GERMAN FEDERAL INSTITUTE FOR RISK ASSESSMENT

BfR
German Federal Institute for Risk Assessment

How FAIR are your research data – and why should you care?

Strategies and tools for more reproducible research

Till Biskup (FSK, FDM)

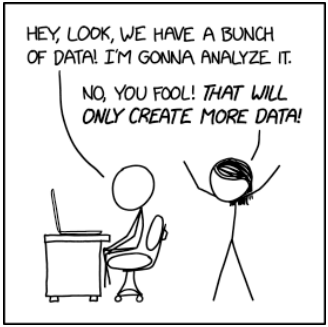
Promovierenden-Seminar
26th April 2023

Reproducible research

Why bother? It did work fine for decades, didn't it?

- Why suddenly all the fuss about 'research data management'? Didn't we do fine without it for decades?

DATA TRAP



'It's important to make sure your analysis destroys as much information as it produces.'

The ‘data deluge’

A parallel from the early days of software development

> *. . . so long as there were no machines, programming was no problem at all; when we had a few weak computers, programming became a mild problem, and now we have gigantic computers, programming has become an equally gigantic problem. [. . .]*
>
> *The increased power of the hardware, together with the perhaps even more dramatic increase in its reliability, made solutions feasible that the programmer had not dared to dream about a few years before. And now, a few years later, he had to dream about them and, even worse, he had to transform such dreams into reality!*
>
> – Edsger Dijkstra

☛ With (at least) exponentially growing amounts of data
we clearly need tools and strategies to cope with the situation.

Dijkstra, *Commun. ACM* **15**:859, 1972

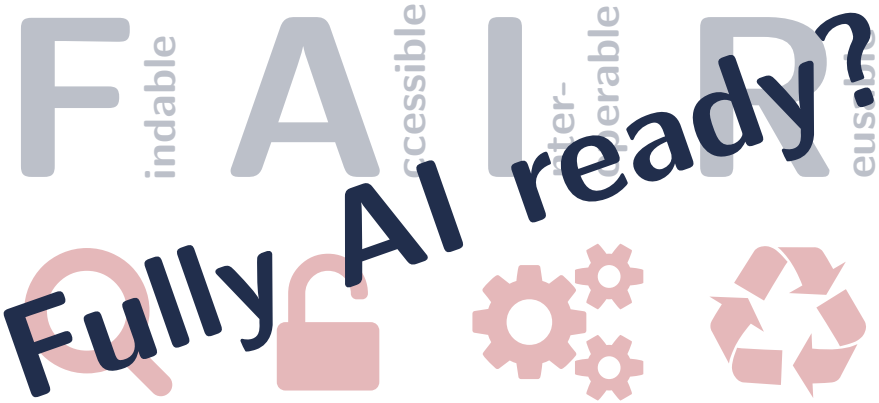
Reproducible research

The FAIR Guiding Principles for scientific data management and stewardship

Wilkinson *et al.*, *Scientific Data* **3**:160018, 2016

Reproducible research

The FAIR Guiding Principles for scientific data management and stewardship



Mons, *Nature* **578**:491, 2020

Reproducible research

Reproducibility is at the heart of the scientific method

> *If I have seen further*
> *it is by standing on ye shoulders of giants.*
>
> – Sir Isaac Newton

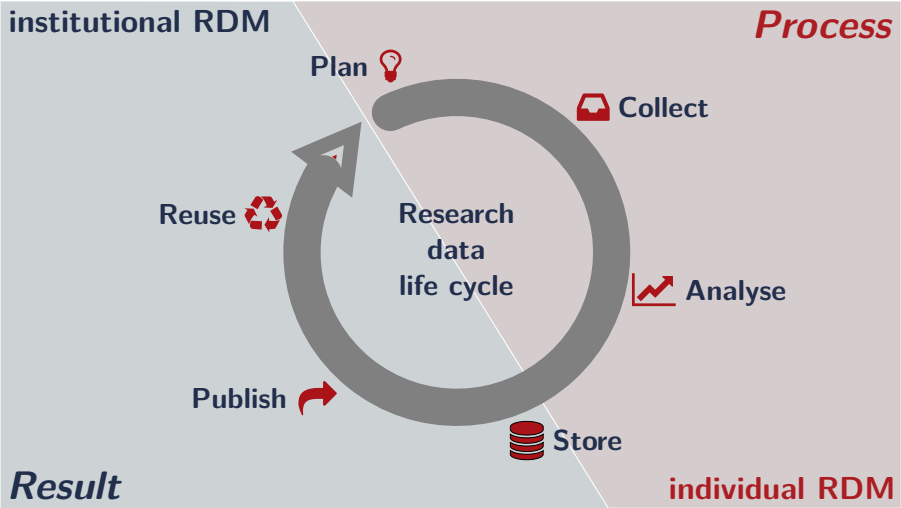
- How can I make sure others can reproduce what I have done?
- For how long do I remember myself what I have done?
- What is the usual length of stay of people in a group?

