NFDI4Chem—A Research Data Network for International Chemistry

by Christoph Steinbeck, Oliver Koepler, Sonja Herres-Pawlis, Felix Bach, Nicole Jung, Matthias Razum, Johannes C. Liermann, and Steffen Neumann

Research data provide evidence for the validation of scientific hypotheses in most areas of science. Open access to them is the basis for true peer review of scientific results and publications. Hence, research data are at the heart of the scientific method as a whole. The value of openly sharing research data has by now been recognized by scientists, funders and politicians. Today, new research results are increasingly obtained by drawing on existing data. Many organisations such as the Research Data Alliance (RDA), the goFAIR initiative, and not least IUPAC are supporting and promoting the collection and curation of research data. One of the remaining challenges is to find matching data sets, to understand them and to reuse them for your own purpose. As a consequence, we urgently need better research data management.

NFDI and NFDI4Chem

In 2018, the federal government of Germany and the federal states agreed to fund a national research data infrastructure (Nationale Forschungsdateninfrastruktur, NFDI). The funding should a) be long-term, b) cover all major areas of science and humanities, and c) be collaborative and coordinated [1]. The funding scheme was implemented by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) in three rounds of applications in three successive years. The consortia of the final round will be announced in November 2022, forming an NFDI with up to 30 consortia. The chemistry consortium NFDI4Chem was one of nine consortia funded in the first round in 2019 based on a funding proposal which was subsequently published in the journal *Research Ideas and Outcomes* [2].

NFDI4Chem working principles

To alleviate the lack of research data in chemistry, a seamless flow of data from the generation in the lab into institutional and public repositories is essential. A key component of NFDI4Chem is therefore the widespread adoption of Electronic Laboratory Notebooks (ELN), tightly integrated with analytical instruments in the lab and respective software and tools for data processing and analysis. In NFDI4Chem we refer to this assembly as Smart Lab. ELNs provide a convenient way for the acquisition of data in a structured, semantically annotated format, making it easy to transfer data into repositories without extra work and publish it there in a FAIR (Findable, Accessible, Interoperable and Reproducible) way for reference and reuse [3]. The NFDI4Chem infrastructure is built on parallel efforts to establish internally accepted standards for open data and metadata standards, terminologies for semantic annotation of data and legal guidelines and policies on licences for data publication.

Along with all new developments, NFDI4Chem develops teaching and training materials for all levels of researchers and students to foster the cultural change in chemistry. Here, tutorials, videos and best practices accompany the NFDI4Chem knowledge base. The federation of repositories can be searched by the overarching search service. And finally, if you need some help or support on how to use the NFDI4Chem services you can contact the NFDI4Chem Helpdesk.

Smart Lab

An important task of NFDI4Chem is to provide systems for efficient digital acquisition, storage and analysis of research data and their metadata. The concept of capturing research data at the earliest possible point in time, *i.e.* parallel to their creation in the laboratory, and the further processing of the data up to the point of publication is combined in NFDI4chem in the Smart Lab work area. The Smart Lab concept envisaged by NFDI4Chem combines the components (1) data transfer, (2) provision of an electronic laboratory notebook with measurement data integration, (3) integration of digital tools for data processing and analysis, and (4) models for transferring data and metadata to repositories. The Smart Lab is developed as a decentralized infrastructure component that is installed and operated by the respective research institutes or individual research groups. The installation will be supported by the NFDI4chem team, if required, so that the barriers to entry into the digital infrastructure are minimized.

The ELN is the main component of the Smart Lab concept and allows the decentralized components to be connected to other infrastructure components, e.g. the NFDI4Chem repositories. This makes it possible for scientists to digitally manage research data within the protected environment of their institution and prepare them for data publication in order to make them available later in repositories with little additional effort. The ELN Chemotion was selected as the

Ensuring that research data is available to use in collaboration with other scientists is a key target of the Nationale Forschungsdateninfrastruktur, (NFDI).

reference implementation in NFDI4Chem. Chemotion ELN already provides essential components of the functionality required by NFDI4Chem at the present time, and further components will be added as part of the further development of NFDI4Chem. In order to support the generation of FAIR data and to make the process of data preparation for subsequent publication as efficient as possible, various processes have been implemented in Chemotion ELN. These include methods for data conversion to open file formats, tools for data visualisation and editing, for data annotation and description of processes as well as data through metadata. Chemotion ELN differs from many other ELNs in all these areas by adapting its functionality to the specifics of chemical research, *e.g.* the need to support a FAIR description of molecular and reaction data. To this end, Chemotion ELN integrates existing standards such as SMILES and InChI identifiers of the stored molecules or jcamp.dx file formats for spectroscopic data. It aims to permanently map the work results of nationally and internationally active communities that develop suitable standards, vocabularies, ontologies and metadata schemes.

