Research Data Management – A practical guide

How to manage scientific data in your daily work

Susanne Boye and Martin Geisler

Leibniz-Institut für Polymerforschung Dresden e.V.
Short Course: Outline

1. **Introduction** – why we need to care about our data?

2. **Basic principles/tools for research data management**
   - ✓ Name your data – Conventions and nomenclature
   - ✓ Describing your data set – Meta data and ReadMe
   - ✓ Structure your data - Folder management
   - ✓ Document your scientific /practical work - Electronical lab notes

3. **Practical „real-life“ example**
There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days.

- Eric Schmidt
  Executive Chairman of Google
You can have data without information, but you cannot have information without data.

- Daniel Keys Moran
  American Computer Programmer and Science Fiction Writer
1. Why we need to care about our data?

Motivation:
1. Why we need to care about our data?

Motivation:

• Data Security
1. Why we need to care about our data?

Motivation:
• Data Security

✓ Backup (notice: USB drives are not a backup!)
✓ 3-2-1 Backup rule:

• 3 copies of data (1 x primary copy of raw data, 2 x copies of backup)
• 2 media (copies of data on at least 2 different storage media)
• 1 external backup (original and backups not at the same location)

1. Why we need to care about our data?

Motivation:

• Data Security
• Part of “Good Scientific (Research) Practices”

• Demand of national / international funders
• Institutional rules and policies for handling research data
• Prevention of scientific misconduct: data fabrication, falsification and plagiarism
1. Why we need to care about our data?

Motivation:

• Data Security
• Good Scientific (Research) Practice
• FAIR principle

... Do I find the data?

... Can I access the data?

... Do I understand the data?

... What I’m allowed to do with the data?
1. Why we need to care about our data?

Motivation:

• Data Security
• Good Scientific (Research) Practice
• FAIR principle

More information about FAIR data: https://www.go-fair.org/fair-principles/
1 Why we need to care about our data?

Motivation:

• Data Security
• Good Scientific (Research) Practice
• FAIR principle
• Data Sharing: Movie clip, open science

https://thepublicationplan.com/2019/05/14/five-essential-factors-for-data-sharing/, https://youtu.be/N2zK3sAtr-4, CC-BY licence
1. Why we need to care about our data?

Motivation:

- Data Security
- Good Scientific (Research) Practice
- FAIR principle
- Data Sharing
- Data Science
1. Why we need to care about our data?

Key Points for Data Management:

• Easier to analyze organized, documented data
• Find data more easily
• Don’t drown in irrelevant data
• Don’t lose data
• Get credit for your data
• Avoid accusations of misconduct

https://twitter.com/datadynamicsinc/status/1522877966090661888?lang=da
2. Basic principles/tools for research data management

Lots of information in your head gets lost over time

You, today

You, in six months (or anyone else)
2. Basic principles/tools for research data management

Describe your data!

And in 6 months!!!

Well-defined DATASET

Data

Structure

Metadata

You, today

2. Basic principles/tools for research data management

Enable integration with other data

DATASET

Data collection

2. Basic principles/tools for research data management

Enable long-standing usage of collected data

Data collection

Select

Use

Data selection for research

2. Basic principles/tools for research data management

✓ Name your data – Conventions and nomenclature
✓ Describing your data set – Meta data and ReadMe
✓ Structure your data - Folder management
✓ Document your scientific /practical work - Electronical lab notes
File Naming

Do’s and Don’ts - Examples

Best practices

- 20240523_ProjectA_SEC_BoyeS_v1.xlsx
- 20240523_ProjectA_MeetingNotes_BoyeS_v2.docx
- 20240523_ManuscriptScience_BoyeS_v525.docx
- ...

https://xkcd.com/1459/
File Naming

Requirements

• Human and machine-readable
• Be consistent
• Use date format ISO 8601: YYYYMMDD
• Include a version number (Creamer et al. 2014)
• Write down naming convention in data management plan

AVOID:
• Special characters: {}<>()!*#`;":?@ ...
• Space
• Cryptic
• Too long
File Naming

Take-Home Tips

✓ 1. Think about your files - What related files are you working with?

