RESEARCH DATA MANAGEMENT: GUIDE FOR INVESTIGATORS
RESEARCH DATA MANAGEMENT: GUIDE FOR INVESTIGATORS

TABLE OF CONTENTS

I. INTRODUCTION

II. RESEARCH DATA MANAGEMENT PLANNING
   A. MANAGING DATA ALONG THE RESEARCH LIFE CYCLE
   B. INTELLECTUAL PROPERTY (IP)
   C. DESIGNING A RESEARCH DATA MANAGEMENT PLAN
      i. PHASES OF THE DATA MANAGEMENT LIFECYCLE
         1. Design
         2. Create/Obtain
         3. Store
         4. Use/Reuse
         5. Publish
         6. Share
         7. Archive
         8. Disposal

III. ADDITIONAL RESOURCES

IV. ATTRIBUTION, SHARING, AND ADAPTING
I. INTRODUCTION

Research Data Management: Guide for Investigators ("Guide") addresses how to plan and manage research data considerations for researchers who expect to work with confidential or sensitive information involving individuals in the course of a research study. The Guide follows the research data management lifecycle, highlighting key issues to consider and a roadmap to follow through the lifecycle of data from the beginning planning phases, through design, and all the way to proper destruction and disposal of the data. It is important to remember that data management planning is an iterative process. As the research study changes, so should the plan.

Since technology risks are always evolving, the Guide is designed to provide a technology-neutral approach toward research data management. We are looking to educate researchers about the various types of issues, considerations, and agreements that must be evaluated throughout the data management lifecycle.

Researchers should only collect, store, and archive the minimum research data necessary to complete the research study. The collection of more data than needed increases the risk of loss of privacy, confidentiality, theft, etc. Researchers should be mindful that Data Use Agreements and other contracts may require data to be disposed of securely. Depending on the type of data, there may be regulatory requirement(s) around data retention.

The Guide does not supersede institution research data management policies and procedures. You are always advised to reach out to your institution research administration, legal, and IT departments with any questions.

For additional resources on how to protect research data, please see the other materials developed by the Harvard Catalyst Data Protection Committee. The available resources have been developed and vetted in collaboration with local institutional research administration, IT, and other departments. To learn more about the resources available, please see the Harvard Catalyst website.

---

1 This document is not intended to be an exhaustive resource. It does not address research integrity, conflict of interest, institutional policies, quality assurance, or FDA submission criteria: 21 CFR Part 11.

II. RESEARCH DATA MANAGEMENT PLANNING

A. MANAGING DATA ALONG THE RESEARCH LIFE CYCLE

The way you structure and collect the data can have practical, legal, and ethical implications. A data management plan (DMP) is a formal document that outlines how data will be handled during your research, including after the study has been completed. The DMP is a planned roadmap on how to manage and protect your data, meet funder requirements, etc. A well-structured and executed DMP can help reduce and mitigate risks to research data, protecting the confidentiality of participant data, and allowing for more efficient use of resources (e.g., amount of time staff, preparation for publications, budgeting for data management from study inception, etc.).

The DMP planning should begin very early in the research process, when the study is being conceptualized, not when the data is being created and collected. Regardless of whether you are prospectively collecting data as part of your research, or acquiring it from a third-party or another researcher, it is essential to think through all of the ways data will become incorporated into your project, including whether you have plans to make this data accessible to others in the future. Your plans should help to assure appropriate use and access as well as the privacy, security, and confidentiality of the data.

To make developing a plan easier, we have broken down the DMP into segments or “elements,” which can be customized to the research study.

Research data lifecycle elements include:

- **Design**- Research study design elements regarding how the data will be documented and whether or not a contract or agreement is needed.
- **Create/Obtain**- When/how the data will be received, acquired, or generated from research participants, third parties, other data sources. *Only collect the data that is needed.*
- **Storage**- Where and how the data will be stored, backed-up, and accessed. If data is required to be reused, the data is analyzed and stored in a way that it can be recreated and reused.
- **Use/Reuse**- Plans for how data will be used and anticipated secondary use.
- **Publish**- Considerations and requirements for what needs to be shared with journals, publishers, and funding agencies.
- **Share**- Plans and documentation methods for how data will be shared.
- **Archive**- Long-term data management and plans for data retention.
- **Disposal**- The plan for disposing of data securely; ensuring the data is destroyed properly. Considerations include the size of the data set and the storage platform.
B. INTELLECTUAL PROPERTY

Governed by federal and state laws, Intellectual Property (IP) is the right to protect the products of human intelligence and creation, such as copyright, trademark, and patents. Determine if the data is protected by copyright and consider IP throughout the data lifecycle including IP rights to the researcher, institution, funder, outlined in research agreements.

