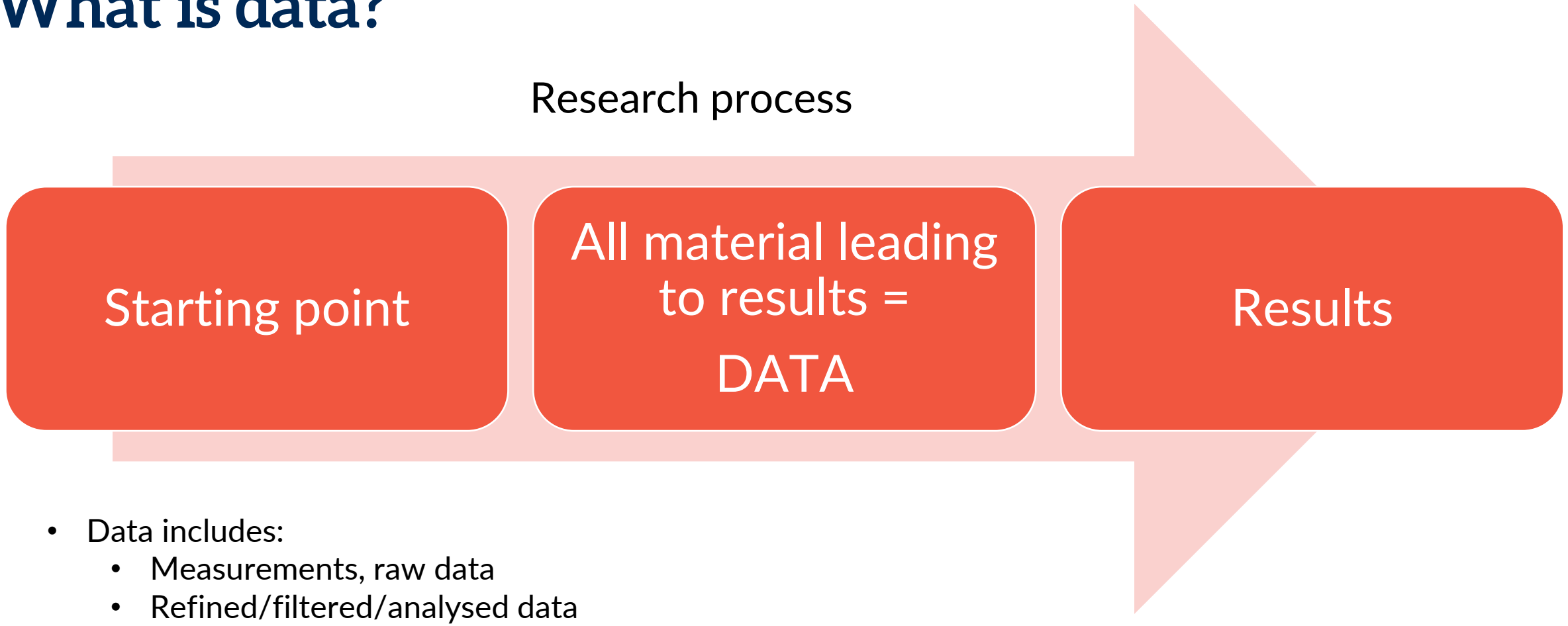
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Open Data – Requirements and How-to

Juuso Marttila, juuso.h.marttila@jyu.fi, Data Management Expert
Research Data Services, Open Science Centre, University of Jyväskylä

What is data?



- Data includes:
  - Measurements, raw data
  - Refined/filtered/analysed data
  - Code related to the production/analyzation of the data
  - Lab books, codebooks, other documentation

Funder requirements on Open Data

- For Research Council of Finland and EU participation in Open Science is not an option, it is a default
  - Most foundations do not have specific stipulations on Open Data, BUT your university or research infracstructure often does
- This means that your data must be available as openly and as soon as possible and to follow FAIR principles
  - at least metadata must be published for all data from the project
- Data should be opened as soon as possible
- Your data management plan outlines your promises on research data
  - Be credible, be realistic, follow your plan

Example: Horizon project agreement on research data

29.3 Open access to research data

Regarding the digital research data generated in the action ('data'), the beneficiaries must:

(a) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:

- (i) the data, including associated metadata, needed to validate the results presented in scientific publications, as soon as possible;
- (ii) other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan';

