Good Research Data Management: From Theory to Practice

Volker U. Schwartze

Outline

1. What is research data management, why is it important and how do I do it?
   - Definitions and concepts
   - Requirements and challenges

2. Where can I get support regarding research data management?
   - Services provided by the FSU Jena
   - Additional information sources

Research Data Management – The What, Why and How

What is (Research) Data?

- No clear/consistent definition (disclipine-specific definitions)

- Set of values/information/findings that are the result of observations, measurements, surveys etc.[1]

[1] Duden (https://www.duden.de/woerterbuch)

What is (Research) Data?

- German Research Foundation (Deutsche Forschungsgemeinschaft, DFG):
  - Measured values
  - Laboratory findings
  - Survey data
  - Audiovisual information
  - Texts
  - Objects from a collection/samples
  - Test methods (e.g. questionnaires, software,…)

DFG (2015) „DFG Guidelines on the Handling of Research Data“

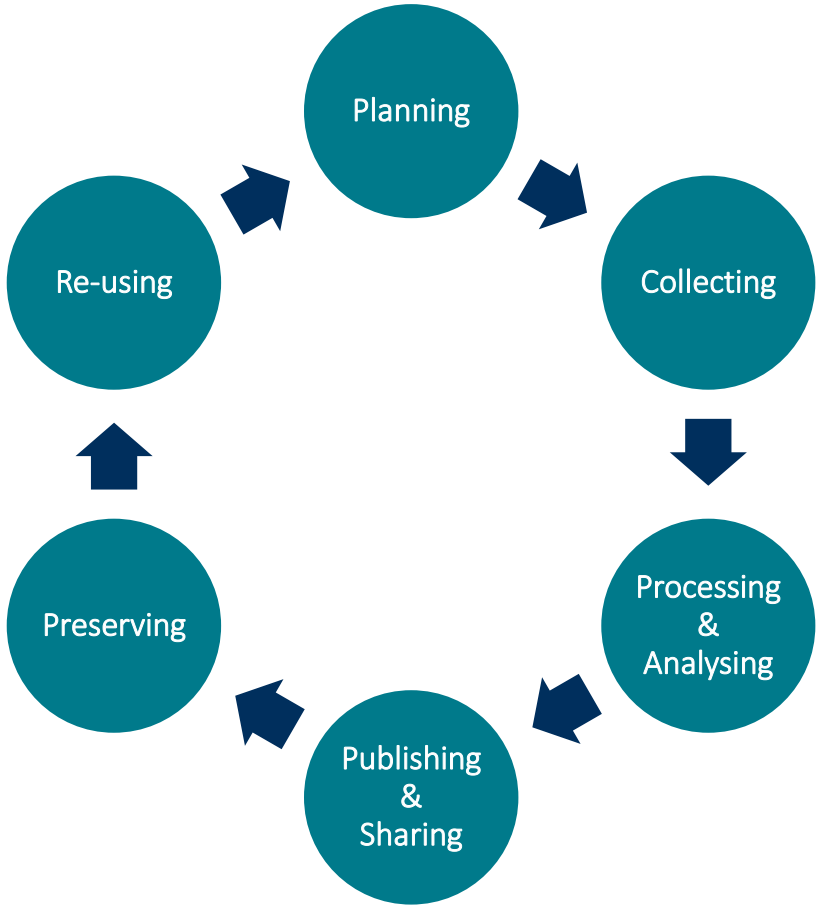
What is Research Data Management?

Data management refers to all aspects of creating, housing, delivering, maintaining, and archiving and preserving data.
It is one of the essential areas of responsible conduct of research.[1]

The management of research data includes the development of a demand-oriented planning to use, re-use, publish and archive data efficiently.[2]

[1] Mantra 2017: 5, [2] Forschungsdaten.info

The Research Data Life Cycle

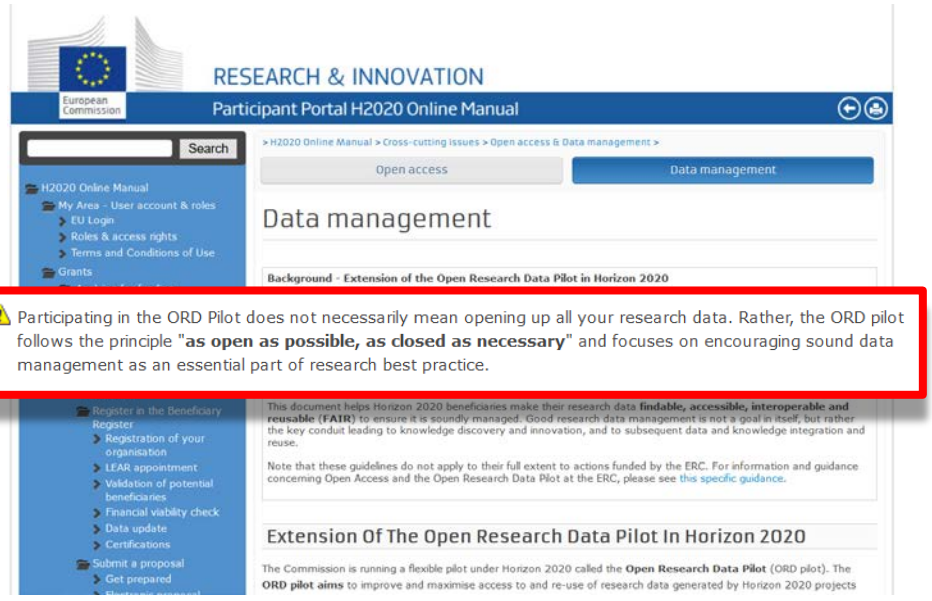


Why Research Data Management?

