Research Ideas and Outcomes 11: e177037
doi: 10.3897/rio.11.e177037

Grant Proposal

Proposal NFDI4Chem 2025-2030 In the National Research Data Infrastructure (NFDI) — Our Vision: All Chemists Publish FAIR Data

Christoph Steinbeck[‡], Nicole Jung[§], Felix Bach[|], Steffen Neumann[¶], Sonja Herres-Pawlis[#], Johannes C. Liermann[¤], Oliver Koepler[«], Christoph Bannwarth[#], Theo Bender[¤], Thomas W. Bocklitz[»], Franziska Boehm[|], Christian Bonatto Minella[|], Frank Biedermann[§], Werner Brack[^], Ricardo Cunha[˅], Paul Czodrowski[¤], Franziska Eberl[‡], Thomas Engel[¦], Albert K. Engstfeld[?], Tillmann G. Fischer[ς], Pascal Friedrich[§], Frank Glorious[¢], Benjamin Golub[ℓ], Christoph Grathwol[৯], Rainer Haag[P], Johannes Hunold[«], Christoph R. Jacob[ℓ], Jochen Johannsen[#], John D. Jollife[¤], Stefan M. Kast[₳], Carsten Kettner[₴], Stefan Kuhn[₣], Giacomo Lanza[₮], Jan Lisec[₦], Georg Manolikakes[₭], Ricardo A. Mata[₲], Jens Meiler[ς], Matthias Müller[#], Ralph Müller-Pfefferkorn[?], Jochen Ortmeyer[#], Wendy Patterson[₴], Jürgen Pleiss[₩], Annalisa Riedel[¤], Jens Riedel[₦], Ulrich Schatzschneider[₸], Leonie Schuster[«], Peter Seeberger[⌗], Johann-Nikolaus Seibert[₭], Peter F. Stadler[ς], Philip Strömert[«], Robert Strötgen[ℓ], Patrick Théato[§], Nicolas Tielker[₳], Pierre Tremouilhac[§], Hans-Georg Weinig[§§], Wolfgang Wenzel[§], Kirsten Zeitler[ς]

‡ Friedrich Schiller University, Jena, Germany
§ Karlsruhe Institute of Technology, Karlsruhe, Germany
| FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Germany
¶ Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany
# RWTH Aachen University, Aachen, Germany
¤ Johannes Gutenberg University Mainz, Mainz, Germany
« TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
» Leibniz Institute of Photonic Technology e.V., Jena, Germany
^ Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany
˅ Institut für Umwelt & Energie, Technik & Analytik e.V. (IUTA), Duisburg, Germany
¦ Ludwig-Maximilians-Universität München, Munich, Germany
? Ulm University, Ulm, Germany
ς Leipzig University, Leipzig, Germany
¢ University of Münster, Münster, Germany
ℓ Technische Universität Braunschweig, Braunschweig, Germany
৯ Deutsche Pharmazeutische Gesellschaft e.V., Frankfurt am Main, Germany
P Freie Universität Berlin, Berlin, Germany
₳ Dortmund University, Dortmund, Germany
₴ Beilstein-Institut, Frankfurt am Main, Germany
₣ Tartu University, Tartu, Estonia
₮ Physikalisch-Technische Bundesanstalt, Braunschweig, Germany
₦ Bundesanstalt für Materialforschung und -prüfung, Berlin, Germany
₭ RPTU University Kaiserslautern-Landau, Kaiserslautern, Germany
₲ University of Göttingen, Göttingen, Germany
? Dresden University of Technology, Dresden, Germany
₩ University of Stuttgart, Stuttgart, Germany
₸ Julius-Maximilians-Universität Würzburg, Würzburg, Germany
‡‡ Center for the Transformation of Chemistry, Delitzsch, Germany
§§ Gesellschaft Deutscher Chemiker e.V., Frankfurt am Main, Germany

© Steinbeck C et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Corresponding author: Christoph Steinbeck (christoph.steinbeck@uni-jena.de), Nicole Jung (nicole.jung@kit.edu), Felix Bach (Felix.Bach@fiz-karlsruhe.de), Steffen Neumann (sneumann@ipb-halle.de), Sonja Herres-Pawlis (sonja.herres-pawlis@ac.rwth-aachen.de), Johannes C. Liermann (liermann@uni-mainz.de), Oliver Koepler (oliver.koepler@tib.eu)

Reviewable v 1

Received: 04 Nov 2025 | Published: 26 Nov 2025

Citation: Steinbeck C, Jung N, Bach F, Neumann S, Herres-Pawlis S, Liermann JC, Koepler O, Bannwarth C, Bender T, Bocklitz TW, Boehm F, Bonatto Minella C, Biedermann F, Brack W, Cunha R, Czodrowski P, Eberl F, Engel T, Engstfeld AK, Fischer TG, Friedrich P, Glorious F, Golub B, Grathwol C, Haag R, Hunold J, Jacob CR, Johannsen J, Jollife JD, Kast SM, Kettner C, Kuhn S, Lanza G, Lisec J, Manolikakes G, Mata RA, Meiler J, Müller M, Müller-Pfefferkorn R, Ortmeyer J, Patterson W, Pleiss J, Riedel A, Riedel J, Schatzschneider U, Schuster L, Seeberger P, Seibert J-N, Stadler PF, Strömert P, Strötgen R, Théato P, Tielker N, Tremouilhac P, Weinig H-G, Wenzel W, Zeitler K (2025) Proposal NFDI4Chem 2025-2030 In the National Research Data Infrastructure (NFDI) — Our Vision: All Chemists Publish FAIR Data. Research Ideas and Outcomes 11: e177037. https://doi.org/10.3897/rio.11.e177037

Abstract

The first funding period of NFDI4Chem established a robust foundation for research data management (RDM) in chemistry by promoting FAIR data principles and creating a cohesive infrastructure to capture well-annotated data early in the lab through electronic lab notebooks (ELNs) and making this data available in public repositories. Key achievements include standardised data formats and metadata, a federated repository environment, and improved data visibility and accessibility. Training programs and outreach have significantly increased awareness and adoption of best RDM practices. In the second funding period, the consortium aims to advance these achievements by consolidating this infrastructure, developing a model for its sustainable maintenance and operation, and fostering cultural change for its widespread adoption. Goals include ensuring seamless data workflows from laboratories to open repositories, enhancing interoperability, and supporting innovative research through AI-ready data. The work plan is organised into six task areas (TAs). TA1 (Management) provides leadership and supports all other TAs in achieving their objectives. TA2 (Smart Lab) aims to develop a fully digital research environment, including an ELN as a modular platform. This environment will support data collection, management, storage, analysis, and sharing. Integrating devices and external resources will enable seamless data transfer to repositories. TA3 (Repositories) will consolidate the repository ecosystem. The goal is to integrate repositories into a federated system for better accessibility and interoperability, ensuring long-term data availability and sustainability. TA4 (Metadata, Data Standards, and Publication Standards) focuses on developing and promoting new data and metadata standards in an international community process. This includes applying ontologies to create a semantic foundation for linking research data, making it machine-readable and enabling knowledge graphs. TA5 (Community and Training) is dedicated to fostering a cultural shift towards digital chemistry through continuous engagement, collecting requirements, and providing extensive training and support through workshops and open education resources. It will promote FAIR-compliant machine learning

applications, embedding RDM into academic curricula to ensure future scientists are well-versed in these practices. TA6 (Synergies and Cross-Cutting Topics) aims to enhance collaboration across NFDI consortia and beyond. This includes developing ontologies, terminology services, the search service, and other cross-cutting solutions, integrating these developments into existing infrastructure, enabling interdisciplinary data harmonisation and fostering machine learning applications.

Keywords

data, infrastructure, FAIR, chemistry

List of Abbreviations

AAI

Authentication and Authorization Infrastructure

AB

Advisory Board

API

Application Programming Interface

CD

Circular Dichroism

ChEMBL

Chemical database of the European Molecular Biology Laboratory

CI/CD

Continuous Integration / Continuous Deployment

CRC

Collaborative Research Centre

CTS

CoreTrustSeal

CV

Cyclic Voltammetry

DB

Database

DBG

Deutsche Bunsen-Gesellschaft für Physikalische Chemie

(D)LS

(Dynamic) Light Scattering

DPhG

Deutsche Pharmazeutische Gesellschaft

DSC

Differential Scanning Calorimetry

ELN

Electronic Laboratory Notebook

EOSC

European Open Science Cloud

EPR

Electron Paramagnetic Resonance

FID

Specialised Information Services (Fachinformationsdienst)

FP1

Funding Phase 1

FP2

Funding Phase 2

GPC

Gel Permeation Chromatography

GUI

Graphical User Interface

HPC

High-Performance Computing

HPLC

High-Performance liquid Chromatography

InChI

IUPAC International Chemical Identifier

IR

Infrared (vibrational spectroscopy)

IUPAC

International Union of Pure and Applied Chemistry

KO

Key objective

LADS

Laboratory and Analytical Device Standard

LTA

Long Term Archiving

LIMS

Laboratory Information and Management System

MI

Minimum Information

MIChI

Minimum Information about a Chemical Investigation

MS

Mass Spectrometry

NMR

Nuclear Magnetic Resonance (spectroscopy)

OA

Open Access

OAI-PMH

Open Archives Initiative Protocol for Metadata Harvesting

OC

Organic Chemistry

OER

Open Educational Resource

OS

Open Source

PAINS

Pan Assay Interference Compounds

PC

Physical Chemistry

PMC

Pharmaceutical and Medicinal Chemistry

PSDI

Physical Sciences Data Infrastructure

RDA

Research Data Alliance

RDM

Research Data Management

RM

Risk and Mitigation

SBOM

Software Bill Of Materials

SC

Steering Committee

SDH

Semantic Data Hub

SMILES

Simplified Molecular-Input Line-Entry System

TA

Task Area

TF

Task Force

TGA

Thermogravimetric Analysis

UI

User Interface

UV

Absorption spectroscopy in the UltraViolet wavelength range

UX

User Experience

1 Preface

This proposal was written and submitted to the funding agency in August 2024. The text published here is a slightly shortened version of the submitted proposal and is structured along the guidelines specified by the funding agency. It outlines a working program for 5 years and assumes the full funding of the requested budget. Upon the decision by the German Research Council (Gemeinsame Wissenschaftskonferenz) in July 2025, NFDI4Chem is facing a shorter period of funding (2025/10 - 2028/12) and a yearly budget 30% lower than originally requested. These shorter resources will certainly influence our work program. Nevertheless, we are grateful for the funding and look

forward to further building and supporting the National Research Data Infrastructure for chemists.

2 Scope and Objectives

2.1 Research domains addressed by the consortium, specific aim(s)

NFDI4Chem is building a research data infrastructure for all chemical fields dealing with the synthesis, analysis and characterisation of molecules (Steinbeck et al. 2020). We are making this distinction to define clear-cut boundaries to those NFDI consortia dealing with materials science. To spend the inevitably limited funding responsibly and effectively, our working principle is to establish infrastructure components for those fields in chemistry with the highest number of users, as defined by the size of the respective divisions in the German Chemical Society.

In the next funding period, NFDI4Chem aims to expand and enhance its scope by addressing several future needs: We will consolidate our repository ecosystem and extend existing ones to cover more subdisciplines and data types and pursue the integration of repositories into a federated system for better accessibility and interoperability. We will develop further standards for data and metadata and promote and establish open data standards in cooperation with international bodies. Additionally, we develop and apply ontologies to create a semantic foundation for linking research data, enabling machine-readable data and creating knowledge graphs. All this enables us to continue the digital transformation of chemistry by using Electronic Lab Notebooks (ELNs) and tools in Smart Lab environments for early digital data and metadata capture, linking these with data repositories and a downstream search service to seamlessly digitise the entire research process. Ultimately, we provide access to semantically annotated, high-quality research data for chemistry research and future AI methods.

Since its inception, NFDI4Chem has impacted the research landscape in several ways: We promoted standardised data formats and metadata, facilitating better data sharing and reuse (Herres-Pawlis et al. 2022). NFDI4Chem established a virtual environment of federated repositories, improving the visibility and accessibility of research data and integrating existing repositories with national and international systems, thus improving the overall data infrastructure for chemical research (Neumann et al. 2023).

Towards a cultural change to digital chemistry, we boosted interest and engagement for the FAIR data principles within the chemistry community through training programmes and outreach activities to promote best practices in research data management (RDM) (Herres-Pawlis 2023, Herres-Pawlis et al. 2023, Pearman-Kanza et al. 2024, Koepler et al. 2020). We established a solid ontological ground for machine-actionable data with well-curated chemistry ontologies provided by the terminology services (Strömert et al. 2022, Strömert et al. 2023).

2.2 Objectives and measuring success

NFDI4Chem's efforts and decisions are guided by our vision that *all chemists publish FAIR data.* Our mission is to *support scientists in their efforts to collect, store, process, analyse, publish, and reuse research data in chemistry.* We develop and maintain a national research data infrastructure for the research domain of chemistry in Germany, and we enable innovative services and science based on research data.

To translate our vision and mission into reality, we have identified several key objectives:

Key Objective 1: Evolve the federation of services and repositories for storing, disclosing, searching and reusing research data. Integrate community standards for the seamless integration of distributed data sources and uniform access to data for innovative data reuse.

Key Objective 2: Lead international community processes to establish minimum information (MI) standards for data and machine-readable metadata to create semantically rich linked chemistry data.

Key Objective 3: Develop and foster the application of Smart Laboratory Environments by promoting the use of digital tools in all stages of research and creating seamless digital data workflows.

Key Objective 4: Foster cultural and digital change within the chemistry community in Germany to create awareness for and foster a community-agreed technically reliable RDM infrastructure for all levels of academia, beginning in undergraduate studies curricula.

Key Objective 5: Collaborate with other consortia and promote cross-cutting developments to extend synergies within OneNFDI and enable interdisciplinary data harmonisation and integration.

Key Objective 6: Facilitate data reuse and enable AI in chemistry by providing AI-ready data, data infrastructures, and high-quality data and metadata through all services.

These key objectives align with the overarching goals of the NFDI, described in the federal-state agreement (Bund-Länder-Vereinbarung) (Gemeinsame Wissenschaftskonferenz 2018). For long-term commitment and the trust of the chemistry community, statements on the long-term availability of the NFDI4Chem infrastructure are imperative. Key players of the consortium are, in alignment with their strategic vision, willing to commit to the long-term provision of repositories and services, which nonetheless require reliable, additional funding. NFDI4Chem recognises its role in the consolidation of the NFDI. The entanglement of our own services with basic services, exemplified by the terminology service, demonstrates our commitment to the consolidation process. We continue to contribute to the cross-cutting topics of the NFDI and metadata standardisation efforts. Our services are tailored to meet the needs of our community. We will continue our strategy and conduct in-depth community surveys

regularly. We have adapted the DFG datasheet and the defined key performance indicators (KPI) to monitor our services and community impact in FP2. We will identify further KPIs relevant to NFDI4Chem and embed these in the organisational structure and operations in FP2 to evaluate our success (T1.3.3). A critical assessment of the KPIs will be conducted as part of the annual consortium meeting. We will participate in the NFDI TF Evaluation & Reporting for the overarching development of NFDI-wide KPIs and utilise recommendations from the NFDI.

3 Consortium

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

NFDI4Chem started in October 2020 with 7 (co-)applicant institutions, many of which contributed with several institutes or departments, and 13 participating institutions. From 2022 to 2024, we gained 4 additional partners for the consortium who had already supported the work in various areas in FP1, specifically in enzymology, material and analytical sciences, didactics, photo- and electrochemistry.

For FP2, we will expand the consortium to include more representatives in physical, theoretical, and computational chemistry, scientific infrastructure, data processing, and AI development With these new partners, the consortium in FP2 will have 7 applicant institutions (many of them including several contributing institutes), 20 participating institutions and 2 participating individuals.

Embedding in the community

From the outset, NFDI4Chem identified the needs of the community with a survey in 2019 (Herres-Pawlis et al. 2020). In spring 2023, a second survey with more than 800 responses (680 from Germany) included professors (22%), senior researchers (24%) and PhD students (33%) from different areas of chemistry. Digital data analysis improved, with 45% using seamless methods and 65% still using some non-digital steps. Metadata description increased to 56% (42% in 2019). ELN use increased to 30% (18% in 2019), particularly in organic chemistry (34%) and materials chemistry (33%). Common ELNs included Chemotion (26%), ELabFTW (20%) and Sciformation / Open Enventory (9%), as well as Word and Excel, which some also classified as ELNs. The use of data repositories increased by 20% (from 13%). 86% felt that research data management in the curriculum would benefit future students and groups. This feedback demonstrates the impact of NFDI4Chem, and further surveys will continue in FP2. (see M5.1).

NFDI4Chem has strong links with the major subdisciplines of chemistry (organic, inorganic, polymer, biochemical, pharmaceutical, physical, computational and analytical chemistry) through learned societies as consortium members (GDCh, DPhG, Bunsen Society, FID Pharmazie) and outreach activities. Collaborative development with specialised communities such as chemical ontologies, NMR, EPR standards,

electrochemistry, coordination chemistry, macromolecules, and enzymes is achieved through regular meetings and workshops. Users can provide feedback via the helpdesk and GitHub repositories. Interaction with the ontology community is particularly strong, as evidenced by the annual Ontologies4Chem workshops with ~50 participants from major chemistry ontology projects (Koepler et al. 2023) and joint activities with IUPAC and international projects such as the Physical Sciences Data Infrastructure (PSDI). NFDI4Chem actively participates in working groups on standards for analytical chemistry methods.

