NFDI4Chem - A National Research Data Infrastructure for Chemistry

Extended Abstract for 1st NFDI Conference, 13-14th may 2019, Bonn, Germany

1. Formal details

Planned title of the consortium
Fachkonsortium Chemie für die Nationale Forschungsdateninfrastruktur

Acronym of the planned consortium
NFDI4Chem

Lead institution or facility
TBA

Name and work address of a contact person (including email address and institutional affiliation)
Dr. Oliver Koepler, Leibniz Information Center for Science and Technology (TIB), contact@nfdi4chem.de

Participants in the NFDI conference (names, institutional affiliation and email address; max. 3 persons)

- Dr. Oliver Koepler, Leibniz Information Center for Science and Technology (TIB), oliver.koepler@tib.eu
- Prof. Dr. Sonja Herres-Pawlis, RWTH Aachen University, sonja.herres-pawlis@ac.rwth-aachen.de
- Prof. Dr. Christoph Steinbeck, Friedrich-Schiller-Universität Jena, christoph.steinbeck@uni-jena.de

Research area of the planned consortium
Research area 31, Chemistry

Participating research institutions (without an address)
- Leibniz Information Center for Science and Technology (TIB)
- Karlsruhe Institute of Technology (KIT)
- RWTH Aachen University
- Friedrich-Schiller-Universität Jena
- Universität Leipzig
- Fritz-Haber-Institut der MPG
- Universität Stuttgart
- Leibniz-Institut für Pflanzenbiochemie Halle (Saale) - IPB
- Johannes Gutenberg-Universität Mainz
- LMU München
- Universität zu Köln

- Georg-August-Universität Göttingen
- Technische Universität Dortmund
- Albert-Ludwigs-Universität Freiburg
- Christian-Albrechts-Universität zu Kiel
- Forschungszentrum Jülich
- Bundesanstalt für Materialforschung -und prüfung (BAM)
- Fraunhofer-Institut für Angewandte Polymerforschung

Participating infrastructure facilities and/or potential information service providers (without address)

- Leibniz Information Center for Science and Technology (TIB)
- FIZ Karlsruhe - Leibniz Institute for information Infrastructure
- Karlsruhe Institute of Technology (KIT), Steinbruch Computer Center (SCC), KIT Library
- Technische Universität Dresden (TUD), Centre for Information Services and High Performance Computing (ZIH)
- RWTH Aachen University
- Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Participating learned Societies
- Gesellschaft Deutscher Chemiker (GDCh)
- Deutsche Bunsen-Gesellschaft für Physikalische Chemie (DBG)
- Deutsche Pharmazeutische Gesellschaft (DPhG)