Good Scientific Practice

Policies

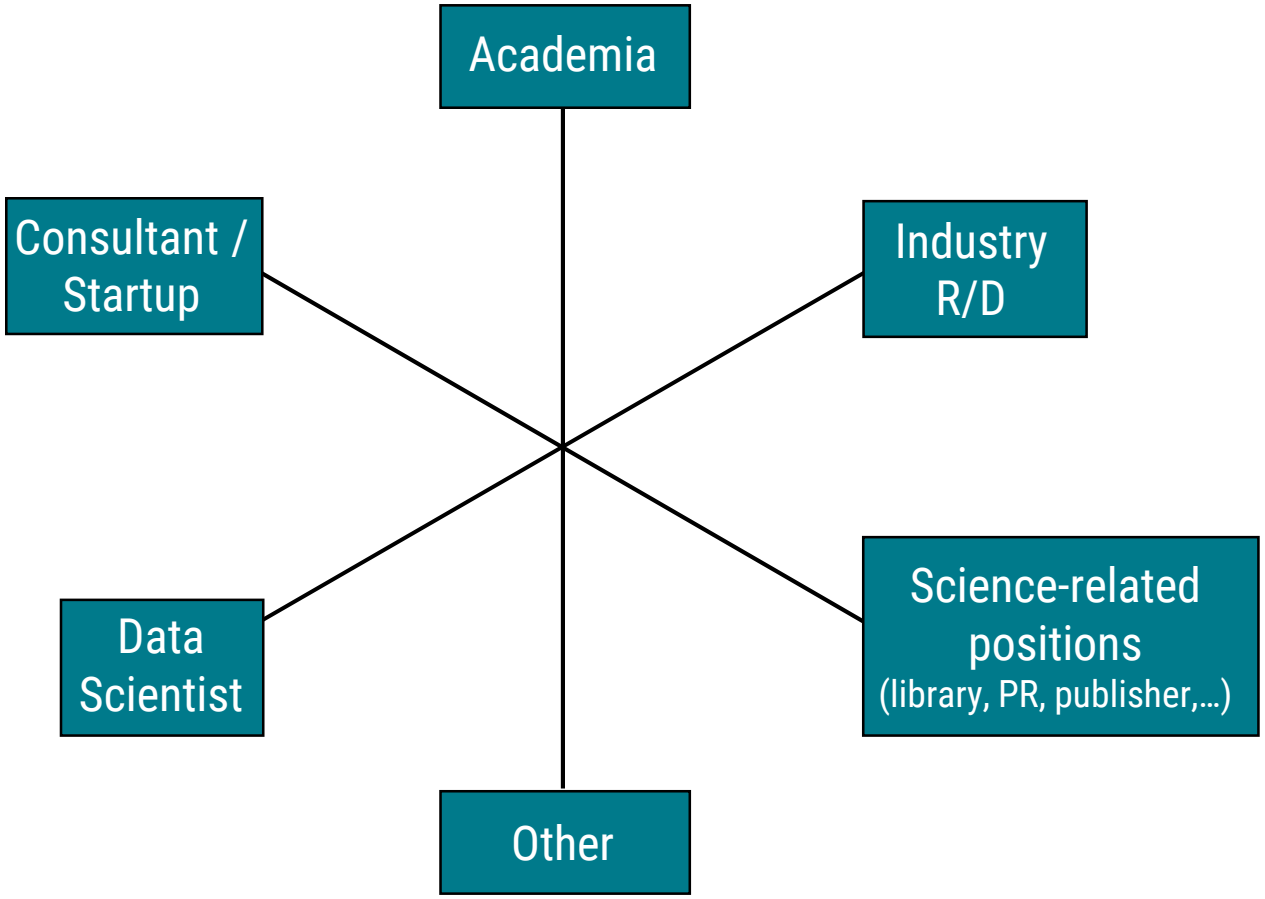
Funding agencies



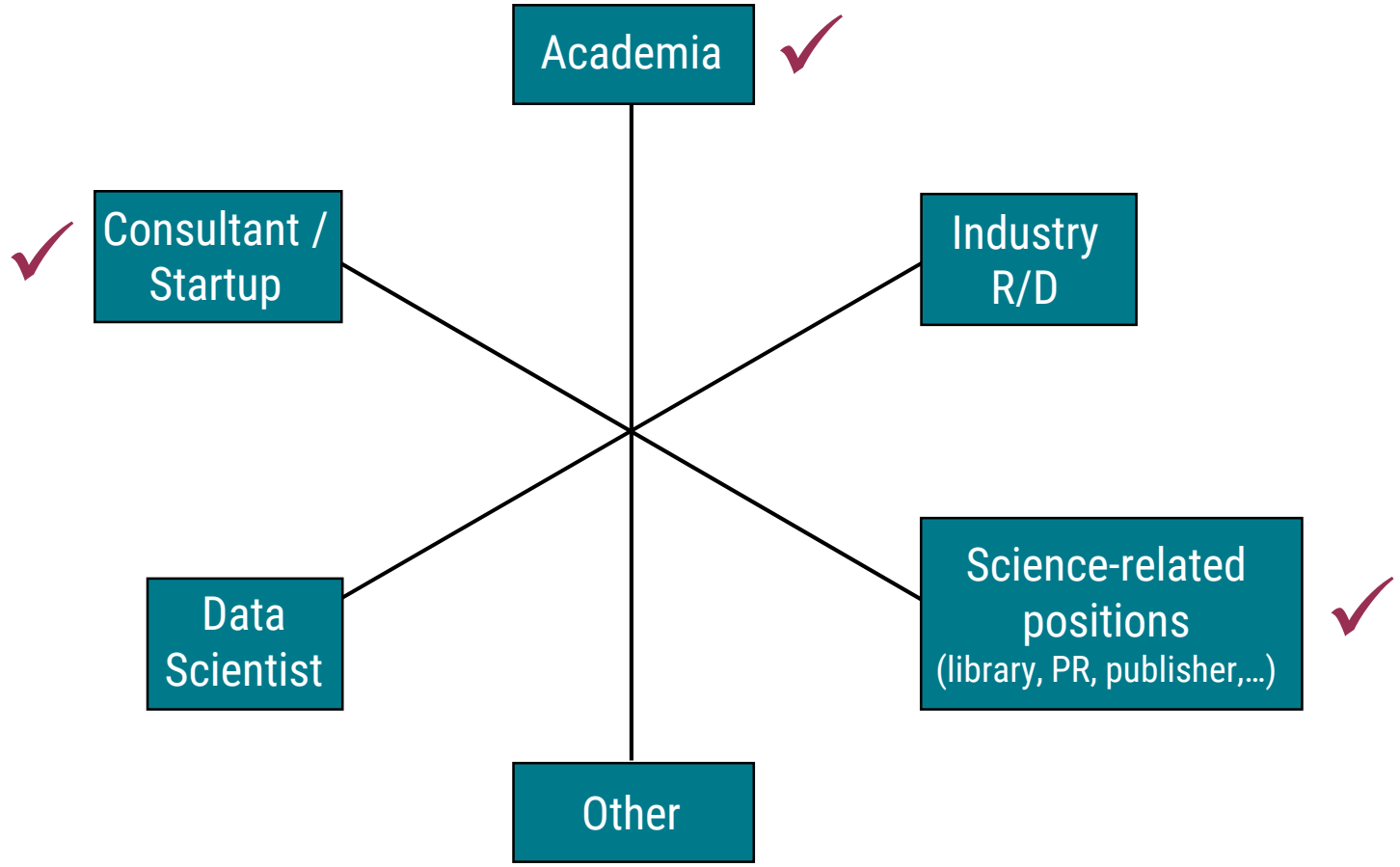
Why Research Data Management?

- Demanded by funding agencies and institutions
- Good scientific practice
- Verifiability and reproducibility
- Saves time and resources
- Data security and prevention of data loss
- Re-use of data (no need for unnecessary repetitions)

