CITI PROGRAM

Alexander Brown
ID 4843136

Data Management (RCR-Basic)

Utah State University - Physical Science Responsible Conduct of Research Course

Switch View

Data Management (RCR-Basic)

Content Author

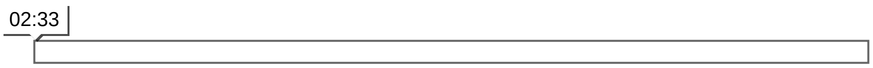
- **Reid Cushman, PhD**

  CITI Program

Introduction

Please review at least one of the videos below before you begin reading the module. Each video is approximately three minutes long.

- Life Sciences - Data Management
- Social/Behavioral/Education Sciences - Data Management



This module will describe strategies that may help prevent some of the challenges illustrated within the videos.

In its classic textbook form, the scientific method relies on a hypothesis-driven experiment. Data are collected, then analyzed and interpreted. The results are communicated to others, who can seek to assess the quality of the work. A sufficient number of hypotheses consistent with the data help to build support for a theory. And so it goes.

In the real world, it is rarely this clean and linear. Moreover, the opportunities for true controlled experiments – with a single intervention tested in a randomized experimental group and compared to a randomized control group – are often limited. Many other quasi-experimental designs are used such as cohort and case-control studies, cross-sectional studies, and individual case analyses. The key requirements, whatever the design, are that the study is described in enough detail so that it can be assessed and, in principle, replicated by others. Studies that are poorly designed or built on a foundation of false or misleading data are at best useless and a waste of resources. At worst, they can cause setbacks and even direct harm if actions are taken on the basis of inaccurate information.

While this module provides an overview of data management issues, it is not a substitute for more formal training on the topic. Understanding the requirements for appropriate data management is critical to a researcher's effectiveness.

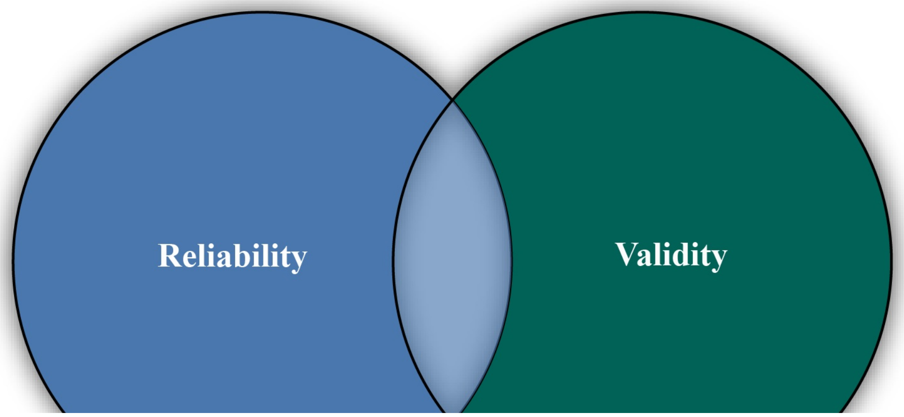
Learning Objectives

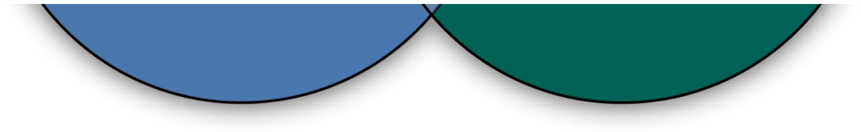
By the end of this module, you should be able to:

- Describe core issues about data management that arise during the research process.
- Discuss methodological, technological, and legal-regulatory considerations that affect data management decisions.
- Describe ethical and compliance issues relating to data ownership, data sharing, and data protection.

The Nature of Data

Data ideally represent information obtained by following a carefully defined research plan. For many types of research, a starting point is **operationalizing concepts**. This process involves defining concepts in such a way that they can be accurately and precisely measured, with the goals of both data reliability and validity.





Reliability

Reliability is usually defined in terms of replication, or whether one obtains the same result on repeated measures of the same phenomenon. In other words, it refers to consistency of outcome if the conditions of things being studied have not themselves changed.

Validity

In this context, **validity** refers to the more complex notion of whether operationalized concepts – using the measurement definitions, devices, and methods of the research plan – actually measure what they **purport to measure**.

