1 Binding letter of intent as advance notification or non-binding letter of intent

This is a binding letter of intent for a proposal in 2019

2 Formal details

Planned name of the consortium

Fachkonsortium Chemie für die Nationale Forschungsdateninfrastruktur

Acronym of the planned consortium

NFDI4Chem

Applicant institution

Friedrich Schiller Universität Jena, Fürstengraben 1, 07743 Jena

Prof. Dr. Walter Rosental

Spokespersons

Prof. Dr. Christoph Steinbeck, christoph.steinbeck@uni-jena.de, Institut für Anorganische und Analytische Chemie, Lessingstr. 8, 07743 Jena

Dr. Oliver Koepler, oliver.koepler@tib.eu, Technische Informationsbibliothek, Welfengarten 1B, 30167 Hannover

3 Objectives, work programme and research environment

Research area of the proposed consortium (according to the DFG classification system)

Research area 31, Chemistry

Concise summary of the planned consortium's main objectives and task areas

The vision of NFDI4Chem is the digitalisation of all key steps in chemical research. 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 aims 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 novel scientific approaches based on re-use of 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 data related to molecules and reactions including data for their experimental and theoretical characterisation.

This overarching goal is achieved by working towards a number of key objectives:

Objective 1: Establish a virtual environment of federated repositories for searching, commenting and exchanging research data across distributed data sources. 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 as well as open data standards in key areas of chemistry, where missing, in order to support the FAIR principles for research data.

Objective 3: Foster cultural and digital change towards Smart Laboratory Environments by promoting the use of digital tools in all stages of research and promote subsequent research data management at all levels of academia, beginning in undergraduate studies curricula.

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 training opportunities for researchers.

Objective 5: Explore synergies with other consortia and promote cross-cutting development within the NFDI

Objective 6: Provide a legally reliable framework of policies and guidelines for FAIR research data management

Those 6 main objectives will be pursued through actions in the following 6 Task Areas (TA):

TA1 Management and Coordination will ensure efficient financial and organisational management of the award.

TA2 Smart Laboratory: TA2 focuses on the implementation and adaptation of existing and development of so far missing IT-components embedded in a flexible work environment, necessary to capture data early in the life cycle and to further manage, analyse and store associated information. TA2 enables a digital change in chemistry by supporting scientists with digital infrastructure of tools, services and repositories interoperable within the NFDI infrastructure.

TA3 Repositories: Archiving all relevant research data at each stage of the data lifecycle is a central aspect of the NFDI as a whole. This includes raw data in diverse formats as well as curated datasets. TA3 will adapt major chemistry repositories and databases to standards and interfaces and facilitate searching, commenting and exchanging reusable research data across distributed data sources.

TA4 Metadata and Data Standards: TA4 creates and maintains the specification and documentation of standards required for archival and exchange of data and metadata on molecule characterisation and reactions, together with reference implementations and data validation. Ontologies are used where possible, and missing terminological artifacts will be added.

TA5 Community Involvement und Training: TA5 arbitrates between community and infrastructure units: the community's requirements are collected, analysed and canalised to suited infrastructures. Equally, dissemination and training on all levels, starting in early undergraduate studies, is organised and training material is developed. TA5 also fosters the awareness of the community for RDM and offers incentives of innovations.

TA6 Synergies: TA6 coordinates the activities of NFDI4Chem with the other NFDI consortia. TA6 Synergies is furthermore responsible for managing the cross-cutting topics, including cross-domain metadata standards, semantic data annotation for cross-domain mapping of ontologies, provision of terminology services as well as legal aspects of FAIR research data management. Standards will be developed in close cooperation with international bodies such as RDA and IUPAC.

Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfil the planned consortium's objectives

NFDI4Chem aims to support the workflow of data from its acquisition in the lab - at the workbench, by analytical instrumentation or its generation by calculation and simulation. The data is captured, managed and analysed via virtual research environments as Laboratory Information Management Systems (LIMS) systems or Electronic Laboratory Notebooks (ELNs) and can be collected, shared and disclosed by repositories and curated databases. At all levels of this workflow, we identified existing infrastructures, tools, services, and in addition digital research environments which will be integrated with NFDI4Chem. The overall concept of the NFDI4Chem includes infrastructures, tools, services and environments (summarized as instruments) that (a) are operating on a national level in the focused domains of the NFDI4Chem and (b) are of importance for the NFDI4Chem while being developed and installed by other NFDI consortia. In addition, instruments that (c) are developed or maintained by international players or (d) have a generic function of interest for the NFDI4Chem were collected. Depending on the type of classification (a-d), the components are to be incorporated to the NFDI4Chem via different measures: Instruments that were identified to be incorporated to the NFDI4Chem are: the NMRShiftDB (NMR data), Chemotion repository (molecule characterisation and chemical reaction data), MassBank (MS data), SupraBank (intermolecular interactions), STRENDA-DB (enzymology data), RADAR (generic data repository), the CSD (Organic Crystal Structures), and ICSD (Inorganic Crystal Structure). The diverse challenges with respect to chemical structures, their registration/documentation, identification, visualization and analysis need special instruments that are absolutely essential to build an efficient infrastructure that meets the FAIR data principles and establishes solid curation and quality assurance measures. The NFDI4Chem will therefore include also traditional software libraries such as Chemistry Development Kit (CDK), RDKit, and OpenBabel, PubChem pug but will also constantly add established and novel IT-components such as Ketcher editor, JSMol, KNIME, ChemScanner, NMRquickCheck, NMRde.org (to be identified on in international level e.g. via the J. of Cheminformatics). The instruments will be embedded or connected to ELNs in order to develop interoperable virtual research environments that can be adapted in a flexible manner to the needs of the different chemistry sub-domains. A suitable Open Source ELN is the Chemotion-ELN but also other ELNs used by the community (e.g. Open Inventory) will be included to the NFDI4Chem strategy to achieve a network for the interoperable use of data. We will further build on multi-disciplinary services like DataCite (via the TIB), ORCID, and Re3Data. Commercial software products beneficial for a highly functional infrastructure will be connected if possible.