- You can use different conventions for different file sets
- Do you have established file naming conventions in your discipline or group?

Example: This convention will apply to all of my AF4 files, from raw data through processed file
File Naming

Take-Home Tips

✓ 1. Think about your files

✓ 2. Identify metadata (e.g. date, sample, experiment) – What information is needed to easily locate a specific file?

- Descriptive with contextual information
- Pick three pieces of metadata
- Names should be human readable: understand what’s in each one

Example: For my AF4 measurements, I want to know date, sample ID, and injection number for that sample on that date

gibbons.de, datamanagement.hms.harvard.edu
File Naming

✓ Take-Home Tips

✓ 1. Think about your files

✓ 2. Identify metadata (e.g. date, sample, experiment)

✓ 3. Abbreviate or encode metadata - Don’t forget to document any codes!

• Standardize the categories and/or replace them with 2- or 3-letter codes

  Example: e.g. 2-letter project abbreviation: project 1 -> P1, project 2 -> P2
  or 3-letter code for technique, SEC, AF4, NMR,...
File Naming

Take-Home Tips

✓ 1. Think about your files
✓ 2. Identify metadata (e.g. date, sample, experiment)
✓ 3. Abbreviate or encode metadata
✓ 4. Use versioning - Are you maintaining different versions of the same file?

• Track versions by adding information on the end of file name
  • Use version numbers („v01“ or „v02“)
  • Use version date (YYYYMMDD)

Example: analysis workflow: _raw and _processed
File Naming

Take-Home Tips

• How you want to sort and search
  • Decide which metadata should appear at the beginning
  • Use default order: alphatically, numerically, or chronologically

Example: AF4 measurement: 5075_2024-02-16 11-48 - wash 400uL

✓ 5. Think about how you will search for files – what comes first?
File Naming

Take-Home Tips

• Determine characters for separation of metadata – many computer systems cannot handle space
• Use dashes (file-name.xx) or underscores (file_name.xx) or capitalize first letter of each word – camel case (FileName.xx)
  • Use default order: alphabetically, numerically, or chronologically
  • Avoid special characters: .,’@#€§$%/()=?{}[]`!

Example: 20240603_Presentation_Course_FFF2024_V01.ppt

✓ 6. Deliberately separate metadata elements – avoid spaces or special characters
File Naming

Take-Home Tips

- Name conventions should be documented that others in your lab/group can follow this standard
  - Document in a README.txt together with your files
  - If the file is moved or shared, users will be able to identify the file from its file name

Example: My file naming convention is [SA-MPL-EID]_[YYYYMMDD]_[###]_[status].[tif]
P1_SEC_PS30k_20240423_2_raw.txt

✓ 7. Write down your naming conventions – Include a top-level README file on how to navigate the structure
Describing your data set – Meta data and ReadMe

Meta data – Enrich your data with meaning!

Meta data helps to understand data

Meta data answers questions:

✓ **Who** created the data?

✓ **Why** was the data created?

✓ **When** was the data created?

✓ **How** was the data created?

✓ **What** is the content of the data?

More information under: [https://data.research.cornell.edu/content/readme](https://data.research.cornell.edu/content/readme)
Describing your data set – Meta data and ReadMe

Easiest way to make data more understandable and reusable:
ReadMe.txt file “Love letter to the Future” – for each folder (ideally)

Content:
✓ Title
✓ Date
✓ Author
✓ Description of folder content, structure and organization of files, abbreviations, coding, etc.
✓ … as much information as possible

More information under: https://data.research.cornell.edu/content/readme
Structure your data – Folder Management

✓ How to organize your data?