- We need to document *in sufficient detail* what we have done.
- Documentation needs to be automated wherever possible.

Sir Isaac Newton: Letter to Robert Hooke, 5th February 1676

Reproducible research

The research data life cycle as an abstract model



Research data life cycle

A few comments to the individual stations

Plan

- estimate kind and amount of research data to be collected
- clarify authorship, contributors, licenses, and IPR

- research data management plan (not only for funders!)
- easily accessible tools for project planning

Collect

- recording metadata *during* data acquisition
- Who has done what with whom when how and why?

- machine-readable and *human-writable*
- recording cannot be (fully) automated

Research data life cycle

A few comments to the individual stations

Analyse

- gap-less protocol of each analysis step
- fully reproducible data processing and analysis

- recording *fully automatically*
- full reproducibility only with scientific workflow system

Store

- (de)centralised storage with centralised backup
- conventions for file and directory names or PIDs

- data safe with automatic data upload from measurement devices
- local PIDs (hint: file paths are not persistent)

Research data life cycle

A few comments to the individual stations

Publish

- ▶ describe data package to be published
- ▶ as complete as possible: data, documentation, analyses, . . .
- ☛ data curation (ensuring data quality)
- ☛ workflow for (automatised) upload to repository

Reuse

- ▶ Overview of available research data
- ▶ direct link to data, alternative: contact details
- ☛ catalogue of (locally) available research data
- ☛ discipline-specific repositories for research data

Strategies and tools for more reproducible research

A focus on those aspects we can do ourselves

> *Do what you can, with what you've got, where you are.*
>
> – Theodore Roosevelt

- We as scientists are primarily responsible for managing our own data
- We know our own data – and only we can answer some crucial questions

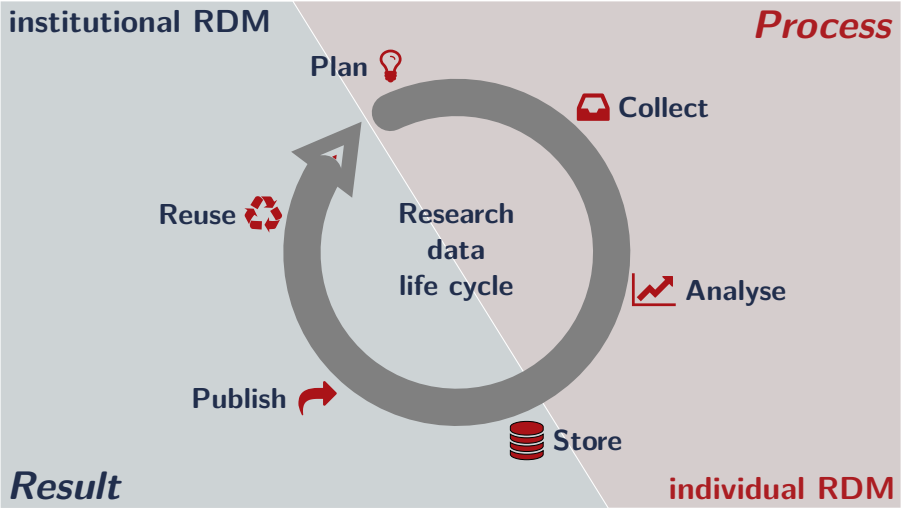
Requirements

- minimal external dependencies
- robust, modular, flexible, (easily) extensible
- sufficiently simple and elegant: tools that JUST WORK™

Theodore Roosevelt: An Autobiography, Macmillan, New York 1913, p. 364

Strategies and tools for more reproducible research

A focus on those aspects we can do ourselves

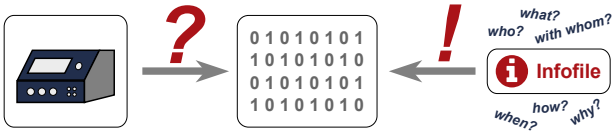


Collect – the Infofile format

A solution for recording metadata during data acquisition

Towards more reproducible and FAIRer research data: documenting provenance during data acquisition using the Infofile format

*Bernd Paulus, Till Biskup**



Digital Discovery **2**:234–244, 2023

Collect – the Infofile format

A solution for recording metadata during data acquisition

Challenges

- Record all relevant metadata *during* data acquisition in a machine-readable way
- Independent of vendor file formats, operating systems, and network infrastructure (including internet access)

Demands

- Human-writable and machine-readable format, with focus on the human side
- Sufficiently simple and elegant to use: should work well for an undergraduate student
- Adapts well to the changing requirements of science: robust, modular, flexible, and (easily) extensible

Collect – the Infofile format

A solution for recording metadata during data acquisition

```
common Info file - v. 0.1.0

GENERAL
Date start:  2020-04-04
Time start:  11:05:00
Date end:    2020-04-04
Time end:    15:50:00
Operator:    John Doe
Purpose:     Kill time

SAMPLE
Name:        Random sample 1
Description: Nicked from bench neighbour

COMMENT
To be or not to be...
```

- Who has done
- what
- with whom
- when
- how
- and why?