ELN and other Smart Lab components are being developed as open source software because NFDI4Chem would like to strongly support the reuse by an international community and promote collaboration with scientists worldwide.

Repositories

Repositories are essential building blocks for the NFDI, as they provide reliable access to research data across all disciplines. NFDI4Chem integrates existing and, if necessary, newly developed repositories into an interoperable, federated RDM infrastructure. The aim is to offer a comprehensive set of interconnected repositories covering all data relating to molecules, their reactions and characterization. They will enable processing, analysis, disclosure and re-use of research data in a FAIR way. The first step was to identify relevant repositories based on the following criteria:

- The repository is suitable for the deposition of molecule related data
- The repository contains either reusable data or functionality that covers the needs of the NFDI4Chem community (such as viewers, editors or analysis tools)
- The repository software is accessible as open source
- The operators of the repositories have declared their willingness to adapt their services to the standards developed by NFDI4Chem including the FAIR data principles
- The repository operators fulfil the formal requirements to be funded in accordance with the guidelines of the NFDI.

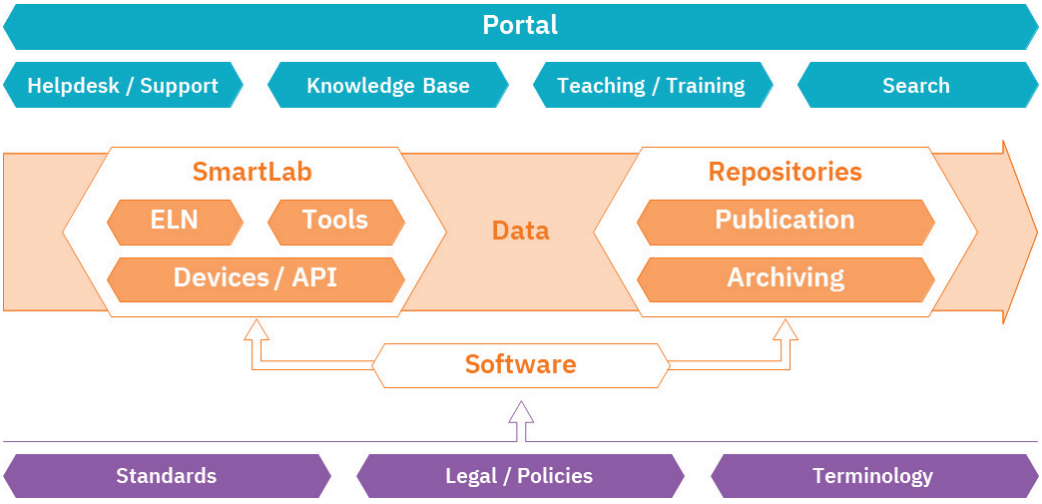


Figure 1: On the way to digitalised chemistry: key components of a digital research process

The resulting list of recommended repositories is available to the chemistry community on the NFDI4Chem portal:

- Chemotion Repository
- MassBank EU
- RADAR4Chem
- STRENDA DB
- Suprabank
- nmrXiv
- NOMAD

Over the course of our work since October 2020, we further developed the Chemotion repository at KIT as well as MassBank EU, which specifically covers data from mass spectrometry. RADAR4Chem at FIZ Karlsruhe was newly established as a "catch-all" repository for data that does not fit well into the scope of the other offers.

To address missing functionality, previously unsupported data types and poor maintainability of systems, we also develop completely new repositories. This includes nmrXiv, a much improved system for rich NMR data, replacing the former nmrshiftdb2 database. Furthermore, VibSpecDB (a database for Raman and IR spectra) is currently being developed and will soon complement our range of spectroscopy repositories. We will continue to include other repositories into the federation that cover additional sub-disciplines of chemistry or important data types.

In addition to those listed so far, however, there are other repositories that do not meet the above criteria but are highly relevant to our community, including CSD (https://www.ccdc.cam.ac.uk/solutions/csd-core/components/csd/), ICSD (https://icsd.fiz-karlsruhe.de/) and the joint CCDC (https://www.ccdc.cam.ac.uk/structures/) Access Structures Service. Here, together with the respective providers, we try to integrate these databases into our federation while fulfilling the FAIR criteria to the greatest possible extent and achieving easy access for our community.

