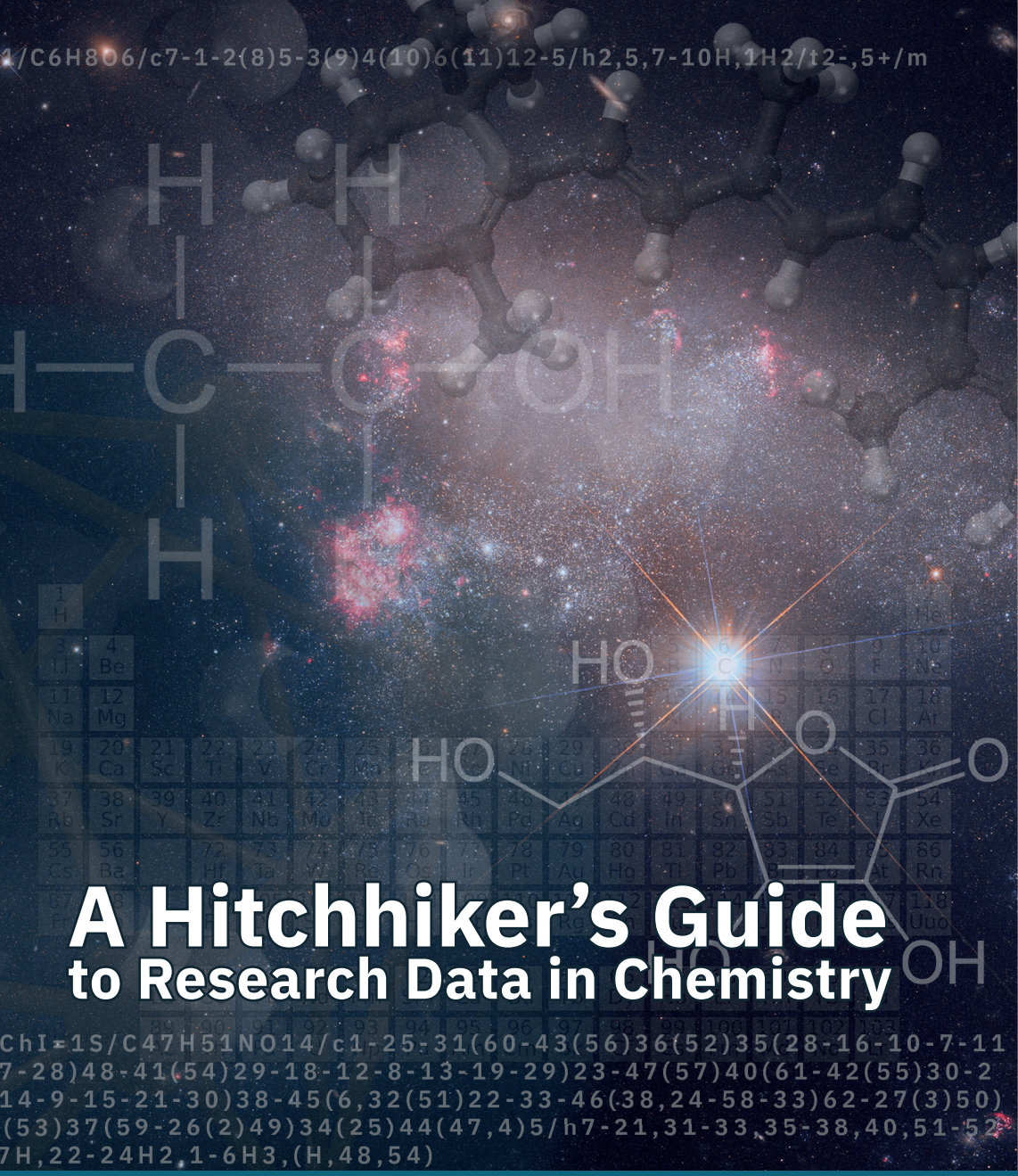
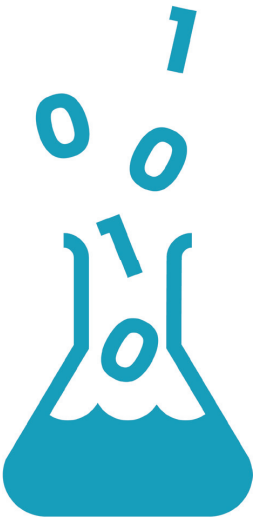


A Hitchhiker's Guide
to Research Data in Chemistry

NFDI4Chem

Research Data Management

Whenever scientists research, they produce data – like measurement data, laboratory values, texts or questionnaires. On top of that, chemists produce specific data, such as analysis data from NMR, MS or IR spectra, but also computational data like simulations and quantum chemical calculations.