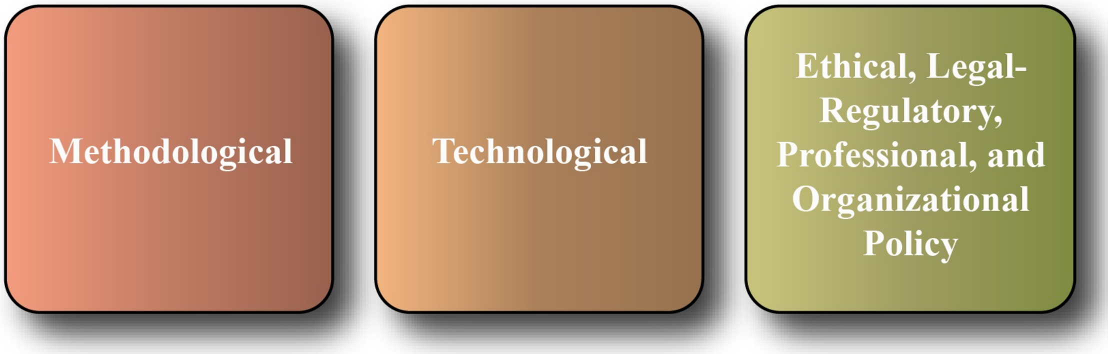
Achieving Reliability and Validity

Achieving reliability and validity can be relatively easy or extremely difficult depending on the context. It can be among the most difficult challenges in designing a study.

Different fields have different standards for the attributes that constitute measures of sufficient quality. **Therefore, researchers need to learn the discipline-specific standards within their realm of research**.

Categorizing Data Management Issues

Data management issues can be categorized in many ways. In this module, they are divided into three main types. Each of these categories is then discussed in the context of the typical cycle of study design, data acquisition, analysis, reporting, and post-study housekeeping.


Methodological
Technological
Ethical, Legal-Regulatory, Professional, and Organizational Policy

Methodological

Methodological issues are those associated with ensuring that a research study is well-designed with respect to statistical analysis. This includes choices about:

- Population and sample selection
- Group assignment
- Data collection
- Data analysis
- Data presentation

Crucial methodological issues include steps used to reduce error, random factors that blur a relationship, and those used to lessen bias, non-random factors that skew a result and can lead to false conclusions.

Technological

Technological issues pertain to use of the available tools and processes for managing information throughout a research project's lifecycle and after the project concludes. This is sometimes called **data lifecycle management**. Its main goals are to ensure data confidentiality, integrity, and availability for as long as the data exist. Achieving such goals requires a combination of measures, which balance prevention, detection, and response to problems.

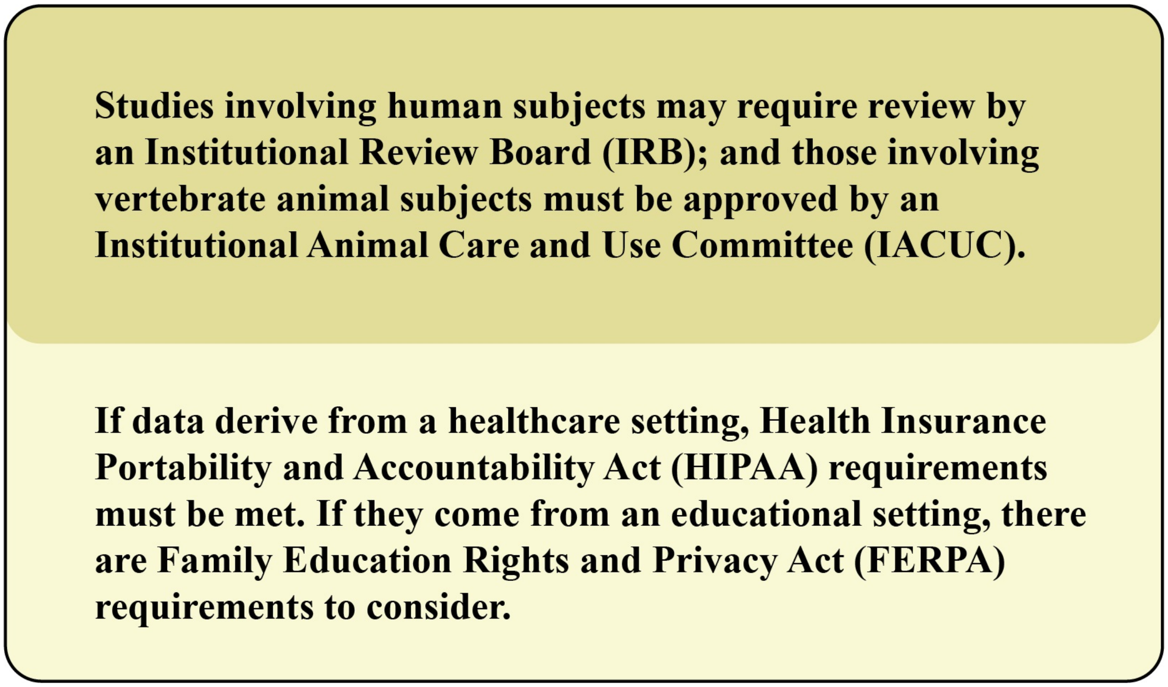
Ethical, Legal-Regulatory, Professional, and Organizational Policy

