Research Ideas and Outcomes 6: e55852
doi: 10.3897/rio.6.e55852

Grant Proposal

# NFDI4Chem - Towards a National Research Data Infrastructure for Chemistry in Germany

Christoph Steinbeck[‡], Oliver Koepler[§], Felix Bach[|], Sonja Herres-Pawlis[¶], Nicole Jung[|], Johannes C. Liermann[#], Steffen Neumann[¤], Matthias Razum[«], Carsten Baldauf[»], Frank Biedermann[|], Thomas W. Bocklitz[ˆ], Franziska Boehm[«], Frank Broda[¤], Paul Czodrowski[ˇ], Thomas Engel[¦], Martin G. Hicks[ʔ], Stefan M. Kast[ˇ], Carsten Kettner[ʔ], Wolfram Koch[ʕ], Giacomo Lanza[₵], Andreas Link[ℓ], Ricardo A. Mata[ϑ], Wolfgang E. Nagel[P], Andrea Porzel[¤], Nils Schlörer[Ⱥ], Tobias Schulze[ⱬ], Hans-Georg Weinig[ʕ], Wolfgang Wenzel[|], Ludger A. Wessjohann[¤], Stefan Wulle[Ғ]

‡ Friedrich-Schiller-University, Jena, Germany
§ TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
| Karlsruhe Institute of Technology, Karlsruhe, Germany
¶ RWTH Aachen University, Aachen, Germany
# Johannes Gutenberg University Mainz, Mainz, Germany
¤ Leibniz Institute of Plant Biochemistry, Halle, Germany
« FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Karlsruhe, Germany
» Fritz-Haber-Institut der MPG, Berlin, Germany
ˆ Leibniz Institute of Photonic Technology, Jena, Germany
ˇ TU Dortmund, Dortmund, Germany
¦ Ludwig-Maximilians-Universität München, Munich, Germany
ʔ Beilstein-Institut, Frankfurt am Main, Germany
ʕ Gesellschaft Deutscher Chemiker e.V., Frankfurt am Main, Germany
₵ Physikalisch-Technische Bundesanstalt, Braunschweig, Germany
ℓ Deutsche Pharmazeutische Gesellschaft, Frankfurt am Main, Germany
ϑ Universität Göttingen, Göttingen, Germany
P TU Dresden, Dresden, Germany
Ⱥ University of Cologne, Cologne, Germany
ⱬ Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany
Ғ Universitätsbibliothek der TU Braunschweig, Braunschweig, Germany

Corresponding author: Christoph Steinbeck (christoph.steinbeck@uni-jena.de), Oliver Koepler (oliver.koepler@tib.eu)

Reviewable v1

Received: 25 Jun 2020 | Published: 26 Jun 2020

Citation: Steinbeck C, Koepler O, Bach F, Herres-Pawlis S, Jung N, Liermann JC, Neumann S, Razum M, Baldauf C, Biedermann F, Bocklitz TW, Boehm F, Broda F, Czodrowski P, Engel T, Hicks MG, Kast SM, Kettner C, Koch W, Lanza G, Link A, Mata RA, Nagel WE, Porzel A, Schlörer N, Schulze T, Weinig H-G, Wenzel W, Wessjohann LA, Wulle S (2020) NFDI4Chem - Towards a National Research Data Infrastructure for Chemistry in Germany. Research Ideas and Outcomes 6: e55852. https://doi.org/10.3897/rio.6.e55852

## Abstract

The vision of NFDI4Chem is the digitalisation of all key steps in chemical research to support 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 chemistry 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 and easy to use 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:

Key Objective 1: Establish a virtual environment of federated repositories for storing, disclosing, searching and re-using research data across distributed data sources. Connect existing data repositories and, based on a requirements analysis, establish domain-specific research data repositories for the national research community, and link them to international repositories.

Key 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. Identify and recommend open data standards in key areas of chemistry, in order to support the FAIR principles for research data. Finally, develop standards, if there is a lack.

Key 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 (RDM) at all levels of academia, beginning in undergraduate studies curricula.

Key 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 RDM and data science into curricula. Offer a wide range of training opportunities for researchers.

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

Key Objective 6: Provide a legally reliable framework of policies and guidelines for FAIR and open RDM.

## Keywords

Research Data Management, Databases, Chemistry, NFDI, NFDI4Chem

## Consortium

### Research domains or research methods addressed by the consortium, objectives

Chemistry is a core natural science influencing and supporting many other research areas such as medicine and health, biology, materials science, engineering, or energy. The long-term preservation and re-use of research data from chemistry therefore also fertilises other disciplines. Research Data Management (RDM) in chemistry is currently not organized systematically and separated solutions of individual institutions lead to a low visibility, accessibility and usability of research results. The lack of (interdisciplinary) use of research data not only causes high costs for society, but also delays national and international developments and thus innovation in central research areas. The added value that emanates from the preservation and study of scientific data in chemistry is particularly high, since the significance of the data is often immortal and older data can also be used for current investigations. In most cases it is even absolutely necessary to access older data, because experimental data or complex simulation data in particular can only be generated with high costs and great effort. A loss of the previously acquired data can be an irretrievable loss of knowledge. **The vision of NFDI4Chem is the provision of a sustainable RDM infrastructure through the application of digitalisation principles to all key steps of research in chemistry.** NFDI4Chem will support scientists in their efforts to collect, store, process, analyse, disclose and re-use research data in Chemistry. Measures to promote Open Science and 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 funding 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:

**Key Objective 1**: Establish a **virtual environment of federated repositories** for storing, disclosing, searching and re-using 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.

**Key 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.

**Key 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 RDM at all levels of academia, beginning in undergraduate studies curricula.

**Key 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 RDM and data science into curricula. Offer a wide range of training opportunities for researchers.

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

**Key Objective 6**: Provide a **legally reliable** framework of **policies and guidelines** for FAIR RDM.

## Composition of the consortium and its embedding in the community of interest

NFDI4Chem 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. This broad community inclusion was achieved through working with our peers in a series of widely advertised workshops, a nation-wide user survey which will be discussed in detail below, as well as through working with our learned societies in Chemistry over the course of two years prior to submitting this proposal. The consortium consists of individuals and groups who shaped open and FAIR RDM in Germany in the past.

NFDI4Chem is supported by the German Chemical Society (GDCh), German Bunsen Society for Physical Chemistry (DBG) and German Pharmaceutical Society (DPhG) - representing approximately 40,000 members - to reach out to the chemistry community as a whole. All learned societies will continue to serve as participants and members of our advisory boards during the implementation phase to ensure a continued deep embedding in our community. Our assumptions, considerations and knowledge of user needs are based on an intensive exchange with the community. For this purpose, we have evaluated past and current surveys of which our latest is summarized in more detail as follows (Technische Informationsbibliothek et al. 2010, Tristram 2019). A first survey (Hausen 2019) performed by the FDM Team of the RWTH university library and IT Center in May/

June 2019 with 427 answers from chemists at North Rhine-Westphalia universities already revealed that 1 out of 4 professors knew the FAIR principles whereas only 1 out of 10 PhD students did so.

Our latest international survey was performed by the NFDI4Chem survey team, of which data from July through September 2019 were analysed. The survey was announced via the newsletter of the German Chemical Society (GDCh), emails to universities, Twitter, and a guest editorial in Angewandte Chemie (Herres-Pawlis et al. 2019). More than 600 responses, thereof 530 from Germany, have been collected from researchers at different levels of their career and working in different subdisciplines of chemistry. The survey covers researchers that are familiar with and apply the RDM rules of their institution but also those who are not familiar with RDM policies (Fig. 1,Fig. 2).

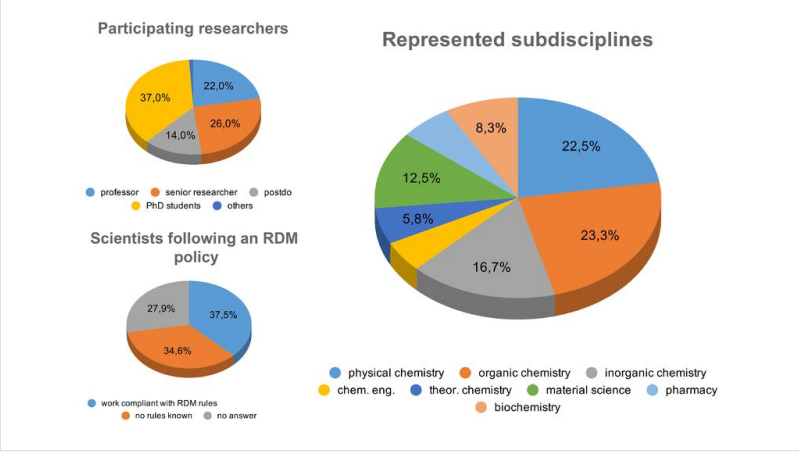


Figure 1.

Key data from the community survey in 2019 with N=530 responses.

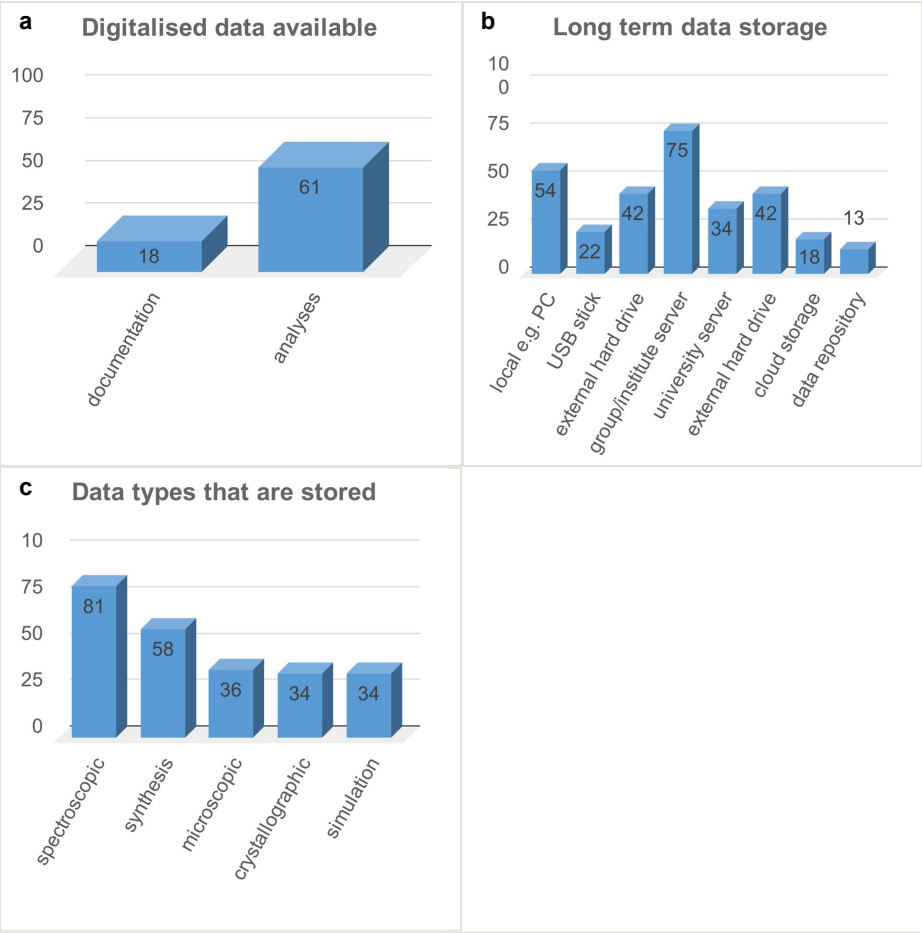


Figure 2.

Community survey: digitalisation status, solutions to store data and data types.

Most important insights with respect to the current digitalisation status, the data storage solutions, and responses on types of data are summarised in Fig. 2. Digital work takes place with respect to the analysis of data but the documentation of research processes is mostly non-digital. Some who claim to use electronic laboratory notebooks (ELN) consider MS Word to be an ELN. **Long-term archival** is important for all scientific projects: 82% claim to successfully follow the DFG rules of good scientific practice but data is very often stored on local computers and group or institute servers, while repositories are only rarely used to store data. Most of the data have their origin in spectroscopy and synthesis, but microscopic, crystallographic, and simulation data are also recorded. The participants also state to **re-use data** from colleagues of the same or other institute/working group and a small group (15%) also re-uses data from a repository. **Sharing** of data seems to be a common practice (sharing practices are summarized in Fig. 3) where email is dominating both internal(!) and external data sharing. Remarkably, only 10% indicate that they do not

share data outside their group. Interestingly, metadata are stated to be useful by the majority of the participants, either for group internal or external re-use of the data or in combination with a publication, but only 42% of the participants describe their collected data with **metadata.** The survey shows that the importance of data publication, especially in data repositories, is not yet anchored in the consciousness of scientists. While 59% of the scientists have published research data in form of supplemental information, only 16% have used repositories for a text-based and 8% for a **data publication**.

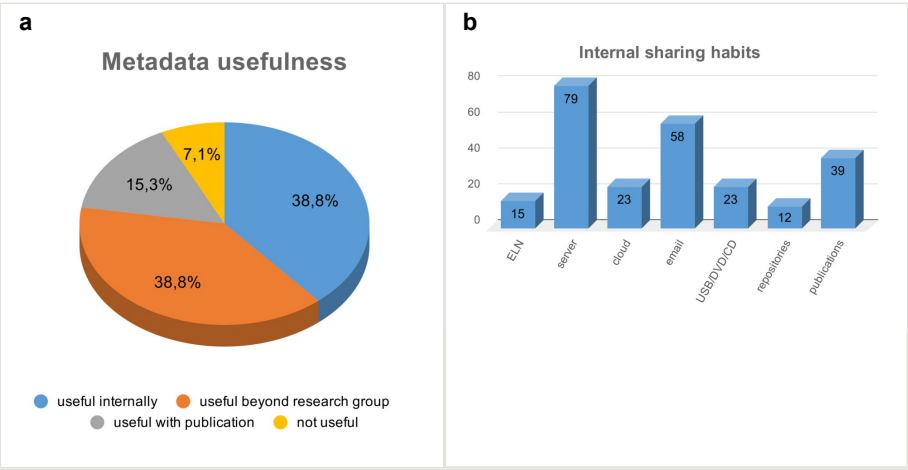


Figure 3.

Community survey: usefulness of metadata and habits of data sharing.

Some desired improvements of the current RDM system were mentioned:

1. general guidelines for data organisation and accepted standards for data formats and metadata annotation,
2. solutions to the time-consuming manual extraction of metadata,
3. need for software which simplifies the process of recording, analysing and saving data. Many participants indicated that they need an ELN capable of handling spectroscopic data. Generally, the participants wished that
4. data should be searchable more easily, e.g. in a national or Europe-wide repository on spectroscopic parameters of molecules including raw data.

The introduction of data stewards was considered desirable due to various problems ranging from non-functioning infrastructure to the lack of compliance with research data management requirements within teams.

**Based on this analysis, we have designed the work programme of NFDI4Chem to achieve a breakthrough in RDM in chemistry.** Our six task areas (TA) each address one or more of the issues identified above. We base our work on integrating the existing lighthouses of RDM in our research domain, fill gaps both in the repository landscape and the underlying standards, develop and disseminate powerful tools to enable early digital

data and metadata capture in the lab, and develop a strong training programme for chemists to understand and adopt the concepts developed in NFDI4Chem. Also, we dedicate a full task area to leveraging the synergies between the NFDI4Chem and the NFDI as a whole:

**TA1 Management and Coordination** will ensure the lean and efficient financial and organisational management of the project.

**TA2 Smart Laboratory (Smart Lab)** 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** enables the reliable storage, dissemination and archival of all relevant research data at each stage of the data lifecycle. This includes raw data in diverse formats as well as curated datasets. TA3 will adapt major existing chemistry repositories and databases to standards and interfaces, thus fostering interoperability and FAIRness as well as facilitating storing, disclosing, searching and re-using research data across distributed data sources.

**TA4 Metadata, Data Standards and Publication Standards** creates and maintains the specification and documentation of standards required for archival, publication 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 are added.

**TA5 Community Involvement and Training** interfaces between community and infrastructure units: the community's requirements are collected, analysed and channeled. Equally, dissemination and training on all levels are organised, starting in early undergraduate studies, and discipline-specific training material is developed. TA5 also fosters the awareness of the community for RDM and offers **incentives for innovations**.

**TA6 Synergies** coordinates the activities of NFDI4Chem with the other NFDI consortia. TA6 is responsible for the coordination of 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 RDM. Harmonisation will be sought by working with international bodies such as the Research Data Alliance (RDA) and the International Union of Pure and Applied Chemistry (IUPAC). TA6 develops an overarching search service and terminology service, that both will be linked to the NFDI.

To fulfil this ambitious work programme, we have assembled a hand-tailored consortium to perform the proposed work:

The resulting NFDI4Chem consortium consists of those institutions, groups, and individuals, who have been previously recognised for developing and supporting electronic

infrastructure for chemistry in Germany and beyond. The consortium is led by **Christoph Steinbeck**, Professor for Cheminformatics at the University of Jena, with 20 years of experience in building open chemistry databases and cheminformatics infrastructure components. For nine years, Steinbeck led the chemistry area at the European Bioinformatics Institute (EMBL-EBI), which is tasked by its 27 member states to perform the open access RDM for the biosciences in Europe. Apart from building production quality open chemistry databases such as ChEBI and MetaboLights, Steinbeck led European e-infrastructure consortia to develop standards in biochemistry (COSMOS, FP7 EC312941) as well as for high-performance cloud computing environments in chemistry-based molecular biology (PhenoMeNal, H2020 EC654241). In this context, Steinbeck founded ELIXIR and goFAIR communities in his areas of working and is active in cheminformatics related working groups of the IUPAC. He leads the INF project of ChemBioSys CRC in Jena and is a PI in the data synopsis work area of the Microverse cluster of excellence in Jena.

**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 Technische Informationsbibliothek. He has nearly 15 years experience in building e-infrastructures and digital libraries for chemistry and research data, fostering open data and metadata standards. His work involved projects like the virtual library of chemistry - personalised information services for chemical research and industry (DFG), the VisInfo project for interactive, graphical retrieval processes for research data (Pakt für Forschung und Innovation, SAW-Programm), and development of digital preservation processes and document delivery service APIs within the Specialised Information Services for Pharmacy - FID Pharmazie (DFG). Oliver Koepler coordinated the TIB IT project to merge the TIB discovery service and the library catalogue, introducing a user-centred design approach. Current projects include the development of a research data and knowledge management system for engineering in the INF-project of the CRC 1153 - Tailored forming and the IUPAC Smiles+ project.

**Nicole Jung** works on method development in synthetic organic chemistry and is heading the projects of the Institute of Organic Chemistry at the KIT in chemoinformatics and laboratory digitalisation. These projects include the development of infrastructure in the form of an electronic lab notebook (Chemotion-ELN) and the establishment of the chemotion-repository (chemotion projects - DFG grant BR1750/34-1). The Chemotion-ELN is developed to support the digital work in chemistry laboratories and provides extensive functionalities and tools for the scientific work of chemists, in particular organic chemists. The chemotion repository is an Open Access repository designed to store, manage, and publish research data that is assigned to molecules, their properties and identification as well as reactions and experimental investigations. These infrastructure activities are supported by the programming of software tools to support digitalisation, and the initiation of data curation and data collection efforts in synthetic chemistry. N. Jung strongly supports the ideas of FAIR and Open Access data and was awarded as SPARC Europe Open Data champion.

**Felix Bach** works on generic RDM and big data analysis and is heading the RDM group at the Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT). He is also part of the joint management of the service team RDM@KIT. Felix Bach concentrates on big data management and analysis - especially in the analysis of large time series and has developed a generic concept and software framework for structural analysis of huge multivariate time series data. He has worked and is still involved in several large scale data management projects including bwDataArchive in which the first generic research data archive infrastructure for long-term bit preservation of hundreds of petabytes of data was built up for the universities of the state of Baden-Württemberg. Additional projects were focused on enabling better data flows between the heterogeneous RDM systems, to make archiving and publishing of data sets easier for scientists of all fields, integrate existing systems as part of the research data cycle and create and install RDM policies and a RDM support team involving all organisational units that play a role in RDM (library, computing centre, legal department, research funding etc.). Felix Bach is engaged in several projects and initiatives that foster open source research software and their sustainability e.g. by creating and installing software development policies for universities within the Helmholtz Community. He is also engaged in the RDA, RDA-DE and regional RDM forums to make research reproducible by advancing FAIR and Open Access research data.

**Matthias Razum** is heading the e-research department at FIZ Karlsruhe – Leibniz Institute for Information Infrastructure. Since 2004, he has been intensively involved in the design and implementation of solutions for RDM, virtual research environments and digital long-term archiving. He has conducted research on these topics in several large BMBF and EU projects (e.g. eSciDoc, SCAPE) and built infrastructures together with partners from science and humanities. Since 2017, he has been responsible for the operation of the generic research data repository RADAR, which is now used by eight universities and research institutions in Germany. Matthias Razum has participated in several working groups on RDM on a national and state level. He is a Steering Committee Member of the Preservation and Archival Special Interest Group and Advisory Board Member of the Generic Research Data Infrastructure (GeRDI). Between 2010 and 2017, he served as a Steering Committee Member of the International Conference on Open Repositories.

**Steffen Neumann** is head of Research Group Bioinformatics and Scientific Data at the IPB, a non-university research centre and a member of the Leibniz Society, dedicated to chemical, and biochemical plant research. He is an expert in the area of statistical mass spectrometry data analysis and metabolite identification. In this context he is pushing Open Data and Open Standards, leading to community-wide e-Infrastructures, and to exploit these for functional annotation through advanced metabolomics analyses. Dr. Neumann was WP lead for "*Standards Development*" in the FP7 COordination of Standards in MetabOlomicS (COSMOS) project and WP lead for "*Tools, workflows, audit and data management*" in the H2020 project PhenoMeNal on computational metabolomics, is PI in the BMBF funded MASH which is part of the german Network of Bioinformatics infrastructures (de.NBI). He is a member of the scientific advisory boards of french and australian metabolomics infrastructures, and associate editor for BMC *Bioinformatics*, MDPI *Metabolites* and Nature *Scientific Data*.

**Sonja Herres-Pawlis** heads the Chair of Bioinorganic Chemistry at RWTH Aachen and combines bioinorganic chemistry (mainly Tyrosinase models and entatic state models) with sustainable polymerisation catalysis. In both fields, she and her team combine advanced synthetic methods, extensive kinetics with a multitude of spectroscopic methods and density functional theory to derive a comprehensive mechanistic understanding. She was one of the founders of the Molecular Simulation Grid (MoSGrid), a BMBF-funded initiative between 2009 and 2012, which focused on the design and development of a user-friendly electronic environment for theoretical workflows in chemistry with implemented search engines for molecules (Krüger et al. 2014, Gesing et al. 2016, Grunzke et al. 2013, Gesing et al. 2012). In MoSGrid, she was caring for the chemical community, collecting the requirements, writing regular newsletters to document the latest developments in the field - far beyond the official BMBF funding period. In the interdisciplinary EU infrastructure project ERflow (2012-2014), she studied multi-level quantum-chemical workflows and their interoperability in different workflow systems. Since 2015, she works together with R. Grunzke and R. Müller-Pfefferkorn on Metadata Management for Applied Sciences (MASI) in the framework of a DFG infrastructure project. Here, a metadata assignment tool for manual metadata definition per spectroscopic method has been established (Arshad et al. 2016, Grunzke et al. 2019).

**Johannes Liermann** is the head of the analytics core facility at the Institute of Organic Chemistry (JGU) and is specialized in NMR spectroscopy of small molecules. He has been cooperating with Nils Schlörer (UzK) in the context of the public domain NMR database nmrshiftdb2 which was originally conceived by Christoph Steinbeck. In cooperation with Nils Schlörer, J. Liermann is PI in two DFG projects for establishing an electronic assignment workflow to improve the quality of published NMR data (IDNMR project, LI2858/1) which strongly involves research data management aspects and offers important elements for an RDM infrastructure for molecular chemistry such as NMR data repositories. In addition, J. Liermann is elected board member of the Magnetic Resonance Discussion Group (FGMR) in the GDCh.

## The consortium within the NFDI

### Thematic Embedding

In preparation for this proposal, NFDI4Chem has extensively cooperated via joint workshops with FairMat, NFDI4Ing, NFDI4Cat, DAPHNE, PAHN-Pan, NFDI4Phys, NFDI4BioDiversity, NFDI4Agri, NFDI4Health, NFDI4Microbiota, and further neighbouring consortia. The cooperation has covered topics like interdisciplinary (meta)data standards, cross-domain search, legal aspects, and access to repositories. Networking with other consortia has been facilitated by the fact that many members of NFDI4Chem are also active in other consortia. These are: NFDI4Biodiversity, NFDI4Medicine, PAHN-PaN, NFDI4Culture, MaRDI, NFDI4MobilTech, NFDI4Agri, NFDI4MSE, FAIRMat, NFDI4Ing, NFDI4Phys, NFDI4Earth, GHGA, DAPHNE.

To emphasise the importance of cross-cutting topics in the NFDI as a whole, 21 NFDI consortia signed the Berlin Declaration (Glöckner 2019), co-authored by the NFDI4Chem

leadership, identifying central topics of general interest for all consortia. We will participate in preliminary workshops to further foster cooperation in preparation of interconsortial working groups. Particularly relevant are our interactions with thematically related consortia where we want to contribute our expertise. We have been in close coordination with the neighbouring consortia from the material and engineering sciences from the very beginning. Discussions with NFDI4Cat showed, that a community-tailored approach is best implemented through agreements on shared tasks in the areas of standards and cross-cutting topics like ontologies, metadata formats and the cross-linking of data repositories. While NFDI4Chem focuses on molecules and their characterisation data, NFDI4Cat covers the areas of technical chemistry and chemical engineering sciences. FAIRmat embraces condensed matter physics which includes soft matter and the (chemical) physics of solids and liquids, addressing a distinct community and research area. Nevertheless, there are partially overlapping areas where NFDI4Chem and FAIRmat will collaborate, for example in the definition and implementation of extended, new metadata standards for quantum chemistry in the NOMAD Repository. With NFDI4Ing, we have discussed options to model metadata and exchanged experiences on digitalisation of workflows for scientific data in chemistry and material science. In life sciences, molecule characterisation data like physicochemical, target engagement, bioactivity, pharmacokinetic, toxicology or safety and regulatory data 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 metadata standardisation and cross-sectional mapping for chemical compound characterisation data in contexts 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 DataPlant we share a common interest in developing training material for data literacy with a special focus on molecule-specific aspects. Together with MaRDI we have identified the potential of new research insights by providing suitable and easy-to-use interfaces to apply mathematical methods of MaRDI on chemical data retrieved from the NFDI4Chem repositories. 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.

We see the following topics as areas where NFDI4Chem would specifically invest effort to coordinate across consortia with the whole of the NFDI:

**General principles of FAIR data management, international networking and awareness-raising**: Key personnell of NFDI4Chem are active in a number of international efforts, such as GO FAIR, 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 customisation 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 standardisation of interfaces and technological platforms across the NFDI which can be customised to individual research domains and application scenarios.

**Catalogue of all services developed by the NFDI**: Following the model of the European Open Science Cloud (EOSC) and enabling easier integration into the same, we suggest the cross-cutting service catalogue enabling the keyword based discovery of services by users. Individual catalogues for exposure on individual consortium portals can then be generated on the fly from the central catalogue. The central catalogue will be designed to feed automatically, if desired, into the EOSC service catalogue.

**Mechanisms and instruments for agreeing on international standards**: Research data can only be re-used when annotated with sufficient metadata 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 efforts for the past 10 years and we 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 lookup terminology services used across the entire 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 terminologies we will describe research data in machine-readable form and adopt and develop research data semantics for properties, methods, units.

**Efficient and harmonised 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 and 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.

## International networking

The community-wide negotiation of standards in chemistry, covering both data and metadata standards, require global efforts and cannot be achieved on a national level. Here, NFDI4Chem leadership and participants are already been entertaining a wide range of activities and have extensive experience in driving global standardisation efforts in cooperation with international bodies for standardisation in chemistry and beyond. In independent standardisation efforts, we were instrumental in establishing Chemical Markup Language (CML) as one of the truly open and versatile data formats in chemistry (Murray-Rust et al. 2001, Kuhn et al. 2007). In order to promote the electronic deposition of NMR data to chemical structures, we have co-developed NMReData (Pupier et al. 2018), which is currently being adopted by *Magnetic Resonance* in Chemistry and other journals for this purpose. In truly international efforts, we contributed to the developments of XML formats for NMR (Schober et al. 2017) and MS (Martens et al. 2010) data. We were further founding members of the Blue Obelisk (Guha et al. 2006, O'Boyle et al. 2011), an international network of chemistry groups promoting Open Data, Open Standards and Open Source (ODOSOS).

NFDI4Chem members are furthermore active in institutionalized international efforts such as:

**The International Union of Pure and Applied Chemistry (IUPAC)** has traditionally played a significant role in creating and maintaining standards in chemistry. In chemical information, the most widely used **open** standards such as the International Chemical Identifier (InChI) and its variants as well as the spectroscopic standards JCAMP are maintained and developed by IUPAC divisions. NFDI4Chem key personnel are involved in the development of those standards. Speaker Christoph Steinbeck is a member of the InChI subcommittee of the IUPAC. Participant Thomas Engel is the delegate of the German Chemical Society (GDCh) to the IUPAC Division VIII committee. Oliver Koepler is engaged in the IUPAC SMILES+ specification project. Participant Patrick Théato is secretary of the subcommittee on polymer education (IUPAC Division IV) and member of the subcommittee on polymer terminology (IUPAC Division IV). All members will link and push the outcomes of the NFDI4Chem TA4 to an international level, for a broader discussion, and finally to influence world-wide recommendations of the organisations. As a former member of the CPEP committee on printed and electronic publications, Christoph Steinbeck contributed to the maintenance of the JCAMP standard. He still maintains the JCAMP reference implementation hosted on SourceForge.