✓ Files and data
✓ Data growth
✓ Change over time
✓ Relationships
✓ Duplication of data

DATA CHAOS

https://www.industr.com/de/wider-das-kabelchaos-loesungen-fuer-die-elektroinstallation-2252844
Structure your data – Folder Management

Data flows

Raw data
↓
Ready-to-analyze data
↓
Processed (computational) results

Figures
Tables
Numerical results

Report
Manuscript
Poster
Presentation
Structure your data – Folder Management

Data hierarchy

✓ One project – one folder
✓ Consistent structure for each project

AVOID:
✓ Overlapping categories
✓ Too large folders and too deep structures
✓ Repetition of information
Structure your data – Folder Management

Document your structure

✓ Make sure to capture metadata about content of folders and files:
  
  Naming conventions
  
  Who made it, when, where???

✓ Create documentation / ReadMe files
✓ Establish standardized structures in your group/department

Take-Home Tips:

✓ Dump older files cluttering your working directory
✓ Delete unneeded files when the project is finished
Short Course: Digitalization in Polymer Analysis (S. Boye and M. Geisler)

Document your practical work - Electronical lab notes

Traditional paper notebooks

Page from a laboratory notebook of Alexander Graham Bell, 1876.

Page from Gale’s lab notes “Breaking Bad”

Document your practical work - Electronical lab notes

Benefits

- Helps you to organize yourself and your research.
- No loose paper on your desk nor any cryptic notes.
- No data loss when researchers move on.
- Simplified summary of your research.
- Documenting helps you to understand old data.
- Easy data sharing.
- Long-term storage and archive.

Can you read and understand it?

Document your practical work - Electronical lab notes

Digitalization of notes:

✓ Is easy to read and to edit – no handwriting
✓ Is versioned, changes can be reverted and tracked
✓ Is searchable and sortable
✓ Use of templates – standardization
✓ Links to: resources, other experiments, raw data or analysis workflows
✓ Application of APIs - Automated analysis + Data science

Sample Table:

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FK_AF4_014</td>
<td>Ferrocen-Psone (0.5 mg BCPint)</td>
</tr>
<tr>
<td>FK_AF4_015</td>
<td>N3-Ferrocen-Psone (0.5 mg BCPint)</td>
</tr>
<tr>
<td>FK_AF4_016</td>
<td>Psone A (0.5 mg BCPint)</td>
</tr>
</tbody>
</table>

Materials / Device:

- Buffer: 1 mL, pH 7.4, + 0.02 % NaN₃
- Channel: LC
- Spacer: 400 μm
- Membrane: RC (cut-off: 10 kDa)
- Pump + AS system: Agilent 1200 series
- FFF device: Eclipse DuoTec (Wyatt)
- LS detector: DAWN-HELOS-II (Wyatt): λ=680 nm
- RI detector: Optlab T-EX (Wyatt): λ=680 nm
- Diode array detector: SPD-M20 (Shimadzu), wavelengths:
  - AUX1 (nm): 254
  - AUX2 (nm): 280
  - AUX3 (nm): 400
  - AUX4 (nm): 500
Research Data Management

Avoid the scientific data nightmare: watch the video

3. Practical „real-life“ example

STEPS TO A DATA MANAGEMENT OF A SCIENTIFIC PROJECT

1. Data sources
   ✓ Idea/Grant
   ✓ WHO
   ✓ Collaboration partners

2. Data collection
   ✓ Experimental Data
   ✓ Discussion
   ✓ Design experiments

3. Data preparation
   ✓ Data Cleaning
   ✓ Data Analysis
   ✓ Structure

4. Manuscript/Report
   ✓ Figures
   ✓ Writing/Revision
   ✓ Submission/Revision

5. Safe & long-term storage
   ✓ Data organization/Read Me Files
   ✓ Supervision
   ✓ Long-term storage, open access
3. Practical „real-life“ example

**STEPS TO A DATA MANAGEMENT OF A SCIENTIFIC PROJECT**

- **Cloud (IPF Cloud)** ➔ Create one folder for each project. The project should be defined upfront.