Importance of Research Data Management Skills

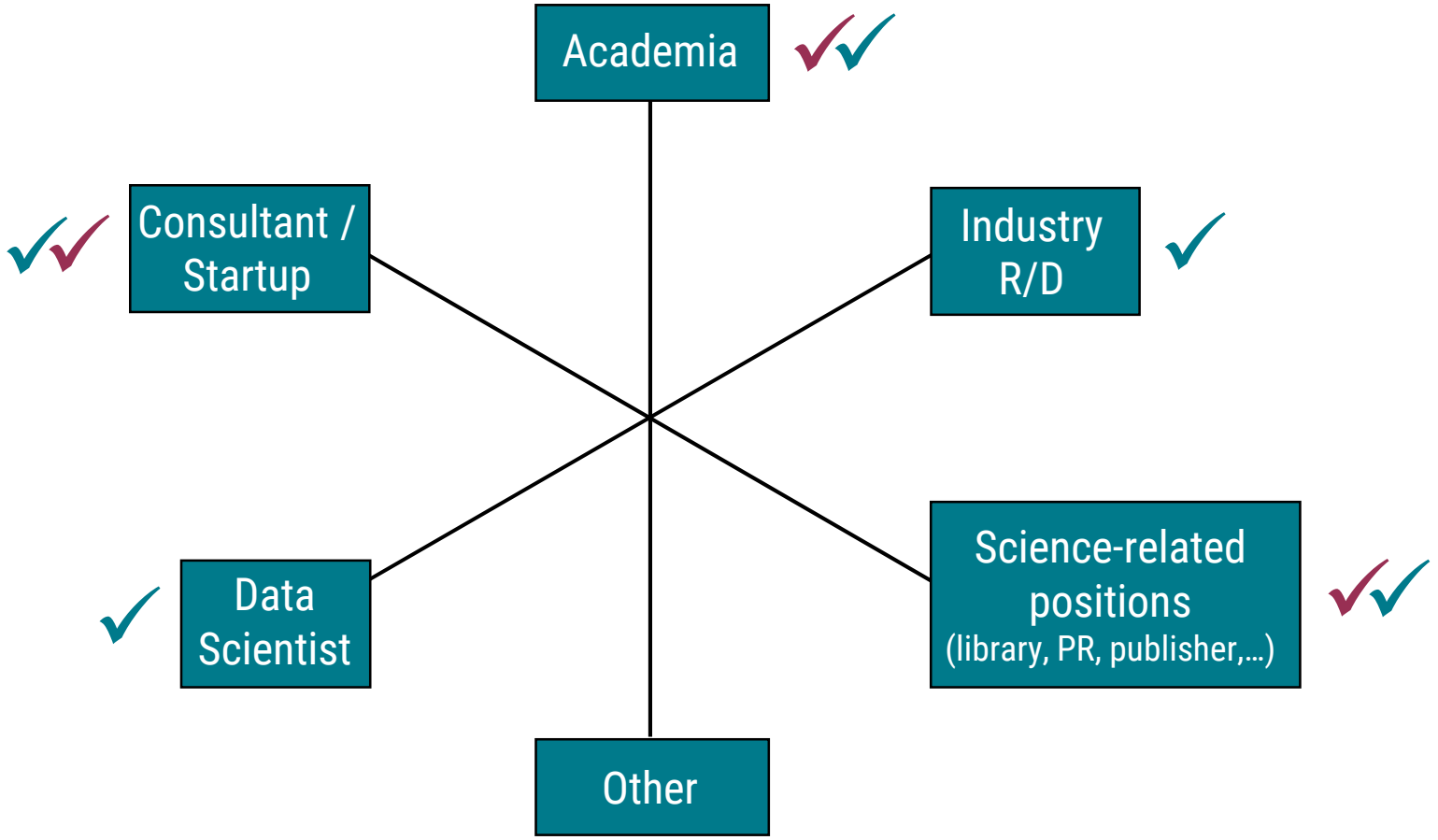


Importance of Research Data Management Skills



Knowledge about Open Data/ Reproducible Research

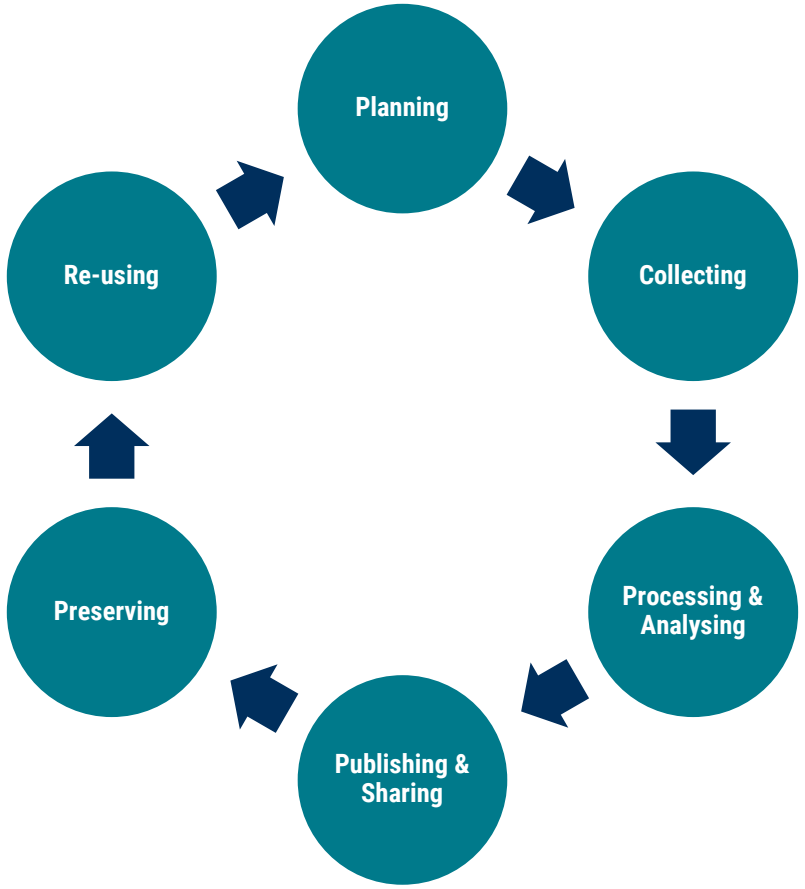
Importance of Research Data Management Skills



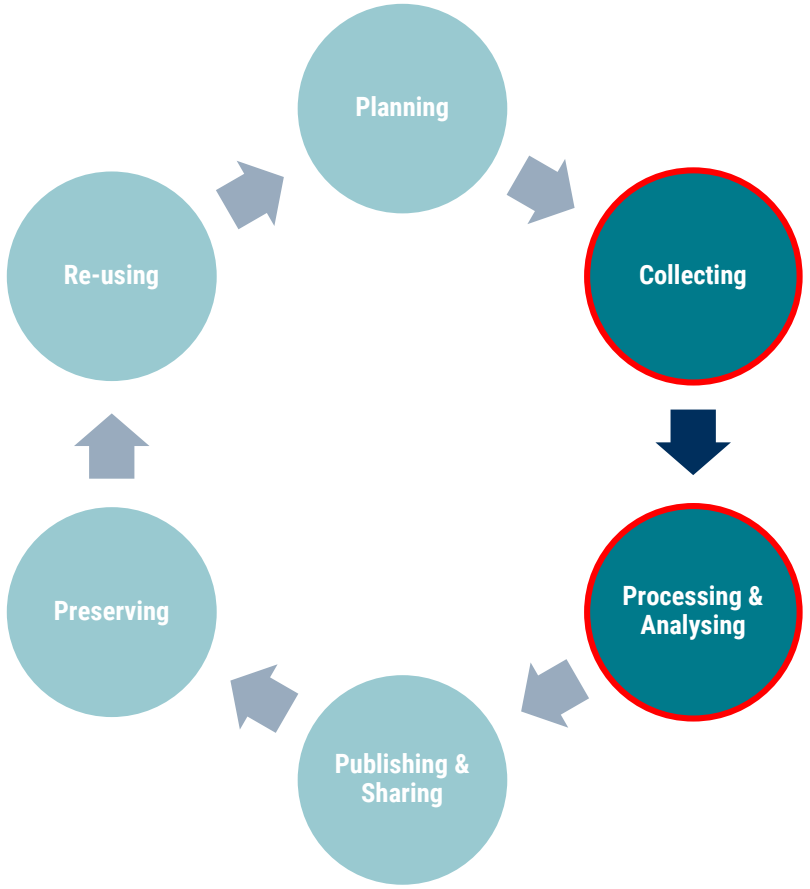
Knowledge about Open Data/ Reproducible Research

Data Management Skills

Research Data Management – When and How?