**European Open Science Cloud (EOSC)**: NFDI4Chem aims to seamlessly integrate all services developed into the service catalogue of the European Open Science Cloud (EOSC). Applicants of this proposal are active contributors to the EOSC. The Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT) participates in the EOSC related projects EOSC-hub, EOSC-Secretariat, EOSC-synergy, EOSC-Pillar as well as in the AAI related work in the GEANT4-3 project. The Leibniz Institute of Plant Biochemistry was partner in the PhenoMeNal project, led by NFDI4Chem speaker Christoph Steinbeck, that build an infrastructure for data processing and analysis for medical metabolomics. All applicants will use their existing expertise to ensure interoperability of the NFDI4Chem services with the functionalities of the forthcoming EOSC. Within EOSC, the Molecular Open Science Enabled Cloud Services project (MOSEX) has expressed its interest to cooperate on all key objectives with NFDI4Chem (see LoS).

**Research Data Alliance (RDA)**: Five key members of NFDI4Chem are engaged in the Chemistry Research Data Interest Group (CRDIG) of the RDA (Chemistry Research Data IG 2015). Additionally, some members of the consortium are engaged in the Storage Service Definitions WG, Research Data Repository Interoperability WG und Long Tail of Research Data IG.

**GO FAIR**: NFDI4Chem Speaker Christoph Steinbeck and co-applicant Steffen Neumann were instrumental in instantiating the GO FAIR implementation network for metabolomics, one of the first implementation network in GO FAIR at all. Metabolomics as an interdisciplinary field has a strong analytical chemistry and cheminformatics component. Members of NFDI4Chem also have strong links to GO FAIR Chemistry Implementation Network (ChIN) (Chemistry - GO FAIR 2019), as the ChIN operates in tandem with the CRDIG of the RDA.

**ELIXIR** is the pan-European infrastructure for biological information (ELIXIR Consortium 2019). The purpose of ELIXIR is to support life science research and its translation to medicine and the environment, the bio-industries and society. The NFDI4Chem partners IPB and FSU are members of ELIXIR-DE. We will coordinate efforts with those parties in ELIXIR who are also handling chemical information, such as the Core Data Resources BRENDA (BRENDA Enzyme Database 2019), ChEBI (EBI Web Team 2019) and ChEMBL (ChEMBL Database 2019), and existing infrastructure efforts like ELIXIR-AAI (ELIXIR Consortium 2013).

## Organisational structure and viability

We have designed the organisational structure of NFDI4Chem to ensure that the work programme can be pursued in an efficient and agile manner, that the decision making process within NFDI4Chem is legally sound and transparent and that the community and our stakeholders are closely attached to our operation. In the following, we describe the various components of our **organisational structure** as shown in Fig. 4.

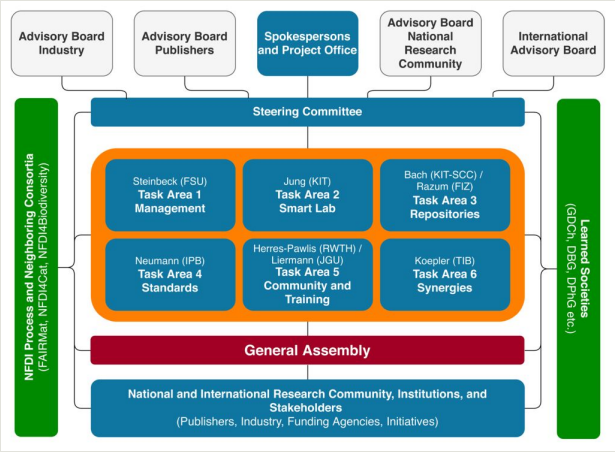


Figure 4.

NFDI4Chem governance structure.

**The General Assembly (GA)** is the central decision making body of the project. The GA consists of all co-applicants and participants (represented by one delegate from each partner), as well as 5 members of the community elected by the GA. The GA will decide on all issues of fundamental importance for the whole project. The GA will be held at the annual NFDI4Chem project meetings or online if urgent matters require that. These meetings will be used for strategic planning, presentation of scientific results, and discussion of major management issues. The GA decides when overall agreement is required in the matters of budget and consortium management.

The **Steering Committee (SC)** is the central body responsible for monitoring and evaluating project progress and supervising project objectives, and which takes the necessary decisions in scientific coordination and administration of the project. Based on the contributions from task areas, the SC prepares periodic reports and the final report. The SC consists of the two speakers, the project manager and the task area leads. The SC is headed by project speaker Christoph Steinbeck.

NFDI4Chem will establish a focused set of **Advisory Boards**, which are consulted regularly to ensure that the consortium is on track and develops and delivers services which are:

- aligned with the mission of the NFDI in general and
- address the needs of the chemistry community.

**Advisory Board *Industry*** comprises companies developing data producing equipment, data management systems and data analysis solutions. They will provide insights on data formats and software, and aim at including the NFDI4Chem recommendations and processes into their software developments.

**Advisory Board *Publishers*** will provide insights on research data associated with scientific publications. They will aim at including the NFDI4Chem recommendations and processes into their guidelines for authors, editors and reviewers. The manuscript submission systems should support compliance with the guidelines.

**Advisory Board *National Research Community*** represent researchers and organisations performing research in Germany. We aim to cover the subject areas reflected in the list of DFG review boards (DFG Fachkollegien) mentioned in the section General Information above.

***International* Advisory Board** complements the advisory boards described above, with a special focus on international organisations and individuals which are lighthouses of collaboration and alignment of efforts to collect, store, process, analyse, disclose and re-use research data.

**Decision Making**: The distributed nature of the NFDI4Chem project necessitates a decentralised administration of execution control for effective decision making. The operational level comprises all the project partners who are responsible for the execution of the strategic work plan detailed in the task areas (TA). The TA leaders will be responsible for keeping track of the measures with the listed deliverables. Any deviations will be brought to the attention of the project office (PO) and discussed latest at the next steering committee (SC) meeting. The TA leaders oversee the budget, technical aspects including quality checks and communications with the project office when required. This process is supported by the detailed project management plan maintained at the FSU. The main scientific controlling and decision making body in the project is the steering committee (SC). The SC is responsible for all decisions regarding project management, distribution, monitoring and re-organisation of specific tasks if necessary and for all cases which do not require the voting of the GA. The SC convenes by electronic communication on a regular basis and on-demand, as organized by the project office. The GA will be the highest decision making body in the project and will be consulted for strategic planning, major management topics and other fundamental issues. The GA will take in particular decisions if overall consensus is required in the matters of inclusion of a new partner, exclusion of an existing partner, major changes in budget or the project strategy and in other unforeseen situations that need discussion or decision making. Decisions of the GA that need voting require a simple majority of the project partners based on the principle “one partner – one vote”. To have a quorum, 75% of the partners have to be present at the GA physically or present through teleconference facilities at the time of decision making and voting. In a stalemate situation, the Project Coordinator will have the deciding vote.

## Operating model

Our operating model follows the unanimous agreement amongst consortia during the governance workshop in Bonn on August 30, 2019, where we were strongly advocating for an overarching NFDI e.V. which consortia can join as dependent legal entities. Until agreement has been reached about the legal model for the NFDI as a whole, we will operate NFDI4Chem under a normal academic consortium model. This will be based on a

consortium agreement (CA) which will clarify the operational dependencies between co-applicants and participants, the mode of operation as described above and in TA1, and allow for the transfer of funds from the applicant institution to the co-applicants and participants.

Special attention will be given to the financial compensation model for participants, where no exchange of goods and services in a commercial sense is currently foreseen. Since all NFDI consortia have to solve these issues, we already agreed to work together to design a CA and contracts that are compliant with the German law. As a strong advocate of the Berlin Declaration, NFDI4Chem looks forward to close collaboration with other NFDI consortia and the NFDI directorate on this topic.

## Research Data Management Strategy

### Research Data Management in Chemistry - a status quo

Chemistry consists of many subdisciplines with a large variety of methods and data. Chemical research dealing with molecules, their reactions and properties can be described by experimental procedures, observations, theoretical models and their computation, and by the resulting data. Data are further processed, analysed, evaluated, and in many cases assigned to a molecule as proof of a hypothesis. The large 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 a plethora of different **data types and formats**, most of them being proprietary. These data often consist of spectroscopy and spectrometry results (such as nuclear magnetic resonance (NMR), mass (MS), IR/Raman, UV). Other data come from elemental analysis (EA), X-ray, electron paramagnetic resonance (EPR), cyclic voltammetry (CV), gel permeation chromatography (GPC) and differential scanning calorimetry (DSC) measurements or thermogravimetric (TGA) or dynamic mechanical thermal (DMTA) analysis, only to mention a few. Proprietary data formats can consist of a single, binary file of unknown structure, to directories populated with several files, some in ASCII text or (rarely) XML. Efforts in the life-sciences towards open MS formats resulted in mzML (Martens et al. 2010), and a toolset converting from vendor formats to mzML via closed source Windows DLLs provided by the vendors. There is no standard raw data format in NMR spectroscopy that covers all data and meta-data aspects, therefore the proprietary Bruker format has become a quasi-standard next to the proprietary formats for processed spectra in NMR analysis software (e.g., ACD Spectrus, MestReNova).

While in theoretical chemistry or computational chemistry **research data workflows** are seamlessly digital, the provenance and context of data produced in experimental laboratory workflows are mostly documented in hand-written lab journals. In our survey only 18% of researchers declared to use an electronic lab journal, with some of them considering text processors such as MS Word to be ELNs. When investigating a research problem, multiple devices generate multiple data sets of various data formats, stored in different locations

within an institution. Keeping track of all these data trails is an enormous challenge for researchers. Efficient systems for the **curation** of user-provided descriptive and contextual data and connection to related device-captured experimental and analytical data are missing. Although many devices provide experimental data in a digital format, accessibility and reuse of data is hampered by missing standards for data exchange formats, as can be seen from the wide-spread use of proprietary data formats in device-generated experimental output. As a result most descriptive metadata about these research data are still non-digital at the beginning. This coexistence of analogue metadata and data next to digital research data causes considerable problems later on, when it comes to the publication of research results and the corresponding research data.

When a researcher reaches the point of preparing a **publication** or even **data publication** based on formerly generated data, the deficits from the previously described points become visible. For instrument data, insufficient ways of presentation are common in chemistry journals. So far, research data are included in very condensed form in the publications, e.g. in the experimental section, in tables or images. Often additional provenance and context information about the research data and representations of the data can be found in the supplements to an article publication. Published as PDF files or bitmap images, these supplements may contain more detailed tables, peak lists or images of spectra, none of which can be re-analysed in their presented form. These existing conventions are in stark contrast to RDM in accordance with the FAIR principles. As our survey described above revealed, so far only 16% of researchers published their research data in data repositories in addition to an article publication at least once. This is certainly also caused by the fact that data repositories for annotated raw data exist only for a few domains in chemistry. More often, there are curated databases in which the derived data from the raw data sets can be found, usually combined with an assignment to a molecule. Those are often curated by human domain experts from the primary literature in a painstaking process. In the end, deposition of data and metadata is currently a rather complex and time-consuming process resulting in low acceptance by researchers.

Nevertheless, data is often exchanged among scientists. This is mostly accomplished through direct contacts. Only 15% of researchers have reused original data files of data that was obtained in their research field. Our survey reveals that only 40% of researchers have defined workflows and curation standards. In most cases curation and final storage procedure is chosen by the individual researcher or by rules of the research group.

Examples of successful RDM exists only in few subdisciplines such as crystallography, where the CIF file format enabled the community for 30 years to share data and embedded this data sharing into the publication process (The Cambridge Structural Database (CSD) 2019).

**Databases and Data Repositories**: Chemistry has a long history in indexing chemical data and creating searchable data collections. The extraction and indexing of chemical data found in literature by Beilstein/Gmelin handbooks or by CAS, resulting in the digital databases SciFinder and Reaxys, are the most prominent examples. In addition to these outstanding commercial databases, there are a number of databases with a narrower focus

on specific data about molecules, reflecting the importance of molecule characterisation data. Rather than a full assessment of the global landscape of chemistry databases, we describe those relevant to the NFDI4Chem strategy described below. Widely used molecule-centred databases like **ChemSpider** (Dabb 2016) or **PubChem** (PubChem 2019) provide information about molecules and their properties aggregated from various resources. Both databases allow to search molecular structures in order to retrieve basic information about molecules like names, identifiers, or physicochemical properties. Entries may even link to actual datasets related to a molecule. **ChEMBL** (Gaulton et al. 2011) is a manually curated chemical database of bioactive molecules with drug-like properties. Further examples are the SDBS (Spectral Database for Organic Compounds (SDBS) 2019) and for mass spectrometry mzCloud (Anonymous 2013) and NIST database (Stein 1990).

However, besides being partly proprietary, these databases lack the availability of the original research data and comprehensive metadata about it, which is often not available from the literature. This level of granularity can be achieved by data repositories where the actual research data file can be deposited with their corresponding metadata providing provenance and context. The **Crystal Structure Database (CSD)** of the Cambridge Crystallographic Data Center (CCDC) is an example of such a curated data repository, and is probably the best known repository for research data in chemistry. The publication of crystal structures via the CSD using the CIF standard is accepted as a standard procedure and well embedded in the article publication process in chemistry. A comparable database is **ICSD** in the field of inorganic crystallography, which cooperates closely with CSD. These repositories may serve therefore as kind of a best practice model in terms of user attraction.

With a national focus, Germany hosts the spectroscopic databases **nmrshiftdb2** (Kuhn and Schlörer 2015) for NMR spectroscopic data, **MassBank EU** (Vinaixa et al. 2016) containing mass spectrometry data and **Suprabank** (Website Suprabank 2019) containing guest-host interactions.

A true data repository with a focus on primary research data in chemistry is the **Chemotion repository** (Anonymous 2018), which collects data on reactions as well as analytical data for molecules and reactions. Besides this repository spanning wide fields in chemistry, subdiscipline specific repositories like **NOMAD** (NOMAD Repository 2019) and **ioChem-BD** (computational materials science data) (ioChem BD 2019), and **Strenda DB** (functional enzymatic data) (Tipton et al. 2014) exist. Generic and multi-disciplinary data repositories like **RADAR** (Kraft et al. 2016) provide (certified) data archiving services which can assist subject-specific offerings in terms of sustainability and reliability. At the same time, they can serve as catch-all repositories.

**Data Flow into repositories**: A key insight from our analysis described in section 2 is that data handling is currently not yet fully digital and that manual steps are often necessary, especially at the beginning of the process of data generation in the lab. This slows down RDM workflows and hinders the publication of FAIR data in repositories. Solutions to this problem have been addressed by the chemical and pharmaceutical industry already in the

1980ies and software as well as cheminformatic processes were developed to foster digitalisation strategies. Those solutions, however, did not find their way into academic data management. Work environments such as Laboratory and Information Management Systems (**LIMS**) and **ELNs** in combination with chemistry specific software and identifiers demonstrate how successful digitalisation of work processes can be achieved. **Data transfer** from devices to a digital management area solved by LIMS allows for control of the data flow, data tracking, and data transfer (STARLIMS 2019, LIMS 2019, Dassault Systèmes BIOVIA 2019b, SampleManager LIMS Software 2019, LIMSWiki. 2019) . A LIMS mitigates the errors that come along with manual data handling, prevents the loss of data, and improves the work efficiency. Disadvantages of a LIMS for academic institutions such as the high initial software acquisition costs and the running costs of user licenses can be overcome by open source solutions (LIMS Software 2019, Open-LIMS 2019, Bika LIMS 2014). Data transfer models that offer LIMS function by integration of devices to a digital workflow were presented by members of the consortium for NMR data by the nmrshiftdb2 (Kuhn and Schlörer 2015) and a data transfer solution based on device dependent open source modules (Potthoff et al. 2014, Potthoff et al. 2019, Lütjohann et al. 2015, Lütjohann et al. 2014). LIMS can be combined with ELNs which is used to collect and manage written information, values, data files and images. Although several ELNs have been developed in the past (Scinote-Web 2019, Dassault Systèmes BIOVIA 2019a, Barillari et al. 2015, Labfolder - Electronic Lab Notebook (ELN) 2019, Adam and Lindstädt 2019), only a few are available for chemists in academia due to the requirement to process chemical structures (Coles et al. 2013, Notebook|PerkinElmer 2019, Rudolphi 2019, GGA Software Services 2019, Day et al. 2015, Willoughby et al. 2014, Rudolphi and Goossen 2011, Tremouilhac et al. 2017), and only four of them are available as open source software, enabling further adaptations to changing research requirements. Despite the manifold benefits of a LIMS or ELN, less than 18% of academic groups run commercial, free, or open source systems according to our survey. Reasons for missing management systems in particular in academia can be found in the often very diverse infrastructure, institution boundaries of the universities, installation problems and/or long-term maintenance responsibilities. Open source ELNs are currently still missing probably because of a lack of awareness of the importance of a systematic change of the work habits to enable a digital documentation of the work. The integration of ELNs requires not only the acquisition of new infrastructure but also the adaptation of existing procedures.

The availability of **software** to analyse and process data is an important factor to achieve re-usability of research data. Here, the heterogeneous situation in the field is reflected by many individual software packages and tools. Important components are open source tools, enabling the creation of FAIR data are chemoinformatic libraries such as RDKit, CDK and OpenBabel which are used for diverse RDM applications in chemistry. Specific software in chemistry is also needed for spectroscopic data (e.g. NMR, MS, IR and UV/Vis), playing a pivotal role with respect to the identification and verification of research results. The current practice is to analyse those data either manually from printed spectra or using stand-alone software. To enable FAIR RDM, the visualisation and processing of spectroscopic data, but also its annotation (assignment) to the research data provenance and context, is important. While web-based solutions to be used embedded to ELNs or

repositories, are necessary, only a few open source, web-based visualisation tools are available (Banfi and Patiny 2008, Zhang and Brüschweiler 2007, Vosegaard 2015, Xia et al. 2012, Mohamed et al. 2016, Beisken et al. 2015). All of the mentioned solutions miss functionalities for comprehensive FAIR data management in chemistry with respect to functions as processing of data, data analysis and data annotation.

The use of **standards** in chemistry is characterised by pragmatic mix of open standards and what one could call publicly known (meta)data formats, where open standards are characterised by an open, public and inclusive development process, whereas publicly known formats have been developed by companies, documented somewhere, can be re-used, but the public has no or little influence on the development and if often not notified of changes, in the case of the commonly used SMILES format. A full account of available standards is outside of the scope of this section. Standards for *structure representation*, for example, are InChI and SMILES which was reworked by the Blue Obelisk community (O'Boyle et al. 2011, Guha et al. 2006) with the intention to provide an open standard. Examples including chemistry tables are the MDL Molfiles, SDFiles, RDFiles, each consisting of the structure information and additional metadata. The formats available for the machine readable presentation for reactions are for example Rxnfiles and Chemical Markup Language (CML) (Murray-Rust and Rzepa 1999). A common problem of all structure representation formats is the partly missing coverage of polymer chemistry and inorganic chemistry.

*Experimental data* is represented by a plethora of data formats where open standards exist only for a fraction of the methods. Documentation for proprietary data formats are often not available. The most prominent example for the long term usage of an open data format is the Crystallographic Information File (CIF) (Hall et al. 1991), which enabled the crystallographic community to share and collect data in electronic form since the early 1990s. For spectroscopic data, the JCAMP format (Lampen et al. 1999, Lampen et al. 2016) was developed and later maintained under the auspices of the CPEP committee of the IUPAC, but not dynamically adapted to the fast moving innovation in all areas of spectroscopy. More recent developments move towards the use of XML for the representation of spectroscopic data (Kuhn et al. 2007, Schober et al. 2017, Martens et al. 2010). *Theoretical Data* are produced by an eclectic mix of open and closed software. Data formats produced by proprietary software may still be well understood, as in the case of the quantum chemistry software Gaussian, but are not open. Open software will often produce output in an open standard. Attempts to capture the information from simulations in open formats are manyfold, see for example Grunzke et al. (2013), Wakelin et al. (2005).

In conclusion, the consistent, **long-term** and safe **storage** in repositories, with a few exceptions like silico chemistry and crystallography, is still rather uncommon. Standardised and normalised data, and descriptive metadata play only a minor role in the local management and storage of data. Journals do not require authors to deposit primary research data in well annotated form in community accepted repositories. NFDI4Chem aims to fundamentally change this situation through a concerted effort combining technical and cultural change with outreach to publishers and community training.

## The strategy of the NFDI4Chem consortium

NFDI4Chem will support the scientists in all steps of the research data management data life cycle. The measures will address the seamless integration of identified key components into NFDI4Chem infrastructure to optimise and facilitate data workflow and capture of metadata. These efforts will be supplemented by the development of standards and policies, resulting in a substantial increase of data quality in the RDM process. Finally, comprehensive community outreach activities will ensure that the initiated digital change will be transformed into an increased and sustainable awareness of RDM in the minds of researchers, thus fostering a cultural change.

The workflows to be supported start with data acquisition in the lab with data captured at the workbench, by analytical instrumentation or data that arise by calculation or simulation. The data is captured, managed and analysed via virtual research environments and can be collected, shared and disclosed by repositories and curated databases. At all levels of this workflow, we selected existing infrastructures, software, and services which will be integrated with NFDI4Chem. Single missing key elements were identified and their development will complement the envisaged infrastructure (Fig. 5).

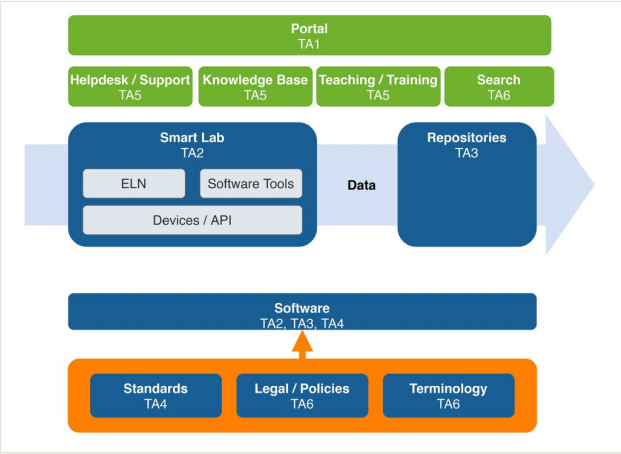


Figure 5.

Key components of the NFDI4Chem and their mapping to task areas.

NFDI4Chem will not only facilitate RDM but will create added values for scientists to accelerate the acceptance of concepts and services by the community. A very important aspect is to minimise additional efforts by data management and to allow a seamless data transfer throughout all components of the NFDI4Chem. By this, scientists will be faster to collect FAIR data, standards can be introduced and kept easier and errors due to manual rework will be avoided. Data and metadata should be transferred without manual interaction from the devices generating data to the virtual work environments to the discipline-specific or generic repositories. All components of the infrastructure that are

necessary for this fully digital workflow should be available to every scientist. Therefore, different measures such as the installation on site or a centralised operation, have to be realised depending on the type of component. For the operation of services and in particular with respect to the hosting of repositories, the NFDI4Chem strategy relies on the distribution of responsibilities to different partners within the consortium. Among those, three infrastructure centres (TIB, KIT-SCC, and FIZ) offer, due to their experience and expertise with chemistry research, the basic infrastructure support. Additional infrastructure centres (e.g., ITC Aachen, ZIH Dresden) further support the strategy directly by agreeing to host selected national repositories operated by participants (UzK, UFZ). NFDI4Chem establishes a strategy based on infrastructure components, software, and services that are operating on a national level but also international solutions will be included to the overall strategy by different measures. The concept of NFDI4Chem offers infrastructure for chemists covering all subdisciplines. To ensure a broad coverage, experts of the subdisciplines organic chemistry, inorganic chemistry, physical chemistry, polymer chemistry, biochemistry, pharmaceutical chemistry and computational chemistry are involved to build NFDI4Chem. In the first funding phase, the components of NFDI4Chem have a strong focus on those aspects of chemical research that have the greatest need for catching up in terms of digitisation. While the areas in chemistry, where theoretical research is predominating the daily work, are already well-positioned at least in terms of digital data availability, the support of experimentally driven subdisciplines includes fundamental digitalisation instruments. The proposed concept is a strongly demand- and user-driven approach, including all components and measures that were identified to be crucial for FAIR data management and those to be essential to allow RDM customised to the needs of the subdisciplines. The latter aspect is the most important prerequisite for the acceptance of the NFDI4Chem by the community. The requirements were identified in the past two years leading to this proposal by the members of the consortium, feedback of additional experts in the field and by the outcome of the requirement analysis from different surveys summarised above. Regular surveys covering the whole community in addition to a constant contact to the users of the NFDI4Chem infrastructure including road shows and workshops in universities are key instruments to constantly improve the service portfolio. The direct communication with scientists via hands-on trainings and live demos ensures the effective knowledge transfer and the awareness of new services and functions of the NFDI4Chem infrastructure. Both activities, consulting and training the community, foster the cultural change in chemistry.

## Metadata standards

The FAIR principles developed by the international FORCE11 initiative (Hagstrom 2014) demand findable, accessible, interoperable and reusable research data. These principles are an internationally accepted framework of minimum requirements of metadata for an effective research data management and the foundation of standards and processes to be developed in NFDI4Chem. Besides human- and machine-readable interoperable metadata, and standards as key elements of these principles, the integration of persistent identifiers is a prerequisite for the establishment of such standards and is addressed by several approaches for quality management concepts. CoreTrustSeal (CTS) – merged

from the former Data Seal of Approval (DSA) and the ICSU World Data System (WDS) Member Certification, the nestor Seal for Trustworthy Digital Archives, and the ISO 16361 Standard form a three-tier global framework of repository certification. All three concepts require persistent provision of metadata. The use of PIDs ensure the quality of metadata through standardisation, disambiguation of the described entities, as well as permanent findability, citation and cross-linking of scientific objects and authors. International standards for referencing have been established in recent years, including the open standards Digital Object Identifier (DOI) and Open Researcher and Contributor ID (ORCID). The international consortium DataCite registers DOI and additionally provides with the DataCite Metadata Schema (DataCite Schema 2019) a core set of domain-agnostic properties enabling an accurate and consistent identification of data for citation and retrieval purposes (Neumann and Brase 2014). To facilitate finding data, the provision of further discipline-specific metadata is mandatory. The RDA metadata directory lists chemistry specific metadata developed so far (Standards 2019). This standardisation process is far from complete and is being further promoted at international level by the collaborative work between IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS) Subcommittee on Cheminformatics Data Standards (SCDS), the Research Data Alliance (RDA) Chemistry Research Data Interest Group (CRDIG), and the GO FAIR Chemistry Implementation Network (ChIN). We will closely work with the respective stakeholders and communities which have initiated the existing metadata standards and their further development. Several members of the NFDI4Chem consortium are already embedded in the aforementioned initiatives. Rudimentary metadata about experimental methods can often be found embedded in the datafile of exchange data formats like JCAMP or JCAMP-DX. In biocatalytic research FAIR “Standards for Reporting Enzymology Data" (STRENDA) for data and metadata have been proposed. For NMR spectroscopic research data NMReDATA is a newly developed standard to report the NMR assignment and parameters of organic compounds (Pupier et al. 2018). It incorporates data, metadata, and chemical shift values, signal integrals, intensities, multiplicities, scalar coupling constants, lists of 2D correlations, relaxation times, and diffusion rates. The file format is an extension of the existing Structure Data Format and is easily readable by humans and computers. With respect to the digitisation of all steps in the research process, the capture of metadata of both generated research data and experimental or computational methods used, is a central challenge to fully describe provenance and context of the data (Willoughby et al. 2014). ISA (Sansone et al. 2012) is a successful metadata standard to capture essential metadata, including the experimental design, the applied protocols, association between samples, data files and the experimental factors for further statistical analysis. ISA consists of the ISA-Tab and ISA-Json formats, and a rich ecosystem of software for creation, management and consumption of ISA metadata.

To guide researchers in reporting metadata, NFDI4Chem will support international processes for the creation of Minimum Information standards in selected areas of chemistry. The accepted consensus for a specification of Minimum Information about a Chemical Investigation (MIChI) will be reached and discussed with the stakeholders (researchers, infrastructures and journals) and will guide the parameterisation of ISA-Tab descriptions of studies.

Extensions developed in the context of metabolomics capture the analytical information such as chemical shift and multiplicity in NMR-based experiments, and m/z, retention index, fragmentation and charge for mass spectrometry. Both SMILES or an InChI and database references are used for reporting metabolites.

NFDI4Chem will embrace these existing data standards and extend them based on requirements in the chemistry subdisciplines. Together with device vendors, publishers and our infrastructure developments these will become an integral part of the digital workflow in chemistry research.

## Implementation of the FAIR principles and data quality assurance

The FAIR principles promote data to be Findable, Accessible, Interoperable and Re-usable. FAIR data management is also a necessary condition for exchange with other disciplines - a key aspect of the NFDI as a whole. The dissemination and application of FAIR RDM services and repositories in chemistry is still at the beginning. Reasons are manifold; a lack of data and metadata standards, insufficient data quality, low data coverage, lack of efficient digital workflows, lack of search functionalities or scientific acknowledgement for data publications resulting in low acceptance of RDM services. 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 and minimising efforts that arise with RDM.

In NFDI4Chem, we will work towards a fully FAIR ecosystem of chemical RDM by following a now well-accepted strategy (Wilkinson et al. 2016).