- parameters as key–value pairs
- grouped into (logical) blocks

☛ Many more blocks/parameters for specific methods

Collect – the Infofile format

Key aspects

- Plain text file, human-writable
  - No external dependencies
- Resides next to the data
  - Independent of vendor formats; data and metadata always together
- Machine-actionable metadata
  - Analysis routines (*e.g.*, ASpecD) can make sense of your data
- Sufficient detail
  - Never forget an important detail
- Modular and extensible
  - Easy to adapt to specific needs

ASpecD: A modular framework for the analysis of spectroscopic data focussing on reproducibility and good scientific practice

*Jara Popp, Till Biskup**

Chemistry—Methods **2**:e202100097, 2022

Analyse – the ASpecD framework

A framework for the fully reproducible Analysis of Spectroscopic Data

Challenges

- Fully reproducible data processing and analysis including a gap-less protocol of each step
- Most scientists have not received any formal training in programming or software development

Demands

- Largely automated ‘scientific workflow system’ capable of handling large amounts of data
- Sufficiently simple and elegant to use: should work well for an undergraduate student
- Adapts well to the changing requirements of science: robust, modular, flexible, and (easily) extensible

Analyse – the ASpecD framework

A framework for the fully reproducible Analysis of Spectroscopic Data

```yaml
format:
  type: ASpecD recipe
  version: '0.2'

datasets:
  - /path/to/first/dataset
  - /path/to/second/dataset

tasks:
  - kind: processing
    type: BaselineCorrection
    properties:
      parameters:
        kind: polynomial
        order: 0
  - kind: singleplot
    type: SinglePlotter1D
    properties:
      filename:
        - first-dataset.pdf
        - second-dataset.pdf
```

```yaml
system_info:
  python:
    version: "3.7.3 ..."
  packages:
    aspecd: 0.6.4
# ...
- kind: processing
  type: BaselineCorrection
  properties:
    parameters:
      kind: polynomial
      order: 0
      coefficients:
      - -0.04609818536259180
      fit_area:
      - 10
      - 10
      axis: 0
  apply_to:
  - /path/to/first/dataset
  # ...
```

☛ Recipe-driven data analysis:
We usually have an idea what we want to happen to our data.

Analyse – the ASpecD framework

Key aspects

- Recipe-driven data analysis
  - No programming skills needed
- Dataset as unit of data and metadata
  - Abstracts away from vendor file formats
- Full reproducibility
  - History is a fully working recipe
- Modular and extensible
  - Focus on the operation, not the infrastructure
- Support for different spectroscopic methods
  - Python packages for dedicated methods available

Analyse – the ASpecD framework

Reliable, high-quality software

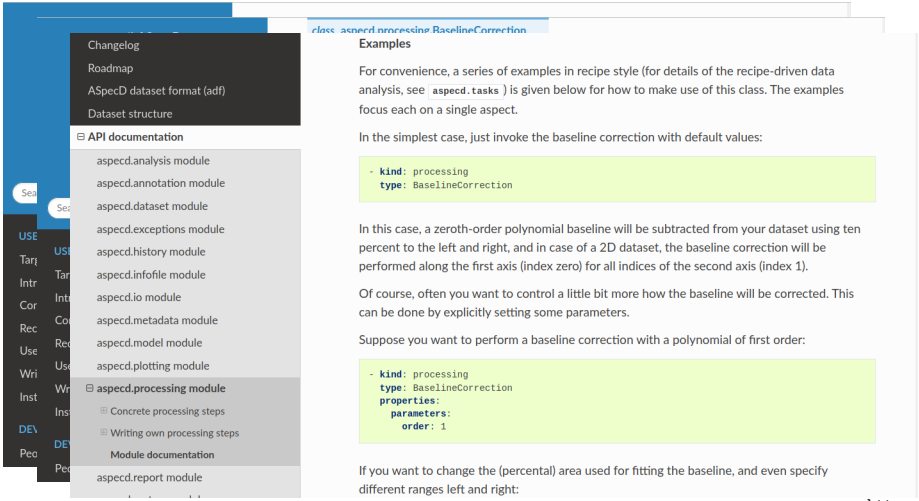
Following best practices in software development, *e.g.*:

- test-driven development
  - high test coverage, better reliability
- clean code
  - readable, expressive, self-documenting
- fully documented
  - `https://docs.aspecd.de/`
- version control system
  - `https://github.com/tillbiskup/aspecd/`
- open source
  - BSD license: everybody is allowed use and modify it

☛ Everybody who can program Python could take over the project.