By federating the repositories, we achieve several goals: we enable cross-searching, support single sign-on for users, provide uniform programming interfaces for integration with ELN and analysis and validation tools, and enable the linking of distributed information to a molecule across repository boundaries.

Furthermore, NFDI4Chem cooperates with major publishers and editors in chemistry to come up with recommendations for suitable data repositories in author guidelines, thus contributing to an improved publication process which substitutes current supplementary information with references to datasets in repositories.

Standards, International scope and collaboration with IUPAC

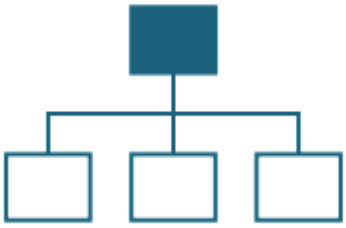
Although the NFDI is funded by German funding agencies, the work towards standardization in any field of science is always a global endeavour. Whenever the need for standardisation arises, an international group of stakeholders can take the lead in designing a draft standard which is then ideally agreed upon by a wider audience in rounds of consultation. NFDI4Chem has pledged to work with the IUPAC, RDA, goFAIR and other international organisations dealing with standards in research data management in this respect.

NFDI4Chem addresses standardization efforts on different levels. First, the representation of molecular structural information using inline notation as SMILES [4] or the later InChI [5]. For organic molecules SMILES and InChI cover more than 95 % of the molecules but for inorganic molecules the situation is not satisfying. Hence, NFDI4Chem intensively contributes to the further development of the InChI, especially in the molecular inorganic subcommittee to solve the problem of the molecular representation of metal-donor bonds. Only when all molecular compounds can be unambiguously identified by the InChI can they be digitally linked and knowledge on their properties and reactions be connected. Secondly, the characterisation of the molecules by various spectroscopic methods also needs minimum information standards. In numerous working groups, NFDI4Chem elaborates minimum information standards. For single crystal X-ray diffraction, cif files and their deposition are established for more than 20 years [6], for NMR spectroscopy and mass spectrometry standards are developed but not yet used or known in the whole community *(vide infra).* For further methods, such as UV/Vis spectroscopy, EPR spectroscopy or cyclic voltammetry, standards are urgently needed. Third, we jointly discuss the extension of metadata standards like DataCite or Bioschemas.org to incorporate chemistry specific metadata fields. These metadata formats play an important role when it comes to data repositories and interdisciplinary re-use of data.

With regards to our cooperation with IUPAC another example is the curation of ontologies, which often includes definitions of the terms in the ontology. Here the IUPAC Gold book as the compendium for chemical terminology and reference for chemists worldwide comes into play. Not only can entries from the Gold book be referenced in term definitions of ontologies, the further development of the Gold book can also benefit from recommendations for missing terms identified in the process of ontology curation. A further example is our continuous contribution to the WorldFAIR initiative of IUPAC and CODATA. Herefore, a cookbook for chemists on FAIR data is developed as a free and curated online resource.

Together with several IUPAC colleagues, we formulated a call to the chemical community to actively involve in the evolution of standards for all methods used in chemistry. Only international agreed standards for every method will open up avenues to digitization in chemistry.

Ontologies, Linked Data, and Knowledge Graphs



The increased availability of Big Data in chemistry demands not only for machine-readable but machine-interpretable data to fully support data-driven research. To get the most out of research data, we need to break down data silos and work towards a full data integration. In conjunction with the development of standards for research data NFDI4Chem fosters the semantification of data. To achieve these goals we need to describe our data comprehensively and unambiguously with metadata. Metadata adds the context of the why, how, when, where and by whom to data. Agreeing on shared concepts to describe entities and relations we can describe our domain in a structured way expressing statements in the form of subject, predicate and object. Ontologies help us to semantically describe data by providing terms, relations and logic to link data and building knowledge graphs. As a first step, Ontologies4Chem provided an overview of ontologies suitable for describing research and research data [7]. From there we will contribute to existing ontologies and develop new ones for new scopes identified within the work of NFDI4Chem. Our activities are embedded in the community. The first international Ontologies4Chem workshop brought together the ontology community, chemists and service developers discussing ongoing developments and future cooperation on the way to a general roadmap for chemistry ontologies for research data management [8].

