Research Data Management

Obrad Vučkovac
University of Belgrade
Vinča Institute of Nuclear Sciences – Library

EIFL Open Science bootcamp

Photo by Rene Böhmer on Unsplash

Training challenges

- Complexity of the topic
  - RDM, DMP, FAIR, Data publishing and licenses, Open Data
- Terminology
- Misconceptions (Open Data, data scooping)
- Good practice, but not always mandatory
- More work for researchers, “administrative burden”
- Infrastructure (in national language)
- Lack of incentives

Research data management

Research data management (RDM) refers to the organization, storage, preservation, and sharing of data that was generated or collected and used in a research project.

Research Data Management = good research practice

Final goal of RDM is that data are:

- secured and preserved;
- easy to find, understand and reuse.

Research data

Research data - information that has been collected, observed, generated or created to validate original research findings

- Depends on audience;
- Digital and non-digital data
  - Non-digital data --> physical samples, 3D digitization; include metadata;

Question: *What needs for data to be understandable and reusable, even after some longer period of time?*
Data needs metadata and other documentation (codebook, software code, lab notebooks, data quality assessments, etc.), and to be in sustainable formats.

Research data and metadata needs to be ready for both

humans & machines

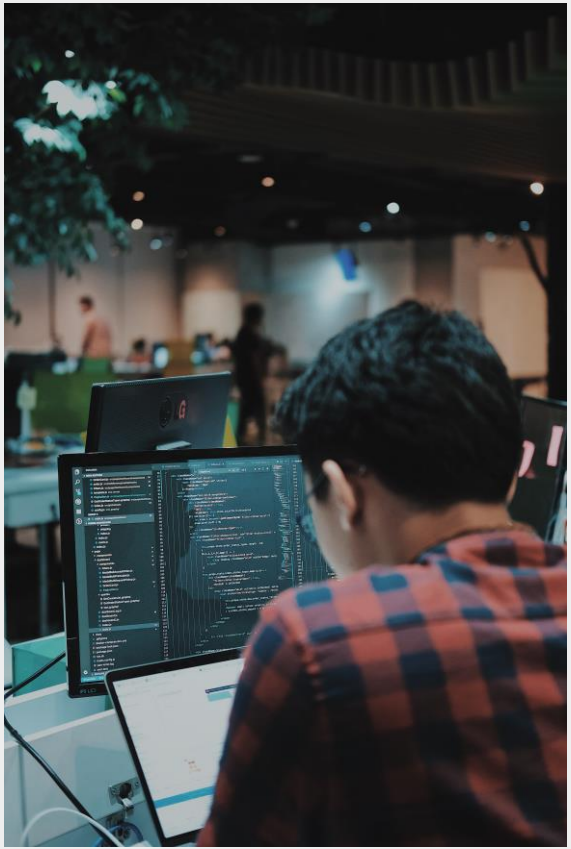


Photo by Arif Riyanto on Unsplash

The need for RDM

Reasons why researchers need a good RDM practice:

- requirements: funder, institution, publisher
  - ex. Horizon Europe, Bill & Melinda Gates Foundation, Wellcome Trust, etc.
- improve scientific communication and cooperation
  - increase citations
- transparency, reproducibility
- prevention of data loss
- reduce duplication of efforts
  - more ethical research