I. DATA CONTENT- The type of data being collected, stored, shared, and how you plan to use and destroy it all need to be considered.
   a. Collect/Obtain- If the data was sourced from another database, you will need to check the original database for IP considerations prior to making your data available. These considerations include attribution, notification of use, redistribution, quality control/standards, and risk.
   b. Storage- Consider what information needs to be stored and the best location to store it for both short-term and long-term access. If the data contains sensitive information, make sure extra precautions are taken to ensure privacy. If you are storing data in the Cloud, check to see where the servers are located and if the laws for that location are still best suited for your research and the data being collected. For example, if you collect research data in Massachusetts and store it on a server located in Ireland, your IP protections and enforcements capabilities may change because the data is located in Ireland.
   c. Sharing- Limit the authorized personnel who have access to the data. Data access should only be granted to personnel who need to have access to complete the research. Always be sensitive to the type of data and who it is being shared with. Sharing data can increase the risk of the data getting intercepted by unauthorized personnel; always take extra precautions when sharing sensitive data.
   d. Use- Consider the type of data when determining the use.
   e. Destruction- Sensitive data may require additional steps to ensure the data has been properly destroyed in a specified time allotment. Check with your IT department, research compliance, and grants offices for any destruction specifications for your research study.

II. LICENSING-
   a. Open Data Commons (ODC): Public Domain Dedication, Attribution License, Open Database License
   b. Creative Commons (CC): CC Zero (CC0), Public Domain Mark (PDM)

III. COPYRIGHT- Data copyright considerations should be made for both the data content and the databases the data is stored, shared, and archived on. Clarify the copyright parameters, making sure you fully understand what is protected by copyright.

IV. FUNDING- Your institution, sponsor, agency, etc. can have their own policies on copyright and data ownership.
   a. Federally funded research- Per the Bayh-Dole Act (37 CFR 401), inventions from federally funded research projects must be reported to the government. The funded institution is permitted ownership under the Act, but the government also has the right to practice the subject invention.
      i. The funding recipient has two months to report the invention to the federal funding agency and one year to file for a patent. They have two years to notify the funding
agency of retention of ownership plans; to provide a license to the funding agency to practice the invention; and to report any commercialization and licensing efforts.

ii. The awardee must also keep the funding agency informed regarding progress in patenting and/or commercializing the invention.

b. State funded research- Most state contracts are issued in accordance with the Commonwealth Master Agreement and statewide contracts. The state government your institution is located in reserves the right to identify copyrightable works.

c. Sponsored/Agency funded research- Each institution handles patent rights to sponsors differently. Check with your institution as to how they engage with sponsors/agencies.

d. Institution funded research- Prior to the start of the research, work with your Office of Sponsored Programs for guidance on how to appropriately include IP in your research agreement.

e. Other scenarios may conflict with IP rights to your research. Always make sure to work closely with your institutional officials to ensure your rights are protected prior to beginning the research study.

V. LOCATION-

a. Always consider the location where your data will be collected, stored, analyzed, etc. If the location for one of those steps crosses state or country borders, your IP protection and enforcement depend on the laws of that location.


c. Multi-Site/Multi-Center Research- United States law does not protect the copyright of factual data. Facts cannot be copyrighted.

VI. ACCESS TO THIRD PARTY DATA OR ASSOCIATED RESEARCH ARTIFACTS- If you grant access to the research data, your research may require additional considerations based on the access agreement.

VII. ADDITIONAL CONSIDERATIONS-

a. Do agreements contain language prohibiting use or publication due to IP considerations?
C. DESIGNING A RESEARCH DATA MANAGEMENT PLAN

**Design:** Research study design elements regarding how the data will be documented and whether or not a contract or agreement is needed

**TYPE OF DATA**
Accurately describing your data and the plan for obtaining and use will be paramount for both your funding proposal and any related IRB application.

The first step in preparing a research proposal or research plan is to clearly determine and define the type of data that you are looking to obtain or create. Data can be categorized to help determine the type of data and mechanisms to capture it. Most common data categories include:

1. **Observational** - captured in real-time (e.g., survey results, sensor readings, images, etc.)
2. **Experimental** - generated results (e.g., gene sequences)
3. **Simulation** - machine generated (e.g., predictive modeling)
4. **Derived/Compiled** - generated from existing datasets (e.g., data mining)

Describing data is an important step to enhance research data security. We recommend using the quick tips and checklist resources: [Data Privacy and Security Planning Checklist](http://catalyst.harvard.edu/pdf/regulatory/DataPrivacyandSecurityPlanningChecklist.pdf) and [Top Ten Research Data Security Tips](http://catalyst.harvard.edu/pdf/regulatory/Top%2010%20Data%20Protection%20Tips%20for%20Researchers.pdf).

