NIDDK Guidance Regarding Repository Selection

Introduction

The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) supports the sharing of data through publicly available repositories. NIDDK encourages researchers to consider the NIDDK-specific guidance and expectations (https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/guidance-writing-dms-plan) and desirable characteristics of data repositories outlined by the National Institutes of Health (NIH) (NOT-OD-21-016) to select an appropriate repository(ies) for sharing scientific data.

To enable researchers in their repository selection, NIDDK has created the "Considerations When Selecting a Repository" tool. NIDDK does not endorse any specific repository, and this list should not be considered exhaustive. Investigators are encouraged to contact their program officer if they are uncertain about the repository selection process.

Tool Structure

This tool presents researchers with a series of questions. Responses will guide researchers through the tree-like model to relevant repository selection resources. Questions are outlined in green diamonds. Repository selection guidance is found in orange text bubbles.

Researchers may need to complete this process for each specific type of scientific data generated. Selected repositories should be included in the Data Management and Sharing (DMS) Plan in the DMS element section "4. Data Preservation, Access, and Associated Timelines."

Data Linkage Across Repositories

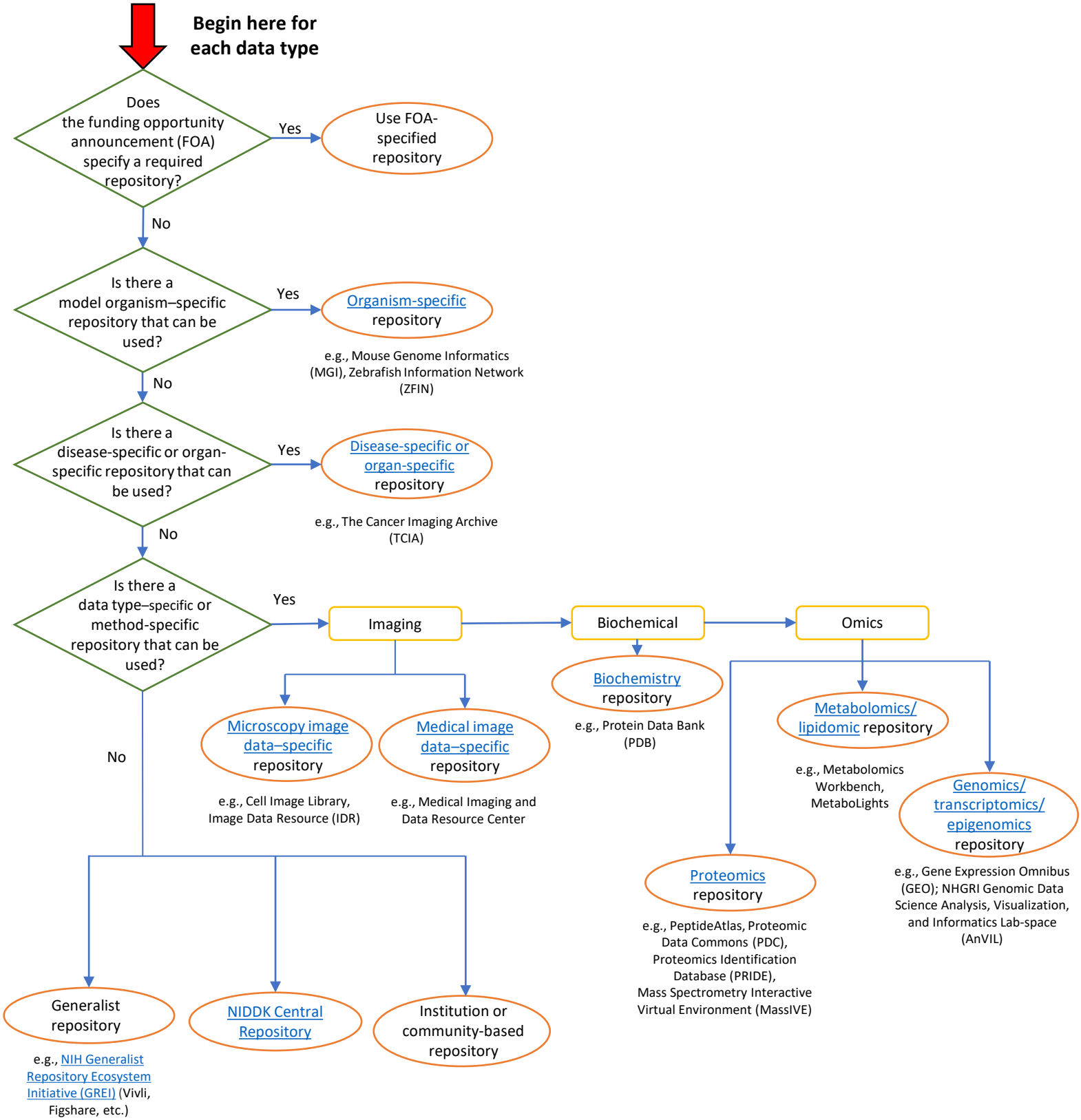
When multiple data types are generated for a given sample or linked within a study (e.g., multiple measures of the same participant or animal), it is important to consider how future users will be able to find and access all the linked study data. Some repositories assign digital object identifiers (DOIs) or globally unique identifiers (GUIDs) that can be reported in the associated study metadata to map and link together related data across multiple repositories. These characteristics should be factored into repository selection.