Ethical, legal-regulatory, professional, and organizational policy issues refer to what is required, recommended, discouraged, or prohibited by governmental entities and other organizations that have oversight over research activities, including an employer's standards.

Study Design

Study design involves planning for all the steps in the research cycle. Careful planning is in each researcher's interests in order to achieve an outcome that meets the methodological standards of the field or profession. However, it is also spurred by compliance with legal-regulatory and organizational requirements.

For example, in the U.S.:



Studies involving human subjects may require review by an Institutional Review Board (IRB); and those involving vertebrate animal subjects must be approved by an Institutional Animal Care and Use Committee (IACUC).

If data derive from a healthcare setting, Health Insurance Portability and Accountability Act (HIPAA) requirements must be met. If they come from an educational setting, there are Family Education Rights and Privacy Act (FERPA) requirements to consider.

Other federal, state, and even local regulations may govern the conduct of research, such as the use of hazardous chemical, biological, or radioactive materials. Beyond materials, certain technologies used in the research may also be subject to regulatory controls.

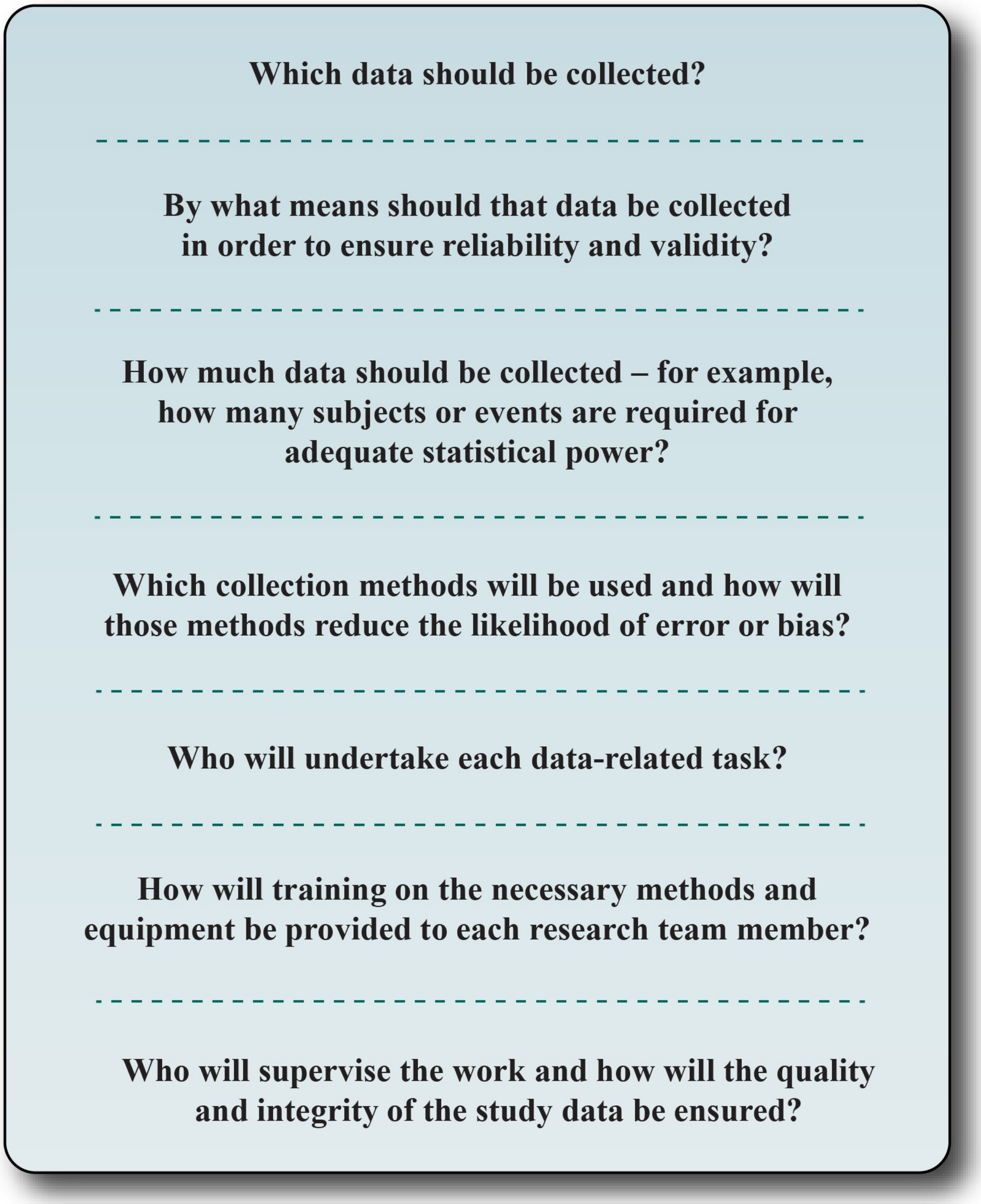
Legal-regulatory rules from other countries may apply depending on which entity funds the research and where the research takes place. For example, if the research is located outside of a researcher's organization, it may require a range of access permissions depending on the jurisdiction.