- **ELN** [https://www.labarchives.com/](https://www.labarchives.com/) ➔ Create one Notebook for each project. All notebooks should have the same structure.

Data sources  Data collection  Data preparation  Manuscript/Report  Safe & long-term storage

---

Electronic Labbook in combination with an inhouse drive or CLOUD service

Data collection  Manuscript/Report  Safe & long-term storage
3. Practical „real-life“ example

STEPS TO AN IDEAL DATA MANAGEMENT OF A SCIENTIFIC PAPER

CLOUD structure *in progress*

**Name of Project**

- *Read me* file (people involved into the project, nomenclature, status of the project, new ideas...)
- *Electronic Lab Book/protocols* (divided by person involved into the project)
- *Raw data* (organized by characterization methods, by operator and by positive/negative results)
- *DataAnalysis*
- *Communication* (Reports, meetings, working plan, emails)
- *Literature*

**Name of Project-Lab Book**

- People involved into the project
- Nomenclature and summary of compounds; Tags
- Meeting Notes; Status of the project
- Experimental Data (can be divided by people involved)
  - SM 20210510 Fabrication of FcPsomes
  - SM 20210514 Stability of FcPsomes
  - SM 20210514 Preparation of FcPsomes for AF4

*When a student finishes:* 1) change the ownership of the notebook, 2) make a PDF file and 3) added it in the ELN folder in the cloud.
### STEPS TO AN IDEAL DATA MANAGEMENT OF A SCIENTIFIC PAPER

#### Data collection/Sharing Data

<table>
<thead>
<tr>
<th>Notebooks</th>
<th>47</th>
</tr>
</thead>
</table>

- **Redox (Fc), pH responsive and crosslinked Psomes**
  - 1. Peroxidase-Like Activity, Host-Guest Interactions and H2O2-Responsive Release Behavior
    - People involved into the project
    - Nomenclature, summary of compounds, softwares and tags
    - Meeting Notes; status of the project
  - **Experimental Data**
    - **Recording Data collection**
    - New...
  - 2. Host-Guest Interactions and H2O2/amylase-Responsive Release Behavior (Special Issue)
  - 3. Immobilization of enzyme-like containing polymeric vesicles on a solid support
    - New...
  - Deleted Items

---

3. Practical „real-life“ example
STEPS TO AN IDEAL DATA MANAGEMENT OF A SCIENTIFIC PAPER

- Paper Preparation
  - Experimental Description Collection
  - Raw Data Collection

Let’s take a look the content!
3. Practical „real-life“ example

STEPS TO AN IDEAL DATA MANAGEMENT OF A SCIENTIFIC PAPER

CLOUD structure *after acceptance publication*

First Author_Name of Paper/Journal_Year

- **Communication**

- **Manuscript** (main text and SI, PowerPoint and pics of the last version of Figures, TOC, Readme File)

- **Scientific Data** (Raw data and Processed Data ordered by Figures)

- **Proof of submission to the library**

The supervisor* is the responsible to check the last version
3. Practical „real-life“ example

**STEPS TO AN IDEAL DATA MANAGEMENT OF A SCIENTIFIC PAPER**

- **First author**
- **Journal**
- **Year**

**Workshop › Moreno_Biomacromolecules_FerrocenePsomes_2022**

- **Paper History**
- **Figures + Text**

**Submission to the library**

- **Figures + Text + Raw Data + Experimental Descriptions**

Let’s take a look the content!
Coding and Version Control: git\textsuperscript{[5]}

Distributed version-control for your source code

- Free and open-source software
- Structured source-code management system
- Easy and simple to use for everyday programming
- Reference your compiled code with unique SHA-hash
- Good scientific practice to comment and archive your code!