Community services by NFDI4Chem

NFDI4Chem develops and offers a comprehensive suite of services accessible through the NFDI4Chem portal. These services incorporate all standards, guidelines and policies that are jointly developed with national and international partners like NFDI, goFAIR, RDA or IUPAC. Services are interlinked, starting from the Smart Lab with the ELNs towards the individual data repositories to the federation of NFDI4Chem repositories searchable by the NFDI4Chem search service. The search service harvests metadata from all repositories in the federation providing a single point of entry for an initial query, i.e. for all available data for a given molecule in the federation. For harvesting, all repositories need to agree and apply common metadata and API standards. Additionally, data semantically annotated and linked

using ontologies, terminologies and controlled vocabulary provided by the terminology service, which will be connected with chemotion ELN and the various data repositories to enable semantic data annotation from the very beginning. The terminology service provides a comprehensive collection of ontologies relevant for chemistry and can be browsed and searched by human users and also by machines using the API. Future developments will add further plugins enabling ontology curation, mapping, and design within the interface of the service. Standing by the side of these technical services are the community services of the Helpdesk and the knowledge base. The knowledge base is a place full of knowledge regarding research data management in Chemistry, where users can find information and further ressources. Besides the NFDI4Chem services, a plethora of smaller tools and widget to support researchers on their daily endeavours with research data are available on github.

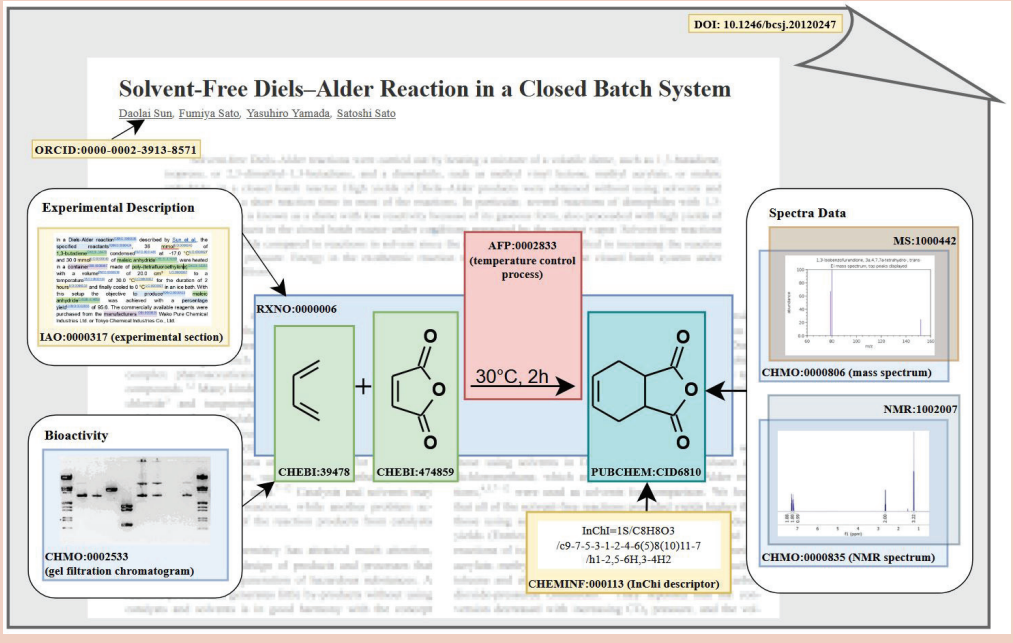
NMR data deposition and standards. A use case for the collaboration of IUPAC and NFDI4Chem

In most chemical subdisciplines, the vast majority of generated data originates from analytical methods like NMR spectroscopy. It uniquely allows the study of properties and interactions of nuclei in chemical systems such

Ontologies4Chem: the landscape of ontologies in chemistry

Philip Strömert, Johannes Hunold, André Castro, Steffen Neumann and Oliver Koepler

Ontologies are a holistic approach to semantically describe the ever growing maze of data, information and knowledge in a domain. They provide terms, relationships, and logic to semantically annotate and link data to create knowledge graphs. While domain experts, by virtue of their training and implicit knowledge, should be able to perceive and interpret the semantics expressed in text, tables, and images of articles and their experimental sections, computers cannot fully do so without fine-grained metadata annotations. Therefore data must be made machine-readable and machine-interpretable right from the start. A proper semantic description of an investigation and the why, how, where, when, and by whom data was produced in conjunction with the description and representation of research data is a natural outcome in contrast to the retrospective processing of research publications as we know it. See full text in ref. 7 or doi.org/10.1515/pac-2021-2007



Semantics hidden in a research article

as molecules or solids and has become a technique with widespread applications in synthetic chemistry, natural products chemistry, physical chemistry, biochemistry, or material science among many others.