The NFDI4Chem training programme is essential to promote RDM awareness. We have developed a two-day interactive course for chemists covering RDM basics and specific scenarios based on the FD Mentor concept (Biernacka et al. 2020) and materials from the research data teams at FSU, RWTH and JGU. Launched in January 2022, the workshop is open to all chemists in Germany, with growing international participation. It was held every two months in 2022 and was consistently well booked. In 2023, we moved to institution-specific workshops in collaboration with local RDM teams. The Chemotion workshop series introduces the electronic lab notebook through a "learning by doing" approach, adaptable to the needs of the participants. A similar format is offered for the generic LabIMotion modules in Chemotion. By July 2024, we have organised 73 workshops (see Fig. 1 for an overview of on-site workshops in Germany), reaching over 800 participants. To meet demand and support the NFDI section EduTrain, we are expanding to include workshops for data stewards and a 1-day Lead-by-example workshop. In addition, we are providing materials for integrating chemistry-specific RDM content into curricula to facilitate teaching culture change.

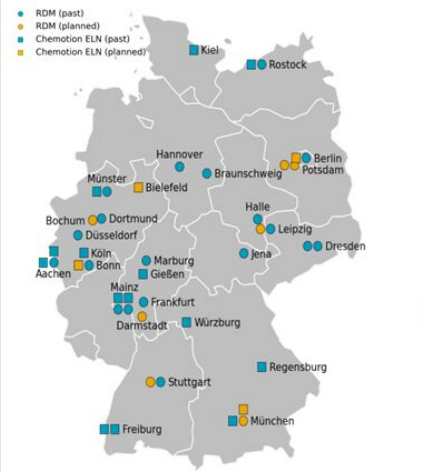

Fig. 1: *In-person workshops on RDM and Chemotion held by NFDI4Chem so far, and still planned in 2024.*

Figure 1. doi

In-person workshops on RDM and Chemotion held by NFDI4Chem until date of submission (2024/08), and still planned in 2024.

To promote RDM awareness among young scientists, we presented the FAIR4Chem award at the GDCh JungChemikerForum (JCF) in 2022, 2023, and 2024. In 2023, we organised the first Chemistry Data Days, a two-day conference on data management for non-RDM experts, with around 100 participants. We engage with the community through monthly 'Stammtisch' discussions on RDM, ELNs, repositories, and ontologies, along with Chemotion Q&A sessions and various RDM and Chemotion workshops. Finally, NFDI4Chem is active on several social media channels: LinkedIn (753 followers, >170 articles), Instagram (168 followers, 148 posts), BlueSky (247 followers, 109 posts), and X (formerly Twitter, 1184 followers, 562 tweets), the latter no longer actively promoted (all figures as of 30/07/2024).

3.2 The consortium within the NFDI and the national academic research system

NFDI4Chem members have actively shaped the NFDI, with our spokesperson leading the Consortium Assembly in 2021/22. NFDI4Chem spokespersons and co-spokespersons played key roles in NFDI strategy workshops (Glöckner et al. 2019, Bierwirth et al. 2020, Ebert et al. 2021, Konsortialversammlung des Vereins Nationale Forschungsdateninfrastruktur (NFDI) e.V. 2022), helping to establish its foundational bodies. Recognising that the success of NFDI hinges on effective collaboration across consortia, NFDI4Chem is deeply committed to addressing cross-cutting topics and creating overarching solutions to common challenges. In NFDI, the sections serve as crucial forums for addressing cross-cutting issues with the goal of creating overarching solutions. NFDI4Chem's commitment to this collaborative approach is evidenced by the election of three consortium members as spokespersons for NFDI e.V. sections (Boehm et al. 2021, Koepler et al. 2021, Herres-Pawlis et al. 2021), i.e. Section “Education and Training”, “(Meta)data, Terminologies and Provenance”, and “Ethical, Legal and Social Aspects. Two of them are heading the sections in their second term of office from April 2024. NFDI4Chem members actively contributed to task forces and working groups within the various sections during the first funding phase. Our contributions include developing standardised methodologies for ontology curation and development processes, using terminology services in research data management tools and services, and establishing NFDI-wide identity and access management services. We also contribute to developing research data management plans, open educational resources for RDM training, support and recommendations for ELN implementations in consortia, legal aspects of RMD, or an NFDI metadata standard in the Task Force Metadata (Hunold et al. 2023). A significant outcome of our collaborative efforts is the close cooperation with NFDI4Cat, FAIRMat, NFDI4Ing, DataPlant, NFDI4Health, NFDI4Culture, and NFDI4Objects (Dolcet et al. 2003). Cooperations cover developing and curating ontologies while enhancing tools and services for ontology curation (NFDI4Chem 2023, Bender 2024, NFDI4Chem 2024e). Furthermore, we are collaborating with FAIRMat and NFDI4Cat to develop and apply electronic lab notebooks (ELN). In FP2, we will closely monitor developments and stay involved in strengthening the sections, contributing to joint solutions for common challenges, and achieving tangible cross-consortia results.

The work of the sections highlights the need for collaborative solutions across NFDI consortia. We, therefore, welcome the Base4NFDI objective to create common services to support RDM in chemistry. NFDI4Chem is committed to fostering the emerging Base4NFDI services and, where relevant for our consortium, to engage with and use future Base4NFDI services. NFDI4Chem has actively participated in the Base4NFDI workshops held on 16 February 2024 and 22 April 2024 to identify and agree on viable contributions of the consortia for the successful integration of Base4NFDI services. In alignment with ongoing developments for the chemistry community, NFDI4Chem is actively contributing to the basic service for Terminology Services TS4NFDI during the initialisation phase, integrating the NFDI4Chem TS into the TS4NFDI backend architecture. We will continue our engagement in the integration phase of TS4NFDI as described in M6.2. During the initialisation phase, we actively contributed to the requirements collection for IAM4NFDI, TS4NFDI, KG4NFDI, and PID4NFDI by completing surveys and attending workshops. NFDI4Chem has engaged with IAM4NFDI in incubator pilots (IAM4NFDI 2024), prototyping the access control to our Terminology Services and Chemotion ELN via Community AAI. In FP2, NFDI4Chem will continue to support structured ongoing communication with Base4NFDI. We will closely monitor and connect to activities in Base4NFDI and provide a dedicated communication channel to Base4NFDI technical and community updates in M6.4.

NFDI4Chem and community members have also contributed to the NFDI infra-talk series and co-organise the Physical Sciences Consortia Joint Colloquium with five other consortia. In addition, NFDI4Chem is strengthening links with universities, RDM service units, and IT centres, having (co-)organised six local networking events in the past (Neumann et al. 2024). NFDI4Chem continues to present its work and services at national RDM events. Targeting FDM initiatives of the federated states, we strive to sustain the integration of NFDI4Chem training measures into academic RDM services (see T5.3.5). Hereby, we empower the next generation of scientists with urgently needed digital skills (Bertelmann et al. 2024). This goes hand in hand with integrating RDM content into curricular teaching (see M5.4).

More details on our continuous involvement in the NFDI are described in M6.4, and in Eberl et al. (2024) concerning collaborations.

3.3 International networking

NFDI4Chem emphasises that developing standards and best practices for research data management requires a global perspective. Thus, NFDI4Chem actively engages with the worldwide community of chemists and research data infrastructure experts. Strengthening collaboration with the International Union of Pure and Applied Chemistry (IUPAC), NFDI4Chem aligns its measures with several IUPAC projects. TA4 and TA6 work with the FairSpec project. In the WorldFAIR: Global cooperation on FAIR data policy and practice project, NFDI4Chem contributed to IUPAC WorldFAIR Chemistry deliverable 3.2 "Chemistry Training Package" (Chalk et al. 2024). TA6 and IUPAC are jointly developing recommendations for using the Compendium of Chemical Terminology

("Gold Book") to define new ontology terms in chemistry-specific ontologies (Strömert et al. 2024). NFDI4Chem representatives participate in various Research Data Alliance (RDA) working and interest groups, i.e., regularly contributing to the Chemistry Research Data Interest Group (CRDIG) sessions at RDA plenary meetings. The interoperable federation of repositories was showcased and discussed with RDM experts at international conferences like FDO2022, RDA plenary 20, RDA-DE2023, IDW2023, SciDataCon, and PIDfest2024. S. Herres-Pawlis is a Board Member of the InChI Trust, guiding inorganic chemistry implementation into the InChI. P. Théato chairs the Sub-Committee on Polymer Terminology, and the NFDI4Chem spokesperson C. Steinbeck serves on the IUPAC Committee on Publications and Cheminformatics Data Standards. Additionally, NFDI4Chem organised three IUPAC Global women's breakfasts, highlighting women's achievements in NFDI4Chem. We recently joined a panel on critical reflections on RDM at the SDG Graduate Schools Alliance midterm conference, emphasising equal opportunities for researchers in the global south. NFDI4Chem has established collaborations with international learned societies such as the American Chemical Society (ACS) with its division of Chemical Information (CINF). NFDI4Chem co-organized two sessions on "Helping Chemists manage their Data" and "Metadata to Knowledge Graphs" at the ACS Fall Meeting 2023. S. Herres-Pawlis is on the advisory board of the UK's Physical Sciences Data Infrastructure (PSDI), and a joint session with PSDI was held during EuChemS 2024 in Dublin. NFDI4Chem regularly collaborates with the European Chemical Society (EuChemS). At the EuChemS congress 2022 in Lisbon and 2024 in Dublin, Chemotion ELN garnered significant interest. We also collaborate with the International and the European Younger Chemists Networks (IYCN and EYCN), key stakeholders for RDM implementation in universities, while GDCh-JCF acts as a multiplier for NFDI4Chem in the German community. For instance, we organised an international RDM workshop at the IUPAC Conference in Den Hague in 2023. NFDI4Chem works with the Royal Society for Chemistry (RSC) on developing and curating ontologies. NFDI4Chem partners are active in the European Life Sciences Infrastructure ELIXIR, contributing to Bioschemas developments and organising projects at the 2022 and 2023 ELIXIR Hackathons. Embedded in the NFDI activities, NFDI4Chem has presented its work at European Open Science Cloud (EOSC) events, such as the presentation of the Terminology Service at the EOSC Symposium 2021 in the session "Metadata and Data Quality." Co-spokesperson Oliver Koepler actively contributes to the EOSC Technical and Semantic Interoperability Task Force.

3.4 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. The organisational structure described below has been established during the FP1 and has served the consortium well. In the following, we describe the various components of our **organisational structure**, as shown in Fig. 2.

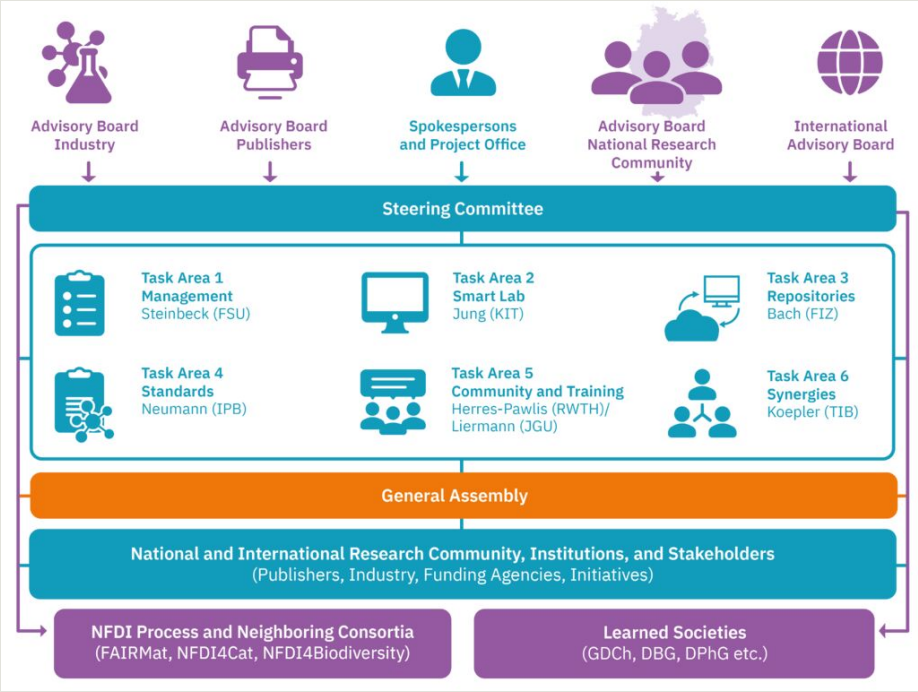

Figure 2. doi

NFDI4Chem governance structure.

The General Assembly (GA) is the central decision-making body of the project. The GA comprises all co-applicants and participants (represented by one delegate from each partner).

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 it. The annual NFDI4Chem meetings will be used for strategic planning, presentation of scientific results, and discussion of significant management issues. The GA decides when an overall agreement is required in budget and consortium management matters.

The **Steering Committee (SC)** is the central body responsible for monitoring and evaluating project progress and supervising project objectives, and it makes the necessary decisions in scientific coordination and administration of the project. Based on the contributions from task areas (TAs), the SC prepares periodic and final reports. The SC consists of the two speakers, the NFDI4Chem project manager, the TA leads and their project managers. Project speaker Christoph Steinbeck heads the SC.

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

a) aligned with the mission of the NFDI in general and

b) address the needs of the chemistry community.

The ABs and their advice proved invaluable during our first funding period.

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

Advisory Board *Publishers* provides insights on research data associated with scientific publications. They aim to include 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* represents 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, focusing on international organisations and individuals. These 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 TAs. 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 at the next steering committee (SC) meeting. The TA leaders oversee the budget and 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 that do not require the voting of the GA. The SC convenes by electronic communication regularly and on-demand, as organised 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 make particular decisions if overall consensus is required in the matters of inclusion of a new partner, exclusion of an existing partner, major budget or project strategy changes, and 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 physically present at the GA or

through teleconference facilities during decision-making and voting. In a stalemate situation, the Project Coordinator will have the deciding vote.

Internal Distribution of Funds: According to our consortium agreement, FSU, as the lead organisation for NFDI4Chem, will conclude forwarding contracts with all co-applicants and participants as part of a normal consortium agreement following well-established models.

3.5 Operating model

NFDI4Chem is committed to building a robust and sustainable open data research infrastructure that is accessible and beneficial to users and providers. Our operating model is based on the following pillars: **User-centric approach**: We prioritise the needs of researchers, educators and other stakeholders by providing easy access to high-quality research data. Our services include data storage, management, sharing, and analysis tools that are user-friendly and reliable. **Collaborative Framework**: Data providers, including academic institutions, research organisations and data centres, work together within the consortium to offer their resources and expertise. This collaborative approach ensures a diverse and comprehensive federation of data repositories serving various research areas. **Open access and interoperability**: We are committed to the principles of open access, ensuring that all data within our infrastructure is freely available to users. In addition, we emphasise interoperability to facilitate seamless integration with other data infrastructures and research tools worldwide. Our operational and financial model has proven effective, benefiting both users and providers. By relying on funding from research funding agencies, we have avoided the pitfalls of commercial interests that could compromise the open-access ethos of our consortium. The current funding spares data providers and users from costs, preventing a drop in scientific engagement and avoiding the overhead of collecting user fees across multiple services. The critical components of our funding model are as follows. **Infrastructure funding**: Primary funding comes from national infrastructure funding. Funding agencies recognise the importance of open research data for scientific progress and innovation and provide grants and subsidies to support our activities. This infrastructure funding allows for the sustainable operation and maintenance of NFDI4Chem after the first two funding periods. **In-kind Contributions**: Consortium members contribute in-kind resources such as computing power, data storage facilities and personnel. This model leverages existing capabilities within member institutions, reducing overall costs and fostering a sense of ownership and commitment among members. **Project-based grants**: We actively seek project-based grants to fund specific orthogonal initiatives related to the consortium. These grants allow us to develop new tools that could not be integrated into the main development plan, expand our data repositories and curation capacities, allow the integration of further community members, and improve user support services. We do not currently charge user fees and have no plans to do so in the future. Our mission is to ensure that financial barriers do not impede access to valuable research data. By securing sufficient funding from research organisations and leveraging in-kind contributions, we maintain a sustainable model that supports free access for all users.

We are implementing the following strategies to ensure the continuity of services and the sustainability of NFDI-wide activities. **Diversified funding sources**: We diversify our funding to include multiple research agencies and project-based grants to reduce reliance on any single funding source. **Regular evaluation and adjustment**: We assess our services and financial health and adapt our strategies to meet evolving needs and the financial landscape. This proactive approach helps us to stay ahead of potential challenges. **Community engagement and advocacy**: We actively engage with the research community to understand its needs and advocate for continued support from funding agencies. Building strong relationships with stakeholders ensures continued relevance and support for our infrastructure. **Collaboration with NFDI**: We collaborate with other NFDI consortia, including Base4NFDI, joining forces and generating synergies to provide a suitable service portfolio, including all required services, without duplicating them.

In summary, our operational and financial model is designed to support an open, collaborative and sustainable research data infrastructure that meets the needs of both users and providers. Through diversified funding, in-kind contributions and a commitment to open access, we ensure our services' long-term viability and relevance.

4 Research Data Management Strategy

4.1 Scientific relevance and quality of the measures

NFDI4Chem aims to provide all scientists with an infrastructure for collecting, managing and publishing FAIR data. At the same time, awareness of this infrastructure and services and the willingness and ability to use them must be established. Therefore, we have drafted a comprehensive strategy to develop and provide infrastructure and services for research data management and training in the chemistry community. Since 2020, we have been pursuing our goals through close cooperation in 6 task areas (TAs) allowing us to work successfully on the measures that we aimed for: TA1 deals with the management of NFDI4Chem, TA2 develops software for digital labs, TA3 establishes and hosts repositories, TA4 develops standards and metadata schemes, TA5 ensures communication with and education of the community, and TA6 provides overarching services for terminologies, (meta)data search and harmonisation.