To be **Findable**, data in all of our resources will have rich machine-readable (meta)data, linked with domain-specific and cross-domain vocabularies. Our ELN and device integration strategy outlined in TA2 will foster the aggregation of rich metadata early in the data generation process. Repositories within NFDI4Chem will register their datasets at DOI services like DataCite assigning globally unique and eternally persistent identifiers (DOIs) to enable indexing in a data catalog to be build for the Search Service in TA6.

To be **Accessible**, all data will be retrievable by their identifier using HTTP(S) as standardised communications protocol, which is open, free and universally implementable. Repositories and services will provide standardised APIs (see TA3). NFDI4Chem components, all being Open Access, will be using a NFDI-AAI service which allows authentication and authorisation procedures, where necessary (see TA6).

To be **Interoperable**, our (meta)data will use a formal, accessible, shared, and broadly applicable language for knowledge representation (see TA4), which uses vocabularies that follow FAIR principles and includes qualified references to other (meta)data (see TA6). Like the minimum information (MI) standards in biology, the chemistry community will develop MI metadata standards to semantically describe experiments and simulations, molecule

characterisations and others. At the same time, NFDI4Chem will promote open data formats.

To be **Re-usable**, our (meta)data will have accurate and relevant attributes, be released with a clear and accessible data usage license supported by legal policies and guidelines defined in TA6, their provenance, and meet domain-relevant community standards. This information is verified in the data curation process (see TA4). We will enable the reuse of data to support domain cross-linking, big-data analysis and future artificial intelligence (AI) methods.

Agreeing on data and metadata standards in a particular research domain is an international effort where NFDI4Chem input will advance the field. Thus, the effort will be pursued through collaboration with scientists and standardisation bodies such as the IUPAC, the Research Data Alliance (RDA) or the GO FAIR initiative. The strong links from NFDI4Chem to those international organisations are outlined in section 2.4 above.

Quality assurance of data in NFDI4Chem is supported by measures in TA2, TA3 and TA4. The definition of MI standards allows for automated checks at all steps of the data life cycle.

At the time of data curation, automatic plausibility checks and data validations can support the reviewers in the (peer) review process. Data in repositories and databases will be curated by a mix of automatic and manual quality checks. Last but not least, the re-use of data enables a quality check by the community, which will be the ultimate corrective measure for data quality in NFDI4Chem repositories. Details of these measures will be discussed in the individual task areas that address them.

### Services provided by the consortium

NFDI4Chem will develop and implement a coherent service catalog (see Fig. 6) for a comprehensive, easy access to all services of NFDI4Chem. Additionally, NFDI4Chem services will be embedded within a NFDI-wide service catalog supporting cross-domain research. Our services will foster the digitisation of key steps in research workflows starting from experiment planning, to data acquisition and management to data publication and thus will promote a digital change towards FAIR data handling. NFDI4Chem services build on a flexible, modular infrastructure of existing components whose interoperability is ensured by policies and standards. Software tools will enable the use of infrastructure components and the operation of the NFDI4Chem services. The services are further promoted by teaching & training activities, documentation and support will be available via helpdesk and a NFDI4Chem Knowledge Base. Policies and guidelines for a legally reliable data management and standards for data, metadata and vocabularies will be the fundamental building blocks. Repositories, databases and Smart Lab as infrastructure components as well as other tools and software components of the NFDI4Chem services will rely on this fundament. The synergy of all services will promote a cultural change in the handling of research data in chemistry.

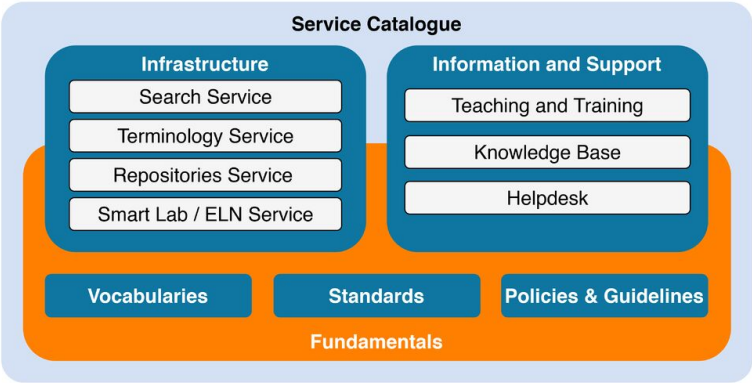


Figure 6.

NFDI4Chem service catalog.

The virtual environment of federated repositories for storing, disclosing, searching and re-using research data across distributed data sources, complemented by a Search Service, Terminology Service and ELN as a service will be the core services of NFDI4Chem, enabling the implementation of further concepts like the Smart Lab or the NFDI4Chem portal.

Databases and repositories that cover the relevant data types used by the NFDI4Chem community will be included into the envisioned federation of repositories shown in Fig. 7. Currently, **nmrshiftdb2** (Steinbeck and Kuhn 2004, Steinbeck et al. 2003, Kuhn and Schlörer 2015), hosted at UzK (Cologne), contains datasets for more than 40,000 molecules whose structures, chemical shifts (so far mainly for nuclei $^{1}H$ and $^{13}C$) and assignments can be accessed for search (e.g. dereplication, similarity, fragments, structure, signals) or prediction. More recently, deposition of raw data and electronic assignments in NMReDATA format (Pupier et al. 2018) were included and entries receive DOI's to facilitate citation in publications. Data can be exported in various formats, including as NMReDATA file. The nmrshiftdb2 database can be installed as local repository in NMR laboratories to improve integrating the workflow of academic chemistry groups via optional LIMS (Kuhn and Schlörer 2015) functionality. Additional tools were developed to assist and evaluate spectra assignment to further reduce the barrier of electronic data processing. It is part of the DFG-funded project IDNMR.

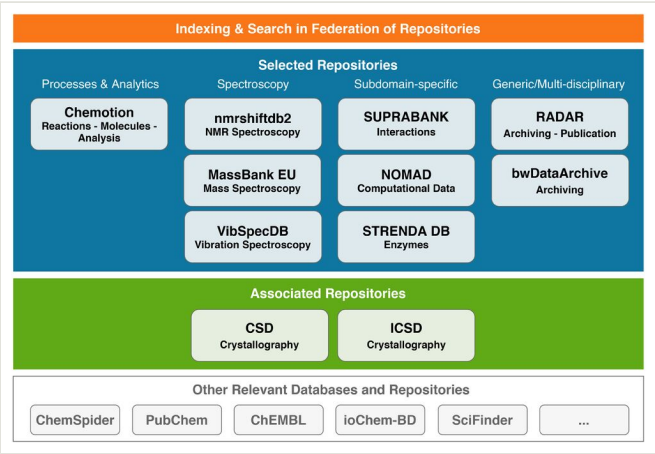


Figure 7.

Existing services forming the nucleus of the envisioned federation of repositories as part of the NFDI4Chem infrastructure.

**MassBank EU**, hosted at UFZ (Leipzig) is the first public repository of mass spectral data for sharing them among the scientific research community (FAIRsharing.org: MassBank 2016). MassBank data are useful for the chemical identification and structure elucidation of chemical compounds detected by mass spectrometry. The MassBank spectral data is hosted in a revision control system with all spectral data and the corresponding meta data in a human readable record format, and continuous integration (CI) checking record integrity for each change. Instances of the web interface are hosted at UFZ and IPB, and can be installed locally as well.

To complement our spectroscopic portfolio we have identified **VibSpecDB**, a currently internally utilised database for vibrational spectra (Raman and IR spectra) that is hosted at FSU (Jena). In the course of this project, VibSpecDB will be integrated into the NFDI4Chem spectroscopy concept and converted to full open access. Currently, the database itself features APIs for programming languages like Python or R, but no GUI based import routines, web-interface or viewers. These functionalities will be developed in the course of NFDI4Chem and a license as well as access-right management system will be added to the database forming a repository for vibrational spectra.

The **Chemotion Repository** covers research data that is assigned to molecules, their properties and identification as well as reactions and experimental investigations and is hosted at KIT (Karlsruhe). Scientists are supported in their efforts to handle data in FAIR manner: The data is stored along with molecule and reaction specific identifiers and DOI-assigned data files are given with distinct ontology-supported metadata (Anonymous 2018, Bräse et al. 2017). The findability of the data is achieved by a text and structure search and its availability via PubChem. The repository is interoperable with the chemotion ELN with respect to data transfer, offers export schemes to other systems and data are curated by automatic checks and a peer reviewing process. The integration of data stored in the

repository to publications was shown with several examples (Jung et al. 2014, Bär et al. 2019, Huang et al. 2018).

**Suprabank**, also hosted at KIT (Karlsruhe), is a curated database that provides data on intermolecular interactions of molecular systems which are not available in other repositories or databases. SupraBank is mainly aimed at supramolecular and physical chemists that deal with binding, assembly and interaction phenomena. Molecular properties are retrieved from PubChem, allowing the correlation of intermolecular interactions parameters to molecular properties of the interacting components. All molecules, solvents, and additives are searchable by their chemical identifiers. At present, the Suprabank stores >1100 curated data sets of intermolecular interaction parameters.

**StrendaDB** is a repository operated at BI (Frankfurt) for enzymology data providing the means to ensure that data sets are complete and valid before scientists submit them as part of a publication. Data entered in the STRENDADB are automatically checked, allowing users to receive notifications for necessary but missing information. Currently, more than 50 international biochemistry journals already included the STRENDA Guidelines in their Instructions for Authors.

The **NOMAD** repository (hosted at FHI, Berlin) enables the confirmatory analysis of calculated materials data, their reuse, and repurposing. It facilitates research groups to share and exchange their results, inside a single group or among two or more.

The repositories **RADAR** (FIZ Karlsruhe) and **bwDataArchive** (KIT, Karlsruhe) provide (certified) data archiving services and serve as catch-all repositories.

For the deposition of crystallographic data, NFDI4Chem will collaborate with the **CSD** (CCDC, Cambridge) for organic structures and for the **ICSD** (FIZ Karlsruhe) for inorganic structures. Both repositories, although being commercial, are established in the community and serve as a standard host for crystallographic data. The interoperability of both infrastructures with NFDI4Chem will be established during the funding period (see LoS).

In addition to the described infrastructure components in Germany, other international repositories and databases such as **ChemSpider** (molecule and physicochemical data), **PubChem** (molecule data, vendor and toxicology information), and **ChEMBL** (bioactivity data) will be connected.

The Repository Services are complemented by a Search Service and a Terminology Service (see TA6). The **Search Service** facilitates the access to all the resources available in the NFDI4Chem core and associated repositories. It will be built on a metadata catalog and provide a semantically harmonized access to the federated repositories based on the standardisation efforts for chemistry disciplines. Chemistry-specific search options include molecular structure and properties search. This approach will be combined with interconsortia harmonisations measures allowing not only cross-repository, but also cross-domain data discovery. The NFDI4Chem portal realises the concept of a single point of entry to NFDI4Chem services and further information. The Search Service, the knowledge base or the helpdesk will be provided via the portal. The **Terminology Service** will provide

machine-readable and human-readable descriptions of research data, thus enabling researchers and components of NFDI4Chem and NFDI to access, curate and update vocabularies for chemistry and related domains. Terminologies will be developed by community driven workshops and will be continuously extended to cover the needs of all relevant subdisciplines. In addition to the Repository Services, the components of the NFDI4Chem as well as the services, in particular the Smart Lab will integrate the Terminology Service to ensure the semantic description of the data with standardised vocabularies.

The **Smart Lab**, as part of the infrastructure, fosters data workflows from devices towards repositories in the federation. It integrates software components that are necessary requirements to build the NFDI4Chem services as well as an ELN that is provided as open source software component and as a service. Crucial building blocks for almost all NFDI4Chem services, but of special importance for the Smart Lab, are software and tools. The drawing, processing and visualisation of chemical structures is an important part of a chemical data infrastructure that enables comprehensive search, identification, quality assurance and curation of datasets in chemistry. Widely used software libraries such as Chemistry Development Kit (CDK), RDKit, OpenBabel support these functionalities. CDK (Willighagen et al. 2017), a Java library for cheminformatics is used in hundreds of software projects around the globe, which originated in the lab of Christoph Steinbeck. Software tools such as Ketcher editor, ChemSpectra, NMReDATA components, JSMol and KNIME will be additional components of the NFDI4Chem services. The Ketcher editor allows to draw chemical structures, supports templates and symbols for planning and processing of solid phase synthesis (Kotov et al. 2018), and tools for drawing complex molecules such as organometallic compounds. For the visualisation of spectral data, the Steinbeck lab developed the SpeckTackle widget and library (Beisken et al. 2015), which is used for spectrum visualisation in MassBank EU und will be used in further spectroscopic developments in NFDI4Chem in combination with ChemSpectra, a software to visualise and analyse spectroscopic data with integrated solutions for IR (Infrared), MS (Mass), and one dimensional $^{1}H$ and $^{13}C$ NMR spectroscopy. Ketcher, ChemSpectra and SpeckTackle (provided by participants) will be improved and adapted during the funding period to be embedded into the NFDI4Chem components and services where appropriate. Device integration and data transfer are also part of the Smart Lab. The first solutions for seamless data transfer from devices to ELN already exist among the consortium members (Kuhn and Schlörer 2015). Solutions for smaller lab devices (e.g. balances) have been elaborated (Lütjohann et al. 2015, Lütjohann et al. 2014). Concepts for the availability of data coming from analytical lab devices, for instance, were proposed (Potthoff et al. 2014, Potthoff et al. 2019). These solutions will be generalised and expanded to more lab devices with the ambitious aim of building a comprehensive library of such device integrations. As this is a cross-cutting topic among lab-based experimental research, NFDI4Chem will join forces with related consortia to achieve this goal.

The **ELN** supports scientists in collecting, managing, storing, analysing, and sharing data as a preparatory step to disclose data via a repository. NFDI4Chem will extend today's notion of ELN by complementing its core functions with additional features (see software

tools and libraries), thus broadening the scope of the ELN to a digital research environment. The ELN will be designed as an open, modular platform, combined with concepts to handle device integration and to transfer data interoperably to repositories or databases. It will offer interfaces to external sources and web services. The ELN to be developed covers the needs of multiple subdisciplines of chemistry. NFDI4Chem has chosen Chemotion ELN as reference service but will also consider other ELN used by the community (e.g. ElabFTW, Open Enventory). Chemotion ELN provides generic management tools as well as extensive functionalities for workflows in organic chemistry (Tremouilhac et al. 2017, Jung et al. 2017). Its interoperability has been demonstrated with well-known databases like SciFinder and PubChem and the embedded visualisation of predicted NMR data via NMRdb.org. The Chemotion ELN is listed by several international services (Electronic Lab Notebooks, HMS 2019, Douglas 2017), is scrutinised in current (not domain-specific) initiatives aiming to review the most relevant ELNs for a broad roll-out (Schultze-Motel 2018), is presented on international conferences (Jung 2019), and was mentioned recently as a best practice example in Adam and Lindstädt (2019). Other projects, such as the BMBF-funded NanoS-QM project, have selected Chemotion-ELN to test a consistent data management in the highly interdisciplinary field of nano-safety research. The Chemotion ELN will be provided as continuously updated open source software with full documentation. All described functions of the ELN are part of the **ELN as a Service**. The centrally hosted offering (KIT-SCC) addresses those scientists working interdisciplinary or in smaller chemistry groups who do not have access to a suitable IT facility or data centre which could run a local ELN instance for them. An ELN as a Service provides these scientists with an easy way to work in accordance with the FAIR principles. At the same time, the ELN as a Service as a Service serves as an instance for teaching and training purposes (see TA5). To support digital data management as early as possible in the scientific career, the ELN as a Service is available for practical courses in all subdisciplines of chemistry, and in particular suitable for those that are involved in the development of the software in M2.3.

A cultural change in data handling and management in chemistry, as envisaged by NFDI4Chem, is necessary to implement the FAIR data principles. This change is an ambitious undertaking, in particular as new software and instruments have to be introduced, new workflows have to be established and the overall behaviour of the daily work has to be changed. NFDI4Chem will support the scientists in this process with a service for **Teaching and Training** and a **Helpdesk** to foster acceptance and use of the community. The teaching and training team is of high importance to foster the general acceptance of the infrastructure and to enable the scientists to use the components of the infrastructure in the right manner. NFDI4Chem will teach scientists about how to use the provided infrastructure for digitalised work according to the FAIR data principles and will raise awareness for its importance. The training team will explain the details of the single components of NFDI4Chem, and will train the use of especially the ELN and the repositories including the data transfer methods. The team will plan roadshows and will be available on request, offering not only theoretical advice but live demo sessions with the centrally hosted ELN and its functions.

Besides NFDI4Chem services, we will further build on NFDI services relevant to NFDI4Chem like the AAI-service for user authentication and authorisation, and on multi-disciplinary services like DataCite (via the TIB), ORCID, and Re3Data.

**Contingency measures**

The reliable and sustainable operation of the services provided by NFDI4Chem is of great importance for their acceptance by researchers. Extensive and multifaceted contingency measures are therefore planned.

Long-term **data availability** is boosted by the use of standardised data formats. In addition to the definition of such data formats, the consortium will also provide data converters and software components for the generation, viewing and verification of these formats. This is supported by the development of curation criteria that ensure data quality (see TA4). Central databases and repositories are supported both technically and organisationally in successfully passing certification (Core Trust Seal) and thus demonstrating their sustainability (see TA3). The repository operators thereby draw on the expertise of experienced infrastructure facilities. These facilities also offer established services for the long-term archiving of research data. At the same time, the consortium develops exit strategies for repositories and provides resources for data rescue in order to preserve important but no longer maintained data sets by transferring them to central repositories.

As software plays an increasingly important role in the digital transformation of chemistry, **sustainable software design and software quality** are important aspects. We will foster a modular software architecture with standardised components for reuse in multiple services provided by the consortium and beyond. We introduce Continuous Integration (CI) and Continuous Delivery (CD) processes in order to provide always workable and tested software in short development cycles even with distributed software development across the many participants. NFDI4Chem will broaden the knowledge of state-of-the-art software design and software quality by organising workshops in cooperation with the German chapter of the Association of Research Software Engineers (see LoS of de-RSE and R. Reussner). All software created in NFDI4Chem is developed as open source software and is published under OSI-compliant licenses on public repositories such as GitHub. Extensive and good documentation both of source code and of services as well as training for developers and users contribute to the growth of active communities that ensure the ongoing maintenance and further development of the services.

## Work Programme

### Overview of task areas

Table 1

Table 1.
Overview of task areas

| Task Area | Measures | Responsible Co-Spokesperson(s) |
|---|---|---|
| Management | M1.1 - M1.4 | Christoph Steinbeck |
| Smart Lab | M2.1 - M2.6 | Nicole Jung |
| Repositories | M3.1 - M3.4 | Felix Bach, Matthias Razum |
| Standards | M4.1 - M4.5 | Steffen Neumann, Christoph Steinbeck |
| Community Involvement and Training | M5.1 - M5.6 | Sonja Herres-Pawlis, Johannes Liermann |
| Synergies | M6.1 - M6.4 | Oliver Koepler |

## Task Area 1: Management

**Description and General Objectives**: TA1 provides adequate and lean leadership and support to all Task Areas in achieving their objectives. The highly collaborative and distributed nature of NFDI4Chem calls for an effective management structure and a sound decision making process to be in place, to ensure: efficient planning and controlling of project activities, seamless communication across partners, robust and transparent decision making, balance multiple responsibilities and competing priorities of consortium partners, prompt reporting and finally, successful project delivery. The overall management structure is illustrated in Fig. 1. The details of the distinct levels and responsibilities of the management are discussed below.

**The Project Office (PO)** is located at the Friedrich-Schiller-University (FSU) in Jena where space and facilities are available for administrative purposes. The PO supports the Project Speaker as well as the steering committee in the day-to-day operational management of the project and handles the administrative management, the compliance to contractual obligations of the Consortium Agreement and the correct dissemination and exploitation of the project results. The PO is also responsible for the appropriate communication with the consortium and the DFG and will handle the financial administration and safeguard the adequate execution of the project budget. The PO will manage and monitor the project progress in order to meet the project objectives, handling time and resource constraints appropriately.

The PO is headed by Prof. Steinbeck and consists of the project management team with a project manager exclusively hired for this project, and experienced staff from the FSU administrative and financial team, the FSU funding coordinator Dr. Margull and the FSU press office as required during the course of the project. The PO members guarantee adequate administrative project controlling, coordination of the reporting, take care of financial and budgetary matters.

To summarize, the objectives of TA1 are:

**O1.1**: Efficiently **manage the consortium activities** to maximise NFDI4Chem's impact. If necessary, handling time and resource adjustments appropriately.

**O1.2: Organise and document all NFDI4Chem services, consortium, advisory board and stakeholder meetings and decision making processes**, as well as regular staff

exchanges between the NFDI4Chem partners in collaboration with our consortium partners.

**O1.3**: Safeguard **compliance with the contractual obligations of the Governance Model** and correct dissemination and exploitation of the project results.

**O1.4**: Manage central funds to react to necessary future project extensions.

These objectives will be pursued through the following measures:

**Measure 1.1: Overall legal, contractual, ethical, financial and administrative management of the consortium**

**Goals**: Ensure the legal and financial operation of the consortium

**Description**: This measure will deal with the management of the project funding and the monitoring of the decision-making procedures, always in compliance with contractual obligations under the consortium agreement.

**Task 1.1.1 Negotiate and conclude a consortium agreement with all partners**

A consortium agreement will be negotiated with all partners to ensure a legally sound distribution of funding from the funder via the FSU to all co-applicants and participants. The consortium agreement will further set the framework for the successful project implementation and sets out the rights and obligations between the partners.

**Task 1.1.2 Transfer of annual budget to partners. Retrieve and collate reports on the use of financial resources by partners**

Based on the transfer agreement negotiated above, the FSU administration will ensure timely and correct transfer of funds to all partners. We will also collect all information for the reporting required by the DFG.

**Task 1.1.3 Maintain communications with the NFDI headquarter and the DFG**

This task comprises regular reporting to and communication with the DFG and the NFDI headquarters. Reporting to the DFG will adhere to their rules of resource usage

**Deliverables**: (D1.1.1) Consortium Agreement Document, (D1.1.2) Report on resource usage by partners, (D1.1.3) Open documentation of NFDI cross-cutting activities related to NFDI4Chem on portal.

**Measure 1.2: Coordination at consortium level of the technical, outreach, training and cross-cutting activities of the project**

**Goals**: Ensure the timely and precise execution of the work plan through effective and agile management of a highly distributed project.

**Description**: The NFDI4Chem speaker, assisted by the project manager for day-to-day management of the project, will be closely monitoring and coordinating the activities of the consortium based on the work plan laid out in this proposal.

**Task 1.2.1 Organise and document consortium and TA meetings**

As part of this task, we will organize monthly tele-meetings (Skype, Hangouts, phone) of the NFDI4Chem steering committee. Discussions and decisions will be minuted. We will invite national and international collaborating PI's to participate if needed. Technical teleconferences of the Task Area participants and leads will be held separately and likewise individually documented.

**Task 1.2.2 Organise and document advisory board meetings for NFDI4Chem**

In close coordination with our learned societies, we will form the Advisory Boards (ABs) described above and maintain communication with the ABs through meetings which will be held at least bi-annually. ABs will be invited to join the annual consortium meetings organised in this TA. AB's will be consulted for advice on questions of strategic importance for NFDI4Chem and their advice will be evaluated within the consortia and with neighboring consortia as well as used to steer the future directions of NFDI4Chem.

**Task 1.2.3 Day-to-day management of the NFDI4Chem project**

This task will be dedicated to the execution of the work **plan** as laid out in this proposal. At the beginning of the project we will produce a **detailed project plan** which will include a list of success indicators to monitor during the whole project, as well as the data we will gather that will help in assessing its impact. These indicators and metrics will be reported at least in the annual reports. It will **organise**, focus, continually motivate, and empower the project staff to do their work. It will perform **controlling** of the project, by tracking the work and comparing it and results against the this work plan. It will use information from these tracking efforts to make changes to plans when the information suggests that a change is called for. It will continuously monitor project risks and mitigate them if necessary.

**Deliverables**: (D1.2.1) Continuously updated protocols and minutes of consortium and TA meetings, (D1.2.2) Minutes of Advisory Board meetings published on portal, (D1.2.3) Open project plan continuously updated on portal

**Measure 1.3: Coordination of knowledge management, Internet Publishing System and other innovation-related activities**

**Goal**: The consortium, the NFDI as a whole, and the chemistry community will be holistically informed about the work of NFDI4Chem

**Description**: Efficient, complete and accessible information about NFDI4Chem's work, progress and results will be maintained and disseminated via our internet publishing system and social media. This will allow our users and stakeholders, the consortium, the NFDI as a whole as well as the wider public to evaluate our progress and best use our services.

**Task 1.3.1 Develop, maintain and host an informative portal**

The NFDI4Chem portal at https://www.nfdi4chem.de/ will be developed and maintained in close coordination with the project's community and training activities (e. g., T5.2.1 and T5.3.1). A corporate design for NFDI4Chem will be developed. **The NFDI4Chem portal is the single point of access for user services and information**. Most prominently, it will provide access to the catalogue of services provided by NFDI4Chem. Secondly, a Knowledge Base (see T5.3.1) with documentations, resolutions, work-arounds and best practices for the NFDI4Chem services and tools will support both helpdesk staff and users. Each tool will be well documented and accompanied by didactically structured online manuals and multimedia web-tutorials.Thirdly, users will find help through a ticketing system by the RDM helpdesk unit (see T5.3.4)

For the consortium, the portal allows for content management by the partners, and employs additional components e.g. for supporting intranet, calendar, portal searches as well as advanced analytics, functional testing, communication via mailing lists.

**Task 1.3.2 Develop and maintain full documentation of all NFDI4Chem activies**

The policies, standards and workflows developed in this endeavour will be formally documented and published in the form of manuals, white papers and recommendations. Any document created under this umbrella will be released under a Creative Commons License to allow for barrier-free dissemination.

**Deliverables**: (D1.3.1) The NFDI4Chem portal, (D1.3.2) Accessible documentation on NFDI4Chem portal

**Measure 1.4: Organise and coordinate centrally managed funds for future project extensions**

**Goals**: Enable NFDI4Chem to react to new and unforseen developments in the national RDM landscape.

**Description**: We will reserve 5 FTE annually as centrally managed funds to react to new developments in the community which need to be integrated into NFDI4Chem. A possible scenario, for example, is the successful funding of a database project by DFG LIS funding, which then needs to be integrated into NFDI4Chem. The project leader can then apply for a number of person months to fund this integration. Calls for lightweight proposals will be held regularly. These proposals for cost-neutral project extensions will be reviewed by the DFG.

**Task 1.4.1 Organise future project extensions**

TA1 will regularly review, in cooperation with all NFDI4Chem partners and the community, the need for cost-neutral project extensions to integrate novel and unforeseen developments in to the NFD4Chem. This will be further supported by the innovation incubator organised by measure 5.5. The resulting proposals to the DFG will be lightweight, the amount of person month granted will typically be suited for the adoption of

an already externally funded project to NFDI4Chem-agreed standards and technologies. Proposals will be evaluated by the DFG.

**Deliverables**: (D1.4.1) Report on required project extensions. List of funded projects published on portal

**Measure 1.5: Overseeing the promotion of Equal Opportunity in the project**

**Goals**: Ensure that minorities and female researchers are provided with equal opportunities in the project.

**Description**: Equal opportunities are a central part of the staff development strategy in modern institutions. This comprises the support of recognised minorities in general and female researchers in particular. Data Science, Cheminformatics, and even more so computer science experience an under-representation of female developers and researchers. The promotion of equal opportunity for minorities in science is therefore an important component of NFDI4Chem management.

**Task 1.5.1 Promote and optimize equal opportunity measures across NFDI4Chem**

We will collect and report information about efforts to improve the provision of equal opportunity across the project. Based on this information we will constantly aim to improve our way of working and advertising towards better provision of equal opportunity.

**Risks and Mitigation Strategies**

This proposal is mostly concerned with the development of infrastructure, software, APIs, containerised tools and workflows in Chemistry, as well as their interaction with infrastructures and services in the NFDI as a whole and beyond (such as the EOSC). The expertise of the scientists within the NFDI4Chem consortium is excellent and suitable for these tasks. The leadership has a strong track record for delivering excellent results in large scale projects on time. The structure of the grant proposal is simple and tight, allowing straightforward assessment of progress and giving confidence that the objectives of the grant will be reached. The tasks described are very concrete and measurable and based on previous well-defined achievements of the partners. Unlike fundamental research, where there are many unknown outcomes, the partners have previous experience in all of the proposed work and are confident that it can be completed in time. We are aware of typical risks in such developments, which are delayed developments, unstable and unsuitable software and the inability of partners to make concerted and well-accepted decisions. The specific technical and management risks have been declared in detail in tables in the individual task areas.

The partners are known to be willing to engage in the culture of Open Science, Open Source, and Open Standards, as well as knowledge, policy and data sharing. The actual development of infrastructure components and services and even whole infrastructures has been exercised by the consortium members.

All partners contribute to the key task areas on repository provision and supporting activities. The challenge is to get the NFDI4Chem services accepted and used by the community. A central component to ensure that our NFDI4Chem is widely accepted and used is TA5, where our partners as well as the wider community will be engaged and supported to use and develop for our services. (Table 2)

Table 2.