Depending on the type of data, it may require special management considerations. Consult with your research data protection contact (for institutions affiliated with the Harvard Medical School CTSA, Harvard Catalyst, please reference the [Harvard Catalyst Regulatory Atlas](http://connects.catalyst.harvard.edu/regulatoryatlas/?mode=clid=51)) if any of the following are true:

- Data will be stored on a secure, password protected, server behind a firewall.
- Data will be stored on a mobile computing device (e.g., laptop, smartphone, iPad, etc.) or removable media (e.g., flash drive, CD/DVD, etc.) during the research study.
- You will be using a cloud vendor (see the [Guidance for Researchers Using Internet Cloud Computing Services and Apps](http://catalyst.harvard.edu/pdf/regulatory/Guidance%20for%20Cloud%20Services%20and%20Apps.pdf)), commercial service, or other third party platform for storage, backup, access, analysis, de-identification, re-formatting, or other service; each vendor or commercial solution may have different requirements, encryption, or fees to be considered.
- Data will be transferred or transmitted to vendor, contractor or processed by a third party for linking, correlation, analysis, etc.

To ensure that your DMP is successful, designate limited study personnel in charge to make sure the project adheres to the plan and changes are updated to the DMP as they occur.

**BUDGETING**
Advanced planning facilitates accurate budgeting; data-intensive projects involving sensitive information may require additional security measures with costs that exceed routine departmental support. Work with

---

7 Harvard Catalyst Data Protection Website: [http://connects.catalyst.harvard.edu/regulatoryatlas/?mode=clid=51](http://connects.catalyst.harvard.edu/regulatoryatlas/?mode=clid=51)
8 Linking data or combining datasets can increase the risks of re-identification, should be considered carefully and fully described in your research protocol.
grant management and clinical trials offices to develop appropriate budgets based on what you will need to manage the data throughout the lifecycle.

Budgeting considerations:
- Cost of obtaining existing datasets
- Rental/storage space (e.g., online data platform)
- Personnel costs to manage the data both during collection and future use
- Incidental costs to change the storage media for the data (e.g., converting and moving the data from CD to USB)
- Long term storage cost for data that will be archived
- Costs for securely destroying the data
- Anticipation of costs for future use or unanticipated secondary use of the data

Check with your funder to clarify which data management plan elements apply, if any. Some recommended guidance documents to consider while developing your DMP include: NSF ENG Data Management Plan Requirements, NIH Data Sharing Policy, NSF Grant Proposal Guide (2013), NSF’s Public Access Plan (2015), NIJ’s Applying for Data Resources Program Funding, and the DMPTool from California Digital Library.

CONTRACTS AND AGREEMENTS

Once the data needed for the research has been identified, determine if receiving, sharing, acquiring, or generating data will require a contract or agreement. These agreements are usually reviewed by the IRB, IT, and the contract office.

Applicable confidentiality and data ownership agreements include:
- Confidentiality Agreements
- Data Use Agreements (DUA)
- Non-Disclosure Agreements (NDA)
- Material Transfer Agreements
- Clinical Trial Agreements
- Certificates of Confidentiality (for more information, see the IRB section)

PROTOCOL DESIGN DEVELOPMENT AND PREPARATION FOR IRB REVIEW

According to the Common Rule, as defined by the US Department of Health and Human Services (HHS) in 45 CFR 46.102(f), human subjects research is research relating to a living individual about whom an investigator obtains:
1) Data, through intervention or interaction with the individual; or
2) Identifiable private information.

If your data involves research involving human subjects as per the HHS regulations defined in 46.102, Subpart A, or individually identifiable defined under the Privacy Rule, the project will require IRB review.

ELEMENTS TO BE ADDRESSED IN THE IRB PROTOCOL:

A. **Staff**- A list of all study staff members, a description of their access to the data, and overall data access plan. See the section on Use>Access and Collaboration for additional considerations.

B. **Collaborators**- A list of any collaborators with whom you anticipate sharing data, the method in which you plan to share, as well as those who will be collecting or reporting data, or have access to personally identifiable information about subjects, including individuals or entities to whom you may transfer data for statistical analysis or de-identification.

C. **Data Source**- The source of each of the datasets; your IRB may request information about other approvals associated with the collection and access to research data.
D. Data-
   a. **Type/Form** - If the research data you are collecting involves direct or indirect human subjects identifiers, the subjects’ authorization and informed consent will most likely be required. See HHS 45 CFR 46.102 and the *HHS Guidance on Informed Consent* for additional information, and check with your IRB to ensure that informed consent documents and authorizations permit the types of data sharing you anticipate (e.g., sharing outside of your institution, publication, or posting in publicly accessible repositories). **NOTE**: Some forms of research are eligible for waivers of the authorization requirements. See *HHS 45 CFR 46.116(d)*. Consult your IRB and IT for guidance and approval prior to collecting research data.
      i. **Collection** - Work with your IRB to determine if a *Certificate of Confidentiality* is appropriate for your study. This allows the investigator and others with access to research records to refuse to disclose identifying information on individual participants in civil, criminal, administrative, legislative, or other proceedings at the federal, state, or local level. The NIH grants such certificates but they have important limitations and are not a “magic bullet.”
      ii. **Design** -
         1. If using more than one dataset, consider how linking data impacts identifiability and risk.
         2. If you wish to change the design or conduct of your study after collecting initial data, you must submit modifications to the IRB for review and approval.
         3. If data will be made available through a registry (e.g., dbGaP), or if future open access of data is planned or likely, indicate how data will be released. Ensure that you understand whether any agreements entered into permit or prevent unspecified future uses of data or unanticipated secondary uses, as these also require IRB review and consideration.
   b. **Consent** - Work with your IRB to determine appropriate language for the informed consent form. Make sure to consider the entire research data lifecycle (collection, storage, sharing disposal, etc.) and to include all applicable elements in the consent document.