Common RDM misconceptions

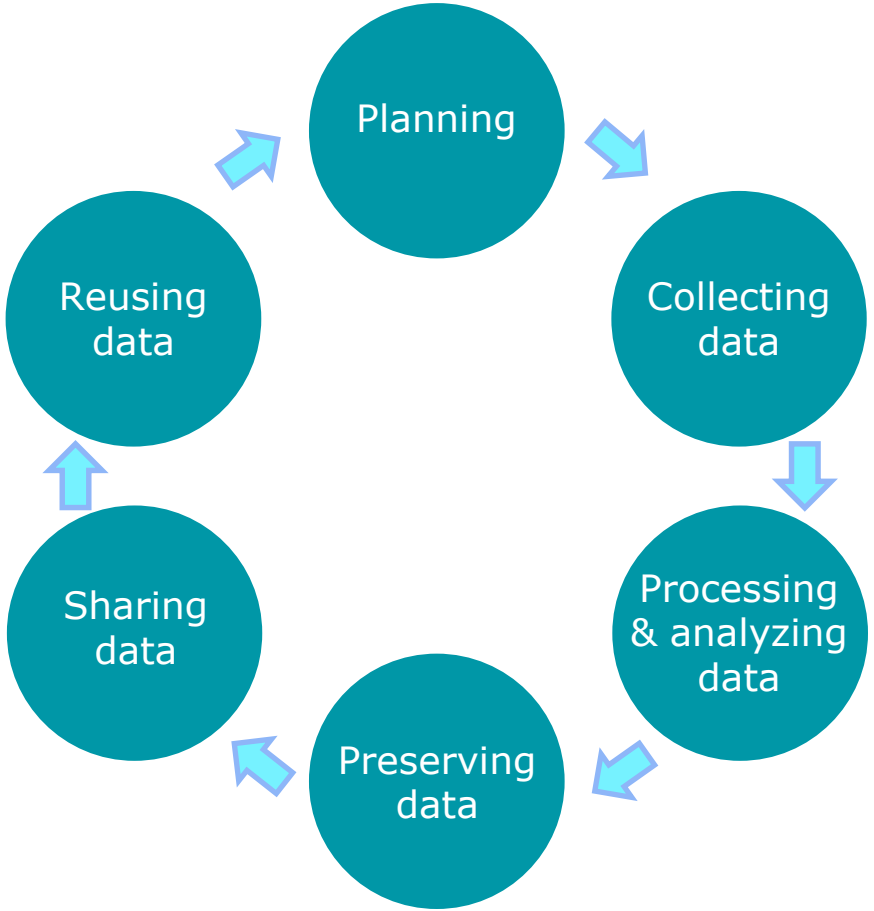
Common misconseptions regarding RDM:

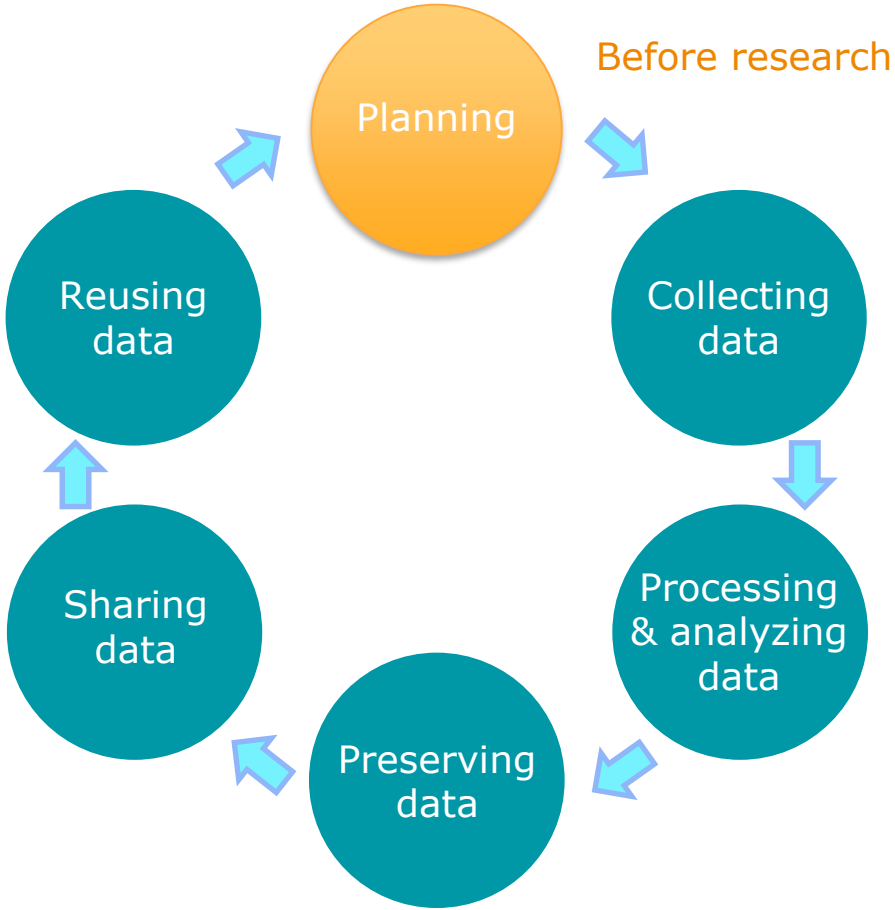
- administrative burden
- data scooping, lack of control over data
  - RDM does not mean Open data
- fear of wrong data interpretation
- IPR issues, data privacy

