

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



2



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.

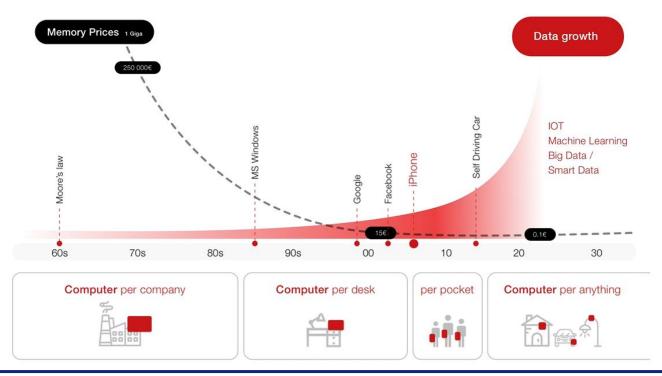
- Daniel Keys Moran

American Computer Programmer and Science Fiction Writer



https://data-flair.training/blogs/data-science-big-data-quotes/

Motivation:





Motivation:

Data Security









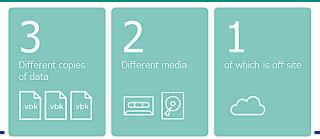




https://www.dailydot.com/unclick/missing-laptop-poster-goes-viral/, https://www.citrusuk.net/laptop-repairs/data-recovered-damaged-laptop/, https://forerunner.com/ worms-trojans-and-spywares/ ID 227809565© Wave Break Media Ltd | Dreamstime.com

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)





Motivation:

Alexander von HUMBOLD1

Data Security

European

Part of "Good Scientific (Research) Practices"

• Demand of national / international funders

• Institutional rules and policies for handling research data

FULBRIGHT Scholar Program

Prevention of scientific misconduct: data fabrication,

falsification and plagiarism

Volkswagen**Stiftung**





DAAD





Deutscher Akademischer Austauschdienst German Academic Exchange Service









Bundesministerium für Bildung und Forschung



Motivation:

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



... Do I understand the data?

do with the data?

Motivation:

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



More information about FAIR data: <u>https://www.go-fair.org/fair-principles/</u>



Motivation:

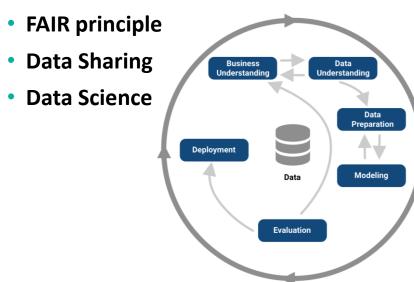
- Data Security
- Good Scientific (Research) Practice
- FAIR principle
- Data Sharing: <u>Movie clip</u>, open science





Motivation:

