VANDERBILT UNIVERSITY

January, 2023

Guidance to VU Investigators: New NIH Data Management and Sharing Plan Requirements

Outline of Slides

- Overview of New NIH Data Management and Sharing Policy
- How to Share your Data
- How to find and use a generalist repository, if necessary
- How to write specific elements of the Data Management and Sharing Plan

Policy Overview

Starting January 25, 2023, all NIH research grant proposals, both new and competing renewals, will need to include a Data Management and Sharing plan, have costs accounted for in the budget, and have progress shared with NIH in annual reports.

https://sharing.nih.gov/data-management-and-sharing-policy

The policy is driven by the F.A.I.R. principles: findable, accessible, interoperable, and reusable.

Background

1
All data must be managed and shared, with certain exceptions.

2
Projects with budgets >$500,000/year already require data management and sharing, although with different policy details.

3
NSF already requires Data Management Plans at time of proposal, although with different policy details.

4
New policy applies to all NIH research grants and requires more detail than previous policies. Does NOT apply to training grants, fellowships, infrastructure, or instrument grants.

NIH Definition of Scientific Data for the Data Management and Sharing (DMS) Plans:

> **Recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications**

NIH DMS FORM: 6 “Elements”

https://grants.nih.gov/sites/default/files/DMS-Plan-blank-format-page.pdf

Element 1: Data Type

A. **Types and amount of scientific data expected to be generated in the project:**
Summarize the types and estimated amount of scientific data expected to be generated in the project.

B. **Scientific data that will be preserved and shared, and the rationale for doing so:**
Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.

C. **Metadata, other relevant data, and associated documentation:**
Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

Element 2: Related Tools, Software and/or Code:
State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

Element 3: Standards:
State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources, and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist.

Element 4: Data Preservation, Access, and Associated Timelines

A. **Repository where scientific data and metadata will be archived:**
Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived; see Selecting a Data Repository).

B. **How scientific data will be findable and identifiable:**
Describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

DRAFT

Element 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse of scientific data:

NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See Frequently Asked Questions for examples of justifiable reasons for limiting sharing of data.

B. Whether access to scientific data will be controlled:

State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

C. Protections for privacy, rights, and confidentiality of human research participants:

If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

Element 6: Oversight of Data Management and Sharing:

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

DRAFT

Included in the policy

All Scientific Data must be managed (see NIH definition of “data”). Researcher decides what constitutes data and how to maximize sharing data and justification for what is and isn’t shared.

Data that led to null findings

Data sets of all sizes

Data generated with SBIR support (a 20-year delay is allowable)

Data for which there is no known repository

Qualitative Data, unless there are justifiable limitations to sharing (For example: field reports and ethnographic writings that contextualize and interpret rich participant-observation data.)

Data that requires a Data Use Agreement for sharing (in other words, data still has to be shared, but with appropriate restrictions on public access).

Excluded from the policy

- Non-research grants: training, fellowship, conferences, and infrastructure,
- Data not necessary for or of sufficient quality to validate and replicate research findings,
- Laboratory notebooks,
- Preliminary analyses,
- Completed case report forms,
- Drafts of scientific papers,
- Plans for future research,
- Peer reviews,
- Communications with colleagues, or
- Physical objects, (e.g., laboratory specimens)

More on required sharing

Data must be shared at time of first publication or at the end of the project period, whichever comes first. . But for competitive renewals, there may be modest flexibility regarding how to handle unpublished data at end funding cycle (12/20/2022 FAQ at NIH DMS Web site)

Unpublished data must be deposited and reported at end of project period even if they will end up in a published paper. But see red font note above.

Data must be reported irrespective of the whether a competitive renewal application for the relevant NIH grant is being prepared or has been approved. But see red font note above.

If the grant cycle includes a period of no-cost extension (NCE), the deadline for posting unpublished data is the end date of the NCE. But see red font note above.

Pre-Prints are not considered “papers” under NIH DMS policy. They may, however, be used as data repositories.

Also Excluded from sharing

- Software and code
- Data already public
- Data for which there are justifiable ethical, legal, and technical reasons for limiting and/or delaying sharing. This includes but not limited to:
  - Informed consent will not permit or limits scope of sharing or use
  - Privacy or safety of research participants would be compromised and available protections insufficient
  - Explicit federal, state, local, or Tribal law, regulation, or policy prohibits disclosure
  - Restrictions imposed by existing or anticipated agreements with other parties

How are plans submitted and reviewed?