Visible success of the consortium’s work: The achievements of NFDI4Chem are tangible in different forms and contexts. For the researching laboratory chemist, the digitalisation of chemistry is becoming imminent through the expanding use of electronic laboratory journals and software tools supporting data capture and analysis. Within FP1, we provided generally applicable, extendable, and freely available software and tools covering the whole research data lifecycle, demonstrated for several use cases (recent application: workflow for cyclic voltammetry) (Herrmann et al. 2023). Meanwhile, 70 local instances (at 68 different locations) of the Chemotion ELN are known, spanning academia (instances in German universities and other countries: DE: 41, NL: 5, CH: 4,

FR: 2, other EU: 6, UK: 1, India: 2, US: 1, Singapore: 1) and first industrial users (5 companies known).

Chemists face the next challenge with integrating FAIR data publication into the publication process of research results in a research article. NFDI4Chem has successfully tackled this challenge on two fronts. Firstly, from the data producer perspective, the smart lab supports users in the downstream publication of data in repositories through a user-friendly and seamless transfer of metadata from ELN to repositories of the NFDI4Chem federation such as Chemotion repository, RADAR4Chem and nmrXiv, with no additional burden for the user. In the evolving landscape of chemical research, a suite of repositories has been established to cover all major data types, subdisciplines, and functions in chemistry. The chemistry community trusts and accepts these repositories, with endorsements from leading journals and funding agencies. RADAR4Chem stands out for optimising FAIRness scores of datasets through the FAIRimpact implementation action (Soltau et al. 2024), integrated archiving for more than 25 years, and allowing complex and big datasets. Chemotion has made significant strides with over thirteen frequent releases, continually integrating new features to meet researchers' and reviewers' needs. Even in its pre-release phase in early 2023, nmrXiv already hosted 81 compounds and 490 spectra, demonstrating its value for the community.

Secondly, NFDI4Chem has initiated a broad-based discourse with all relevant chemistry publishers to integrate data publications into the article submission processes. The central aim is to ease the data publishing processes so that researchers can lower the barrier to publishing data. Our two Editors4Chem Workshops were attended by 25 editors, including many editors-in-chief, covering main publishers in chemistry, such as Wiley-VCH, RSC, Springer Nature, ACS Publications, Thieme, Beilstein, PLoS, and MDPI. Projects included the adoption of data availability statements and the recommendation of trusted chemistry-friendly repositories, highlighted in the NFDI4Chem Knowledge Base and implemented e.g. by Angewandte Chemie. Moreover, TA4 brought the idea of Minimum Information Standards into the Community (Herres-Pawlis et al. 2022) and elaborated on standards for NMR, among others. As a molecular standard, the InChI stagnated over many years until NFDI4Chem joined many InChI subgroups, pushing it forward towards opening the core development to GitHub and implementing inorganic chemistry (Herres-Pawlis et al. 2024).

With a perspective on machine-readable data and in preparation for AI applications, TA6 has established a further cornerstone with chemistry ontologies for the semantic annotation of data. The NFDI4Chem Terminology Service (TS) provides 40 high-quality ontologies for both human and service-to-service use, the last one impressively demonstrated by over 400,000 API calls to the TS backend by now. The services and offers of NFDI4Chem quickly find their way into the community through the engaged work of TA5. A dedicated team of trainers delivered 73 workshops on research data management to more than 800 chemists. A newly established knowledge base with now 62 individual articles in English and German was created by 24 contributors and attracts

400-500 visitors per month. Moreover, NFDI4Chem could present the first examples of embedding RDM content into chemistry curricula (Fink et al. 2023).

Review of measures in FP1 and adjustments in FP2: Overall, our general strategy for the organisation of TAs and measures was well-balanced and efficient, so we are proposing the same structure for FP2. Nevertheless, we adapted some organisational aspects of our work in comparison to previous plans: In addition to the individual focus topics in the TAs, we have recognised an increasing need for regular dialogue on the collaboration between all TAs on a joint strategy for the development and implementation of standards to cover the perspectives from data producers to data re-users and the intermediary services over the course of FP1. We, therefore, introduced cross-TA meetings on a biweekly basis, supported by regular workshops. This formerly not planned meeting routine allows us to discuss dependencies of tasks across the TAs, adjust priorities, and accelerate decisions required to plan and conduct the measures.

We have to recognise that standardisation work is complex and is becoming more so. This concerns standards for metadata, data, and ontologies through coordination in our international network with IUPAC, RDA, or EOSC, as well as the standardisation of devices with formats and interfaces. Standardisation will, therefore, continue to be an integral part of FP2 work. With the establishment of NFDI consortia in the second and third rounds, NFDI4Chem consolidated activities with consortia close to its scientific scope. Consequently, the integration of NOMAD, now funded by FAIRmat, shifted within NFDI4Chem. NOMAD is now associated with NFDI4Chem, enabling collaboration and support for the theoretical and computational chemistry community. Similarly, NFDI4Chem will consolidate Suprabank into the Chemotion repository (see M3.4). We will review current and new services for their potential integration to avoid duplication and to ensure compatibility with existing NFDI4Chem infrastructure. We will expand our perspective for the review of potential consolidation candidates to NFDI basic services and the ongoing work on the overarching NFDI service architecture drafted by the NFDI and its sections in FP2. The first steps of our approach are outlined in Chapter 3.2, which includes the NFDI4Chem terminology service and the NFDI basic terminology service. With the establishment of the ELSA section, we will continue the discussion of the legal aspects of RDM at the NFDI level.

In FP1, the consortium gained strong support from previously less-involved chemistry sub-disciplines, including electrochemistry, photochemistry, environmental chemistry, and theoretical and computational chemistry. This support allowed systematic inclusion in all NFDI4Chem areas, formalised by adding new partners and adjusting measures in TA2-TA6 to meet community needs in FP2.

NFDI4Chem emphasises the importance of quality assurance for generating high-quality data and maintaining a trustworthy infrastructure. While initial efforts in FP1 focused on establishing and expanding services and data available, FP2 will prioritise quality assurance measures to raise the maturity and quality of the infrastructure and provide (meta)data. Quality measures will cover data and metadata validation in software services (T2.3.5, T3.3.1, and T3.3.4.), but also underlying standardisation processes for

Minimum Information about Chemical Investigations MIChI and derived metadata schemas (T4.1.1., T4.3.2), the integration of terminologies (T6.1.1, T6.2.2) and the schema validations and crosswalks (T6.3.3).

In the course of FP1, the wealth and granularity of chemical information obtained from all areas of chemistry, particularly through the smart lab, has increased significantly as a result of the systematic digitisation of processes. This information needs to be expressed in rich chemistry metadata to be available in downstream services and data reuse scenarios. We, therefore, adjusted our metadata strategy to support the development of interoperable metadata schemas based on discipline-specific MIChIs (see Chapter 4.2) The newly defined Metadata Schema Service (MSS) will provide the necessary tools to support the creation and application of these schemas ensuring interoperability and mapping to more generic schemas (M6.3).

In summary, the planned measures and adjustments allow NFDI4Chem to provide FAIR data on a large scale and take actions to ensure data reuse in all areas, collectively extending the infrastructure towards machine-actionable and AI-ready data. Building upon standardisation and semantically rich chemistry metadata, the Chemistry Knowledge Graph allows us to harness the potential of machine-actionable data. The expansion of ML-focused expertise in FP2 enables NFDI4Chem to demonstrate data reuse in chemistry-specific use cases and make it accessible for other future partners and models. Furthermore, the integration of AI models can directly create benefits for users of the NFDI4Chem infrastructure (e.g., in ELN or repositories). These incentives for our users will help gain additional supporters in FP2. In the following sections we describe the individual measures and outcomes of TAs in more detail, especially with respect to results that are achieved to prepare our work plan for FP2.

TA1 (Management): based at the applicant institution, manages the technical, financial and administrative processes with the support of the partner institutions. The cooperation agreement of November 2020 serves as the legal basis for the cooperation and the transfer of funds. OpenProject (an enterprise licence that has been purchased together with other consortia) is being used to monitor the progress of work at the consortium level, in addition to a meeting and reporting system that ensures project control. TA1 organises two consortium meetings per year. Additionally, smaller retreats within and between TAs have been effective in driving developments. Four Advisory Boards (National, International, Industry, and Publishers) have been established to provide advice and feedback at the consortium meetings. Communication within the consortium is based on regular meetings, mailing lists, and the chat tool Rocket.Chat. A strategic communication concept was initiated early on, including a corporate design and the web site, which was launched in Q4 2021 and is frequently updated and expanded. It serves as a single point of information for the community and provides an overview of the consortium and all its services.

TA2 (Smart Lab): Chemotion ELN and other tools in NFDI4Chem underwent substantial improvements in functionality and development approaches, facilitating the distribution and fostering their acceptance. Key achievements were reached with the establishment

of a feasible Docker container approach for all institutions using the ELN (KIT Karlsruhe Institute of Technology 2024) and the adaptation of the ELN to the needs of different sub-disciplines. Highlights consist of new entities (Chemotion ELN contributors 2022), modules for inorganic and polymer chemists, and flexible modules (LabIMotion) (Huang and Lin 2024). TA2's software development follows a defined workflow: planning, community consultation, coding, testing, and frequent releases (Chemotion core: 13 releases, NMRium wrapper and further tools: 28 releases in 2020-2024). Feedback via helpdesk and GitHub (582 resolved GitHub issues in 2020-2024) allows transparent activity tracking, requirement discussions and task allocation. TA2 improved and extended device integration into the ELN (Starman 2023), improved data conversion from device-generated files to open, standardised formats with the Converter software, and developed and harmonised software for reading, processing, visualising, and analysing multiple data types (Huang et al. 2021, Patiny et al. 2023). A comprehensive online documentation describes features, methods, and videos for easy reuse of components. The data transfer and publication from the ELN to the repositories Chemotion and RADAR (Chemotion ELN contributors 2021, Tremouilhac et al. 2020) was established, and similar processes for other NFDI4Chem repositories are partly implemented (ComPlat 2024) and will be continued in FP2. The ELN developments are or will also be used by scientists in the consortia NFDI4Cat, DAPHNE4NFDI (Dolcet et al. 2003), and FAIRmat. Additionally, the work on ELN interoperability and data exchange will be continued to enable interdisciplinary work. In FP2, TA2 will add additional focus on enabling the re-use of the data in a machine-readable and actionable way, e.g. for AI projects. This investment will be returned to the ELN users who will benefit from ELN-enabled AI tools, creating incentives through advanced tools that support the scientists' work.

TA3 (Repositories) successfully established a virtual environment of federated repositories that covers all subdisciplines and data types in chemistry with a focus on molecule-related data. The services are feature-rich, interoperable, trusted by the community, and recommended by journals and funding agencies. TA3 integrated important existing repositories that were selected based on quality criteria like functionality for users, suitability, open source accessibility, adherence to standards, and funding requirements. The repositories currently part of the NFDI4Chem federation are listed in TA3. Detailed information is summarised on our website and in Bach et al. (2023). Our knowledge base article How to Choose the Right Repository guides researchers in choosing the right repository for their data type and needs. In addition, we evaluated other international repositories through re3data using criteria (Bonatto Minella et al. 2023) selected by TA3 in collaboration with other TAs. For FP2, we aim to grow the federation, and we have signed LoIs with CSD, ICSD, and CCDC Access Structures Service for crystallography for their integration with NFDI4Chem. All NFDI4Chem repositories were optimised for operational fitness, interoperability, metadata standards (developed by TA4) and harvesting, dataset landing pages and interfaces to services (e.g. NFDI4Chem Search Service). Concepts were developed to connect repositories with scientists' workspaces, enabling data transfer from the Chemotion ELN to the repositories. The repotracker software was developed to record and monitor data transfer

processes. In collaboration with TA4, we decided on minimum information standards (Herres-Pawlis et al. 2022) for the metadata of datasets in the repositories of our federation, based on DataCite. The FIZ OAI-PMH provider and backend software were published open source on GitHub and integration help was made available to repositories. The relevance of TA3's work can be seen in rising data publication numbers and recommendations by journals and funders, e.g., Angewandte Chemie, ChemBioChem, Journal of Natural Products, DFG GSP (in German) (Wiley-VCH 2024, Proteau 2023).

TA4 (Metadata, Data Standards and Publication Standards) focuses on the development and harmonisation of minimum information (MI) standards and metadata for chemical research data, as well as data standards for molecules and reactions, including experimental and theoretical characterisation. We reviewed data standards and file formats (Rauh et al. 2022), and contributed relevant standards and formats to the EDAM ontology and to the FAIRsharing catalogue. In FP2, these can be the basis of chemistry software content in the Base4NFDI project nfdi.software. The development of MIChI guidelines is an ongoing process involving discipline-specific workshops, consensus on journal reporting standards and technical interoperability. In workshops, we have developed MIChIs for polymers and NMR, reports of which are available on the NFDI4Chem website. Based on these, TA2's LabIMotion extension allows the implementation of corresponding documentation templates, which are hosted on a template hub (Karlsruhe Institute of Technology (KIT) 2024) and may become documentation standards. Internationally, we work with IUPAC, the InChI Trust, and TA5 on the WorldFAIR initiative on a FAIR chemistry cookbook. Our involvement in the InChI Trust Organometallics Working Group has led to important developments in the international open molecular representation standard, including a non-disconnection approach. S. Herres-Pawlis currently serves as the leader of the unified InChI organometallics/inorganics working group and is a member of the InChI Trust Board. In FP2, we need to incorporate these improvements on the InChI for inorganics into our services. NFDI4Chem assisted the chemistry community and partners in preparing standards-compliant data publications in the Lead-by-Example measure. They are described in our knowledge base and serve as test data for service development (Fischer et al. 2023). These activities were successful and will be ramped down since self-service training material reduced the need for individual support. The Editors4Chem workshops with IUPAC in 2021 and 2023 involved publishers to integrate FAIR data recommendations into journal author guidelines. Our survey results on author guidelines in chemistry journals show (Parks et al. 2023, Parks et al. 2024) increased focus on research data guidelines. Feedback encourages us to continue with the Editors4Chem activities.

In **TA5 (Community and Training)** we made significant efforts to facilitate a cultural shift towards FAIR RDM within the scientific community. We conducted two major surveys, the second with over 800 participants in 2023, which showed a gradual increase in the use of ELNs and strong support for the inclusion of RDM in curricula (for details, see 3.1 and Herres-Pawlis et al. (2020)). We also gathered community requirements through conference booths and presentations on national and international conferences and

workshops. A comprehensive communication strategy was implemented, including a regularly updated website, newsletters, social media outreach, and monthly digital “Stammtisch” (regulars’ table) meetings for discussion of RDM topics. Extensive training and support initiatives were launched, including the development of the NFDI4Chem Knowledge Base (400-500 visitors per month from all over the world, s. 4.4), bespoke RDM training modules, OER RDM materials and a helpdesk. Those measures with multiplicative effects will remain important in FP2, and the materials flow directly into the section EduTrain and the knowledge-graph-based DALIA platform for OER RDM materials. Curricular integration of RDM content was promoted through practical exercises in labs and lectures, and best practices were highlighted through use cases and flagship labs. The annually presented FAIR4Chem award highlights excellent FAIR data practices from the community for the community. We also initiated blind prediction challenges to improve reproducibility in experimental and computational workflows.

TA6 (Synergies and Cross-Cutting Topics) aims at a holistic use of the NFDI4Chem infrastructure and services with a special focus on data integration and semantics. During FP1, significant progress was made in cooperation with TA2, TA3 and TA4 in developing a joint strategy for ontologies and metadata standards to ensure seamless data integration within the NFDI4Chem service federation. The implementation of the metadata standards is closely related to the search service (TIB-Leibniz Information Centre for Science and Technology 2024a), which harvests and indexes metadata from currently 140,000 datasets in the NFDI4Chem repository federation. Initial harvesting included utilisation of the DataCite and Bioschemas JSON-LD formats. For the NFDI4Chem ontology collection, we conducted a thorough evaluation of ontologies in the chemistry domain and defined selection criteria (Strömert et al. 2022). We are continuously curating ontologies such as CHMO, MOP, RXNO, CHEMINF or the IUPAC Goldbook in collaboration with RSC, IUPAC and the OBO Foundry, and created the new Vibrational Spectroscopy Ontology (VIBSO Strömert et al. 2024c). We organised Ontologies4Chem workshops (Strömert et al. 2022a) in 2022 and 2023, with an average of 50 international participants. TA6 further developed the Terminology Service (TS Jupp et al. 2015 During FP1, the TS was extended to become a central ontology management platform for chemistry ontologies (Strömert et al. 2023). TA6 further ensured NFDI4Chem’s participation and contributions to NFDI measures, NFDI basic services, and international networking, as described in chapters 3.2 and 3.3. TA6 evaluates its measures through KPIs like visits, API calls, and indexed repositories and datasets. The NFDI4Chem ontology collection is assessed against self-defined criteria, and its impact is measured by curation issues and pull requests. We also evaluate Ontologies4Chem output by participants and represented ontology projects. Additionally, published recommendations, best practices, and policies are assessed. In FP2, TA6 will integrate terminologies, the metadata schema service and chemistry knowledge graphs into the Semantic Data Hub as logical steps towards machine-actionable AI-ready chemistry data.

