HPC.NRW

Research Data Management

Challenge or Chance?

DAS KOMPETENZNETZWERK FÜR HOCHLEISTUNGSRECHNEN.

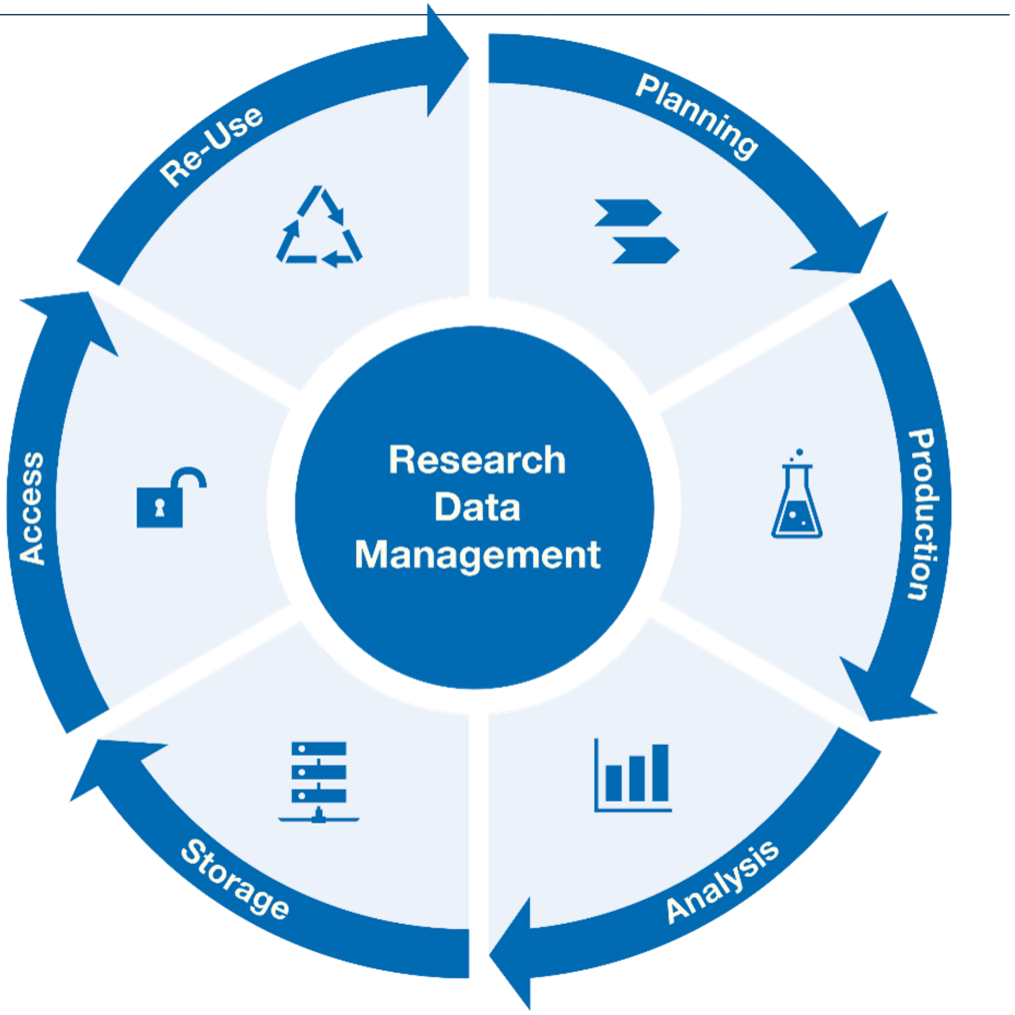
Goal of RDM

Establish **measures** for the

- organization
- preservation (long term storage) and
- documentation of the data