- Data Security
- Good Scientific (Research) Practice



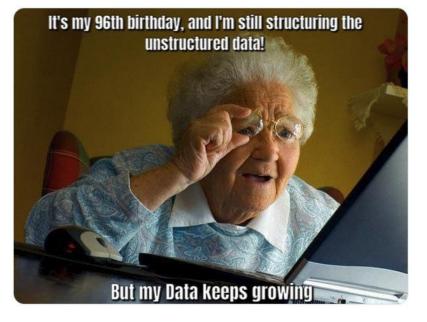




12

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



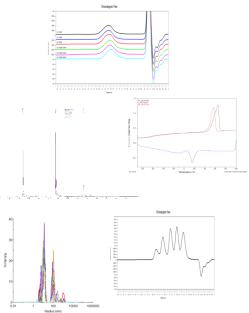


2. Basic principles/tools for research data management



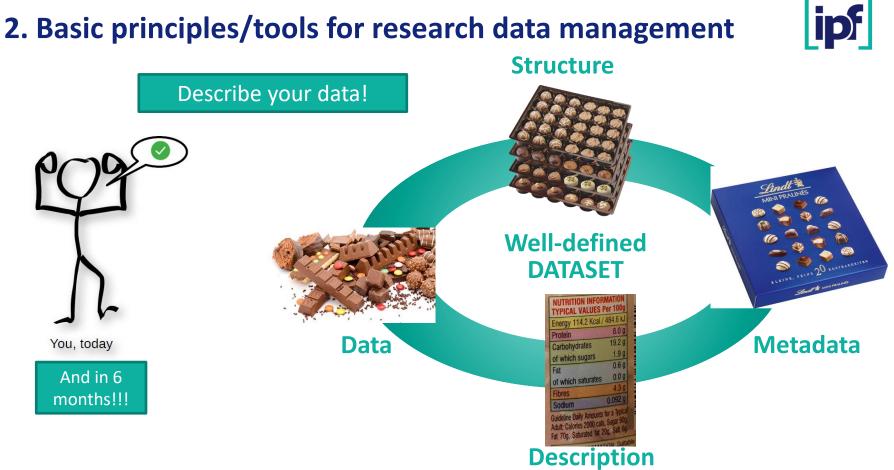
Lots of information in your head gets lost over time











15

https://www.wunderweib.de/schoene-schokoideen-pralinen-selber-machen-3375.html, https://www.lauensteiner.de/de/nougat-pralinen-p7844/, https://www.lieferello.de/Pralinen/Lindt-Mini-Pralin-s-9St-44g.html, https://www.krusteundkruemel.de/so-liest-der-low-carber-die-label-auf-nahrungsmitteln/



DATASET

2. Basic principles/tools for research data management



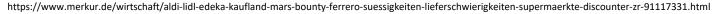
Enable integration with other data





Data collection

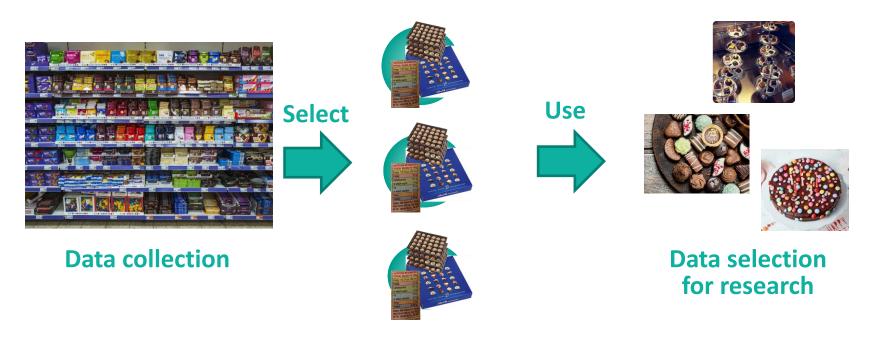
16





2. Basic principles/tools for research data management

Enable long-standing usage of collected data



17

https://relaxdays.de/wohnen/kuche/serviergeschirr/servierstander/etagere-porzellan-10035404.html, https://www.pinterest.de/pin/wir-wnschen-allen-finalistendes-food-blog-award-2015-im-rahmen-der-berlin-food-week-heute-viel-erfolg-zu-den-finalisten-gehren--419819996492652425/



ipf

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



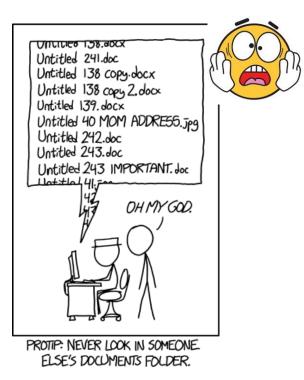
. . .

19

Best practices

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

ipf





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 displine or group?

Example: This convention will apply to all of my AF4 files, from raw data trough 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



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,...

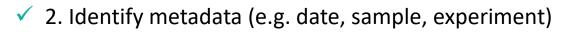


File Naming

Take-Home Tips

24

1. Think about your files



- 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: alphatically, numericall, or chronologically
 - Avoid special characters: .,~!@#€§\$%&/()=?{}[]`!

Example: 20240603_Presentation_Course_FFF2024_V01.ppt

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



File Naming







 Name conventions should be documented that others in your lab/group can follow this standard

- Document in a REDME.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

27



 $gibbons.de,\,datamanagement.hms.harvard.edu$

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?

28

What is the content of the data?