Research Data Management – When and How?



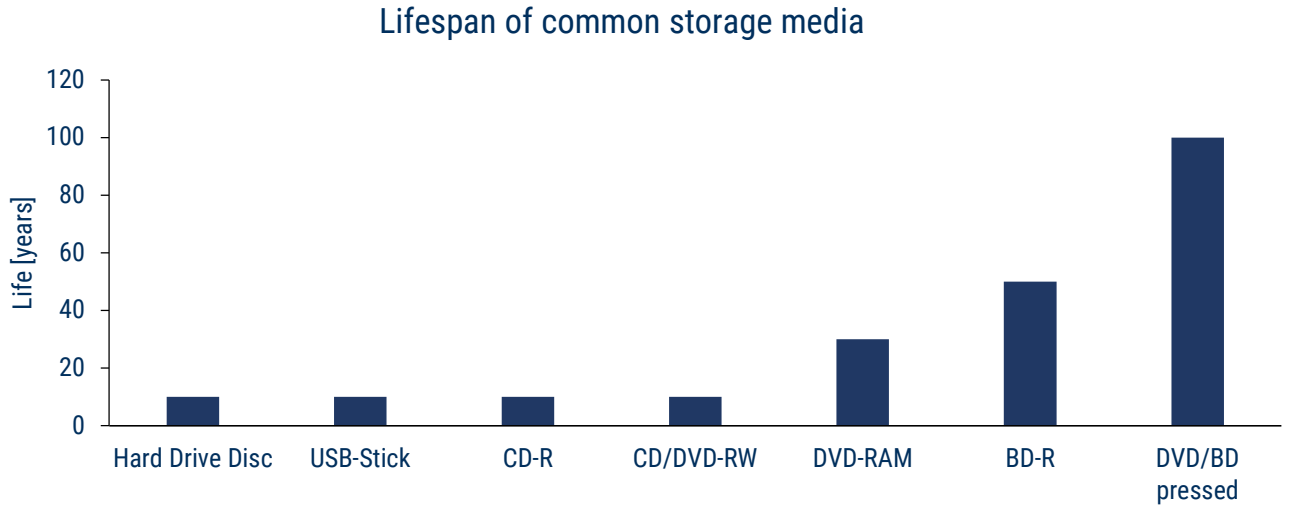
During the project

How will the data be managed during the project?

- Storage capacity & back-ups (frequency, media)

Storage and Backup

- 3-2-1 rule (3 copies, 2 types of hardware, 1 offsite)
- Regular backups
- Lifetime of storage restricted → ever 2-5 years new hardware

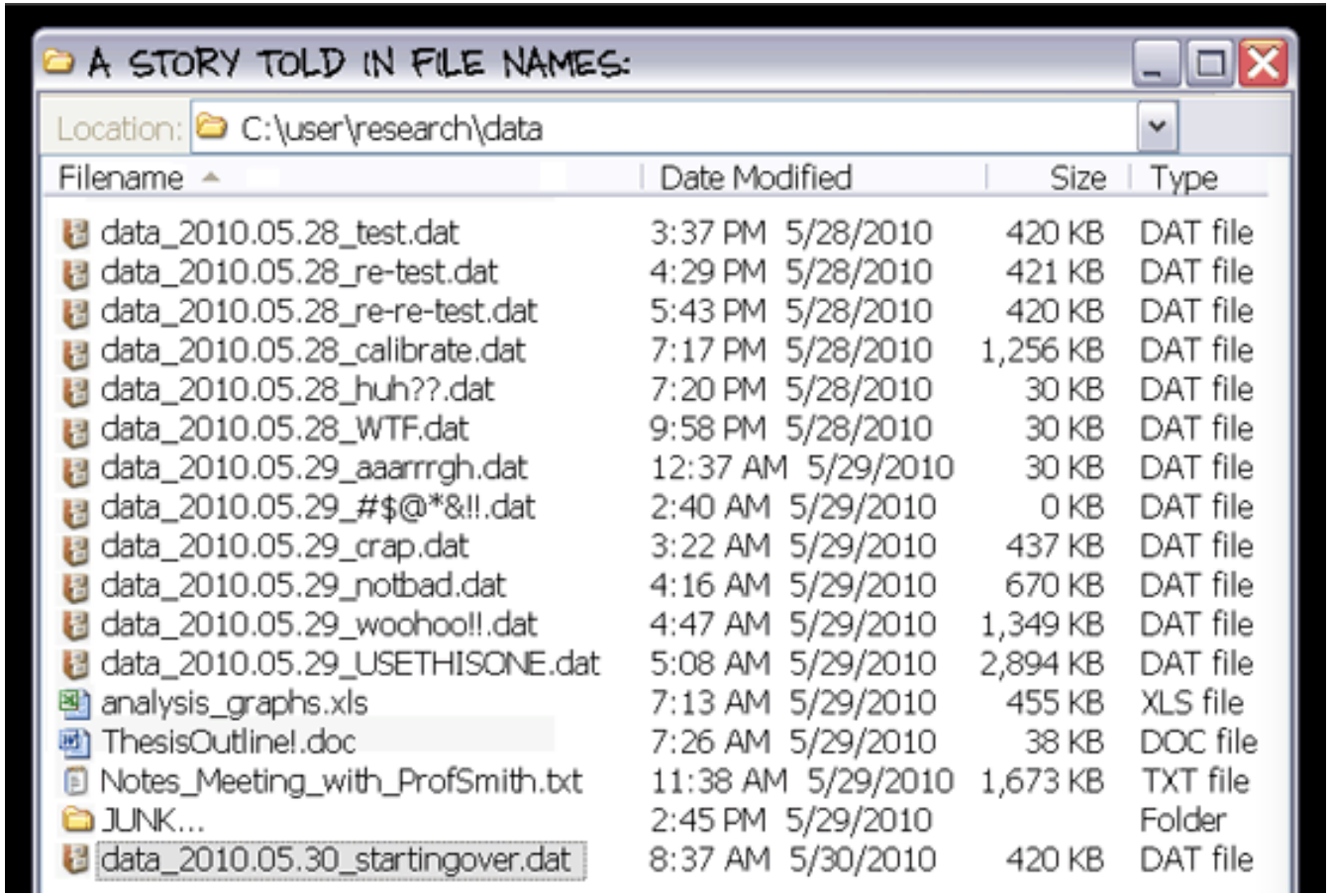


J. Rex (2019), DOI: 10.5281/zenodo.2579580

How will the data be managed during the project?

- Storage capacity & back-ups (frequency, media)
- Data organisation and structure (files, database)
- Naming convention for files and folders

File Naming gone wrong!



How will the data be managed during the project?