Table to Task Area 1: Management.

| Description of management risks | Proposed risk-mitigation measures |
|---|---|
| RM1: **Ambitious scale**. Coordinating a project requiring many different expertises across institutions needs in a vivid scientific field. A loss of focus could lead to poor delivery.<br>**Likelihood**: Low | Frequent internal electronic and video communication of the SC, the Annual review process at the stakeholder meetings and review by the NFDI and DFG, which will track progress.<br>**TA(s) involved**: All |
| RM2: **Competitive labour market.** Difficulties in hiring of skilled staff in a highly competitive job market might be a risk. Computer Scientists with expertise in infrastructure development and distributed computing are under very high international demand.<br>**Likelihood**: Medium | The project will need to have enough time between the funding decision and its start to scout actively for suitable candidates, and international job advertisements in traditional channels and social media. In addition to that several of the partners have already identified possible candidates for their staff positions<br>**TA(s) involved**: TA2, TA3, TA4. |

## Task Area 2: Smart Lab

### Description and Objectives

The overarching aim of the task area 2 “Smart Lab” is to develop and provide a modular virtual lab environment of concepts, services and software for fully digitalised research data management workflows within NFDI4Chem. The modular approach of the software design with open interfaces is a key aspect enabling flexible combinations with alternative NFDI4Chem or external services. TA2 will also strongly feed into the subsequent task areas through the development of software components and design decisions that will be adopted by TA3 services. TA2 will help to investigate requirements of subdisciplines in chemistry which fill fertilise developments both in TA3, TA4 and beyond. Procedures documented in TA2 will feed into the overall knowledge base developed in TA5. We have therefore given TA2 more room to exemplify the careful and detailed approach that we will take in all NFDI4Chem developments. The work of TA2 will span solutions from data acquisition with different devices, the collection of experimental or computational data, the capture of metadata as well as management and analysis of data and its preparation for further storage in repositories or a publication in scientific journals (Fig. 8). TA2 will aim to solve the technological challenges that currently hinder the digital availability, storage and seamless transfer of data in chemistry laboratories. The availability of FAIR data, enables the transfer of the data to repositories and databases for the scientist through standardised interfaces. All measures will be elaborated in close collaboration among participating scientists, developers and institutions. The Smart Lab with full device integration requires a decentralised architecture and institutional hosting. The results of the measures of TA2 are

therefore rolled out as software packages (see T2.4.1) to be installed and operated in the different research institutions. This rollout is supported by a technical helpdesk (second level support) for on-site or remote technical support (see M2.2) and a teaching and training team (see T5.3.2, T5.3.3). Alternatively NFDI4Chem will provide a centrally hosted virtual lab environment, an ELN as a service (see M2.4) where institutional hosting is not viable.

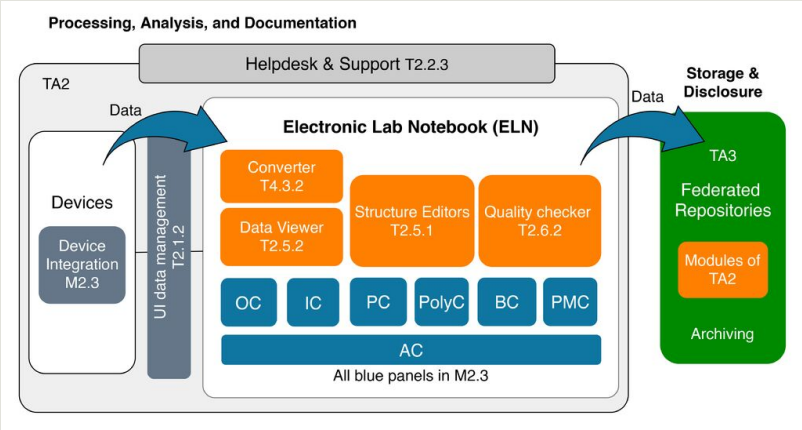


Figure 8.

Components of TA2, Smart Lab, and connection to TA3: Seamless data transfer from devices via the ELN to repositories. Grey panels: modular software plugins, orange panels: modular software plugins for ELN and repositories, blue panels: software developed for the ELN. OC, IC, PC, PolyCC and AC represent the modules of the subdisciplines to be developed (see list of abbreviations in the last section of the Grant Proposal and the table to it).

To summarize, the objectives of TA2 are:

**O2.1**: Enable digitalisation of processes in chemistry as a prerequisite to obtain FAIR data.

**O2.2**: Development of a digital research environment (Smart Lab) that provides all functions to collect, manage, store, analyze and share data.

**O2.3**: Offering an ELN (as part of the Smart Lab), designed as a modular platform that allows to integrate devices, to use external sources of information and to transfer data to repositories.

**O2.4**: Development and installation of tools for all subdisciplines in chemistry to facilitate digital RDM according to the FAIR data principles.

**O2.5**: Making the components of the Smart Lab available to all scientists in Germany which are currently missing suitable means to manage data in a FAIR manner.

TA2 will contribute in particular to key objective 2: *minimum information (MI) standards for data and machine-readable metadata*, and key objective 3: *Smart Laboratory Environments by the following measures*:

**Measure 2.1: Device integration and management**

**Goals**:

1. Develop flexible and modular Open Source solutions to integrate scientific devices which are added to a digital workflow.
2. The developments should be available in all research institutes that miss suitable solutions for device integration so far.

**Description**: Devices play a pivotal role in an experimental work environment in chemistry labs. They are used for the conduction of experiments, the recording of data and the analysis of results. Common examples are gas and liquid chromatography, NMR, UV or IR spectroscopy and mass spectrometry. Besides, each subdomain in chemistry uses specific equipment. The diversity of the used devices is large and their integration into a digital workflow is challenging. Many instruments used in academia are legacy devices with outdated software/operating systems or missing webinterfaces.

The depicted challenges will be met by three strategies addressing risks and uncertainties in the adoption of solutions by users, the dynamics of such an adoption as well as support, or lack thereof, that we receive from device manufacturers for our work: **Strategy A** will propose a solution for the desired device integration in the short term independent of other stakeholders. **Strategy B** will rely on the installation of standardised processes for the seamless transfer of data from devices to other units of the digital lab environment. Important stakeholders such as manufacturers of devices will be involved. **Strategy C** complements the solutions towards the NFDI4Chem workflow by the integration of open source LIMS systems.

**Task 2.1.1 Strategy A, part 1: Central data availability**

The access to centrally stored data is a prerequisite to further transfer, convert, analyze and manage data. By Strategy A, a data transfer strategy from devices to a central institute's server will be defined and elaborated. The processes will work for a diverse set of possible devices that may be connected to different networks depending on the institution where the devices are located. Due to the very different requirements of the necessary solutions, the outcome will be a set of categories and processes depending on the digitalisation level that the device allows. The level categories will range from 0, i.e. no integration to digitalisation strategy, to 10, full integration possible, showing the ability for a seamless integration and the processes to be applied. The results will be disseminated via a service catalogue, a registry for open source solutions established by M6.3. The solutions to be developed will be set up in a similar manner with respect to the work described previously (Potthoff et al. 2019, Lütjohann et al. 2015), providing improved, reworked processes and a comprehensive incorporation of all workflows. In particular, the device and vendor-dependent options to store data on different locations need to be extended systematically to cover the currently applied systems in chemistry labs. The solutions will be developed in a generic manner, if possible. If customized solutions have to

be developed for special devices and vendors, the TA2 team will prioritize efforts based on the evaluation of the NFDI4Chem user surveys.

**Task 2.1.2 Strategy A, part 2: Data management**

Routines to monitor new incoming data on the registered data server(s) and assignment of the data to the responsible scientists will be implemented. A clear and diverse UI offering a fast organisation of the incoming data will be developed, as well as an optional function to automatically assign the data not only to scientists but also directly to the experiment, including the generation of an additional, manipulation-resistant backup of the transferred data. The developments will be embedded into the UI of the ELN developed in T2.2.1, to provide a user friendly application according to the principles of LIMS.

**Task 2.1.3 Strategy B: Development of standards and interfaces for data exchange with devices**

NFDI4Chem will push the process of standardisation of data storage, data transfer and open interfaces with members of TA2, TA4, and TA6 and in cooperation with e.g. the RDA, Pistoia Alliance, and Allotrope Foundation. By bringing together the important stakeholders on a national level and proposing solutions to the processes that are important in the context of NFDI4Chem, T2.1.3 will prepare solutions for device integration on an overarching level. Within Strategy B (‘device interoperability’, T2.1.3), the installation of standardised processes for the seamless transfer of data from lab devices including simple balances and complex analytical instruments will be fostered.

**Task 2.1.4 Strategy C: Using synergies by connecting open source LIMS**

Strategy C will integrate open source LIMS solutions to the NFDI4Chem workflow to be used complementary to the more general approach to data availability and management described in T2.1.1 and T2.1.2. Suitable LIMS systems will be selected after a short evaluation phase of the currently available open source systems according to criteria such as device integration and interoperability with other software like the ELN-software to be used in NFDI4Chem (Mu 2018). Centrally managed funds (see T1.4.1) will be used to support the best model to integrate either already available LIMS or a novel innovative approach to meet the requirements of the users. The software can be used on the one hand to manage devices, e. g. to place and queue analysis requests and to track samples (remote functions & automation) and, on the other hand, to transfer and manage the results in form of analytical information.

**Deliverables**: (D2.1.1) List of devices of the subdomains that are covered by the data transfer protocol, (D2.1.2) Management system for device data is released to GitHub, (D2.1.3) Important manufacturers agree on a common strategy for data storage and transfer via standardised protocols, (D2.1.4) A management system for sample interactions and transfer visualizing the status of all samples is implemented.

**Measure 2.2: Establishment of the Electronic Lab Notebook (ELN)**

**Goals**:

1. Extension and improvement of the functionality of the Chemotion ELN as a component of the open, modular virtual lab environment.
2. Efficient coordination and teaching of the decentralised ELN development team and management and review of all source code contributions.
3. Adapt the current open source code basis to FAIR data management.

**Description**: Within M2.2, the core software of the ELN will be defined, improved, extended and continuously upgraded. The ELN will be developed based on the existing open source software Chemotion-ELN (Tremouilhac et al. 2017, Kotov et al. 2018) which was chosen after an evaluation of the currently existing open source solutions for RDM (Adam and Lindstädt 2019, Tremouilhac et al. 2017). The developments will ensure, along with the subdiscipline-specific adaptations of the ELN software in M2.3, the suitability of the ELN for a broad chemistry user community and its acceptance by the scientists. Furthermore, the work of M2.2. will include all tasks to guarantee the sustainability of the created software and the quality of the outcome of the source code which has to be merged from the contributions of many programmers. The M2.2 team will implement and improve the main features to enable FAIR data management and the capture of metadata, provenance and context of research data as early as possible in the data process by using the proposed ELN. The resulting ELN platform serves as a basis for all further improvements and additions by the community or other NFDI4Chem working groups and is understood as an offer to those scientists that are currently searching for an ELN. NFDI4Chem develops this tool as a best practice solution fostering FAIR data management in every lab. In addition, the proposed ELN offers a fast solution to publish and share data in NFDI4Chem supported repositories.

**Task 2.2.1 Basic ELN structure, coordination of sub-tasks**

Here, the basic ELN structure will be created and maintained, the sub-tasks elaborated by the subdisciplines will be coordinated core developer team at the KIT and merged and general implementation for all subdisciplines will be developed. The results of the M2.1, M2.3, M2.5, and M2.6 will be reviewed and merged to the main code. T2.2.1 additionally ensures the quality of the developments in TA2 and the establishment and preservation of the necessary expertise within the development team. The T2.2.1 ELN team will be responsible for the training of programmers who will become part of the developer's group. Furthermore, the core ELN team will coordinate and merge the documentation of the developments ensuring a sustainable software development. The integration of an NFDI-wide AAI into the services of NFDI4Chem will be covered as soon as the consortia agreed on this cross-cutting topic. In parallel, a detailed role management including options for support and possible administration by NFDI4Chem staff will be developed in cooperation with T6.3.2. An advanced definition of user settings for UI, data storage options, computing resources for the creation of a "customized" ELN will be designed. Further, the ELN will support the terminology service developed in M6.1 to ensure the semantic enrichment of the data in the ELN for all subdomains. The usability of the ELN will be improved by a data file versioning tool that allows the versioning of data directly in the ELN even if proprietary

software is used to open and edit the files. The trustability of the ELN will be fostered by the implementation of electronic signatures for all processes to be managed in the ELN. A comprehensive audit trail recording will be integrated.

**Task 2.2.2 Generic implementations relevant to all subdomains**

In this task, the ELN software will be complemented by advanced RDM tools for non-standardised processes such as experiments in interdisciplinary projects or the integration of computational data. To enhance customization of the ELN, user-defined additional entities (main working elements of the ELN), will be established. According to the user's requests, additional open source tools, such as an annotation tool for images have to be included. Data converters developed in T4.3.2 will be integrated, including UI and backend processes, to convert incoming data automatically according to TA4 standards. The standardised data and metadata can be transferred to and used by additional storage and computing resources. To this end, the necessary UIs and protocols will be developed. With respect to the connection of storage options, interfaces to NFDI4Chem and international repositories are of importance. An automated repository advisor will propose the transfer of data to a suitable repository. Also chemistry specific requirements, such as the coordination and collection of a well-curated reagents database for all subdomains in chemistry including toxicity data, safety sheets and physical data that is available by extracting or connecting external databases are elaborated in T2.2.2. To enable communication between different users of the ELN messenger, notification and commenting tools are added and a feedback button for the communication of ELN-users with the development team will be created (to be implemented for ELN as a Service, see M2.4 and T5.1.1). Over the whole funding period, the UI of the ELN will be further adapted to the most common browsers and non-standard screen resolutions. In addition, helper tools have to be added to guide new users through the functionality of the ELN.

**Task 2.2.3 Technical helpdesk (second level support)**

A second level support will be formed by the technical helpdesk in TA2 receiving user requests by the general helpdesk installed in TA5. The technical helpdesk will solve all technical issues (installation, set up, maintenance and use) of the admins or users with respect to the ELN, the device integration modules and will answer questions with respect to software requirements and issues. The technical helpdesk will support by remote assistance and -if necessary- by on-site support, guaranteeing that the concept of NFDI4Chem will be successful, independent of the expertise of individual institutions with respect to technical issues.

**Deliverables**: (D2.2.1) Ten releases of ELN code, (D2.2.2) documentation is available via Knowledge Base (TA5) and can be re-used by teaching team, (D2.2.3) Establishment of the technical helpdesk (second level support).

**Measure 2.3: Development and extension of subdomain specific ELN functions**

**Goals**: Adaptation of the NFDI4Chem proposed ELN by developing modules and associated features that are necessary for a broad adoption of ELN in subdomains of chemistry research.

**Description**: The extension of the ELN to the subdisciplines of chemistry has to reflect the needs to enable FAIR data management with respect to the standards developed in TA4, the overall strategy and dependencies of NFDI4Chem and the interoperability of all services (see Figs 6, 8). The RDM and scientific extensions will be elaborated by working groups formed by different scientific users and advisors. The subdisciplines represented in this TA were selected based on the number of members in the respective divisions in the German Chemical Society and by maximizing the diversity of their domain-specific requirements.

One developer, assigned to one of the scientific members of the working group, will realize the required extensions. The scientific users consist of representatives of at least three groups (additional members of the project groups see LoS) from different institutions with a different research focus to ensure an equal representation of the defined needs. Additional requirements to be included will be continuously gained by the outcome of the community surveys (see TA5). The working group is organized through regular meetings where the ELN developer team (see M2.4) as well as TA4 and TA5 are involved. The developments will be designed, wherever possible, as a library or plug-ins allowing the re-use in other software applications and services. All new modules will be fully compatible to enable a flexible combination of the available modules (guaranteed by software testing and CI). The preliminary developments and parallel additional support by other projects such as the science data center (Bräse and Nestler 2019) will accelerate the availability of “user-appreciated features” on top of the mandatory infrastructure, creating an ELN software that scientists enjoy to use. Advanced functions will create incentives for the user such as faster access to data or access to additional calculated, predicted or retrieved information (examples see T5.5.2) (Materials Project 2019). The working groups’ interactions will be coordinated by T2.2.1. All subdisciplines will include measures for quality control based on the already available quality control features of the ELN that will be closely coordinated with measure 4.1. The requirements analysis performed for defining the needs of measure 2.3 will inform the developments of MI standards for the respective subdisciplines and vice versa.

**Task 2.3.1 Subdiscipline Organic Chemistry**

For the subdiscipline Organic Chemistry (OC), data visualisation, assignment and annotation options will be generated for reactions that include several reaction steps. This feature is important for the generation of oligomeric structures with well-known but often repeated reactions with certain building blocks that are used in peptide or oligonucleotide synthesis. It is also necessary for the documentation of reactions with unknown or labile intermediates not to be represented as a single molecule. Additionally, a workflow scheme has to be designed that summarizes the reaction path of multistep reactions (summary of a synthetic sequence such as a total synthesis) in a flexible manner, giving a schematic overview to be annotated (and exported) for internal documentation or publication. The

basic calculation tools referring to the information in the reaction table and the analytical results have to be extended by calculation tools. Parameter descriptions will be amplified (e.g., temperature single value/range to be extended to a temperature profile with annotations). An analytical data manager will be developed, reflecting the needs to manage different sets of analyses for different samples of the same target molecules.

**Task 2.3.2 Subdiscipline Inorganic Chemistry**

The needs of Inorganic Chemistry (IC) differ substantially from OC in the requirements for structure representations (with respect to the application in catalysis or organometallic complexes with multiple bonding schemes) and for specific ELN functions e.g. with respect to kinetic investigations. The current manual documentation tradition in IC does not allow a consequent standardised structure representation, reactions consisting of many steps including time resolved data acquisition are hard to record and a correct annotation of the data without digitalised workflows is not feasible. An adaptation of the ELN includes therefore

1. the extension by specific documentation, calculation and visualisation requirements in catalysis,
2. the implementation of a module for kinetic measurements and
3. the improvement of the open source structure editor Ketcher (see T2.5.1) integrated in the ELN.

Predefined template sheets for substrate and condition screening experiments, advanced tagging methods, easy to visualize time-vs-conversion plotting tools will be developed, as well as a multi-spectra UI for spectra of kinetic measurements or calculated spectra and to assign them to the reaction progress.

**Task 2.3.3 Subdiscipline Physical Chemistry**

For many applications in Physical Chemistry (PC), devices play a central role in the scientific process. Device data are not only used for the identification of compounds or the confirmation of a result but are often a scientific result in themselves. The ELN will be extended with functions reflecting the special requirements resulting from experiments that strongly depend on device settings, device configuration and different setups. RDM in PC does currently not allow a standardised record and annotation of the setup and device configuration which hinders reproducibility and comparison of experiments. Modules for PC will allow the configuration of a device setup (incl. self-built instruments) either as flexible tool with drawing and annotation features or as predefined elements registered to the ELN (e.g. microscopy, spectroscopy or laser table setup). The module for PC will further provide a comprehensive section to store information on measurement parameters, outcome parameters and descriptors, to be defined as predefined or flexible information fields after a detailed analysis of the scientists needs (defined by TA5, surveys). We will implement a module for equations and plotting areas to visualize analyzed data such as fitting models and also computational results, as well as a workflow designer combining the experimental investigation incl. several instruments, software and algorithms. For the establishment of

an efficient physical chemistry information system, connection to open source software for calculations (e.g. R) as well as image processing toolkits such as ImageJ will be enabled.

**Task 2.3.4 Subdiscipline Polymer Chemistry**

This working group will use results of the Science Data Center MoMaF as a basis and will add additional requirements. The work will include the extension of the Ketcher editor with respect to Polymer Chemistry (PolyC) relevant standards and recommendations (see T3.1) in accordance with the IUPAC nomenclature. This issue requires constant approval and comprehensive rework due to ongoing discussions about polymer representation, processing to identifiers (Audus and de Pablo 2017) and recording of undefined or partly defined parts in a machine readable manner. The PolyC working group will include calculation tools in form of a separate PolyC reaction table to plan and evaluate experiments including the calculation of polymer-relevant parameters such as molecular weight distribution from size exclusion chromatography (SEC). A special measurement module for results and values of analytical PolyC-investigations such as GPC, TGA, DSC and others will be created and correlated with the evaluation of different types of polymers and their fabrication processes to assign and manage the relevant metadata. The work will be coordinated with measure M2.1, where the devices are connected and data transfer procedures are established. The needs of the community surveys (see TA5) will be included to the work of the PolyC ELN development group.

**Task 2.3.5 Subdiscipline Biochemistry**

Here, functions for storage and representation of processes required for Biochemistry (BC) are added to the ELN. Processes include isolation and purification of small molecules or biopolymers and the enzyme-catalyzed conversion of natural and synthetic molecules. The functionality provided should support the assignment of single samples/batches including the processing status of analytical confirmation to a certain point of the workflow, given in list format and in a graphical summary. With respect to an extension of the existing entities of the ELN-database, the new entity “biomolecules” allows the storage and annotation of large molecules such as small peptides, peptide analogs and small nucleic acid sequences. The module will extend the available main entities in the ELN and the depending information fields to information describing folding, conformational and hydrolytic stability, and aggregation properties. Furthermore, biomolecules without sequence information will be stored in the ELN together with relevant details for their identification and characterisation as well as homology models and molecular dynamics simulations. Important database IDs and identifiers in particular of the NCBI as well as links to the relevant databases (e.g. RCSB PDB, BRENDA, UniProt) will be given and relevant software has to be included or connected (e.g. BLAST).

**Task 2.3.6 Subdiscipline Pharmaceutical/Medicinal Chemistry**

For Pharmaceutical/Medicinal Chemistry (PMC)-related work, entities like molecules or biopolymers have to be stored along with bio-relevant parameters. To this aim, the panel bio-params will be developed: The data model of the ELN has to be extended to medicinal

chemistry metrics such as logP, $pK_a$ and $pK_b$, solubility, aggregation properties, polymorphism, and PAINS relevance from experimental measurements or computational predictions. The ELN will store experimentally determined and predicted or imported reference values of physico-chemical properties. The PMC group will extend the modules available for assays description of the basic version of the ELN to record the most important experimental details and results of biochemical investigations based on chemical entities. The modules will be extended to store and visualize results from investigations that include multiple molecules or targets. The user interface of the ELN will allow a flexible way of visualisation and summary of the obtained results including efficient filter functions and categorisation tools as well as advanced search tools for combined parameters. The ELN will allow entity-pairing of experimental results with chemicals and biological targets or docking/molecular dynamics results of biological targets and small molecules. Standardised export and import functions in accordance with common plate reader formats as well as data exchange formats for bioassay data repositories such as ChEMBL will be provided.

**Task 2.3.7 Subdiscipline Analytical Chemistry**

The Analytical Chemistry (AC) sub-group will focus on functions for the storage, representation and visualisation of analytical information and processes in the ELN. AC plays a pivotal role in all subdisciplines of chemistry and is used to characterise molecules, materials and the success of synthetic procedures. The most widely used analytical techniques include MS, NMR, IR as well as UV spectroscopy. The second most prominent data type is chromatographic information, which can be combined with spectroscopic information. The resulting data belongs to the larger data types in chemistry, with metabolomics datasets being in the order of 30 GB on average and up to 1.9 TB. This task will enable the ELN to handle complex analytical data types (such as 2D NMR or hyphenation data) to be read by embedded data viewers and processors (see T2.5.2) and to be navigated and visualised in the ELN UI. Metadata provided with the analytical data will be extracted and stored in standardised machine readable format (see TA4). A close collaboration with the other subdisciplines is required to enable the availability of relevant data and metadata as well as suitable presentations for results of spectra analysis in OC, IC, PC, PolyC, BC and PMC. The available analytical data and metadata will be processed to suitable standardised data exchange formats to be transferred to the relevant data repositories of NFDI4Chem (MassBank, nmrshiftdb2, VibSpecDB, Chemotion) as well as international repositories or databases (MetaboLights).

**Deliverables**: (D2.3.1) Five releases of ELN code including subdiscipline specific changes on GitHub, (D2.3.2) Documentation of developments is provided via Knowledge Base, (D2.3.3) Documentation of features are provided on a GitHub wiki.

**Measure 2.4: ELN as a service and distribution of updates**

**Goals**:

1. Ensure the availability of the NFDI4Chem supported ELN in production mode for all scientists in Germany.
2. Provide access to other open source ELNs ensuring the best support and advice for the scientists.

**Description**: Many users, working groups, or institutions will want to host the ELN under their own control and responsibility. Others prefer an out-of-the box solution with minimal cost of ownership and maintenance effort. For those, and to enable a low entrance hurdle for as many users as possible, we will provide and operate an ELN as a service as part of the infrastructure services of the NFDI4Chem. For this ELN as a service, the update and upgrade of all instances will be organized and recorded centrally. This will avoid compatibility and dependency issues after a certain runtime. Users will be liberated from the identification and management of new updates. The access to the centrally hosted ELN managed in a transparent, non-bureaucratic way, ensuring that the ELN as a service is used in accordance with the NFDI4Chem operation model. Other open source ELNs will be installed as a centrally hosted ELN demo version to allow a use for testing and teaching purposes.

**Task 2.4.1 Distribution of updates**

The source code and its documentation will be available on code repositories such as Github for the dissemination of the source code to developers. The offer of updates and upgrades for instances of the ELN running decentralized at German universities will be managed. Latest bug-fixes and additional developments will be merged to the base code and a systematic integration with the deployed instances will be managed. The versions, dependencies and changes have to be disseminated based on a CI model via Github and via additional documentation on the Knowledge Base (see T5.3.1).

**Task 2.4.2 ELN as a Service - centrally hosted instance**

KIT-SCC will install a centrally hosted ELN as a service. The task includes the definition of a general applicable procedure for an access model implemented with the AAI infrastructure (see cross-cutting topic M6.3.2). The host institution KIT-SCC will care for the recent updates and upgrades of the software which in that case cover the developments of all subdomains. In addition, a user model will be defined allowing the use of computing resources. As the benefit of using the ELN increases with better external resources and web-services as well as additional algorithms embedded, KIT-SCC will manage the implementation of features like searching in external databases SciFinder or Reaxys. The constant access to the ELN as a service is overseen by a support team at the host institution, providing their best effort to solve user problems and mediate bug fixes. High availability of the ELN service is ensured by a fail-safe hosting using a professional redundant server architecture and virtualisation.

**Task 2.4.3 Service for open source ELNs**

KIT will install demo instances for other open source ELNs than the NFDI4Chem supported one. The ELNs can be tested for their suitability for researchers that are looking for alternative solutions or want to discover features to be implemented to the NFDI4Chem ELN without the obstacle to set up these ELN on their own. The choice of open source ELNs to be integrated will be made after the next user survey. Already identified ELNs are OpenEnventory, eLAbFTW, and SciNote. The teaching and training team will include the open source ELNs to their documentation portfolio.

**Deliverables**: (D2.3.1) Distribution model for ELN updates and portal report on latest updates and new instances, (D2.3.2) Access model for users and release of terms of use, (D2.3.3) ELN as a service starts registration process for users, (D2.3.4) Agreements with external databases and license holders, (D2.3.5) Demo instance for diverse chemistry relevant open source ELNs to be tested by the community.

**Measure 2.5: Viewers, processors and editors for structures and data**

**Goals**:

1. Improve available open source data viewers, processors, and structure editors be embedded to the NFDI4Chem infrastructure.
2. Enable independence from commercial or closed source software.
3. A modular design enables the use as stand alone, client-side application as well as its incorporation into repositories and the ELN.

**Description**: Viewers and editors for data and structures are a fundamental part of the Smart Lab concept (ELN) *and* chemistry repositories. Wherever possible, NFDI4Chem aims to support scientists with free and open source tools to work with and re-use research data. Therefore, viewers and editors may also serve as a valuable contribution to NFDI4Chem as stand-alone version offered by a web-service. An easy-to-use structure editor for the input of chemical structures is the prerequisite for the success of NFDI4Chem as the generation of correct and machine-readable structures are the basis of the chemists' work with a digital environment. Ketcher (Kotov et al. 2018) is selected amongst other open source editors due to its high level of functionality. The additionally necessary functions to be integrated will be developed with experts of the different subdomains (OC, IC, PolyC) to ensure a correct subdomain specific representation of the molecules and other chemical entities. Besides chemical structures, specific data like NMR or MS spectra require the support by modern tools to process, visualize, and analyse the given information. Therefore, a non-proprietary viewer-processor software, open to a community-driven applicability extension, is a cornerstone for successful data management. The components of existing data viewer-processors will be combined to a modern web-based universal data viewer for the web services offered by NFDI4Chem. IDNMR (DFG funded: SCHL 580/3) will provide a comprehensive tool for the processing, analysis and electronic assignment of 1D and 2D NMR spectra which converts extracted information into a machine-readable and archivable format (Pupier et al. 2018). ChemSpectra (Huang 2019), a software to visualize and analyze spectroscopic data (NMR, IR, mass), based on the previously developed software tools NMRglue and SciPy, as well as SpeckTackle, an established

software e.g. for the visualisation of mass spectra, will amend the IDNMR viewer by additional tools and workflows to include other spectra types. An improved, multipanel UI, being able to deal with different analyses types will enable a flexible integration of the requirements from different subdomains. All components improved in this measure will be of equal importance for all infrastructure services, in particular for the repositories and databases described in TA3 which have similar requirements for the search, visualisation and analysis of their stored research data.

**Task 2.5.1 Rework of Structure Editor Ketcher**

**Organic Chemistry**: The Ketcher editor will upgraded. The usability of Ketcher will be improved with regard to the options to draw structures such as actions to move or rotate (sub-) structures. The templates section needs to be reorganized and a structure clean-up has to be added. The editor will be improved with respect to currently missing functions like the implementation of automated recognition of structures with planar chirality and the analysis and correct assignment of stereogenic centres in sugars and other complex molecules. We will improve Ketcher to deal with reactions including the automatic assignment of all molecules to a defined function in the ELN as starting material, reagent, or product and it has to be extended to work with textual information. **Inorganic Chemistry** : With regard to the bonding schemes, classical Lewis structures very often fail in organometallic chemistry. Here, a general consensus must be found in the community for an ideal graphical representation. Correct recognition of relative stereo-information in complexes and the addition of functionality to create complexes including pi- or even delta-bonds to represent allylic or sandwich complexes. **Polymer Chemistry**: Definition of polymeric structures in the editor which are translated correctly to InChI and which allow a suitable presentation in the UI of the ELN. Dependencies of OpenBabel and RDKit/CDK have to be considered.

