ORNL/LTR-2023/2739

An Approach to Data Management Planning for Protected Data Projects

Ian Goethert
Katie Knight
Franciel Linares

March 2023

OAK RIDGE
National Laboratory

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

DOCUMENT AVAILABILITY

Reports produced after January 1, 1996, are generally available free via OSTI.GOV.

Website www.osti.gov

Reports produced before January 1, 1996, may be purchased by members of the public from the following source:

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone 703-605-6000 (1-800-553-6847)
TDD 703-487-4639
Fax 703-605-6900
E-mail info@ntis.gov
Website http://classic.ntis.gov/

Reports are available to US Department of Energy (DOE) employees, DOE contractors, Energy Technology Data Exchange representatives, and International Nuclear Information System representatives from the following source:

Office of Scientific and Technical Information
PO Box 62
Oak Ridge, TN 37831
Telephone 865-576-8401
Fax 865-576-5728
E-mail reports@osti.gov
Website https://www.osti.gov/

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/LTR-2023/2739

IT Services Division

AN APPROACH TO DATA MANAGEMENT PLANNING FOR PROTECTED DATA PROJECTS

Ian Goethert
Katie Knight
Franciel Linares

March 2023

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

Table of Contents

CONTENTS ..... *iii*

ABSTRACT *4*

1. Introduction *4*

1.1 Overview 4

1.2 Data management form 2

2. Data Storage and Organization *2*

2.1 Enterprise data warehouse 3

2.2 Institutional Repository 3

2.3 Data Lakes 3

2.4 Databases 3

2.5 Flat Files 3

3. Data Lifecycle and Pipelines *3*

3.1 Data ingestion, Quality Control, and Updating 4

3.1.1 Data Receipt and Responsibilities 4

3.1.2 Data Archiving and Preservation 4

3.1.3 Data Quality Control 4

3.1.4 Data Updates and Versioning 5

3.2 Documentation and metadata 6

3.2.1 Onboarding and Knowledge Transfer 6

3.2.2 Metadata 6

3.3 Backup and disaster recovery 7

4. Conclusion *7*

Appendix A. Data management plan form *2*

AppendiX B. Additional Data Management Information *5*

Appendix C. Outline *6*

ABSTRACT

This report describes what is required in a data management plan for data that needs to be protected in some fashion. Here we provide an overview of what a data manager should consider, including the data ingestion and various extract-transform-load processes, metadata considerations and documentation, to what might need to be accounted for in the event of data loss. In addition, this report includes two appendices: forms that, when filled out, make the user compliant with DOE data management requirements as well as additional requirements for handling protected data at ORNL.

1. Introduction

This document is designed to help you make a data management plan for your project. The purpose of a data management plan is to help make clear your plans for the management, description, analysis, storage, sharing, and preservation plans (or lack thereof) for your data. Specifically, here you will read about considerations as to where data will be stored and backed up, what metadata you are using to describe your data, if there are any sensitivities to the data, and if there are any specific sharing and/or licensing requirements surrounding the data.

The bulk of this document explains what is necessary in drafting a good data management plan. At the end, we have provided two templates (Appendices A and B) that include questions and spaces for answers that, when filled out, will provide you with the necessary document for a clear, concise, and complete data management plan. Appendix A aligns with the data management requirements as established by the Department of Energy, and Appendix B allows users to provide additional information not covered by the DOE requirements. When something is not applicable to your project, simply document that it does not apply and, if necessary, give some justification.

This document is purposely technology-agnostic. Some data management plans may require systems architecture planning and documentation as a separate reference document.

1.1 Overview

For your project, you will need to explain how data is Findable, Accessible, Interoperable, and Reusable. This lines data up with the FAIR principles (more information about this can be found on https://www.go-fair.org/fair-principles/).

1. Findable: How data is described?

    a. What metadata are you and your team using?

    b. Do you use a controlled vocabulary for your data elements (e.g., temperature, medications, procedures), or is it ad-hoc and unstandardized?

2. Accessible: How is data accessed and potentially shared?

    a. Do data use an open protocol for authentication, authorization, and retrieval?

    b. Does data include unique identifiers?

    c. What are the storage speed requirements for access? (e.g., hot, warm, or cold storage)

3. Interoperable: What type of language are you using to represent the data

    a. Do you reference other data sets, internally or externally?