A **process** to keep the data

- accessible
- understandable
- reusable



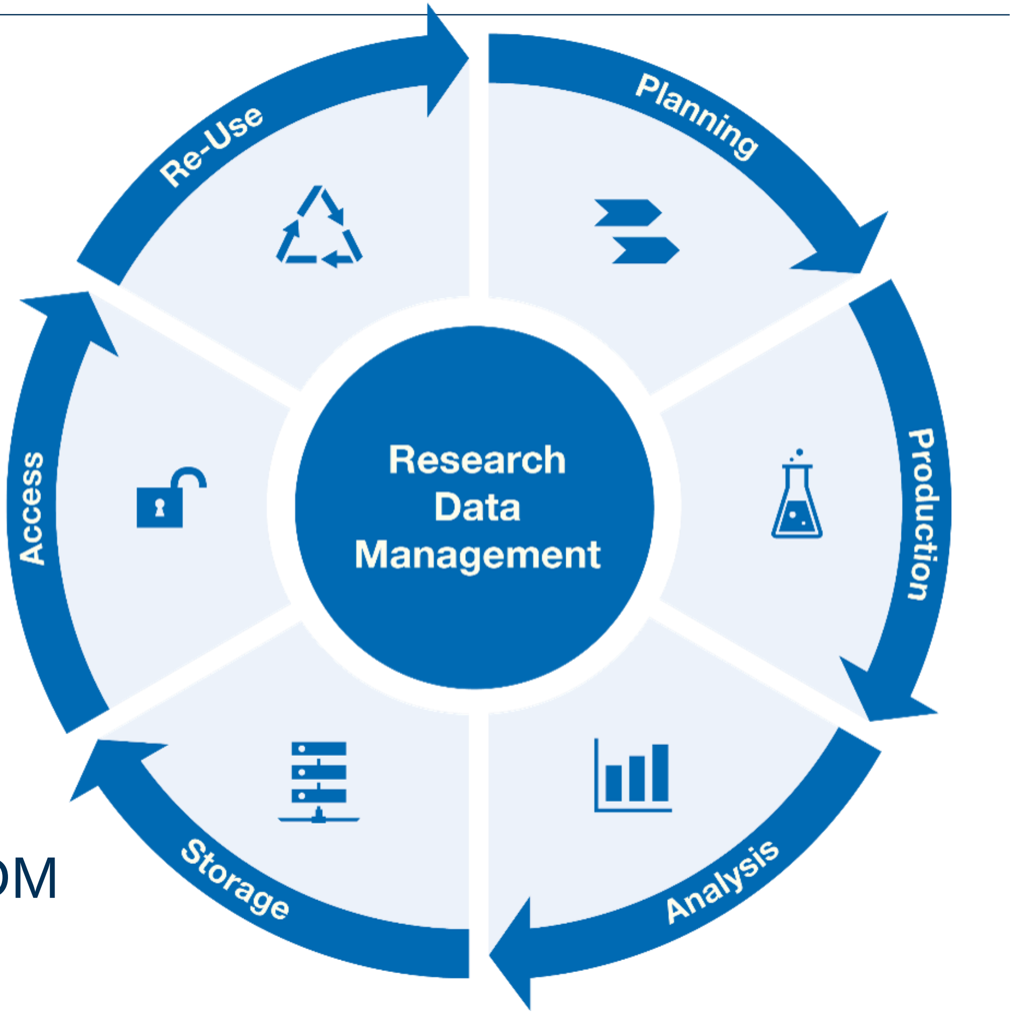
Goal of RDM

Establish **measures** for the

- organization
- preservation (long term storage) and
- documentation of the data

A **process** to keep the data

- **Not Open Access**
- RDM is needed for OA but OA is not required for RDM
- Everything can be „closed“ but organized



Common Issues

Common issues that arise due to missing RDM measures:

- Storage
- Documentation
- Low reusability
- File format unknown / unreadable

Common Issues

Common issues that arise due to missing RDM measures:

– Storage:

  - Insufficient data redundancy (one copy?)
  - Data corruption (storage not suitable)
  - Loss of data due to missing back up
  - No data policy: unclear when to store data where

Data gets corrupted or lost

Common Issues

Common issues that arise due to missing RDM measures:

– Documentation:

  - Missing
  - Unclear how data is structured, versioned, …
  - Missing information (units, dimensions, abbreviatiations ...)
  - Unclear who and how data was produced
  - Unclear what data can be used for
  - No input scripts
  - Many folders with unclear structure and no clear naming (old, new, ...)