E. Regulations and Legal Considerations-

Research data is subject to many kinds of regulations, legal constraints, and institutional policies. Your IRB will be the best authority on which regulations and legal considerations you should be aware of for your particular research project. If your research study involves participants or additional sites outside the state in which your institution is located, you will need to consider federal laws. If your research is not confined to the US, or if you have international collaborators, international laws and regulations will need to be considered. International law can have employment, privacy, liability and tax ramifications that are separate from US federal law.  

* Federal Law-

   o The federal policy governing the protections of human subjects is: *Health and Human Services Code of Federal Regulations, HHS 46 CFR part 46*  

---

* See the Additional Resources section below for additional regulatory references
- **Subpart A** or the “Common Rule” - Basic HHS Policy for Protection of Human Research Subjects

- **Subpart B** - Additional Protections for Pregnant Women, Human Fetuses and Neonates Involved in Research

- **Subpart C** - Additional Protections Pertaining to Biomedical and Behavioral Research Involving Prisoners as Subjects

- **Subpart D** - Additional Protections for Children Involved as Subjects in Research

  - Health Insurance Portability and Accountability Act (HIPAA) - HIPAA applies to HIPAA “covered entities” and HIPAA “business associates.” In rare circumstances, a researcher may be acting as a HIPAA covered entity if he or she is providing health care and conducting certain electronic transactions for which the Department of Health and Humans Services has developed a standard such as payment claims (e.g., billing insurance companies). A researcher may also be subject to HIPAA standards if he or she is using or disclosing Protected Health Information (PHI) on behalf of a covered entity. To learn more about PHI, visit the HHS site: [Summary of the HIPAA Privacy Rule](https://www.hhs.gov/). The terms under which such uses or disclosure on behalf of a covered entity are customarily set forth in a business associate agreement.

- **State Law** - You must consider state laws for both the state you are located in/conducting the research in and any other states where the data is being shared/reviewed. Many state privacy laws impact biomedical research. These laws often times apply to health information concerning a specific disease or area of illness deemed to be particularly sensitive, such as mental health, substance abuse, HIV/AIDS, sexually transmitted infections, and developmental disabilities. Such laws may require that extra precautions be taken to protect the privacy of individuals participating in research that could reveal their status with respect to these diseases, disorders, or conditions.

- **International Research (research outside the contential United States)** - Research data protection differs across the globe, with some countries or regions having robust measures and others essentially none (see map below). The IRB can provide guidance on regulatory compliance and also advise on ethical considerations for the protection of research subjects in the absence of regulations.

  - The [ClinRegs](https://www.clinregs.niaid.nih.gov/) website, a service of the National Institute of Allergy and Infectious Diseases (NIAID), provides clinical regulations for countries around the world and has recently undergone several functionality upgrades.

  - DLA Piper’s [Data Protection Laws of the World Handbook](https://www.dlapiperdataprotection.com/#handbook/world-map-section/c1_FR) is a resource to help make it easier to compare regulations in countries around the world.

---

10 [https://clinregs.niaid.nih.gov/](https://clinregs.niaid.nih.gov/)

11 Source: [https://www.dlapiperdataprotection.com/#handbook/world-map-section/c1_FR](https://www.dlapiperdataprotection.com/#handbook/world-map-section/c1_FR)

Accessed 9-13-16
**F. Contractual Obligations:** When collecting, reviewing, storing, and sharing research data, the data being applied may encounter contractual obligations. For example, if you collect research data on an online survey, download the data to a wireless shared server, and then share with others affiliated with the research study, you must check the contractual obligations of each media. This includes the survey software, the wireless network used for the shared server, the shared server itself, etc.

To navigate the contractual obligations and considerations (e.g., conflicts of funding, institutional policy, etc.) that may apply to research data use and production, consult the departments in your institution responsible for negotiating such arrangements; these offices might be called technology transfer office, grants and contracts, clinical trials office or IRB. Your institution's research offices can provide you with contact information.
Once the IRB has reviewed and approved all contracts, agreements have been finalized, and IT has reviewed technology and servers, arrangements should be made for the secure collection or receipt of data.