4. Reusable: How data is stored and preserved

    a. Where is the data stored and how is it maintained? Are there any costs involved, and/or commercial or open-source software in use?

    b. How often is the data migrated and updated, if at all?

    c. What is the data format?

    d. Are regular backups made of the data? Where and how is that process documented and, and who or what is responsible for this?

    e. Are you using any tools to interpret and/or read your data (special code and/or software)

5. In addition, you must document sponsor's and researcher's requirements to implement these principles. Documentation may include:

    a. Sponsor-required vocabularies (e.g., SNOMED, OMOP, ICD codes, etc.)

    b. Researcher access restrictions or other IRB-related information that may affect how individuals may access and reuse the data

    c. Licensing that needs to be considered, either with the data itself or with the software/code used to read/interpret the data (e.g., SAS, Tableau, data use agreements)

    d. Regulatory requirements such as HIPAA for PHI/PII, Official Use Only, etc.

1.2 DATA MANAGEMENT FORM

At the end of this document are two forms for you to use (Appendix A and B). All the questions on these forms will help you answer what is outlined in this report. Please feel free to use these forms to clarify your data management requirements and share these with others.

2. DATA STORAGE AND ORGANIZATION

It is important to decide how and where your data will be kept. It is also important to consider any storage requirements, including all intermediate forms such as user-managed "scratch" storage for data analysis or other "sandbox"-type work. Does it use a Relational Database Management System (RDBMS) or some other kind of platform? Is it a directory of Excel spreadsheets? The following lists some examples of places where data might be stored, though it is not exhaustive.

2.1 ENTERPRISE DATA WAREHOUSE

This is a database that primarily stores large amounts of business data. These are typically used for business intelligence, where data is loaded into some type of form suitable for analytics as well as predictive modeling. Data is usually operational and can hold both structured, unstructured, and be labeled.

2.2 INSTITUTIONAL REPOSITORY

Institutional repositories are digital archives for preserving open-access, high-value datasets and possibly other documents (publications, images, code) for long-term use and access. To ensure discoverability, they employ structured metadata, controlled vocabularies, and multi-step workflows to ensure the files and metadata are understandable to human and machines. Datasets are usually structured and labeled. Additionally, the structured descriptive metadata is harvestable via a standard protocol and available for ingest into a data catalog.

2.3 DATA LAKES

A concept closely tied with storage solutions like Apache Hadoop, a data lake is a large data repository built to assist with capturing, refining, and archiving raw, unstructured, semi-structured, or multi-structured data. This large repository is then available for access and analysis by multiple users. Data lakes may accept a variety of data formats and schemas, such as text, CSV, JSON, Parquet, RDF triples or relational data: anything may be stored. Moreover, data lakes can handle high volumes of input data since ETL is not necessary on ingest. Most significantly, data lakes store unprocessed data, and therefore are markedly different from data warehouses, which store structured, pre-processed data.

2.4 DATABASES

Relational databases store data in relations, viewed as tables; relations are composed of records and fields, and each record in the table is associated with an identifier and a field that contains some type of unique value.

Non-relational databases, like NoSQL data stores, are useful when handling large volumes of data, as they are designed to scale across many servers, allowing for large read/write operations in a short time. Still, these systems make use of attribute-values, albeit by using different data structures.

2.5 FLAT FILES

Data that is stored in a directory, or “folder” (or series of directories or “folders”). Folders can be local, remote, or cloud based. Do you need to account for directory structure changing or evolving over time?

3. DATA LIFECYCLE AND PIPELINES

Data pipelines may include the ingestion process, quality control measures, and updates to the data. How data will be archived/preserved, shared and used or modified, and finally disposed of need to be considered.

3.1 DATA INGESTION, QUALITY CONTROL, AND UPDATING

Consider how you wish to import data into your chosen environment, how you plan to preserve and archive your data, as well as what you will be using to document these preservation requirements and reproducibility for later publications or other assertions of the work you are doing. Consider what extra safeguards are needed for gold standard data.

3.1.1 Data Receipt and Responsibilities

Understanding the process by which data is brought into the project is key: researchers, engineers, and other data stakeholders need to understand this process to help.

3.1.1.1 Data Interfaces

Consider how data is transferred or obtained in your environment. This could be a onetime ingress, nightly ETL (Extract/Transform/Load) process, or streaming data from a sensor. Interfaces may be needed to ensure quality or uniformity of data. Transforming data may require additional quality control or validation-type steps.

3.1.1.2 Data Ingress

What are the procedures for ingress of new data? Who sees new data before being released to research/ingestion pipelines? This is important in the even that a sponsor has sent incorrect and potentially sensitive data. What is the procedure for rejection of new data? Examples are data that does not meet quality standards, corrupt data, improperly formatted data.