- Plans recommended, but not required, to be 2 pages or less
- DMS Plan format page will be added to list of Format Pages and incorporated into FORMS-H application instructions
- Submit plans to “Other Plans” as a single pdf attachment.
- No hyperlinks can be listed in plan
- NIH program staff will review determine if plan is acceptable or unacceptable
- Unacceptable plans are returned to the PI during the Just-in-Time period for revision and resubmission by the PI and re-review by the NIH PO. Funding of grants may be delayed if an acceptable DMS plan is not submitted and approved in a timely manner.
- Peer reviewers only consider if budget is reasonable
- Calling your program officer (if known) prior to grant submission to discuss your DMS plan may be wise.

Budget for Plans

- Cost specific to the project are allowable
  - Curating data, developing supporting documentation, metadata, and formatting for repository deposition
  - Preserving/sharing data through repositories (data deposition fees)
  - Local data management considerations

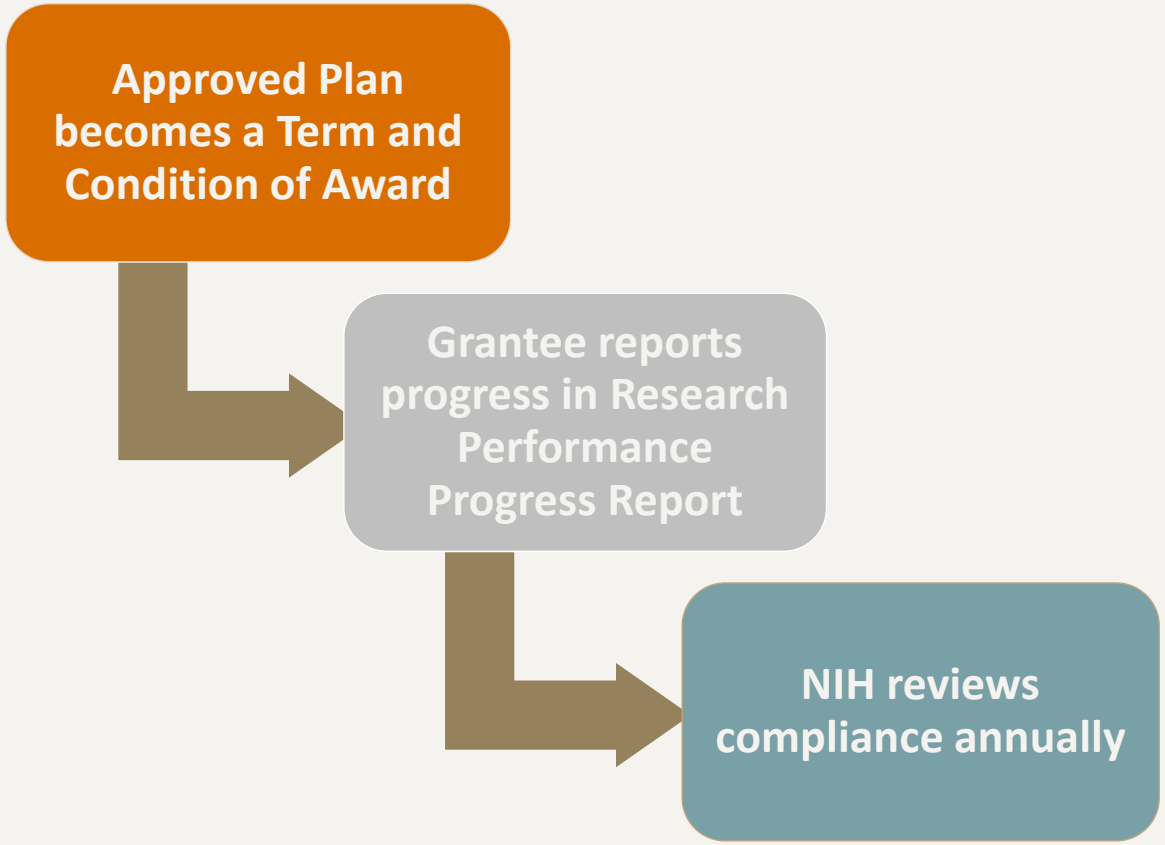
- Cost must be incurred during the performance period (you can only spend grant funds while the grant is active or in the no-cost extension period).