- Storage capacity & back-ups (frequency, media)
- Data organisation and structure (files, database)
- Naming convention for files and folders
- Version control (e.g. Git)
- Security & access control
- Metadata creation

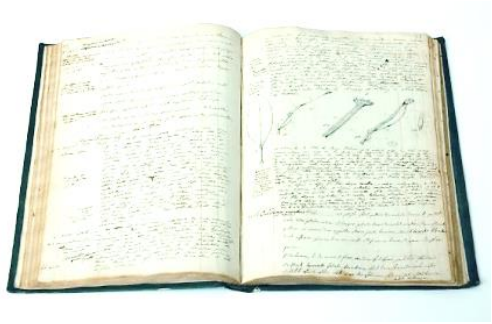
What are metadata?

- Data about data
- Contextual information about dataset:

  - Who?
  - What?
  - When?
  - Where?
  - Why?
  - How?

Document your work/data,
while you work on them!

How are metadata recorded?



Lab book

⬇

Human readable

```
<?xml version="1.0"?>
<dwr:DarwinRecordSet
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://rs.tdwg.org/dwc/dwcrecord/  http://rs.tdwg.org/dwc/xsd/tdwg_dwc_classes.xsd"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
    xmlns:dwr="http://rs.tdwg.org/dwc/dwcrecord/">
    <dcterms:Location>
        <dwc:locationID>http://guid.mvz.org/sites/arg/127</dwc:locationID>
        <dwc:country>Argentina</dwc:country>
        <dwc:countryCode>AR</dwc:countryCode>
        <dwc:stateProvince>Neuquén</dwc:stateProvince>
        <dwc:locality>25 km al NNE de Bariloche por Ruta 40 (=237)</dwc:locality>
    </dcterms:Location>
    <dwc:Occurrence>
        <dcterms:type>PhysicalObject</dcterms:type>
        <dcterms:modified>2009-02-12T12:43:31</dcterms:modified>
        <dcterms:rightsHolder>Museum of Vertebrate Zoology</dcterms:rightsHolder>
        <dcterms:rights>Creative Commons License</dcterms:rights>
```

Digital metadata (XML)

⬇

Human and machine-readable

+ searchable
+ structured
+ standardised

Is there a specific standard for metadata?

Is there a specific standard for metadata?

Broadly applicable standards

- ISO 639 Code for the representation of the names of languages. e.g. eng – English; ger/deu – German
- ISO 8601 Codes for the representation of date and time. e.g. 2014-06-19T13:15:30Z

Is there a specific standard for metadata?

Broadly applicable standards

- ISO 639 Code for the representation of the names of languages. e.g. eng – English; ger/deu – German
- ISO 8601 Codes for the representation of date and time. e.g. 2014-06-19T13:15:30Z

Domain-specific standards

- Consistent terms/definitions/nomenclature
- Domain-specific vocabularies and ontologies

The bottom line is that ...

- …the metadata accompanying a data set should be written for a user 20 years into the future.
- …clear storage concept (& back-up), folder structures and file naming conventions increase the findability of your data.

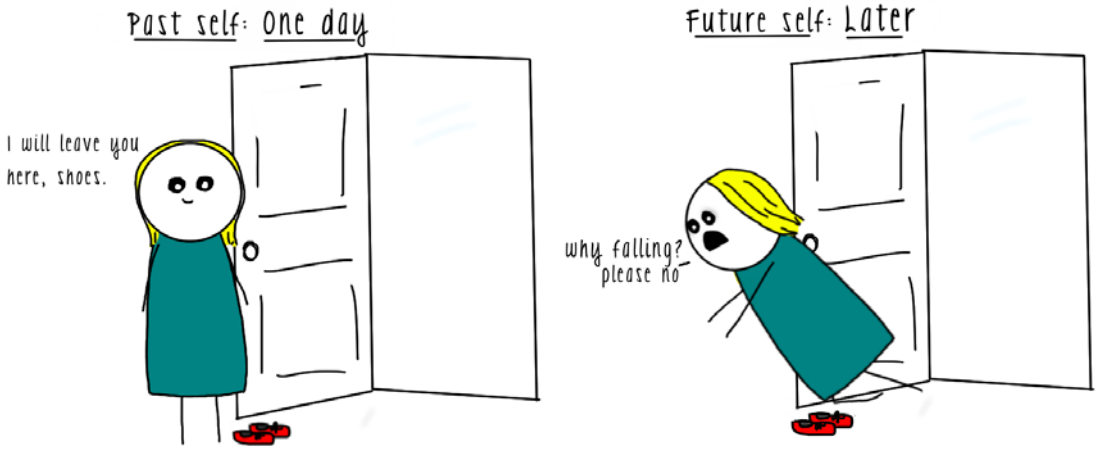
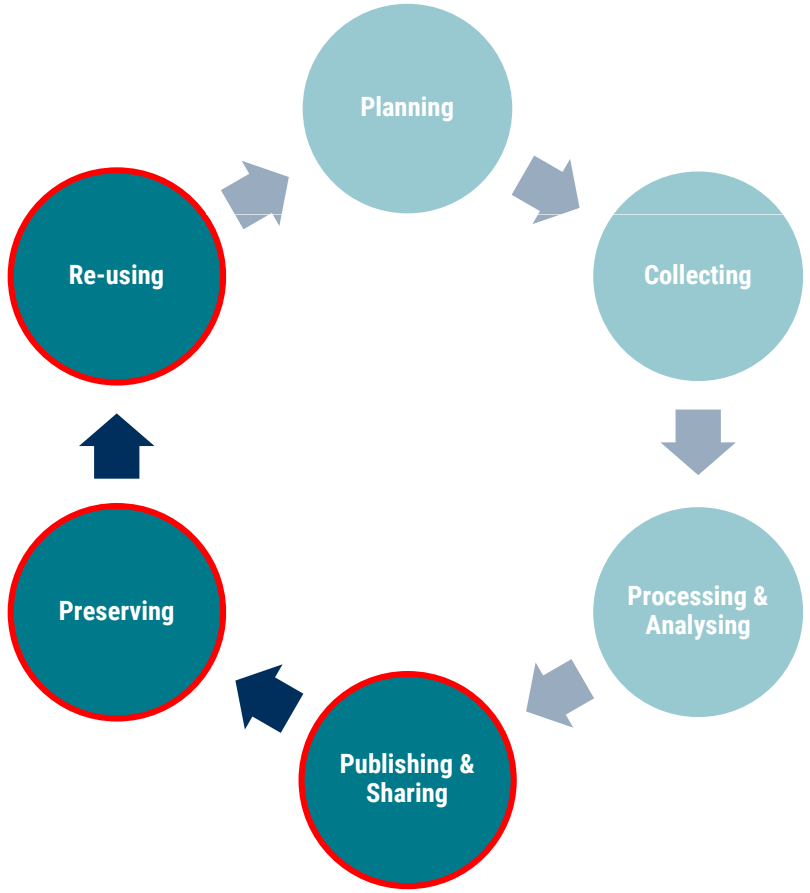


Image CC-BY-NC-ND: past self and future self, by www.thisisnotthatbog.com

Research Data Management – When and How?



After the project

How will the data be managed after the project?

