Other Data Repositories

Amelia Harrison

FAIR Data Practices for Omics Analysis
Presented: April 22, 2022

Before selecting a data repository…

- Know what data is required to be submitted to a repository
  - Submission of data may be required by grants and journals
  - Example: Raw sequencing data
- Determine what (meta)data needs to be supplied as supplementary data
  - Generally decided during the manuscript writing process
- Determine what (meta)data contains barriers to reproduction
  - Computationally expensive to produce, required proprietary software, etc.
- Gather any bespoke code
  - Anything required for reproduction that cannot be easily understood and recreated
- Determine if any other (meta)data is required for reproduction

Selecting a data repository

Important things to consider:

- Does the grant or journal require data to be submitted to specific repositories?
- Are there other requirements for submitting your data?
- Is there a specialty database for the (meta)data?
- How big is the (meta)data?

Desirable characteristics for databases (NIH)

- Unique, persistent identifiers
- Long-term stability
- Free and easy access
- Broad and measured reuse
- Clear use guidance
- Security and integrity
- Common format
- Retention policy

https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository

Domain-specific databases/repositories

Some types of data have dedicated databases

- SRA - raw (unassembled) sequencing data
- PDB - protein crystal structures
- GEO - functional genomics data (RNA-seq, ChiP-seq)
- GenBank - genomes, assembled sequences, plasmids, etc.
  - Less specialized, but still only takes certain kinds of data

GitHub

- Best place to store code and scripts
  - Should submit
- Not a great place for data
  - File size limit of 100mb
  - Metadata is usually okay (remember, it’s often smaller than you main data)
- Owned by MicroSoft

GitHub

- In 2020, GitHub introduced the GitHub Archive Project
- “Mission is to preserve open source software for future generations by storing your code in an archive built to last a thousand years”

General repositories

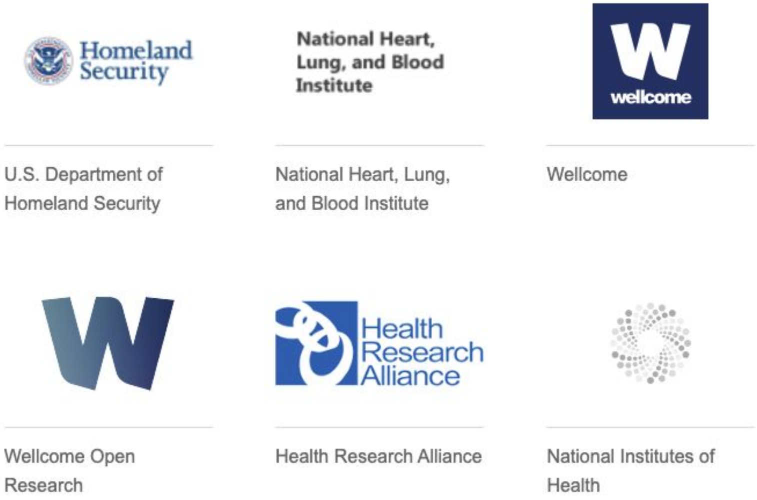
- Collect data regardless of type
- Good for any (meta)data that does not have an obvious home elsewhere
- Zenodo and Figshare are the most common
    - Zenodo size limit: 50GB per project
    - Figshare size limit: 20GB per account
- Both have stable support
- Both require you to make an account and apply a data license

zenodo

- Government funded
  - Funded by CERN, OpenAIRE and the EU

figshare

- In 2013, Figshare partnered with journal publisher PLOS to help make data from publications available
- Funded by several US Government agencies



U.S. Department of Homeland Security

National Heart, Lung, and Blood Institute

Wellcome

Wellcome Open Research

Health Research Alliance

National Institutes of Health

Other options

- Small datasets can generally be stored by a journal
    - Articles submitted to PubMed Central can have datasets ≤ 2 GB in size
- There are many other domain-specific databases/repositories
    - These are often supported by a single funding source or lab
    - Be aware that these databases often disappear
        - In other words, you cannot rely on them for permanent data access
    - Does not mean you cannot contribute to them

Case Study: The Pacific Ocean Viromes (2013)

- “All sequences were deposited to CAMERA (http://camera.calit2.net) under the following project accessions: CAM_P_0000914 and CAM_P_0000915”
- At the time, CAMERA was a publically available, searchable database
  - Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis
- Later on, CAMERA ran out of funding and it no longer exists
- Now, the data is hosted by iVirus in the CyVerse Data Commons
- Technically, the data is still publicly available, but hard to find

Which accession numbers do I provide?

- Provide a different accession number for each repository
- Example: For a whole-genome shotgun sequencing project you have...
    - A BioProject
    -