3.1.2 Data Archiving and Preservation

Planning for data preservation is a multi-step process. Some things to consider when you are planning out how data will be stored for later use:

1. What format is the data in (spreadsheet, HDFS, PDF, image files)
2. Does the data need to be stored in a specific way (e.g., will other users need access to a series of nested directories)
3. How often is data backed up?
4. When can data be deleted, if ever?
5. What are the storage costs after the project has ended?
6. Who is the responsible owner of the data?
7. Is there a disposition plan for data? When it is OK for data to be deleted/destroyed?

3.1.3 Data Quality Control

What are the processes for assuring validity of data for all the steps in your data pipeline? For instance, do you distinguish between unprocessed data, data that is in a state of QA, and production-level data? The below graphic provides such an example.

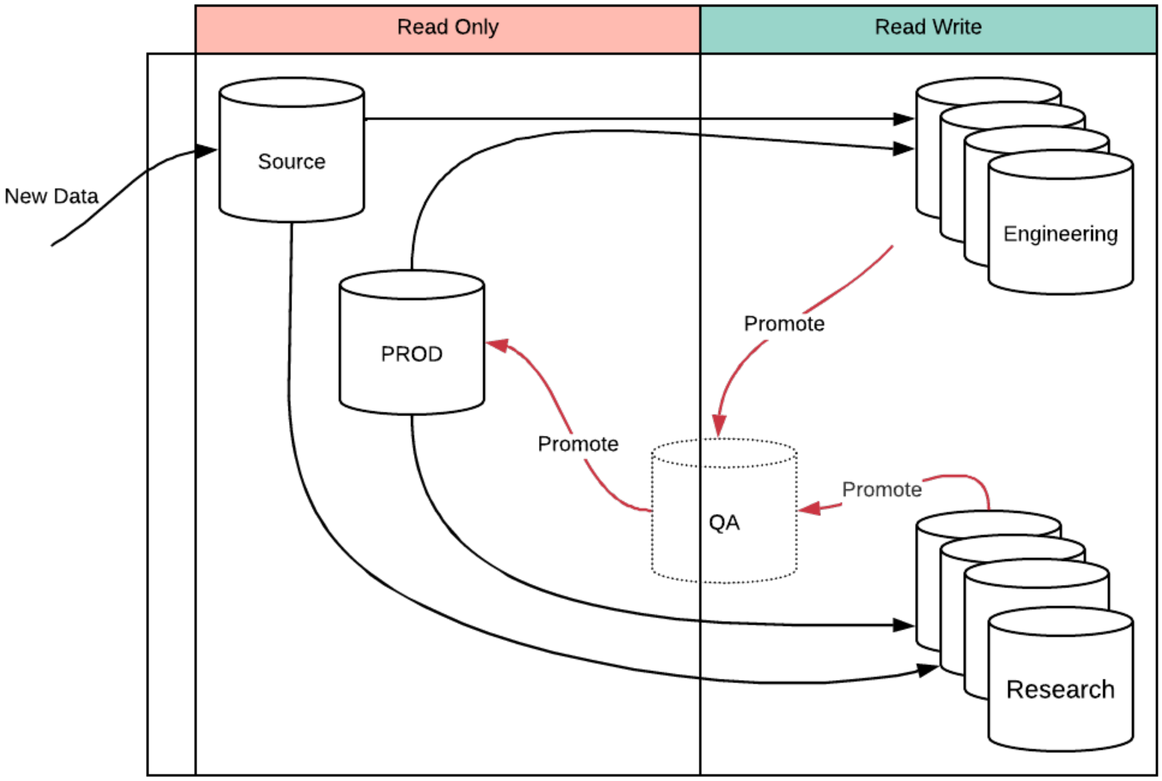

Figure 1: Example Data Staging for Ingestion, Quality Control, and Production of Engineered and Research-derived Data Products

In the process illustrated in Figure 1, new data is ingested from a given source (a stream, a hard drive, a database), and either made directly accessible to researchers or is sent to an engineering process (some type of extract-transform-load process). During either, data may be considered in a state of "QA", where either engineers or researchers are modifying the data in some fashion. Finally, data is promoted to a Production stage after either researchers deem data in a stable state (e.g., data that was used to train a machine learning model and that needs to be shared along with the model) or engineers have finished with the designated ETL process (e.g., data that was taken from several sources, blended, and then determined appropriate to share in some fashion).