Organizations typically create resources to guide their researchers through methodological, technological, and legal-regulatory complexities. In general, organizations do not want to halt or delay research. However, they want to ensure that organizations do not want to halt or delay research. However, they want to ensure that their research activities do not violate any important rules or cause harm. Researchers have a fundamental obligation to familiarize themselves with the guidance their organization provides with regard to study design and conduct.

Case Study **Using Electronic Databases**

Data Collection

A collection plan should govern the acquisition of data. Issues to consider include:



Which data should be collected?

By what means should that data be collected in order to ensure reliability and validity?

How much data should be collected – for example, how many subjects or events are required for adequate statistical power?

Which collection methods will be used and how will those methods reduce the likelihood of error or bias?

Who will undertake each data-related task?

How will training on the necessary methods and equipment be provided to each research team member?

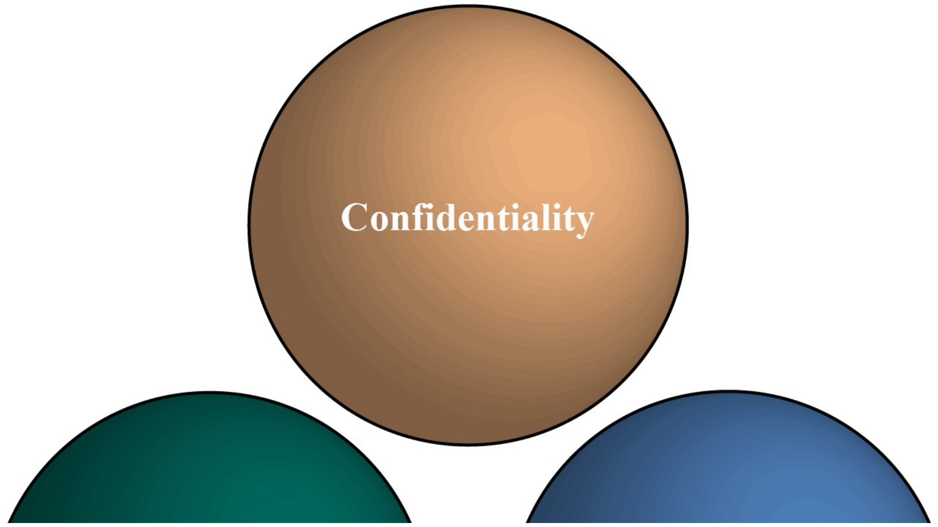
Who will supervise the work and how will the quality and integrity of the study data be ensured?

Data collection planning allows researchers to consider proactively when and where adjustments are necessary in the methods, procedures, or quantity of data to collect. As changes might become necessary, for example, collecting fewer data points than originally planned given acquisition costs, researchers must proceed cautiously. There is a line beyond which the quality of a study is altered so much that it is no longer worth undertaking. Making these choices well in advance and coherently across all study dimensions is likely to lead to more objective and trustworthy determinations.