Person that knows the data leaves = data becomes less/un-usable

Common Issues

Common issues that arise due to missing RDM measures:

- Low Reusability:
  - Researchers typically leave after time period
  - No contact possibility
  - Follow up projects might be impossible
  - Re-doing research, waste of time an resources

Person that knows the data leaves = knowledge drain

Common Issues

Common issues that arise due to missing RDM measures:

– File format unknown / unreadable

  - Old and outdated
  - Files cannot be opened
  - Data cannot be reused
  - No responsible person
  - Research software outdated

Person that knows the data leaves = data is not accessible

RDM – in HPC

Common Issues on HPC Systems

- Data is created using **self written** scripts or code
- Data is stored in **personalized accounts**
- migration and storage concept → **individual**
- Data can consist of **several types**: code, input, script, output, logfile, metadata, raw, …
- No common Metadata scheme
- In general many formats and tools
- Data creation involves SW and HW, reproducibility

RDM – in HPC

Common Issues on HPC Systems

- Data is created using **self written** scripts or code
- Data is stored in **personalized accounts**
- migration and storage concept → **individual**
- Data can consist of **several types**: code, input, script, output, logfile, metadata, raw, …
- No common Metadata scheme
- In general many formats and tools
- Data creation involves SW and HW, reproducibility

RDM – Why do it?

- It saves a lot of time in the long run
  Easier to build on existing work in the group

- Less time wasted in trying to understand organization
  More time available for research

- Reduced data loss and reduced duplication of work

- Data can be handled by everyone, instead of only the person that left

RDM – Why do it? - Save Time and Resources

- It **saves** a lot of **time** in the long run
  Easier to build on existing work in the group

- Less time wasted in trying to understand organization
  More time available for **research**

- **Reduced** data loss and Reduced **duplication** of work

- Data can be handled by everyone, instead of only the person that left

RDM – Why do it? - Required

- Required by many funding agencies and more and more journals
- DFG (Codex, checklist, FAIR, …)
    - Prepared so that the data can be reused
- BMBF (FAIR, checklist, ...)
    - relevant repositories as soon as possible, but not later than six month after the end of funding
- EU (FAIR, DMP, ...)
    - as open as possible as closed as necessary

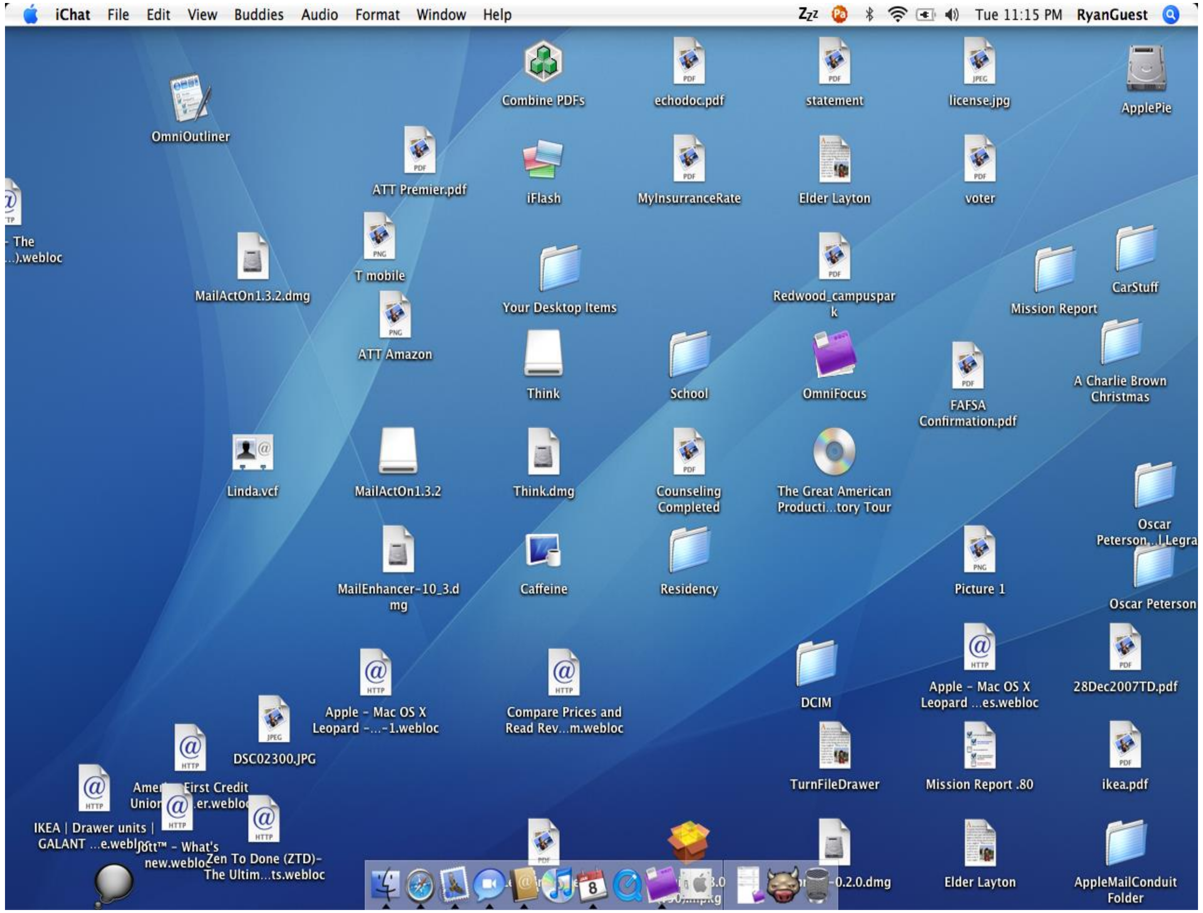
RDM – How does it Help?