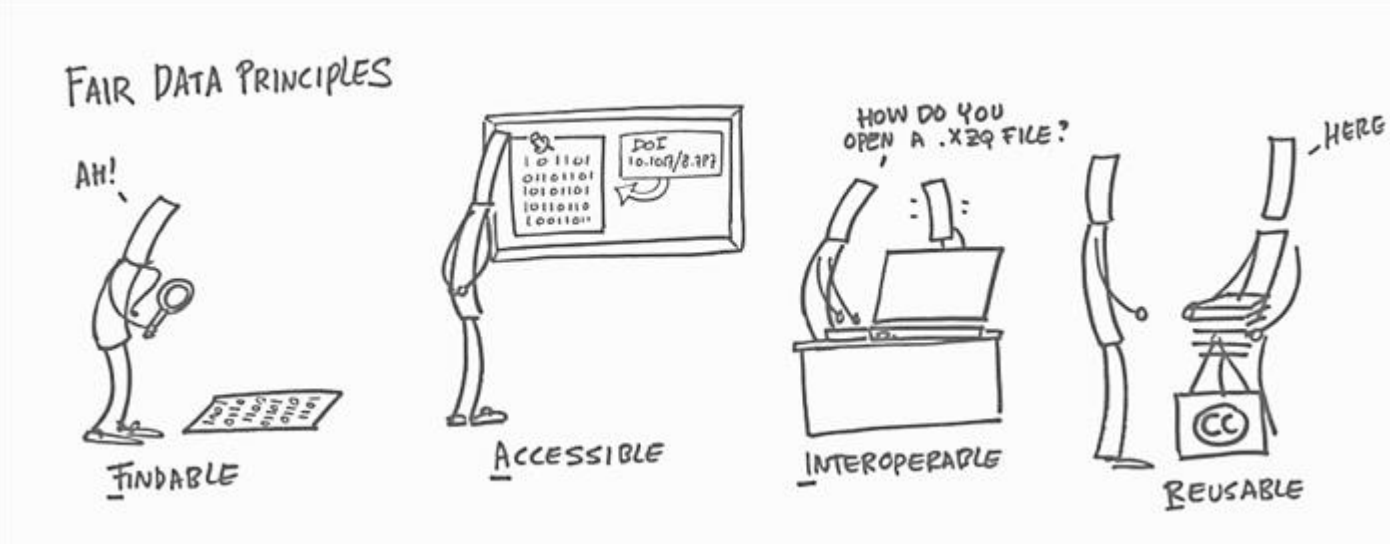
(b) provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).