- Do not include general infrastructure costs not associated with the specific project or costs associated with gaining access to research data.

- Will need to write a budget justification for non-modular grants.

Compliance

- NIH reviews the progress of plans as presented by PI in annual progress reports (RPPR). NIH is revising the RPPR templates to include DMS reporting.
- NIH Program Managers are still learning how to implement the policy
- Section 6 of the plan asks how plans will be monitored at the institution.
- Will there be DMS compliance audits by NIH?



Courtesy of NIH

NIH Resources

Website: NIH Scientific Data Sharing

FAQs: DMS Policy FAQs

Draft Format Page

NIH-supported repositories

Human Subjects and Genomics info

Email Box: Sharing@nih.gov

NIH Grant Program Managers

Vanderbilt Resources

Library – data curation and repositories

Research IT – local storage for data and consulting for their research data needs

SPA – Data Use Agreements

Research Integrity & Compliance – interpreting the policy and best practices

Human Research Protections Program – address data sharing in the informed consent

NIH Data Management and Sharing Plans: Summary

1. Applies only to grants for NIH proposals submitted Jan. 25 or after.

2. All NIH research grant proposals (but not training grants and fellowship proposals) will have to include a DMS form page that addresses 13 different points.

3. NIH Program Officers will review the plans. An unsatisfactory plan will need to be addressed during the Just-in-Time period for the grant to get funded.

4. PI reporting on compliance with DMS plans will become part of the yearly RPPR Process.

5. Data and metadata that need to be deposited into a publicly-accessible repository:
    - Data related to published papers but not included in the paper or the supplementary information (SI) section.
    - Unpublished data that meets the NIH's definition of scientific data at the end of every grant cycle.

6. A VU-based DMS plan manager needs to be named in the DMS plan.

7. You can budget for any anticipated cost in your proposal.

8. Be careful to propose a plan that you truly intend to implement. You will be held accountable for following the plan you propose.

Complying With DMS Data Deposition Requirements

Two Data Classes Will Need to be Deposited to Satisfy DMS Policy Requirements

1. Data for published papers
   satisfied by the data in publications or linked supporting/supplementary materials. Other data can be deposited in specialized repositories (e.g., protein databank for 3D structural coordinates) or generalist repositories.

2. Data not published (please see NIH definition of "data")
   by the end of the grant cycle (including any period of "no-cost extension")

   Option A: Use online pre-print (e.g., BioRxiv) and linked supporting/supplementary materials.

   Option B: Deposit in specialized repositories (e.g., protein databank for 3D structural coordinates) or generalist repositories.

   Option C: (For data generated by a graduate student). Include data in their Ph.D. dissertation (including appendix), which eventually will be assigned a DOI.

   Option D: Deposit in generalist repositories. (see below)

   (The above options are not mutually exclusive. Often, multiple options will need to be used).

If you are likely to eventually submit a competing renewal application, you may try proposing that all data from the current grant period will be either published or deposited within, say, 1-2 years of the end of the grant period (including any NCE). If your program officer finds this unacceptable s/he will let you know and you can revise!

For Data not in Publications or Their Associated Supporting or Supplementary Data Sections: Where to Share and for How Long?

- When possible, use established data type-specific repositories (PDB, Genbank, etc.). NIH supports many Scientific Data Repositories: https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/repositories-for-sharing-scientific-data
- Some grant programs, Institutes, Offices, or Funding Opportuniuty Announcements (FOAs) may indicate specific data repositories to be used – follow any special instructions.
- Prioritize the use of discipline or data-type specific repositories to make it easy for people in your field to find
- Otherwise, use “generalist” repositories
- Data needs to be stored and made available for the full duration of the grant (including possible future renewal plus 3 years.