As you can see, it is easy to get lost in this. But there is hope: good research data management (RDM) helps you to make your data usable throughout the entire data life cycle. It helps you collect, create, analyze, store, publish, and re-use your data.

Benefits

- RDM helps you get your data in order.
- RDM supports you over the entire data life cycle.
- RDM covers legal and organisational issues.

Open data, FAIR data

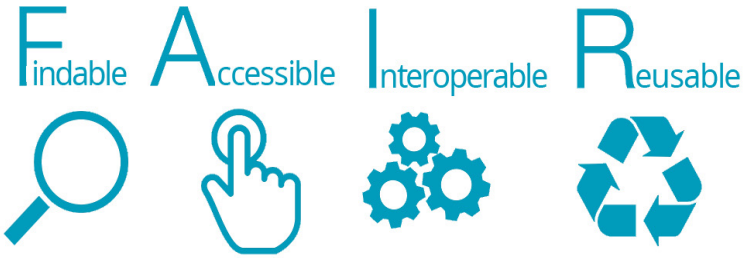
Major funders such as the DFG and the EU require researchers to include information about how research data is handled during the project as a mandatory part of a proposal. And although you still often read “data available upon request” in journals, publishing raw data is becoming more common.

Christoph Steinbeck is Professor for Analytical Chemistry, Cheminformatics, and Chemometrics at the Friedrich Schiller University in Jena, Germany.

He is Spokesperson and leads TA 1 – Management.
ORCID: 0000-0001-6966-0814

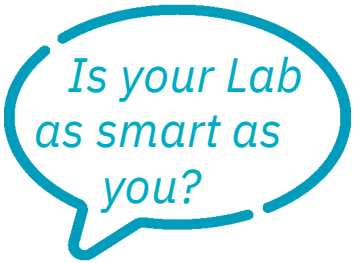
Chemists build upon the work of their predecessors. While information has been widely available since the invention of printing, data does not necessarily meet the FAIR principles, as it is frequently concealed and/or inaccessible to others.

Unpublished data, different file formats and proprietary standards prevent data from being easily reused. The FAIR principles serve as guidelines for researchers for sustainable handling of research data.



Smart Lab

A smart lab represents a holistic approach to data management. All steps within your workflow are interconnected in a digital way. Experimental data from devices are directly imported by an ELN (see next pages), that assigns all the necessary metadata and then export data seamlessly to a repository.



Sounds great? Well, the next pages will show you how to get there.

Dr. Johannes Liermann is head of the routine NMR spectroscopy lab at the Department of Chemistry of Johannes Gutenberg University Mainz.

He is Lead of TA 5 – Community Involvement
ORCID: 0000-0003-2060-842X

Data Management Plans

Sometimes a team member leaves in the middle of something. Are you prepared?

Data management plans (DMPs) help to outline how research data will be handled during and after a project to ensure preservation of the data and their metadata.

So DMPs are particularly helpful when created prior to data collection to guarantee consistent organisation, annotation, and formatting of the data. They can be used as part of quality and project management and are updated throughout the project. A number of research funders require DMPs, e.g. the ERC.

Benefits:

- A DMP supports your research because the data is findable and understandable.
- DMPs help you think ahead and consider issues of confidentiality, ethics, security, and copyright at the very beginning of the project.
- They provide continuity of knowledge and its sharing as new project members join.

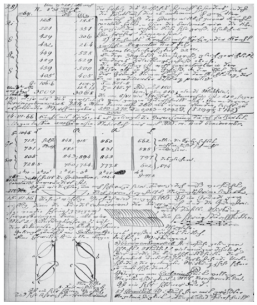
Matthias Razum is Vice President e-Research at FIZ Karlsruhe – Leibniz Institute for Information Infrastructure.

He is Lead of TA 3 – Repositories.
ORCID: 0000-0002-5139-5511

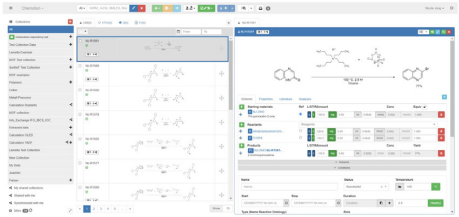
Electronic Lab Notebooks

The way research data is documented – literally with pen and paper – has made little progress over the last century. But your data has.

ELN-software supports you in your daily lab work – from documentation of the molecules used, to reactions and analysis data. Interfaces (APIs) allow data from measurement devices, such as NMR and IR instruments, to be fed directly into the ELN. ELNs help assign rich metadata to experiments, making collected data traceable and reusable.