Other institutions
- Beilstein-Institut

Planned proposal submission date (2019, 2020, 2021)
2019

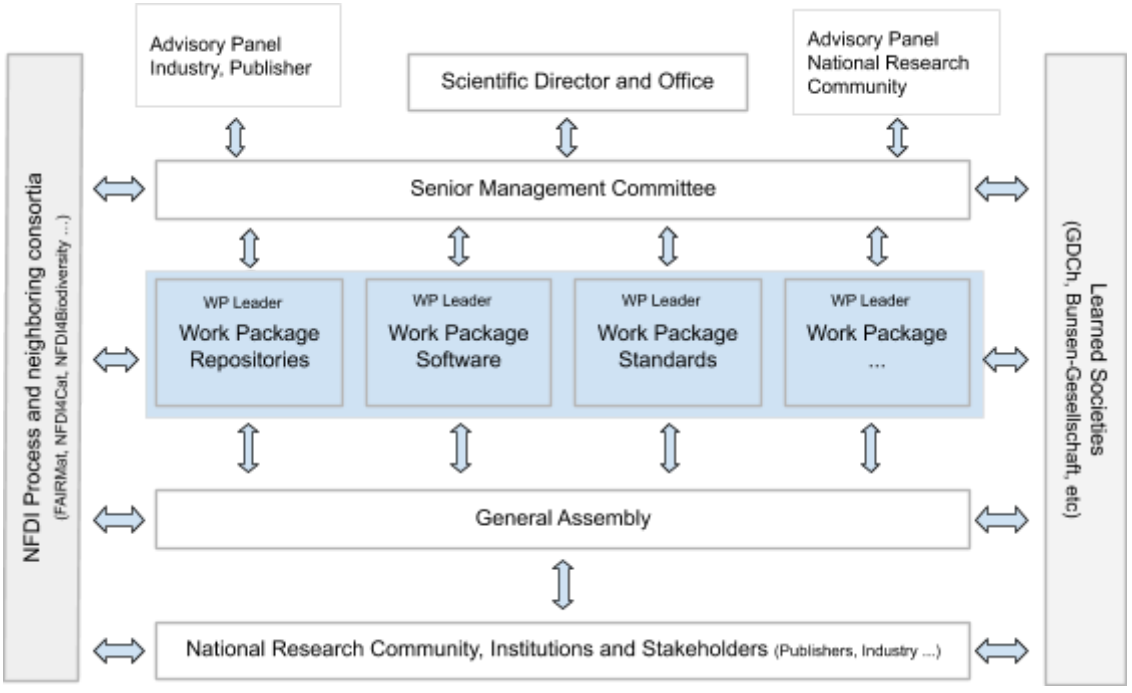

Figure: Draft organisational chart for the NDFI4Chem consortium

2. Subject-specific and infrastructural focus of the planned consortium

2.1. Key questions/objectives of the consortium

The vision of NFDI4Chem is the application of digitisation principles to all key steps of research in chemistry. NFDI4Chem supports scientists in their efforts to collect, store, process, analyse, disclose and re-use research data. Measures to promote Open Science and Research Data Management (RDM) in agreement with the FAIR data principles are fundamental aspects of NFDI4Chem to serve the community with a holistic concept for access to research data. To this end, the **overarching objective** is the development and maintenance of a national research data infrastructure for the research domain of chemistry in Germany, and to enable innovative services and science based on research data. NFDI4Chem intends to represent all disciplines of chemistry in academia. We aim to collaborate closely with thematically related consortia. In the initial phase, NFDI4Chem focuses on molecules and data for their characterisation and reactions, both experimental and theoretical.

This overarching goal is achieved by working towards a number of key objectives:
Objective 1: Connect existing data repositories and, based on a requirements analysis, build one or multiple domain-specific research data repositories for the national research community, and link them to international repositories.
Objective 2: Initiate international community processes to establish minimum information (MI) standards for data and machine-readable metadata, where missing, in key areas of chemistry, as well as missing open data standards, in order to support the FAIR principles for research data in NFDI4Chem.
Objective 3: Foster the development and adoption of Electronic Laboratory Notebooks (ELN), software, tools and Application Programming Interfaces (APIs) between commonly used instrumentation and software towards an embedded, digital information architecture to help researchers to capture research data in well-annotated electronic form at the earliest possible point in time in the research process.
Objective 4: Engage with the chemistry community in Germany through a wide range of measures to create awareness for and foster the adoption of FAIR data management. Initiate processes to integrate research data management (RDM) and data science into curricula. Offer a wide range of NFDI4Chem-related training opportunities for researchers.
Objective 5: Maintain a close relationship with neighbouring NFDI consortia to avoid duplicate development and exploit synergies.
Objective 6: Engage with experts to explore the legal aspect of FAIR research data management, design and develop the NFDI4Chem accordingly, and to offer advice for the research community.