**Task 2.5.2 Web-based data viewer/editor**

The available components of IDNMR, ChemSpectra and SpeckTackle will be merged. In accordance with the results of the data standards working group (see M4.3), the combined editor will be extended to read the most important proprietary and open spectroscopic data and to support the storage and export of the respective standard formats. Until a final standardisation is achieved, the currently available open formats JCAMP, mzML and NMReData will be actively supported as output format besides other open formats to be defined by the user community (see M5.1). In the long term, the combination of data converters used to generate standard open formats and the data viewer-editors to read, visualize and edit files in open formats, should be achieved. The NFDI4Chem spectra editor will be provided as stand-alone software and will be adapted to the NFDI4Chem supported services ELN and repositories.

**Deliverables**: (D2.5.1) Ketcher Version 3.0 including features of OC, IC, PolyC is provided on GitHub, (D2.5.2) Well-elaborated material/descriptions on how to add structures correctly to allow a correct registration of the generated structures in databases is provided

by the Knowledge Base, (D2.5.3) Combined JS Data (Spectra) viewer is available on GitHub, manuals delivered to the training team (M5.3).

**Measure 2.6: Additional software instruments**

**Goals**:

1. Development of extensions to the most important cheminformatic libraries necessary to support the NFDI4Chem infrastructure efficiently.
2. Define and develop models to curate chemical data in different phases of the data life cycle.

**Description**: To run and modernize the NFDI4Chem infrastructure in the long term, additional software to be used as web-service or implemented to the ELN and repositories will be developed, maintained or extended. Current needs were identified for chemoinformatic libraries and curation or quality control tools: CDK, RDKit and OpenBabel are important cheminformatic libraries/toolkits that are necessary to process chemical data. The community-driven open source projects provide a basic functionality that is used for many applications and services in chemistry. As the community-driven work does not cover all the needs of the progress planned in the NFDI, the cheminformatics toolkits will be improved to allow a comprehensive processing of chemical structures and the retrieval of computed information. Data curation is a basic instrument to maintain the usefulness and re-usability of data collections. While current data collections are basically curated manually due to missing automatic data curation tools, the ratio of automatic data curation should be increased step by step by the development and implementation of reliable mechanisms to relieve the traditional peer reviewing process. Data curation tools will be developed for spectroscopic data files and textual analysis given traditionally in the supporting information or methods part of a publication.

**Task 2.6.1 Contribution to Chemoinformatic libraries**

Development of extensions to OpenBabel, RDKit and CDK covering the required functions to be used within NFDI4Chem. Already identified essential extensions (applicable to at least one or all libraries/tools) are the support of Molfile (MDL) V3000 (converting MDL V3000 between chemical formats), the improvement of "clean-up" features to be able to visualize in particular higher molecular 2D-structures, the option for SVG depiction for reactions and polymers and the development of conversion features to CDX data. The definitions and standards elaborated in polymer chemistry should be incorporated into the chemoinformatic libraries.

**Task 2.6.2 Software for data curation, plausibility, completeness, and quality check**

Based on the already existing software and workflows developed currently in a decentralised manner for different analytical applications, NFDI4Chem will build quality control and curation software that include the available and forthcoming tools of the NFDI4Chem but also of international developments. These include tools for file conversion (see T4.3.2), data processing and visualisation (see T2.5.1 and T2.5.2) as well as

(semi)automated data evaluation. Best practices for data handling and publishing workflows defined in TA4 will be combined with the necessary functionality to analyze the provided information on a textual but also data file level. Already implemented quality control and curation models will be extended to a broader field of application in analytics, referring to an assessment on the FAIRness of data and a check for completeness and plausibility (see M4.5). The assessment, which is only applicable to single areas will be applied to the combined analytical methods allowing the evaluation of research results in a comprehensive manner. The corresponding quality control and curation software will be developed in a modular manner and will be embedded into the NFDI4Chem infrastructure services to support with standardised processing, analysis, evaluation and reviewing of data. Also, the software will be designed to become part of a new electronic publication procedure for scientific journals in chemistry, thus facilitating both, the review process during submission equally for referees and submitters, while ensuring the quality of data sets submitted to repositories (M4.5 and T3.3.5).

**Task 2.6.3 Peer Review Organiser**

In this task, a peer review management framework, a “Peer Review Organiser”, that can be used for the different repositories within the NFDI4Chem consortium, is developed. The framework will manage the choice and assignment of reviewers to entities of the relevant repositories. It will serve as a central service to enable a standardised peer reviewing of datasets across repositories in a professional manner, compatible to peer reviewing for publications in scientific journals. The framework will implement authentication, authorisation and roles based on the NFDI AAI model defined in T6.3.2. The peer reviewing organizer manages the selection of reviewers according to a standardised protocol and assigns the reviewing tasks to those peers being expert in the respective research area. The reviewing organizer will function as connection to the journals in a second step, providing automatically generated reports of the peer review and also the automated plausibility checks. The software will be implemented into selected repositories in M3.3.

**Deliverables**: (D2.6.1) Contribution of selected features to open source GitHub code of RDKit, CDK and OpenBabel, (D2.6.2) Web-Service that offers data curation features as demonstrator, (D2.6.3) Peer Review Organizer is published in peer reviewed journal.

**Risks and Mitigation strategies** - Table 3

Table 3.

Table to Task Area 2: Smart Lab

| Description of risks | Risk-mitigation measures |
|---|---|
| **R1**: The new instruments, workflows and measures are not or only slowly accepted by the community<br>**Likelihood**: Low | Development of the infrastructure according to the users’ needs and preferences, based on user surveys and continuous feedback.<br>**TA(s) involved**: TA5, TA6 |

| | |
|---|---|
| **R2**: The development and international agreement on (new) standards is not fast enough and the NFDI4Chem consortium is missing basic models on how to generate interfaces, report, export, and transfer data<br>**Likelihood**: medium | Solutions to the given risks are described in TA4 and TA6 (Chamanara et al. 2019). NFDI4Chem starts with reasonable models that are not agreed on internationally but can be extended in a flexible manner. Adaptation to international agreements in a second step.<br>**TA(s) involved**: TA4, TA6 |
| **R3**: Inadequate interfaces to related NFDI services could affect the project.<br>**Likelihood**: low | The interfaces will be part of cross-cutting topics and discussions.<br>**TA(s) involved**: TA6 |
| **R4**: Manufacturers of devices do not agree on common open interfaces and open formats to facilitate the data transfer and data readability.<br>**Likelihood**: high | Data transfer from devices is realized via three alternative strategies. If the manufacturers will not collaborate, the NFDI4Chem will focus on the strategies which work without the manufacturers' commitment with respect to data availability (T 2.1.1. and T 2.1.2) and conversion (T4.3.2).<br>**TA(s) involved**: TA2, TA4 |
| **R5**: The implemented data viewer(s) will not cover the whole range of possible data formats of the community: The variety of data formats and also currently missing standards (allowing a limitation of output files) may complicate the process.<br>**Likelihood**: medium | TA2 will focus on the most important file formats for each analysis type and experimental setup (surveys TA5). For uncommon file types and proprietary files, a common strategy towards standardised file formats will be developed in TA4.<br>**TA(s) involved**: TA2, TA3, TA4 |
| **R6**: The installation of the required AAI for operation of the ELN as a service is delayed<br>**Likelihood**: medium | A user access model with limited authorisation functions will be developed to serve as an intermediate solution to be replaced by the NFDI overarching AAI as soon as available.<br>**TA(s) involved**:TA6 |
| **R7**: Security risks and data leakage for the decentrally installed ELNs and ELN as a Service<br>**Likelihood**: low | The ELN software will undergo regular security checks as part of the CI process and penetration tests. KIT-SCC will provide guidelines for a secure runtime environment.<br>ELN as a service is hosted in a professionally operated data centre at KIT-SCC. It benefits from a sophisticated security architecture with firewalls, intrusion detection systems and virus scanners maintained by experienced administrators.<br>**TA(s) involved**: TA2 |
| **R8**: The productive operation of the ELN as a service is not possible because the requirements for the availability of the service are too high.<br>**Likelihood**: low | The operation takes place in a professional data centre with great experience in the provision of redundantly designed services to achieve high reliability. In addition, the operation can be mirrored in another data centre (FIZ) if required.<br>**TA(s) involved**: TA2 |

## Task Area 3: Repositories

### Description and Objectives

Establishing and maintaining an interoperable network of domain-specific FAIR research data repositories and integrated tools (see Fig. 9) for the national research community in Germany is at the heart of NFDI4Chem. NFDI4Chem has a holistic approach to support researchers from the very beginning of data generation to the deposition and publication of data in suitable repositories. Services of TA2 cover the early stage of research data, providing a seamless handover of the research data to the repository services of TA3 in which they can store and manage all their data (raw data in various formats as well as curated datasets). Researchers must

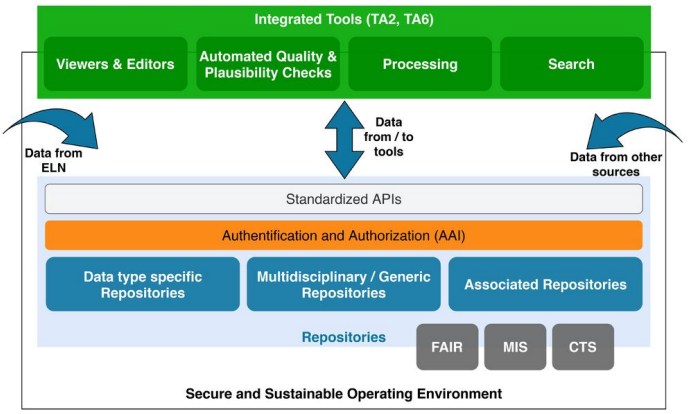


Figure 9.

Data flow between NFDI4Chem components via standardised APIs.

1. be able to quickly and easily identify suitable repositories that support specific data or functionality,
2. be able to efficiently store their data, and
3. easily disclose and re-use relevant datasets across disciplines.

Appropriate tools must support them in adequately describing the research data in their scientific context in order to ensure both their interoperability and reusability, even for scientists of other fields.

Thus, TA3 will establish a **virtual environment of federated repositories** for collecting, storing, processing, analysing, disclosing and re-using research data as part of the NFDI4Chem infrastructure. It will support all relevant repositories in converging into sustainable and interoperable services in reliable operating environments. The federation will be realized by implementing standards regarding metadata and API defined in TA2 and TA4 as well as vocabularies and ontologies addressed in TA6. This allows for their deep integration with Smart Lab environments including ELN as described in TA2 and ensures connectivity towards advanced tools and services, e.g. for data comparison, analysis, and visualisation, including tools of other scientific fields to enable interdisciplinary use cases that involve molecule related data. TA3 will foster and strengthen the FAIRness of data in all major fields of chemistry. Based on repeated surveys of requirements conducted in M5.1, lacking functionality and service offerings will be augmented. Repositories, ideally, take into account processes for data quality control and include all necessary functionality to describe and evaluate research data as well as to keep them findable, accessible, interoperable and reusable as well reproducible in the long term. A worldwide overview of chemistry-related databases and repositories can be found at re3data.org (re3data.org 2019). In Germany, the availability and use of chemistry repositories is currently still at a limited level. Most repositories do not yet meet common criteria such as certification or the use of persistent identifiers. At the same time, some of them already play an important role

in chemistry today. The following list of criteria were developed to identify those repositories which can form the nucleus for the envisioned virtual federation:

1. the repository is suitable for the deposition of molecule related data,
2. the repository contains reusable data or functionality that covers the needs of the NFDI4Chem community,
3. the repository software is open source,
4. the operators of the repositories have declared their willingness to adapt their services to the standards developed by NFDI4Chem including the FAIR principles and
5. the repository operators can be funded in accordance with the funding guidelines of the NFDI (main operator is based in Germany and is a non-profit organisation).

Applying these criteria leads to the following list of repositories and data collections (from now on referred to as “selected repositories”): **Chemotion Repository**, **nmrshiftdb2**, **MassBank EU, VibSpecDB**, **Suprabank**, **NOMAD**, and **STRENDA** (see section Research Data Management in Chemistry - a status quo for descriptions). We will raise the selected repositories to a common level of sustainability, trustworthiness and interoperability and subsequently integrate them into the virtual environment of federated repositories. New or currently emerging repositories will be periodically evaluated and, as soon as they have reached a level of maturity, considered for inclusion (see M3.3).

Further databases and data repositories (see section Research Data Management in Chemistry - a status quo) do not fulfill the criteria listed above. Therefore, they are not considered for the following measures which focus on the “selected repositories”. However, we will seek an ongoing exchange on interoperability issues and encourage them to participate in the development of NFDI4Chem standards and interfaces. Two databases ( **CSD**, **ICSD**) have already stated their commitment to do so as in kind contribution (see LoS) and are included as “associated repositories”. All standards and interfaces will be published and can be implemented by other repositories in order to integrate them into the NFDI4Chem Infrastructure. Additionally to the chemistry-specific repositories, we consider the generic/multidisciplinary data repositories **bwDataArchive** and **RADAR** as relevant, because they fulfill three of the four criteria and can play an important role both as catch-all repositories and data archiving services. Additionally, they can offer long-term archival functionality for the other selected repositories where needed.

In order to achieve the goals of this Task Area, we will focus on measures targeted at standards compliance and interoperability of repositories (API and minimal information standards), their integration into the aforementioned virtual environment of federated repositories, adaption of NFDI-wide concepts (such as authentication and authorisation infrastructure, metadata standards) and the operational sustainability and long-term preservation of research data. Additionally, functional improvements based on identified user needs will be addressed.

To summarize, the objectives of TA3 are:

**O3.1**: Researchers can publish and archive their data on a low-threshold and re-usable basis via a network of standards-based and quality assured repositories that are hosted in trusted data centres to ensure long term accessibility of datasets.

**O3.2**: Researchers can ingest, search, annotate and exchange research data or metadata across distributed data sources by means of a virtual environment of federated repositories.

**O3.3**: Support findable, accessible, interoperable, reusable (and therefore reproducible) Open Data in chemistry (FAIR).

TA3 contributes in particular to the key objective 1: *virtual environment of federated repositories*.

**Measure 3.1: Establish repositories as a core part of the NFDI4Chem Infrastructure**

**Goals**:

1. Making selected repositories interoperable and integrating them into the NFDI Infrastructure;
2. easing data deposition, data publication, and metadata exchange by implementing standardised interfaces (APIs).

**Description**: NFDI4Chem will develop standards for metadata (in TA4) and programming interfaces (TA6 together with TA2 and TA3). TA3 will implement these specifications in the selected repositories. To this end, we will determine which parts of the standards are applicable to a specific repository. Subsequently, we will adopt the metadata schemas of the individual repositories and develop relevant API endpoints. We will migrate already existing datasets in the selected repositories regarding their descriptive metadata to comply with the new standard.

**Task 3.1.1 Adapt metadata schemas of each selected repository**

TA3 will evaluate the Minimal Information Standards as defined in TA4 and identify applicable parts for each selected repository based on supported data types and offered functionality. KIT-SCC and FIZ will support all affected repositories in implementing the applicable minimal information standards. Further, TA3 will support the terminology service that is created in M6.1 to ensure the semantic enrichment of data in the selected repositories. In order to improve the interoperability with international data repositories, KIT-SCC and FIZ, together with the repository operators, will implement the OpenAIRE Guidelines for Data Archives. This includes the support of the metadata schema specified by OpenAIRE and the provision of information via an OAI provider. TA3 will validate the results in order to maximize compliance with the standards, thus ensuring a high degree of interoperability.

**Task 3.1.2 Design and implement Application Programming Interfaces (API)**

The idea of Smart Labs requires the in-depth integration of laboratory equipment, ELNs and repositories. Especially the seamless data exchange between repositories, ELNs and databases needs well defined interfaces, which must be provided by all involved components. In close cooperation with TA2, TA4 and TA6, we will specify application programming interfaces (API) that allow a standardised metadata exchange between repositories, ELN and (international) databases. As stated in the Berlin Declaration (Glöckner 2019), this task will also cooperate with other NFDI consortia via communication channels established in TA6 to identify cross-cutting aspects for repository API, thus increasing the interoperability across disciplinary boundaries. At the same time, this will strengthen the interoperability with important databases and repositories on an international level. We will evaluate existing protocol standards like the lightweight protocol for depositing content (SWORD) (SWORD 2019) and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Lagoze and Van de Sompel 2003) for their suitability with regard to the requirements of NFDI4Chem.

KIT-SCC and FIZ, together with the operators of the selected repositories, will implement the necessary changes. In close cooperation with M4.3, KIT-SCC and FIZ will set up an (automated) test suite in order to help repository operators in validating their API endpoints against the API specifications.

**Task 3.1.3 Integration with Authentication and Authorisation Infrastructure (AAI)**

The Authentication and Authorisation Infrastructure (AAI) solution that will be set up as a cross cutting topic for all NFDI services has to be integrated with the NFDI4Chem infrastructure (see T6.3.2). The roles and rights that are necessary to operate the federated repositories need to be designed and implemented with respect to the AAI system. This includes access rules for users, administrators, reviewers and data providers and has to take into account hierarchies within organisations and working groups and projects. KIT-SCC and FIZ, together with the operators, will ensure the full integration of all selected repositories into the NFDI-wide AAI.

**Task 3.1.4 Setting up the federation of interoperable NFDI4Chem services**

After the selected repositories have implemented the necessary technical requirements for integration into the NFDI4Chem infrastructure (Minimal Information Standards, metadata schemas (see T3.1.1), API (see T3.1.2) and AAI (see T3.1.3)), the actual integration will take place. KIT-SCC and FIZ will test the connection of repositories with ELN and the metadata search via the standardised API in close cooperation with TA2 and TA6. These steps will include practical tests of the interaction of the individual services (functional tests, integration tests, load tests, penetration tests). The integration of the services will thus establish the technical basis of the NFDI4Chem infrastructure.

**Task 3.1.5 Migrate all data with respect to the new metadata schema**

The implementation of the minimal information (see M4.1) and data format (see M4.2) standards in the selected repositories will require adaptations of the descriptive metadata and formats of existing datasets. In this task, the operators of the repositories will be

supported in transferring their content to the new standards, thus making them interoperable (in cooperation with T4.4.1). The repository operators will then perform quality assurance measures, adjust the ingest processes and update the documentation accordingly.

**Deliverables**: (D3.1.1) Operational virtual environment of federated repositories, (D3.1.2) Completed data migration of all relevant repositories

**Measure 3.2: Ensure sustainable repository operation within the NFDI**

**Goals**:

1. Prepare all selected repositories for secure and sustainable operation,
2. guarantee reliable data publication, and
3. ensure long term data archiving and accessibility.

**Description**: A central task of the NFDI is the reliable provision of services and the sustainable storage of and access to research data. This applies in particular to the selected repositories, which are the result of concrete needs of the chemical community. In this measure, we will enable all selected repositories for sustainable operation (T3.2.1 - T3.2.4). Additionally, we will rescue relevant orphaned datasets by importing them into reliable repositories (T3.2.5).

**Task 3.2.1 Analyse operational fitness**

In this task, we will determine the maturity level of the selected repositories with regard to secure, reliable and sustainable operation based on the ISO 16290:2013 standard (Technology Readiness Level). Desired characteristics include a modularised software architecture with well-defined interfaces to storage and long-term archival systems, fail-safe hosting, use of standard open source components and libraries that are maintained by a broad community, well described operational processes, good user interfaces and documentation, as well as adequate file and container formats following guidelines developed in TA 4. Furthermore, TA 3 will analyse the security of the software with static code analysis and penetration tests. KIT-SCC and FIZ will conduct the assessment of each selected repository and identify necessary optimisations.

**Task 3.2.2 Enable hosted operation**

The aim of this task is to enable efficient and professional operation of all selected repositories. Many repositories have grown out of scientific initiatives in the community, which is why they are often operated by the researchers themselves. To be able to offer the services in NFDI4Chem efficiently and resiliently and to relieve the burden on researchers, TA 3 will make the relocation of operations to data centres possible, thus ensuring features such as backups, redundancy and fast network connectivity. At the same time, these changes will enable to easily replicate repositories in other data centres or to operate them in cloud infrastructures. Based on the analysis produced in T3.2.1, TA3 will assist repository operators in optimising the software architecture and technical properties

of their services with regard to the operation in hosted environments. This includes actions such as containerisation, the replacement of proprietary technologies by standard open source components and the implementation of interfaces to data-centre-grade storage and archival systems. KIT-SCC and FIZ will provide experienced research software engineers for this task. Additionally, KIT-SCC and FIZ provide long-term archiving services for repositories which lack a digital preservation strategy (bwDataArchive, RADAR).

**Task 3.2.3 CoreTrustSeal certification**

The CoreTrustSeal (CTS) certification of research data repositories is regarded as a strategic step, aiming to increase trustworthiness and transparency for users who need to decide on suitable research data infrastructures for their data deposit. Thus, FIZ and KIT-SCC will support repository operators in certifying their services by offering individual advice, by helping with the CTS application and by providing services (e.g. hosting including long-term archival, see M3.4) to improve their quality.

**Task 3.2.4 Data rescue and exit strategy**

In this task, KIT-SCC and FIZ will take care of data collections or repositories whose contents are regarded as important for the community and whose further maintenance is no longer guaranteed by the previous operators. In these cases, we will try to identify a suitable repository (otherwise RADAR as a multi-disciplinary repository solution) and to migrate the data at risk to it. This involves mainly manual actions such as converting metadata and data, adopting to container formats for long-term archiving, as well as moving DOIs for published datasets. This task also serves as an "exit strategy" for the relevant repositories in case the further operation of one of them is not feasible any longer.

**Deliverables**: (D3.2.1) Operational Fitness Report, (D3.2.2) Software releases of selected repositories with improved software architecture, (D3.2.3) CTS certification for selected repositories

**Measure 3.3: Create added value for researchers by improving repository functionality**

**Goals**: Covering all disciplines or data types in chemistry with adequate publication and archiving facilities; implement functional improvement and extensions for supported repositories.

**Description**: In this measure, we will carry out a gap analysis of the existing relevant repositories in order to identify voids in the coverage regarding data types or disciplines and to close these gaps through adjustments or, if necessary, new developments. We will evaluate existing repositories and active projects in order to identify those with interesting data or functions and to integrate them into the NFDI4Chem infrastructure, assuming a certain degree of maturity and operational fitness.

**Task 3.3.1 Gap analysis and implementation strategies**

In this task, we will generate an overview of the functional requirements, analytical methods and associated specific data types of the subdisciplines represented in NFDI4Chem and compare them with the existing functions of the relevant repositories. At the same time, we will evaluate the user-friendliness of the relevant repositories. The resulting compilation of identified deficits will be prioritised. The result is a list of recommendations for action for existing repositories or the development of new ones. In the case of smaller optimisations of high priority, KIT-SCC and FIZ will support repository operators either with experienced research software engineers or with applying for NFDI4Chem centrally managed funds (see M1.4). In case of larger developments, KIT-SCC and FIZ will support the repository operators in acquiring appropriate third-party funding.

As an example, we have already identified a gap with respect to the availability of repository functions for reactions and processes in inorganic chemistry, polymer chemistry, and physical chemistry. The adaptations that are made in the ELN development T2.3 by different working groups will be integrated to the chemotion-repository data model to offer a basic coverage of the available information by an extended repository functionality. Another example is the joint storage of experimental and theoretical data, such as measured and simulated Raman or UV spectra. Until now, such data are only given as graphs in supplemental information of publications. The joint storage with joint metadata annotation will ensure that the experimental and simulated data will be found together which facilitates reuse. In numerous cases, experimental data can only be understood in combination with theoretical peak assignment.

TA3 supports interested parties in defining NFDI4Chem-compliant specifications and formulating project ideas. The Innovation Incubator (see T5.5.4) assists in allocating appropriate funding for the implementation. KIT-SCC and FIZ integrate resulting new components and services into the NFDI4Chem infrastructure.

**Task 3.3.2: Integration of data processing tools**

With the expected growth of data, calculation and analysis tools will be increasingly important as they allow the generation of added value for scientists. To enable and automate computational data processing and analysis, legacy data has to be normalised or common data formats have to be used. This can be achieved by using chemoinformatic libraries which handle structural data in a machine-readable way. Data converters have to be employed to transform data into standard formats (see M4.3).

In this task, the relevant repositories will integrate necessary chemoinformatic libraries and data converters. Examples have been shown by the chemotion project, where molecule data was used directly to calculate theoretical characteristics of desired synthesis products using Turbomole calculations. Another application can be shown by the implementation of established machine learning models (see T5.5.2) which support the scientists in their scientific work (e.g reaction prediction and retrosynthesis, or small molecule identification from MS data).

**Task 3.3.3 Integration of data viewers and structure editors**

Data viewers and structure editors allow the entry and view/analysis of chemistry specific entities. Depending on the need of the single repository, the choice of the distinct software component differs (for spectra viewers/processors, ChemSpectra, IDNMR-viewer and Spectackle are used for Chemotion, nmrshiftdb2, and MassBank EU). With Luc Patiny from EPFL, MassBank EU is currently developing a new viewer system for easier handling of repository data and to be able to compare to to external data. This task deals with the integration of data viewers and structure editors to selected repositories. As soon as the improved structure editor (in year 3) and the combined data viewer (in year 4) developed in M2.5 are available, the repository operators will integrate these software components into the selected repositories.

**Task 3.3.4 Versioning of datasets**

Data modifications, such as fixing errors in data and metadata, adding additional information or data elements to a data collection, or jointly preparing a dataset for publication, require that changes are tracked to ensure traceability. All versions of published datasets need to be accessible and linked, so that references via persistent identifiers like DOIs remain valid. Throughout the curation process (see T4.5.2), editorial staff from the repository operator as well as automated processes for data validation and quality checks (see T3.3.5) might trigger modifications of metadata which needs to be tracked as well and may lead to new versions of a dataset. Especially with peer reviewers or contributors from the community, the tracking of changes and the possibility to roll back modifications are important.

Together with the repository operators, TA3 will design a functional concept for the introduction of versioning. The operators will then technically implement the concept in the respective repositories. KIT-SCC and FIZ will provide experienced research software engineers to support the repository operators with this task.

**Task 3.3.5 Integration of automated plausibility and quality checks**

Quality assurance is a comprehensive process along the life cycle of research data. This process can already be supported by automatic checks during data input. At the time of data curation, automatic plausibility checks and data validations can support the reviewers in the (peer) review process. These functions allow a fast screening of the submitted data, giving an evaluation that can be used for a tagging of data prior to manual peer reviewing. The checks will validate all repository specific data with respect to completeness, the availability of data in open standard formats, and its quality concerning different aspects such as the presence of relevant signals in spectroscopic data. TA2 will develop tools for automated plausibility and quality checks (see T2.6.2). These tools will be based on the outcome of the curation criteria and workflows designed in TA 4 (see T4.5.2). In TA 3, repository operators will integrate the tools into their services where applicable.

**Task 3.3.6 Integration of Peer Review Organiser**

In this task, we will integrate the Peer Review Organiser developed in T2.6.3 into selected repositories by the respective operators. The Peer Review Organiser will allow for a standardised and efficient review of datasets, either by appointed reviewers or by a community-driven process. As a prerequisite for the successful integration of the Peer Review Organiser (see T2.6.3) into the individual repositories, processes and data access protocols have to be defined within the repositories. This will be done in T3.1.2. Additionally, viewers have to be made available (see T3.3.3). KIT-SCC and FIZ will support the repository operators with research software engineers in integrating the Peer Review Organiser into their services (in year 5).

**Deliverables**: (D3.3.1) Gap analysis report for selected repositories, (D3.3.2) software release of selected repositories with integrated processing and viewer functionality, (D3.3.3) software release for selected repository with added peer review functionality

**Measure 3.4: Productive Operation**

**Goals**:

1. Transfer selected repositories to data centres to
2. ensure the reliable, trusted and secure operation as sustainable services,
3. set up contracts that define roles and service level agreements for hosting, data publication, long term data archiving and accessibility.

**Description**: The secure, performant and reliable operation of the selected repositories requires a professionally operated runtime environment. Several partners of NFDI4Chem (e.g. KIT-SCC, FIZ, TUDr) provide such environments. Selected functional components (e.g. long-term archival) or complete repositories can be replicated or moved to such hosted environments in order to increase the service quality and data security while the scientific and editorial responsibility for the repositories will remain with the current providers. All NFDI4Chem hosted services will be published on the central NFDI4Chem portal (see T1.3.1).

**Task 3.4.1 Contractual arrangements**

Operation in computer centres requires contractual regulation of various legal topics. These include Service Level Agreements (SLA) and hosting contracts, where liability, data protection (GDPR) and the necessary granting of rights (licenses) must be clarified. In this task, KIT-SCC and FIZ will draft corresponding standard contracts and negotiate them with the parties concerned.

**Task 3.4.2 Service migration**

In this task, we will support repositories that have so far been running in non-optimal operational environments in migrating to professional data centres. First, KIT-SCC and FIZ will analyse the service requirements for each affected repository. This will be followed by the provision of a suitable runtime environment, integration of the transferred services into existing processes for data backup and security (backups/snapshots, firewalls, intrusion

detection systems) and finally optimisations such as virtualisation, parallelisation and redundant service configuration. This task builds on the outcome of T3.2.2.