Left: Page from the laboratory notebook of the father of experimental electrophysiology, Emil Dubois-Raymond (1818 – 1896). Right: Pages from a contemporary laboratory notebook.



Screenshot of a modern ELN Software, e.g. Chemotion. All important data are at a glance.

Benefits:

- Secure data by storage and rights management.
- Standardised metadata in human and machine-readable formats.
- Linking experimental descriptions to collected data.
- Data are findable, accessible and available, even after a change of personnel!
- Seamless transfer from ELN to the repository.

Dr. Nicole Jung is working at the Compound Platform (ComPlat) as part of the group of Stefan Bräse at KIT-Campus North.

She leads TA 2 – Smart Lab
ORCID: 0000-0001-9513-2468

Repositories

Good scientific practice wants you to save your data for at least 10 years. But did you know that some hard drives don't last that long? USB-sticks can get lost, even cloud storage cause issues, when they are abroad or providers stop the service.

So the best way to store your data for the long term is in a repository, a platform to acquire, store, archive, publish, curate, preserve, and access data, run by scientific facilities.

There are different types of repositories (generic, domain-specific, institutional), so it can be hard to find the right one.
Take a look at our overview: https://t1p.de/y8fuq

Benefits:

- Comprehensive functionality for the collection, preparation and re-use of data.
- Discipline-specific methods and data processing tools.
- Automated procedures to ease data curation.
- Functions for a seamless publication and citation of deposited data.
- Embargo settings. You decide who sees your data!

Dr. Felix Bach is head of the Research Data Department at FIZ Karlsruhe and works on repositories, data archiving and electronic lab notebooks (ELN).

He is Lead of TA 3 – Repositories
ORCID: 0000-0002-5035-7978

Metadata

Data about data is something everybody knows. *Author* and *book title* in a catalogue for instance, or *musician*, *song title*, *genre* in a playlist are metadata, which describe what content you can expect. So good metadata are necessary to find datasets in a database or other storage.

Research data is just a bit more complicated, as there are some domain-independent and some domain-specific metadata. But metadata play an important role in making your data FAIR.

```
    <subjectArea>
        <controlledSubjectAreaName>Engineering</controlledSubjectAreaName>
    </subjectArea>
</subjectAreas>
<resource resourceType="Dataset">NMR, GPC and Raman experimental data</resource>
<geoLocations>
    <geoLocation>
        <geoLocationCountry>GERMANY</geoLocationCountry>
        <geoLocationRegion>Aachen</geoLocationRegion>
    </geoLocation>
</geoLocations>
<dataSources>
    <dataSource dataSourceDetail="Instrument">Bruker Avance II (400MHz) </dataSource>
    <dataSource dataSourceDetail="Instrument">Bruker Avance III (400MHz) </dataSource>
    <dataSource dataSourceDetail="Instrument">GPCmax VE-2001 from Viscotek: two Malvern Visco
```

Extract from an XML file with metadata. These are machine and human readable.

Machine-readable metadata should be provided in a standardised format, while the metadata entities should be well-documented. Good to know: ELNs assist you to add metadata to your research data.

Benefits:

- Increases findability and interoperability of research data.
- Can be organised by using schemas (like Dublin Core).
- Minimum information standards set guidelines regarding which metadata is required when reporting data.

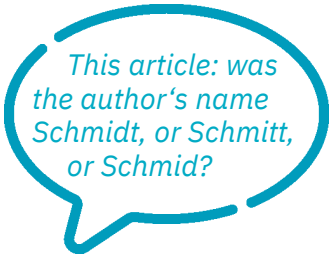
Dr. Steffen Neumann is head of the Research Group Computational Plant Biochemistry at the Leibniz Institute of Plant Biochemistry in Halle (Saale).

He leads TA 4 – Metadata & Standards
ORCID: 0000-0002-7899-7192

Persistent Identifiers

Names can be mixed up. (Web-)Addresses can get lost. Identifiers help you to identify resources – persistently.

Here's an example :
https://doi.org/10.1000/182



Persistent identifiers (PID) such as digital object identifiers (DOI) refer to published articles and research datasets in repositories. ORCID IDs refer to researchers with their affiliation.

Chemistry-specific identifiers help to make chemical structures machine-readable, like the InChI code, or can be assigned to each chemical substance, like the CAS Registry Number.

In terms of the FAIR Data Principles, identifiers should link scientific publications to their associated dataset(s). For example, a related identifier in dataset metadata references the associated journal article DOI. This helps to describe and contextualise the data.

Benefits

- Help you to identify resources.
- PIDs are permanent, no loss of links.