- Helps with File naming

https://phdcomics.com/comics/archive.php?comicid=1531

RDM – How does it Help?

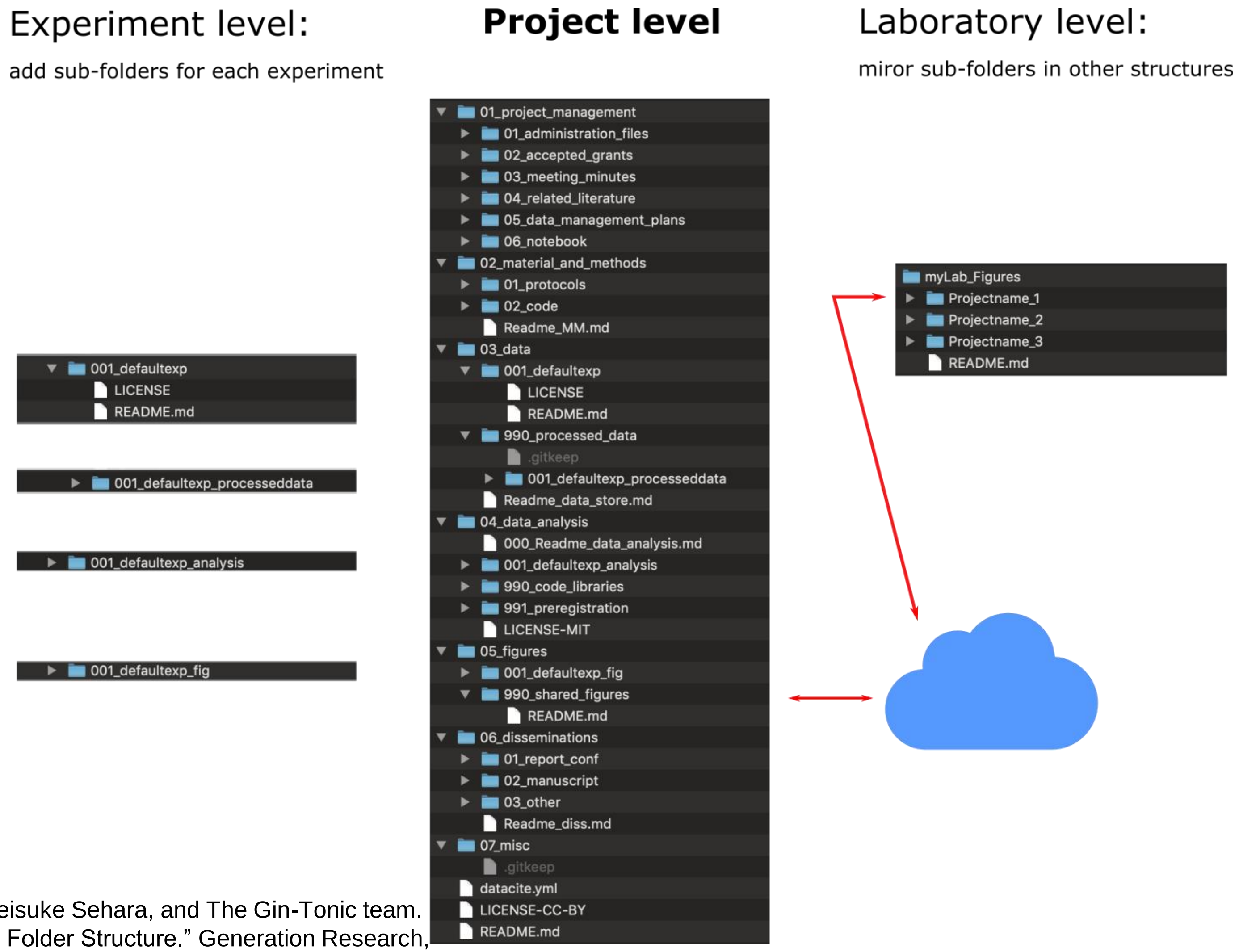
- Helps with File naming
- Folder Structures
- Documentation aspects
- Handling of large files