Desirable Repository Characteristics

- ✓ Unique Persistent Identifiers
- ✓ Long-Term Sustainability
- ✓ Metadata
- ✓ Curation and Quality Assurance
- ✓ Includes user Dashboard
- ✓ Free and Easy Access
- ✓ Broad and Measured Reuse
- ✓ Clear Use Guidance
- ✓ Security and Integrity
- ✓ Confidential
- ✓ Common Format
- ✓ Provenance
- ✓ Retention Policy

Guidance on what **generalist data repositories** we can refer our faculty to

Recommended Generalist Repositories

- Nine “generalist” repositories mentioned on the NIH website
- Five meet the basic requirements of the NIH and are suitable to recommend
  - Harvard Dataverse, Dryad, Figshare, Open Science Framework (OSF), and Zenodo
- Three particularly recommended because of their ease of use
  - Harvard Dataverse, Figshare, and OSF
- OSF is distinct in that it is both easy to use and provides for planning throughout the data life cycle.

https://heardlibrary.github.io/digital-scholarship/manage/repository/

Repositories Recommended for Ease of Use

Harvard Dataverse

Pros

- Free
- Vanderbilt Institutional login is available
- Each dataset can associate with personal identifiers (a persistent digital identifier (ex. ORCID iD) that you own and control)
- Very detailed User Guide
- Great dashboard for tracking your datasets
- Retain complete control over your data. Create customizable access to your data (metadata is always public no matter what level of restriction you set for your data. People can search and find your data. When restricted, they can request access, and you can grant or deny them access)
- Automatically generates a data citation to use in publications

Cons

- File size limit $\leq$ 2.5 GB
- Total storage limit of 1 TB/user

Figshare

Pros

- Dedicated dashboard
- Can create projects and collections
- Has extensive tutorials and How-To articles for help using the service
- Integration with lab archives (electronic lab notebook)
- Has Figshare+ for larger datasets >20 GB (5 TB file size limit)

Cons

- File size limit 20 GB
- Free dataset limit of 20 GB

Figshare+ Pricing

- One-time cost, a data publishing charge, based on the amount of storage needed
- Includes deposit support and dataset review
- Pricing is based on storage with a 'up to 100GB' tier and then tiered in increments of 250GB ($585 per 250GB) + a review fee ($160) per deposit.

100GB	250GB	500GB	750GB	1TB	1.25TB
$395	$745	$1,330	$1,915	$2,500	$3,085
1.5TB	**1.75TB**	**2TB**	**3TB**	**5TB**	**5TB+**
$3,670	$4,255	$4,840	$7,180	$11,860	get in touch

A disadvantage of Figshare is that you need to make a purchase for each new dataset to be deposited.

Recommended All-In-One Project Management System

OSF (Open Science Framework)

Pros

- Curation of data from the start
- Integration with preprint services
- Compatible with some reference management services (Mendeley, Zotero)
- Can add-on storage applications to avoid storage caps (dropbox, figshare, etc.)

Cons

- File size limit < 5 GB
- The private data storage limit is 5 GB, and the Public storage limit of 50 GB
- Necessary to have add-on storage applications due to total dataset limits of 50 GB for public storage and 5 GB for private storage

COS calculates one-time storage fees up to 1 TB, with additional storage tiers available upon request.

Determine the appropriate amount of OSF storage for your project or proposal using the table below. Please contact our team at billing@cos.io to request a quote should you anticipate your project will exceed 1 TB of data.

Storage Amount	Tier Cost
0 - 100 GB	$500
101 - 250 GB	$1,000
251 - 500 GB	$2,000
501 - 750 GB	$3,000
751 GB - 1 TB	$4,000
Above 1 TB	Request cost

Other Recommended Repositories

Dryad

Pros

- Quality control and assistance (internal curators)
- Linked to ORCiD (must log in with these credentials)
- Requires the addition of a “README.md” file for each dataset

Cons

- The interface isn’t intuitive
- Not free (fee covers curation and preservation)
- Data publishing charge of $120/dataset ≤ 50 GB, $50 per additional 10 GB