De-identification and Limits on Data Sharing

Many repositories require data to be de-identified prior to repository submission to protect participant privacy prior to repository submission. If data cannot be sufficiently de-identified, the choice of repository may be limited. While researchers are expected to maximize scientific data sharing, there may be factors that limit sharing of some data. Researchers should consult justifiable limitations to sharing and their project officers on such limitations.

National Institute of Diabetes and Digestive and Kidney Diseases

Considerations When Selecting a Repository



After initial repository selection, investigators should confirm repository eligibility criteria and data requirements (e.g., data access type, preservation duration, storage capacity, possible data linkage strategies, acceptable file format)

Content last updated on March 15, 2023

Example Use Case 1: Companion experiments with different data

A study is proposed to generate the following scientific data:

1. Genotype data generated by sequencing human biospecimen from a clinical study.
2. High-throughput functional screening in the zebrafish model organism, including images.

Questions to address prior to using the Considerations When Selecting a Repository tool:

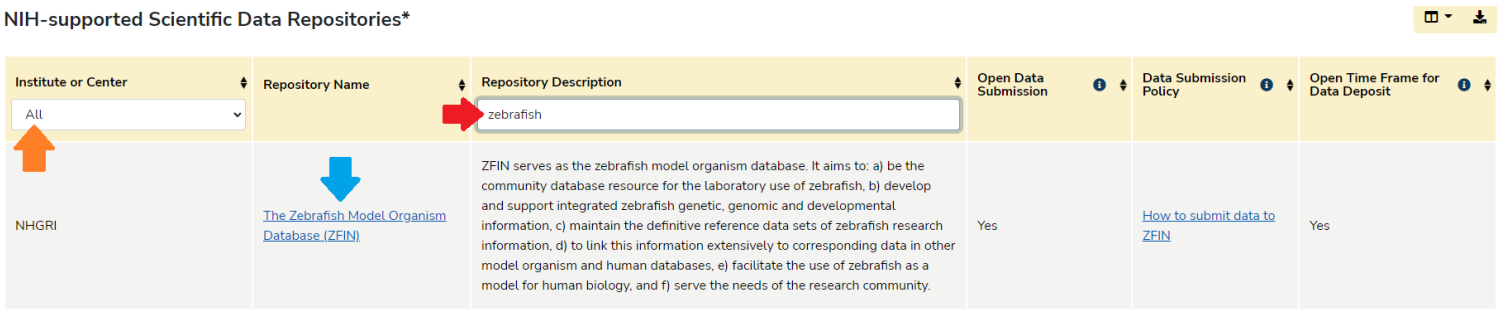
- Does the study include any human data (including cell lines)?
  - Yes, for item 1. Ensure de-identification or any limitations on use have been addressed
  - No, for item 2.
- Does the study include data that needs to be linked or cross-referenced within the study data?
  - No. Items 1 and 2 are not linked, and we expect all data from item 2 to be in a single repository.