2.2. Known needs/current status of research data management in chemistry

2.2.1. From a research perspective

Chemistry consists of many subdisciplines with a large variety of methods and data. Thus, it has very heterogeneous needs concerning RDM. The huge diversity of experimental and theoretical methods (e.g. NMR, IR, UV/VIS, MS, HPLC, Electron Microscopy, bioactivity assays, quantum and force-field calculations, cheminformatics approaches) results in many different data formats, most of them being proprietary. Open data and metadata formats which are essential for the exchange between measurement devices, analysis programs and repositories, are rare, not widely supported and mostly do not cover all functionalities of the proprietary formats. Missing machine-readable metadata standards are a fundamental barrier

to find and access data. In addition to the complex situation of data formats, there are still several non-digital working steps in the research process using pen and paper. Most labs still use handwritten lab notebooks to document research. In addition to the technical hurdles, there is a strong reluctance among researchers to publish their research data in repositories. Research data are an essential part of scientific articles and these data are presently not shared with others before the publication and only a small fraction afterwards. On the other hand, chemistry is a very fundamental discipline that delivers data and insights to many other disciplines (biology, medicine, materials etc.). This stresses the need for sustainable data management to provide multiple opportunities for exchange with other disciplines. To address the technical challenges and to foster acceptance by scientists, data acquisition in FAIR and open data formats need to be established continuously over the research data lifecycle, beginning at the earliest point in time in the research process at the lab bench. Electronic Laboratory Notebooks (ELN), software, tools, repositories, and interfaces to instruments form an information architecture for the exchange of data through open APIs. ELN can operate as a central hub within the architecture to manage data acquisition, analysis, annotation with metadata and export to repositories. Like the minimum information standards in biology, the chemistry community has to develop metadata standards to semantically describe experiments and simulations, molecule characterisations and others. These standards need to be a) complete and comprehensive enough to allow others to understand experiments/simulations and their boundary conditions to allow for re-use and b) concise enough to avoid overload of researchers who have to produce these data annotations, beyond what is possible automatically, in the first place. To support domain cross-linking, big-data analysis and future artificial intelligence (AI) methods, metadata has to be machine-readable and interpretable. Agreeing on data and metadata standards in a particular research domain is an international effort which can partly be stimulated by NFDI4Chem but ultimately only is achieved through collaboration with scientists and standardization bodies such as the IUPAC, the Research Data Alliance (RDA) or the GO FAIR initiative.

2.2.2. In terms of available information providers and services

In general, the dissemination and application of FAIR RDM services and repositories in chemistry are still at the beginning, especially in Germany. Reasons are manifold; missing data and metadata standards, insufficient data quality, low data coverage, missing search functionalities or missing scientific acknowledgement for data publications resulting in low acceptance of RDM services. Nevertheless, successful RDM exists in few sub-disciplines such as the Cambridge Structural Database (CSD) for small-molecule crystal structures, a database well accepted within the community and the publication process. In three preparatory workshops in April 2018, October 2018, and March 2019, we identified existing national components (repositories, databases, data standards, software including ELN) on which we propose to build the initial NFDI4Chem information architecture. Most of those components exist as island solutions built to serve a particular purpose and need to be developed in order to operate under FAIR principles in the NFDI4Chem ecosystem. First components identified are, for example, the Chemotion repository (molecule characterisation and chemical reaction data), NMRShiftDB (NMR data), MassBank (MS data), STRENDA-DB (enzymology data) and RADAR as a generic data repository. The assessment does not only hold for databases and repositories but also software libraries such as the Chemistry Development Kit (CDK) and RDKit or workflow frameworks where inevitably new code will need to be developed.

2.3. Summary of the planned research data infrastructure that is specifically intended to address the needs of research users in their respective work processes

Establishing and maintaining an interoperable network of domain-specific FAIR research data repositories and supporting services for the national research community in Germany is at the heart of NFDI4Chem. We will address this core mission with several objectives, which will ultimately lead to a cultural change in chemistry. On an infrastructure level, we will connect existing open data repositories and, based on the requirements analysis, establish one or multiple domain-specific research data repositories for the national research community in Germany. On a data level, we promote the use of open data and metadata standards. Where missing, we will initiate international community processes to develop and establish MI standards for open data and metadata in key areas of chemistry. Quality assurance and curation of data is an essential part of an information infrastructure. We will advance the development of vocabularies and ontologies to semantically annotate research data as key elements to connect and search heterogeneous data repositories. On a software level, we promote the development and adoption of Electronic Laboratory Notebooks, repositories, software, tools, instrumentation, connected through interfaces. The goal is to move towards an embedded, digital information architecture within the lab environment to help researchers to capture research data in well-annotated electronic form at the earliest time point in the research process. ELN constitute central tools within the workflow to manage data acquisition, analysis, annotation with metadata and export to repositories. While the focus will be on open source software and open data formats, the NFDI4Chem information infrastructure aims to be used side by side with any software and tool already in use by the community.