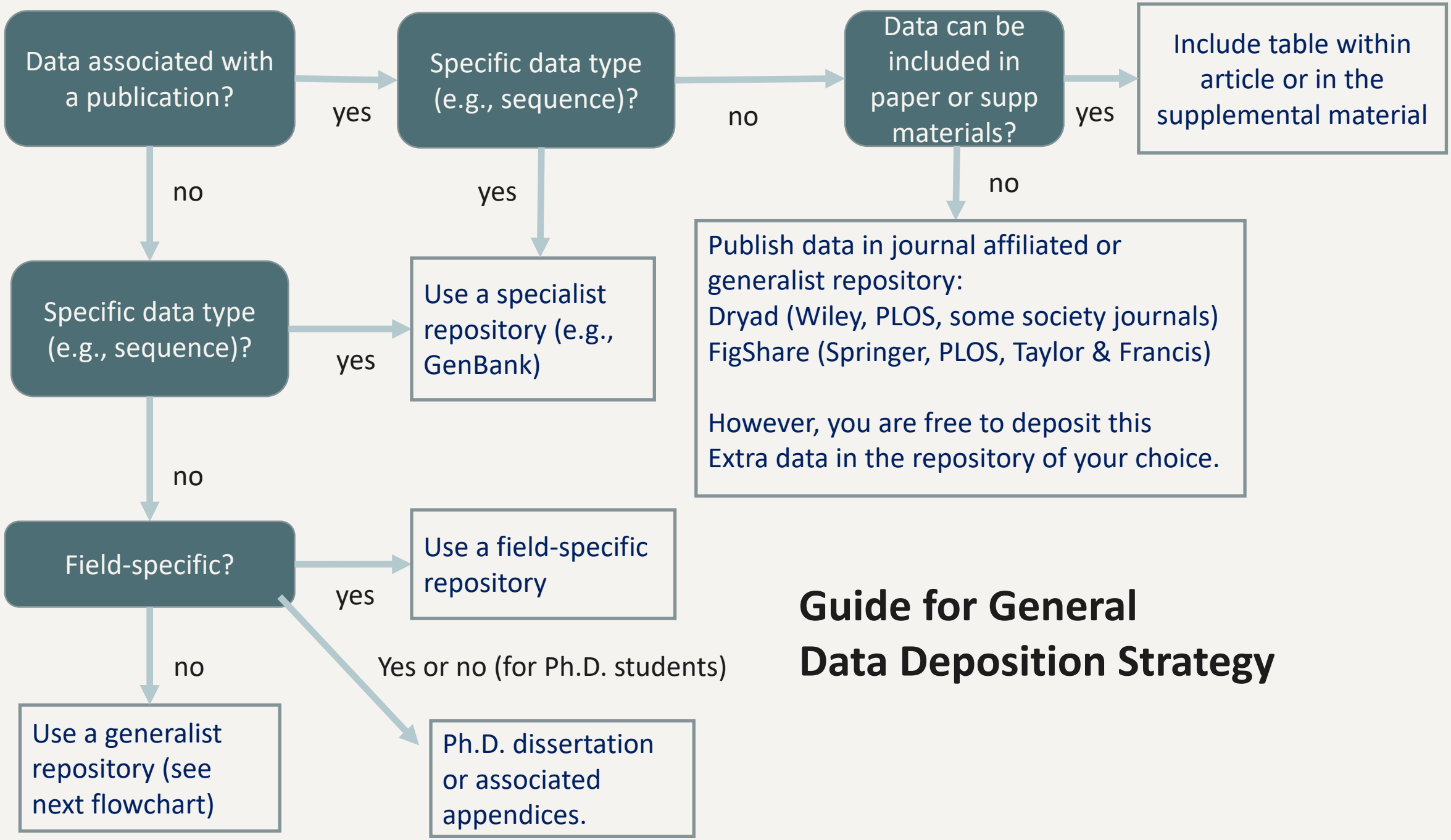
Zenodo

Pros

- Free data-hosting initiative associated with the European Organization for Nuclear Research (CERN)
- Lots of communities to choose from to upload your data into. Some are OA journals and discipline-specific curated communities as well.
- Will notify funding agencies if you enter the grant number