- ALL data have to be stored for at least 10 years (Good Scientific Practice)

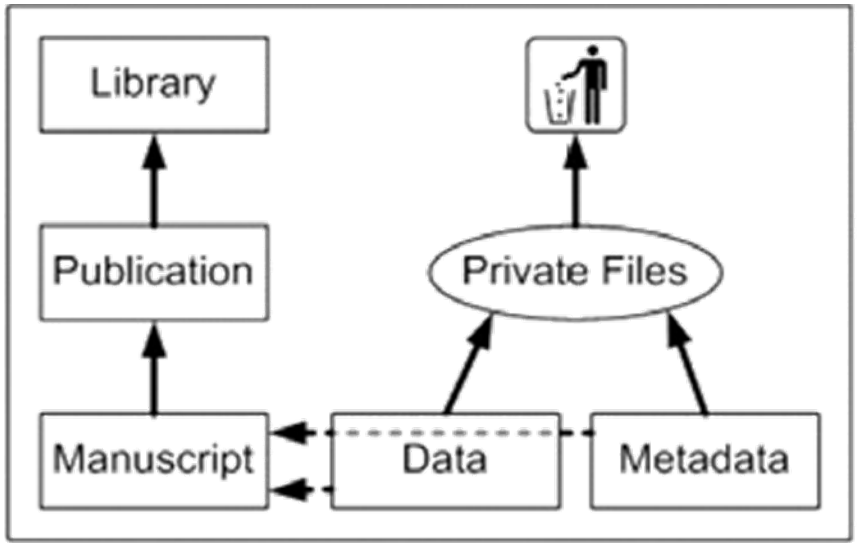
- Data selections for long-term archiving (for eternity):
  - Data that can not be reproduced (unique measurements/observations in time and space)
  - Data that can only be reproduced with a lot of effort (e.g. cost/computer intensive analysis)
  - Data that are the basis for publications

How and where do I preserve my data?

- Data publication or data archiving?
- Maintenance of existing infrastructure?
- Migrating data into a (public) data repository?
- Access level of data (open, limited user group, closed) → sensible/personal data!
- Embargo periods?

➡ Funding organisation encourage/request data publication

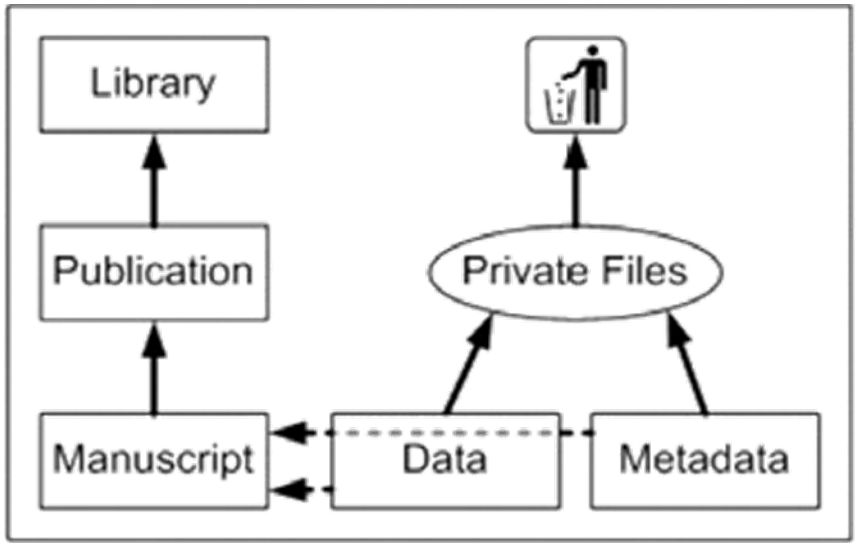
Data Preservation/Publication: Ideal vs. Reality



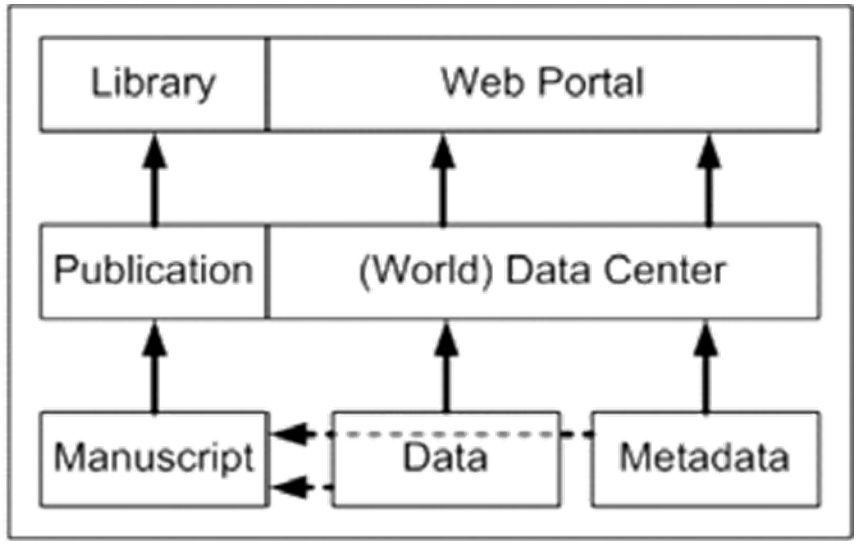
Reality

Rümpel 2011:26

Data Preservation/Publication: Ideal vs. Reality



Reality



Ideal

Rümpel 2011:26

Data Preservation/Archiving: Things to keep in mind

- At least archive folder, better in a specific archive server
- Challenges in long-term archiving:

Maintenance of data content	• Bitstream preservation
Maintenance of functionality	• Migration • Emulation
Maintenance of usability	• Documentation • Metadata

J. Rex (2019)

How can I publish my data?

- Supplement of an article in scientific journals
- Data journal
- Data repository

Which repository should I use?

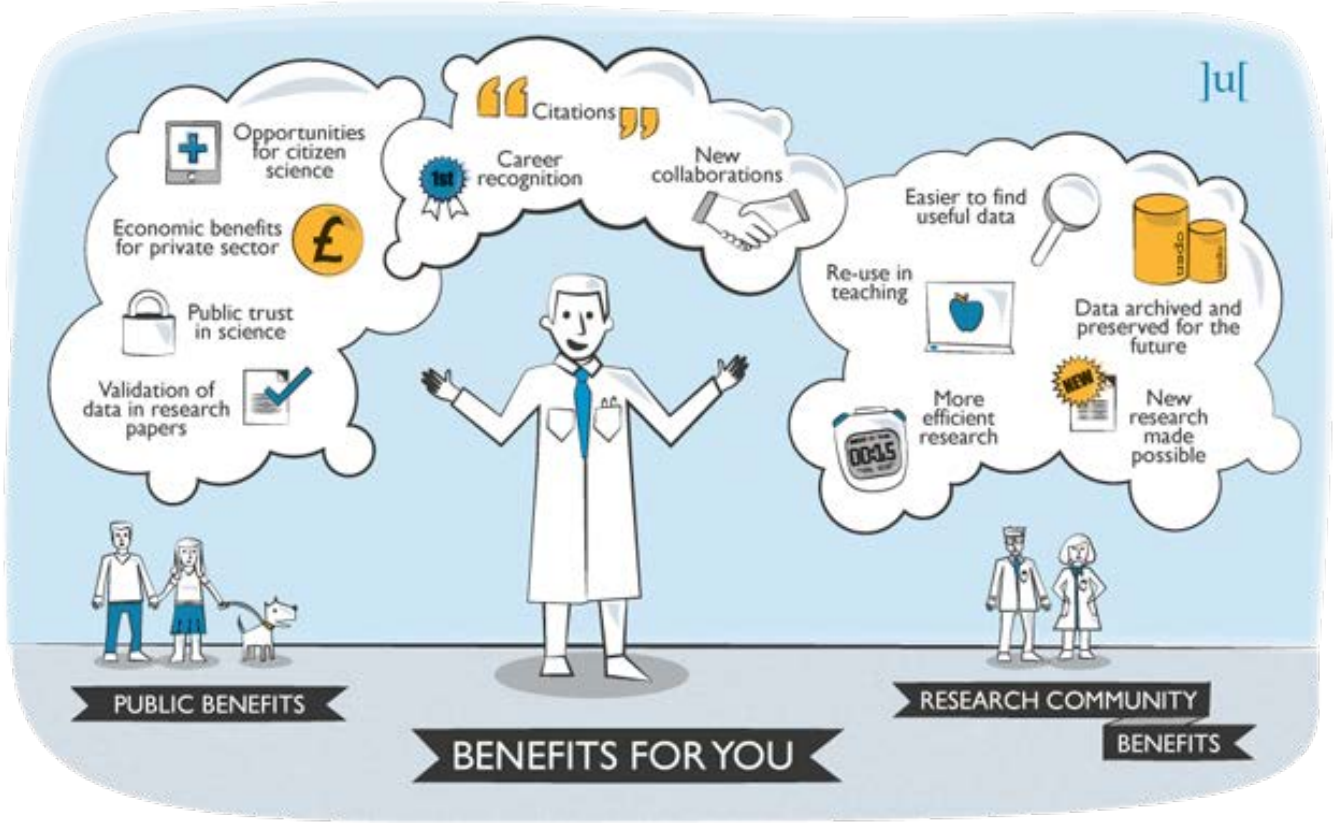
- Search for suitable repositories (e.g. re3data.org)

