FAIR Data

Adil Hasan Uninett/Sigma2

Nicest2 Hackathon 11/Mar/21

Motivation

Motivation

- Problem is that it may be obvious what the content is and how to use it, but a little later it may not be so clear.

Motivation

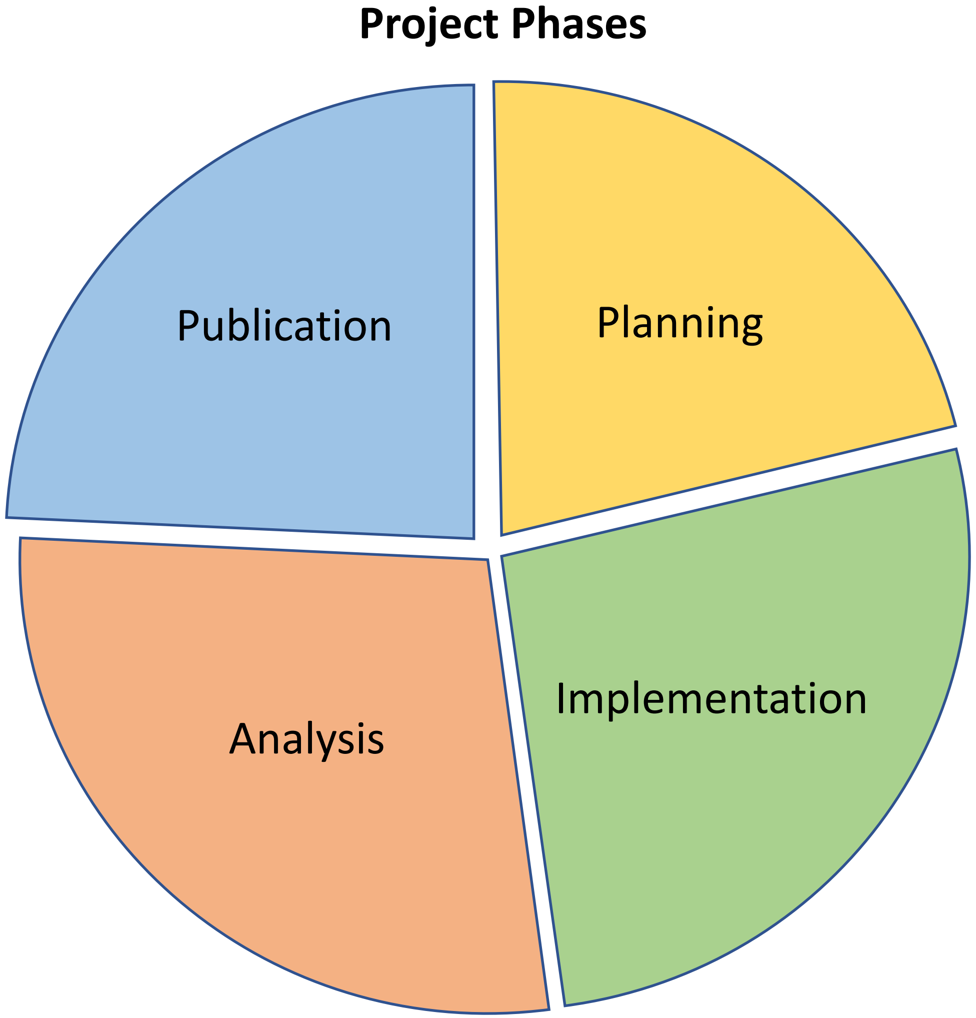
- Lots of data exists that researcher are *already* reusing.
- FAIR aims to provide researchers with some guidelines to make it easier for others to reuse their data in a machine-actionable manner.
- FAIR paper:
  - https://www.nature.com/articles/sdata201618

What is FAIR?

- Findable – metadata, identifier
- Accessible – AAI, licenses, documentation, metadata
- Interoperable – documentation, metadata
- Reusable – documentation, metadata
- It applies to *outputs*
  - Data internal to your project that only you need to use doesn’t need to be FAIR (as long as it’s useable by you and your intended audience).

