Data Repository Submission Guide for Research Administrators

Repository Selection Checklist

Use this table to evaluate whether a data repository is suitable for your project:

Consideration	Guiding Questions	Examples	Lessons Learned
Grant Requirements	Does the grant require a specific repository?	NIH grants may mandate repositories like dbGaP, AnVIL, or BioData Catalyst.	Non-NIH funded studies may need to find a sponsoring institute or center to use NIH-funded repositories.
Security & Access	Are you submitting controlled access data?	There are specific security requirements for controlled access data and only certain NIH data repositories meet those requirements.	Reach out early to confirm that the data repository can accept controlled access data.
Security & Access	Does the repository allow data with PHI/PII?	AnVIL and other repositories do not accept PHI or PII due to additional security requirements for this type of data.	Reach out early to confirm what types of data the repository can accept.
Capacity	Can the repository handle your data size?	Long read genomic sequencing data has a larger data footprint than other data types and repositories may not be able to take on the complete dataset.	Confirm with the data repository that they would be able to take on a dataset based on the estimated total size. If needed, negotiate with the repository and share only what is most relevant for reproducibility and scientific value.
Data Type Compatibility	Does the repository support your data formats/models?	Repositories offer differing support for specific data types. For example, imaging data or domain-specific data types (e.g. electrophysiology data) may not be well supported by certain repositories.	Negotiate with your program officer if it is possible to submit to a domain-specific repository that will make better re-use and sharing of the data.
Journal Requirements	Does your target journal require a specific repository?	For example, *Nature* has preferred repositories by data type.	Review journal policies early in the manuscript drafting process.

Coordinating a Successful Data Submission

Understand Submission Timing

- Check submission deadlines or release requirements (e.g., NIH GDS Policy)
- **Lesson Learned:** Don't wait until the end of the grant when funding has ended—build in lead time for study registration (up to several months) and data submission (3 months).

Plan for Multi-site or Collaborative Submissions

- Is your study single-site or multi-site?
- **Lesson Learned:** Designate a data coordination lead and aim for a unified, centralized submission.

Review Consent and Data Use Limitations

- What are the terms of participant consent?
- **Lessons Learned:**
  - Broad data use terms like "General Research Use" allow for more general re-use of the data.
  - Overly narrow data use, additional data use modifiers, or segmentation of a study into many consent codes may lead to negotiations with the Genomic Program Administrators to modify the study registration for optimizing sharing while respecting the original participant consent.
  - Avoid parent/sub-study registration formats in dbGaP where possible; though there are certain study designs where this type of registration is optimal.

Confirm Data Submission Requirements

- What are the repository's data modeling requirements?
- **Lesson Learned:** Contact the repository early. Many have user guides, templates, and onboarding support.

Train and Support Submitters

- Do staff know how to submit their data?
- **Lesson Learned:** Recommend to PIs that they train several staff members on data submissions and develop internal SOPs if this is a routine step in their group. Data repositories like AnVIL developed tools for self-service data ingestion that will guide submitters step by step on their submission.

Manage PI Changes

- Has the PI changed during the project?
- **Lesson Learned:** Notify where the study is registered and the repository of any changes and assign a new PI.

Additional Resources for Data Management

- **DUOS for sharing NIH or non-NIH funded data:** https://duos.org/
- **Data Sharing Technology & Policy blog**: https://duos.blog/
- **Are you compliant with NIH's updated Genomic Data Sharing policy? Terra can help!**: https://terra.bio/nihs-updated-genomic-data-sharing-policy/
- **The AnVIL platform is an NHGRI-supported data commons:** https://anvilproject.org/