Data may be "mined" from existing sources or newly collected for a study. In general, pre-existing data have the advantage of being cheaper and faster to acquire; in some cases, obtaining them may be nearly costless. However, a key disadvantage is that the original collector or owner of the data may impose strict terms and conditions for using the data. Another potentially more serious disadvantage is whether the original data collection methodology was appropriate for a different study.

Data Management and Information Technology

Planning out a research study must also include the purely technological information technology aspects of data management. As mentioned above, it is common to pursue three main goals: confidentiality, integrity, and availability.



Confidentiality

Confidentiality refers to preventing inappropriate users or uses of the data. It is usually discussed within the context of protecting the privacy of research subjects, which is a regulatory and ethical requirement if those subjects are human beings. Confidentiality also involves protecting intellectual property related to the research.

Integrity

Integrity refers to the responsibility to record, store, and preserve data appropriately during the full lifecycle of a study, which helps to improve the accuracy of data analysis.

Availability

Availability refers to ensuring that all appropriate users have access to data whenever necessary. As with integrity, availability concerns can extend past the formal end of the study, to ensure access by others who wish to replicate the work.

Upholding confidentiality, integrity, and availability may require very different approaches, depending on the type of data and the technologies used to record it. For example, the requirements to maintain paper notebooks differ from those for calibrating instruments that generate electronic data.

Researchers must adhere to legal-regulatory, organizational, and disciplinary-professional requirements, and make sure every member of the research team is professional requirements, and make sure every member of the research team is appropriately trained to follow them. This is particularly important when complex data acquisition or storage instruments and methodologies are used. In addition to training, it will be necessary to specify procedures for periodic testing of collection and storage devices, and confirming data backups. Researchers should always have the ability to recover data in an uncorrupted and unaltered form.

An increasing number of U.S. funding agencies, such as the National Science Foundation, U.S. Geological Survey, and National Oceanic and Atmospheric Administration, require an **overall data management plan** for projects that they fund. A catalyst for the increase in these federal requirements was a memo issued in February 2013 by the director of the Office of Science and Technology Policy, directing U.S. agencies that fund over $100 million dollars annually in research and development expenditures to increase public access to the results of the research.

Researchers must have a plan and document that the plan has been honestly followed. When deviations are discovered, they should be corrected as soon as possible, and the corrective actions, or plan revisions, should be formally recorded. In some cases, a significant deviation from the research plan may need to be reported to an oversight body, such as an Institutional Review Board.

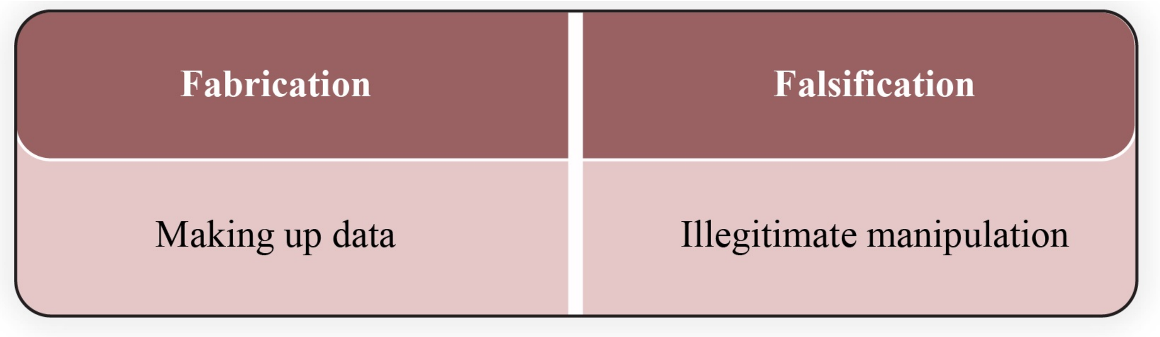
Data Analysis

Although differences in specific disciplines may exist, in general, the choice of methods, statistical or otherwise, used to analyze data is ideally made before data collection begins. However, unexpected results may sometimes necessitate the use of analytic methods that were not contemplated during study design. It may be appropriate to use new tools, if there is a sound justification for doing so.

Researchers must always have logical and justifiable reasons for selecting a particular approach; otherwise, the researcher may appear to be "venue shopping" in the hope that something of apparent significance will appear if enough different approaches are tried.

Oversight bodies, such as an Institutional Review Board, may sometimes detect and address methodological problems during the review process. However, this role is not typically the principal focus of oversight bodies. In general, such entities only question research methodology within the context of upholding regulations or organizational policies. Therefore, researchers should not rely solely on an oversight body to identify methodological problems for them.