Analyse – the ASpecD framework

Extensive documentation available online: `https://docs.aspecd.de/`



`https://docs.aspecd.de/`

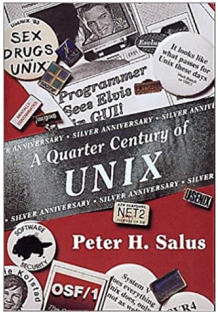
Sensible requirements for a digital (research) infrastructure

A decades-old – and proven – answer

❓ What are the reasonable requirements for a digital infrastructure that covers large parts of the research data life cycle?

The Unix Philosophy

- Write programs that do one thing and do it well.
- Write programs to work together.
- Write programs that handle text streams, because that is a universal interface.



Peter H. Salus (1994): A Quarter Century of UNIX. Reading (MA), Addison-Wesley; S: 53.

Strategy: Automation

Learn how to program to automate the ‘boring stuff’

> *Civilization advances by extending the number of important operations which we can perform without thinking about them.*
>
> – Alfred North Whitehead

- This is not a plea not to think . . .
- Think through and formalise recurring processes in order to keep your mind free for the important things.
- Learn how to program and how to use your computer (beyond Word and Excel)

S. Allesina and M. Wilmes: Computing Skills for Biologists. Princeton University Press, Princeton & Oxford 2019

A. N. Whitehead: *An Introduction to Mathematics*, Dover, Mineola 2017 (orig. 1911)

The FAIR Principles revisited

How to apply the FAIR Principles to the small (and individual) scale?

Findable

Do I know where exactly this one dataset is located I remember to have recorded and urgently need NOW for a publication (or my thesis)? And if not, would I have any chance to successfully search for it?

Accessible

Do I have access to the data, or is it on someone else's computer, hard drive, memory stick or – beware – at my old institution (without me being in the position to phone somebody up and ask for help)?

The FAIR Principles revisited

How to apply the FAIR Principles to the small (and individual) scale?

Inter-operable

Did I remember to export the data into a format I can work with – without access to the instrument control software of this old device retired five years ago – or alternatively far away or at my old institution?

Reusable

Did I record all necessary information, *i.e.* metadata, to answer all the questions I may have now – with much more experience and context and for the first time really looking at the data?

The FAIR Principles revisited

How to apply the FAIR Principles to the small (and individual) scale?

> *Note that there is no way to email yourself in the past to ask for clarifications.*
>
> – Allesina and Wilmes 2019, p. 2

☛ Be fair to yourself – and think of your ‘future self’.

Key takeaways

- The FAIR Principles can be applied to your local context.
- Reuse first and foremost means: reuse by your future self.
- FAIR is just a tool to ask (and answer) the right questions.

S. Allesina and M. Wilmes: Computing Skills for Biologists. Princeton University Press, Princeton & Oxford 2019

Summary

What to (hopefully) take home

- Traceability and reproducibility are key aspects of the scientific method.
- The amount of (research) data we create increases exponentially. Hence, we need to develop strategies to manage these data.
- We need to document *in sufficient detail* what we have done. Documentation needs to be automated wherever possible.
- Reproducible research requires a digital research infrastructure consisting of modular, robust, interoperable and extensible tools.
- Only systems that are sufficiently easy to use and whose use promises obvious advantages will be used.

Resources

What to read and where to go for some more information

- Wilkinson *et al.*, The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**:160018, 2016
- Popp und Biskup, ASpecD: A modular framework for the analysis of spectroscopic data focussing on reproducibility and good scientific practice. *Chemistry–Methods* **2**:e202100097, 2022.
- Paulus und Biskup, Towards more reproducible and FAIRer research data: documenting provenance during data acquisition using the Infofile format. *Digital Discovery* **2**:234–244, 2023.

- S. Allesina and M. Wilmes: Computing Skills for Biologists. Princeton University Press, Princeton & Oxford 2019

- `https://zenodo.org/communities/fdm_bfr/`
- `https://www.till-biskup.de/de/lehre/programmierkonzepte/`
- `https://www.reproducible-research.de/`

GERMAN FEDERAL INSTITUTE FOR RISK ASSESSMENT

Thank you for your attention

Till Biskup (FSK, FDM)

German Federal Institute for Risk Assessment
Max-Dohrn-Straße 8–10 • 10509 Berlin • GERMANY
Telefon +49 30 - 184 12 – 0 • Fax +49 30 - 184 12 – 99 0 99
bfr@bfr.bund.de • www.bfr.bund.de/en