Using the Considerations When Selecting a Repository tool:

- For the data outlined in item 1 above:
  - Does the Funding Opportunity Announcement (FOA) specify a required repository? – No
  - Is there a model organism–specific repository that can be used? – No
  - Is there a disease-specific or organ-specific repository that can be used? – No
  - Are there existing data type- or method-specific repositories? – Yes, this study generates Omics data. Searching "genotype" in the NIH-supported Scientific Data Repositories table returns database of Genotypes and Phenotypes (dbGaP). Checking the repository website confirms that dbGaP will accept the genotyping data.
- For the data outlined in item 2 above:
  - Does the FOA specify a required repository? – No
  - Is there a model organism-specific repository that can be used? – Yes, searching "zebrafish" in the NIH-supported Scientific Data Repositories table returns the Zebrafish Information Network (ZFIN). Checking the repository website confirms that ZFIN will accept the phenotype data generated by the functional screen and associated images.

Please note that when searching the NIH-supported Scientific Data Repositories:

- Researchers should search using keywords in the "Repository Description" text entry box (red arrow in image) rather than the funding agency in the "Institute or Center" dropdown menu (orange arrow in image) to find potential repositories (blue arrow in image).
- Researchers are encouraged to consider repositories from all NIH resources (orange arrow in image) rather than filtering to just "NIDDK" when selecting repositories.



Example Use Case 2: Multifaceted experiment

A study is proposed to generate the following scientific data:

1. Single nucleotide polymorphisms (SNPs) from genomics analysis of participant blood.
2. Metabolomics from participant stool samples.
3. Peptides from participant stool samples.

Questions to address prior to using the Considerations When Selecting a Repository tool:

- Does the study include any human data (including cell lines)?
  - Yes, for items 1, 2, and 3. Ensure de-identification or any limitations on use have been addressed.
- Does the study include data that needs to be linked or cross-referenced within the study data?
  - Yes. Data coming from items 1, 2, and 3 should be linked. Look for repositories that issue IDs (such as DOI or GUIDs) to link the data from the same participant.

Using the Considerations When Selecting a Repository tool:

- For the data outlined in items 1–3 above:
  - Does the Funding Opportunity Announcement (FOA) specify a required repository? – No
  - Is there a model organism–specific repository that can be used? – No
  - Is there a disease-specific or organ-specific repository that can be used? – No
  - Are there existing data type- or method-specific repositories? – Yes, each of the scientific data described in items 1–3 are types of "Omics" data.
- For the data outlined in item 1 above:
  - What omic type? SNP (genomics). Searching "SNP" in NIH-supported Scientific Data Repositories returns Single Nucleotide Polymorphism Database (dbSNP). Checking the repository website confirms that dbSNP will accept the human genotyping data in .vcf format.
- For the data outlined in item 2 above:
  - What omic type? Metabolomic. Searching "Metabolomic" in NIH-supported Scientific Data Repositories returns the Metabolomics Workbench. Checking the repository website confirms that the Metabolomics Workbench will accept the quantified metabolite concentrations in .csv format and raw mass spectrometry data in .mzML format.
- For the data outlined in item 3 above:
  - What omic type? Peptide (proteomics). Searching "peptide" in NIH-supported Scientific Data Repositories returns PeptideAtlas. Raw files and metadata will be submitted to PRIDE/ProteomeXchange, which will then pass through to PeptideAtlas, as indicated on the website.