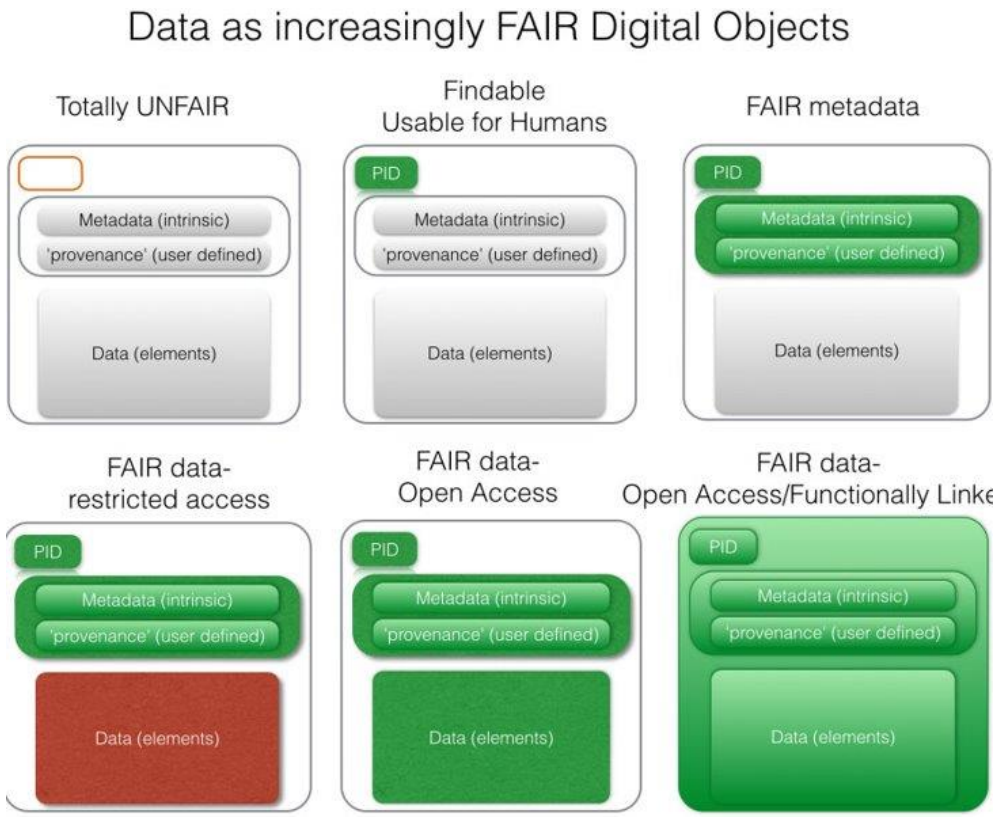
Rerporting data to funder

- Research data and its availability will be reported in end reports for EU and RCF
  - Also note that EU projects are often lump sum fundings of which part can be revoked if promised deliverables (published datasets) are not produced
- RCF asks you to list all used and created research datasets, how they have been managed and how re-use of dataset has been made possible
  - Also list of all datasets suitable for re-use and where they can be downloaded (PID)
  - If only metadata is available, list metadata landing pages with their PIDs
  - If dataset is not openly available, you must explain why

FAIR principles in the core of Open Science

- **Data cannot always be open, but it can always be as FAIR as possible**





As open as possible, as closed as necessary – acceptable reasons for not publishing data openly

- Never publish/open
  - Data containing personal information
  - Data on which a third party holds IPR rights
- Use a *reasonable* embargo for
  - Data necessary for patents in progress
  - Data for your own upcoming publications
  - Data that you're going to exploit yourself in forseeable future
- No need to disclose data unsuitable for any reasonable reuse
  - BUT also, remember the possibility to show your zero results by publishing data when paper cannot be done due to lack of results
- Remember that some data can have restricted (e.g. only for research purposes) access, if that makes publishing it easier

Funder requirements for open data in practice

Not sufficient

- Witholding data without justified cause
- Publishing files on your own web page (no PID = no open data)
  - PID = persistent identifier, e.g. DOI, URN
- Publishing without valid open licence

Sufficient

- Publish data in open, proper data repository that gives PID and licence to the dataset, OR
- Publish metadata in proper data repository/catalogue AND
  - Argue in DMP/report on why whole dataset could not be opened
  - Bare minimum for ALL datasets!

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Open Data – What, where, when…

What data you should make openly available?

- All data or just parts needed as supplements for publications?
  - Is the data only for validation of your results or are there some possible re-use?
  - All at once (in the end) or in smaller bits per publication?
    - Remember that funders are interested in all data, not just those that supplement publications
  - Remember possibility to make linkages between multiple datasets
- Raw data or analyzed/filtered data?
  - Is raw data usable/understandable as such?
- Remember, whatever you publish, it must be well-organized and well-documented and usable as such

Different data repositories for opening data

- Discipline specific repositories are the golden standard
  - E.g. HEPData (Repository for *publication-related* High-Energy Physics Data), Cambridge Crystallographic Data Centre (CCDC)
- Some journals have their own repositories for supplementary data
- Institutions may have their own repositories (JYX for JYU, Datapankki for HY
  - You can also use CSC's IDA storage + ETSIN publishing platform via your institutions
  - Datapankki and IDA are for especially large datasets
- There are also generic repositories (e.g. Zenodo, Figshare, Open Science Framework), but most of these are not curated
- If you're publishing only metadata, you can use JYX in JYU or Data Catalogue in HY. Qvain and Zenodo also make this possible.

Publishing code

- Code should always be maintained in some version control environment, e.g. Github or (institutional) Gitlab
  - Remember sufficient documentation, code-specific licences etc. alongside the code
- Github has an export option to Zenodo
- From institutional Gitlab you can take a snapshot (zip) of the code repository and publish that in institutional or any other repository.
- If you want to point a live repository, you can make a metadata publication that points to the Gitlab repository
- Also remember that MATLAB and R have own communities for publishing code as standalone sets: The Comprehensive R Archive Network, rOpenSci, MATLAB File Exchange

Checklist for opening data

- Take opening into account when planning data management
  - Extra important to agree on ownership, use rights and opening data
    - Any third party rights must be considered
  - Plan what is appropriate timing for opening the data
- Create dataset with sufficient documentation and arrangement so that others may understand and use it (file naming, readme-files, lab diaries etc.)
- Choose appropriate repository and level of openness
- Choose and give appropriate licence for the data (e.g. Creative Commons licences). It tells recipient on how the data may be used and safeguards your rights to be cited when data is used.

More information and help

- JYU: https://www.jyu.fi/en/research/research-data-management/guide/opening-research-data
  - researchsupport-osc@jyu.fi
- HY: Checklist for providing open access to research data (1.0). **University of Helsinki**. https://doi.org/10.5281/zenodo.17698438 (FI, EN, SV)
  - datasupport@helsinki.fi