Research data management

Obrad Vučkovac
University of Belgrade
Vinča Institute of Nuclear Sciences

Training challenges

- Complexity of the topic
  - research data lifecycle, data management plan, FAIR principles, Data publishing and licenses, Open Data
- Terminology
- Misconceptions (Open Data, data scooping)
- Good practice, but not always mandatory (policies)
- 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

The aim of data management is data that are:

- secured and preserved;
- findable, understandable and reusable.

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 need to be ready for both

humans and machines

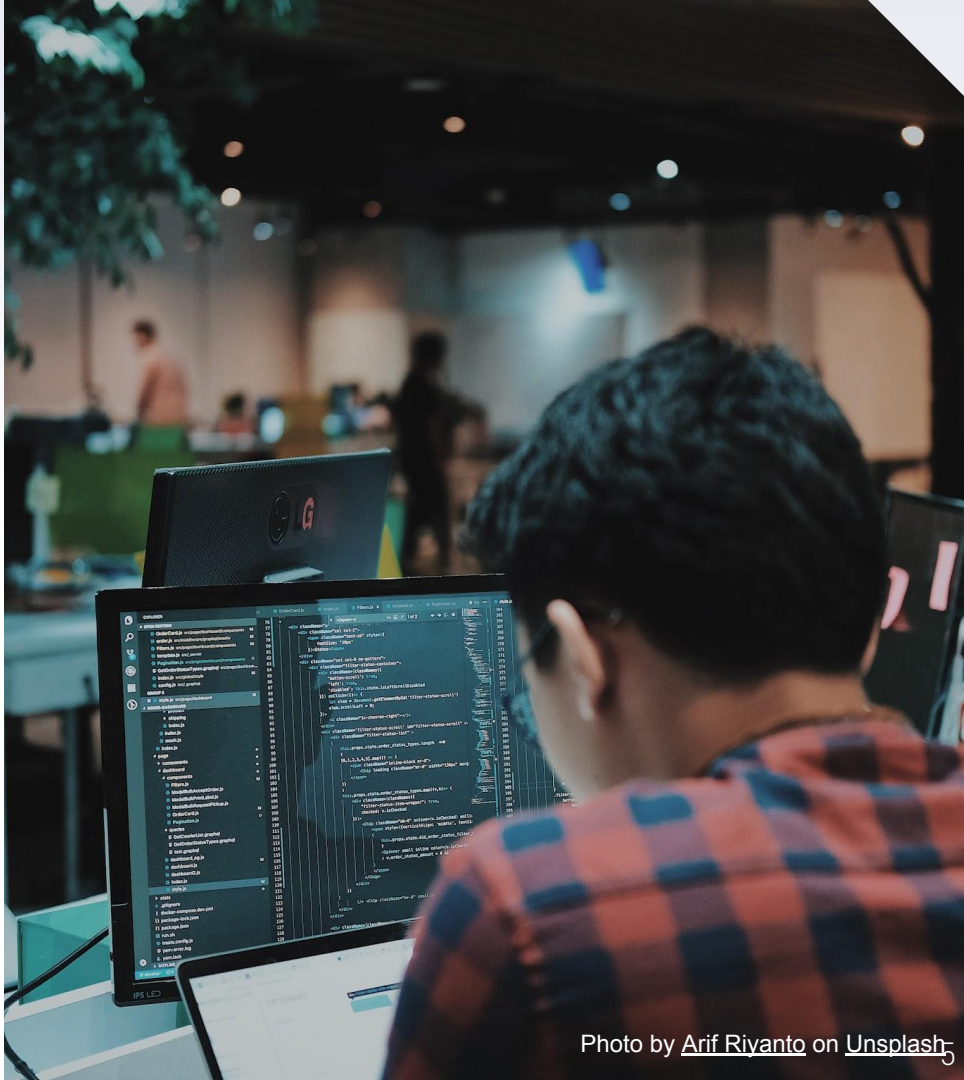

Photo by Arif Riyanto on Unsplash

Benefits

Reasons why researchers need a good RDM practice:

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

Common misconceptions

Common misconseptions regarding RDM:

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

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

During research (active phase)

After research

Research Data Lifecycle

Planning phase

Data Management Plan (DMP)

: a document specifying how data will be managed during and after the research project.

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

During research: Organizing data

Organizing data

- file and folder naming
  - agree on terminology
  - separate words with _ (underscore)
  - agreed date format (e.g. YYYYMMDD - ISO 8601 standard)
- folder structure
- version control
  - automatic or manual versioning system
  - include version number: v1, v2_1, final

EXAMPLES!

File formats

Best practice:

- non proprietary / open formats (e.g. CSV over XLSX; ODT over DOCX)
- common usage by research community
- lossless compression

DANS File formats

Recommended formats - UK Data Service

During research: Documentation

Documentation

: any descriptive and contextual information needed to find, understand, and (re)use research data.

- (electronic) lab notebooks (e.g. Jupyter Notebook)
- README file
  - https://guides.lib.uci.edu/datamanagement/readme
- codebooks, instruments info, calibration etc.
- gitHub, GitLab
  - can be used for version control
- … anything that can provide additional information on data

Metadata

: 'data about data', used to describe and annotate data.

A highly structured, machine-readable form of data documentation.

Use Library OPAC or database with faceted search as an example of information discovery.

- use standardized metadata whenever possible
  - standards: FAIRsharing, RDA metadata directory
- tools: Dublin Core generator, DCC Metadata Tools etc.
- controlled vocabularies

During research: Data cleaning and analysis

Data analysis and interpretation

- Data cleaning, validation and quality checking
  - OpenRefine
- Anonymize sensitive data
  - Amnesia (OpenAIRE)
- Describing and documenting
- Storing, organizing, version control
- **Include these processes in RDM costs**

Photo by Paul Calescu on Unsplash

During research: Data preservation

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

Preservation during the active phase:

- backup plan
  - e.g. 3-2-1 rule
  - recommend cloud storage (Google Drive, OneDrive etc.)

Backup and long-term preservation are NOT the same

- access control
  - password protection
  - encryption

After research: Data preservation

Long-term preservation:

- Data selection and appraisal
  - specify how long to preserve data
  - specify what data to preserve
  - specify access and responsibilities
- Migrating to open and sustainable formats
  - DANS list of preferred formats
- Archive data and documentation
  - metadata is available
- Specify terms of use

Photo by Paul Calescu on Unsplash

Photo by Ula Kuźma on Unsplash

After research: Data sharing and publishing

Publish data in a repository

- Select a repository:
  - domain (discipline) specific
    - re3data.org
  - general repository (e.g. Zenodo, Figshare)
  - institutional data repository
- Enrich with metadata
- Publish in data journal
- Choose a licence

Criteria for repository:

- persistent identifier
- guaranted long-term preservation
- costs (yes or no)
- type of access does it allow restricted access?
- licence choice?
- is it certified? (e.g. CoreTrustSeal)

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.

THANKS!

Any questions?

Obrad Vučkovac

University of Belgrade

Vinča Institute of Nuclear Sciences - Library

ORCID: 0000-0001-5616-2680

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Illustrations by Sergei Tikhonov
- Photographs by Unsplash