4.2 Metadata standards

NFDI4Chem adopts and develops metadata standards associated with chemical research data, including information about molecules, reactions and processes, as well as data for their experimental and theoretical characterisation. As a fundamental molecular standard, IUPAC's InChI is used to link data across repositories and as identifiers in molecular metadata. Moreover, the InChI key makes molecules searchable through generic search engines. NFDI4Chem has been intensively involved during FP1 in the development of the next version of the InChI by making the InChI FAIR and opening development via a public GitHub repository. The InChI code is guarded by the InChI Trust, of which S. Herres-Pawlis is a board member. For FP2, we will contribute to the implementation of inorganic stereochemistry into the InChI and the PolymerInChI (see T4.2.1). We are active in the European ELIXIR Bioschemas community, where S. Neumann has become co-leader of the chemicals working group. New Bioschemas profiles, such as for reactions, are under discussion to be specified and submitted to become an accepted standard. F. Bach is a member of the Technical Specification & Implementation Group of the international FAIR Digital Objects Forum. We contribute to curating the ELIXIR Toxicology Community collection on FAIRsharing.

Looking at FAIR metadata with a special focus on Findability and Interoperability, NFDI4Chem utilises metadata standards in two ways: The integration of metadata into landing web pages of repositories or other services to ensure findability by data or general search engines utilising common schema.org or Bioschemas standards and the availability of structured, machine-readable metadata using various metadata content and serialisation schemes like DataCite, Dublin Core, DCAT accessible through a standard API like Open Archives Initiative Protocol for Metadata Harvest (OAI-PMH).

At the beginning of FP1, RADAR4Chem and Chemotion Repository provided agnostic DataCite and Dublin Core metadata with limited chemical information implemented via an OAI-PMH endpoint for metadata harvesting. Later on, we extended the XML-based OAI-PMH architecture (using the FIZ-OAI provider) to provide JSON-LD format via OAI-PMH service, where it can be queried both as JSON-LD and via a crosswalk as OAI Dublin Core (Castro et al. 2023). MassBank embeds Bioschemas into the landing pages of the datasets, enabling web crawlers to access them. NmrXiv, the latest addition to the NFDI4Chem repository federation, provides experimental, extended Bioschemas to provide the chemistry-specific metadata. The search service applies different harvesting agents to gather metadata from these repositories and integrate it into the central search index. Additionally, the Search Service API provides all metadata of the federation in DCAT.

Proceeding from this situation, further metadata integration and harmonisation strategy was drafted in collaboration between TA2, TA3, TA4 and TA6. Aspects covered were data structures and data export capabilities of ELN systems and data repositories, as well as the downstream data integration and indexing of the search service. For the various data types in the scope of NFDI4Chem, we facilitate the creation of Minimal Information of

Chemical Information of Chemical Investigations (MIChI), integrating elements of general-purpose schemas followed by domain-specific additions.

In the course of the MIChI development across the subdisciplines of chemistry, we identified and described various data structures with detailed information and the chemical context of the research data. With more and more MIChI evolving, we recognised chemistry core elements across the individual MIChI profiles. This led to the idea of defining a base profile for general chemical information that can be extended with discipline-specific modules. This modular design will enable NFDI4Chem to develop more tailored and adaptable metadata schemas, better serving the needs of various chemistry subdisciplines. With the goal to transform MIChI profiles, including ontological concepts from our NFDI4Chem ontology collection, into commonly used schemas like Bioschemas, we recognised that a significant amount of the newly defined detailed chemistry information could not be mapped adequately or at all. Both schema.org and the extended Bioschemas are, in the end, general schemas or a set of vocabularies with types and properties. In order to be able to meaningfully use the Bioschemas for our needs, it would therefore be necessary to extend it with new types and properties within the limits of its own semantic framework. To avoid these limitations (i.e., losing the expressivity of MIChI) during the integration, we have adjusted our general metadata strategy to harness the full potential of semantically rich chemistry-specific metadata schemas within the scope of chemistry and related disciplines and map to common metadata schemas like Bioschemas for broader dissemination of metadata across further communities to promote data discoverability. To adequately support discipline-specific metadata schemas for the community and metadata interoperability, we will introduce the Metadata Schema Service (MSS) in FP2. The service will fulfil several purposes: it will enable the MIChI teams to transform MIChI profiles into actionable metadata schemas, utilising semantic concepts from the terminology service into the schemas. Acting as a registry, the MSS stores MIChI-based schemas and makes them searchable. At a later stage, the MSS enables crosswalks between schemas, supported by the mapping services that use mappings between concepts.

This strategy enables us to create and apply discipline-specific metadata schemas for chemistry research data. We utilise the full potential of semantically rich data to build chemistry knowledge graphs and apply reasoning on chemistry ontologies with extended logical axioms and still map these schemas to cross-domain ones like DataCite or Bioschemas. On the other hand, the strategy enables us to easily support common metadata standards like DataCite, Bioschemas or schema.org and broader dissemination across disciplines taking into account the recommendations on the Cross-Domain Interoperability Framework (CDIF) (Gregory et al. 2024).

In FP2, we will further focus on machine actionability of data and metadata and follow national (e.g. SMART principles described in the *German Standardization Roadmap on Artificial Intelligence* (Wahlster and Winterhalter 2022), developed jointly by the German Institute for Standardization (DIN), German Commission for Electrical, Electronic & Information Technologies of DIN and VDE (DKE) and Federal Ministry of Economic Affairs and Climate Action (BMWK).

4.3 Implementation of the FAIR principles and data quality assurance

NFDI4Chem endorses FAIR (meta)data across all task areas and services, ensuring consistent implementation through standards, policies and recommendations. Additionally, education and training (TA5) support community acceptance and understanding of the concept of FAIR data.

To ensure **Findability**, all our services aim to provide rich, machine-readable (meta)data linked to domain-specific and cross-domain vocabularies. Our ELN and instrument integration strategy in TA2 promotes early metadata collection during data generation. Most NFDI4Chem repositories register their datasets with DataCite and assign globally unique and persistent DOIs. Most dataset landing pages of NFDI4Chem repositories already provide HTML-embedded metadata following Bioschemas (Neumann et al. 2023) and FAIR signposting (Van de Sompel et al. 2023) to increase findability via search engines such as Google Dataset Search.

To ensure **Accessibility**, NFDI4Chem data are retrievable via persistent identifiers using HTTPS. All components are available under open-access models. For functions requiring login, established authentication protocols like OpenID and Shibboleth are used, with a planned migration to NFDI-AAI. NFDI4Chem repositories support OAI-PMH for programmatic access, and standardised APIs are provided or will be implemented (see TA3).

To ensure **Interoperability**, our metadata and data adhere to existing standards. We address gaps by developing solutions discussed with the chemistry community, such as MIChI-based metadata standards for experiments, simulations, and molecule characterisation. NFDI4Chem promotes and extends open data formats (Rauh et al. 2022), implementing them systematically in our services. All components, from ELN to repositories, automate the collection of standardised data and metadata, ensuring interoperability. We use community-accepted toolkits like RDKit, CDK, and OpenBabel, as well as terms from established ontologies such as CHMO and RXNO. Automated conversion processes store data in standardised formats like JCAMP-DX. Integrated data editors ensure even proprietary inputs are stored in standardised, readable formats, producing interoperable information reusable in other systems, including repositories.

To ensure **Reusability**, our (meta)data have accurate, relevant attributes conforming to domain-specific standards. We use established ontologies and develop our own, such as VIBSO, ensuring data is machine-readable. We prioritise openly licensed, well-maintained ontologies and integrate standards like ROR and GND into NFDI4Chem services. Data are released with clear, accessible usage licences supported by legal policies. NFDI4Chem repositories curate data and metadata to ensure reusability, with curation levels varying. For example, the Chemotion Repository is highly curated, while RADAR4Chem is mainly automatically checked. We also provide targeted datasets for machine learning and other support, facilitating data reuse.

NFDI4Chem ensures data and metadata quality by different **quality assurance** measures.

1. The **Minimum Information for Chemical Investigations (MIChI)** define requirements on how to fully describe the various data types and serve as a reference.
2. NFDI4Chem services capture or process data according to schemas derived from MIChI profiles. All data are **captured in a structured manner**, so both the presence of the data elements and their values can be verified. Wherever possible, data is collected applying subject-specific standards (e.g., MolFile and InChI for structures), allowing inputs to be assessed at least formally and often also in terms of their values. Additional tools (e.g., structure editors) are often required for data entry according to these standards; these tools are integrated into NFDI4Chem services, and their use within the services is harmonised.
3. NFDI's services include automated data audit protocols. This allows the structured and standardised data to be tested for readability and reusability through processing and validation. By incorporating cheminformatics tools such as CDK or RDKit, the data captured or provided in the services (e.g., ELN or repositories) can be verified through simple plausibility checks using various inputs. Special checks can be carried out in, for example, Chemotion or nmrXiv, where developed tools compare the obtained experimental data to, e.g., simulated data. These and other verification procedures are particularly accelerated and improved in FP2 due to the rapid development of AI and the involvement of relevant experts in NFDI4Chem.
4. NFDI4Chem services **support multi-level data curation**. Curation in NFDI4Chem repositories is partly organised by subject-specific and organisational responsibility (RADARChem) or by manual curation of the complete dataset by a fixed team of curators in Chemotion. All repositories already support or will support curation and commenting by the community, allowing for the control and correction of published data. This is accompanied by systematic versioning of datasets and metadata, either in whole or in parts.
5. Users of NFDI4Chem services are supported by extensive training, primarily through TA5, preventing misuse, incorrect data input or annotation, and thus improving data quality. The knowledge base offers general data preparation guidance, while individual services provide specific instructions according to chemical community standards.

4.4 Services provided by the consortium

Following the definitions in Amelung et al. (2023), these services are developed and provided by NFDI4Chem:

The **NFDI4Chem Knowledge Base** (N4C-KB) (NFDI4Chem 2024d), launched in late 2021, involves 21 contributors and offers various entry points based on the viewer's discipline, role, or specific interests. It covers a wide range of topics, from basic RDM

concepts to more in-depth articles. The N4C-KB assists users in selecting the right data repository for their research data needs. Hosted at JGU and built using the open-source framework Docusaurus, the platform allows all content to be stored in a GitHub repository using simple Markdown syntax. This approach makes it easy for authors to contribute without requiring web programming skills. The website is automatically updated with every change to the repository. The N4C-KB team actively supports contributors and accepts content in a variety of formats. Since 2023, N4C-KB has employed Matomo for privacy-friendly web analytics, showing that it is accessed by 400-500 persons per month.

The **Terminology Service** (TS) (TIB-Leibniz Information Centre for Science and Technology 2024b), hosted by the TIB, is a comprehensive repository and curation platform for ontologies, terminologies and vocabularies in chemistry and related disciplines. As of August 2024, it contains 40 terminologies, selected based on criteria established by our Ontologies4Chem overview (Strömert et al. 2022b) and community approval (Strömert et al. 2022a). The TS provides advanced search, browse and access capabilities within these terminologies, providing rich metadata and information. Users can explore terminologies through graph or tree visualisations and access development and curation features. It facilitates connections to original terminology repositories, enabling term requests and comments as a step towards a comprehensive terminology curation platform. The TS plays a pivotal role in generating semantically annotated, machine-actionable data and provides a comprehensive API (TIB-Leibniz Information Centre for Science and Technology 2024d) for other NFDI4Chem services to integrate terminologies into their data annotation workflows, such as ELNs or data repositories. In addition, NFDI4Chem's Ontology Elements web components Venkata et al. (2023) provide an easy way to implement semantic annotation widgets using these terminologies. The TS is part of the basic service TS4NFDI backend architecture.

NFDI4Chem drives the development and establishment of **ELNs** (NFDI4Chem 2024c) as a key requirement to achieve systematic digitalisation. While the developed ELN software is offered to users as source code to be hosted locally, there are three additional ELN-based services within NFDI4Chem:

1. For IT staff and admins, a Docker container service is provided to easily install the ELN with all required dependencies (KIT Karlsruhe Institute of Technology 2024). Management of single instances or multiple instances is supported by providing a command line interface (Chemotion ELN contributors n.d.).

2. NFDI4Chem offers the hosting of four (five until the end of 2023) ELNs as test instances that can be easily used by scientists looking for the right ELN solution. From more than 60 available ELN software tools, the four (five) most important OS solutions for chemists have been selected and are hosted for testing, comparison and educational purposes (NFDI4Chem 2024f).

3. Chemotion ELN will be hosted as a service for individual researchers and small groups. The service is currently being set up. Initial pilot use cases with individual

users and small groups are already underway in preparation for the full service. The service includes the migration of content to a local instance when a sustainable user group size is reached - which could be successfully completed with a first pilot user group in 2023.

The **federation of core repositories** (NFDI4Chem 2024b) comprises seven German-hosted repositories, each covering essential content in key subdisciplines of the chemical community. These repositories are developed and provided as individual services tailored to specific discipline-specific processes and functionalities driven by their respective communities. Within the NFDI4Chem federation, existing repositories adapt, and emerging repositories develop workflows, standards and functionalities to create a harmonised data infrastructure. This infrastructure aims not only at data interoperability but also at collaborative interaction between repositories and other NFDI4Chem services. The selected core repositories can handle different data types, chemical processes, analytical data and specialised methods, thus supporting the entire data landscape. The roles and requirements of these core repositories vary, with the first funding period focusing on strengthening existing repositories through strategic source code improvements for efficient development and stable hosting.

To date, five NFDI4Chem repositories are in operational use. Of these, we describe RADAR4Chem, Chemotion repository, nmrXiv, and MassBank EU in more detail, as they are currently relevant to the widest user community, and major changes have been released.

RADAR4Chem (FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH 2024) is a feature-rich repository covering complex datasets from all domains of chemistry. Launched in March 2022, it offers free, versatile publication and storage options for any kind of chemistry-related data type. It has been developed by adapting the existing RADAR service (DFG project 2013-2016). RADAR4Chem is crucial to filling gaps in discipline-specific repositories, allowing the publication of complex, heterogeneous, and large datasets, as well as long-term data preservation. A seamless data transfer from the Chemotion ELN to RADAR4Chem allows the publication of data collected in the ELN with just a few clicks.

The **Chemotion Repository** (Karlsruhe Institute of Technology (KIT) 2024b) deals with data related to chemical reactions and chemical substances and was established at KIT in 2015, initially serving a narrow range of scientific data. Since then, it has evolved within NFDI4Chem to meet different scientific needs. It serves as a pilot for data transfer from ELN to repositories and will be adapted to manage data transfer to other suitable repositories for interoperability. Key milestones have been achieved in 2022 and 2023, including the development of tools such as the repo-tracker and repo-downloader software. It provides enhanced reporting capabilities and supports the publication of different data types with discipline-specific templates. NFDI4Chem funds its development, and the content of the repository is curated by in-kind contributions.

The **nmrXiv** repository (nmrXiv Project Group 2024), hosted at FSU, is a new NMR spectroscopy data repository and analysis platform built from the ground up. It builds on the experience of its predecessor, nmrshiftdb2. nmrXiv is open, FAIR and consensus-driven, preserving both raw and processed NMR data. In July 2024, shortly after its release, it already contained 459 compounds with 2704 spectra organised in 45 projects. It provides DOIs, web UI and REST APIs (Open API, DataCite, Bioschemas, NMRium). nmrXiv follows the DataCite metadata schema, enhanced with InChI and SMILES. It uses two-factor authorisation and single sign-on with popular social network logins, including ORCID. Storage capacity is provided in-kind by FSU. nmrXiv is recommended for the deposition of NMR data associated with publications by a growing number of journals, including Angewandte Chemie and the Journal of Natural Products of the ACS.

MassBank EU (MassBank consortium 2024), hosted at the UFZ, is the first public repository of mass spectrometry data, facilitating its sharing with the scientific community. Since 2021, its compound dataset has grown to 15,075 (from 14,788) and its spectra to 90,190 (from 86,576) in 2023. MassBank uses GitHub for AAI (open read access, limited write access) and uses GitHub issues for curation tracking, which are managed by the MassBank record validator. Spectral data and metadata are stored in a human-readable record format within a revision control system, and continuous integration ensures record integrity with each change. NFDI4Chem funding has enabled a modern software overhaul, with a first development release using a JS-based front-end and a REST-based back-end.

Suprabank (Biedermann Labs 2024) is a specialised database, hosted at KIT since 2019, offering unique data on intermolecular and supramolecular interactions. It primarily addresses supramolecular and physical chemists as well as biologists in organic chemistry, focusing on binding, assembly and interaction phenomena not found in other repositories.

STRENDA DB (Beilstein-Institut zur Förderung der Chemischen Wissenschaften 2024), established in 2016 and operated by BI, is a well-established repository for enzymology data. It collaborates with over 55 international biochemistry journals and has integrated the STRENDA guidelines into its author instructions. The database ensures the completeness and validity of enzymology data prior to submission for publication. It primarily contains functional enzymology data, including kinetic and experimental data. STRENDA DB is an in-kind contribution.

In addition to the services available in production mode, the **VibSpecDB** repository, focusing on Raman and IR spectra, is under development and will publish the first prototype in FP1.

NOMAD (Scheidgen et al. 2023) plays a pivotal role in theoretical and computational chemistry, bridging the gap to material sciences. Although they are no longer represented by an institutional partner in NFDI4Chem, we continue to collaborate closely to evaluate data exchange interfaces and co-design functionalities (see LOS NOMAD).