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

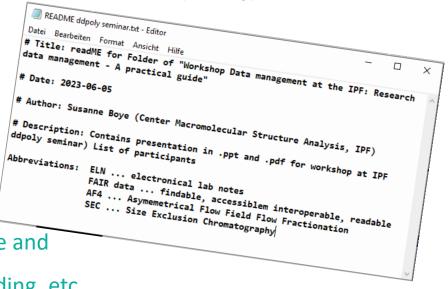
🗸 Date

Author

Description of folder content, structure and

organization of files, abbreviations, coding, etc.

✓ ... as much information as possible





Structure your data – Folder Management

✓ How to organize your data?

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

DATA CHAOS



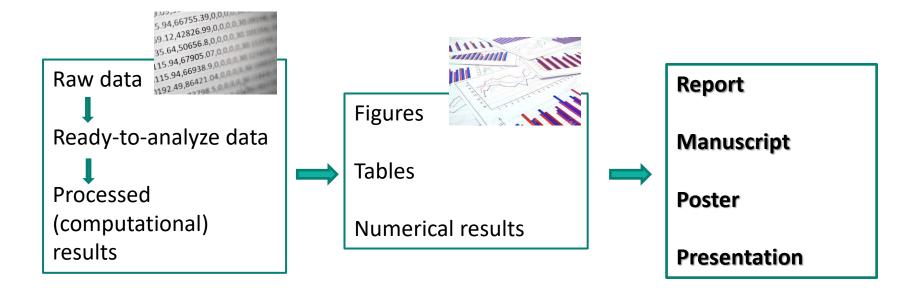




Structure your data – Folder Management



Data flows





Structure your data – Folder Management

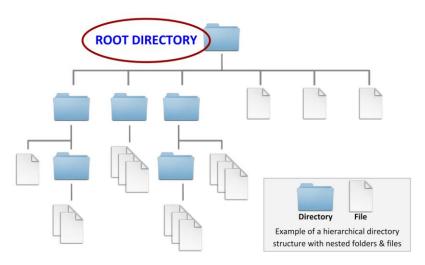
Data hierarchy

- ✓ 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:

33

- Dump older files cluttering your working directory
- Delete unneeded files when the project is finished







Document your practical work - Electronical lab notes



Traditional paper notebooks

March 10th 1876 To my delight he came and delan asked him to releast the words month piece M . That articulate sounds proceeded an affect was loved but indictinet and sumpled If I had read beforehand The passage gis I W The plate W- Wation I should have recommend les it was I could not make out The sense - but word here and there was quite distinct to and out " and "for thee" and cloself his other The sectionce " Mr. Bell Le your and finally The Transiting Latenach understand what I say? Do- you - un der - stand - what - I - say mite clearly and intelligitly Then shouted into M the following 1. 71. "W" Water - Come here - I want to

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



Page from Gale's lab notes "Breaking Bad"



www.eppendorf.com/pipetting



https://en.wikipedia.org/wiki/Lab_notebook, https://www.therpf.com/forums/threads/gale%E2%80%99s-lab-notes-breaking-bad.305036/

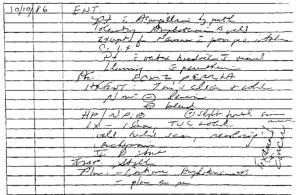
Document your practical work - Electronical lab notes

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

					-
	Samples:				
	Acronym	Description			
	FK_AF4_014	Ferrocen-Psome	(0.5 mg BCP/ml)		
	FK_AF4_015		me (0.5 mg BCP/ml))	
	FK_AF4_016	Psome A (0.5 mg	BCP/ml)		
d					
				Susanne Boye - Apr 18, 2023, 2:31 PM MESZ	
	Materials / Device				
				Susanne Boye - Jan 11, 2023, 12:49 PM MEZ	<u>.</u>
	Buffer: 1 mM, pH 7.4, + 0.02 % NaN ₃				
	Channel: LC				
or	Spacer: 490 µm				
	Membrane: RC (cut-off: 10 kDa)				
	Pump + AS system: Agilent 1200 series				
	FFF device: Eclipse Dualtec (Wyatt)				
	LS detector: DAWN HELEOS-II (Wyatt): λ=660 nm				
	RI detector: Optilab T-rEX (Wyatt): λ=660 nm				
	Diode array detector: SPD-M20 (Shimadzu), wavelengths:				
	AUX1 [nm]	AUX2 [nm]	AUX3 [nm]	AUX4 [nm]	
	254	280	400	500	