**Task 3.4.3 Operations**

This task comprises the ongoing operation of the selected repositories by their operators and the involved data centres. This will require continuous service monitoring, the regular installation of updates to the repository software, but also the underlying components such as the operating system and virtualisation software, the elimination of technical faults and the implementation of standard operational routines (e.g. backups). In addition, the repository operators will assure the ongoing ingest and curation of new datasets in accordance with the agreed quality and curation criteria.

**Task 3.4.4 Providing second level service helpdesk**

We will establish a distributed second level helpdesk for the selected repositories (including hosting and digital preservation) regarding support in technical and content questions. A central unit set up in T5.3.4 will receive enquiries via a ticket system and forwards them to the second level helpdesk according to the thematic focus (service interruption, bug report, subject matter enquiry).

**Deliverables**: (D3.4.1) Templates for Service Level Agreements and hosting contracts, (D3.4.2) Performed migration with improved sustainability and service quality of selected repositories

**Risks and Risk Mitigation** - Table 4

Table 4.

Table to Task Area 3: Repositories

| Description of risks | Risk-mitigation measures |
|---|---|
| **R1**: The finalisation of community standards regarding minimal information standards takes too long. **Likelihood**: Low | TA3 provides for an iterative approach in which also partial results can be implemented. **TA(s) involved**: TA3, TA2, TA4 |
| **R2**: The implementation of API and minimal information standards are impeded by missing resources of the repository operators. **Likelihood**: medium | In addition to the personnel resources earmarked for this purpose for the repository operators, KIT-SCC and FIZ provide additional development capacities. **TA(s) involved**: TA3 |
| **R3**: Compliance with CTS certification requirements is proving difficult for repositories. **Likelihood**: medium | TA3 supports the repositories with digital long-term archiving services and the provision of suitable hosting environments, in particular to address aspects of sustainable and reliable operation. **TA(s) involved**: TA3 |
| **R4**: Important datasets or databases in the field of chemistry are no longer maintained by their current hosts. **Likelihood**: high | TA3 has earmarked resources for the transfer of data sets and databases to appropriate repositories in order to preserve them for research. **TA(s) involved**: TA1, TA5 |

| | |
|---|---|
| **R5**: Security risks and leakage of unpublished data<br>**Likelihood**: medium | TA3 will assess the technical architecture of all selected repositories (cf. task 3.2.1) and support repository operators to improve the operational fitness of their services, including security aspects. Additionally, repositories are upgraded to be easily transferable to more secure hosting environments where necessary.<br>**TA(s) involved**: TA3 |

## Task Area 4: Metadata, Data Standards and Publication Standards

### Description and Objectives

TA4 creates and maintains the specification and documentation of standards required for archival, exchange and reuse of data and metadata on characterisation of molecules and reactions, together with reference implementations and data validation.

Our standardisation efforts will integrate with international standardisation efforts, such as IUPAC or domain-specific efforts, e.g. Metabolomics Standards Initiative (MSI), thus contributing to our key objective 2: *Minimum information (MI) standards for data and machine-readable metadata*. By coordinating all standards developments and support in one TA, we ensure that they result in a set of modular, coherent and interoperable standards across the NFDI4Chem consortium.

An essential component for the implementation of the FAIR principles is the use of Persistent Identifiers (PIDs). PIDs unambiguously identify all scientific output. The obligatory and standardised metadata connected with PIDs make research data findable, accessible and citable. The interoperability is supported through standard vocabularies and links to other PIDs in the metadata record of a PID.

For the specification and agreement of minimum reporting standards we will follow successfully established processes, such as the *Minimum Information for Biological and Biomedical Investigations* - MIBBI (Taylor et al. 2008), and work by the applicants on spectroscopic data types (mzML, nmrML, JCAMP and NMReDATA) and reporting standards for (bio)chemical reactions.

By adopting and using generic metadata we will allow to record general provenance of data. Especially for cross-domain applications data needs to be unambiguously semantically annotated, both for humans and machines. Ontologies are used as an integral aspect of the standards where possible, and missing terminological artifacts will be created, and available through the terminology services (see M6.1). Using discipline-specific terminology we will describe research data in machine-readable form and semantics like properties, methods and units without the ambiguity that comes with free-text description. Through reference implementations we will ensure that data are machine-readable, and data pass validation.

Extensive documentation about using the standards, the available tool sets and getting-started tutorials will be part of the M5.3 training activities and available from the NFDI4Chem portal. This TA will also contribute information about the standards itself to

repositories of standards, such as the fairsharing.org, a curated resource on data and metadata standards, inter-related to databases and data policies, to make the standards themselves FAIR.

Finally, we will jump-start the adoption and lead-by-example, as well as working with publishers to embed standards-compliant RDM into the scholarly publication process.

The work in this TA will be in conjunction with TA2 and TA3 where the standards will be used, TA6 where cross-cutting aspects of the infrastructure are developed and TA5 to ensure community training and user acceptance, which supports our key objective 4: *Engage with the chemistry community in Germany*. We will also work together with other NFDI consortia, especially to harmonise metadata capture to support the key objective 5: *Explore synergies with other consortia and promote cross-cutting development within the NFDI*.

To summarize, the objectives of TA4, designed to support FAIRness in NFDI4Chem, are:

**O4.1**: Provide a set of modular, coherent and interoperable metadata and data standards in key areas of chemistry through national and international processes

**O4.2**: Ubiquitous use of (persistent) identifiers for data, instrumentation, protocols and terminology, all with informative metadata

**O4.3**: Increase adoption and integration of standards through generic reference software implementations and integration with scholarly publication processes

**O4.4**: Contribute standards and identifiers on (bio-)chemical entities to entire NFDI

These objectives will be pursued through the following measures:

**Measure 4.1: Development of Minimum Information metadata standards for Chemical Investigations**

**Goals**: Community accepted consensus about (minimum) information that has to be reported about a chemical investigation, the data and metadata that has to be deposited in a repositorium.

**Description**: Sufficient and harmonised metadata is fundamental to put the chemical information data records into context, to make it findable in our search service (M6.2), and interpretable. The descriptive domain-independent and domain-dependent metadata must be in a human and machine readable format to facilitate cross-linking for value added services. To encompass minimum information for synthetic, analytical and theoretical work, the minimum information requirements will be collected from the subdisciplines (in collaboration with M5.1) for all relevant chemical methods. By workshop series, an accepted consensus for Minimum Information about a Chemical Investigation (MIChI) will be reached and discussed with the stakeholders (researchers, infrastructures and journals, in collaboration with M5.6) and participants of TA2 and TA3. Especially in the case of the

plethora of spectroscopic methods, minimum information about spectroscopic methods have to be elaborated. Here is also a close link to the use case collection in T5.5.1.

**Task 4.1.1 Domain-independent metadata**

This task will provide the specification of generic information about datasets. Generic metadata are widely used in research communities of all disciplines. Their use facilitates easier interoperability between systems and disciplines, enables standard-based quality management and in this way provides reliability for re-use of data. We will survey and select from existing standard models such as DataCite and make a selection for the infrastructures in TA2 and TA3. The DataCite metadata schema is mandatory for all institutions registering DOI via the German registration agencies. The schema is a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes. The metadata schema supports openness and extensibility by collaborating with the Dublin Core Metadata Initiative (DCMI), Science and Metadata Community (SAM) to maintain a Dublin Core Application Profile for the schema. While DataCite's metadata schema has been expanded with each new version, it is, nevertheless, intended to be generic to the broadest range of research datasets, rather than customized to the needs of any particular discipline. However it offers a reliable method to link the community specific metadata with the dataset metadata. Extensions of generic metadata standards for NFDI4Chem purposes should be integrated according to shared methods, research objects, software or devices which demands close coordination between subject field experts and infrastructure providers in order to balance the requirements of low-threshold, practicable metadata standards on the one hand and added value through domain specific minimum information standards on the other hand. Data container formats such as the RDA recommended BagIt specification will be evaluated as standard archival recommendation.

**Task 4.1.2 Domain-specific minimum information requirements**

In this task, we will develop and maintain specifications for minimum information (MI) standards about specific subdomains in chemistry. Subject-specific metadata are tailored to the special features of the data and standards used and provide very specific information beyond the generic schema. It is therefore essential for consistent and high-quality metadata that the consortium provides specifications and guidance for the integration of discipline-specific metadata in a generic schema. Some disciplines have a very active community and are already very advanced, examples include the information on enzymology with the STRENDA committee and STRENDA-DB and minimum information requirements for enzyme-catalyzed reactions.

**Task 4.1.3 Workshop series on minimum information requirements**

Domain-specific MI standards will be developed in a series of 5 *international* workshops in 5 subdisciplines of chemistry (e.g. organic, inorganic, physical/theoretical chemistry, polymer chemistry and biochemistry/pharmaceutical chemistry) facilitated and organised by NFDI4Chem. Each of the developed standards will be a **modular recommendation** for

Minimum Information on Chemical Investigations due to the sheer size and variability of the field. Prior to publication of the standards, there will be a Request for Comments (RFC) as part of a public review of the standards. The reasoning for this strategy is that NFDI4Chem can not single-handedly facilitate the standard development for chemistry as a whole but can demonstrate the feasibility and strategy for the procedures in different levels of maturity. The task area leads have extensive experience in MI standard development at the boundary between chemistry and biology.

**Deliverables**: (D4.1.1) Report of Workshops on MIChI, (D4.1.2): Guidance document on MIChI, (D4.1.3) Process documentation for standards development. Series of peer-reviewed publications on MI standards for 5 selected domains of chemistry.

**Measure 4.2: Development and maintenance of standards for data exchange & archival**

**Goals**: Develop and specify standards for specific data types in chemical research

**Description**: Vendor-independent data exchange and long-term archival require open formats for the actual research data, raw instrument output, as well as molecule and reaction standards. Existing data formats will be examined from a broad perspective addressing the needs of a heterogeneous user community with special emphasis on research data (from both experimental and computational sources). The key goal is the definition of open-format minimal common denominators for related data types/groups for a broad user base to maximize general acceptance and which allow for a systematic extension/differentiation according to specific sub-community needs.

**Task 4.2.1 Molecule and chemical reaction standards**

Work with experimental chemists, software developers and repositories on specification of data standards for chemical data on molecules and chemical reactions, such as Chemical Markup Language (CML), representation of reactions and molecule characterisation for data exchange between e.g. the ELN and repositories (Chemotion) and intermolecular interactions (e.g. in SupraBank). This includes enzyme-catalyzed reactions and describing experimental results (time course of starting materials, intermediates, or products) and modelling results (kinetic parameters). An important aspect will be the flexible data model to link, via PIDs, between different standards, e.g. annotation of a subset of atoms or bonds.

**Task 4.2.2 Analytical data standards**

Work with software developers, instrumentation companies, publishers, IUPAC, and repository operators on specification of data standards for spectral data such as NMReData, nmrML, mzML and JCAMP. This includes work on raw and metadata extraction, ensuring unique and persistent instrument identifiers. Data sources include measurement devices such as NMR, mass, X-ray diffraction, UV/Vis, IR, Raman, EPR and fluorescence spectroscopy.

**Deliverables**: (D4.2.1) Specification document of molecule and chemical reactions standards (D4.2.2) Specification document of analytical data standards

**Measure 4.3: Implementation and support of software components for creation, validation and consumption of standardised data formats**

**Goals**: Develop and implement software support for standards in M4.1 and M4.2.

**Description**: An indispensable aspect of standards development is the availability of reference implementations and software to validate compliance of data. If possible, reader and writer software should be available for important programming languages (e.g. Python, Java) and knowledge representation (e.g. Resource Description Framework, RDF) that are commonly employed by the user community.

**Task 4.3.1 Reference implementation of standards reader / writer / validation libraries**

We will make sure that reference implementations are available for all standards developed and used in NFDI4Chem. We will either make existing implementations available as open source and improve them as appropriate or newly develop them as open source software. Particular attention will be paid to validate data files with regard to their specification. Where possible, additional semantic validation (like in many of PSI-developed standards) will be performed. The integration of these data handling libraries into community developed data processing and analysis software will be supported through dedicated Hackathon workshops. Larger integration efforts by researchers will be handled through centrally managed funds for future project extensions in NFDI4Chem. The methods will be implemented in a way that enables them to be integrated into the repositories in a seamless way.

**Task 4.3.2 Reference implementation of raw data converters**

In tight collaboration with M2.2 and M 3.3, we will implement converters between instrument raw data and other data sources into the standard formats adopted and developed in M4.1 and M4.2. The converters are essential fundamental building blocks in NFDI4Chem as they foster the integration of the ELN with a wide variety of laboratory environments. Additionally, file conversion to standard formats allows the application of e.g. spectra viewers and processors to a large number of converted files. Examples include a generic data converter from diverse mass spectroscopy file formats (e.g. RAW) to mzML, or XWinNMR to NMReData and JCAMP.

**Deliverables**: (D4.3.1) Published libraries on code hosting repository, (D4.3.2) Published software on code hosting repository

**Measure 4.4: Lead-by-Example: Creation and deposition of substantial standards compliant data sets**

**Goals**: Create/convert and deposit a substantial number of compliant data sets.

**Description**: The development of standards encompasses the guidelines and specifications, reference implementations and already some (usually reduced) example instances of data encoded in a standard-compliant manner.

This measure aims to provide a large body of real data encoded in a standard-compliant manner through the consortium members. Together with the early adopters of the standards, we will make extra efforts to e.g. publish supplemental data beyond today's reporting guidelines of the journals, and make use of the TA2 Smart Lab components and databases and repositories developed in TA3. In addition to the data sets itself, this effort will document the process to FAIRify the research data, surface practical issues and suggestions for improvements.

**Task 4.4.1 Select, prepare, maintain, adapt and deposit data sets**

Herein, representative as well as substantially complex real data sets will be identified for all subdisciplines. In the beginning of the project, the seamless workflow shown in Fig. 5 will not exist yet, and some manual processing will be required. This in turn feeds back into developments in TA2, TA3 and elsewhere in TA4 to improve systems and processes.

Until the standards, systems and processes are finalised, the lead-by-example data sets will require regular adaptation and effort. They will be used for testing, validation, stress-tests, but also for educational purposes and community outreach. Together with project partners we will include data sets from organic chemistry, inorganic chemistry, physical chemistry, polymer chemistry, pharmaceutical/medicinal chemistry and analytical Chemistry. A participant will contribute data records of complex environmental samples derived from different compartments (e.g water, sediments, biota) and with different analytical strategies (e.g. multi-layered effect-directed analysis or spatio-temporal collection of samples during environmental surveys). Another participant will contribute data focussing on complex inter- and transdisciplinary multimodal data sets from its diverse instrumental infrastructure. So-called *Datathon* workshops will allow intense discussions and direct feedback on the process.

**Deliverables**: (D4.4.1): 50 Experimental processes and data publications using NFDI4Chem infrastructure

**Measure 4.5: Data standards for quality control, curation and publishing**

**Goals**:

1. Scientific publications having high-quality research data must become the norm, rather than an optional addition. Therefore, processes need to be established for peer review of data sets, including data quality control and review of their FAIRness.
2. Integration into scholarly publication processes.

**Description**: While good scientific practice and author guidelines for publications already mandate for all relevant information supporting a scientific manuscript, the majority of

supplemental information are not FAIR, not complete and highly diverse across the publishing landscape. Reviewers do seldom check the quality of supporting data and even less so its form or standards compliance. Through this measure, we will work with publishers and editors to improve the quantity and quality of research data associated with scientific literature. Curation processes, quality control, and FAIRness assessments constitute an integral part of this measure. This measure also has great synergy potential with other NFDI consortia. This includes processes to encourage, validate, and review deposited and associated data.

**Task 4.5.1 FAIR assessment**

The FAIRness of datasets is not a black-and-white decision. Several metrics for FAIRness have been, and are still being, developed Wilkinson et al. (2018). This task assesses the FAIRness of a range of representative datasets published in the chemistry research literature with respect to the provided description of metadata, protocols and research data and will:

1. identify common anti-patterns with opportunities for improvement and
2. provide examples of best practices for data description and publication. The output of this task also feeds into M5.3 - M5.5 to raise the appreciation of good data publication and the measures where NFDI4Chem eases the process to solve remaining FAIRness gaps.

**Task 4.5.2 Curation and data quality standards**

Data quality control is an essential component of FAIR RDM. A data quality assurance strategies includes the annotation of rich and up-to-date metadata, best practice PID services, license information, policies, quality control methods and support, hence generating trust in the access to and reusability of research data. The curation of data sets can be performed by human curators and reviewers, and also through automated processing pipelines. Automated data validation tools can be used for local, curated repositories at academic institutions, as well as for public databases. An important aspect of such a quality strategy for repositories is the storage of quality control data along with the raw data. This quality control data might be utilized for the computational validation of the data's compliance.

As an example, the implementation of standards derived for curation and data quality control in NMR spectroscopy of small organic molecules can be considered: Here, a work flow for a data evaluation protocol for publications is developed and an example implementation is produced for the IDNMR project. It is intended to facilitate:

- submission of real data,
- review processes and
- data access/useability by scientists after publication while at the same time
- securing a higher standard for published spectroscopy data including the control/ curation by experts in the field.

### Task 4.5.3 Integration with scholarly publication processes

We will work with publishers and editors in our advisory boards (see letters of support from Wiley-VCH, MDPI) to improve the quality of research data associated with scientific literature. This includes the integration of recommendations into journal author guidelines for data repositories, templates, and examples to overcome traditional supplemental information. Where necessary, the guidelines for reviewers and editors also need to be augmented with checks for standards compliance as part of the peer review process. The benefit in form of simplifying meta-studies for e.g. analysis of quality of experimental data and processes and of modelling processes can then be showcased in M5.5.

**Deliverables**: (D4.5.1) Report on FAIRness of data standards and datasets published by the community, and recommendations for integration into the NFDI landscape, (D4.5.2) controlled Terminology on curation and data quality concepts, (D4.5.3) Report on adoption of NFD4Chem facilitated standards by publishers

**Risks and Mitigation strategies** - Table 5

Table 5.

Table to Task Area 4: Metadata, Data Standards and Publication Standards

| Description of management risks | Proposed risk-mitigation measures |
|---|---|
| R4.1: Inability by the community to **agree on minimum reporting standards** due to incompatible assumptions. **Likelihood**: Medium | We aim for a modular set of recommendations for the subdomains, avoiding that one standard has to fit all subdomains. **TA(s) involved**: TA3, TA2, TA4 |
| R4.2: **Unwillingness** to share data required to lead by example. **Likelihood**: low | Consortium is fully dedicated to push FAIR Open Data, and numerous datasets are available. **TA(s) involved**: TA3, TA4, TA5 |
| R4.3: Technical hurdles to decode proprietary vendor formats, licensing issues for proprietary Windows DLLs. **Likelihood**: Medium | Positive experience engaging with mass spectrometry vendors. Training of users to include Open Formats in tender requirements. **TA(s) involved**: TA4, TA6 |
| R4.4: **Objection by publishers** to adopt stringent author guidelines ensuring good scientific practice, especially those with a business model based on APC. **Likelihood**: Medium | We are targeting high-quality journals, where good scientific practice and reproducible research have the highest priority. Other journals will follow the trend. **TA(s) involved**: TA3, TA4, TA5 |

## Task Area 5: Community Involvement and Training

### Description and Objectives

Until now, the chemistry research community is hardly aware of RDM, neither of its requirements nor of its possibilities and opportunities. Not surprisingly, existing RDM tools like ELNs or repositories are unknown for the most part and not included in everyday research workflows.

Therefore, the implementation of a national RDM infrastructure **requires a bold cultural change** (Fig. 10) and community effort that is the paramount objective of this task area.

This requires intense **training** of all participants in chemical research at all levels: university leadership, principal investigators, and graduate students. Undergraduate students must also be trained in RDM as soon as possible during their education to anchor RDM in the awareness of coming generations of chemical researchers.

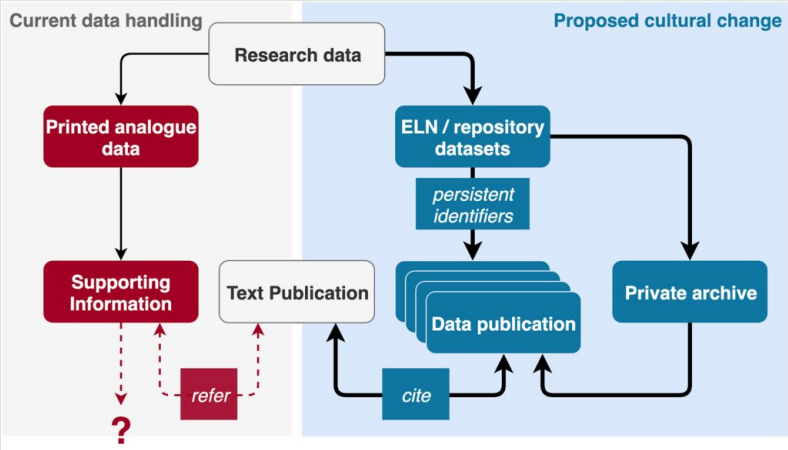


Figure 10.

The cultural change in chemical RDM is supported through training and community involvement in TA5 as well as by the NFDI4Chem infrastructure.

The success of a national RDM infrastructure will strongly depend on specifically listening to the researchers' needs and directly facilitating their everyday research workflows. Since the NFDI4Chem consortium will inevitably be constituted mostly of RDM experts or at least persons with already a strong affinity to digital tools, the specific needs for chemistry research have to be continuously collected and compiled into requirements for the other task areas.

To summarize, the objectives of TA5 are:

**O5.1**: Mediating a cultural change to the community

**O5.2**: Collecting requirements from the community

**O5.3**: Training of the community and the next generation of scientists

TA5 will contribute in particular to key objective 2: *minimum information (MI) standards for data and machine-readable metadata*, key objective 4: *engage with the chemistry community* .

**Measure 5.1: Community Requirements**

**Goals**: Obtaining a detailed overview of the community's RDM needs and their transformation into feasible tasks for the other TAs.

**Description**: Knowing the community's needs in RDM is key to obtain a general acceptance for the tools and standards developed in NFDI4Chem. In this measure, the technical possibilities and the scientists' needs will be synchronised and canalised to the right participants. The prioritisation and technical feasibility of the implementation has to be planned with all participants.

**Task 5.1.1 Continuous collection of requirements and needs from the community**

During the whole project phase, we will continuously assess the chemistry community's needs with regard to RDM. Once every year, we plan to conduct a nationwide online survey. To complete the picture, we will explore more granular formats for interaction with the community like a feedback button in the user interface of ELN (see T2.2.1). These will include an online "suggestion box", regular challenges to write use case stories, as well as dedicated discussion panels on community events (organized in measure 5.6). The surveys will also address the incorporation of other open source ELNs as described in T2.4.3. In addition to these broader community events, we will organize round tables with the local RDM units to collect requirements and challenges in an informal way (every 3 months). In parallel to general chemistry requirements, pharmacy-specific requirements for RDM will be assembled via DPhG and FID Pharmazie. Requirements specific for environmental chemistry will be collected by UFZ in cooperation with the NORMAN Network (NORMAN Network 2019).

**Task 5.1.2 Transformation of requirements into feasible tasks for other TAs**

We will continuously assemble the needs identified in T5.1.1 and break them down into assignments for the NFDI4Chem task areas. The feasibility and details of those tasks will be discussed, prioritised and assigned at the regular steering committee teleconferences. Universities, research institutions, and libraries increasingly have staff that do consulting and/or implementation regarding RDM and its application and services (FDM-Kontakt 2019). We plan to identify and contact such facilities to inform them about NFDI4Chem and establish two-way communication channels. This way, they can incorporate the NFDI4Chem infrastructure when they consult their local chemical researchers. This has the potential to be a significant multiplier to facilitate the adoption of the NFDI4Chem network.

**Deliverables**: (D5.1.1) Requirement analysis and breakdown of tasks to TAs, (D5.1.2) Second requirement analysis, (D5.1.3) Comparison of achieved technical possibilities with requirement analyses.

**Measure 5.2: Awareness**

**Goals**: Raising awareness for the digital change in chemistry.

**Description**: The digital change in chemistry is a long-term process since the scientists start with highly differing starting conditions: as the user survey has shown some institutes already possess RDM policies but a large number of working groups does not even have a group server nor a RDM policy or agreement. In order to encourage the desired

developments, digital workflows, tools, and infrastructures will be constantly brought to scientists' minds by M5.2.

**Task 5.2.1 Portal contents, newsletters, and social media in cooperation with societies**

In this task, we will promote the digital change by different means and on several levels, via the NFDI4Chem portal (see M1.3), links on the homepages of the participating universities/institutes, email newsletters in close cooperation with associations (e. g., GDCh, DPhG, DBG) as well as member journals of learned societies and social media channels. Events such as local colloquia will also enable frequent face-to-face exchange with scientists (see T5.3.3).

**Task 5.2.2 Conference booths and presentations**

The digital change in chemistry shall also be promoted through presence at scientific conferences, both in poster and talk presentations as well as in information booths. Those conferences comprise the GDCh conferences (e.g. Wissenschaftsforum, but also annual or biannual meetings of the subgroups like the German Conference of Chemoinformatics, EuChemS Chemistry Congress), DPhG and DBG annual meetings. A familiarity effect shall be achieved by using the NFDI4Chem corporate layout (see T1.3.1) at all activities. Moreover, information material will be developed in a modular way on different knowledge levels to foster the promotion on conferences, along with promotional materials and give-aways. This material will be continuously extended.

**Task 5.2.3 Roadshows**

We will perform regular roadshows with on-site chemistry-specific seminars at universities across the nation. These roadshows will be performed by a mixed team consisting of chemists, computer scientists, and ideally persons from the local RDM service unit (see T5.3.4). They will get in touch with scientists, mostly doctoral students, during one-day workshops at universities and research facilities to advise them on the scope of RDM and to learn face-to-face about the community's requirements, questions, and preoccupations.

**Deliverables**: (D5.2.1) Establishment of regular processes for news feeds/newsletters on all channels, (D5.2.2) Establishment of a NFDI4Chem conference material set, (D5.2.3) Establishment of a roadshow routine

**Measure 5.3 Training, dissemination, documentation, and support**

**Goals**: Development of training and documentation materials, dissemination to the community, as well as counseling and support

**Description**: Training of the community and dissemination of technical possibilities in RDM has to be performed on several levels. Validated information has to be presented in many ways, starting with a NFDI4Chem homepage as credible information source, ranging over dissemination via the regular colloquia of the learned societies which are well visited in universities' daily life. When the scientists work with the new tools, a helpdesk is needed

for support. Additionally, RDM ambassadors will be trained to contribute to the dissemination of the NFDI4Chem ideas. They will be recruited from the group of actively working PhD students with high interest in RDM.

**Task 5.3.1 Establish a knowledge base**

The community outreach of this TA feeds into a comprehensive online resource on the NFDI4Chem portal (see M1.3) with information regarding all questions of chemistry RDM. This includes standards, recommendations, and best practice examples. In addition, this knowledge base will aggregate all software and service documentation elaborated by the other task areas. Other formats like dedicated blogs and video snippets for best practise examples are planned (in link to M5.2 and M5.5).

**Task 5.3.2 Create training materials**

In order to promote the standardised digital workflows developed by the different TAs, we will create suitable training materials. These training materials will exploit all feasible media formats, including traditional manuals and guidelines but also digital formats such as e-learning resources (massive open online courses, MOOC, moodles), webinars, and video tutorials. Training resources shall include both comprehensive documentations as well as short information snippets (e. g., tip of the month) to stimulate researchers to explore the scope of RDM on their own. Suitable materials for the different user groups will be developed together with the consortium's subcommunity work groups.

**Task 5.3.3 Dissemination through local colloquia**

We also aim to stay in direct contact with the research community. By organising local talks, seminars, and roadshows (see T5.2.3), we do not only seek to promote RDM but also to keep in close touch to the community's needs. This will for instance be performed in the context of the well-established GDCh colloquia (ca. 650 per year) throughout the 60 GDCh local chapters allover Germany. Also the JungChemikerForum (JCF) will help to disseminate the cultural change into the next generation. Moreover, the 150 DPhG colloquia throughout 23 local chapters will help to spread the ideas.

**Task 5.3.4 Establish an infrastructure helpdesk**

TA2 and TA3 will be mainly concerned with development tasks. However, the users will require hands-on assistance to use new digital infrastructures. We envision a central service unit providing not only software support for the RDM tools but also ad-hoc counselling on general and chemistry-specific questions of RDM and on the portfolio of the available NFDI4Chem services. This will include how to deal with highly varied and complex RDM situations within the NFDI4Chem infrastructure itself.

The first part of this support unit will be a first level helpdesk for the software tools offered by NFDI4Chem. This central helpdesk will by run by two helpdesk managers located at FSU and TIB. Their task will be to react quickly to external requests and either reply or dispatch to responsible staff in the second level technical helpdesks for ELN (see T2.2.3)

and repositories (see T3.4.4). To enable an efficient workflow, the helpdesk unit will use a ticketing system.

**Task 5.3.5 Training of community ambassadors**

We also plan to train dedicated community ambassadors. These would be scientists with IT knowledge, able to bridge the gap between users and developers. They will be in close touch with institutional RDM units to propagate the adoption of chemistry-specific RDM aspects. Institutional RDM units will then serve as multipliers at their institutions. We will assist the institutional RDM units in providing chemistry-specific DMP templates that include NFDI4Chem services.

**Deliverables**: (D5.3.1) Establishment of a credible information source in chemistry for all questions related to RDM, (D5.3.2) Development of first set of training materials, (D5.3.3) Development of a second set of training materials, (D5.3.4) Development of a third set of training materials, (D5.3.5) Establishment of an infrastructure helpdesk unit.

**Measure 5.4: Curricular Teaching**

**Goals**: Integration into curricular teaching (for the training of the upcoming scientists).

**Description**: Whereas M5.3 focuses on the current generation of researchers and their RDM training, M5.4 targets on the teaching of the current and upcoming generation of BSc. and Master students. We will develop curricular recommendations which will be implemented at the participants' universities and recommended to all universities by the learned societies. Together with RDM units at the applicants' universities, training materials on all curricular levels will be developed and made accessible publicly.