USE GIT!!!

\[\text{[5] https://git-scm.com/}\]
Use IDE: Jupyter and JupyterLab[6]

- Jupyter: Interactive Computing and Data Processing
  - Free and open-source software
  - Native Python binding for easy usage (also R)
  - Documentation/Latex using Markdown within the source
  - Collaboration and exchange of jupyter notebooks


Data Management Plan (DMP)

➔ Which data will be produced (type, format)?
➔ How large will be the data set probably?
➔ How and where shall the data be stored during and after the project duration? Backups?
➔ Is it planned to publish the data, and if yes, under which conditions?
➔ Could legal or ethical problems occur in collecting, analyzing and publishing the data?
➔ How much does it cost?
➔ Who is responsible for all the processes of research data management?

Collaboration Tools and Step-by-Step Survey:

[8] https://dmptool.org/
Connecting data: 5-star-linked-data\textsuperscript{[9]}

Design Pattern for Open Data

\textbullet\: data is openly available in some format.

\textbullet\: data is available in a structured format, such as Microsoft Excel file format (.xls).

\textbullet\: data is available in a non-proprietary structured format, such as Comma-separated values (.csv).

\textbullet\: data follows W3C standards, like using RDF and employing URIs.

\textbullet\: all on the other, plus links to other Linked Data sources.

For data publishing use public repositories e.g. Zenodo:

https://zenodo.org/

\textsuperscript{[9]} https://5stardata.info/en/
Connecting data: Example 1

Dynamic matrices with DNA-encoded viscoelasticity for cell and organoid culture

Yu-Hsuan Peng, Syuan-Ku Hsiao, Krishna Gyata, André Ruland, Günter K. Auernhammer, Manfred F. Maitz, Susanne Boye, Johanna Lattner, Claudia Gerri, Alf Honigmann, Carsten Werrner & Elisha Krieg

Nature Nanotechnology 18, 1463–1473 (2023) | Cite this article

15k Accesses | 12 Citations | 93 Altmetric | Metrics

An Author Correction to this article was published on 31 January 2024

This article has been updated
Connecting data: Example 1

Data availability

Supplementary Information containing materials and methods, supplementary figures, tables, datasets and accession numbers for biological materials are available with this paper. Additional datasets and materials generated during and/or analysed during the current study are available from the corresponding author on reasonable request. The supplementary data and code supporting the findings of this study are openly available on figshare (https://figshare.com/projects/Dynamic-mattresses-with-DNA-encoded-viscosity-for-cell-and-organoid-culture/768981). Source data are provided with this paper.

Code availability

A Python script for the thermodynamic calculations of CCL interactions is available as Supplementary Code 1 and at figshare (https://doi.org/10.6084/m9.figshare.2330942). A Python script for the statistical simulation of the maximum percentage of intermolecular crosslinks as a function of CCL complexity is available as Supplementary Code 2 and at figshare (https://doi.org/10.6084/m9.figshare.2399263).

Source data

Source Data Fig. 2
Simulation results, rheological data, source data for Fig. 2d and thermodynamic prediction.

Source Data Fig. 3
Rheological data.

Source Data Fig. 4
Rheological data.

Source Data Fig. 5
RFU value from qPCR gel volume quantification, statistical source data of cell viabilities and immune response.

https://static-content.springer.com/.../41565_2023_1483_MOESM12_ESM.xlsx

MS Excel sheet with numerical data
Connecting data: Example 2

The FAIR principle:

- **Findability**
- **Accessibility**
- **Interoperability**
- **Reusability**
Connecting data: Example 2

Database of Raman and ATR-FTIR spectra of weathered and biofouled polymers

FAIR4Chem Award 2024
Laureate Robin Lenz
IPF Dresden
Thank you for your attention! Questions???

If you have further questions, just contact us:

Dr. Susanne Boye (boye@ipfdd.de)
Dr. Martin Geisler (geisler@ipfdd.de)