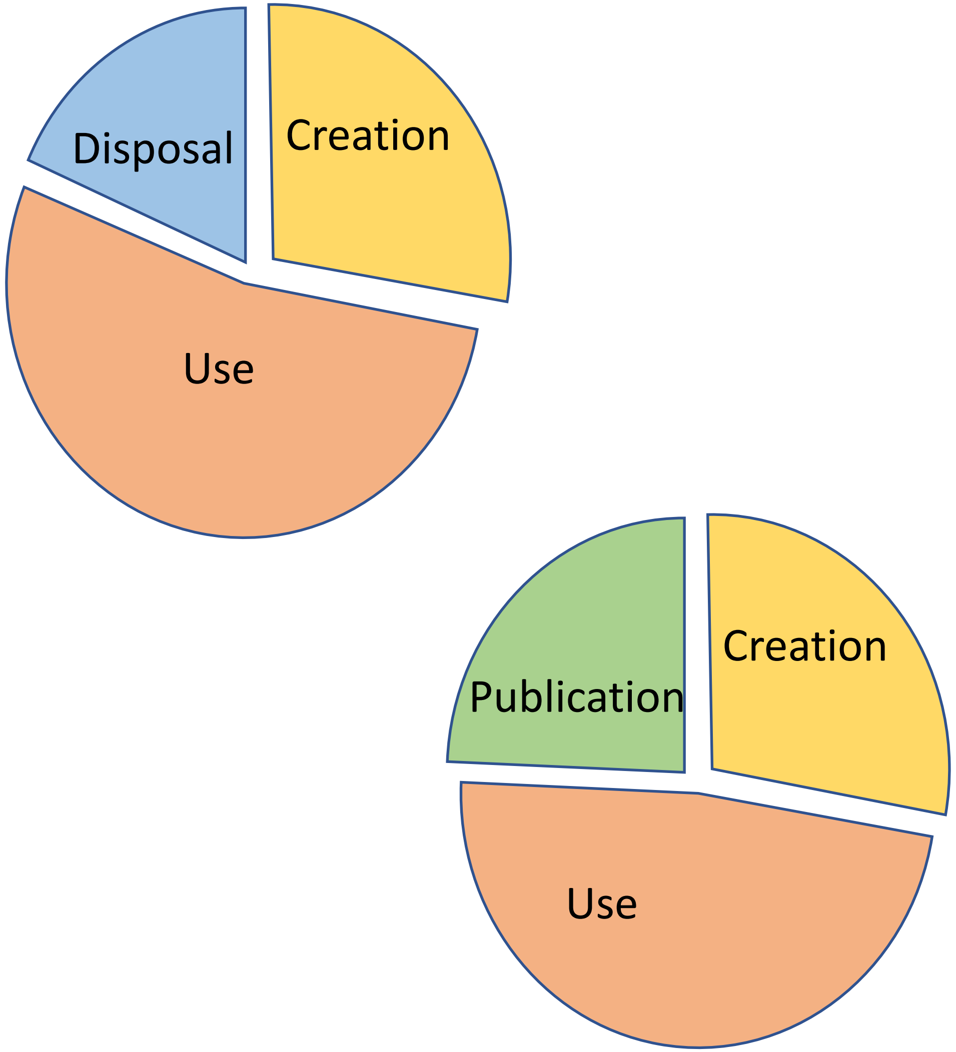
Where does it apply?

- Highly idealised
- Actually, after planning all phases may overlap
- **Each phase may produce outputs**!
  - Software, documentation, data, publications, tutorials, etc.
  - These outputs may be useful to people outside of your project.



Data Lifecycle

- At each phase digital objects may be produced.
- Each digital object has a lifecycle
- Some of these will live beyond the end of the project, some not.



Data Lifecycle

- Creation – object created or object updated, metadata created, storage, documentation, provenance, licenses, etc.
- Use – object may be moved, packaged with other objects.
- Deletion – object is no longer useful and deleted.
- Publication – object packaged for use by others (incl license, AAI, documentation, metadata, etc).
- What about data reuse?
  - Skip the Creation step

FAIR – what should I do?

- Think about your audience – what do they need to use to reuse your digital content programatically?
- Findable
    - How can your audience find your data?
        - Give digital content a static or persistent identifier (e.g. DOI). Include persistent id in your publication.
        - Make it possible for people to search for your content, put in some registry that can be searched, put on commonly used services.
        - Use commonly used terms or standard terms and metadata, include licenses describing access and reuse.

FAIR – what should I do?

- Accessible
  - Provide documentation on what researchers need to do to access the content.
  - Ideally, the documentation should enable machine access.
  - Ideally make the content as free from restrictions as possible. It makes programmatic access much simpler.
  - Licenses
  - Make sure there are standard or community agreed, composable ways to access the data
- Interoperable
  - Use standards for content and metadata, or widely used community formats and terms.
  - Make sure content arrangement is documented.
  - Licenses need to be flexible enough to enable interoperation.

FAIR – what should I do?