**Collect/Create**

A. Method
   a. Only the data described in your research proposal and IRB application should be collected/obtained. Collection of additional data than is necessary increases the risk of the data being obtained by an unauthorized user.
   b. Electronic survey data should only be collected through IRB and Information Security approved third parties.

B. Metadata is the description of your data characteristics to help make it easier when identifying and reusing data. It is important to structure your metadata so that it will support long-term discovery and preservations of your research data. Some examples of metadata are:
   a. When and where the data was generated
   b. When the data was last edited
   c. What was used to generate the data

C. Data Keys are a strategic asset when collecting and creating your research data to protect it. A data key is a variable value applied to text to encrypt or decrypt the data; the data can only be opened by the data key.

**Obtain** (existing data)

- If existing data are being obtained for your research and are able to be transferred electronically, please consult with your information security contact. Institutions have established methods for secure transfer of datasets.
- Hard copy media should never be sent via a method without tracking or signature requirements or delivered to a general area like an unstaffed mailroom.
- If duplicate/multiple copies will be shared with collaborators, vendors, or across multiple institutions, ensure that each site that will receive data has appropriate safeguards in place.

**Analysis and Processing of Data**

- Seek out institutional resources (e.g., guidance on encryption, protecting wireless networks, security and malware software, etc.) to create secure research computing environments if your research involves sharing, electronic data transfers, or multi-site analysis of your data with external collaborators.
- Describe your statistical method and analysis plan, including sample size and its scientific rationale.
STORAGE AND RETENTION
The type of access required, and the number of people accessing the data may help determine the manner of storage and the level of security controls. Some data require special management considerations. Data that is sensitive, part of a more than minimal risk study, includes PHI, etc. is subject to several restrictions.

You should think carefully about the following:

- **Data Storage:**
  - **Where:** Where the data will be stored? Will the data be stored with a third party such as the cloud? How will access to the server be secured? Consider the level of access granted to each person (e.g., read, edit/write, delete).
  - **How:** How will the data be stored? On a secure password protected server, behind a firewall? How will remote access to the system be secured?

- **Data Protection:**
  - How will the data be protected? Will the data be protected by password or encryption? Who will be authorized access? What are the access restrictions? Who manages access?

- **Mobile Devices:**
  - If using a mobile computing device (e.g., laptop, smartphone, iPad) or removable media (e.g., USB) for any part of your study, determine how the data-containing PHI will be stored and backed up in case the mobile device is lost or stolen. Consider adding tracking and remote locking and deleting options to these devices.

- **Storage of PHI:**
  - Who will be storing and managing PHI? Will PHI will be stored by a collaborator or vendor-owned platform? What devices and safeguards will be implemented? How will remote access to the system be secured?

- **Storage Contracts/Agreements:**
  - Work with your institution to determine if it is appropriate to contract with a third-party vendor to store and back-up data. Consider the additional vetting and precautions that must be conducted. Your institution may have a list of approved technologies to be used in research; when possible, use technologies already approved by your institution. These vetted technologies are often most recommended by the institution as best choices. Using these devices also avoids the time needed to vet a new technology.

- **Backing-Up Data:**
  - You should always back-up your data in a second location so that if the data is lost due to technical or human error it can still be recovered. How will the data be backed up? How secure is the backup system? Who has access? How long is backup data kept? Will the data key be kept with the backup files? What is the procedure if the backup system becomes obsolete? Who will manage and determine the changes needed?

- **Dataset Size:**
  - Estimate the size of datasets that will be collected and produced, and whether the amount and/or formats of data will change over time. For example, data stored on a cassette tape has become obsolete as that technology is rarely used, and the machines needed to read the tapes may no longer be available.

- **Costs:**
  - Consider the costs of storing data for long periods of time. These costs may include the storage device/media and personnel to safeguard the data. Additional costs may be incurred if the format the data is stored in needs to be changed due to the media no longer being available or a more protective media having been developed. Consult with your IT and grant departments on estimating storage costs. The grant department can help you...
budget for these costs and look into ways to bring in additional funds if needed. IT departments should be informed of anticipated large data sets in order to support back up.

Check with your institution about policies for data storage and retention. Applicable agreements for data storage and retention may include:

- Sponsor’s policy on study data retention
- Cooperative Research and Development Agreements (CRADAs)
- Clinical Trial Agreements
- Data Use Agreements
- Business Associate Agreements
- Disposal contracts
- Lease agreements that provide for return of equipment or media containing research information
While developing your data management plan, consider who will access the data, where the data will reside, and how it will be backed-up may affect data use and access. Consult your institution’s research data protection resources and contacts.

**ACCESS AND COLLABORATION**

Be able to describe with specificity who has access to the data and the manner of access. Adhere to the minimum necessary principle (also known as least privileges), meaning only those with a legitimate research, business, or operational need should have access. For example, a research partner who is only reviewing output and co-authoring a paper might not need access.