- Domain/topic-specific repositories (preferable)
- Generic or institutional repositories (if no specific repository is suitable)

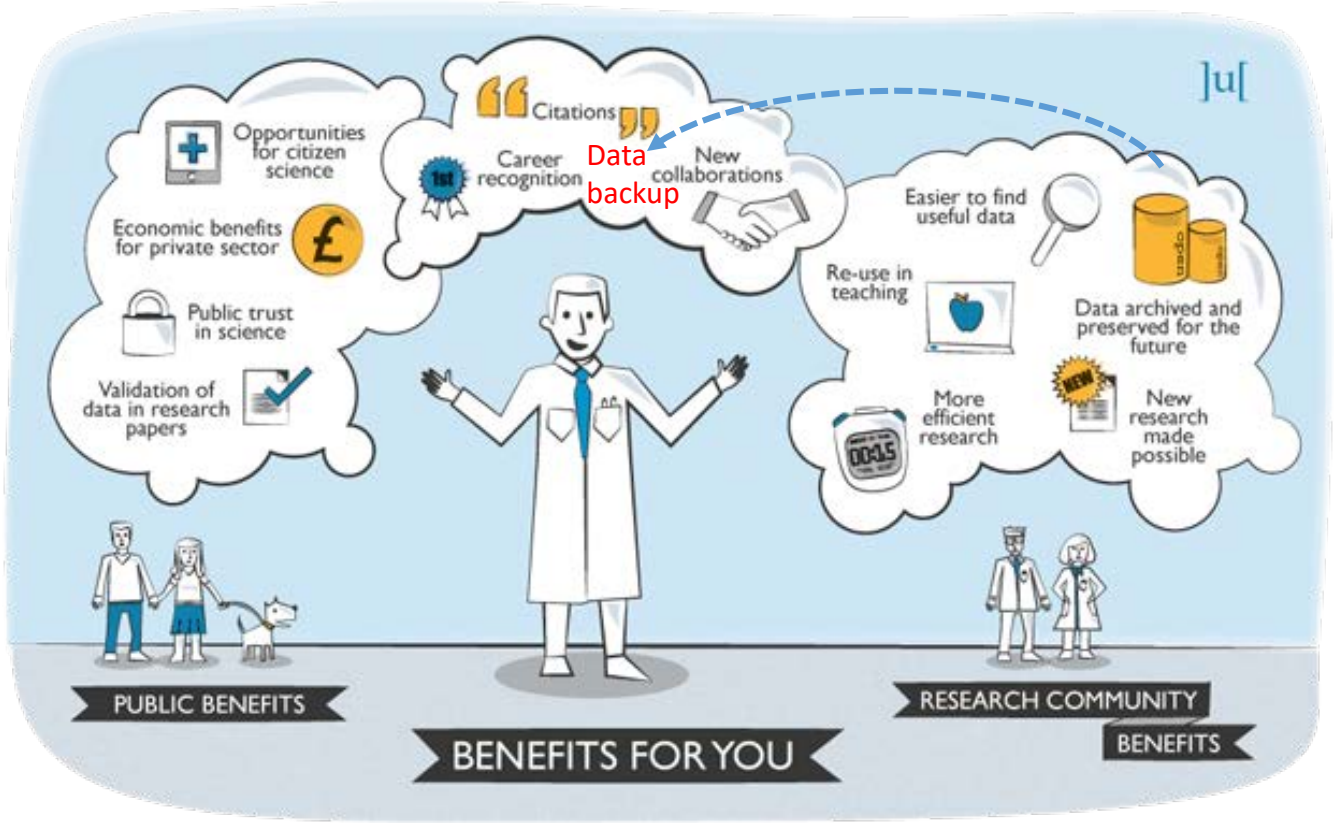
zenodo

dbt
Digitale Bibliothek Thüringen

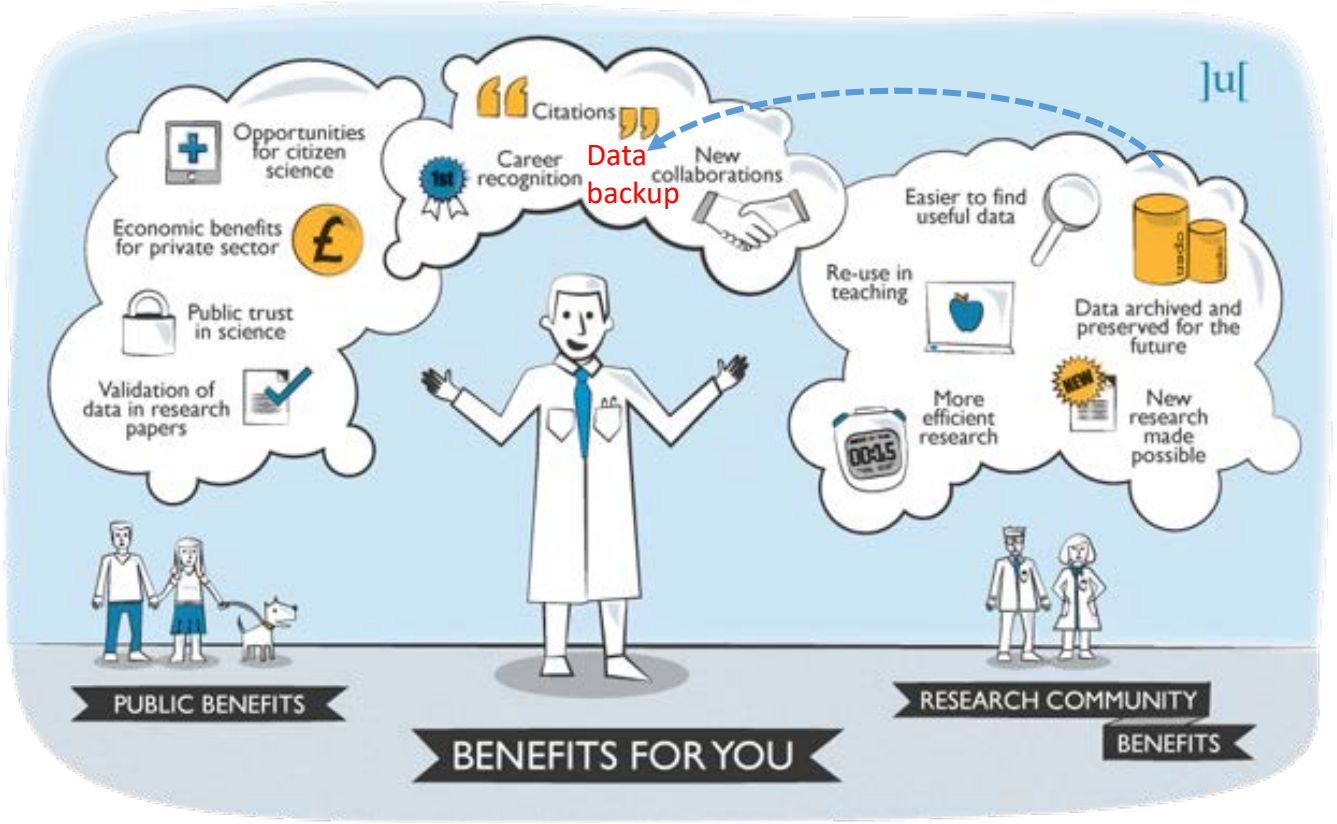
Why should I publish my data?



