

BioData.pt

Ready for BioData Management?



Advanced Data Management Topics

Daniel Faria, João Cardoso



<https://rdm.elixir-europe.org>

Learning Outcome 1:

Understand the options available to fill in the various sections of a DMP and/or where to look for them.

FAIR Data Documentation

- **Metadata Standards**

- Specify the metadata fields that must be filled in to enable data interpretation
- Can be looked up in [FAIRsharing](#)
- Not all domains have metadata standards
 - Adapt a similar standard: core metadata fields are essentially common to all domains
 - Adopt a generic standard: probably not rich enough for FAIR data
 - You can extend it

FAIR Data Documentation

- **Controlled Vocabularies & Ontologies**

- Used to fill in the metadata fields and/or to annotate the data itself
- Can be looked up in [BioPortal](#)
- Check metadata standard for ontology recommendations
- Choose domain-specific ontologies when able

- **Metadata Capturing**

- Automatically by experimental equipment or software
- Manually, in electronic lab notebook or other *in silico* solution
- Manually, on paper
- **Daily**, throughout the project

Data Quality

- Equipment calibration and verification practices
- Equipment-provided quality assessment (e.g. nucleotide sequencers)
- Service provider's quality assurance (e.g. ISO certification)
- Use of controlled vocabularies
- Data validation
 - Upon entry
 - *A posteriori*
- Data cleaning
 - Remove outliers
 - Handle missing values

Storage

- Personal (group) storage
 - Acquired through the project (in budget)
 - Pre-existing (maintenance costs in budget)
- Institutional storage
 - Acquired through the project (in budget)
 - Pre-existing (usage costs covered by overheads)
- National
 - e.g. FCCN, BioData.pt
- Cloud
 - e.g. Google, Amazon

Backups

- DIY
 - Redundancy:
 - Physical: redundant data server, hard-drive, tape
 - Virtual: virtual machine
 - Periodicity:
 - Triggered: periodic/automatic check of changes
 - Manual: upon changes
 - Periodic: e.g. hourly, daily
- Other
 - Check institutional or service provider's backup practices and assurances

Security & Protection

- Malicious attacks and accesses are essentially impossible to prevent
- Accidents happen: fire in a data center, early hardware failure
- Sensitive data should **always** be encrypted, so that when access happens, it is not compromised
- All data should be backed-up in a separate physical location (or the cloud) so that when accidents or malicious attacks happen, nothing substantial is lost
- Access protocols: who will have access and how access is controlled
- Consult IT experts in your institution or elsewhere

Legal & Ethical Requirements

- **Personal Data**

- Carefully review [GDPR checklist](#)
- Data anonymization policy for sharing amongst project partners and/or publication
- Personally identifying information is sensitive and should **always** be encrypted
- Research subjects always need to sign consent forms
- Research subjects have the right to request their data and ask you to remove it any point
- Assign a person responsible for overseeing personal data

Legal & Ethical Requirements

- **Intellectual Property**

- Typically owned by the host institution
- Make sure to check your institutional policies and your contract with them

- **Code of Conduct**

- Avoid gender bias and discrimination
- Avoid bias and discrimination towards minorities
- Handle occurrences of inappropriate behaviour
- Typically deferred to the host institution
 - There must be agreement between projects spanning multiple institutions and countries

Data Sharing & Preservation

- **Data sharing**

- All non-sensitive data should be made public to comply with FAIR principles
- You can have an embargo if needed to secure publication of research articles
 - This should not exceed 2 years after the end of the project

- **Data preservation**

- You are legally bound to preserve research data for a number of years (check institutional, national and funders' policies)
- Data that is shared in a public repository is essentially preserved (but you may still be legally required to host it as well)
- Data from failed experiments due to faulty materials, protocols, etc, can be erased

Final Remarks

- Consult institution experts (IT, policy, ethics, etc) and ask for their contribution on the DMP
- Consult national experts if your institution doesn't have in-house expertise
- Consult data management toolkits
 - e.g. [ELIXIR RDM Toolkit](#) (under construction)