The following elements may need to be considered:

- **Staff**
  - The number of people who will collect or work on the data, and whether any are external to your institution; all authorized personnel must be approved study staff trained on the type of data specific to the study.
  - Research staff turnover; when staff comes and goes, how will access be granted/terminated? Who will manage staff access?

- **Access**
  - Whether data will be accessed remotely; if so, how will protections be put into place to protect data from remote locations (e.g., securing servers, securing email, etc.).
  - Whether plans are needed to make the data accessible to other users in the future; if so, what measures will be taken to meet assurances of privacy, security and confidentiality (e.g., if you plan to provide data and images on your website, will the website contain disclaimers, or conditions regarding the use of the data in other publications or products?)

- **Sharing**
  - Whether a data sharing agreement, institutional policy, law or regulation imposes restrictions related to the use or sharing of the data; agreements with industry sponsors can often contain special restrictions on secondary uses.
  - How data sharing will be tracked and documented; what names will be given to the data files? Will there be a key to understand the names and most current version?

**FUTURE USE OR UNANTICIPATED SECONDARY USE OF DATA**

Check with your institution/sponsor/vendor about policies regarding future use/secondary use of data, sponsor, or vendor future use of data.

Applicable agreements for future use or unanticipated secondary use of date may include:

- Data sharing agreements (e.g., waivers, consents, etc.)
- Clinical Trial Agreements (e.g., waivers, consents, etc.)
- Data Use Agreements
- Business Associate Agreements
Replicating data for Publishers, Journals, and Funding Agencies

- Make sure to review the requirements of any journals, publishers, and funding agencies.
- Most often, researchers are required to deposit replication datasets into a public repository. When choosing a public repository we recommend you confirm that it can support the replication of datasets, and that the datasets are easily discoverable for other researchers without having to contact the original research data investigators.
  - A suggested public repository is Dataverse. This is an open source data repository software developed at Harvard University’s Institute for Quantitative Social Science (IQSS) along with many collaborators worldwide. To learn more about replication of data we recommend Gary King’s article, *Replication, Replication.*

Reporting, Publication, and Public Access

You must take a variety of considerations into account as you plan for the reporting and publication of research data.

**Reporting** - All research data that you report must be accurate and not fabricated. Reference all research agreements for reporting requirements. Often funding agencies will have specific requirements for what information should/should not be reported, to whom, when, etc. These important steps must be followed, otherwise you may risk losing rights to your research data and/or be penalized for breaking the agreement terms.

**Publication** - All research data that you report must be accurate and not fabricated. A publisher may request additional information to prove your research data is accurate and your conclusions match your data. Depending on where you are submitting your data to be published, the company may have their own rules for publishing your research data.

**Public Access** - Consider whether your sponsor or journal publisher requires public access to the data and ensure that your agreements, consents, and approvals allow for the specific types of sharing, posting, or other secondary uses required. If you are placing your data on a publicly available website, you may want to consider the impact on individuals if your data is combined with other publicly available sources.

You may also want to consider the impact on individuals if your data were to be obtained through Freedom of Information Act (FOIA) requests. “FOIA is a law that gives you the right to access information from the United States government.”

---

THIRD PARTY (e.g., outside vendor, external researcher, government agency, etc.)
Check with your institution regarding policies about working with a third party source.
Applicable agreements for third party sources may include:
- Service Provider or Data Storage Agreements (e.g., Amazon Cloud)
- Cooperative Research and Development Agreements (CRADAs)
- Memoranda of Understanding (MOU)
- Clinical Trial Agreement
- Consortium Agreement
- Data Use Agreements (DUA)
- Business Associate Agreements (BAA)
- Internal Vendor Risk Assessment

TRACKING/DOCUMENTATION
- Describe in your DMP how data sharing will be tracked/documentated and who will be managing this process.
- When documenting and tracking your research data, some key factors to keep in mind include:
  - the data file format(s) (e.g., jpeg, .doc, sas, etc.)
  - data key (to help with analyzing the data set(s) and being able to recreate the data explained in the journal, etc.)
  - size of the data file(s) (to ensure all the data is in the file)
  - consistent naming conventions.
- Consider how and when to audit and monitor.

OTHER SHARING CONSIDERATIONS:
- Accurate statistical documentation of the data to be shared will make it of the most use for any secondary analysis. Metadata (set of data that explains other data) and data documentation should be kept in a separate file from the source data (primary location the data comes from) to ensure it does not get destroyed with the source data.
Research study close-out is typically thought of as merely an administrative exercise at the end of activity supported by a particular sponsored research award. However, the conclusion of any study involving research data needs to take additional steps to protect the data for future use. The disposition of the data at the end of the project should be documented and become part of the research record.