Interfaces to other proposed NFDI consortia: brief description of existing agreements for collaboration and/or plans for future collaboration

NFDI4Chem is open for collaborations with any chemistry related consortium or consortia with similar cross-cutting topics. In the foundation phase of NFDI4Chem first contacts with FairMat, NFDI4Ing, NFDI4Cat, DAPHNE, PHAN-Pan, and NFDI4Phys have been established and deepened during joint workshops covering topics like interdisciplinary (meta)data standards, cross-domain search, and access of repositories. Intensive consultations with neighbouring consortia from the material and engineering sciences, in particular with NFDI4Cat, showed that a community-tailored approach is best implemented through agreements on shared tasks with these consortia in the areas of standards and cross-cutting topics. With NFDI4Ing, we have exchanged our experiences on digitalisation of workflows for scientific data in chemistry and material science. Further exchange of the development of (meta)data standards and open formats is planned. The NFDI conference offered the opportunity to identify more consortia with mutual visions like NFDI4Healh, DeBioData or NFDI4NanoSafety. In life sciences molecule characterisation data like physicochemical, target engagement, bioactivity, pharmacokinetic, toxicology or safety and regulatory data play have been identified as linking elements. With NFDI4BioDiversity we will collaborate on integrated data access across the consortia, and development of data management tools for smart Lab environments, here metabolomics data is of particular interest for the biodiversity community. Together with NFDI4Health and NFDI4Microbiota we will discuss (meta)data standardisation and cross-sectional mapping for chemical compound characterisation data in context such as medication, dietary factors or metabolome data. Synergies between NFDI4Chem and NFDI4BIMP have been identified for spectral and spectrometric imaging data along with ‘pure’ image data like atomic force microscopic (AFM) imaging data and will be further investigated. With consortia from the life sciences, namely DataPlant, we share a common interest in developing training material for data literacy with a special focus on molecule-specific aspects.

In the discussions with all consortia mentioned above, the consensus for collaborative measures in dealing with molecule data became apparent. NFDI4Chem aims to coordinate these efforts with the NFDI.

4 Cross-cutting topics

General principles of FAIR data management, international networking and awareness-raising: Key personnell of NFDI4Chem is active in a number of international efforts, such as goFAIR, RDA interest groups, ELIXIR implementation networks, the European Open Science Cloud (EOSC) and more, which promote FAIR data in both the chemical as well as biomedical domain. We will aim to harmonize those existing efforts with FAIR data aspects across the whole of NFDI and engage in international networking with generic and specialized bodies promoting RDM and standards. As leaders and participants in collaborative research and excellence clusters in Germany, we will help to promote and implement the principles of FAIR data management in our local community, gather requirements and promote the adoption of the NFDI.

Repository technology and customization toward individual domains: Repository technology will be at the heart of virtually any NFDI consortium's implementation plan. To foster the interoperability of a potentially diverse portfolio of repository technologies, NFDI4Chem wants to promote standardization of interfaces and technological platforms across the NFDI which can be customized to individual research domains and application scenarios.

Mechanisms and instruments for agreeing on international standards: Research data can only be re-used when annotated with sufficient meta-data adhering to community agreed standards. New standards required for the NFDI cannot be negotiated at a national level but require extensive and long-term international consultations. The NFDI4Chem leadership has been engaged in such effort for the past 10 years and want to contribute to agreeing on common best practises for international development of standards within the NFDI.

Ontologies, terminology services: Once agreed, controlled vocabularies and ontologies will ideally be managed through common lookup services re-used across the whole NFDI.

Machine-readable data, data validation: Especially for cross-domain applications data needs to be unambiguously semantically annotated, both for humans and machines. Using discipline-specific terminology we will describe research data in machine-readable form and with less ambiguity research data semantics like properties, methods, units.

Efficient and harmonized materials and measures for outreach and training across NFDI: 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 throughout the NFDI. 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. NFDI4Chem would like to promote concerted efforts with the NFDI towards those goals.

Legal aspects of research data management, data sharing: NFDI4Chem participants have expertise to address legal aspects of RDM and provide support for the NFDI community on e.g. legal questions 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 those fundamental issues.

Unified and interoperable governance models across NFDI: NFDI4Chem leadership and participants have extensive experience in building international research data infrastructures in the biomolecular and chemical domain and beyond and will happily share this knowledge during discussion across NFDI domains.