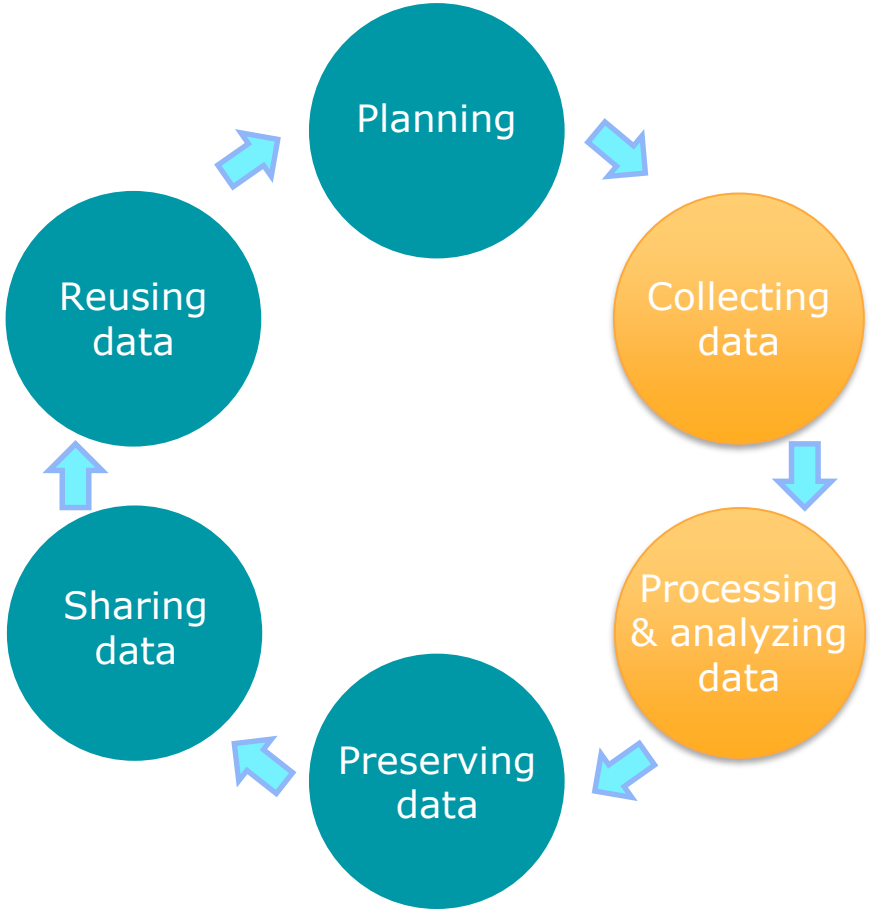
In fact, all these issues can be solved with good RDM practice.

Tips: be prepared; have “data champion” with you; design a website (with Q&A); promote RDM use cases on workshops, website and/or social media