**Task 5.4.1 Develop curricular recommendations**

The next generation of scientists needs **data literacy** as a meta-competence and key skill of "managing data". Moreover, on a more advanced skill level, **data science** consists of the overlap of programming, mathematics/statistics, and domain knowledge. By fostering data literacy among chemistry students, chemistry can become an essential component in data science. Vice versa, enhanced data science skills offer chemistry students new employing perspectives as these competences are in increasing demand in the chemical and pharmaceutical industry. Moreover, the efficient inclusion of RDM standards in curricula opens up new possibilities in the interconnection of courses and even between student cohorts in different institutions (e.g., comparison of synthetic procedures among different lab years).

To that end, curricular recommendations will be developed in this measure in direct connection to local projects of participating universities (e.g., “Data Literacy Education.nrw”) but also a step-by-step concept for teaching on Bachelor and Master level as well as for the graduate / research group level. After the joint development of curricular recommendations, these recommendations will be disseminated (see M5.3) and illustrated with working examples.

As industry demands and curricular reality in chemistry diverge strongly in terms of Data Science and RDM skills of students, we pursue the definition of key elements of these disciplines to be integrated into the chemistry core curricula in cooperation with the Advisory Board Industry.

**Task 5.4.2 Preparation of teaching materials in cooperation with task 5.3.2**

The recommendations developed in T5.4.1 will be implemented in lecture material aimed at students at all levels. Subdiscipline-specific material will be created based on already existing material of the participating universities and other projects (e.g., FDMentor). This material will comprise classical media such as slides, script material and supplemental material for lectures but also media of modern didactics (e.g. small videos, tutorials, e-learning material, e-tests). The involvement of the GDCh JungChemikerForum and the Doktorandentagung of the DPhG will promote these efforts.

**Deliverables**: (D5.4.1) Manifest on curricular recommendations, (D5.4.2) Provision of a first set of teaching material, (D5.4.3) Provision of a second set of teaching material, (D5.4.4) Provision of a third set of teaching material.

**Measure 5.5: Best Practice and Innovation**

**Goals**: Collection of best practice examples of RDM from both the consortium and the community.

**Description**: Chemistry and related disciplines use a multitude of different workflows to collect experimental and theoretical data. Every lab has evolved individual solutions for its scientific problems. To guide the scientists with real-existing examples from all subdomains in NFDI4Chem, we will collect best practice examples in this measure. A special focus is on studies which combine experimental data with theoretical computations since these benchmark analyses can be tools for validation of theoretical models.

Moreover, we plan to showcase benefits of well-managed scientific data through meta-studies. As an example, the implementation of standards derived for curation and data quality control in NMR spectroscopy of small organic molecules can be considered. Experiences and interdisciplinary exchange may lead to more and where possible, generalized, schemes for research data handling.

**Task 5.5.1 Assemble Use cases, visions, success stories**

We want to encourage scientists to acquaint themselves with RDM by presenting use cases, visions and eventually success stories where good RDM “pays off”. This task has a strong connection to M4.4 (lead-by-example), which focuses on the dataset creation itself.

As the development of better theoretical and computational models for chemical problems hinges critically upon the availability and findability of well-curated experimental data, use cases for NFDI4Chem instruments will be developed covering two components: wet/dry lab-integrating blind prediction challenges and training of students involved in these tasks to maximize sensibility and skills for advanced and sustained RDM. Strong collaboration

with e.g. the RTG 2455 (Göttingen-Öffentlichkeitsarbeit, Georg-August-Universität 2019) and the D3R/SAMPL initiative (D3R 2019, Mobley 2019) will be established to implement NFDI4Chem developments and to feed-back use case results to the NFDI4Chem consortium. Conferences will be opened to NFDI4Chem to this end; e.g. the so far US-hosted “SAMPL Workshop” will for the first time be organized by GDCh-CIC in Europe as a satellite meeting to GCC 2020 in order to establish this link.

**Task 5.5.2 Enabling machine learning applications as use cases derived from NFDI4Chem repositories**

In a cooperation between chemists and machine learning experts, state-of-the-art artificial intelligence incubator use cases based on NFDI4Chem data will be explored. Selected pilot machine learning models will be trained and integrated with NFDI4Chem in order to showcase a large potential for added value. Preliminary ideas include application to predictive property modeling using convolutional or graph-based neural networks, chemical design based on “generative adversarial networks” and “autoencoder” strategies for chemical design (Putin et al. 2018, Gómez-Bombarelli et al. 2018), or neural network-guided retrosynthesis planning (Segler et al. 2018) based on suitable NFDI4Chem data. This shall be done in cooperation with the BMBF competence center for Big Data ScaDS Dresden/Leipzig that is led by TUDr.

**Task 5.5.3 Select and present flagship labs with exemplary RDM implementations**

To encourage acceptance of RDM infrastructures and to inspire the community, we will select laboratories and institutes with exemplary RDM. We will work with these early adopters to harmonize their RDM implementation with NFDI4Chem compliant workflows and standards and showcase them as flagship labs. Lessons learnt during these efforts will inspire our training events and road shows. We will produce videos introducing those flagship institutions.

**Task 5.5.4 Establish an Innovation Incubator**

In this task, we will support TA2 and TA3 in initiating innovative new projects. The Innovation Incubator assists parties with convincing project ideas relevant for NFDI4Chem to apply for centrally managed funds for future project extensions (see M1.4) or other funding opportunities. We will additionally assemble best practice recommendations for software development and CI.

We will propagate the principles of User Centred Design (UCD) and User Driven Development (UDD) in the creation of all services of NFDI4Chem in order meet the requirements and needs of scientists. For this purpose, we will organize UX workshops throughout the lifetime of NFDI4Chem.

**Deliverables**: (D5.5.1) Set of best practice documentation on the NFDI4Chem portal, (D5.5.2) Set of best practice videos from flagship labs on the NFDI4Chem portal.

**Measure 5.6: Communicating to Community Stakeholders**

**Goals**: Connecting NFDI4Chem with important community stakeholders.

**Description**: This measure promotes the standards developed by NFDI4Chem for data and metadata formats in close exchange with international stakeholders and European initiatives. In the long term, these standards will be made mandatory for publication, again in close collaboration with stakeholders such as learned societies and publishers.

**Task 5.6.1 Continuous exchange between community stakeholders to ensure the implementation of NFDI4Chem recommendations**

An active exchange with important stakeholders like chemical societies, GO FAIR ChIN, funding agencies, publishers, and software/instrumentation vendors is essential for the dissemination and acceptance of scientific standards. The learned societies will use their networks and close ties to national, European and international organisations and initiatives (e.g. MOSEX, see LoS) in order to raise awareness on RDM at all stages and in order to avoid parallel and/or contrary developments from the beginning. IUPAC as the world's leading organisation for standardisation in chemistry will be a main partner. In addition, publishers welcome this process and shall at a later stage, as soon as related NFDI4Chem recommendations are available, support their implementation e.g. by adopting corresponding mandatory guidelines for authors.

**Task 5.6.2 Making meta-data standards mandatory through IUPAC and the relevant special sub-committees**

In M4.1 and M4.2, existing data formats will be examined from a broad perspective addressing the needs of a heterogeneous user community with special emphasis on differentiating between research data and appropriate metadata standards. This task ensures the communication of the well defined standards to the community via regular white papers as well as the integration into IUPAC recommendations. Hereby, these standards will become binding in a long-term process. This integration with IUPAC Division VIII (Structural Representation and Chemical Nomenclature) will be supervised by T. Engel and additionally supported by P. Théato (IUPAC Division IV - Polymer Chemistry) and C. Steinbeck (InChI subcommittee).

Furthermore, the Advisory Boards Industry / Publishers, GDCh, IUPAC, Beilstein, and society divisions (e.g. GDCh-CIC, GDCh-FGMR, AGTC) will explore and exchange the needs of scientific meta-/data standards. The commonly arranged definitions ensure a broad acceptance in the community.

**Deliverables**: (D5.6.1) Establishment of regular communication channels to all community stakeholders (“jour fixe”), (D5.6.2.) Recommendation of well defined standards (white papers).

**Risks and Risk Mitigation** - Table 6

Table 6.

Table to Task Area 5: Community Involvement and Training

| Description of risks | Proposed risk-mitigation measures |
|---|---|
| **R5.1: Biased or incomplete requirement assessment.** Assessment of RDM needs may not reach a representative fraction of the community and lead to biased or incomplete results<br>**Likelihood**: Medium | Results must constantly be reviewed and discussed by the National Research Community Panel.<br>Research communities must get involved in the targeting of the surveys.<br>**Involved TA(s)**: TA1, TA5 |
| **R5.2: Developed tools, standards and materials are not adopted by the community**, either by a mismatch of expectations or by ignorance.<br>**Likelihood**: High | Development must constantly be monitored and discussed by the National Research Community Panel.<br>Development activities must include early stage hands-on testing by PhD students.<br>Development activities and results must be massively and repeatedly promoted.<br>**Involved TA(s)**: TA1-5 |
| **R5.3: Premature frustration of the community.** Lack of usability or practicality of developed products may propagate reluctancy for challenges involved with RDM in everyday research.<br>**Likelihood**: Medium | Careful testing of all products with representative user groups.<br>User-friendliness by design<br>**Involved TA(s)**: TA2, TA5 |

## Task Area 6: Synergies

### Description and Objectives

The vision of the NFDI is the cross-linking of infrastructure components and services, as well as the creation of new, comprehensive services for the creation of a national research data infrastructure. The acceptance and utilisation of NFDI4Chem goes hand in hand with a cultural change in chemistry. This change is supported by the added value and new possibilities while developing NFDI4Chem. TA6 addresses issues and actions for a holistic use of the NFDI4Chem infrastructure and services. TA6 ensures the harmonisation of the usage of existing and new components and adds overarching infrastructure services (Fig. 11). The great opportunity for the NFDI is the creation of interdisciplinary approaches and services. Joint strategies for cross-cutting topics will be explored in inter-consortia working groups. NFDI4Chem envisages molecules as linking elements between data collections of chemical-related consortia within the NFDI to investigate and discuss interdisciplinary research questions.

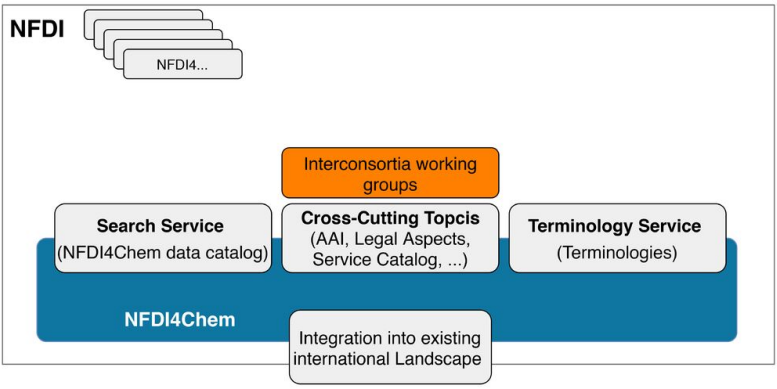


Figure 11.

TA6 will manage the synergies and interfaces to other NFDI consortia and the NFDI as a whole. It will ensure that cross-cutting topics such as NFDI-wide services are properly addressed and their results incorporated into NFDI4Chem.

The NFDI4Chem search service provides a semantically harmonized data catalogue of all NFDI4Chem repositories and data sources. TA6 develops and provides a terminology service for NFDI4Chem. TA6 coordinates the participation of NFDI4Chem in inter-consortia working groups on cross-cutting topics. TA6 ensures the integration of NFDI4Chem and its components into the landscape of existing infrastructures relevant for the chemistry community.

To summarize, the objectives of TA6, designed to support FAIRness in NFDI4Chem, are:

**O6.1** Users can search for research data across distributed data sources of federated NFDI4Chem repositories and linked repositories of the NFDI.

**O6.2** Scientists can use and access NFDI4Chem's services and components in a uniform and user-friendly manner.

**O6.3** The outcome of NFDI4Chem is available on international online platforms.

**O6.4** Users have the possibility to share, to publish, and to reuse research data in a legally sound manner.

TA6 will contribute in particular to key objective 1: *Virtual environment of federated repositories*, key objective 2: *minimum information (MI) standards for data and machine-readable metadata*, key objective 5: *Synergies and key objective* 6: *Legally reliable policies and guidelines*

**Measure 6.1: Development and Implementation of a NFDI4Chem Terminology Service**

**Goals**: The NFDI4Chem terminology service will enable researchers and components of NFDI4Chem and NFDI to access, curate and update terminologies for chemistry and related domains.

**Description**: The exploitation of data by academia and industry continues to be inefficient. The ambiguity of natural language words and phrases used to annotate data with their meanings is perhaps the primary obstacle. Semantics encoded in such form are inaccessible to machines and often ambiguous even to human experts. The consequent lack of semantic data interoperability and under-performing machine support in tasks ranging from discovering to cleaning and integrating data is a root cause for why data are under-exploited. In response to these concerns, the research community is adopting the FAIR data principles. Some of the principles have seen good progress in their implementation. A good example is the resolvable persistent identification (PID) of data to ensure their findability and accessibility. The key remaining concern, however, is the practice of using ambiguous words, phrases or even incomprehensible abbreviations to encode data semantics and the elusive principle of data interoperability. To address this concern, we will develop, deploy and sustainably operate the NFDI4Chem terminology service. The NFDI4Chem Terminology Service will enable semantic data interoperability, discovery and exploitation across disciplines and thus supports user-centric scientific applications. The terminology service is instrumental for establishing a comprehensive, integrated internet of FAIR data and services through semantic representation, terminology-based interoperability and accessibility and interlinking of research data. The Terminology Service will make use of existing vocabularies like CHMO (RSC Ontologies 2019a, RXNO (RSC Ontologies 2019b), CheBI (EBI Web Team 2019) or CHEMINF (Chemical Information Ontology - CHEMINF 2019) and newly developed terminologies within NFDI4Chem.

**Task 6.1.1 NFDI4Chem Terminology service**

The Terminology service will build on global and open standards, such as languages for knowledge representation recommended by W3C. The design and implementation of the Terminology services functionalities are built upon inclusive stakeholder participation from the different disciplines of chemistry. Stakeholders are data providers, curators, researchers and other consumers. We will evaluate, test and finally deploy the most suitable tools for reliable management of subject-specific vocabularies. The web-based working environment VoCol (Halilaj et al. 2016) is one possible tool to be evaluated. VoCol is an open development environment for the agile, joint curation of vocabularies on the basis of Git version management. The basis is the representation of vocabularies with semantic technologies such as RDF and OWL. Knowledge engineers can model with VoCol and collaborate with experts for testing and quality assurance on the basis of the Git versioning. A RESTful API will be implemented to provide access to terminology in a uniform way regardless of their degree of complexity.

**Task 6.1.2 Development and Curation of Terminologies**

In small-scale research projects researchers often create their own terminologies based on their research context. In these cases, the integration of the data into a well-defined terminological environment is often challenging if not impossible. We aim to offer a set of tools for supporting the development, curation and publication (or integration) of such terminologies. This set of tools include, for instance, transformation tools from textual and tabular documents into a semantic format, a linked data interface, terminology integrity checks and validation, etc. For terminology curation, an indexing of content (e.g. vocabularies, ontologies) and a common schema for the Terminology Service output will be defined. A mapping to this schema is required for every underlying terminology or connected external service in order to achieve a harmonized API output of the Terminology Service. Within the VoCol tool for example, a Git version control is used to orchestrate the collaboration with Branching/Merging, Push-Pull Requests and others. VoCol also integrates a variety of services for curation, visualisation, publication, documentation, issue tracking, quality control and validation. Using VoCol, the developed vocabularies can be maintained and adjusted in a continuous, flexible process. Each chemistry subdiscipline may provide discipline and context specific requirements on the terminology service. With the terminology service various stakeholders like data providers, curators, ontology engineers and researchers can develop and curate terminologies collaboratively and continuously. Individual requests from other community stakeholders for new terminologies or updates to existing terminologies can be submitted to a ticket-based help-desk, which serves to ensure quality assurance of terminologies in NFDI4Chem.

**Task 6.1.3 Notification service**

We will develop and implement a tool to enable a semantic and interdisciplinary terminology subscription and notification service for the Terminology Service. The subscription tool will recognise new matching terminologies according to the specifications of a user, activates defined pre-processing of the terminology if requested and informs the subscriber through email that new data is available with a link to access it. The tool will also notify users when terminologies of interest are available or have been changed.

**Task 6.1.4 Workshop series on Terminology development**

The development, adoption and curation of terminologies is a collaborative effort of various stakeholders like data providers, curators, researchers and other consumers. To foster the development of missing terminologies for NFDI4Chem a series of workshops will be facilitated and organised by NFDI4Chem. Each chemistry subdiscipline may provide discipline and context specific requirements on terminologies and the Terminology Service.

**Deliverables**: (D6.1.1) Well-defined human and machine interfaces to access terminology and single-point-of-entry for a chemistry terminology in the NFDI, (D6.1.2) Ticket-based change requests system for community stakeholder, (D6.1.3) Subscription-based notification service, (D6.1.4) Report of Workshops on terminology development, (D6.1.4) NFDI4Chem derived vocabularies

**Measure 6.2: Development of an overarching NFDI4Chem search service**

**Goals**: NFDI4Chem search service provides a semantically harmonized data catalogue of all NFDI4Chem repositories and data sources. Users can search in the data catalogue for research data across NFDI4Chem. The search platform will integrate domain specific search tools like chemical structure search.

**Description**: NFDI4Chem search service aims to provide a semantically harmonized access to the subdiscipline specific repositories. Even within the chemistry domain we have to accept diversity and heterogeneity of data and metadata due to specific requirements of subdisciplines and characteristics of data types. NFDI4Chem will develop and need to agree on shared principles and requirements towards metadata and semantic descriptions for all repositories to enable FAIR data discovery. Nevertheless each repository can provide more detailed metadata and deeper data descriptions for their areas of application. On the next level of a federation of chemistry-related consortia we need to agree on shared metadata semantic descriptions for chemistry-focused searches across these consortia. Finally, for the whole NFDI we need to agree on a core set of metadata and semantic descriptions of data. This approach will allow not only a cross-repository, but also a cross-domain data discovery, supporting FAIR data principles.

For NFDI4Chem, we plan to adapt Comprehensive Knowledge Archive Network (CKAN) (CKAN - Data Management System 2019) as data catalog. CKAN is widely used in government data catalogs, such as the UK's data.gov.uk (Find Open Data - Data.gov.uk 2019), the United States government's Data.gov (Lakhani et al. 2010), and the Australian government (Australian Government 2019) to share open data with the public. In combination with CKAN we further plan to adapt the well-established DCAT Application Profile for data portals in Europe (DCAT-AP), a specification based on the Data Catalogue Vocabulary (DCAT) (DCAT Application Profile for Data Portals in Europe | Joinup 2019) developed by the W3C. It specifies metadata records to meet the specific application needs of data portals while providing semantic interoperability with other applications on the basis of reuse of established controlled vocabularies and mappings to existing metadata vocabularies like Dublin Core (Dublin Core Metadata Initiative 2013). Taking into account chemistry specific requirements the search platform will additionally provide chemical structure search.

**Task 6.2.1 Definition, adaptation and implementation of a NFDI4Chem data catalogue vocabulary**

For NFDI4Chem we plan to adapt the DCAT Application Profile DCAT-AP for the chemical domain. Alternatively the current draft CiteDCAT-AP (DataCite to DCAT-AP Mapping 2019), a DCAT-AP-compliant representation of DataCite metadata may be available for further investigations. Together with the NFDI4Chem repositories we will extend DCAT RDF vocabulary for a standard exchange of metadata in the chemical domain. The new CHEM-DCAT-AP defines additional chemical concepts, attributes and relations of data sets.

**Task 6.2.2 Enable metadata harvesting**

In close cooperation with the NFDI4Chem repositories APIs for metadata harvesting will be developed and implemented to provide the data for the search service. Alternatively ingest functions of the repositories into the data catalogue may be evaluated and implemented.

**Task 6.2.3 Overarching NFDI4Chem search**

The NFDI4Chem search will be realised by a customized CKAN distribution. Customisations include visualisation of chemical structures and an interface for chemical structure search. The search service provides faceted search and semantic tagging. The semantic tagging functionality will be linked to terminology service to obtain tags from the registered vocabularies.

**Task 6.2.4 Implementation of a chemical structure search for the NFDI4Chem search service**

Searching chemical information is often performed by drawing a chemical structure with a respective tool and use the parsed structure as query. open source cheminformatics toolkits (e.g. CDK, RDKIT, see T2.6.1) will be employed to store structural information adequately and enable sophisticated structure, substructure and similarity searches. For the user interface drawing tools developed in T2.5.1 will be implemented.

**Deliverables**: (6.2.1) Approved chemistry specific data catalogue vocabulary, (D6.2.2) Repositories provide APIs to harvest metadata, (D6.2.3) Web portal for searching in metadata catalogue of NFDI4Chem repositories, (D6.2.3) Integrated chemical structure search in harvested chemical structure data, providing chemical structure information

**Measure 6.3: Coordination of and contribution to cross-cutting topics activities within the NFDI**

**Goals**: Members of NFDI4Chem are involved in the activities concerning cross-cutting topics relevant for the consortium and shape processes in our best interest. Results of the inter-consortia work are successfully embedded into the development of the NFDI4Chem.

**Description**: During the initial phase of the NFDI several cross-cutting topics have been identified and discussed with different consortia (see 2.3.1). As a matter of fact, the vision of the NFDI will be shaped by collaborative work on these topics among the consortia. With this measure, we will ensure that NFDI4Chem will identify or initiate and participate in relevant inter-consortia working groups to foster cross-cutting topics. We have already identified cross-cutting topics of outstanding importance for our consortium, to which we want to contribute to the NFDI and inter-consortia working groups: a NFDI-wide service catalogue, an AAI service and a legally reliable framework of policies and guidelines for FAIR research data management.

**Task 6.3.1 Coordination of cross-cutting topics activities within the NFDI**

Together with other consortia, cross-cutting topics will be identified and working groups established. Activities and discussions needed to be monitored and forwarded to TA leads when necessary. NFDI4Chem delegates will then participate in relevant working groups

and decision making processes. The results of the working groups will be embedded in the infrastructure and services of NFDI4Chem. A first workshop has already been discussed with NFDI4Ing and further consortia.

**Task 6.3.2 User Identity Management for Authentication and Authorisation**

In the spirit of FAIR data principles, access to data and services should be as open and interoperable as possible. Nevertheless there are several reasons why both resources and services require researchers to sign in. Especially sensitive data and licensed services need strong security regarding access and rights management. For the many services in the NFDI4Chem, the number of individual, separate logins quickly becomes difficult to manage. To overcome this burden a NFDI wide Authentication and Authorisation Infrastructure (AAI) needs to be implemented. A successful implementation of an AAI service is e.g. the ELIXIR-AAI (ELIXIR Consortium 2019). Building on the DFN-AAI (DFN-AAI 2019) users are enabled to use their existing university logins. Alternatively, organisation credentials or community identities like ORCID or OpenID may be used to sign in and access data and services they need. Furthermore the AAI service allows service providers to control and manage access rights of their users and create different roles and access levels for single researchers or research groups. For NFDI4Chem we will address this cross-cutting topic in a way that users of NFDI4Chem and the NFDI can easily access resources and services by using single-sign-on and well-defined authorisation mechanism. We will contribute to the development of a NFDI-AAI Service. We will collect the specific requirements of the diverse NFDI4Chem repositories and systems such as the ELN. Using this knowledge, we will participate in the NFDI wide AAI discussions and workshops in order to 1) help further the AAI infrastructure and 2) represent the interests of NFDI4Chem and 3) coordinate the integration and implementation the NFDI wide AAI for NFDI4Chem systems

**Task 6.3.3 Legally reliable framework of policies and guidelines for FAIR research data management**

The operation of an infrastructure as the NFDI has to build on rules, to that individual users commit themselves. To ensure the confidence in the NFDI, these rules have to be in accordance with German and if possible international legal status quo. At the same time, industry, professional associations, publishers and other rights holders are important, established partners for chemical research. It is a key to provide a science-friendly legal framework that balances that great variety of interests through guidelines and data policies, terms of use, licensing standards, data property. Of course this furthermore includes the legal support of the involved researchers and research managers in the universities and other research institutes. Within this task we address the necessary steps to provide a legally reliable framework of policies and guidelines for FAIR RDM for researcher in the NFDI4Chem.

**Development of guidelines and policies for liability regulations for service and infrastructure providers**: Operation models of service and infrastructure providers have to address the liability regime. Together with relevant service and infrastructure providers

liability relevant constellations and conditions especially for operating in the NFDI4Chem and NFDI infrastructure will be identified and evaluated. Based on derived use cases guidelines and policies for liability regulations in the context of use conditions for NFDI4Chem services and infrastructures will be worked out.

**Development of guidelines and policies for legally reliable research data management**: With the implementation of the FAIR data principles, researchers need a reliable legal framework when publishing their data in repositories. Legal considerations about data ownership and possible licensing models are essential in this regard. In this task we develop guidelines and policies for researchers to choose appropriate license models. The analysis will cover the legal relationships between institutions and their researchers and further stakeholders like funding institutions or industrial partners. Legal assignment standards must also be taken into account, as well as revised DFG guidelines on good scientific practise. From these, legal suggestions concerning doctoral rules, supervision agreements, institutional rules and other scientific and higher education relevant areas will be distilled.

**Deliverables**: (D6.3.1) Documentation of relevant cross-cutting topics for NFDI4Chem, (D6.3.2) NFDI4Chem delegates are members of inter-consortia working groups relevant to NFDI4Chem, (6.3.3) Reports on the integration of working groups results into NFDI4Chem, (D6.3.4) Guidelines and policies for legal reliable implementation of FAIR data principles for researchers and research institutions, (D6.3.5) Guidelines and policies for liability regulations for services within NFDI4Chem

**Measure 6.4 Integration into the landscape of existing infrastructure**

**Goals**: Cross-linking NFDI4Chem with existing research data infrastructures, services and projects of the aforementioned areas

**Description**: In order to increase the findability and reuse of research data provided by the NFDI, cooperation with and connection to existing national and international infrastructures and infrastructure projects should be taken into account. We are aware of several infrastructure projects which have been pursuing similar aims. Many of these projects have already been invited to collaborate in TA2 and TA3. In this measure, we will seek to identify further projects relevant for the NFDI4Chem network and seek collaboration and integration. This measure will be conducted in close collaboration with TA2 and TA3. For existing infrastructure collaborations to increase the findability and reuse of NFDI4Chem data will be explored and implemented.

**Task 6.4.1 Identify, contact and invite further infrastructure projects to collaborate with NFDI4Chem**

Beyond already known infrastructures and projects we will look for further initiatives as matching partners for NFDI4Chem. Shared interests and joint approaches on NFDI4Chem key objectives will be discussed to evaluate possible cooperations and cross-linking of infrastructures and services.

### Task 6.4.2 Increase the findability of NFDI4Chem data in the pharmaceutical domain

The PubPharm (PubPharm 2019) search portal of Fachinformationsdienst (FID) Pharmazie specifically meets the search requirements of the various pharmaceutical disciplines, e.g. searching for pharmaceutical relevant journal articles in more than 55 million references. As a joint effort we aim to integrate metadata from NFDI4Chem search service in PubPharm and to extend the PubPharm search space to include chemical research data, so that PubPharm could offer a central pharmacy-specific access for these research data. In cooperation with PubPharm we will agree on metadata suitable for PubPharm and develop an automated metadata exchange.

### Task 6.4.3 Increase the findability of NFDI4Chem data in chemistry web portals

We have identified several well-established web resources for chemical information as possible partners to increase the findability of NFDI4Chem data beyond the NFDI initiative on the international level. Portals like ChemSpider from the Royal Society of Chemistry (RSC), the open chemistry database PubChem at the National Institutes of Health (NIH) and CompTox Chemicals Dashboard (Williams et al. 2017) at the US Environmental Protection Agency (US EPA) provide workflows for metadata ingest from external data sources like NFDI4Chem. We will negotiate the registration of NFDI4Chem as data source for these portals and ingest data to enrich these data collections. Additionally we will investigate the enrichment of Wikidata (Wikidata 2019) items of chemical entities (molecules) with relations to research data of NFDI4Chem.

**Deliverables**: (D6.5.1) Reports of further infrastructure projects relevant to NFDI4Chem, (D6.5.2) Cooperation agreements with further infrastructure projects where possible, (D6.5.3) NFDI4Chem metadata ingests into PubChem, Chemspider, Wikidata

**Risks and Risk Mitigation** - Table 7.

Table 7.