- Reuse
  - Documentation and metadata on how to use the content – ideally with some examples or reference outputs.
  - Did we mention licenses?
  - Try to make sure the docs and metadata are understandable by reusers.
  - Make sure there’s a way for your audience to get help (Data Steward/content expert contact).

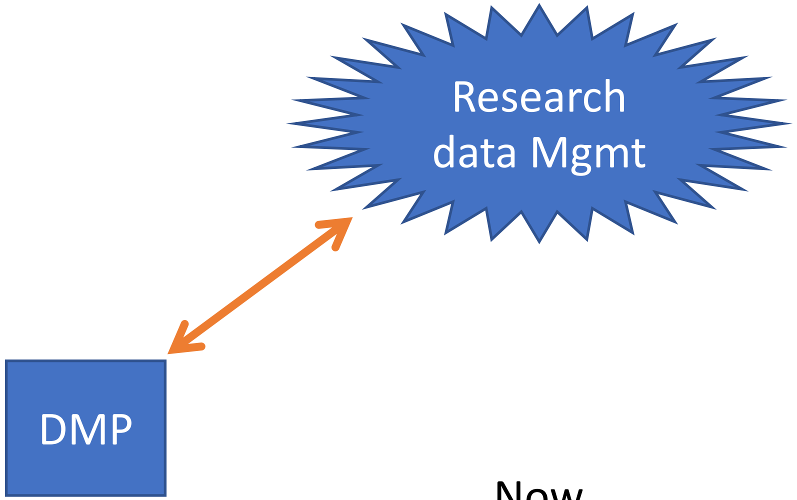
FAIR – all well and good, but this is research…

- Difficult to identify potential users of the data, so difficult to include all relevant metadata.
    - Best solution is to make sure there’s a contact person familiar with the data. That’s ok for the short term, but longer term you can only have domain experts who may be able to answer some questions, but not all.
        - They may not know about anomalies, e.g. ‘why is there a gap in the data here?’, or ‘why were these intervals chosen?’
        - So, anomalies and any design decisions are very much worth documenting at the time the content is being created.

Can I measure how FAIR my data is?

- There exists a tool which is under active development: F-UJI: https://f-uji.net/
    - Code is on github with an API: https://github.com/pangaea-data-publisher/fuji
- It’s under active development and results only give an indication of the FAIRness. It should really only be used as a guideline.
- The most reliable test of the FAIRness of your data is how easily your intended audience can find and reuse your data.

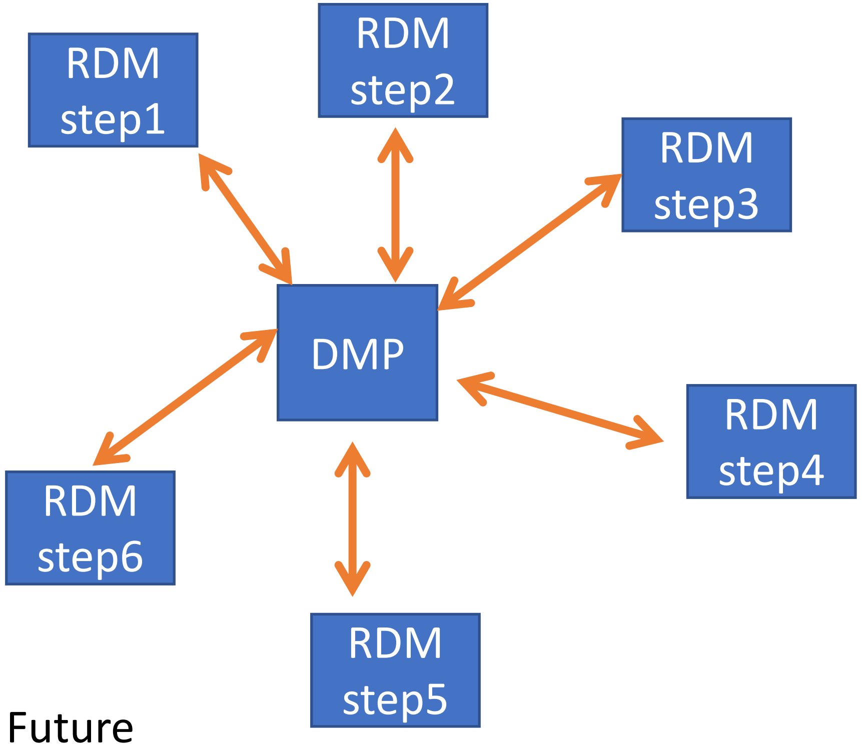
Data Management Plans



Research data Mgmt

DMP

Now



RDM step1

RDM step2

RDM step3

DMP

RDM step4

RDM step6

RDM step5

Future