H3ABioNet

Pan African Bioinformatics Network for H3Africa

Process of making data FAIR and OPEN

Ziyaad Parker

moz://a

Background

- More and more research data are being generated
- Researchers are tackling difficult questions on large data sets
- Data Management is important because it requires **sharing and reusing** of data

Problems with current approach

- Data is generated and nobody is able to find the data sets
- If it is online,
  - It is difficult to use because no documentation is available
  - Nobody is aware of the fact that it is online
- When researchers code, sometimes the code and data is kept in private and not well documented
  - What happens when the person responsible?
    - Resigns
    - Accident
    - Losing of hard drive, formatting because of virus found
    - Password Protected, etc
- Difficult to document every step. What to share? What not to share?
- There is low motivation and capacity in the community to share and document software and data properly, often due to **lack of information** and incentives

Problems with current approach

- **Research data needs to be easily discoverable.**
- Very few data repositories are doing a good job of this such as Space Physics Data Facility and Genbank.

GODDARD SPACE FLIGHT CENTER
Space Physics Data Facility

RELEASE
100.0
GenBank
NCBI
National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health

Good Research Management means to ...

- **Plan**
- **Collect**
- **Describe**
- **Analyse**
- **Preserve/Publish**
- **Access**

- Data should be **easily available to download** with no complicated software to use

FAIR means lowering barriers and not "reinventing the wheel" in order to enable meaningful (novel) research

Importance of FAIRifying data

- “Percentage of time spent finding and organising data according to research data specialists: 79%” (#RDAPlenary).
- The way to move forward with big data is to share it.
- FAIRness should be assessed before and after the work with data is done.

79%

What is FAIR Data?

- ‘As open as possible, as closed as necessary’
- There are different degrees of FAIRness, as research disciplines, resource types (e.g. data and software) and their requirements are strongly varied - but the shared goal is **quality & good scientific practice**

What is FAIR Data

- The FAIR Data Principles by the **FORCE11**
  - **To be:**
    - **Findable**
    - **Accessible**
    - **Interoperable**
    - **Reusable**

FAIR Data Principles

- **To be Findable**
  - F1: Metadata are assigned a globally unique and eternally **persistent identifier**
  - F2: Data are described with rich metadata
  - F3: Metadata are registered or indexed in a searchable resource
  - F4: Metadata specify the data identifier

Findable - Role as a researcher

- Assign a globally unique PID upon publication (or draft upload)
- provide metadata schema in human- & machine-readable format
    - PID, author names, subject areas, etc.
    - support structured input of metadata (submission forms or XML schema)
    - index (meta)data to enable effective searching
    - allow metadata upload & assign corresponding PID

Findable - What researchers need to know?

- Digital Object Identifier (DOI) is an international and recognised standard
  - Most researchers use doi’s for paper publications
  - Few researchers use doi’s for research data
  - http://doi.org/10.5524/100336
  - Project, Instrument, Experiment, Runs, Physical Samples, etc

PID 101

1. A PID is a “long lasting reference to a digital resource“
2. There are different sorts of PIDs & different uses, (e.g. for articles, data, persons, organizations, ...)
3. PIDs are offered by organizations - Ask your institute/library
4. You do NOT have to pay for PIDs (by yourself)!
5. PIDs are mostly used for (persistent) citation – All published resources should have one
6. A correct citation always includes a PID → look in your citation manager
7. Metadata behind a PID are most important – please take care when providing them
8. PIDs are not perfect (they are issued by organizations, aka humans!)
9. PIDs are really useful & fun – they make yourself & your work more visible!

Findable - What researchers need to know?

- PID - Persistent Identifier
- Provenance means validation and credibility - a researcher should comply to good scientific practices and be sure about what should get a PID (and what not).
- Metadata is central to visibility and citability - metadata behind a PID should be provided with consideration.
- Policies behind a PID system ensure persistence in the WWW - point. At least metadata will be available for a long time.
- Machine readability will be an essential part of future discoverability - resources should be checked and formats should be adjusted (as far as possible).
- Metrics (eg: altmetrics) are supported by PID systems.

Findable - Role as a scientist

- Check datasets that you use for a PID and cite it
- Ensure that your datasets get published with a PID
  - Choose repositories that automate this
  - Report this requirement to repos that don't
- Add rich metadata (describe dataset's context, quality, condition and characteristics)
  - Should be understandable by researchers from different discipline (ask a friend to proofread)

Data Citation using PIDs

Paper

Koen Kole, Rik G.H. Lindeboom, Marijke P.A. Baltissen, Pascal W.T.C. Jansen, Michiel Vermeulen, Paul Tiesinga, Tansu Celikel (2017):
Proteomic landscape of the primary somatosensory cortex upon sensory deprivation,
GigaScience Volume 6, Issue 10, 1 October 2017, Pages 1–10. DOI
https://doi.org/10.1093/gigascience/gix082

Note in the paper

“Availability of the supporting data Data supporting this work are available in the GigaScience repository, GigaDB [14]. The raw mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository [15] with the dataset identifier **PXD005971**”

Reference

[14] Kole K, Baltissen M, Lindeboom R et al. Supporting data for “Proteomic landscape of the primary somatosensory cortex upon sensory deprivation.”
GigaScience Database 2017.http://doi.org/10.5524/100336

FAIR Data Principles

- **To be Accessible**
  - A1: Metadata are retrievable by their identifier using a **standardized** communications protocol
    - A1.1 The **protocol is open, free** and universally implementable
    - A1.2 The protocol allows for an **authentication and authorization** procedure, where necessary
  - A2: **Metadata remain accessible**, even when the data are no longer available

Accessible - What researchers need to know?