This cooperation will significantly benefit both communities, facilitating interdisciplinary use cases and advancing scientific research.

The **Search Service** by TIB (TIB-Leibniz Information Centre for Science and Technology 2024c), which was released in the summer of 2022, acts as a central hub for searching the federated repositories of NFDI4Chem. It currently includes ~140.000 datasets from the Chemotion Repository, MassBank, nmrXiv and RADAR4Chem. The integration of the chemistry sub-collection of DaRUS marks the first integration of datasets from a generic data repository. The service regularly harvests and indexes metadata, handling different metadata models and protocols It enhances them with chemical metadata, enabling searches by chemical structure codes like InChI, molecules and analytical methods. The metadata and search capabilities are continuously adapted to the latest MIChi and metadata schema releases to increase the granularity of discipline-specific metadata.

The **NFDI4Chem Helpdesk** (NFDI4Chem 2024a) serves as a central hub for community requests. It provides efficient support for all NFDI4Chem services and RDM topics. Basic issues and common questions are handled by first-level support, while specialised second-level teams of the corresponding services handle complex cases. The Helpdesk streamlines communication with our user community, collects common queries in the N4C-KB for proactive solutions and is hosted by TIB, supported by teams from JGU, FSU, TuBr, KIT and RWTH.

NFDI4Chem offers **workshop series on general chemistry RDM and Chemotion ELN / LabIMotion** (NFDI4Chem 2024h, NFDI4Chem 2024i, NFDI4Chem 2024j), which can be tailored to the needs of working groups and institutions. During FP1, 73 workshops were conducted with over 800 active participants (for details, see 3.1).

The **Metadata Schema Service** will be newly developed in FP2, acting as a repository for MIChI-based metadata schemas. The service will enable users to create, store, browse and re-use schemas. A corresponding API allows services to perform metadata validation and schema crosswalks.

The **Chemistry Knowledge Graph** will be a new service developed in FP2. The knowledge graph will be fed by the search service providing harmonised, semantic rich data streams from the NFDI4Chem service federation. Besides SPARQL endpoints the service will provide an easy-to-use visual interface to query Chemistry Knowledge Graph.

5. Work Programme

The consortium's work is divided into six primary task areas (TAs) with several measures each (see Table 1), focusing on specific aspects of the infrastructure development and implementation (see Fig. 3 below).

TA1 Management oversees the administrative, financial, and technical management of the consortium. It coordinates consortium activities, manages funds, ensures compliance with legal and ethical standards, and facilitates communication among partners.

Table 1.
Overview of the six task areas (TA) of NFDI4Chem.

Task area	Measures	Responsible Co-spokes- person
TA1 Management and Coordination	**M 1.1**: Overall legal, contractual, ethical, financialand administrative management of the consortium **M 1.2**: Coordination at consortium level of the technical, outreach, training activities and their future sustainability **M 1.3**: Coordination of long-term knowledge management, internet publishing system and other innovation-related activities **M1.4**: Overseeing the promotion of equal opportunity in the project	Christoph Steinbeck
TA2 Smart Laboratory	**M 2.1**: Device integration and management **M 2.2**: Establishment and maintenance of the electronic lab notebook **M 2.3**: Viewers, processors, and editors for structures and data **M 2.4**: Development and extension of topic-related ELN functions **M 2.5**: Services provided to the community **M 2.6**: ELN as part of a digital ecosystem	Nicole Jung
TA3 Repositories	**M 3.1**: Enhancement and sustainable operation and maintenanceof the federation of chemistry repositories **M3.2**: Advanced FAIRification: optimise interoperability,data reuse, AI-readiness and user experience **M3.3**: Consolidation, Harmonisation and Standardisationof the federation of repositories **M3.4**: Developing and implementing sustainable service operating models **M3.5**: Extension of the federation of repositories	Felix Bach
TA4 Metadata, Data Standards, and Publication Standards	**M4.1**: Development and Harmonisation of Minimum Information Metadata Standards **M4.2**: Development and maintenance of standards for data exchangeand archival **M4.3**: Implementation and support of software components for creation, validation, and consumption of standardised data formats **M4.4**: Improving re-usability of scientific data through data standards **M4.5**: Integration with scholarly publishing	Steffen Neumann, Christoph Steinbeck
TA5 Community and Training	**M5.1**: Community requirements **M5.2**: Awareness **M5.3**: Training and support **M5.4**: Curricular teaching **M5.5**: Best practice **M5.6**: Community stakeholders	Sonja Herres-Pawlis, Johannes Liermann
TA6 Synergies and Cross-Cutting Topics	**M6.1**: Ontology development, curation and harmonisation**M6.2**: Terminology service, mapping service, pattern/shape service**M6.3**: Semantic data hub**M6.4**: NFDI4Chem within the NFDI and basic services**M6.5**: Integration into the landscape of existing infrastructure and initiatives**M6.6**: Enable machine learning	Oliver Koepler

TA2 Smart Lab develops and implements digital tools and environments for efficient data capture and management in laboratories necessary to capture data early in the life cycle. It creates and enhances electronic lab notebooks (ELNs), integrates laboratory

instruments, and develops workflows that facilitate seamless data transfer and interoperability within the infrastructure.

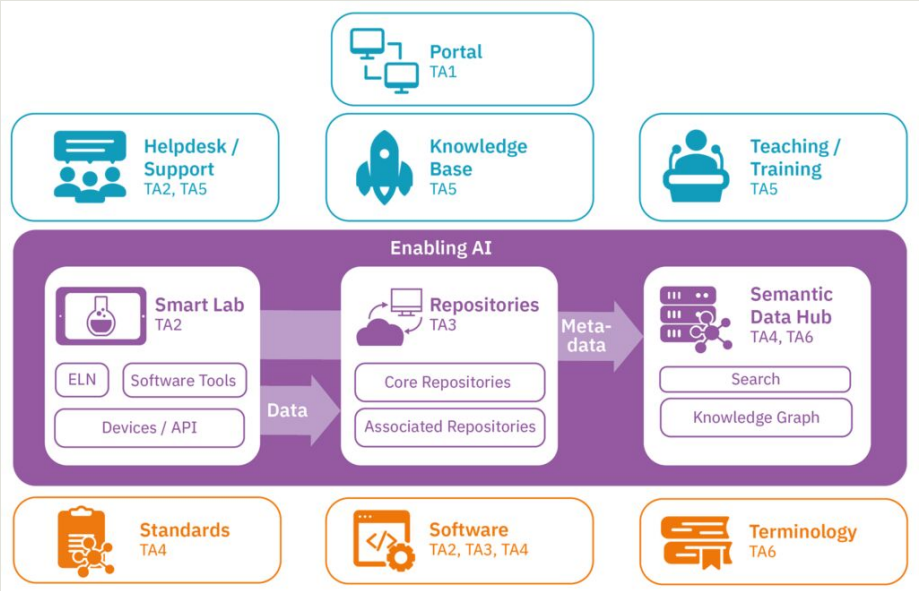

Figure 3. doi

The role and relationship of services and task areas in the user-facing (top), underlying (bottom) and infrastructure layer (middle) of the NFDI4Chem work program.

TA3 Repositories develops and evolves a federated system of repositories to store, share, and preserve chemical data. This includes raw data in diverse formats as well as curated datasets. It develops and maintains core repositories like RADAR4Chem, Chemotion, and nmrXiv, ensuring they meet the needs of various subdisciplines in chemistry. It facilitates data deposition, retrieval through standardised protocols, and re-using research data across distributed data services.

TA4 Metadata, Data Standards, and Publication Standards defines and implements standards for metadata and data formats to ensure consistency and interoperability, together with reference implementations and data validation. Ontologies are used where possible, and missing terminological artefacts are added. It develops minimum information standards, harmonises metadata practices, and collaborates with international organisations like IUPAC to adopt and foster global standards.

TA5 Community and Training fosters a culture of effective data management practices within the chemical research community. It conducts training workshops, develops educational materials, and engages with researchers to promote the adoption of RDM practices and offers incentives for innovations. It is present at chemical conferences with booths and talks and maintains multiple active communication channels to disseminate information and collect the community's requirements, needs and feedback.

TA6 Synergies and Cross-Cutting Topics develops and provides the Terminology Service, Metadata Schema Service, Search Service and Chemistry Knowledge Graph. It develops, curates ontologies and promotes their use in conjunction with semantic technologies to enhance data interoperability, machine-actionability and AI-readiness. Through these means, TA6 ensures the harmonisation of various components of the NFDI4Chem infrastructure. It further coordinates with other NFDI consortia, sections, and working groups on cross-domain metadata standards, cross-domain development, ontologies mapping, and terminology services development. TA6 ensures collaboration with international bodies such as the Research Data Alliance (RDA) or the International Union of Pure and Applied Chemistry (IUPAC) and relevant projects like the Physical Sciences Data Infrastructure (PSDI) in the UK.

The NFDI4Chem consortium is making significant progress in building a national research data management infrastructure for chemistry in Germany. These efforts are already advancing chemical research and ensuring that data are managed efficiently, shared openly, and utilised to their full potential.

5.1 TA1 Management

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, balance multiple responsibilities and competing priorities of consortium partners, prompt reporting and finally, successful project delivery. The overall management structure is illustrated in Fig. 3. 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 administrative management, compliance with 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's progress to meet the project objectives and handle time and resource constraints appropriately.

Prof. Steinbeck heads the PO and consists of the project management team with a project manager exclusively hired for this project, 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 guarantees adequate administrative project control, coordination of reporting, and care of financial and budgetary matters.

In the consolidation phase of the project, TA1 will focus on integrating the progress and achievements of the past years into a cohesive framework that ensures sustainability and long-term impact. This involves harmonising the work of various TAs, streamlining processes, and reinforcing the collaborative infrastructure established during the initial phase. By consolidating our efforts, we aim to enhance the efficiency and effectiveness of project activities. Additionally, consolidation will facilitate the transfer of knowledge and best practices across the consortium, enabling continuous improvement and innovation. This phase is crucial for maintaining momentum and preparing for future project extensions, allowing us to address new challenges and opportunities with a robust, unified approach.

To summarise, the **objectives** of TA1 are:

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

O1.2: Organise and document all NFDI4Chem services, consortium, advisory board, stakeholder meetings and decision-making processes, and 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, exploitation and monitoring of the project results.

These objectives will be pursued through the following **measures**:

M1.1: 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 and the DFG regulations.

T1.1.1 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 set out the rights and obligations between the partners.

T1.1.2 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 the timely and correct transfer of funds to all partners. We will also collect all information for the reporting required by the DFG.

T1.1.3 Maintain communications with the NFDI headquarters 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.

M1.2: Coordination at consortium level of the technical, outreach, training activities and their future sustainability

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 and harmonise the work of the TAs.

T1.2.1 Organise consortium and TA meetings and harmonise work among TAs

As part of this task, the PO will organise weekly tele-meetings of the NFDI4Chem steering committee. Discussions and decisions will be minuted and, if applicable, communicated to the consortium or other partners involved. We will invite national and international collaborating PI's to participate if needed. Technical teleconferences of the TA participants and leads will be held separately and likewise individually documented. Those meetings will be organised by the respective project manager of the TA. Joint meetings of two or more TAs will increase in number and intensity in order to harmonise the work of the TAs and ensure effective long-lasting cooperation efforts in overarching topics. For day-to-day communication among NFDI4Chem participants, a chat platform is used, which is self-hosted at FSU to ensure cost-effective and sustainable use of this communication platform.

T1.2.2 Organise and document advisory board meetings for NFDI4Chem

The composition of the Advisory Boards (ABs) as formed during the first funding period will be critically reviewed and, if necessary, adapted. We will maintain communication with the Advisory Boards through meetings which will be held at least 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 neighbouring consortia as well as used to steer the future directions of NFDI4Chem.

T1.2.3 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 quarterly flash reports and in the annual meetings, and will be fed into the reporting system described in Task 1.3.3. The project plan 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 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.

M1.3: Coordination of long-term knowledge management, internet publishing system and other innovation-related activities

Goal: The chemistry community, the consortium, the NFDI as a whole, and other stakeholders 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 NFDI4Chem central web portal, service portals, 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 in the long term.

T1.3.1 Maintain and host an informative portal

The NFDI4Chem portal at https://www.nfdi4chem.de/ will be maintained in close coordination with the project's community and training activities (e.g., M5.2 and M5.3). **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, the Knowledge Base (see T5.3.1), which contains documentation, resolutions, workarounds, and best practices for the NFDI4Chem services and tools, which support both helpdesk staff and users, will constantly be expanded. Each tool will be well documented and accompanied by didactically structured online manuals and multimedia web tutorials that allow our users to learn how to use our tools themselves. This ensures the long-term usability of our tools when the expected increase in the number of users can no longer be fully instructed through in-person training.

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. calendar, portal searches, as well as advanced analytics, functional testing, and communication via mailing lists.

T1.3.2 Maintain full documentation of all NFDI4Chem activities

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 and long-term availability for our target communities.

T1.3.3. Develop and maintain a sustainable reporting system

A sustainable reporting system will be established to continuously improve the consortium's services, monitor our own governance structure and work progress, and meet the increasing quantitative evaluation requirements from external stakeholders. To this end, automated read-out of indicators will be established for all NFDI4Chem services. The data will be aggregated into a database set up in close collaboration with the NFDI e.V. and all other consortia. A central publicly available library with all written publications, media and software items of NFDI4Chem will allow consortium members and our target audience to view the consortium's output at a glance.

M1.4: Overseeing the promotion of equal opportunity in the project

Goals: Ensure that minorities, e.g. 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 underrepresentation of female developers and researchers. The promotion of equal opportunity for minorities in science is, therefore, an important component of NFDI4Chem management.

T1.4.1 Promote and optimise 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.

Table 2

5.2 TA2: Enabling digitalisation for chemists: Smart Lab

The overarching aim of **TA2** is to develop and provide a modular lab environment composed of concepts, services, and software to fully digitalise research data management. The key aspect of work in TA2 is a modular approach to the software design with open, well-described interfaces, enabling a flexible combination of all developments with different NFDI4Chem or external services. In FP1, most of the basic features for the envisaged Smart Lab concept have been identified, designed, developed and deployed. In FP2, we will improve, adapt, extend, and add to the bundle of these developments to complete and optimise the Smart Lab functionalities. We will continue to establish a robust infrastructure, including state-of-the-art scientific tools, and will implement concepts for continued updates with less dependence on NFDI4Chem and institutional funding. TA2 includes all software developments and solutions that are close to the community in the sense that they form the work instruments that are used for the scientists' daily work. To this end, work packages from the other TAs are implemented within TA2 and its work spans the full range of solutions: data acquisition from different

devices, collection of experimental and computational data, capture of metadata, management and analysis of data, and finally, its storage in repositories. With this, TA2 aims to solve the technological challenges that currently hinder digital availability, storage, and seamless data transfer in chemistry laboratories. Further, enabling the transfer of FAIR data to repositories and databases through standardised interfaces, a prerequisite for automated data processing applications, promotes innovative applications such as developing machine learning (ML) approaches. Smart Lab with full device integration requires decentralised architecture and institutional hosting, though it can also take the form of a centrally hosted virtual lab environment i.e. ELN-as-a-service usage model, if institutional hosting is not viable.

Table 2.

Potential risks of the TA1 work programme and measures for their mitigation.

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. **Likelihood**: Low	Simple and tight structure in the proposal allows straightforward assessment. Frequent internal electronic and video communication of the SC,quarterly flash reports, the annual review process at the consortium meetings and review by the NFDI and DFG, which will track progress. Ultimately, the governance structure has been optimised throughout the first funding **TA(s) involved**: All
RM 2: 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. **Likelihood**: Medium	The project will need to have enough time between the funding decision and the start of the 2nd funding phase in order to prolong contracts of existing staff or advertise new positions where applicable. A long-term career perspective in the NFDI system would increase the attraction of jobs in the consortium but is a matter of political decisions. **TA(s) involved**: all.
RM: Budget cuts. Due to budget cuts, we can not onboard additional stakeholders as partners, limiting the expansion of the consortium and will lose highly trained, dedicated staff. **Likelihood**: High	We will need to prioritise measures and tasks and decide on measures to be dropped to ensure the overall stability of operations. **TA(s) involved**: all

TA2 was and will be deeply embedded into the work of NFDI4Chem: TA2 strongly collaborates with the services in TA3 to provide the data in a FAIR way, adapted to the specific needs of the repositories. It provides fundamental concepts and information for TA4 to accelerate MIChI development and modelling of metadata (see M4.1), and it implements the results of TA4 in the form of standardisation processes. TA2 feeds its new developments and solutions into TA5, providing input for teaching and training materials (see M5.3). Finally, TA2 implements the work of TA6, e.g., by translating ontologies into ELN workflows. All these measures are and will be elaborated in close collaboration among participating scientists, developers, and institutions. At present, TA2 gets its requirements from and designs the suggested concepts with the support of scientists from the subdisciplines of

1. organic,
2. inorganic,
3. polymer,
4. physical,
5. medicinal and pharmaceutical,
6. biological, and
7. analytical chemistry. The subdisciplines
8. photo- and electrochemistry,
9. environmental chemistry, and
10. theoretical chemistry became strongly engaged during FP1 and will be especially strengthened through tasks in FP2.

TA2 will contribute to the key objectives KO1, KO2, KO3, and KO6. More specifically, the **objectives** of TA2 are:

O2.1: Development of a digital research environment (Smart Lab) that provides all functions to collect, manage, store, analyse, and share data. The Smart lab includes an ELN as a modular platform for all subdisciplines in chemistry that allows for the integration of devices, use of external resources, and transfer of data to repositories.