Why should I publish my data?

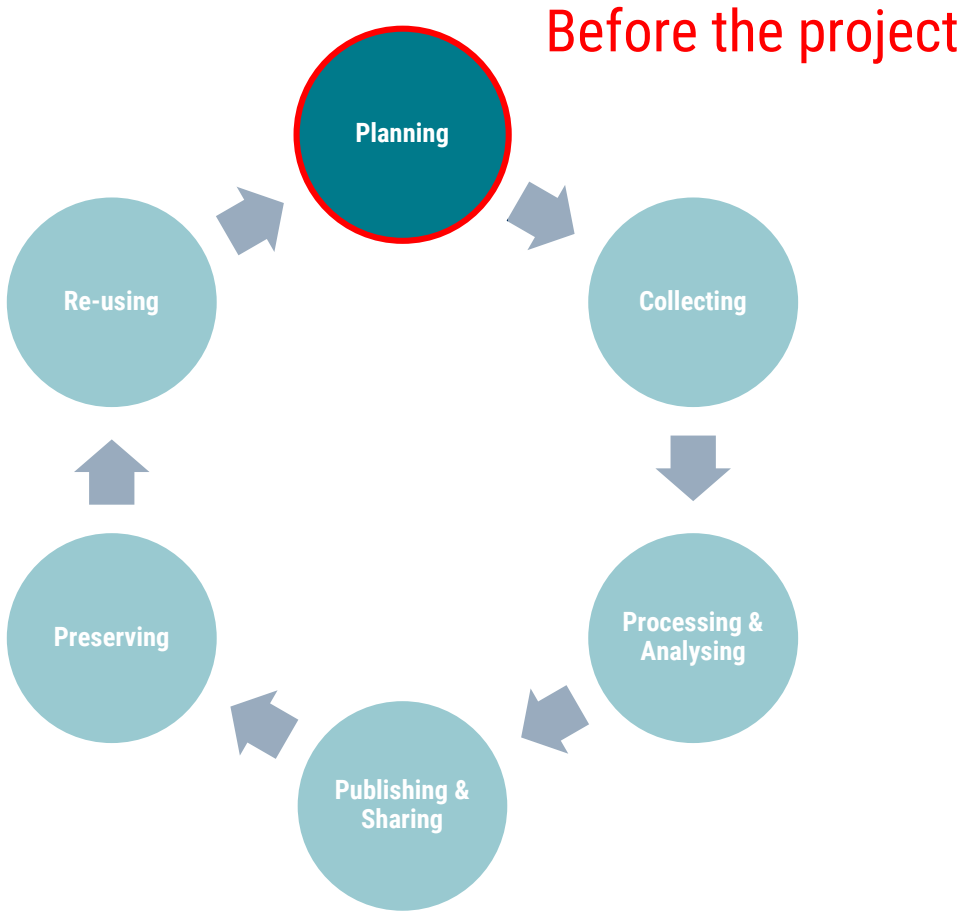


Why should I publish my data?



& required by funding agencies, institutions and journals

Research Data Management – When and how?



What needs to be planned before the start of the project?

To ensure that no aspect of research data management

is forgotten during the project

Research Data Management Plan (DMP)

What is a DMP?

- (Formal) document that outlines how you will handle data during and after the project (incl. financial and legal considerations)
- Information about all stages of the data life cycle (incl. required resources)
- *Living document*

Why is a DMP needed?

- Provides a guideline during the project
- Defines responsibilities
- Several funding agencies request it as part of a grant proposal (e.g. EU)

What is FAIR Data?

Findable

- Persistent identification
- Metadata include identifier of data
- Registered or indexed in searchable resource

Accessible

- Retrievable by identifier (e.g. repository)
- Metadata available (even if data are not available anymore)
- Different options (e.g. open, embargo, limited user group, closed)

Interoperable

- Use of standards (vocabularies, metadata standards)
- Use of standard file formats

Re-usable

- Clear licensing

Modified after Sangya Pundir (CC BY-SA) https://www.go-fair.org/fair-principles/

Where can I get support regarding research data management?

Where can I get support regarding research data management?

Research Data Management Helpdesk

- Advice on all topics of RDM (before, during and after projects)
- Information (material) regarding different aspects of RDM
- Trainings
- Support in development of data management plan
- Networking with other institutions within and outside the FSU Jena

Where can I get support regarding research data management?

Research Data Management Helpdesk

- Advice on all topics of RDM (before, during and after projects)
- Information (material) regarding different aspects of RDM
- Trainings
- Support in development of data management plan
- Networking with other institutions within and outside the FSU Jena

Universitätsrechenzentrum (computer centre)

- Storage space and backup
- Cloud services
- Software

Rechtsamt (legal office)

- Legal advice and forms (e.g. protection of personal/patient data)

Thüringer Universitäts- und Landesbibliothek (ThULB)

- Publication of research data (DBT)

The Research Data Management Helpdesk

- Graduate academy courses:
  - Humanities (winter semester)
  - Life-sciences (summer semester)
- Workshop and coaching (upon request)
- Individual consultation
- Information material on our website

Where can I get further information?

German speaking RDM website

European Collaborative Data Infrastructure

Research Data Management Training

Data Management Short Course for Scientists

Thank you for your attention!

Pleas quote as follows

Gerlach, Roman, Färber, Bettina, König-Ries, Birgitta, Schröter, Annett, Rzymski, Christoph, Schwartze, Volker, Steiner, Petra & Olena Tykhostup (2020) Good Research Data Management: From Theory to Practice. Presentation by Volker Schwartze, 30.01.2020, Promovierendentag der Friedrich-Schiller-Universität, Jena.