A major challenge in making NMR data FAIR-compliant is the lack of up-to-date data formats and standards. IUPAC committees propagated the development of JCAMP-DX [8,9] for NMR and MS data. Despite being outdated and not adequately defined for multidimensional NMR data, JCAMP-DX is still unsurpassed by other attempts at standardised formats for raw and processed NMR spectra. While the Bruker NMR data format may easily be interpreted, it is an awkward choice for vendor-independent deposition of NMR spectra as it has a complex folder structure and contains much technical information of little interest to the public.

Moreover, not only in the recorded spectra but also their analysis (*i.e.* the assignment of chemical shifts, couplings, as well as other correlations and interactions detectable by NMR) has a value on its own. While the current IUPAC recommendations for reporting NMR data essentially date back to the 1970s [9] [10], *a* promising step in the right direction is the development of NMReData [11]. It is a markup format describing molecular structures along with their NMR properties like chemical shifts, coupling constants, or two-dimensional correlations.

NFDI4Chem aims at improving the handling of digital NMR and spectral analysis data on different levels. We are involved in many NMR-related projects (NMRium: an open source web-based NMR visualizer, nmrXiv: a modern NMR repository, the IUPAC project FAIRspec). On the technical level, we work on automatic conversion services to enable the seamless integration of different digital NMR tools. Also, the development of standards and terminologies for NMR is an important topic of our work.

The different layers of NMR research data can serve as showcases where both technical or practical implementations as well as standardisation efforts are required, thus being an excellent use case for the close collaboration between NFDI4Chem and IUPAC.

Conclusion

NFDI4Chem is building a national research data infrastructure for chemistry in Germany. Neither the use of the resulting infrastructure nor the development of standards needed for it is restricted to this country. Scientists from all over the world can use the terminology service, the knowledge base about research data management, repositories such as Chemotion or the respective ELN and standards used in all NFDI4Chem products are developed through international collaboration.

References

1. Hartl N, Wössner E, Sure-Vetter Y: Nationale Forschungsdateninfrastruktur (NFDI). *Informatik Spektrum* 2021, 44:370–373.
2. Steinbeck C, Koepler O, Bach F, Herres-Pawlis S, Jung N, Liermann J, Neumann S, Razum M, Baldauf C, Biedermann F, et al.: NFDI4Chem-Towards a National Research Data Infrastructure for Chemistry in Germany. *Research Ideas and Outcomes* 2020, 6:e55852.
3. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al.: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016, 3:160018.
4. Weininger D: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Model* 1988, 28:31–36.
5. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D: InChI, the IUPAC International Chemical Identifier. *J Cheminform* 2015, 7:23.
6. Hall SR, Allen FH, Brown ID: The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallogr A* 1991, 47:655–685.
7. Strömert P, Hunold J, Castro A, Neumann S, Koepler O: Ontologies4Chem: the landscape of ontologies in chemistry. *Pure Appl Chem* 2022, doi:10.1515/pac-2021-2007. (see insert)
8. Strömert P, Hunold J, Koepler O: 1st Ontologies4Chem Workshop – Ontologies for chemistry. 2022, doi:10.25798/FRNP-SN04.
9. Recommendations for the Presentation of NMR Data for Publication in Chemical Journals. *J Macromol Sci Part A Pure Appl Chem* 1972, 29:625–628.
10. Presentation of NMR data for publication in chemical journals - B. conventions relating to spectra from nuclei other than protons. *Pure Appl Chem* 1976, 45:217–220.
11. Pupier M, Nuzillard J-M, Wist J, Schlörer NE, Kuhn S, Erdelyi M, Steinbeck C, Williams AJ, Butts C, Claridge TDW, et al.: NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. *Magn Reson Chem* 2018, 56:703–715.

Christoph Steinbeck <christoph.steinbeck@uni-jena.de>, Analytical Chemistry - Cheminformatics and Chemometrics and Vice President for Digitalisation of the Friedrich-Schiller-University Jena, Germany
http://orcid.org/0000-0001-6966-0814

Oliver Koepler <Oliver.Koepler@tib.eu>, TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
https://orcid.org/0000-0003-3385-4232