There are two main types of repositories available for research data: discipline-specific (e.g., disciplinary) repositories and general use repositories.

Discipline-specific repositories

These repositories are designed for housing specific types of data for a given domain of research. There are many disciplinary repositories for all types of datasets. Researchers are usually aware of the most commonly used repositories for their field. These repositories typically have specific requirements for file formats and metadata standards, and may have access restrictions for deposit or download of datasets. The requirements for data type, format, and metadata submission mean that disciplinary repositories are often better at aggregating similar data types and allowing for aggregation of datasets. This enables searching as well as meta-analysis.

For example, one of the most widely known disciplinary repositories is GenBank. GenBank houses genetic data and has strict standards for how the data should be formatted, as well as the types and format of metadata accompanying the dataset. Because of these standards, GenBank is a well-known resource for synthesizing genetic data and making new discoveries.

General use repositories

General use repositories may have more relaxed requirements for data submission, often accept a wide range of formats, and have less stringent metadata requirements. General use repositories may be owned by non-profit groups, commercial entities like publishers, or institutions. Institution-owned repositories (IRs) are often housed in the institution's library and can be important for archiving the entire package of research outputs associated with a project, which is critical for reproducibility.

Selecting a repository

All data should be archived in a repository, but choosing the right repository for a dataset can be challenging. If data are an obvious fit for a discipline-specific repository, then depositing data there will ensure it is discoverable and reusable by researchers from the corresponding discipline. However, many research projects have more than one type of data (e.g., genetic data, morphological data, and behavioral data for a species of elephant, and R code for analyzing the data). General use repositories would be acceptable for all of these types of data, which promotes reproducibility and tells a more complete story of the research that was conducted. Rather than selecting only one of these two options (disciplinary or general), researchers should consider (1) depositing all discipline-specific data with a clearly identifiable relevant repository in that disciplinary repository, (2) depositing a data package with all research outputs that do not have a home in a general repository, and (3) providing links within the metadata and readme files of that data package that provide the location of the data deposited in disciplinary repositories.

Repository software

As the need for archiving data grows, repository software is becoming more uniform and widespread. Many repositories are migrating from their custom software systems to open source software projects like Fedora, DSpace, and Dataverse. These projects have begun to develop communities that contribute code and enhancements, increasing the value of the platforms for data archiving and sharing.

USE AND REUSE

Identification and Linking of Datasets

For data to be used by others, there needs to be standardized ways to identify, cite, and link datasets and parts of datasets.

Identifiers

An identifier is a string of characters that uniquely identifies an object. The object might be a dataset, software, or other research product. Most researchers are familiar with a particular type of identifier, the digital object identifier (DOI®). These have been used by the academic publishing industry for uniquely and persistently identifying digital versions of journal articles for at least fifteen years, and recently their use has expanded to other types of digital objects such as posters, datasets, and code. Although the DOI is the most widely known type of identifier, there are other identifier types. Researchers do not necessarily need to understand the nuances of identifiers, however, since the data repository often chooses the identifiers to be used.

Identifiers are generally included in a metadata record, which also describes the location of the dataset. This means that identifiers are only useful if the metadata are kept up-to-date. If attributes of the dataset change, especially the dataset's location, the metadata must be updated to reflect this. Often repositories are responsible for ensuring accurate metadata associated with the identifiers they issue, but researchers should be aware that an identifier does not guarantee that others will be able to find a dataset.

More complex issues associated with data citation

Because data citation is based on a model developed for journal articles, there are some nuances of citing data for which best practices are still being developed. Deep citation— the need for referring to only part of a dataset, rather than the dataset in its entirety—is a commonly identified challenge. . Some databases can be extremely large; citing the entire database would not help others reproduce and/or reuse the specific data discussed in an article. Dynamic datasets are another challenge. If a dataset is constantly updated (e.g., streaming satellite data), there must be a way to timestamp or otherwise subset the streaming data. The research data community is working on developing solutions for these and other complications associated with data citation.

Citations

Once a dataset is shared via a repository, the creator and others must be able to point to that dataset. For journal articles, this is done via citation and the practice of citing data is now being promoted as a way to provide attribution for and point to datasets and other digital objects. Data citations look similar to journal article citations, with common core elements such as authors (often called

"creators" for datasets); date; dataset title; publisher (usually the repository housing the data or the institution associated with the dataset creators); and an identifier. (See *Identifiers* below.)

There is still debate within the scholarly communication community as to how to deal with data citations. They are not usually accepted as part of the traditional bibliography, although some publishers have begun to alter their policies on this practice. Other journals prominently display the associated dataset's citation before the abstract on the article's landing page. More generally, there is recognition in the community that providing access to datasets via data citation is an important way to ensure reproducibility of the journal's results.