O2.2: Creating feedback-driven, easy-to-use software solutions that accelerate scientific work and foster the use of the NFDI4Chem infrastructure by scientists.

O2.3: Support the provisioning of high-quality FAIR data to enable reuse cases that scientists can benefit from and which generate incentives such as data comparison, systematic analysis, and aiding AI initiatives.

O2.4: Making the components of the Smart Lab available to all scientists in Germany and beyond who currently lack suitable means to make their data FAIR.

Fig. 4

M2.1: Device integration and management

Goals:

1. Extend the applicability of our existing concepts to integrate scientific devices into a digital workflow using modular, open-source solutions available to all research institutions.
2. Inclusion of standards for data management, such as LADS to reduce customisation.

Description: Devices play a pivotal role in the experimental work environment of a chemistry laboratory by conducting experiments, recording data, and analysing results. In the FP1, many device types, including legacy devices, have been made accessible, such as those for HPLC, NMR, UV, or IR, MS, and many others (Jung et al. 2024). The developed solutions were made available by provisioning software at different

universities and documenting the processes (Herrmann et al. 2023). In FP2, we will extend the available methods to meet the requirements of techniques in different fields of chemistry. We will refactor and extend the established mechanisms, simplifying integration, improving stability and increasing the performance of these mechanisms to transfer data (T2.1.1). New methods allowing the transfer and inclusion of information from large datasets need to be designed and developed (T2.1.2). With regard to user applications, we will allow LIMS-like data management and remote control of devices by a comprehensive interface (T2.1.3).

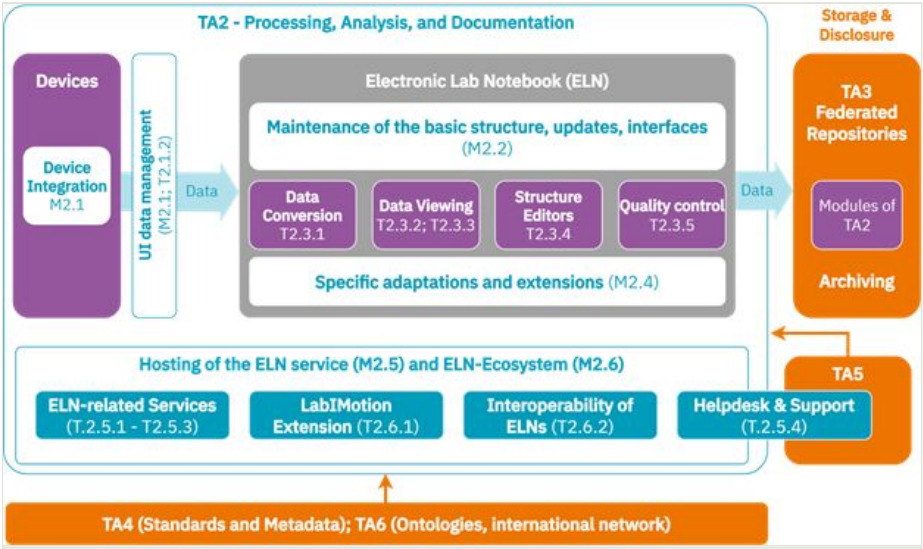

Figure 4. doi

Components of TA2, Smart Lab, and its connections/interactions to TA3-TA6. Orange: other TAs of NFDI4Chem; purple: modular, general software plugins, usable in diverse NFDI4Chem services; white: software developed for the ELN; blue: tasks related to the smart lab but with considerable content from other stakeholders/projects.

T 2.1.1 Data availability and data transfer

A suitable data transfer strategy to access and transfer data to an institutional server in a flexible and extendable manner was developed in FP1. In FP2, the approach will be improved to implement a less resource-intensive process that extends the capabilities of data transfer while augmenting the functionality of the current system. The middleware, which is used to collect data before assigning it to an ELN, will be enhanced so that it makes use of the ELN's API to transfer the data directly. We will continue to develop data transfer solutions for all required devices in chemistry, describing the individual process and the generic approach that allows the reuse of developed protocols, strategies, code, and documentation. The efforts will be prioritised based on the evaluation of user requests. In parallel, we will support data transfer standards (OPC Foundation 2024) recommended by diverse stakeholders in academia and industry - allowing for simplified solutions for an increasing number of devices in the long run.

T2.1.2 Metadata strategies for large data

Current strategies for the work with data and metadata in TA2 of NFDI4Chem are based on data which is of small to medium size (up to 1GB per data file), but increasingly, scientists need a systematic strategy to also manage large data (several GBs to TBs per dataset) in a suitable way. Currently, this is problematic in terms of resource intensity and performance. A solution is to link large data to the original source and a parallel approach to capture the relevant metadata. Such an approach needs concepts that enable the extraction of metadata (in close interaction with M4.2) and the provision of selected metadata (files) depending on data type, instrument, and its software. In FP2, the OS software ChemConverter (work of FP1) will be systematically extended to support the required routines for large data either on local institutional instruments or external clusters in FP2.

T2.1.3 D ata management

Within FP1, routines to monitor new incoming data on the registered data server(s) and its subsequent assignment to responsible scientists and the semi-automatic assignment to experiments in a lab were implemented. As a result of user feedback in FP1, options to transfer data collected using mobile devices were developed. In FP2, we will

- generalise and consolidate our current strategies to extend the capability from our first use case (i.e. scale recording) to different techniques (e.g. elemental analysis date),
- build a data manager directly in the ELN, and
- introduce a sophisticated UI that can be used to plan, record, and visualise data analysis workflows for samples.

M2.2: Establishment and maintenance of the electronic lab notebook

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 management of the ELN development team and the code contributions.

Description: Within M2.2, the core software of the ELN Chemotion will be improved, extended, and continuously upgraded. This measure was the basis of further work on the ELN in FP1 and will remain an important aspect of the work in TA2 in FP2 as a prerequisite for an up-to-date secure codebase. Additionally, continued maintenance of the core system greatly enhances its stability, flexibility and performance in the long run. The work within M2.2 will include the tasks that need to be developed and supported in a general manner to ensure the success of all ELN modules. The results will be parts of the main software i.e. components that are adaptable and usable in different scenarios. The seven tasks are designed to achieve sustainable software, especially T2.2.1 and T2.2.2, which is necessary for a code of the size and importance of Chemotion ELN. We also aim

to deliver present and future (during FP2) user feedback, with features already in the design stage.

T 2.2.1 Basic ELN structure, coordination of functions and final code

In T2.2.1, the basic ELN structure will be maintained by planning and incorporating changes and adaptations, and the work of all tasks related to Smart Lab topics will be managed. The key aspects of T2.2.1 are the coordination and harmonisation of the activities of all developers at different sites in Germany, the planning of resources, the assignment of coding tasks, the organisation of code review and the supervision of the quality of the code. During FP1, a systematic approach of accepting code into the main codebase after review and then following its impact on a subset of users (alpha/beta testing) was put in place. In FP2, we will improve this process by using the industry standards of continuous integration and continuous delivery (CI/CD). Code is managed publicly via GitHub, including not only communication and planning between developers but also the collection of feedback from users and tracking of bugs & issues. T2.2.1 ensures the onboarding and training of programmers within the group as well as maintaining ELN's user-facing documentation reflecting the changes and new features.

T2.2.2 Maintenance, upgrades and updates

The ELN codebase has dependencies in the form of other OS software and libraries, such as the backend framework called *Ruby on Rails* and various frontend JavaScript libraries, which are used to buttress the usefulness of the ELN. These need to be maintained, routinely upgraded and updated to ensure the software's functionality, security, longevity and performance. The steps for the required upgrades (at least two per year) include automated monitoring and, when possible, fixing of security issues as well as a manual intervention to ensure proper integration and functionality, integration of major changes to the codebase, e.g. those brought in by the addition of new features, thorough testing, as well as additional maintenance tasks to ensure the stability and security of the ELN. Additionally, the ELN relies on the database software PostgreSQL, for which updates should be applied promptly, including reviewing and potentially modifying queries and schemas. Those changes need to be transcribed into the ELN's codebase, and the related automation scripts need to be provided to the ELN user/ administrator community.

T2.2.3 Coding of extensions for basic functions

The basic functions for the ELN have been continuously improved and extended during FP1, e.g., with single sign-on options (v1.6.0 Chemotion ELN contributors 2022), a simple Python API client (Starman et al. 2024), seamless integration of the LabIMotion extension (v1.8.0 Chemotion ELN contributors 2022), and an advanced search interface for users (v1.9.0 Chemotion ELN contributors 2022). In FP2, the work plan includes

- extending the admin interface into a more fine-grained roles-based model to support diverse AAI functions,

- adaptation of the ELN according to the needs of practical courses and other special environments,
- implementing a comprehensive tracking system,
- designing, developing, and implementing an additional ELN-embedded database for measurement values enabling enhanced searchability of analytical results and better provisioning of data for AI models.

These subtasks are dictated by users' requirements which we continuously continue to monitor in FP2 to include newer requirements into the development strategy with the support of TA4 and TA5.

T2.2.4 Rework of UI and UX

An intuitive user interface (UI) and a well-designed user experience (UX) are the keys to successful software. During the FP1, the functions and representations within Chemotion ELN increased in almost all aspects, bringing complexity and new UI constraints that need to be resolved to improve UX. Due to the already visible limitations of the current UI framework and, again, driven by user feedback, we started to upgrade the existing libraries and designed a new UX concept with the support of external UX experts for the most relevant aspects of the ELN. The implementation of the resulting design and extending its scope to all areas of the ELN, alongside continuous fixing of UI and UX issues, will be part of FP2.

T2.2.5 Interfaces

APIs allow the ELN to communicate with other internal and external services using a well-documented versioned schema. Chemotion ELN already uses different APIs to exchange data with additional toolkits (Biostats, PhysStats, TLC-App), databases (PubChem, SciFinderN, CAS Common Chemical API), and services (TIB's terminology service, RADAR4Chem repository, Chemotion repository, CrossRef). We will have four main subtasks in T2.2.5 during FP2:

1. Consolidation of all existing and new interfaces to a strategic subtask within T2.2.5,
2. inclusion of new APIs to provide functionalities and information in the ELN by the access to third-party applications (e.g. Jupyter notebooks and Shiny apps), the interaction with theoretical chemistry workflows (e.g. DFT calculations), the interaction with databases (examples include PDB and CSD), and the connection to AI-based tools and services.
3. provision of a comprehensive Python API Client for Chemotion ELN so that the ELN's content and data can be accessed for AI projects (in close interaction with M4.4), and
4. further support APIs from ELN to repositories e.g. by repository to repository transfer (see T3.2.1).

M2.3: Viewers, processors, and editors for structures and data

Goal: Improve OS data viewers, processors, and structure editors to embed them into NFDI4Chem infrastructure or use them as standalone tools.

Description: Viewers and editors for data and chemical structures are a fundamental part of the Smart Lab concept and chemistry repositories (among other systems such as databases, inventories, and chemistry services), as the creation of FAIR machine-readable data depends on them in a digital research environment. This also means that the limitations of viewers and editors, in terms of their functionality, applicability, accessibility, or performance, have a direct impact on the availability of FAIR data in chemistry. Improving OS viewers and editors also serves as a valuable contribution to the wider NFDI4Chem community, which can use them in their standalone versions. 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 searching, visualising and analysing the research data they store.

T2.3.1 ChemConverter

During FP1, we developed ChemConverter, which allows us to convert data from myriad formats into a few standardised file formats, which is important for using data in digital systems such as ELNs and guaranteeing sustainable storage and access to the data. We designed and built a standardised way to

1. read (meta)data from various file formats,
2. map metadata to community-defined metadata schemas, and
3. convert data into the standard ‘JCAMP-DX’ format with support for single files as well as file collections in the form of ‘BagIT’ containers.

In FP2, the concepts will be extended to more complicated data, such as multidimensional data and the usage of additional readers designed by external communities. ChemConverter aims to support a wider range of possible input formats, also including proprietary formats. To comply with other systems and strategies for RDM, ChemConverter will be extended (also see M4.3) to export additional standards such as HDF5, RO-crate, frictionless data, and JSON-LD.

T2.3.2 NMRium

NMRium is an open-source React component developed by Zakodium for processing, viewing, and analysing NMR spectra (Davies and Patiny 2021, Patiny et al. 2024). NMRium has already been successfully integrated into Chemotion ELN and nmrXiv (s. T3.1.3). During FP2, the integration will be enhanced by improving the interoperability of the component. Specifically, the file format will be fundamentally refactored to separate spectral data, annotations and metadata. This will streamline callbacks between NMRium and its environment and greatly improve performance, as only small JSON chunks will be passed instead of the entire dataset. In addition, the internal filter handling will be

refactored to allow modular extensions. The most important planned extensions will be a 2D NMR Fourier transformation, a non-uniform sampling (NUS) reconstruction, and improved usability with respect to touch device support.

T2.3.3 ChemSpectra

ChemSpectra (Huang et al. 2021) is the default data visualisation tool of the ELN, being able to process and visualise the standard file format 'JCAMP-DX' or 'BagIt' containers. ChemSpectra can be embedded into the ELN functions or used as a standalone version. In FP2, the currently supported workflows, layouts and functions, e.g., NMR, MS, IR, UV-Vis, CV, SEC, and TGA, will be continuously extended in accordance with T2.3.1 to support all techniques required by the users and to improve and extend the existing technologies. The functions will include the requirements of all applicable chemistry subdisciplines. Examples from our user feedback to be implemented are EPR, LSS, DLS, CD, porosimetry, rheometry, two- and multidimensional- HPLC and GC with fluorescence matrix as well as combined techniques such as GC-MS and HPLC-MS. ChemSpectra is a flexible tool with the ability to analyse a variety of measurements, offering a broad use case that is indispensable for sustainable, standardised, and affordable FAIR RDM.

T 2.3.4 Implementation and rework of Structure Editor Ketcher (KIT)

A structure editor called Ketcher is of growing importance for NFDI4Chem, as there are only a few other structural editors available that allow professional work with chemical structures (ChemDraw or Marvin). In the last few years, commercial software vendors have increased their licence fees in a way that many institutions cannot afford anymore. At the same time, the providers restrict the use of the licensed products in digital tools and thus prevent the use of the licensed software in software other than their own. Ketcher is the only sophisticated OS structure editor, and its further development should be supported to cover the needs of NFDI4Chem. In FP2, the additional necessary functions for advanced inorganic chemistry, catalysis, and materials sciences will be developed with experts from the different subdisciplines and in collaboration with experts from EPAM Systems' development team.

T 2.3.5 Enhancing data and metadata quality control

Measures to control data and metadata quality are important to increase the acceptance of digital systems and to make use of their benefits for data producers and data reusers. In FP1, quality control features of ELN content were achieved, referring to the completeness of (meta)data, its readability and re-usability, as well as its consistency and scientific plausibility (Huang et al. 2024). The developed functions were implemented based on OS components described in T2.3.2 - T.2.3.4 and additional AI tools - showing the past and future importance and scope of those developments. In FP2, the work on automated curation and quality control features will continue to provide robust and trustworthy data control methods and tools.

M2.4: Development and extension of topic-related ELN functions

Goals: Further development, adaptation, and extension of the ELN Chemotion (reference implementation for NFDI4Chem) to meet the requirements of the scientific community in all subdisciplines of chemistry.

Description: In FP1, the requests referring to the further development of the ELN were included according to different subdisciplines. In FP2, we will consolidate the work on the requirements of different subdisciplines (see description of TA2 above) by discussing and designing new functions with broad applicability, suitable for different areas of the ELN, independent of the user community. Adaptation of the ELN with such functions that combine the efforts of several needs and use cases will be designed in the form of working groups bridging different subdisciplines and technology experts. The working groups' measures for quality control will be closely coordinated with all software developers in NFDI4Chem and advised by colleagues in Research Software Engineering (RSE) [see LOS RSEde].

T2.4.1 Ontologies, semantic enrichment, and standards implementation, InChI

In FP1, the use of ontology terms from the NFDI4Chem's Terminology Service was enabled in different modules of the ELN e.g. hard-coded elements (cell lines, enzymes by the end of 2024) or extensions (see T2.6.1). In FP2, we will include a general approach to flexibly integrate ontologies and/or their terms in all areas of the ELN, supporting the standardisation of terms and the export of well-defined schemas (see also T2.6.2). As a prerequisite to systematically enrich the data with semantics, to support linked data, and to provide machine actionable FAIR data, we will implement MIChI-derived metadata schemas (see M4.1 and M6.3), and make ontologies assignable to generic forms to support interdisciplinary work (see M6.1, M6.2). Further chemistry standards, such as different InChI types (see T.4.2.1) and EnzymeML, will be included and continuously updated (see T2.4.2) in collaboration with TA4 and TA6.

T2.4.2 Coding of new entities and ELN wide entity embedding

During FP1, TA2 implemented several new main entities within the ELN according to the requests from users in different domains: Altogether four hard-coded entities (devices, sample types, cell lines, and enzymes (planned until end of 2024) and 11 generic elements providing versioned templates for new entities were provided (using LabIMotion, see T2.6.1). In FP2, the new entities proteins, samples for materials applications, environmental chemistry, and electrochemistry workflows (support by the Cluster ETOS, see LOS) are scheduled, including the definition of community-relevant identifiers, the inclusion of information queried from external community-relevant databases and the implementation of common reporting standards for these entities. Further entities will be designed and developed based on the community's feedback in FP2 and prioritised according to the user group, depending on the new entities.