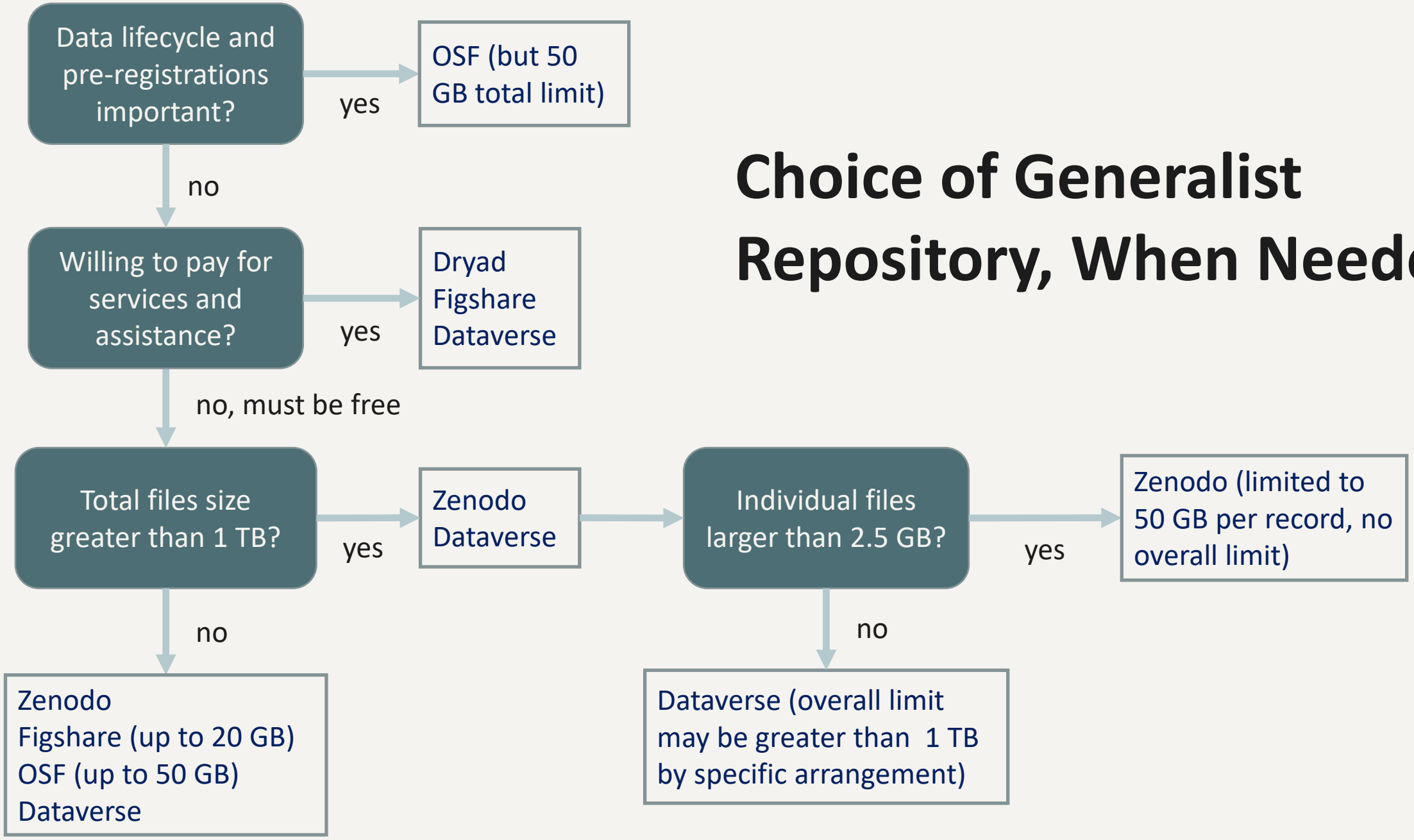
Cons

- No obvious dashboard for your data
- File size limit 50 GB (no upper limit)



Guide for General Data Deposition Strategy

Choice of Generalist Repository, When Needed



Common Metadata Required to Support Deposited Data

Minimum Metadata Required

- **Title**: a succinct summary of both the data and study or focus (usually 8-10 words that adequately describe the content of the dataset)
- **Author(s)**: Name, email, institutional affiliation of the main researcher
- **Abstract**: Brief summary of the structure and concepts of the dataset (should focus on the information relevant to the data itself)
- **Research domain**: primary research domains or drawn from OECD Fields of Science and Technology classification
- **Journal Name** (if associated with a manuscript)

Metadata Recommended

- **Funding information**: funder, grant number
- **Keyword(s)**: minimum of 5 descriptive words to help with data discovery (more is better)
- **Methods**: special chemicals or specific antibodies/reagents necessary to replicate dataset
- **Usage Notes**: programs and/or software required to open the files
- **Related Works**: resources associated with the data (publications, related datasets, etc.)