Table to Task Area 6: Synergies

| Description of risks | Risk-mitigation measures |
|---|---|
| **R1**: Scientists do not participate actively enough in the development of vocabularies for the subdisciplines. The cooperation between scientists and ontology engineers does not work optimally.<br>**Likelihood**: Low | Vocabulary development will be promoted by community measures in TA5 and through ontology workshops supported by stakeholders from TA2, T3 and TA4.<br>**TA(s) involved**: TA2, TA3, TA4 |
| **R2**: Repositories of NFDI4Chem insufficiently support the overaching search service.<br>**Likelihood**: low | Close cooperation with TA3 and TA4 in the development of necessary metadata standards and APIs.<br>**TA(s) involved**: TA3, TA4 |
| **R3**: Necessary results from cross-cutting topic activities are not available in time for NFDI4Chem tasks.<br>**Likelihood**: medium | NFDI4Chem will actively initiate, contribute and lead interconsortia discussions in the topics most relevant to NFDI4Chem. NFDIChem will organise interconsortia workshops. TA leads will coordinate and communicate their needs towards NFDI.<br>**TA(s) involved**: All TAs |

| | |
|---|---|
| **R4**: Legal policies and guidelines are not adopted by scientists sufficiently and do not lead to the desired acceptance by scientists and have no increase in the willingness for data publication. **Likelihood**: low | Close involvement of the community to collect requirements from the very beginning of the measure. Cooperation with community measures in TA5. **TA(s) involved**: TA5 |

## Abbreviations

Table 8

Table 8.
Abbreviations

| | |
|---|---|
| AAI | Authentication and Authorization Infrastructure |
| BC | Biochemistry |
| CD | Continuous Deployment |
| ChEMBL | Chemical database of the European Molecular Biology Laboratory |
| CI | Continuous Integration |
| CRC | Collaborative Research Centre |
| DB | Database |
| DBG | Deutsche Bunsen-Gesellschaft für Physikalische Chemie |
| DPhG | Deutsche Pharmazeutische Gesellschaft |
| DSC | Differential Scanning Calorimetry |
| ELN | Electronic Laboratory Notebook |
| GDCh | Gesellschaft Deutscher Chemiker |
| GPC | Gel Permeation Chromatography |
| HPC | High Performance Computing |
| HPLC | High Performance Liquid Chromatography |
| IC | Inorganic Chemistry |
| InChI | IUPAC International Chemical Identifier |
| incl | including |
| IR | Infrared vibrational spectroscopy |
| IUPAC | International Union of Pure and Applied Chemistry |
| LIMS | Laboratory Information and Management System |
| MI | minimum information |
| MIChI | Minimum Information about a Chemical Investigation |
| MS | mass (spectrometry) |
| NFDI4BIMP | NFDI4 Biological Imaging and Medical Photonics |
| NMR | Nuclear Magnetic Resonance spectrscopy |

| | |
|---|---|
| OC | Organic Chemistry |
| OS | Open Source |
| PAINS | Pan Assay Interference Compounds |
| PC | Physical Chemistry |
| PMC | Pharmaceutical and Medicinal Chemistry |
| PolyC | Polymer Chemistry |
| RDA | Research Data Alliance |
| RDM | Research Data Management |
| SC | Steering Committee |
| SMILES | Simplified Molecular-Input Line-Entry System |
| TGA | Thermogravimetric analysis |
| UCD | User-ceneterd design |
| UDD | User-driven development |
| UI | User Interface |
| UV | Absorption spectroscopy in ultraviolet wavelength range |
| UX | User Experience |

## References


- Adam B, Lindstädt B (2019) Elektronische Laborbücher Im Kontext von Forschungsdatenmanagement Und Guter Wissenschaftlicher Praxis - Ein Wegweiser Für Die Lebenswissenschaften. ZB MED - Informationszentrum Lebenswissenschaften https://doi.org/10.4126/FRL01-006415715
- Mistrik R, Lutisan J, Huang Y, Suchy M, Wang J, Raab M (2013) “mzCloud: A Key Conceptual Shift to Understand ‘Who’s Who’in Untargeted Metabolomics.”. A Key Conceptual Shift to Understand 'Who’s Who’in Untargeted Metabolomics. Metabolomics Society 2013 Conference, Glasgow, July 2013.
- Jung N, Tremouilhac P, Braese S (2018) Scope of ELNs and repositories to improve scientific documentation and reporting: Examples taken from the Chemotion-ELN and Chemotion-Repository. CINF – 29. American Chemical Society.
- Arshad J, Hoffmann A, Gesing S, Grunzke R, Krüger J, Kiss T, Herres-Pawlis S, Terstyanszky G (2016) Multi-level meta-workflows: new concept for regularly occurring tasks in quantum chemistry. Journal of Cheminformatics 8 (1). https://doi.org/10.1186/s13321-016-0169-8
- Audus D, de Pablo J (2017) Polymer Informatics: Opportunities and Challenges. ACS Macro Letters 6 (10): 1078-1082. https://doi.org/10.1021/acsmacrolett.7b00228
- Australian Government (2019) https://data.gov.au/. Accessed on: 2019-10-11.
- Banfi D, Patiny L (2008) www.nmrdb.org: Resurrecting and Processing NMR Spectra On-line. CHIMIA International Journal for Chemistry 62 (4): 280-281. https://doi.org/10.2533/chimia.2008.280

- Barillari C, Ottoz DM, Fuentes-Serna JM, Ramakrishnan C, Rinn B, Rudolf F (2015) openBIS ELN-LIMS: an open-source database for academic laboratories. Bioinformatics 32 (4): 638-640. https://doi.org/10.1093/bioinformatics/btv606
- Bär RM, Heinrich G, Nieger M, Fuhr O, Bräse S (2019) Insertion of [1.1.1]propellane into aromatic disulfides. Beilstein Journal of Organic Chemistry 15: 1172-1180. https://doi.org/10.3762/bjoc.15.114
- Beisken S, Conesa P, Haug K, Salek RM, Steinbeck C (2015) SpeckTackle: JavaScript charts for spectroscopy. Journal of Cheminformatics 7 (1). https://doi.org/10.1186/s13321-015-0065-7
- Bika LIMS (2014) Bika Open Source LIMS project. https://www.bikalims.org/. Accessed on: 2014-10-12.
- Bräse S, Jung N, Kotov S, Tremouilhac P (2017) Chemotion-Repository. https://www.chemotion-repository.net
- Bräse S, Nestler B (2019) Science Data Center Soll Datenaustausch Für Molekulare Materialforschung Erleichtern. no. 2 (February). Laborpraxis URL: https://www.laborpraxis.vogel.de/science-data-center-soll-datenaustausch-fuer-molekulare-materialforschung-erleichtern-a-800464/
- BRENDA Enzyme Database (2019) https://www.brenda-enzymes.org. Accessed on: 2019-10-11.
- Chamanara J, Kraft A, Auer S, Koepler O (2019) Towards Semantic Integration of Federated Research Data. Datenbank-Spektrum 19 (2): 87-94. https://doi.org/10.1007/s13222-019-00315-w
- ChEMBL Database (2019) https://www.ebi.ac.uk/chembl. Accessed on: 2019-10-11.
- Chemical Information Ontology - CHEMINF (2019) Semanticchemistry. https://github.com/semanticchemistry/semanticchemistry. Accessed on: 2019-10-11.
- Chemistry - GO FAIR (2019) https://www.go-fair.org/implementation-networks/overview/chemistryin. Accessed on: 2019-10-10.
- Chemistry Research Data IG (2015) https://www.rd-alliance.org/groups/chemistry-research-data-interest-group.html
- CKAN - Data Management System (2019) https://ckan.org. Accessed on: 2019-10-11.
- Coles SJ, Frey JG, Bird CL, Whitby RJ, Day AE (2013) First steps towards semantic descriptions of electronic laboratory notebook records. Journal of Cheminformatics 5 (1). https://doi.org/10.1186/1758-2946-5-52
- D3R (2019) https://drugdesigndata.org/about/samp. Accessed on: 2019-10-10.
- Dabb S (2016) ChemSpider: Search and share chemistry. Abstracts of papers of the American Chemical Society., Vol. 251
- Dassault Systèmes BIOVIA (2019a) Electronic Laboratory Notebooks. https://www.3dsbiovia.com/products/unified-lab-management/biovia-electronic-lab-notebooks. Accessed on: 2019-9-01.
- Dassault Systèmes BIOVIA (2019b) Laboratory Information Management System. https://www.3dsbiovia.com/products/unified-lab-management/biovia-lims. Accessed on: 2019-9-01.
- DataCite Schema (2019) http://schema.datacite.org
- DataCite to DCAT-AP Mapping (2019) https://ec-jrc.github.io/datacite-to-dcat-ap. Accessed on: 2019-10-11.
- Day A, Coles S, Bird C, Frey J, Whitby R, Tkachenko V, Williams A (2015) ChemTrove: Enabling a Generic ELN To Support Chemistry through the use of transferable plug-ins

and online data sources. Journal of Chemical Information and Modeling 55 (3): 501-509. https://doi.org/10.1021/ci5005948

- DCAT Application Profile for Data Portals in Europe | Joinup (2019) https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe. Accessed on: 2019-10-11.
- DFN-AAI (2019) https://www.aai.dfn.de. Accessed on: 2019-10-10.
- Douglas S (2017) Chemotion ELN - LIMSWiki. https://www.limswiki.org/index.php/Chemotion_ELN. Accessed on: 2017-6-10.
- Dublin Core Metadata Initiative (2013) Website Dublin Core Metadata Initiative. https://www.dublincore.org
- EBI Web Team (2019) Chemical Entities of Biological Interest (ChEBI). https://www.ebi.ac.uk/chebi. Accessed on: 2019-10-11.
- Electronic Lab Notebooks, HMS (2019) Electronic Lab Notebooks. https://datamanagement.hms.harvard.edu/electronic-lab-notebooks. Accessed on: 2019-8-18.
- ELIXIR Consortium (2013) Authentication and Authorisation Infrastructure. https://elixir-europe.org/services/compute/aai. Accessed on: 2019-10-10.
- ELIXIR Consortium (2019) A Distributed Infrastructure for Life-Science Information. https://elixir-europe.org. Accessed on: 2019-10-11.
- FAIRsharing.org: MassBank (2016) https://fairsharing.org/FAIRsharing.dk451a. Accessed on: 2016-10-18.
- FDM-Kontakt (2019) https://www.forschungsdaten.org/index.php/FDM-Kontakte. Accessed on: 2019-10-10.
- Find Open Data - Data.gov.uk (2019) https://data.gov.uk. Accessed on: 2019-10-11.
- Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2011) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research 40 https://doi.org/10.1093/nar/gkr777
- Gesing S, Grunzke R, Krüger J, Birkenheuer G, Wewior M, Schäfer P, Schuller B, Schuster J, Herres-Pawlis S, Breuers S, Balaskó Á, Kozlovszky M, Fabri AS, Packschies L, Kacsuk P, Blunk D, Steinke T, Brinkmann A, Fels G, Müller-Pfefferkorn R, Jäkel R, Kohlbacher O (2012) A single sign-on infrastructure for science gateways on a use case for structural bioinformatics. Journal of Grid Computing 10 (4): 769-790. https://doi.org/10.1007/s10723-012-9247-y
- Gesing S, Krüger J, Grunzke R, Herres-Pawlis S, Hoffmann A (2016) Using science gateways for bridging the differences between research infrastructures. Journal of Grid Computing 14 (4): 545-557. https://doi.org/10.1007/s10723-016-9385-8
- GGA Software Services (2019) Indigo. https://github.com/ggasoftware/indigo. Accessed on: 2019-9-01.
- Glöckner FO, et al. (2019) Berlin Declaration on NFDI Cross-Cutting Topics (Version 1). Zenodo https://doi.org/10.5281/zenodo.3457213
- Gómez-Bombarelli R, Wei J, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel T, Adams R, Aspuru-Guzik A (2018) Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science 4 (2): 268-276. https://doi.org/10.1021/acscentsci.7b00572
- Göttingen-Öffentlichkeitsarbeit, Georg-August-Universität (2019) RTG 2455 Benchmark experiments for numerical quantum chemistry. https://www.uni-goettingen.de/en/587836.html. Accessed on: 2019-10-10.

- Grunzke R, Breuers S, Gesing S, Herres-Pawlis S, Kruse M, Blunk D, de la Garza L, Packschies L, Schäfer P, Schärfe C, Schlemmer T, Steinke T, Schuller B, Müller-Pfefferkorn R, Jäkel R, Nagel W, Atkinson M, Krüger J (2013) Standards-based metadata management for molecular simulations. Concurrency and Computation: Practice and Experience 26 (10): 1744-1759. https://doi.org/10.1002/cpe.3116
- Grunzke R, Hartmann V, Jejkal T, Kollai H, Prabhune A, Herold H, Deicke A, Dressler C, Dolhoff J, Stanek J, Hoffmann A, Müller-Pfefferkorn R, Schrade T, Meinel G, Herres-Pawlis S, Nagel W (2019) The MASi repository service — Comprehensive, metadata-driven and multi-community research data management. Future Generation Computer Systems 94: 879-894. https://doi.org/10.1016/j.future.2017.12.023
- Guha R, Howard M, Hutchison G, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen E (2006) The Blue Obelisk—Interoperability in Chemical Informatics. Journal of Chemical Information and Modeling 46 (3): 991-998. https://doi.org/10.1021/ci050400b
- Hagstrom S (2014) The FAIR Data Principles. FORCE11. URL: https://www.force11.org/group/fairgroup/fairprinciples
- Halilaj L, Petersen N, Grangel-González I, Lange C, Auer S, Coskun G, Lohmann S (2016) VoCol: An Integrated Environment to Support Version-Controlled Vocabulary Development. Lecture Notes in Computer Science 303-319. https://doi.org/10.1007/978-3-319-49004-5_20
- Hall SR, Allen FH, Brown ID (1991) The crystallographic information file (CIF): a new standard archive file for crystallography. Acta Crystallographica Section A Foundations of Crystallography 47 (6): 655-685. https://doi.org/10.1107/s010876739101067x
- Hausen DA (2019) Forschungsdaten in der Chemie (Version 1). Zenodo. https://doi.org/10.5281/zenodo.3415059
- Herres-Pawlis S, Koepler O, Steinbeck C (2019) NFDI4Chem: Digitalen und kulturellen Wandel in der Chemie gestalten. Angewandte Chemie 131 (32): 10880-10882. https://doi.org/10.1002/ange.201907260
- Huang Y, Nguyen A, Gräßle S, Vanderheiden S, Jung N, Bräse S (2018) Addition of dithi(ol)anylium tetrafluoroborates to α,β-unsaturated ketones. Beilstein Journal of Organic Chemistry 14: 515-522. https://doi.org/10.3762/bjoc.14.37
- Huang Y (2019) React-Spectra-Editor. https://github.com/ComPlat/react-spectra-editor. Accessed on: 2019-10-03.
- ioChem BD (2019) https://www.iochem-bd.org. Accessed on: 2019-10-14.
- Jung N, Grässle S, Lütjohann DS, Bräse S (2014) Solid-supported odorless reagents for the dithioacetalization of aldehydes and ketones. Organic Letters 16 (4): 1036-1039. https://doi.org/10.1021/ol403313h
- Jung N, Tremouilhac P, Kramer C, Potthoff J (2017) heiBOOKS. In: Kratzke J, Heuveline V (Eds) E-Science-Tage 2017: Forschungsdaten Managen. Universitätsbibliothek Heidelberg, 127–135 pp. URL: http://books.ub.uni-heidelberg.de/heibooks/catalog/book/285
- Jung N (2019) The chemotion model for digital data management and its benefits for scientists and community. 47th World Chemistry Congress of IUPAC. Le Palais des Congrès of Paris, 7 November 2019. Paris
- Kotov S, Tremouilhac P, Jung N, Bräse S (2018) Chemotion-ELN part 2: adaption of an embedded Ketcher editor to advanced research applications. Journal of Cheminformatics 10 (1). https://doi.org/10.1186/s13321-018-0292-9

- Kraft A, Razum M, Potthoff J, Porzel A, Engel T, Lange F, van den Broek K, Furtado F (2016) The RADAR Project—A Service for Research Data Archival and Publication. ISPRS International Journal of Geo-Information 5 (3). https://doi.org/10.3390/ijgi5030028
- Krüger J, Grunzke R, Gesing S, Breuers S, Brinkmann A, de la Garza L, Kohlbacher O, Kruse M, Nagel W, Packschies L, Müller-Pfefferkorn R, Schäfer P, Schärfe C, Steinke T, Schlemmer T, Warzecha KD, Zink A, Herres-Pawlis S (2014) The MoSGrid Science Gateway – A Complete Solution for Molecular Simulations. Journal of Chemical Theory and Computation 10 (6): 2232-2245. https://doi.org/10.1021/ct500159h
- Kuhn S, Helmus T, Lancashire R, Murray-Rust P, Rzepa H, Steinbeck C, Willighagen E (2007) Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML vocabulary for spectral data. Journal of Chemical Information and Modeling 47 (6): 2015-2034. https://doi.org/10.1021/ci600531a
- Kuhn S, Schlörer N (2015) Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2 - a free in-house NMR database with integrated LIMS for academic service laboratories. Magnetic Resonance in Chemistry 53 (8): 582-589. https://doi.org/10.1002/mrc.4263
- Labfolder - Electronic Lab Notebook (ELN) (2019) https://www.labfolder.com. Accessed on: 2019-9-01.
- Lagoze C, Van de Sompel H (2003) The making of the Open Archives Initiative Protocol for Metadata Harvesting. Library Hi Tech 21 (2): 118-128. https://doi.org/10.1108/07378830310479776
- Lakhani KR, Austin RD, Yi Y (2010) Data.gov. Harvard Business School Case
- Lampen P, Lambert J, Lancashire RJ, McDonald RS, McIntyre PS, Rutledge DN, Fröhlich T, Davies A (1999) An extension to the JCAMP-DX standard file format, JCAMP-DX V.5.01. Pure and Applied Chemistry 71 (8): 1549-1556. https://doi.org/10.1351/pac199971081549
- Lampen P, Hillig H, Davies A, Linscheid M (2016) JCAMP-DX for Mass Spectrometry. Applied Spectroscopy 48 (12): 1545-1552. https://doi.org/10.1366/0003702944027840
- LIMS L (2019) Software Für Menschen Aus Forschung Und Analytik. https://www.limsophy.com. Accessed on: 2019-9-01.
- LIMS Software (2019) Simple LIMS Software Solution for Chemical Laboratory. https://www.simplelimssoftware.com/chdemo.html. Accessed on: 2019-9-01.
- LIMSWiki. (2019) https://www.limswiki.org/index.php/Main_Page. Accessed on: 2019-9-01.
- Lütjohann D, Nicole J, Pierre T, Bräse S (2014) Das Internet Der Dinge Erobert Das Chemische Forschungslabor. GIT Fachzeitschrift Für Das Laboratorium
- Lütjohann D, Jung N, Bräse S (2015) Open source life science automation: Design of experiments and data acquisition via “dial-a-device”. Chemometrics and Intelligent Laboratory Systems 144: 100-107. https://doi.org/10.1016/j.chemolab.2015.04.002
- Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang W, Römpp A, Neumann S, Pizarro A, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz P, Deutsch E (2010) mzML—a Community Standard for Mass Spectrometry Data. Molecular & Cellular Proteomics 10 (1). https://doi.org/10.1074/mcp.r110.000133
- Materials Project (2019) https://materialsproject.org. Accessed on: 2019-9-09.

- Mobley DL (2019) The SAMPL Challenges. https://samplchallenges.github.io. Accessed on: 2019-10-10.
- Mohamed A, Nguyen CH, Mamitsuka H (2016) NMRPro: an integrated web component for interactive processing and visualization of NMR spectra. Bioinformatics 32 (13): 2067-2068. https://doi.org/10.1093/bioinformatics/btw102
- Mu H (2018) Top 10+ Open Source Laboratory Management Systems - LIMS. https://medevel.com/top-10-foss-lims. Accessed on: 2018-12-11.
- Murray-Rust P, Rzepa H (1999) Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. Journal of Chemical Information and Computer Sciences 39 (6): 928-942. https://doi.org/10.1021/ci990052b
- Murray-Rust P, Rzepa H, Wright M (2001) Development of chemical markup language (CML) as a system for handling complex chemical content. New Journal of Chemistry 25 (4): 618-634. https://doi.org/10.1039/b008780g
- Neumann J, Brase J (2014) DataCite and DOI names for research data. Journal of Computer-Aided Molecular Design 28 (10): 1035-1041. https://doi.org/10.1007/s10822-014-9776-5
- NOMAD Repository (2019) https://nomad-repository.eu. Accessed on: 2019-10-14.
- NORMAN Network (2019) Welcome to the NORMAN Network. https://www.norman-network.net. Accessed on: 2019-10-10.
- Notebook|PerkinElmer (2019) https://www.perkinelmer.com/category/notebook. Accessed on: 2019-9-01.
- O'Boyle NM, Guha R, Willighagen EL, Adams SE, Alvarsson J, Bradley J, Filippov IV, Hanson RM, Hanwell MD, Hutchison GR, James CA, Jeliazkova N, Lang AS, Langner KM, Lonie DC, Lowe DM, Pansanel J, Pavlov D, Spjuth O, Steinbeck C, Tenderholt AL, Theisen KJ, Murray-Rust P (2011) Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on. Journal of Cheminformatics 3 (1). https://doi.org/10.1186/1758-2946-3-37
- Open-LIMS (2019) The Open-Source Laboratory Information Management System. http://www.open-lims.org. Accessed on: 2019-9-01.
- Potthoff J, Lütjohann D, Jung N (2014) Trustworthy Laboratory Automation. In: Laux F, et al. (Ed.) The Sixth International Conference on Advances in Databases, Knowledge, and Data Applications. IARIA. 98-103 pp.
- Potthoff J, Tremouilhac P, Hodapp P, Neumair B, Bräse S, Jung N (2019) Procedures for systematic capture and management of analytical data in academia. Analytica Chimica Acta: X 1 https://doi.org/10.1016/j.acax.2019.100007
- PubChem (2019) https://pubchem.ncbi.nlm.nih.gov. Accessed on: 2019-10-10.
- PubPharm (2019) https://www.pubpharm.de. Accessed on: 2019-10-10.
- Pupier M, Nuzillard J, Wist J, Schlörer N, Kuhn S, Erdelyi M, Steinbeck C, Williams A, Butts C, Claridge TW, Mikhova B, Robien W, Dashti H, Eghbalnia H, Farès C, Adam C, Kessler P, Moriaud F, Elyashberg M, Argyropoulos D, Pérez M, Giraudeau P, Gil R, Trevorrow P, Jeannerat D (2018) NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. Magnetic Resonance in Chemistry 56 (8): 703-715. https://doi.org/10.1002/mrc.4737
- Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A, Zhavoronkov A (2018) Reinforced Adversarial Neural Computer for de Novo Molecular Design. Journal of Chemical Information and Modeling 58 (6): 1194-1204. https://doi.org/10.1021/acs.jcim.7b00690

- re3data.org (2019) http://www.re3data.org. Accessed on: 2019-10-10.
- RSC Ontologies (2019a) https://github.com/rsc-ontologies/rsc-cmo. Accessed on: 2019-10-11.
- RSC Ontologies (2019b) https://github.com/rsc-ontologies/rxno. Accessed on: 2019-10-11.
- Rudolphi F, Goossen L (2011) Electronic Laboratory Notebook: The Academic Point of View. Journal of Chemical Information and Modeling 52 (2): 293-301. https://doi.org/10.1021/ci2003895
- Rudolphi F (2019) Sciformation ELN - Sciformation Consulting GmbH. http://sciformation.com/sciformation_eln.html?lang=de. Accessed on: 2019-9-01.
- SampleManager LIMS Software (2019) https://www.thermofisher.com/order/catalog/product/INF-11000. Accessed on: 2019-9-01.
- Sansone S-A, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L, Begley K, Booth T, Bougueleret L, Burns G, Chapman B, Clark T, Coleman L, Copeland J, Das S, de Daruvar A, de Matos P, Dix I, Edmunds S, Evelo CT, Forster MJ, Gaudet P, Gilbert J, Goble C, Griffin JL, Jacob D, Kleinjans J, Harland L, Haug K, Hermjakob H, Ho Sui SJ, Laederach A, Liang S, Marshall S, McGrath A, Merrill E, Reilly D, Roux M, Shamu CE, Shang CA, Steinbeck C, Trefethen A, Williams-Jones B, Wolstencroft K, Xenarios I, Hide W (2012) Toward interoperable bioscience data. Nature Genetics 44 (2): 121-6. https://doi.org/10.1038/ng.1054
- Schober D, Jacob D, Wilson M, Cruz J, Marcu A, Grant J, Moing A, Deborde C, de Figueiredo L, Haug K, Rocca-Serra P, Easton J, Ebbels TD, Hao J, Ludwig C, Günther U, Rosato A, Klein M, Lewis I, Luchinat C, Jones A, Grauslys A, Larralde M, Yokochi M, Kobayashi N, Porzel A, Griffin J, Viant M, Wishart D, Steinbeck C, Salek R, Neumann S (2017) nmrML: A Community Supported Open Data Standard for the Description, Storage, and Exchange of NMR Data. Analytical Chemistry 90 (1): 649-656. https://doi.org/10.1021/acs.analchem.7b02795
- Schultze-Motel P (2018) Helmholtz Open Science Workshop ‘Elektronische Laborbücher. URL: http://gfzpublic.gfz-potsdam.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:3862890
- Scinote-Web (2019) https://github.com/biosistemika/scinote-web. Accessed on: 2019-9-01.
- Segler MS, Preuss M, Waller M (2018) Planning chemical syntheses with deep neural networks and symbolic AI. Nature 555 (7698): 604-610. https://doi.org/10.1038/nature25978
- Spectral Database for Organic Compounds (SDBS) (2019) https://sdbs.db.aist.go.jp. Accessed on: 2019-10-12.
- Standards (2019) http://rd-alliance.github.io/metadata-directory/standards. Accessed on: 2019-10-07.
- STARLIMS (2019) Laborinformationsmanagement-System (LIMS). https://www.informatics.abbott/int/de/. Accessed on: 2019-9-01.
- Steinbeck C, Krause S, Kuhn S (2003) NMRShiftDB Constructing a free chemical information system with open-source components. Journal of Chemical Information and Computer Sciences 43 (6): 1733-1739. https://doi.org/10.1021/ci0341363

- Steinbeck C, Kuhn S (2004) NMRShiftDB – compound identification and structure elucidation support through a free community-built web database. Phytochemistry 65 (19): 2711-2717. https://doi.org/10.1016/j.phytochem.2004.08.027
- Stein SE (1990) National Institute of Standards and Technology (NIST) Mass Spectral Database and Software. Version 3.02. NIST, Gaithersburg.
- SWORD (2019) About SWORD – SWORD. http://swordapp.org/about/. Accessed on: 2019-10-10.
- Taylor CF, Field D, Sansone S, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz P, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Novère NL, Leebens-Mack J, Lewis SE, Lord P, Mallon A, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert CJ, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nature Biotechnology 26 (8): 889-896. https://doi.org/10.1038/nbt.1411
- Technische Informationsbibliothek, FIZ Chemie, Universität Paderborn (2010) Vernetzte Primärdaten-Infrastruktur Für Den Wissenschaftler-Arbeitsplatz in Der Chemie : Konzeptstudie. Technische Informationsbibliothek, Hanover. https://doi.org/10.15488/5515
- The Cambridge Structural Database (CSD) (2019) The Cambridge Crystallographic Data Centre (CCDC). https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd. Accessed on: 209-10-11.
- Tipton K, Armstrong R, Bakker B, Bairoch A, Cornish-Bowden A, Halling P, Hofmeyr J, Leyh T, Kettner C, Raushel F, Rohwer J, Schomburg D, Steinbeck C (2014) Standards for Reporting Enzyme Data: The STRENDA Consortium: What it aims to do and why it should be helpful. Perspectives in Science 1: 131-137. https://doi.org/10.1016/j.pisc.2014.02.012
- Tremouilhac P, Nguyen A, Huang Y, Kotov S, Lütjohann DS, Hübsch F, Jung N, Bräse S (2017) Chemotion ELN: an Open Source electronic lab notebook for chemists in academia. Journal of Cheminformatics 9 (1). https://doi.org/10.1186/s13321-017-0240-0
- Tristram F (2019) bwFDM -Offenes Ergebnismaterial. https://bwfdm.scc.kit.edu. Accessed on: 2019-10-09.
- Vinaixa M, Schymanski E, Neumann S, Navarro M, Salek R, Yanes O (2016) Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. TrAC Trends in Analytical Chemistry 78: 23-35. https://doi.org/10.1016/j.trac.2015.09.005
- Vosegaard T (2015) jsNMR: an embedded platform-independent NMR spectrum viewer. Magnetic Resonance in Chemistry 53 (4): 285-290. https://doi.org/10.1002/mrc.4195
- Wakelin J, Murray-Rust P, Tyrrell S, Zhang Y, Rzepa HS, García A (2005) CML tools and information flow in atomic scale simulations. Molecular Simulation 31 (5): 315-322. https://doi.org/10.1080/08927020500065850
- Website Suprabank (2019) http://suprabank.org. Accessed on: 2019-10-12.
- Wikidata (2019) https://www.wikidata.org. Accessed on: 2020-10-10.

- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18
- Wilkinson M, Sansone S, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M (2018) A design framework and exemplar metrics for FAIRness. Scientific Data 5 (1). https://doi.org/10.1038/sdata.2018.118
- Williams A, Grulke C, Edwards J, McEachran A, Mansouri K, Baker N, Patlewicz G, Shah I, Wambaugh J, Judson R, Richard A (2017) The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. Journal of Cheminformatics 9 (1). https://doi.org/10.1186/s13321-017-0247-6
- Willighagen E, Mayfield J, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O, Torrance G, Evelo C, Guha R, Steinbeck C (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. Journal of Cheminformatics 9 (1). https://doi.org/10.1186/s13321-017-0220-4
- Willoughby C, Bird C, Coles S, Frey J (2014) Creating Context for the Experiment Record. User-Defined Metadata: Investigations into Metadata Usage in the LabTrove ELN. Journal of Chemical Information and Modeling 54 (12): 3268-3283. https://doi.org/10.1021/ci500469f
- Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS (2012) MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis. Nucleic Acids Research 40 https://doi.org/10.1093/nar/gks374
- Zhang F, Brüschweiler R (2007) Robust Deconvolution of Complex Mixtures by Covariance TOCSY Spectroscopy. Angewandte Chemie 119 (15): 2693-2696. https://doi.org/10.1002/ange.200604599