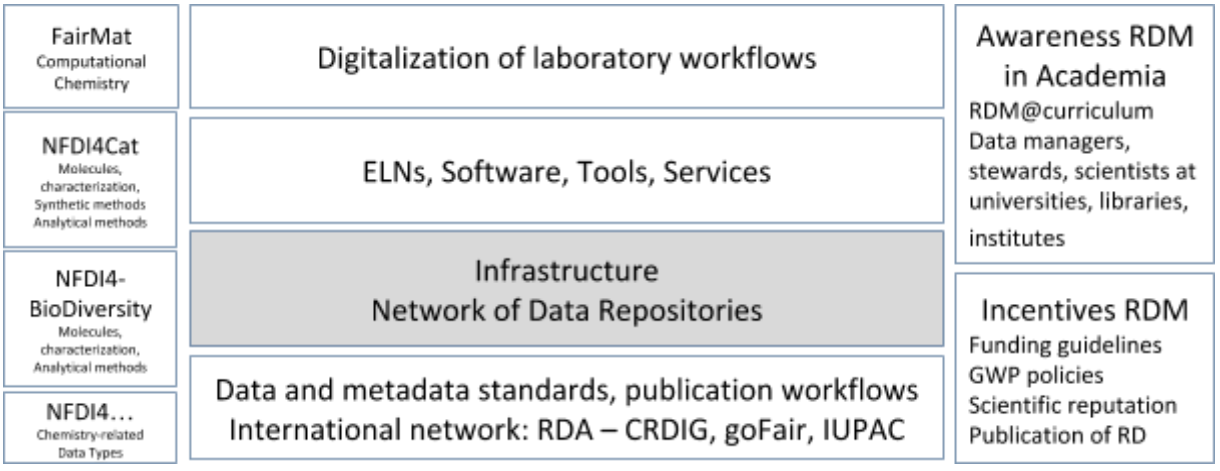

Figure: Working areas of NFDI4Chem, The left column lists related consortia followed by overlapping topics.

The outlined infrastructure enables thorough digitisation of all essential steps in everyday laboratory work and handling of research data. The technological advancement will increase the awareness and acceptance of RDM in the research community. This change in scientific culture and mindsets must not be underestimated. We will, therefore, engage with the chemistry research community in Germany through a wide range of measures and participation models to create awareness for and cultivate the adoption of FAIR RDM. NFDI4Chem provides training opportunities for best practices for RDM. Together with publishers, we will initiate a process towards a transition in the scientific publication process including a combined publication of research data in data repositories, linked with traditional article publications. Within the NFDI infrastructure, we stimulate the enhancement of funding policies and good scientific practice towards open science. Long-term measures are the

implementation of teaching RDM skills at universities, new roles of data managers, data stewards, and data scientists at universities, libraries, and research facilities.

2.4. Description of data types

Chemistry consists of numerous subdisciplines with varying types, amounts, and size of data. The large diversity of experimental and theoretical methods in chemistry results in many different data formats, most of them are proprietary. Software for analysis is often proprietary as well, occasionally provides export to standard file formats. Theoretical data are produced in various file formats that are mostly interoperable. In addition, a large amount of data are still documented manually in text files. Molecular structures are commonly described using proprietary data formats, typically SDF files, but open file formats such as Chemical Markup Language exist. Spectroscopic methods such as NMR, MS, UV/Vis, EPR, IR, Raman employ mostly proprietary machine-typical formats which need to be converted for further analyses and exchange. Common open standards for data storage need to be defined in close cooperation with leading instrument vendors and software developers. JCAMP/JCAMP-DX has been established as an exchange format for chemical and spectroscopic information, but originates from the 1980s and is limited in scope. Modern developments for NMR spectroscopy are NMReData, a reporting standard, and nmrML, a data exchange format, which is not broadly used yet. In biocatalytic research FAIR “Standards for Reporting Enzymology Data" (STRENDA) for data and metadata have been proposed. Overall open data formats will enable the exchange of data and metadata between ELN, modelling platforms, and public repositories.