**RETENTION**
Consider what data should be retained at your institution, lab, in archives, or in other local repositories. Institutional library and data specialists may assist in planning or establishing processes for archiving data, including aiding in the selection of formats and media. Familiarize yourself with publication requirements and institutional guidelines for data retention.

Identifiable data should be held for the minimum amount of time necessary to conduct the research (and meet any access requirements). For example, data that is collected in a corporate sponsored clinical trial might have contractual obligations regarding how long the data must be retained. Data collected in federal or state funded projects, or when using large health care data sets, may require public access to data and therefore may have specific requirements regarding retention, disposal and archiving. It is essential to understand such requirements and proactively plan so that, at the end of a project, data is properly retained, disposed of, shared, or securely archived.

Research record data documentation should include:
- Source of data
- Size of data set
- Number of records
- Variables
- Format of data
- Data key
- Final disposition of data
- Data management personnel

Check with your institution about data archiving policies.
RETURNING -
If data sets were obtained through an agreement with an outside provider or institution, you may be required to return the raw or source data sets at the end of the project. Check with your IT department and review the agreement/contract.

DESTROYING -
If data sets were obtained through an agreement with an outside provider or institution, you may be required to destroy the raw or source data sets at the end of the project. Check with your IT department and review the agreement/contract. Some providers write over the data instead of destroying it, which invokes the risk of special software/programs designed to uncover the original identifiable information. Work with your IT department to ensure that any data that needs to be destroyed is done so correctly, so that the data cannot be recreated.

ARCHIVING DATA -
Being able to separate a raw data set from an analytical data set is an important part of project documentation. If an underlying data set contains personally identifiable information about research subjects and a separate de-identified data set has been created, there may be an obligation to destroy the data set containing identifiers.

HARD COPY MEDIA –
Data may also exist as paper, portable storage media, or removable hard drives. It is just as important to know where these items are stored, and equally important to ensure they are destroyed or returned to the generator/provider. The surest means of destruction is shredding by a bonded vendor. Many vendors have services to shred on site, in the presence of someone who can then attest to the data generator/provider.
III. ADDITIONAL RESOURCES

HIPAA REFERENCES

a. Privacy Rule Requirements: When the researcher is using PHI protected by HIPAA, rules in addition to the Common Rule may apply. HIPAA governs uses and disclosures of PHI by a HIPAA “covered entity” which means a health plan, health care providers that electronically transmit data in a HIPAA transaction, and health care clearinghouses. (45 CFR § 160 and subparts A and E of § 164)

b. Permitted uses and disclosures: Covered entities can use or disclose PHI for research purposes in the following circumstances (per 45 C.F.R. § 164.512(i)):

   o Authorization: The researcher obtained specific written authorization from the research participant. (45 CFR 164.508)

   o Preparatory Research: The researcher asserts that the use or disclosure of PHI is "solely to prepare a research protocol or for similar purposes preparatory to research, that the researcher will not remove any [PHI] from the [CE], and representation that [PHI] for which access is sought is necessary for the research purpose." (45 CFR 164.512(i)(1)(ii) of the Privacy Rule)

   o Documented Approval: An IRB or Privacy Board approves a waiver of research participants’ authorization for use/disclosure of information about them for research. (45 CFR 164.512(i))

   o Research of Decedents’ PHI: The research focuses solely on decedents’ information. (45 CFR164.512(i)(1)(iii))

   o Limited Data Set: The CE and researcher enter into a data use agreement, pursuant to which the CE may disclose only a limited data set to the researcher for research, public health, or health care operations. A limited data set excludes certain direct identifiers of the individual, relatives, employers, and household members. The covered entity providing the data and research must sign a Data Use Agreement that "(1) describes the permitted uses and disclosures of the information and (2) prohibits any attempt to re-identify or contact the individuals." (45 CFR 164.514(e))

   o De-identified: If PHI is de-identified the health information is no longer PHI or subject to the Privacy Rule. 45 CFR 164.514(a)-(c). A CE can always access, use and disclose for research purposes health information that has been de-identified in accordance with 45 CFR 164.502(d), and 164.514(a)-(c) of the Rule without needing to follow the Privacy Rule. Data can be identified either through (1) stripping certain specified elements from the data, or (2) having an expert determine through statistical analysis that there is a “very small” risk that an individual could be identified based on the data.

c. HIPAA Waivers: If your research involves PHI, it may be eligible for a waiver of the requirement of authorization. These waivers must be reviewed by an IRB or Privacy Board.

d. HIPAA Security Rule: The HIPAA Security Rule (45 CFR Part 160 and 164, subparts A and C) establishes national standards to “protect individuals’ electronic personal health information that is created, received, used or maintained by a covered entity.”
**OTHER REGULATORY REFERENCES**

a. **ClinicalTrials.gov:** Federal law requires that certain trials (and their results) be registered on clinical.trials.gov. Determine if you must register your study and find information about how to submit study data [here]. Additional information is available [here].