Research Data Management

37

Avoid the scientific data nightmare: watch the video



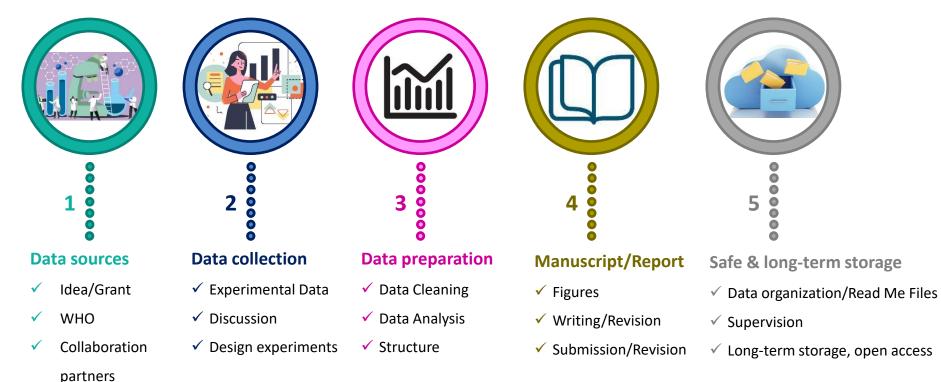




3. Practical "real-life" example



STEPS TO A DATA MANAGMENT OF A SCIENTIFIC PROJECT





38



STEPS TO A DATA MANAGMENT OF A SCIENTIFIC PROJECT

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

Data sources	Data collection	Data preparation	Manuscript/Report	Safe & long-term storage
		c Labbook in c ouse drive or C	ombination with LOUD service	

 ELN <u>https://www.labarchives.com/</u> → Create one Notebook for each project. All notebooks should have the same structure.

Data collection

Manuscript/Report

Safe & long-term storage





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 chracterization methods, by operator and by positive/negative results)
- DataAnalysis
- **Comunication** (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

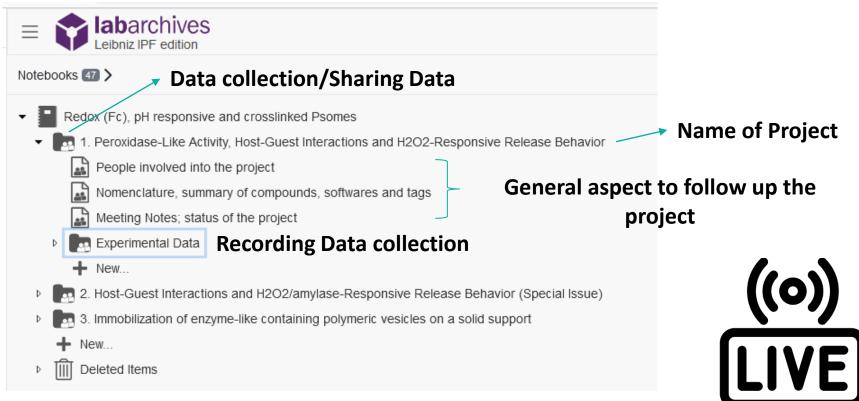
SM 20210514 Stability of FcPsomes

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.