Most data depositions require an associated README.txt file to explain the data in the directory.

- Should be no passcode restrictions
- In English
- No personal or sensitive data
- Example README.txt file to the right.
- Some repositories refer you to README file examples

```
This readme file was generated on [YYYY-MM-DD] by [NAME]
<help text in angle brackets should be deleted before finalizing your document>
<[text in square brackets should be changed for your specific dataset]>


GENERAL INFORMATION

Title of Dataset:

<provide at least two contacts>
Author/Principal Investigator Information
Name:
ORCID:
Institution:
Address:
Email:

Author/Associate or Co-investigator Information
Name:
ORCID:
Institution:
Address:
Email:

Author/Alternate Contact Information
Name:
ORCID:
Institution:
Address:
Email:

Date of data collection: <provide single date, range, or approximate date; suggested format YYYY-MM-DD>

Geographic location of data collection: <provide latitude, longiute, or city/region, State, Country>

Information about funding sources that supported the collection of the data:
```

Additional information from sample Readme.txt file

```
SHARING/ACCESS INFORMATION

Licenses/restrictions placed on the data:

Links to publications that cite or use the data:

Links to other publicly accessible locations of the data:

Links/relationships to ancillary data sets:

Was data derived from another source?
If yes, list source(s):

Recommended citation for this dataset:


DATA & FILE OVERVIEW

File List: <list all files (or folders, as appropriate for dataset organization) contained in the dataset, with a brief description>

Relationship between files, if important:

Additional related data collected that was not included in the current data package:

Are there multiple versions of the dataset?
If yes, name of file(s) that was updated:
Why was the file updated?
When was the file updated?


METHODOLOGICAL INFORMATION

Description of methods used for collection/generation of data: <include links or references to publications or other documentation containing experimental design or protocols used in data collection>

Methods for processing the data: <describe how the submitted data were generated from the raw or collected data>

Instrument- or software-specific information needed to interpret the data: <include full name and version of software, and any necessary packages or libraries needed to run scripts>

Standards and calibration information, if appropriate:

Environmental/experimental conditions:

Describe any quality-assurance procedures performed on the data:

People involved with sample collection, processing, analysis and/or submission:


DATA-SPECIFIC INFORMATION FOR: [FILENAME]
<repeat this section for each dataset, folder or file, as appropriate>

Number of variables:

Number of cases/rows:

Variable List: <list variable name(s), description(s), unit(s) and value labels as appropriate for each>

Missing data codes: <list code/symbol and definition>

Specialized formats or other abbreviations used:
```

Citing Deposited Data

- Elements
  - **Author:** Name(s) of each individual or organizational entity responsible for the creation of the dataset.
  - **Date of Publication:** Year the dataset was published or disseminated.
  - **Title:** Complete title of the dataset, including the edition or version number, if applicable.
  - **Publisher and/or Distributor:** Organizational entity that makes the dataset available by archiving, producing, publishing, and/or distributing the dataset.
  - **Electronic Location or Identifier:** Web address or unique, persistent, global identifier used to locate the dataset (such as a DOI). Append the date retrieved if the title and locator are not specific to the exact instance of the data you used.
  - These are the minimum elements required for dataset identification and retrieval. Fewer or additional elements may be requested by author guidelines or style manuals. Be sure to include as many elements as needed to precisely identify the dataset you have used.

- *APA (6th edition)*
- Smith, T.W., Marsden, P.V., & Hout, M. (2011). *General social survey, 1972-2010 cumulative file* (ICPSR31521-v1) [data file and codebook]. Chicago, IL: National Opinion Research Center [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. doi: 10.3886/ICPSR31521.v1
- *MLA (7th edition)*
- Smith, Tom W., Peter V. Marsden, and Michael Hout. *General Social Survey, 1972-2010 Cumulative File*. ICPSR31521-v1. Chicago, IL: National Opinion Research Center [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2011. Web. 23 Jan 2012. doi:10.3886/ICPSR31521.v1
- *Chicago (16th edition) (author-date)*
- Smith, Tom W., Peter V. Marsden, and Michael Hout. 2011. *General Social Survey, 1972-2010 Cumulative File*. ICPSR31521-v1. Chicago, IL: National Opinion Research Center. Distributed by Ann Arbor, MI: Inter-university Consortium for Political and Social Research. doi:10.3886/ICPSR31521.v1