Data citation is a relatively new topic of discussion among researchers and academic publishers, and several groups have formed to help address the issues around citation. The Research Data Alliance, DataCite, and other international groups are working to develop standards associated with citing data to help increase access to datasets.

Publishing data

The traditional model of scholarly communication involves publishing research results in a journal article. As data become a more important scholarly output, this model is being adopted as a way to communicate data as well. There is much debate within the research data community as to whether "data publication" is the best way to make research data accessible. However, some common themes have emerged that are generally acceptable by the research data community.

Available

Publishing data implies that it is available. The strict definition of "publish" is to make something public; therefore publishing data requires that it be publicly accessible. This can be loosely interpreted to mean the data is in the supplemental materials on the journal publisher's website, or is available on the creator's website, but these are not ideal scenarios for providing long-term access to datasets. Instead, publishing data is most often accomplished via depositing data into a trustworthy repository.

Citable

Publishing data implies that it is citable. Data available to the public must be easily referenced and located; a data citation helps to accomplish this. Data citations include the publisher, which can be used to identify the dataset's location. The identifier can also direct people to the published dataset. (See *Use and Reuse* above for more on data citation.)

Validated

Perhaps the most controversial aspect of publishing data is validation. The norm in the academic community is to publish research results in a peer-reviewed journal. Publications that do not undergo some type of peer review process are less trustworthy and less likely to be referenced.

By extending the publishing model to data, there is a suggestion that peer review (i.e., validation) should be a part of the publishing process.

Data poses interesting challenges associated with peer review. How is the importance of a dataset determined and how can its future impact be predicted? How can a dataset be validated without spending inordinate time and energy reviewing, using, and/or analyzing it?

One suggestion for parsing data peer review is distinguishing between technical and scientific review. Technical review focuses on easily checked attributes, primarily dataset and metadata completeness. This type of review does not require intimate knowledge of the associated discipline, and is therefore more easily accomplished. Scientific review, however, is more intensive. This often involves evaluating methods, examining the dataset for plausibility, and generally determining the reuse potential of the dataset, all of which require domain expertise.

Data publication models

Although the concept of data publication is still debated within scholarly communication groups, new models are emerging to fit the need for providing access to and credit for data. Some publishers are introducing data journals devoted solely to publishing datasets and accompanying descriptions. Other publishers are allowing data articles to be submitted as a new article type to their existing publications. But not all publishers are taking on the role of data publication. Some are deferring to existing data repositories, requesting that authors who are submitting traditional articles publish their data in a repository and provide citations to that dataset.

As data become a more important and recognized independent scholarly output, the data publishing model will likely evolve quickly. Interesting possibilities are emerging from new groups, such as FORCE11 and the Research Data Alliance, which will push the envelope of providing access to data via publishing. It is unlikely, however, that the publishing metaphor will be dropped completely since it is the familiar framework within which scholars are already working.

Credit and incentives

The current system

Until recently, the traditional model of credit in academia centered on publications. This system has been in place for hundreds of years and was the only available means of disseminating work. The system of publishing results has evolved to accommodate the need to evaluate a researcher's impact, and the adage of "publish or perish" has become a common refrain. To its detriment, the community has focused on quantity of publications (and number of citations), especially in high-impact journals, rather than things like enhanced collaborations, novelty of the work, and other more nuanced factors.

At the center of this system is the impact factor (IF). The IF was devised in the 1970s as a tool for research libraries to judge the relative merits of journals when allocating their subscription budgets. It is now being incorporated into the system to evaluate researchers, who are under pressure to choose the journal with the highest IF rather than the one most suited for the work. Few academics would argue that the credit and incentive structure for research is ideal, but what are the alternatives?

There are many ways for researchers to impact their field, including establishing new methods, writing blog posts, interacting with the general public to promote science, and engaging with other disciplines via social media such as Twitter. One of the most important research outputs in terms of impacting future research is the data: by providing access to well documented publicly available data, a researcher's work may be invaluable in years to come. The incentive system of publishing does not, however, encourage data sharing. Some community members are working to establish new paradigms for credit in academia, which would reward the types of impacts that are not captured with publications and their citations.