- Using Git (GitHub, GitLab, etc) for making the data publicly and easily available
  - Not only about making the data available, but also the code that was used to generate the data
  - In a specific format that the other researcher should be able to easily download and use
- What is Git?
  - Git is used to manage a project or set of files as they change over time
  - This gets stored in a data structure known as a repository
  - Web-based graphical interface www.github.com
    - It provides access control and several collaboration features, such as a wikis/documentation tools and basic task management tools for every project
    - Tracking changes in code across multiple versions
    - Markdown and to showcase your work online
    - Can build websites on GitHub (free hosting and subdomain)

Accessible - Role as a scientist

- Access data programmatically whenever possible (web services + R packages and Python modules)
- Email requests sometimes necessary and also FAIR (sensitive data)
    - If granted: secure access possible? Password manager => unique passwords!
- Metadata can help to plan research (especially replication)
- Request these features from the repositories (recommended)

FAIR Data Principles

- **To be Interoperable**
  - I1: Metadata use a **formal, accessible, shared, and broadly applicable language** for knowledge representation
  - I2: Metadata use **vocabularies** that follow FAIR principles
  - I3: Metadata include **qualified references** to other metadata

Interoperable - What researchers need to know?

- Provide machine-readable metadata with well-established formalism structured, using disciplined-established vocabularies / ontologies / thesauri (RDF extensible knowledge representation model, OWL, JSON LD, CSV, XML, schema.org)
- Support referencing metadata fields between datasets via schema (relatedIdentifier, relationIdentifier)

Interoperable - What researchers need to know?

- Data needs to be in the appropriate folders and as readable as necessary

  - **Organized by File Type**
  - Example A
    - Code
      - Step 1
      - Step 2
    - Data
      - Processed
      - Raw
    - Results
      - Figure 1
      - Figure 2
      - Models
    - Readme.txt

  - **Organized by Analysis**
  - Example B
    - Figure 1
      - Code
      - Data
      - Results
    - Figure 2
      - Code
      - Data
      - Results
    - Readme.txt

Interoperable - Role as a scientist

- Provide as precise and complete metadata as possible
- Look for metrics to evaluate the FAIRness of a controlled vocabulary / ontology / thesaurus
    - Often do not (yet) exist
    - Develop controlled vocabulary/ontology/thesaurus
- Clearly define relationships between datasets in the metadata (eg: “is new version of”, supplements to”, “relates to”, etc.)

FAIR Data Principles

- **To be Reusable**
  - R1: Metadata have a plurality of **accurate and relevant attributes**
    - R1.1 Metadata are released and with a clear accessible **data usage licence**
    - R1.2 Metadata are associated with their **provenance**
    - R1.3 Metadata meet domain-relevant **community standards**

Reusable - What researchers need to know?

- Provide metadata schema in human and machine readable format
- Request relevant general and / or subject-specific metadata from researchers
- Offer license file upload or references
- Implement discipline-specific metadata standards if necessary

Reusable - Role as a scientist

- Be as detailed as possible when adding metadata to provide useful context
  - Should be understandable by researchers from a different discipline
  - It needs to be proofread
  - Purpose of data creation / collection, date, conditions, parameter settings, etc.
  - Raw or processed data or both?
  - Explain variable / column / parameter names, if not self-explanatory already or vocabulary-defined
  - Document and cite datasets and software (+version) that you used
    - Cite it using one of the reference managers (BibTeX, Citavi, EndNote, Mendeley, Zotero etc.)
  - Data life cycle, data dissemination, tools used, workflows

Reusable - Role as a scientist

- Set a licence and a provide a link to the licence
    - If applicable, provide information on additional legal conditions
- Specify provenance (your role in collecting / generating the data), citation wish
- Use community standards for data archiving and publication or explain other choices
- Request that repositories collect these details

Benefits of data sharing/publication in good data repositories

- Data are kept safe in a secure environment
- Data are regularly backed up and preserved (long-term) for future use
- Data can be easily discovered by search engines and included in online catalogues
- Intellectual property rights and licencing of data are managed
- Access to data can be administered and usage monitored
- The visibility of data can be enhanced
- Enables more use and citation
- Citation of data increases researchers scientific reputation

Some recommendations:

→ look for the usage of PIDs

→ look for the usage of standards (DataCite, Dublin Core, discipline-specific metadata)

→ look for licences offered

→ look for certifications (DSA / Core Trust Seal, DINI/nestor, WDS, ...)

Any Questions???

Exercise

Get into groups of four. Download any data set. Use any language you are comfortable with (R, Python, Excel, etc).

1. Create a graph (line, bar, etc)
2. Write a few lines of analysis (in word or txt file)
3. Make this data FAIR by using the FORCE 11 principles process.
   a. Upload the files into Github
   b. Cite where you got the data from as well as the PID (i.e: on readme file)
   c. Put all the files in an orderly manner for any person to read
   d. Include rich metadata, specify provenance, etc
   e. Add a licence, to specify your conditions if any
   f. Ask another group to proofread it and see if they are able to download and use it