Information from iassistdata.org

Other Data Management Considerations

- Lab PIs will need to establish rigorous scientific record-keeping practices in their laboratories to satisfy DMS requirements.
- PIs may want to consider switching to the use of electronic lab notebooks (ELN) in their lab to facilitate the organizing of data they aren't publishing

Compare and contrast info for available ELNs:

- 2021 Review of the Best Electronic Laboratory Notebooks | Labs Explorer
- ELN-Scorecard.pdf (labfolder.com)

How to Address Element 3 of the DMS Form: Standards

"State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist."

Data standards help to support the exchange of accurate information and are developed to ensure that data is collected similarly to guarantee the interoperable aspect of the FAIR principles. Standards may be applied in four broad areas:

Standard Metadata schemas for describing datasets

- Example: Dublin Core – Dublin Core Metadata Initiative, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/)

Standard Terminologies, Controlled Vocabulary, and Ontologies

- Example: NIH Health Data Standards, https://www.nlm.nih.gov/healthit/index.html

Content / Encoding Standards, including for storing and transmitting data

- Example: DICOM – Digital Imaging and Communications in Medicine, https://www.dicomstandard.org/concepts

Common Data Elements

- Example: 0 - 10 Numeric Pain Rating Scale. NIH hosts a Common Data Elements repository, https://cde.nlm.nih.gov/home

Open Science Framework explains each of these data terms.

When there is no standard, indicate that no consensus standards exist. Then a detailed data dictionary describing the data fields and format of the data should be provided (typically in a README.txt file).

How to Address Element 6 of the DMS Form:

Oversight of Data Management and Sharing

“Describe how compliance with this Plan will be monitored and managed, frequency oversight, and by whom at your institution (e.g., titles, roles).”

Considerations Regarding Completing “Element 6”

- NIH Instructions regarding Element 6 are pretty much non-existent.

- We recommend that the “compliance manager/overseer” should be either the PI or another permanent senior member of the lab (for example, the Lab Manager or a Ph.D. level staff member).

- If this individual is not the project’s PI, consider whether their biosketch might need to be included as key personnel and/or whether to describe their role within the application, even if there is no effort allocated to that individual in the budget.

- Remember that you can make budget allocations to pay for the effort of whoever will curate and manage the data.

- If you collaborate with other labs/institutions, will you combine data for management or manage pieces separately from lab to lab? It would be best if you spelled this out for multi-PI grants or grants with data-generating collaborators. You will want to ensure your project team is all on the same page about how data will be managed and shared when you write your proposal. It may be helpful to contact your program officer for guidance ahead of time. Note that NIH has made it clear in DMS FAQ that they only want a single DMS Plan per proposal even if multiple investigators and/or institutions are involved.

Recommended Template for Element 6:

[Name of Grant PI or senior member of lab (give title)] will be responsible for verifying management, storage, retention, and dissemination of project data. [NAME] has completed a formal training module on NIH Data Management and Sharing Plan policies and practices, as presented by the Vanderbilt University Office of the Vice Provost for Research. [NAME] will review data management and sharing activity annually and compare it to this plan. If discrepancies are noted, [NAME] will adjust study procedures or submit a revised Data Management and Sharing Plan to NIH.

Note that in cases where there are multiple investigators included in the project (and possibly a subcontract), this statement will have to be altered and extended to explain who will be responsible for managing the data generated by each participating lab and who will be responsible for checking on compliance.

Questions? Contact:

Chuck Sanders, Associate Dean for Research, Vanderbilt SOM – Basic Sciences
chuck.sanders@vanderbilt.edu

Selene Colone, Assistant Dean for Research Logistics and Compliance, Basic Sciences
selene.colon@vanderbilt.edu

Liane Moneta-Koehler, Assistant Provost for Research Integrity and Compliance, VU
liane.monetakoehler@vanderbilt.edu