Beyond methodological errors lies the territory of data fabrication and falsification, which are two forms of research misconduct.



Fabrication	Falsification
Making up data	Illegitimate manipulation

Fabrication, or making up data, is obviously unethical and condemned by the research community. The boundary between illegitimate manipulation, or falsification, and legitimate "enhancement" is not always clear, such as when considering how to present data or whether to discard data outliers. In part, this turns on whether there is intent to deceive, which is a component of a research misconduct finding. However, even without the intent to deceive, sloppiness, ignorance, or other detrimental research practices can compromise the research record (National Academies 2017).

Researchers should learn the standards for data analysis in their field and avoid sloppy research practices. Regardless, the expectation is that whatever is done with respect to data "adjustment" will always be fully documented so that others can view and assess the legitimacy of the approach used.

Reporting Results

In whichever venue or format they choose to present their results, researchers have an obligation to be:

Clear

Honest

Unbiased

This includes providing a sufficiently detailed description of the study design and execution, and information about if and why any data were excluded.

Researchers also need to describe their methods thoroughly, so that the study can be assessed and, in principle, replicated by others. Moreover, data must be represented accurately in charts, tables, graphs, or other formats. For example, it is inappropriate to distort how data are displayed in order to generate a conclusion that appears stronger distort how data are displayed in order to generate a conclusion that appears stronger than it actually is.

For some types of research, the range of journals, conferences, and other venues in which to report findings can be very broad. For others, the nature of the topic and discipline may severely constrain that choice. In most cases, researchers desire to publish their findings as soon as possible to avoid being scooped by someone else. However, a delay can sometimes be advantageous, especially when intellectual property considerations are present, such as a pending patent.

Choices about timing and venues for reporting research are largely a matter of personal strategy, but there are limits. For example, if a collaborative team is conducting the research, then team member preferences will need to be discussed. Also, project funders may have views about when and where to report; indeed, timelines for reporting may be specifically established in the funding agreement.

Assuming the research has value, to delay publication may deny a significant benefit to society or at least to the other researchers in the field. Alternatively, some research has national security implications, and in that situation, publishing too quickly could lead to harmful consequences. Therefore, delay may not only be appropriate, but legally required.