At every stage in this process, it is helpful to consider how this is documented. Is this process made clear to your research team and others who may want to reuse your data? For any of these processes, are you relying on specific tools and/or metadata vocabularies that are easily interpreted by others?

3.1.4 Data Updates and Versioning

What is the process for bringing in new data to the project? Considerations about how to ensure that data does not arrive corrupted or, for entire updates, does not alter legacy data in any way should be made.

Consider if data standards or equipment requirements necessitate migrating or transforming the data to a new format. If data can't be migrated, can data be normalized between formats to ensure compatibility?

How is data versioned within an iterative development process or in the lifecycle of the data? As seen in Figure 1, a common approach is to version data when it's promoted to production or promoted to QA. You may choose to persist versions and document identifiers (e.g., DOIs or some other identifier for a specific dataset) in order to keep track of versions.

3.2 DOCUMENTATION AND METADATA

3.2.1 Onboarding and Knowledge Transfer

Consider the overall scope of the data in your project. What are the master data requirements at hand? For instance, if you intend for your data to be used by others, how will this happen (will they need access to certain kinds of software? Will certain caveats about the data need to be shared?)

Consider also how you are documenting data storage and versioning. If data is stored in an RDBMS, is the lookup or dimension data (data that describes other data) documented somewhere that is accessible and easily understandable by others outside of the project?

Finally, are there any data crosswalks? If so, ensure that this is documented as well, and consider how this might be interpreted by others.

3.2.2 Metadata

How the data is described is of utmost importance: this will determine if and how someone else can find and reuse the data, as well as how easy it will be to blend this data with other data, if desired.

Some considerations regarding the collection and storage of metadata:

1. Will metadata be ingested automatically, manually, or a combination of both?
2. How will metadata be stored and what schema are you using (some controlled vocabulary or not).
3. Will others be able to read and understand the metadata later?
4. Does the metadata make measures and concepts explicit to the end user (for instance, if metadata includes a field called "temperature", is it clear what units temperature is measured in?)

3.2.2.1 Data terminology

How will you develop a terminology for describing data that is clear and understood to all who use the data? Where will that be documented? This can include terms that describe both state and contents such as raw data (which might be defined as data that has not been touched by engineering or other processes), clean data (data that has been subject to some type of transformation process to make it easier for others to use), and "gold standard data", or data that is considered exemplary for some specific process or use. These terms are examples, and these definitions are not standard definitions.

Terminology should also be developed so that, for all who use the data, it is clear and easily understood. For instance, consider how you might define the following, as it relates to your project:

- Sensitive data
- Licensed data
- Public data
- Raw data

- Clean data
- Gold standard data

Finally, consider what the requirements are for separating these different data classifications of data (e.g., licensing requirements: is it clear what is “open” and what is not?)

3.3 BACKUP AND DISASTER RECOVERY

As a final consideration, it is important to account for how data may need to be backed up and, if ever lost, recovered or accounted for if data is irretrievably lost. Some questions to consider as you write your data management plan:

1. What is the project’s tolerance for lost or damaged data?
2. If data is lost, how long does it take to re-create or restore data and how does this impact timelines?
3. What are plans to mitigate lost or damaged data; that is, what is the recovery procedure for data that gets lost or damaged?

4. CONCLUSION

These outlined considerations are by no means exhaustive for every single approach to data management. However, considering each of these elements individually will ensure that an appropriate level of due diligence was exercised by the data manager or management team.

APPENDICES

APPENDIX A. DATA MANAGEMENT PLAN FORM

The following has been created in accordance with the requirements listed by the DOE here: https://www.energy.gov/datamanagement/doe-policy-digital-research-data-management#Requirements%20and%20Guidance
[site last accessed on 2/9/2023]

Project Name:

Principal Investigator Name:

Principal Investigator Email:

Data Types and Sources

Please provide a broad, brief description of the data to be generated or used through the course of the proposed research

Of the data described above, which data may be used to validate your research findings? *Validation could be accomplished by reproducing the original experiment or analyses, comparing and contrasting the results against those of a new experiment or analyses, or by some other means*.

Sharing and Preservation

□ YES, data will be shared without restrictions.
□ YES, data will be shared but with certain restrictions.
□ NO, data will not be shared.

If YES, please describe the means for sharing and any additional applicable contact information. If there are restrictions, please add the rationale for any restrictions on who may access the data and under what conditions.

If YES, please state the *minimum* length of time the data will be available (e.g., 1 month, 5 years, etc.):

If YES, please describe any anticipated delay to data access after research findings are published (if there is no anticipated delay, write "no anticipated delay").