2.5. Description of underlying data processing and data analysis methodologies

Processing and analysis of data in chemistry occur on multiple levels of data science from early data processing via data cleaning to full-fledged data analysis and machine learning.
For some methods such as single crystal X-ray diffraction, a fully electronic workflow has been established several years ago: data are electronically collected, analysed, refined, and finally deposited in the CCDC or in the ICSD after curation. Other methods such as mass spectrometry partly work non-digital, analysed manually on paper, and the results typed into text files. For theoretical applications, fully electronic workflows are established, although still requiring standardisation. Hence, a large range of workflows, also including simulation and cheminformatics tools, need to be addressed by NFDI4Chem in order to reflect the requirements of the community. Besides this workflow centric approach, molecules are *the* chemistry-specific key to link different data processing and data analysis methodologies.
In synthetic chemistry molecules are characterised by various methods, using several spectroscopic methods to unambiguously identify a molecule structure. Such accompanying information is organised in molecule-centric datasets. Overall, different perspectives must always be rendered possible such as the view on the molecule with all available data belonging to this entity and being stored in a repository, or the view on the method which allows to build up curated databases for X-ray data, NMR data etc.
Ultimately, molecules are *the* unique metadatum in chemistry and the bridge to other molecules-based communities like NFDI4Cat, NFDI4Ing, NFDI4genome, FAIRMat and more.

2.6. Planned implementation of the FAIR principles and information about any existing policies or guidelines in the relevant discipline

FAIR principles will be the driving force behind all of the consortium's activities. The FAIR principles are applied in the design of the NFDI4Chem information architecture, development of software and APIs, discussion and use of open data and metadata standards and recommendations for data management plans. NFDI4Chem will provide guidelines and best practices to the community on how to apply FAIR principles. This includes questions about the handling of research data during and after the end of a project, what data will be

collected, processed and/or generated, which methodology and standards will be applied, whether data will be shared/made open access and how data will be curated and preserved. Wherever open standards do not exist, proprietary data formats will be accepted but dedicated processes will be initiated on an international level to convert such data into open formats over time. Senior management members of NFDI4Chem are involved in two implementation networks (Chemistry, Metabolomics) in the GO FAIR initiative. We will exploit these connections to work efficiently with the international RDM community towards full support of FAIR principles in NFDI4Chem.

2.7. Planned measures for user participation and involvement

NFDI4Chem has started as a grassroots initiative driven by experts in the field after the first position paper by the German Council for Scientific Information Infrastructures (RfII) to establish a national research data infrastructure for Germany. It has therefore already been maximally inclusive and consulted a wide range of user communities in chemistry in Germany. NFDI4Chem is supported by the German Chemical Society (GDCh), German Bunsen Society for Physical Chemistry (DBG) and German Pharmaceutical Society (DPhG) to reach out to the chemistry community as a whole. Apart from established outreach instruments such as workshops, conferences, tutorials and training material, feedback mechanisms ranging from electronic surveys via issue trackers to social media elements will be explored. We further expect public policy, funders and learned societies to increase their demand for FAIR and open data management which will naturally increase the incentive for users to engage with these ideas. An iterative approach is envisioned to bring innovations to users: early-adopters in the consortium test new developments, while at a later stage beta-users from the NFDI4Chem community evaluate more stable software versions. Final releases are disseminated to the whole community through regional workshops and tutorials ("NFDI4Chem on Tour"). To teach the next generation of chemists, concepts and best practices of RDM, NFDI4Chem promotes the implementation of learning units into Bachelor and Master curricula. This will be supported by learned societies with experience in curricular recommendations.