Altmetrics

This move towards alternative metrics, or "altmetrics," can be framed as providing information (metrics) about "alternative" research outputs, such as data, code, software, blog posts, and social media engagement. With the upswing in website analytics, blogs, Twitter feeds, and academic social sites like Mendeley, Zotero, and CiteULike, it is easier to capture information about the dissemination and uptake of these outputs. Examples of altmetrics are:

- Number of dataset downloads from a repository
- Number of views of a blog post
- Number of retweets of the link to an article
- Number of saves for an article on Mendeley
- Number of views of a presentation on the web

There is understandable apprehension in the researcher community about altmetrics. Traditional researchers are concerned that a social media presence will become more important than doing quality work. However, the intent is for altmetrics to supplement the existing credit model for publication, not replace it. Providing ways to measure impact that go beyond the traditional systems expands the options for evaluating work. If the number of tweets an article receives is irrelevant to a tenure and promotion committee, they should not be obligated to take this into account. But having the option to consider dataset downloads might be an important step in incentivizing data sharing.

Credit for data

Many researchers spend far more time collecting, cleaning, analyzing, and re-analyzing data than writing papers. But only the final step in the research life cycle is currently considered an

important research output. Preparing data for sharing requires time and effort; if there is no incentive structure to reward doing this work, then researchers will continue to keep datasets private, thereby inhibiting reuse and reproducibility.

Changing the system requires movement on several fronts: data citation standards must be established, researchers must be more aware of the benefits of good data management and best practices for handling data, and tenure and promotion committees must consider research impact beyond publishing in high-impact journals. Requirements from funders for data management and sharing are laying the groundwork for bringing data to the forefront as an important component of research. However, without positive incentives to accompany the mandates, it is unlikely that high-quality, well-documented data will be shared.

CONCLUSION

The guidelines in this primer give insight into how best to plan, document, and preserve datasets responsibly so that they are easier to use and share, as well as making the opportunities for collaboration with other researchers less difficult. While there are still challenges that the research data community is working on, progress is being made and those who collect, use, and store data have more tools and resources than ever before to ensure that their datasets will be available for future reuse.

APPENDIX A: RESOURCES

General Resources

- Tools for creating data management plans
  - Data Management Planning Tool (DMPTool): http://dmptool.org
  - DMP Online: https://dmponline.dcc.ac.uk/

- Data Management Primers
  - ICPSR Guide to Social Science Data Preparation and Archiving. ICPSR is one of the premier social science data repositories. Their handbook on preparing data for archiving is extremely thorough, and will ensure high quality data.
  - UK data archive. The UK Data Archive has an excellent knowledge base for the creation and management of data. These guides are an excellent place to start for data project management issues.
  - DataONE primer on data management. This PDF covers the basics of data management, arranged in the context of the data lifecycle and geared towards researchers.
  - Data Management for Libraries: A LITA Guide. A short guide on data management topics, focused on helping libraries understand data management to better provide services for researchers that they support.

- Training Materials
  - Digital Curation Training for All: a collection of slides, resources, and materials that cover the basics of research data management from the Digital Curation Centre.
  - MANTRA Course. MANTRA is focused on providing educational materials for researchers on important data issues. These lessons are a great start for librarians looking to provide data management workshops.
  - DataONE data management education slide decks
  - UK Data Archive training resources on managing and sharing data

Projects and software referenced in the text

- EML: https://knb.ecoinformatics.org/#external//emlparser/docs/index.html
- FGDC19115: https://www.fgdc.gov/metadata/geospatial-metadata-standards
- Version control
  - Git: http://git-scm.com/doc
  - GitHub: http://github.com
  - Bitbucket: http://bitbucket.org
  - Mercurial: http://mercurial.selenic.com/
  - Subversion: http://subversion.apache.org/
- Software
  - R: http://www.r-project.org/
  - Taverna: http://www.taverna.org.uk/
  - iPython: http://ipython.org/
  - Kepler: https://kepler-project.org/
  - LaTeX: http://www.latex-project.org/
- Licenses
  - Creative Commons: http://creativecommons.org/
  - Open Data Commons: http://opendatacommons.org/
- Storage
  - Dropbox: http://www.dropbox.com
  - Google Drive: https://www.google.com/drive/
- Repositories
  - GenBank: http://www.ncbi.nlm.nih.gov/genbank/
  - Fedora: https://getfedora.org/
  - DSpace: http://www.dspace.org/
  - Dataverse: http://dataverse.org/

- Initiatives
  - Research Data Alliance: https://rd-alliance.org/
  - FORCE11: https://www.force11.org/

- Identifiers
  - DOI: http://www.doi.org
  - DataCite: http://datacite.org

- Reference management
  - Mendeley: https://www.mendeley.com/
  - Zotero: https://www.zotero.org/
  - CiteULike: http://www.citeulike.org