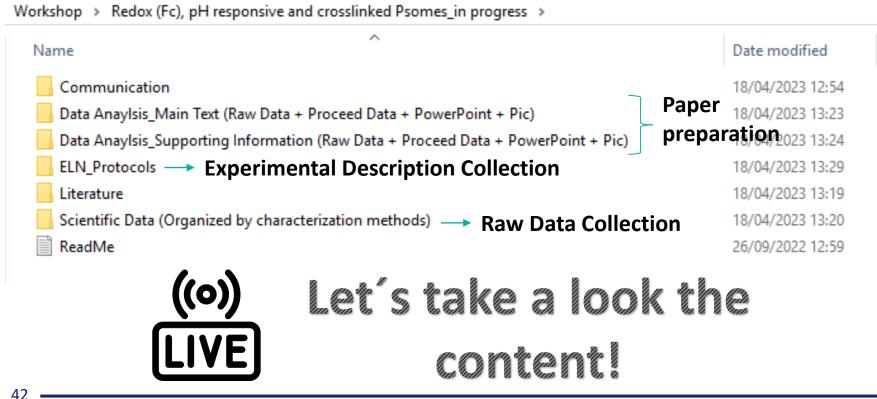
















CLOUD structure *after acceptance publication*

First Author_Name of Paper/Journal_Year

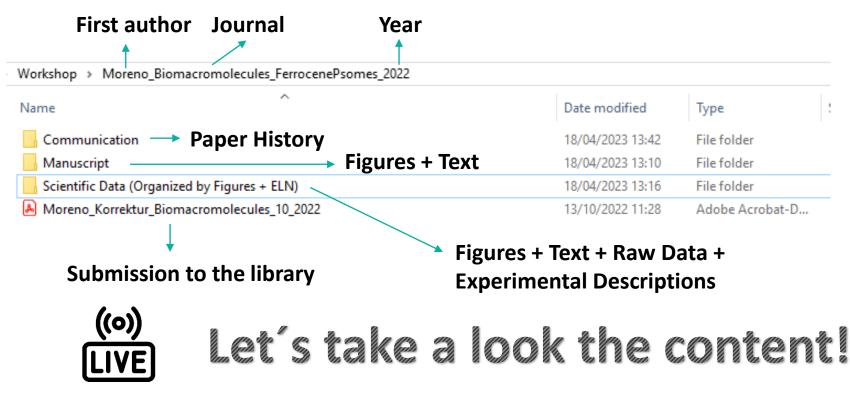
- Communication
- Manuscript (main text and SI, PowerPoint and pics of the last version of
 - Figures, TOC, Readme File)
- Scientific Data (Raw data and Processed Data ordered by Figures)
- Proof of submission to the library

The supervisor* is the responsible to check the last version











Coding and Version Control: git^[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!







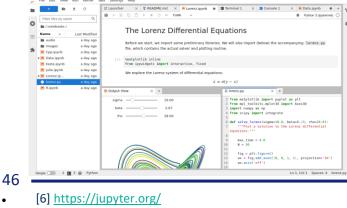
USE GIT!!!

45

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





https://jupyter.org/hub



Linear Chain Wheel		
UnearChain_conei	_rc2.4: × +	
) → ଫ 🏠	localhost:8888/notebooks/LinearChain_EShell_rC2.451.ipynb ···	☆ II\ 🖸 📽 📮 🗏
💭 jupyter	LinearChain_EShell_rC2.451 Last Checkpoint: 07/03/2020 (unsaved changes)	n Logout
File Edit	View Insert Ceil Kernel Widgets Help	Trusted Python 3 O
B + % (h 🚯 🛧 🔸 H Run 🔳 C 🗰 Markdown 🔹 📼	
In [2]:	from graceplot import a import many as no import math s True ==> interactive Mairace-Flots s Truits =>> just Batch-Made and no window MadandracePlotsTakse True	
	Setup for Linear Chain Output Fluctuation Radius of Gyration $dR_{g}^{-1}(v)de = dR_{g}^{-1}(T)/dT \cdot (-\frac{v_{g}}{1-v_{g}}) = dR_{g}^{-2}(T)/dT \cdot (-\frac{T^{-1}}{1-v_{g}})$	VS E
In [3]:	numMonomers = np.array([32, 64, 96, 128]) print("Monomers: ", numMonomers)	
	Monomers: [32 64 96 128]	
	Loading the WL-Data	
In [4]:	Loading the WL-Data	
In [4]:	Loading the WL-Data TInteractionShell = 'EShell_r(2.451'	

Tools for DMP: DSWizard^[7] and DMPTool^[8]



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:







Build your Data Management Plan



Connecting data: 5-star-linked-data^[9]

Design Pattern for Open Data

- **?**: data is openly available in some format.
- **??**: data is available in a structured format, such as Microsoft Excel file format (.xls).
- Image: data is available in a non-proprietary structured format, such as Comma-separated values (.c
- **????**: data follows W3C standards, like using RDF and employing URIs.
- **For data publishing use public repositories e.g. Zenodo:**

https://zenodo.org/







48



nature nanotechnology	View all journals Q Starch Log in
Explore content 🗸 About the journal 🗸 Publish with us 🗸	Sign up for alerts 💭 RSS feed