T2.4.3 High throughput and parallel experimentation

In FP2, TA2 will include workflows and representations to allow the documentation of high throughput experimentation, which has special importance in experimental areas such as but not only (in)organic chemistry (screening and catalysis) and biochemistry (assay conduction and analysis). While the basic requirements were already planned and partly implemented in FP1 (e.g. adaptation of well-plate format options; reaction variations tab), the coding of user input options for the structured planning, e.g., as a result of the design of experiment approach, design of experiments (DOE), the representation of multiple information based on matrix-like setups and the input, processing and representation of analytical data needs to be continued in FP2.

T2.4.4 Data analysis and representation - data overview

In FP1, the focus of data analysis and representation functions in the ELN lay on the implementation of routines for single datasets. In the next step, the ELN needs options to analyse and represent data on a higher level of the software, enabling

1. overarching analysis of analytical data,
2. a combination of calculated and experimentally obtained data, and
3. graphical visualisation of calculated properties in ELN-spanning 2D and 3D visualisation plots.

The analysis and representation tools need the implementation of different options such as filters and a flexible adaptation of the plotting actions. The newly gained data collections need to be organised and managed in parallel to the originally stored source data.

T2.4.5 Beyond experimental investigation

The focus of the work of TA2 in FP1 lies in solving challenges in laboratory environments. FP2 will include the systematic inclusion of workflows and features needed for chemists who work in theoretical chemistry, computational chemistry, and the cheminformatics community (see also M4.4). The first requirements were assessed by existing partners in computational chemistry and by a working group consisting of representatives of five theoretical chemistry groups (Jacob et al. 2024) and include, amongst others,

1. a JSON import and export for entities and subentities,
2. an API to the content of the ELN on different levels e.g. for Jupyter Notebooks,
3. the implementation of methods for quantum chemistry calculation,
4. the representation of spectra from the theoretical output,
5. and the implementation of simple analysis routines for theoretical chemistry such as calculation of energy differences.

T2.4.6 Extensions referring to reported issues

The ELN team needs to work on additional requirements and bug fixes that are encountered during the project's runtime. Addressing unforeseen issues, upcoming new

topics, or reported challenges is an important instrument for enabling community-driven code development. In FP1, we received a total of 393 issues via GitHub (in addition to many requests via helpdesk and email). 209 issues were solved until 07/2024, while more protracted tasks were included as suggested changes in the FP2 program. In FP2, we will continue this approach of continuous support of requests in combination with a strategic alignment and integration of major themes.

M2.5: Services provided to the community

Goals:

1. Ensure the availability of the Chemotion ELN for all scientists in Germany.
2. Provide access to other OS ELNs, ensuring the best support and advice for the scientists.

Description: For institutions or user groups of NFDI4Chem that want to host the ELN Chemotion under their own responsibility using their own resources to have full control over data and access options, TA2 provides a Docker image and a command line interface (CLI) for an easy installation, update, and maintenance of the ELN (T2.5.1). For single users or small groups that need an out-of-the-box solution with minimal cost and maintenance effort enabling low barriers to entry, we will continue to offer an ELN as a service (T2.5.2). A technical helpdesk supports users and system admins with information and hands-on support for all TA2 relevant systems (T2.5.4) and an openly accessible server for OS ELNs allows the comparison of available ELN systems in the domain of chemistry for testing and teaching purposes.

T 2.5.1 Distribution of updates, Docker, CLI/GUI tool

The source code and its documentation will be available on code repositories such as GitHub for interested developers. T2.5.1 will manage the Docker image generation for each main, minor and most important patch releases. The availability of updates and upgrades for self-hosted instances of the ELN (at different German universities) will be managed and communicated via different channels, e.g. admin and user mailing lists. Driven by user feedback, the CLI will be further extended by a GUI to simplify the management and configuration of the ELN instance(s). 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).

T 2.5.2 ELN as a service/ hosting

A centrally hosted ELN Chemotion as a service was prepared and run in a test period with selected users within FP1. A generally applicable procedure for an access model by different users implemented with the AAI infrastructure (see cross-cutting topic M6.3.2) was defined and implemented. TA2 will care for the recent updates and upgrades of the software and will assign the needed resources. FIZ will manage the implementation of features like searching in external databases whenever allowed according to available licences. The constant access to the ELN as a service is overseen by a support team,

which provides its best effort to solve user problems and mediate bug fixes. High availability of the ELN service is ensured by fail-safe hosting using a professional redundant server architecture and virtualisation.

T 2.5.3 Service for open source ELNs

KIT will install demo instances for OS ELNs other than the NFDI4Chem-supported one. In FP1, five (+1) ELNs (OpenEnventory, openBIS, eLabFTW, Herbie, SciNote, Chemotion, planned: RSpace) were supported in the form of an up-to-date test environment, allowing scientists to test different systems for the suitability to their requirements without the obstacle to set up these ELN on their own. A new survey on additional systems is currently ongoing (examples are: sample.de, AI4green), giving the requirements to include further systems in FP2.

T2.5.4 Technical helpdesk (second-level support)

In FP1, 53% of the users' requests (2022-2024: 730 tickets in total) were assigned to topics related to TA2 and especially the use, installation and configuration of the ELN. The second-level support of TA2 handled and answered those requests in FP1 (passed by the general helpdesk installed in TA5; see T5.3.4) and will continue its activities and support for the requests in FP2. The technical helpdesk will solve all technical issues (installation, setup, maintenance and use) of the admins or users with respect to the ELN device integration modules and will answer questions with respect to software requirements and issues.

M2.6: ELN as part of a digital ecosystem

Goal(s): Enable the sustainable handling of data, independent of the used software

Description: Measure 2.4 will adapt the applicability of Chemotion ELN to a broader scientific scope to fill the current gaps between the work of chemists and other disciplines that require either an application-related focus and/or the inclusion of chemistry aspects for analytical and interdisciplinary work with 3rd party funding (T2.6.1). Further, we will enable the interoperability of the ELN Chemotion with other systems that are not part of the NFDI4Chem federation yet via different means, such as the support and implementation of standards that are developed within the most important open source ELNs (T2.6.2).

T2.6.1 LabIMotion extension

During FP1, the ELN team received an increasing number of requirements, each of which would have necessitated field-specific ELN functions and forms at different levels of information within the ELN's structure. To offer solutions to the user community of NFDI4Chem, which often overlaps with communities from other NFDI consortia and their work, we designed and implemented LabIMotion, enabling the design of new ELN entities at different structural levels assigned to well-defined metadata without the need

for hard-coded changes. These activities will be maintained and are planned to be extended by applying for further 3rd party funding.

T2.6.2 ELN interoperability

The interoperability of ELN content will be gained by the support of the initiative 'ELN consortium', which started the development of common standards for ELN-interoperability based on RO-crate and the file format *.eln. The team of TA2 will support this approach by implementing and the design of JSON-LD schemas for reuse in RO-crates, enabling the specific, detailed inclusion of scientific information into the interdisciplinary community. The work will be well-negotiated and supported by working groups in the Helmholtz community (coordinated by KIT and strengthened in TA6, T6.4.1).

Table 3

Table 3.

Potential risks of the TA2 work programme and measures for their mitigation.

Description of management risks	Proposed risk-mitigation measures
RM1: 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, and to export, and transfer data. **Likelihood**: medium	Solutions to the given risks are described in TA4 and TA6 and were successful in FP1. 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. **TA(s) involved**: TA4, TA6
RM2: Security risks and data leakage for the decentrally installed ELNs and ELN as a Service. **Likelihood**: low	The ELN software will undergo regular security checks as part of the CI process and penetration tests. The ELN as a service is hosted in professionally operated data centres at KIT-SCC and FIZ with established security architecture, intrusion detection systems and virus scanners maintained by experienced administrators. **TA(s) involved**: TA2
RM3: Risks in the case of budget cuts. **Likelihood**: high	See general description of this risk in TA1. **TA(s) involved**: all task areas

5.3 TA3: Repositories

TA3 focuses on maintaining and enhancing the federated research data repository system in chemistry, established in FP1, to ensure trust and cooperation for seamless data exchange and interoperability. This federation includes high-quality repositories trusted by the chemical community and recommended by journals and funding agencies, e.g., Angewandte Chemie Int. Ed., ChemBioChem, Journal of Natural Products, DFG GSP (in German), which covers all relevant subdisciplines and data types in chemistry. We foster the publication of high-quality, open-access FAIR chemistry data to guarantee access and easy reuse. TA3 operates and maintains reliable services with interfaces that connect various systems, including data ingestion (directly or via an ELN), terminology

services, and external tools. Adhering to NFDI4Chem's high standards and FAIR principles is crucial. Key repositories were connected to our central search. We will expand this to the entire federation to enhance disciplinary findability.In FP2, we aim to sustain, optimise, and consolidate all federated repositories, enhancing user interfaces, disciplinary metadata and MIChI support in cooperation with TA4 and TA6, interoperability, and machine-actionability for AI applications as shown in Fig. 5. Key workflows for publication, curation, review, and versioning will be streamlined to improve data quality and service usability. Sustainable operating models will be developed, and new repositories meeting our criteria will be onboarded. We will track data reuse and citations to evaluate our goal of promoting high-quality FAIR data publication and reuse by the whole chemistry community.

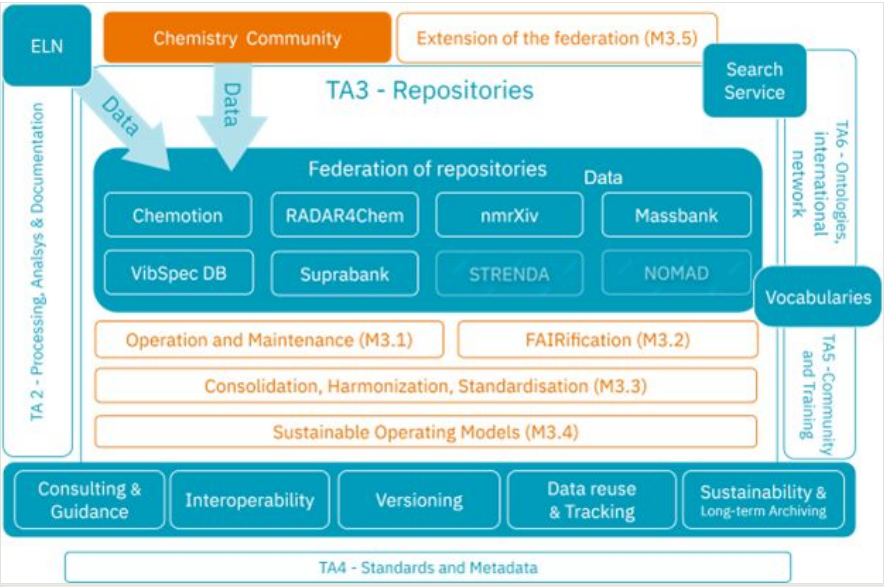

Figure 5. doi

The structure of TA 3, Federation of repositories, and its connections/interactions with other task areas.

TA3 will contribute to NFDI4Chem's **key objectives 1, 2, 3, 5, and 6**, which are also reflected and differentiated in the **objectives of TA3**, which are:**O3.1** Expanding and consolidating a sustainable and trusted federation of standards-based and quality assured data repositories covering the needs of the chemistry community.

O3.2 Enabling researchers to easily ingest, search, annotate, exchange, publish, archive and reuse chemistry research data and metadata across internationally distributed data services.

O3.3 Enhance seamless interoperability of services using state-of-the-art technologies and providing FAIR, high quality chemistry data for interdisciplinary research, data science and AI.

O3.4 Ensuring long-term operation and accessibility of repositories and chemistry data.To reach our goals, we defined **five measures** covering key elements for the enhancement and sustainable operation of the federation of interconnected repositories, including:

1. trust and cooperation, exchanging metadata and data,
2. compliance with our standards for data and metadata (including MIChI), data quality, interoperability, reliable operation, and long-term availability,
3. covering essential needs of major chemistry subdisciplines,
4. onnectivity between repositories, ELN, processing systems, terminology and ontology services, and tools,
5. metadata harvesting allowing disciplinary findability in NFDI4Chem's search and other systems,
6. integrating functions and tools needed by chemists (converters, viewers, editors),
7. tracking data reuse and citations,
8. providing user guidelines (e.g., how to choose the right repository), and
9. seamless journal publication workflows identified in Editors4Chem.

M3.1 Sustainable operation, enhancement and maintenance of chemistry repositories

Goal(s):

1. Ensure sustainable operation of each core repository
2. Ensure secure and up-to-date research software maintenance by patches and minor and major operating system updates
3. Guarantee fully operational and reliable services and technical infrastructure

Description: We describe

1. functionalities developed in FP1 that require
2. maintenance and operational tasks for each repository in the form of seven levels: (A) Required security updates by continuous delivery of patches; (B) Major updates and upgrades of the main backend and frontend components and/or framework(s); (C) Update and upgrade of the database; (D) Update, upgrade, and replacement (if needed) of included software libraries, supported by software bills of materials (SBOM); (E) Adaptation of code required by changed dependencies; (F) Operating system updates and patches; (G) Resilient technical operation and backup strategies.

T3.1.1 Chemotion repository

In FP1, the Chemotion repository was extended referring to strategic requirements such as:

1. the integration of NFDI4Chem recommended AAI options and ROR,
2. the improvement and fine-granular definition of the authors' and reviewers' rights and the connected available functions for different user roles,

3. the versioning of reactions, samples and their analytical data,
4. the support of JSON-LD the most relevant content of the repository, and the integration of embargo options for data collections and their support with collection DOIs. The repository was
5. systematically extended with all requirements and updates available from the developments of Chemotion ELN, which were integrated successively respecting the repository's specific dependencies. The greatest advance was gained with
6. the integration of modules that come with the functionality of LabIMotion (see T2.6.1), allowing not only the adaptation of the content of the repository at different levels to diverse subdisciplines of chemistry but also the standardisation of the content in ELN and repository across different instances with an
7. integrated, version-including template hub of which the content is aligned with the developments of TA4.

Maintenance and Operation (applicable to levels A-G) will cover the above-described improvements of FP1 and further developments aimed for in FP2 (see M3.2-M3.5), as well as the curation of the submitted datasets.

T3.1.2 RADAR4Chem

In FP1, the RADAR4Chem Repository was launched, based on the RADAR repository operated by FIZ Karlsruhe since 2017. RADAR4Chem is a reliable, user-friendly catch-all repository with a low entry threshold (currently free). It allows the publication of large, complex datasets of all chemistry subdisciplines and data types. Access to published data is guaranteed for at least 25 years, with a DOI and comprehensive landing page for each dataset.Developments focused on ensuring FAIRness and increasing interoperability through software releases:

1. Metadata schema update: Integration of Research Organization Registry (ROR) and Integrated Authority File (GND),
2. Optimisation of Schema.org conformity: Improved mapping, markup, embedding JSON-LD and Turtle format serialisations on landing pages,
3. Enhanced Linked Open Data (LOD) framework and extension of the RADAR Knowledge Graph (KG) based on the Schema.org ontology,
4. Optimisation of landing pages according to the FAIR signposting approach, improving machine readability and actionability,
5. Daily creation of a RADAR4Chem knowledge graph based on Schema.org ontology,
6. Support of a SPARQL endpoint for querying the knowledge graph in a standardised and machine-accessible way,
7. Participation in the "FAIRness assessment challenge" organised by the EU project FAIR-IMPACT, increasing the FUJI-score of RADAR4Chem to 87%,
8. Introduction of support for WebDAV for flexible dataset transfer and larger data volumes,
9. Dockerisation of the software application for efficient deployment and maintenance.

Maintenance and operation (tasks A-G) will continue in FP2, guiding further developments (see M3.2 - M3.5). We will also continue to operate RADAR4Chem free of charge, expand dataset curation, enhance subject-specific metadata annotation (and MIChI), and improve functionality.

T3.1.3 nmrXiv

NmrXiv´s ultimate goal is accelerating broader coordination and data sharing among researchers by creating a platform for managing, sharing, and analysing raw and processed NMR spectral data. The milestones achieved by the repository during FP1 are:

1. advanced search,
2. API for data searching and retrieving,
3. faceted search capabilities,
4. management of access permission levels,
5. restricted data access mode via roles,
6. multiple download options in various formats and metadata export as JSON,
7. DOI assignment to datasets,
8. integration of NMRium for data format conversion,
9. mapping of NMR data associated with compounds to third-party databases through identifiers,
10. standardised protocols (e.g., RESTful APIs, OAI-PMH) for data exchange,
11. citation export,
12. data versioning,
13. recommendations covering Liquid-State NMR Experiments of Small Molecules to extend MIChI.

Maintenance and operation (tasks A-G) will continue in FP2, guiding further developments (see M3.2 - M3.5). Additionally, we implement RO-Crate support to create self-describing data packages and (4) support the ISA data model.

T3.1.4 MassBank

During FP1, several achievements were made for MassBank:

1. multiple data releases were deployed, adding spectra from MSSJ, Eawag, "Shin-MassBank", the Federal Institute of Hydrology, Agilent, the University of Antwerp, UFZ, the University of Birmingham, the US EPA, and Stockholm University.
2. Introduction of tombstone pages for deprecated records and metadata improvements through Bioschemas updates and mapping to PSI MS ontology
3. System updates included server versions 2.2.3 and 2.2.4 in 2023 and the new MassBank Accession schema in August 2022.
4. Community engagement was enhanced by participation in key meetings such as the Metabolomics Society meeting in Osaka (June 2024).

Maintenance and operation (tasks A-G) will continue in FP2, with the primary site at UFZ and the backup site at IPB. Further developments (see M3.2 - M3.5 and PID optimisations) will follow.