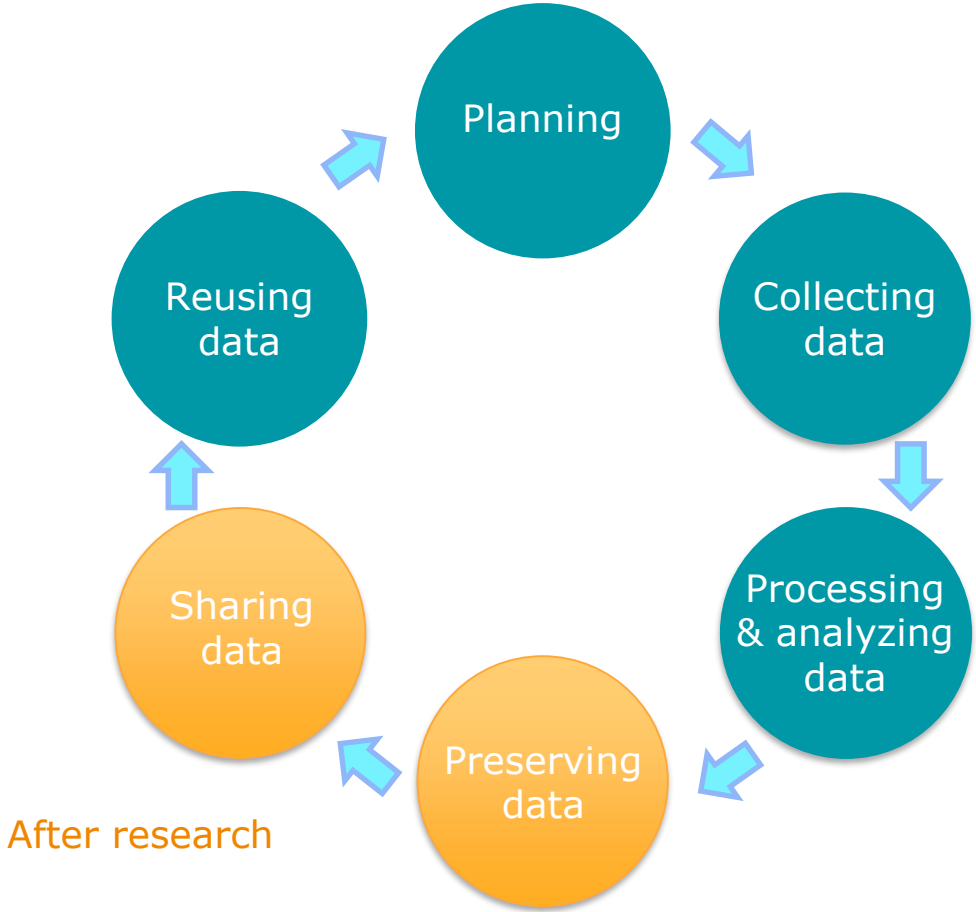
Research data lifecycle

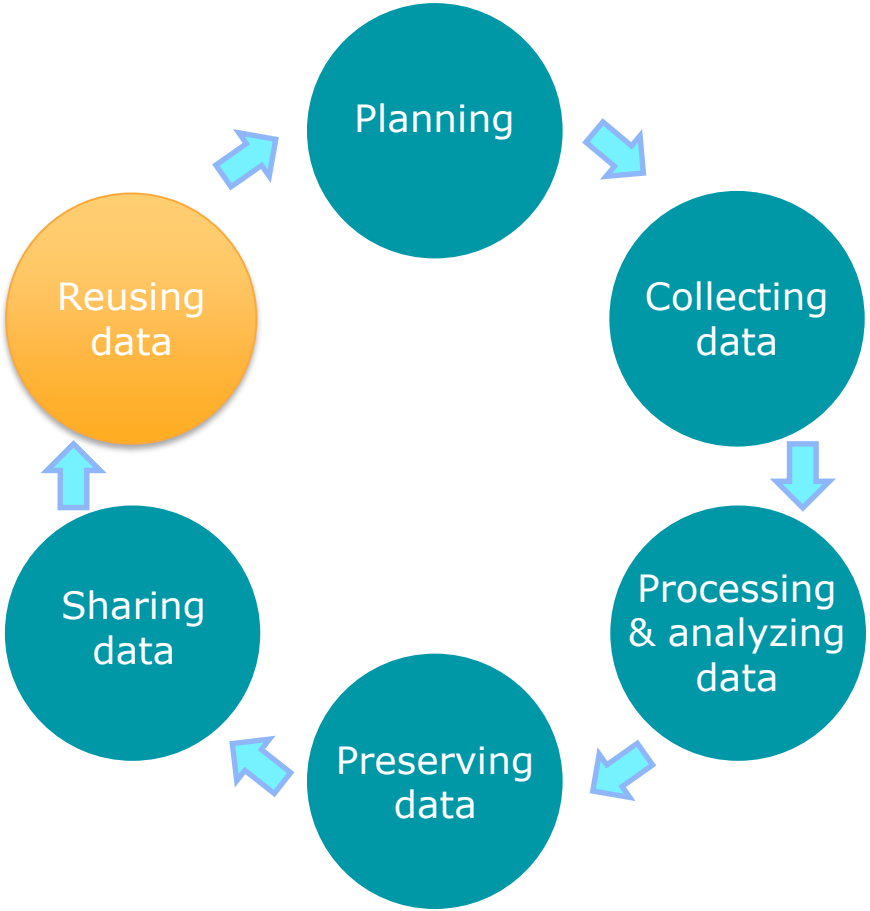



Before research
Planning
Collecting data
Processing & analyzing data
Preserving data
Sharing data
Reusing data



During research


Planning
Collecting data
Processing & analyzing data
Preserving data
Sharing data
Reusing data
After research


Planning
Collecting data
Processing & analyzing data
Preserving data
Sharing data
Reusing data

Data lifecycle: Planning phase

Data Management Plan (DMP)

- Design research
- Policy compliance
- Identify existing data sources
- Data collection and processing
- Data security measures
- Data management cost
- Long-term preservation and sharing

Photo by Jess Bailey on Unsplash

Data lifecycle: Collecting data phase

Collecting and/or generating data

- Organizing data
  - file and folder naming
    - agree on terminology
    - separate words with _ (underscore)
    - date format - ISO 8601: YYYYMMDD
  - folder structure
  - version control
    - include version number: v1, v2_1, final

Data lifecycle: Collecting data phase

- Metadata
  - standardized, structured, human- and machine-readable
  - resources: RDMkit (Elixir-Europe), Metadata MOOC
  - standards: FAIRsharing, RDA metadata directory
  - tools: Dublin Core generator, ENA checklist, …
- Documentation
  - ELN – electronic lab notebooks (*Jupyter Notebook*)
  - README file
    - https://guides.lib.uci.edu/datamanagement/readme
  - GitHub
    - can be used for version control
  - anything that can provide additional information on data

Data lifecycle: Processing & analizing data phase

Analysis and interpretation of data

- Anonymize sensitive data
  - Amnesia (OpenAIRE)
- Data cleaning, validation and quality checking
  - OpenRefine
- Describing and documenting
- Storing, organizing, version control

Data lifecycle: Preserving data phase

Data needs to be safe and preserved during and after the research.