CC-BY-SA 2.0 *My Messy Desktop* von saaby

RDM – How does it Help?

- Helps with File naming
- Folder Structures
- Documentation aspects
- Handling of large files



Experiment level:

add sub-folders for each experiment

Project level

Laboratory level:

miror sub-folders in other structures

```
▼ 001_defaultexp
      LICENSE
      README.md

  ▶ 001_defaultexp_processeddata

▶ 001_defaultexp_analysis

▶ 001_defaultexp_fig
```

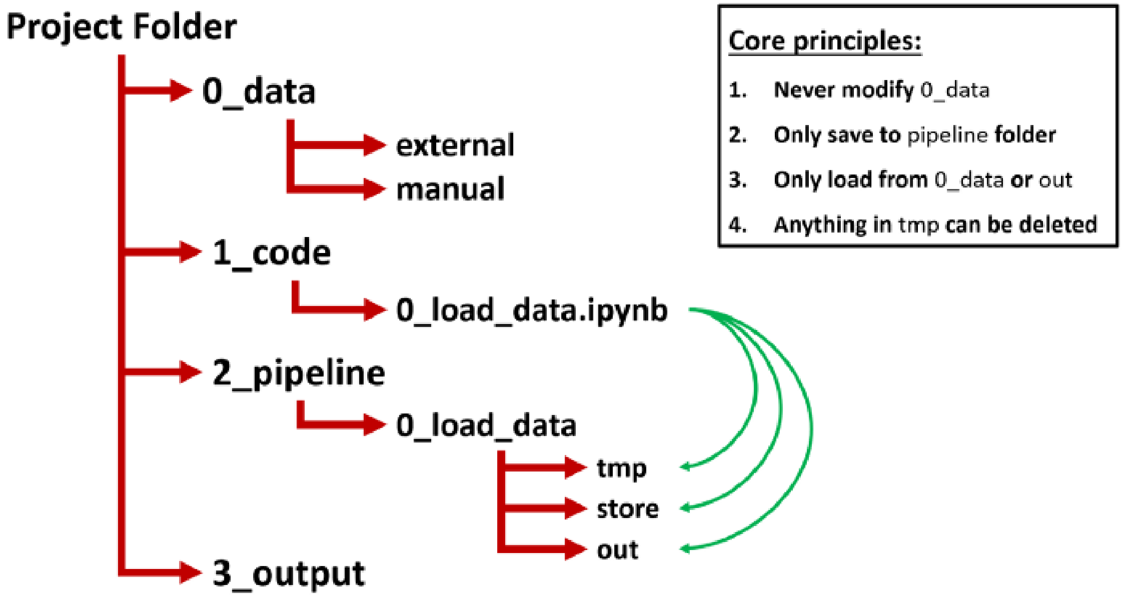
```
▼ 01_project_management
  ▶ 01_administration_files
  ▶ 02_accepted_grants
  ▶ 03_meeting_minutes
  ▶ 04_related_literature
  ▶ 05_data_management_plans
  ▶ 06_notebook
▼ 02_material_and_methods
  ▶ 01_protocols
  ▶ 02_code
    Readme_MM.md
▼ 03_data
  ▼ 001_defaultexp
      LICENSE
      README.md
  ▼ 990_processed_data
      .gitkeep
    ▶ 001_defaultexp_processeddata
    Readme_data_store.md
▼ 04_data_analysis
    000_Readme_data_analysis.md
  ▶ 001_defaultexp_analysis
  ▶ 990_code_libraries
  ▶ 991_preregistration
    LICENSE-MIT
▼ 05_figures
  ▶ 001_defaultexp_fig
  ▼ 990_shared_figures
      README.md
▼ 06_disseminations
  ▶ 01_report_conf
  ▶ 02_manuscript
  ▶ 03_other
    Readme_diss.md
▼ 07_misc
    .gitkeep
  datacite.yml
  LICENSE-CC-BY
  README.md
```

```
myLab_Figures
▶ Projectname_1
▶ Projectname_2
▶ Projectname_3
  README.md
```

Colomb, Julien, Thorsten Arendt, Keisuke Sehara, and The Gin-Tonic team. “Towards a Standardized Research Folder Structure.” Generation Research, 2021. https://doi.org/10.25815/WCY6-M233.