Data Sharing

Case Study

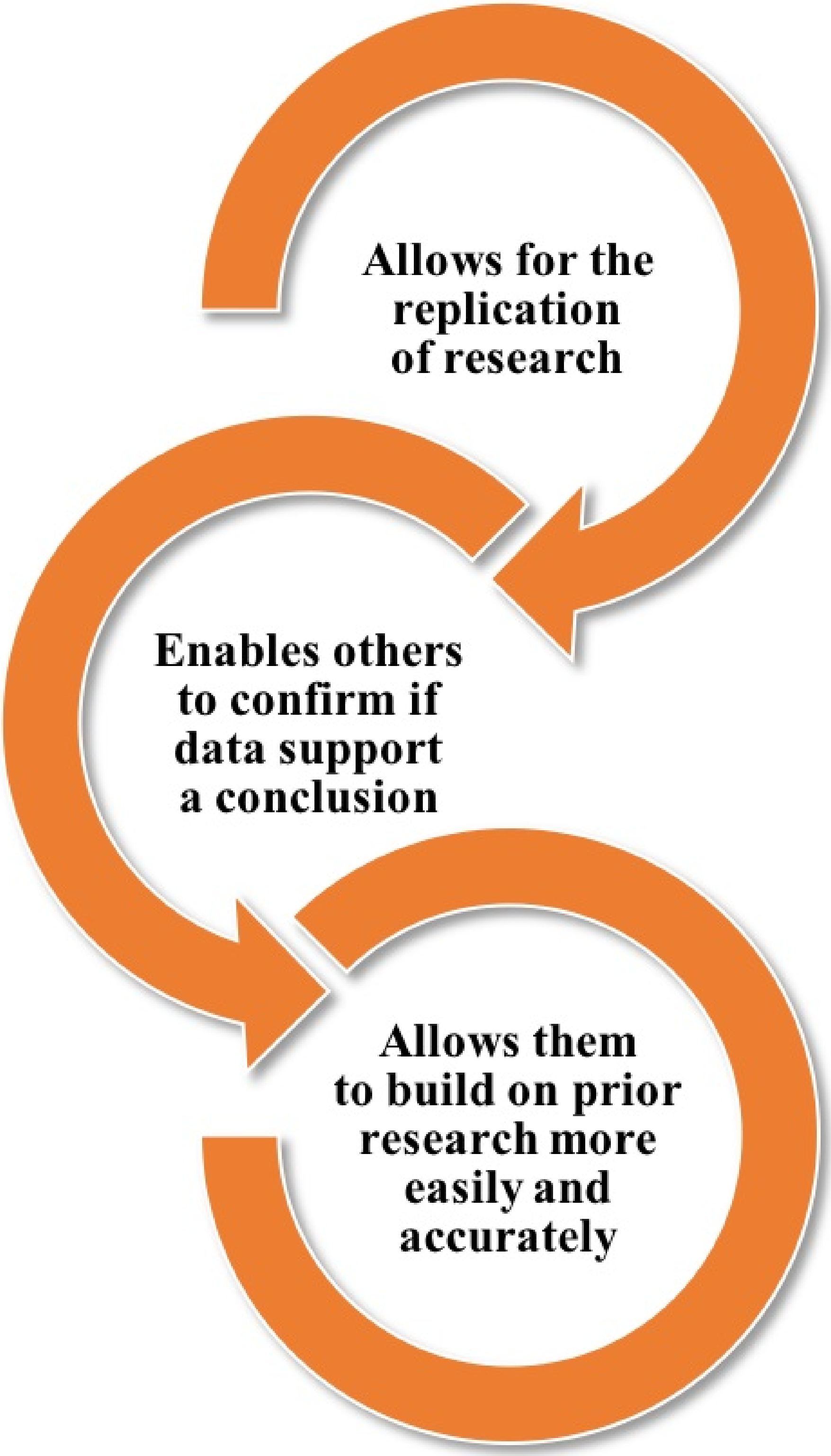
Sharing Data Internationally

Some types of data, particularly those relating to human subjects research, carry with them an obligation to prevent inappropriate disclosure. That is, there may be a duty to avoid sharing the data, at least in identifiable form, with persons who are not supposed to have access to them.

Intellectual property considerations can also impose obligations on researchers to avoid sharing the data outside a defined set of persons. **Non-disclosure agreements**, if signed by researchers or others representing the researcher's organization, can specify these obligations in great detail. For example, a private company might require review and/or approval before the results from research that it funds are published, because of commercial concerns. National security concerns may also constrain data sharing or disclosure – for example, with research that has **military or "dual-use" applications**.

Conversely, entities that fund research, including U.S. federal agencies such as the National Science Foundation (NSF) and the National Institutes of Health (NIH), either encourage or require data sharing. Journals may also impose obligations to share data for any article that they publish.

Arguably, the principles of being a responsible researcher alone impose obligations to share data with others, provided there are no legal or regulatory constraints that apply to the research. The primary motivation for sharing is the same:


Allows for the replication of research

Enables others to confirm if data support a conclusion

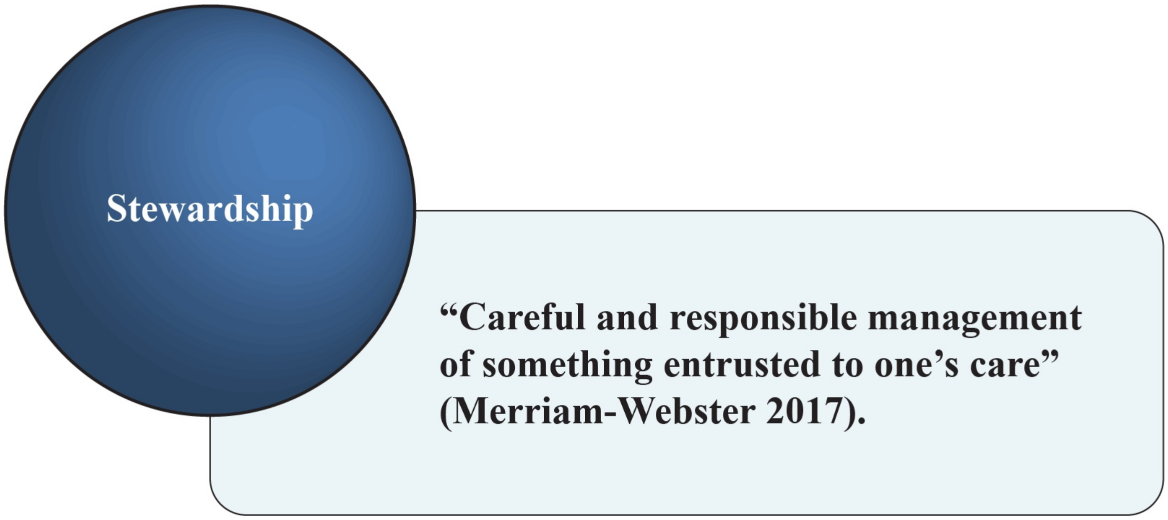
Allows them to build on prior research more easily and accurately