b. **Computerized Systems Used in Clinical Investigations [FDA Non-Binding Guidance]:** This guidance on 21 CFR Part 11 compliance provides recommendations to sponsors, contract research organizations, data management centers, clinical investigators and institutional review boards regarding the use of computerized systems in clinical investigations.

c. **Genomic Data:** If your study involves genotypic or phenotypic data, you should consider whether your data must be submitted to the, database of Genotypes and Phenotypes (dbGaP). Data submission requirements can be found [here].

   a. For NIH funded research that generates large-scale human or non-human genomic data is subject to the NIH Genomic Data Sharing Policy.

   b. Information about the Genetic Nondiscrimination Information Act (GINA) may apply to research requests for genetic information. Researchers should consider including information about GINA in informed consent documents.

d. **Genetic association studies:** Federally funded genetic association studies may require that data sets be deposited in the GWAS Central repository.

e. **Investigational New Drug Applications and Investigational Device Exemptions:** These regulations (21 CFR Part 312) and Investigational Device Exemptions (21 CFR Part 812) specify data collection and maintenance requirements when conducting a clinical investigation of products unapproved by the FDA.

f. **NIAID Requirements:** Data collected in federal or state funded projects or when using large health care data sets, may require public access to data and therefore may have specific requirements regarding retention, disposal and archive (e.g., NIAID Requirements).

g. **NSF Public Access:** For NSF funded research public access requirements apply.

h. **Publication:** For clinical trial data, publications are requiring clinical trial registrations as a condition for publication. (e.g., the International Committee of Medical Journal Editors registration recommendations).

i. **Public Access:** In February 2013, the Office of Science and Technology Policy issued a broad mandate to the major federal agencies supporting research to develop access plans for all federally funded research that generates large-scale human or non-human genomic data is subject to the NIH Genomic Data Sharing Policy.

---

15 Harvard Catalyst Clinical Trial Registration information http://catalyst.harvard.edu/programs/regulatory/clinical-trial-reg.html
17 NCBI: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.cgi
18 NIH Genomic Data Sharing Policy: http://gds.nih.gov/03policy2.html
19 GINA factsheet for researchers: http://www.genome.gov/Pages/PolicyEthics/GeneticDiscrimination/GINAInfoDoc.pdf
20 GWAS: http://www.gwascentral.org/
22 ICMJE: http://www.icmje.org/about-icmje/faqs/clinical-trials-registration/
funded scientific research [to be] made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data (More Information).23

j. **PubMed:** All NIH-funded investigators must submit to PubMed Central an electronic version of their final peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication. (For more information, please see PubMed)24

k. **Return of Results:** If your research involves human subjects, your IRB may require that you make summary results of your research available to research participants; be sure to check with your IRB regarding such requirements.

---

24 Public Access Resources: https://www.countway.harvard.edu/menuNavigation/libraryServices/nihPublicAccess.html#steps
IV. ATTRIBUTION, SHARING AND ADAPTING

We encourage you to:

- **Request** — Email us and request the materials
- **Share** — Copy, distribute, and transmit the work
- **Adapt** — Adapt the work to suit your needs

Under the following conditions:

- **Attribution:**
  For any reuse or distribution, you must make clear to others the terms of this work. In freely using the materials, we require that you acknowledge Harvard Catalyst as the publisher and that you give appropriate credit to any named individual authors.

- **Feedback:**
  We are interested in gathering information regarding who is using the material and how they are using it. We may contact you by email to solicit information on how you have used the materials or to request collaboration or input on future activities. Please share suggested improvements to the tool with us so that we may learn and improve our materials as well.
CONTACT US

Copies of all materials are freely available. Please send your requests, questions and comments to regulatory@catalyst.harvard.edu and visit the Harvard Catalyst Data Protection Committee page here.

CORE WRITING GROUP

This material is the work of the Harvard Catalyst Data Protection subcommittee of the Regulatory Foundations, Ethics, and Law Program. This work was conducted with support from Harvard Catalyst | The Harvard Clinical and Translational Science Center (National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health Award 8UL1TR000170-05 and financial contributions from Harvard University and its affiliated academic health care centers). The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic health care centers, or the National Institutes of Health.

CONTRIBUTORS

The Harvard Catalyst Data Protection subcommittee of the Regulatory Foundations, Ethics, and Law Program wishes to recognize those members that contributed specific content, templates, or examples included within this guidance document, including: Kris Bolt, Lisa Gable, Pam Richmond, Joanna Myerson, Ian Poynter, Sabune Winkler, and Joe Zurba. The subcommittee also utilized materials provided by Mercè Crosas (Chief Data Science & Technology Officer Harvard IQSS) and Sarah Demb (Sr. Records Manager, Archivist Harvard Libraries).