RDM – How does it Help?

- Helps with File naming
- Folder Structures
- Documentation aspects
- Handling of large files



Core principles:

1. Never modify 0_data
2. Only save to pipeline folder
3. Only load from 0_data or out
4. Anything in tmp can be deleted

Ties de Kok (2018): [How to keep your research projects organized, part 1: folder structure]

RDM – How does it Help?

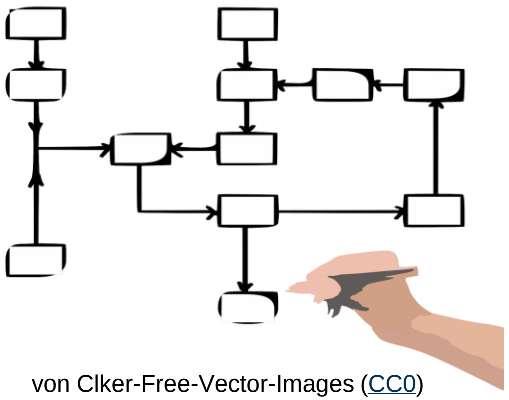
- Helps with File naming
- Folder Structures
- Documentation aspects
- Handling of large files
- Versioning

https://git-scm.com/ (MIT Licence)

Versioning via file name
data_ver_1.0.dat
data_ver_1.1.dat
data_ver_1.1.2.dat

RDM – How does it Help?

- Helps with File naming
- Folder Structures
- Documentation aspects
- Handling of large files
- Versioning
- Metadata
  - Find data and handle large data quantities
  - Increase efficiency



von Clker-Free-Vector-Images (CC0)

Mandatory	Recommended	Optional
Identifier	Subject	Language
Creator (+ nameIdentifier)	Contributor	AlternateIdentifier
Title	Date	Size
Publisher	RelatedIdentifier	Format
PublicationYear	Description (+ methods, technicalInfo)	Version
ResourceType	GeoLocation	Rights
		FundingReference

Manual: https://schema.datacite.org/

RDM – How to do it? - Guides, Policies, Routines, Automation

5S-Data	
sort	Check folders and remove unneeded files
set in order	Design folder structures and file naming conventions
shine	Establish regular routines, document and control procedures
standardize	Document best practices, guidelines and rules, develop joint standards with colleagues, clarify responsibilities
sustain	**Maintain the system and pass it on to your colleagues**

Source: In Anlehnung an Siiri Fuchs, Tanja Lindholm, Juuso Ala-Kyyny, Mari Elisa Kuusniemi, Ville Tenhunen (2020): Organizing data folders with #5SDATA method,verfügbar unter https://www.rd alliance.org/organizing data folders 5sdata method und Lang, Kevin; Roman Gerlach; Jessica Rex; Annett Schröter; Nadine Neute: Coffee Lecture Slides: 5S Data -Organisation is not a 4-letter word! (Coffee Lecture 27.01.2021), verfügbar unter https://zenodo.org/record/4454596#.YWRbWLgzY2w

RDM – How to do it? - Talk to Experts

- Central RDM Service Providers at all Universities

https://www.rwth-aachen.de/cms/root/Forschung/~lnaw/Forschungsdatenmanagement/
https://www.uni-bielefeld.de/ub/digital/forschungsdaten/index.xml
https://www.ruhr-uni-bochum.de/researchdata/de/
https://www.forschungsdaten.uni-bonn.de/de
https://www.tu-dortmund.de/forschung/forschungsdatenmanagement/
https://www.uni-due.de/rds/
https://www.fdm.hhu.de/
https://fdm.uni-koeln.de/en/rdm-services
https://www.uni-muenster.de/Forschungsdaten/
https://www.uni-paderborn.de/forschung/forschungsservice-und-beratung/forschungsdaten
https://www.ub.uni-siegen.de/benutzung-und-service/forschungsdatenmanagement/
https://fdm.uni-wuppertal.de/de/

- Contact for help regarding
  - Use of storage infrastructure
  - Repository selection
  - Organizational questions
  - Policy establishment
  - Handling of data
  - Proposals
  - ...