Obligations to share data with other research teams do not necessarily imply it needs to be done immediately. Each research team needs time to complete its work and publish the results before sharing with others. It is also understood that the costs of data sharing may be recouped.

Ownership, Stewardship, and Data Protection

Ownership can be formally defined as a set of rights, privileges, and responsibilities with respect to an object. The object may be a piece of tangible property, like a car, which can be "possessed" by only one person at a time; or it may be a less tangible construct, like a data set, where copies can potentially be possessed by many individuals at the same time.

Because of the different nature of data, it is common to refer instead to **data stewardship**.



“Careful and responsible management of something entrusted to one’s care” (Merriam-Webster 2017).

Ownership or stewardship of data can include rights to allow or deny access. It usually also implies responsibilities with respect to that data, such as upholding confidentiality, integrity, and availability.

However, just because researchers work with a data set, it does not necessarily follow that they own the data or are its designated primary stewards. For example, when the U.S. government funds a grant, it is normally issued to the organization, not the researchers, and the data are typically owned by the organization. Academic institutions and other research-producing organizations often assign responsibility for being stewards of the data to the research team. The research team may possess some intellectual property rights over the data depending on organizational or funder policies. Each team should learn what the ownership standards are for their specific project.

Both ownership and stewardship of data implies obligations for long-term management. This means that data are protected throughout the data lifecycle (from creation to destruction), so that appropriate sharing can occur and unauthorized sharing is prevented.

As part of this, data lifecycle management considerations drive the technical choices about devices, media, and processes to protect data over the long term. For example, if a storage device becomes obsolete, retrieving data from it may become difficult or impossible.

Because most researchers are not information technology specialists, consulting with experts may be necessary to formulate an appropriate data management plan for the project duration and even afterward. Long-term retention requirements for data are common. Funders, journals, a researcher’s organization, or even government entities may all have data retention provisions that apply.

Summary

The management of data generated along the road to discovery requires researchers to develop systematic plans and processes, so that they can eliminate errors and bias to the greatest extent possible to ensure the quality of the research. They must also exercise appropriate data stewardship. Research does not always proceed in a neat and orderly fashion, and accordingly, data management plans are always potentially subject to change. However, even the simplest project needs a plan.

Acknowledgments

This module is based on an earlier version co-authored by Reid Cushman and Deborah Barnard. Jason Borenstein also provided comments and advice for the module.

References

- Merriam-Webster. 2017. “Stewardship.” Accessed August 4, 2017.
- National Academies of Sciences, Engineering, and Medicine. 2017. *Fostering Integrity in Research*. Washington, DC: National Academies Press.
- National Institutes of Health (NIH). 2015. “Dual-Use Research.” Accessed July 26, 2017.
- National Science Foundation (NSF). 2017. "Dissemination and Sharing of Research Results: NSF Data Management Plan Requirements." Accessed July 23, 2017.

Additional Resources

- Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard. 2014. *Managing and Sharing Research Data: A Guide to Good Practice*. Sage Publishing.
- National Academy of Sciences, Engineering, and Medicine. “Board on Research Data and Information.” Accessed March 29, 2019.
- National Institutes of Health (NIH). 2007. “NIH Data Sharing Policy.” Accessed March 29, 2019.

- U.S. Department of Education (ED). 2018. "Family Educational Rights and Privacy Act (FERPA)." Accessed April 6, 2018.
- U.S. Department of Health and Human Services (HHS). 2018. "Health Information Privacy." Accessed April 6, 2018.

Original Release: June 2014

Last Updated: April 2019

This module has a quiz.

Return to Gradebook

Take the Quiz

SUPPORT

888.529.5929

8:30 a.m. – 7:30 p.m. ET

Monday – Friday

Contact Us

LEGAL

Accessibility

Copyright

Privacy and Cookie Policy

Statement of Security Practices

Terms of Service

CITI PROGRAM