Dr. Oliver Koepler is head of the Lab Linked Scientific Knowledge, part of the Digital Library and Data Science research group of Prof. Sören Auer at TIB.

He is Spokesperson and Lead of TA 6 – Synergies.
ORCID: 0000-0003-3385-4232

Data Publication

In chemical research, we strive to share results with others, commonly through articles in renowned scientific journals. To be able to actually work with and build upon these results, the scientific community also requires the data that the results were based on.

Publishing and therefore sharing these research data in a FAIR manner adds value to the research results and enables discovery and reuse.

Therefore you should consider all aspects such as persistent identifiers, rich metadata, data format standards for analytical data, and by providing machine-readable chemical structures, license and provenance information.

Images of measurement data are not research data. FAIR means that you additionally publish your raw data.

Benefits:

- Raw data publications increase your career recognition.
- It enables new collaborations.
- It provides a citation advantage compared to articles without an associated dataset in a research data repository.

Prof. Sonja Herres-Pawlis holds the Chair of Bioinorganic Chemistry at RWTH Aachen University. She is Board Member of the InChI Trust.

She leads TA 5 – Community Involvement.
ORCID: 0000-0002-4354-4353

How to start

If your are an absolute beginner, a good way to start dealing with your data is a Research Data Management (RDM)-Workshop.

After that, you could build up your personal/teamwide Data Management Plan. It helps you get a structure in your data work. And a DMP is not static, you can adapt your DMP to changing conditions. A good tool to start with your first DMP is RDMO, an open-source RDM-Plattform offered by many universities: https://rdmorganiser.github.io/

Get a good storage. Remember the 3-2-1-rule: 3 Copies on 2 different storage media with 1 backup at an external location. Many universities offer external storage space. Ask your local RDM-Team.

Be sure to have an ORCID-ID. Get familiar with PIDs and DOIs.

As a next step you could look for an ELN and a repository that fit the special needs of your work. They support you with sample management, signatures, statistics and metadata. To find the right ELN, take a look at https://t1p.de/ns0t7.

Repositories you find at https://t1p.de/ewkpz

If you‘ve made it this far: congratulations! Remember to always publish FAIR, and enjoy the increasing amount of FAIR data from other researchers.

About us

NFDI4Chem is a non-profit scientific consortium consisting of chemists to build an open and FAIR infrastructure for Research Data Management in chemistry.

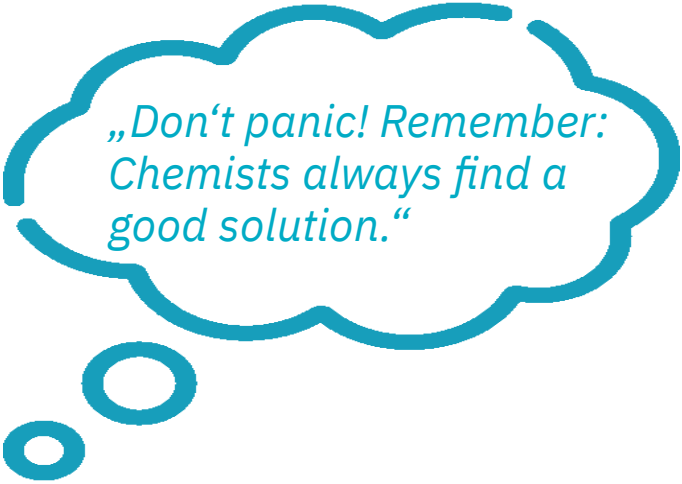
NFDI4Chem supports chemists in their efforts to collect, store, process, analyse, publish and re-use research data. We enable innovative services and science based on research data.

Our offers

- Workshops online and at universities.
- The Chemotion ELN.
- A search service for more than 90,000 chemical data sets.
- The repositories Chemotion, RADAR4Chem & nmrXiv.
- A terminology service.
- An extensive knowledge base.

Any questions?

- Contact helpdesk@nfdi4chem.de
- Visit www.nfdi4chem.de
- Find us on X (fka twitter), linkedin, insta.



NFDI4Chem

ENHANCE YOUR DATA.

Credits: P1: T. Bender with material from NASA & pixabay; P. 3 FAIR: SangyaPundir, CC BY-SA 4.0. P. 4, 6, 12: Freepic, P.5: Staatsbibliothek Berlin, 1865–1868, XIII, 22. VII. 65–9. VI. 68, reproduced with permission , Dirnagl U. and Przesdzing I., P.7: Herres-Pawlis, DOI: 10.22000/579, P. 11: Bierwagen, TIB.

Printed on 70 % PEFC, DC-COC-001015

Funded by

DFG Deutsche Forschungsgemeinschaft
German Research Foundation