Community Standards, NFDIs, ...

Real World: Continuing a Project

- From the desk of a student continuing a project of two prior students:
- "I tried my best to summarize all that, please contact me if something remains unclear.
- The data is somewhat diluted, but still structured. Note that part of it is archived.
- Most of the names are self-explanatory."
- " ... (at least I believe that the column is also in these units)."
- "Some comparison along these lines can be found in folder1 there are two dirs named data-set-1 and data-sets-2.
  Frankly speaking God knows what is inside… Seemingly there is some analysis and some non-standard setup.
  I believe only person XYZ can tell what they were made for."
- " Im afraid I cannot tell you why they are called this way."
- " I believe that's it. Might be there exist some other data that I missed."

Real World: Continuing a Project

- From the desk of a student continuing a project of two prior students:
- "I tried my best to summarize all that, please contact me if something remains unclear.
- The data is **somewhat diluted**, but still structured. Note that **part** of it is archived.
- **Most** of the names are self-explanatory."
- " ... (**at least I believe** that the column is also in these units)."
- "Some comparison along these lines can be found in folder1 there are two dirs named data-set-1 and data-sets-2.
  Frankly speaking God knows what is inside… **Seemingly** there is some analysis and some non-standard setup.
  I **believe only person XYZ** can tell what they were made for."
- " Im afraid **I cannot tell** you **why** they are called this way."
- " I believe that's it. **Might** be there **exist some other data** that I missed."

Real World: Publishing

Retractionwatch.com: https://retractionwatch.com/

Papers retracted due to **missing data**

“...the raw data are **no longer available to validate** the information.” https://retractionwatch.com/2016/02/23/we-are-living-in-hell-authors-retract-2nd-paper-due-to-missing-raw-data/

"Unfortunately, the values of the questioned variables could **not be confirmed** because the original research **records** were **unavailable**.“ https://retractionwatch.com/2017/01/20/boom-headshot-disputed-video-game-paper-retracted/

"...the validity of the data and reported findings in this paper are flawed and cannot be independently verified.“ https://retractionwatch.com/2017/07/31/study-social-media-retracted-authors-cant-provide-data/#more-51243

"...and the authors were **not able to supply raw data** in all instances.“

https://retractionwatch.com/2018/03/14/cancer-biologist-retracts-five-papers/#more-62936

Real World Disputes: The War over Supercooled Water

- How a hidden coding error fueled a seven-year dispute between two of condensed matter’s top theorists.
  https://physicstoday.scitation.org/do/10.1063/PT.6.1.20180822a/full/
- Simulations were used, with custom code written (not published)!
- Chandler: nothing changes
  Debenedetti: HD, LD water emerges
- "The Princeton team’s repeated requests for the Berkeley code went unanswered for more than two years."
- Berkeley: "What he didn’t have, he says, was the time or personnel to prepare the code in a form that could be useful to an outsider"
- It took only a week or so for him to skim the basic structure of the code and identify three or four places where a bug might be. Then it was just a matter of testing each one, a process that took Palmer, Debenedetti, and their team of coworkers a few months. By summer, they had pinpointed the error."
- “if this had been disclosed, this saga might not have gone on for seven years.”

Real World Disputes: The War over Supercooled Water

- How a hidden coding error fueled a **seven-year dispute** between two of condensed matter’s top theorists.
  https://physicstoday.scitation.org/do/10.1063/PT.6.1.20180822a/full/
- Simulations were used, with custom code written (not published)!
- Chandler: nothing changes
  Debenedetti: HD, LD water emerges
- "The Princeton team’s repeated requests for the Berkeley code went unanswered for more than two years."
- Berkeley: "What he didn’t have, he says, was the **time** or **personnel** to prepare the code in a form that could be **useful to an outsider**"
- It took only **a week** or so for him to skim the basic structure of the code and identify three or four places where a bug might be. Then it was just a matter of testing each one, a process that took Palmer, Debenedetti, and their team of coworkers **a few months**. By summer, they had pinpointed the error."
- “if this had been disclosed, this saga might not have gone on for seven years.”