<u>nature</u> > <u>nature nanotechnology</u> > <u>articles</u> > article

Article Open access Published: 07 August 2023

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

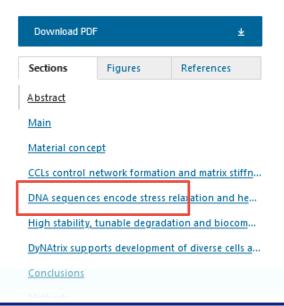
Yu-Hsuan Peng, Syuan-Ku Hsiao, Krishna Gupta, André Ruland, Günter K. Auernhammer, Manfred F. Maitz, Susanne Boye, Johanna Lattner, Claudia Gerri, Alf Honigmann, Carsten Werner & Elisha Krieg

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

15k Accesses | 12 Citations | 93 Altmetric | Metrics

An <u>Author Correction</u> to this article was published on 31 January 2024

This article has been <u>updated</u>





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 matrices with DNA:

encoded viscoelasticity for cell and organoid culture/168281). 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 (<u>https://doi.org/10.6084/m9.figshare.23309429</u>). A Python script for the statistical simulation of the maximum percentage of intramolecular crosslinks as a function of CCL complexity is available as Supplementary Code 2 and at figshare (<u>https://doi.org/10.6084/m9.figshare.23592636</u>).

Source data

Source Data Fig. 2

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

Source Data Fig. 3

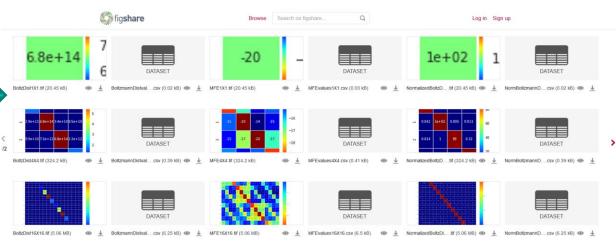
Rheological data.

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





scientific data

Explore content 👻 About the journal 👻 Publish with us 💙

nature > scientific data > comment > article

Comment Open access Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baals, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, ... Barend Mons

Scientific Data 3, Article number: 160018 (2016) Cite this article

727k Accesses | 5435 Citations | 2221 Altmetric | Metrics

An Addendum to this article was published on 19 March 2019

Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

The FAIR principle:

- Findability
- Accessibility
- Interoperability
- Reusability





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

You are here: MicroPlastiX SpecDB Dataset: MicroPlastiX SpecDB RADAR Metadata Content Statistics Technical Metadata Alternate identifier: (URL of the source repository of the project) https://github.com/robna/MPX_specDB (Citeable release snap shots of the source repository) https://doi.org/10.5281/zenodo.8314801 Lenz, Robin Ohttps://orcid.org/0000-0003-4156-7380 Pre Leibniz Institute of Polymer Research Creator/Author: Fischer, Franziska @https://orcid.org/0000-0002-2317-6784 Per Leibniz Institute of Polymer Research Arnold, Melinda ROR Leibniz Institute of Polymer Research Fernández-González, Verónica 💿 https://orcid.org/0000-0002-6890-6154 [PPR University of A Coruña] Moscoso Pérez, Carmen María () https://orcid.org/0000-0002-2451-3535 [## University of A Coruña] Andrade-Garda, José Manuel 😳 https://orcid.org/0000-0003-1020-5213 [ROR University of A Coruña] Muniategui-Lorenzo, Soledad 💿 https://orcid.org/0000-0001-5946-3366 [🕅 University of A Coruña] Fischer, Dieter [ROR Leibniz Institute of Polymer Research] Contributors: (Other) Frias, João (2) https://orcid.org/0000-0001-9943-1035 [*** Atlantic Technological University] (Other) Casotti, Raffaella (Dhttps://orcid.org/0000-0002-9876-4601 [PPR Stazione Zoologica Anton Dohrn] (Other) Pedrotti, Maria Luiza () https://orcid.org/0000-0002-9963-0402 [*** Sorbonne University] (Research Group) The JPI Oceans Project MicroPlastiX consortium members



FAIR4Chem Award 2024 Laureate Robin Lenz IPF Dresden





If you have further questions, just contact us:

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