2.8. Existing and intended degree of networking of the planned consortium

2.8.1. Networking Nationally (in particular with other, potential future consortia or existing state-level initiatives)

During the NFDI preparation phase since Autumn 2018, we have identified a number of related consortia with thematic overlap. These include but are not necessarily limited to NFDI4Ing, NFDI4Cat, FAIRMat, NFDI4BioDiversity, NFDI4genome, NFDI4life and related. Furthermore, there are general requirements for an RDM infrastructure such as data and metadata standards, procedures to reach community agreement on data, and legal aspects of research data, which should be discussed among all NFDI consortia. We will engage both with thematically related consortia as well as with the NFDI process in general, to grant a data infrastructure for interdisciplinary research disciplines located in between the scope of NFDI4Chem and other consortia. The objective of these processes is to inform each other, agree upon best practices in common areas and to avoid duplicate developments and efforts.

2.8.2. Networking internationally

To avoid national island solutions, a number of aspects of building a national research data management infrastructure need to be coordinated at an international level. This applies in particular to the development and implementation of open data and metadata formats. Wherever movements for the development of standards can be identified internationally (e.g. Pistoia), NFDI4Chem will engage with the relevant parties and help drive the development in a meritocratic way. Such developments will be fostered through project groups formed

specifically for the task at hand (for example to develop a markup language for the representation of NMR data) or through bodies dedicated to standardisation (such as IUPAC) or research data in particular (such as RDA). In concrete terms, NFDI4Chem has links to the Committee on Publications and Chemoinformatics Data Standards (CPCDS, IUPAC) and to learned societies worldwide (EuChemS, RCS, ACS, CCS) via GDCh, and the Chemistry Research Data Interest Group within the Research Data Alliance (CRDIG, RDA). Wherever RDM is offered by well-established and community-accepted repositories and data services, NFDI4Chem will not duplicate those efforts but rather engage in potentially bilateral data exchange and in the negotiation and establishment of open interfaces with such international offers. Integration with the European Open Science Cloud (EOSC) will be closely investigated.

2.8.3. Networking between the infrastructure facilities and the research community

Participants of NFDI4Chem reflect a good balance between the research community and infrastructure institutions. Networking with the community will be deepened and broadened by user surveys, interviews and elaborated use cases. NFDI4Chem collects the concrete requirements of the researchers and forwards these to the infrastructure facilities. After implementation of new services and tools, these will be operated, maintained, and further developed by the consortium according to further requirements and user feedback. NFDI4Chem organises regular tutorials and workshops to communicate the latest developments to the community. The communication in all paths is facilitated by learned societies like the GDCh, Bunsen-Gesellschaft or DPhG. Industry forms an important part of the research community and industry best practices will be integrated into the learning process in NFDI4Chem.

2.8.4. Major networking topics

Chemical concepts and methodologies are integral parts of multiple further disciplines. Chemistry itself branches out into many sub-disciplines, all connected by the concept of molecules. NFDI4Chem uses this domain-linking concept to identify and describe “research data commons” with related consortia, e.g. FairMat, NFDI4Cat, NFDI4BioDiversity or NFDI4Ing. The science data centre MoMaF is an interdisciplinary approach that will have an impact in NFDI4Ing and NFDI4Chem and could be adapted by others. Semantically annotated data provide a solid fundament for cross-domain data integration services, establishing a common understanding of data and capture domain-specific semantics by defining concepts, associated attributes and relations. Data mappings to vocabularies enable data integration (e.g. data networking, federated access) and new explorations (semantic search, visualisation). Creating a sustainable infrastructure can only be achieved in a joint effort by all related consortia.

2.8.5. Additional information

NFDI4Chem aims to address legal aspects of RDM and provide support for the chemistry community, e.g. legal questions of the researchers about data ownership, legally compliant operation of the NFDI infrastructures, and the development of science-friendly guidelines for RDM. We assume that there will be similar legal issues in other consortia at a higher level. We propose a joint approach to the fundamental issues, offering support to other consortia.