Preservation during the active phase:

- backup plan
  - 3-2-1 rule
  - recommend cloud storage

backup and long-term preservation are NOT the same

- access control
  - password protection
  - encryption

Image by John from Wikimedia Commons under CC-BY license

Data lifecycle: Preserving data phase

Long-term preservation :

- Data selection and appraisal
- Migrating to open and sustainable formats
  - DANS list of preferred formats
- Archive both data and documentation

Data lifecycle: Sharing data phase

- Choose discipline specific or generic data repository,
  - rather than publishing data as supplementary in journal articles
  - re3data.org
  - certified repositories: CoreTrustSeal, ISO 16363, Nestor Seal;
  - long standing and solid user base – Zenodo
- Provide metadata and documentation
- Choose license and access control
  - look for funder requirements
  - licenses for data and for software
- Data paper

Training tips

- Avoid teaching everything about RDM in one event.
    - break it down in smaller pieces (DMP, FAIR, data publishing);
- have in mind final goals of RDM: data is secured and preserved, easy to find, understand and reuse;
- as soon as possible deal with misconceptions, especially about unauthorised data usage and open data;
- Data champions. Use the researchers that had previous experience with data management.
- Explain terminology used in RDM;
- Present use cases;
- Have recommended materials with you. DIY (in national language);
- stay informed about the policies and requirements;
- stay informed on new developments.

Questions?

Institut za nuklearne nauke Vinča

Obrad Vučkovac

University of Belgrade

Vinča Institute of Nuclear Sciences - Library

ORCID: 0000-0001-5616-2680

Except as otherwise noted, this presentation is licensed under the Creative Commons Attribution 4.0 International Licence.
To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.