If NO, please provide a brief explanation of why and how results could be validated if data are not shared.

□ YES, special software is necessary to access OR interpret this data

Name of software:

□ NO, special software is not necessary to access OR interpret this data

Please describe or provide a link to any applicable policies, provisions, and licenses for re-use and re-distribution, and for the production of data derivatives. If not applicable, write "not applicable."

Please indicate how data and any derived data products should be cited:

Please list any additional resources and capabilities (equipment, connections, systems, expertise, etc.) requested in the research proposal that are needed to meet the stated goals for sharing and preservation (This could reference the relevant section of the associated research proposal and budget request.) If not applicable, write "not applicable."

□ YES, data will be preserved after the project funding ends.

If YES, please state where (e.g., ORNL Institutional Data Repository, ORNL DAAC, etc.)

□ NO, data will not be preserved after the project funding ends.

If applicable, please describe any other future decision points regarding the management of the research data, including any plans to re-evaluate the costs and benefits of data sharing and preservation. If not applicable, write "not applicable."

Protection

A statement of plans, where appropriate and necessary, to protect confidentiality, personal privacy, personally identifiable information, and U.S. national, homeland, and economic security; recognize proprietary interests, business confidential information, and intellectual property rights; and avoid significant negative impact on innovation and U.S. competitiveness.

Rationale

Please describe what impact this data will have within the immediate field and/or in other fields, and any broader societal impact.

APPENDIX B. ADDITIONAL DATA MANAGEMENT INFORMATION

This form provides additional information about how data will be managed, but in a separate form than what is required by the Department of Energy (DOE).

Data Ingestion

Describe briefly how you send data to your designated project space (e.g., streaming data, mailing a hard drive, nightly automated updates to an on-premises database, etc.)

□ YES, my data ingestion process includes a customized process (specialized code or software, in-the-loop ETL processes prior to arrival on premises)

□ NO, my data ingestion process does not include any customized process.

If YES, please explain what this is and how others will have access to it.

Data Quality Control and Versioning

Briefly describe any special processes or procedures that you use or will use for creating data products (e.g., the merging of different datasets and how these data products will be saved, shared, and versioned).

Data Backups and Recovery

Briefly describe how data loss and/or recovery will be managed (or state if this is not a necessary part of the plan).

APPENDIX C. OUTLINE

- A. Overview and Requirements
  - a. Project Needs - What are your team's needs related to data? Use the FAIR data principals: Findability/Accessibility/Interoperability/Reusability
  - b. Sponsor - What requirements has the sponsor laid out for the project?
  - c. Compliance - Are there any compliance or regulatory requirements such as HIPAA (PII/PHI) that must be addressed?
  - d. Other requirements, such as standard vocabularies, IRB, licensing, regulatory, and separation?
    - i. Can data X and data Y be stored together? Can they be blended?
- B. Storage and Organization
  - a. Data sizes and associated requirements
    - i. Disk space and growth planning
    - ii. Disk performance requirements - Hot vs Warm vs Cold
  - b. How will the data be stored
    - i. Managed or Enterprise data warehouse
    - ii. Institutional Repositories
    - iii. Data Lakes
    - iv. Databases
    - v. Flat files
- C. Lifecycle and Pipelines
  - a. Data receipt and responsibilities
    - i. Procedures for ingress of new data
    - ii. Who sees new data before being released to research/ingestion pipelines?
    - iii. Rejection procedure for new data
  - b. Interfaces for obtaining data, ETL, and additional transformations
  - c. Archiving and Preservation
    - i. Reproducibility for publications or assertions
    - ii. Protecting gold standard data
  - d. QA - Processes for assuring validity of data for all steps in the pipeline (DEV/QA/Production Systems)
  - e. Updates and Versions
    - i. Change processes for new data formats, data schemas, new types, etc
    - ii. How should data be versioned withing the lifecycle and/or QA process?
  - f. Documentation and Metadata
    - i. Internal team requirements for onboarding and knowledge transfer, master data management, and crosswalks
    - ii. Automated and manual gathering of metadata
    - iii. Accessible and understandable
    - iv. Data Terminology for describing data that is clear and understood to all who use the data, for example:
      - . Raw Output vs Clean Output vs Gold Output
      - a. Sensitive Data vs Licensed Data vs Public Data
- D. Backup and Disaster Recovery
  - . Project tolerance for lost or damaged data
  - a. Plans to mitigate lost or damaged data
  - b. Recovery procedure for lost or damaged data