T3.1.5 VibSpec DB

In FP1, the first pre-release version of the VibSpecDB, which specialised in vibrational spectroscopy data, was deployed as a test instance on a Kubernetes cluster. It provides basic functionality, such as:

1. file upload,
2. user's profile pages,
3. single sign-on and registration,
4. admin console,
5. spectral and file tree viewer,
6. export functionality.

Users can set up minimal required or custom metadata and tags and create search queries. A separate python microservice was created to parse input data and metadata of some common file formats. A workflow and infrastructure to deploy the software (servers, file storage, database) were developed. By the end of FP1, the data publishing functionality needs to be created, and the first public version will be released. In FP2, the Vibrational Spectroscopy Ontology (see TA4) will be integrated, and standards and minimal requirements for spectra and metadata will be developed and implemented. Additional user interfaces are planned: VibSpecDB CLI and a collection of documented APIs, which will be used in a UI-free repository. Maintenance and operation of the public instance (tasks A-G) will continue in FP2, guiding further developments (see M3.2 - M3.5).

T3.1.6 Suprabank

In FP1, SupraBank was enhanced with features such as

1. ROR integration,
2. user rights and roles management,
3. DataCite integration for DOI-based references,
4. embargo options for pre-publication datasets,
5. peer-review process for dataset curation,
6. CrossRef integration for linking datasets with publications,
7. enhanced metadata for better machine readability and sharing, and
8. new compounds like nanoparticles and materials, including external API integration.

In FP2, SupraBank will enter a consolidation phase to merge with Chemotion, preventing parallel development and double maintenance. The integration concept will be

completed by the end of FP1, allowing the design and implementation phase to start in FP2. The planned functions

1. raw data upload and processing,
2. dataset versioning,
3. collections for user search and curation, and
4. API expansion will be covered by Chemotion after the consolidation.

M3.2 Advanced FAIRification: optimise interoperability, data reuse, AI-readiness and user experience

Goal(s):

1. Unfolding the full potential of the federation of repositories by optimising interoperability and interfaces, connecting additional services
2. enabling NFDI4Chem services to track data transfer, -citation and re-use scenarios
3. FAIR-compliant provision of data collections and metadata for machine agents (machine-actionability)
4. provide optimised user interfaces
5. exchange with international FAIR implementation experts, enhance FAIR principles implementation in chemistry and consult repositories.

Description: The federation of repositories benefits from a common strategy that includes data exchange, storage, and provision. Distinct interfaces enable interaction within the federation and with overarching services (e.g., the Search Service). In FP1, NFDI4Chem developed a sharing policy and concept for implementing the federation. Repositories are gradually integrated into the data exchange and interoperability concept (T3.2.1, T3.2.3), which began in FP1 and continues in FP2. Information exchange and reuse are tracked by systems identifying data flow in various scenarios, including unpublished (repotracker, T3.2.2) and published data (search service, T3.2.4 and citation tracker, T3.2.2). APIs ensure seamless background exchange, while user interfaces provide standardised access via LLMs and SPARQL queries. The federation concept is managed and controlled by collaborative (T3.2.5) and central means (T3.2.6).

T3.2.1 Data and service interoperability

T3.2.1 will optimise the following aspects for each repository:

1. Interfaces for data transfers (a) between ELN and repositories and (b) between repositories. As shown in Fig. 5, repositories within NFDI4Chem are fed with data by manual upload or transfer from services used earlier in the research data life cycle, such as ELN Chemotion (see T2.2.5, TA2) or Repository Chemotion (T3.2.2). Data transfer from the repository Chemotion to the specialised repository nmrXiv is going to be realised until the end of FP1. In FP2, this reference implementation will be extended stepwise to other repositories within the federation.

2. Support for widely accepted data formats in the chemistry community, such as NMReData, JCAMP-DX, and SMILES. This includes implementing community-recognized metadata standards based on the MIChI developments in M4.1 for comprehensive dataset descriptions and transformation via the MSS to schemas like DataCite, schema.org, and Bioschemas for interoperability and findability through external search services,
3. assignment of persistent identifiers, such as DOIs, Handles, ROR, RAiD, ISGN, and InChI, to metadata and digital objects as needed,
4. integration of TA6 services to ensure consistent use of controlled vocabularies, ontologies, and taxonomies specific to chemistry,
5. machine-actionable metadata: Ensure digital objects are accompanied by machine-readable metadata and detailed contextual documentation, in conjunction with T2.3.3 techniques,
6. optimisation of APIs and data exchange protocols, such as OpenAPI for RESTful APIs and OAI-PMH for metadata harvesting, to enable seamless data integration and retrieval across different systems.

T3.2.2 Data reuse tracking

Data reuse tracking will be implemented for different scenarios: data transfer from ELNs to repositories and between repositories will be tracked by the 'Repotracker' (developed in FP1). The Repotracker, currently used as monitoring for the same dataset in different repositories, will be extended to track detailed interactions with individual repositories in FP2. Further, the reuse of data sets of the federated repositories will be tracked using the DataCite Citation Corpus and similar approaches. We will develop the CitationTracker-SW capable of identifying and counting data citations and how data is re-used to be fed e.g. into the NFDI4Chem knowledge graph (T6.3.3).

T3.2.3 Enhanced machine-actionability and user-friendly database queries

To streamline data retrieval from large or nested datasets, the federated repositories will implement:

1. machine-actionable landing pages (e.g., using the FAIR signposting approach),
2. SPARQL endpoints, and
3. LLM support (or equivalent technologies). We will simplify SPARQL query generation with AI tools like LLMs, enabling user-friendly query results (see also T6.3.2).
4. Data and metadata can be stored in an RO-Crate.

Each file and dataset will have a persistent identifier, integrating the FAIR Digital Objects (FDO) approach for robust interaction with digital objects. RO-Crates will be validated and indexed, with APIs and web interfaces provided for accessing FDOs. Implementations will vary by service and user needs, with consultation by the TA3 core team. We will also support repositories with the integration into the Semantic Data Hub (M6.3).

T3.2.4 Connections of the federation of repositories to other services

The federation of repositories aims to connect repositories with several other services:

1. to the NFDI4Chem and other data search services and the Semantic Data Hub (M6.3) to enable seamless discovery of chemical data and related scientific information
2. to the massively scalable (OAI-PMH) provider offered by FIZ to collect and harmonise metadata from federated repositories, ensuring consistent and up-to-date information across the federation,
3. to service overviews and marketplaces like re3data, FAIRsharing, or the EOSC marketplace.

T3.2.5 Continuous design and adaptation of the main common work principles ("TA15")

T3.2.5 facilitates ongoing, long-term collaboration among the PIs of main services (TA 2+3+4+6, "TA15") to optimise the core infrastructure. This involves elaborating, investigating, and negotiating new tasks for NFDI4Chem services. For FP2, key aspects include:

1. designing interoperability concepts for unaddressed repository entities,
2. exploring innovative concepts like FAIR signposting and FDOs,
3. developing a work program for fully implementing the "I" in FAIR (knowledge representation, vocabularies, and metadata references),
4. connecting with external services, and
5. evaluating licensing options for research data and collections. Many of these tasks require cross-task area discussions within NFDI4Chem.

T3.2.6 International FAIR technology assessment and consulting for NFDI4Chem repositories

By engaging with international experts (e.g., Research Data Alliance (RDA), European Open Science Cloud (EOSC)), TA3 will shape FAIR principles and implementation standards, transferring knowledge to the federation's repositories. We aim to ensure our services comply with FAIR principles, set benchmarks and provide recommendations for improvement. We will present and discuss our approaches at international conferences, showcasing our impact and gathering feedback. This will strengthen international collaboration, enhance visibility and credibility globally, and ensure optimal FAIR implementation in our services.

M3.3 Consolidation, harmonisation and standardisation of the federation of repositories

Goal(s):

1. raise data quality and adapt existing descriptions and schemes according to the already developed and forthcoming standards (TA4 and beyond);
2. harmonise dataset versioning;

3. connect to NFDI base services;
4. increased operational fitness;
5. optimal embedding into journal publishing processes.

Description: M3.3 emphasises thorough quality assurance and adaptation to NFDI4Chem standards, ensuring careful management of versioning, data history, and reuse tracking (e.g., data citations). We are committed to seamlessly integrating the repositories into NFDI core services and other overarching infrastructures to enhance collaboration and operational synergies. This comprehensive approach optimises efficiency while maintaining robust data management practices throughout our projects.

T3.3.1 Data quality assurance and adaptation to our standards

In FP2, all repositories, especially Chemotion and RADAR4Chem, will

1. adapt descriptions and schemes to MIChI standards to cover profiles by at least one service. This includes
2. various levels of standardisation, such as (a) relations between investigations, (b) reactions, (c) other chemical processes, (d) entities like samples and devices, and (e) measurement types.
3. Implementation will provide metadata based on MIChI-derived schemas for the NFDI4Chem search and the Semantic Data Hub (M6.3) and optionally support schema.org or DataCite-compliant schemas to facilitate reuse and harvesting (T3.2.1).
4. Templates for user input forms to describe data types will be harmonised or developed, connected to controlled vocabulary and ontologies (M6.2). FP2 will also focus on
5. data quality assurance and
6. data curation mechanisms, depending on data types and available or planned functions for manual or automated processes.

Scientific teams will support manual curation for peer review and diverse data types, while automated curation will be supported by tools developed in TA2, using specialised software like AI.

T3.3.2 Dataset versioning - a harmonised implementation and adaptation

Data versioning means tracking and managing changes to datasets over time. Modifications have to be uniquely identifiable and retrievable, much like version control in software development and allow citing dedicated versions. Existing data versioning frameworks have the drawback of depending on journal paper citations to make new data versions discoverable (González–Cebrián et al. 2024). Versioning is a key requirement for the CoreTrustSeal (CTS) certification, which establishes to report any changes to data and metadata. Thus, we will improve and harmonise it on different levels, promoting transparency of data provenance and reproducibility, enhancing the recognition of data curation efforts (T3.3.1) and allowing citing datasets as well as their individual versions.

T3.3.3 Integrating repositories with NFDI base services

The NFDI4Chem repositories will integrate NFDI base services, starting with IAM4NFDI (AAI infrastructure, SSO) through a phased approach:

1. initial repository integration,
2. handling complex integrations in projects supported by Base4NFDI, including advanced IAM for group membership management across services (ELN, repositories), harmonisation of user representation, roles and rights.

We will evaluate and adapt the virtual organisation concept, and repository operators will receive support for implementation. Other services to be connected include the NFDI Knowledge Graph (KGI4NFDI) and DMP tools like RDMO (DMP4NFDI).

T3.3.4 Operational quality, policies and licences for research data

Operational fitness is crucial for a sustainable and reliable service. Certification, such as CTS, strengthens repositories and builds trust. In FP1, only a few repositories applied for certification due to challenges like the need to be fully operational, application effort, and restrictions to specialised repositories. In FP2, we will enable our repositories to achieve CTS certification or equivalent, first assessed in FP1. We will establish quality measures that meet certification requirements, ensuring repositories operate with a commitment to transparency, reliability, and trust, documented by accessible policies. We will harmonise commitments to meet general standards, including data protection, ethical principles, and adopting specialised licences for research data (e.g., ODbL, CDLA).

T3.3.5 Peer-review process harmonisation for data publications

NFDI4Chem promotes publishing research data alongside traditional papers and established a dialogue with publishers in FP1, leading to a mutual understanding of data publishing. However, aligning peer review workflows for journal and data publications is challenging due to varied submission processes, fear of releasing data before manuscript acceptance, cultural biases, and maintaining reviewer anonymity. To integrate the publication process for both papers and data, repositories must support a transparent, anonymous, and efficient peer review process involving

1. creating/reserving a PID before data publication for referencing in manuscripts,
2. enabling anonymous access to data and metadata for publishers and reviewers, and
3. making data temporarily immutable during review.

Task T3.3.5 will address these needs in coordination with TA4 (M4.5) and Editors4Chem (Parks et al. 2023, Parks et al. 2024), extending publisher recommendations (e.g. Wiley).

M3.4 Developing and implementing sustainable service operating models

Goal(s): Ensuring

1. reliable long-term archiving (LTA),
2. a functional exit strategy,
3. data migration paths,
4. reliable hosting of central software tools as services,
5. sustainable operating models for the federated repositories.

Description: M3.4 aims to establish a comprehensive LTA solution for federated repositories. This includes designing a robust LTA blueprint (based on RADAR's model), implementing it in repositories, and developing exit strategies for discontinued repositories with support for data migration. It offers centralised hosting of key software tools to improve efficiency and promotes sustainable operating models to ensure the long-term reliability and scalability of data services.

T3.4.1 Design of a blueprint for Long-Term Archiving (LTA)

In T3.4.1, we will design an LTA solution as a blueprint for repositories without LTA. We propose using RADAR's well-established design, which can serve directly or as a blueprint for a more specific solution. Data transfer uses standardised protocols and is maintained on tape storage libraries with redundant copies in three locations. The storage architecture is technology-independent. Data is ingested for temporary, then permanent storage. Initially, data is transferred via HTTPS to RADAR's temporary memory, checked for integrity, packaged (BagIt) and transferred via SFTP (TAR file) to three permanent archive tape storage systems. During retention, data packages are immutable and preserved physically. FIZ Karlsruhe (TA3) manages operations, including data distribution, virtual machine maintenance, software development, updates, and patching.

T3.4.2 Long-Term archiving implementation plan

We will follow a structured process to implement the LTA solution proposed in T3.4.1. Hardware and software components will be procured, and the virtual environment will be prepared. Data will be assessed and inventoried for archiving. The archiving solution will be installed and configured for seamless integration, and storage solutions will be established. Testing and validation will be performed. Documentation and policies for data archiving, retention, and access control will be established. Continuous monitoring and maintenance will be ensured.

T3.4.3 Defining an exit strategy for each core repository

TA3 will develop an exit strategy for repositories that can no longer manage their data sustainably. Relevant data collections or repositories should be migrated to a suitable repository or transferred to LTA storage (T3.4.4). Operations like metadata and format

preservation, PID transfer, and ensuring long-term accessibility and usability of data need to be organised, independent of the repository software stack.

T3.4.4 Data migration to other repositories

TA3 supports data migration when relevant repositories are discontinued or require integration into the NFDI4Chem infrastructure. Data will be transferred to the most suitable repository to ensure long-term availability for the chemistry community. This includes mapping database contents, converting data for the target repository, and adapting metadata and licences. Original service functions will be maintained whenever possible. In FP2, the first use case for such a migration will be the consolidation of the Suprabank and Chemotion repositories.

T3.4.5 Hosting sustainable software on central infrastructure

To streamline software tools for chemists, we aim to centralise hosting for minimal costs and efficient maintenance. Identified candidates include the FIZ-OAI provider, viewers, editors, converters for various file types, and ELN instances (T2.5.2). Centralising these tools enhances interoperability and resource efficiency. A flexible deployment infrastructure using container technology (e.g., Docker) will allow updates from multiple sources. Continuous DevOps support will ensure seamless integration and connectivity, providing up-to-date, optimised software across services.

T3.4.6 Sustainable operating-, funding- or business-models

TA3 will advise core services on sustainable operations for long-term reliability, focusing on optimising processes, scalability, and continuity. While services are currently free, securing sustainable funding is essential. RADAR's business model could be used by other repositories to ensure longevity. We prioritise reliable service providers and propose transition solutions to ensure minimal disruption if sustainable models are not feasible (T3.4.3 - T3.4.5).

M3.5 Extension of the federation of repositories

Goal(s):

1. Continuously monitor the global landscape of chemistry repositories.
2. Identify and catalogue candidate repositories for onboarding based on functionality, governance, and user engagement.
3. Integrate new chemistry repositories into the federation, ensuring compatibility and interoperability.
4. Support repositories during the integration phase.

Description: M3.5 monitors global chemistry repositories, assessing their functionality, governance, and user engagement while staying updated on technological and policy changes. Selected repositories are onboarded, meeting technical benchmarks and integrating with consortium services like ELN and central search. Regular forums ensure

knowledge exchange on FAIR technologies, adherence to standards, and smooth integration.

T3.5.1 Ongoing analysis of the international repositories landscape

In T3.5.1, TA3 will continuously monitor global chemistry repositories. Initially focused on national repositories, we began internationalisation with FAIRsharing in FP1.In T3.5.1, TA3 will continuously monitor global chemistry repositories. Since NFDI was initiated as a national call, in FP1, we focused on national chemistry repositories (Bonatto Minella et al. 2023a). In FP1 the internationalisation process was initiated by collaborating with FAIRsharing: a collection of standards and repositories (called NFDI4Chem) was created. The goal is to assess the current technological state, trends, and emerging patterns in repositories, covering additional subdisciplines or needs in chemistry. This ongoing task involves identifying and cataloguing repositories that fit our federation, considering subdiscipline coverage, data, functionalities, governance, technology, and user engagement, while monitoring new developments, policy changes, and technological advances.

T3.5.2 Onboarding of additional chemistry data repositories

We will

1. select international repositories that
2. meet our technical standards and can be interoperable within our federation. Based on (a) M3.1 specifications for technical interoperability and standard compliance and (b) benefits for the chemistry community, our goal is to
3. embed these resources in NFDI4Chem and enhance their accessibility and interoperability. New repositories will be
4. connected to our services, such as our (a) knowledge base, (b) OS FIZ-OAI, (c) Terminology and Vocabulary Services and (d) ELN-interface, as well as to new services of NFDI4Chem. Candidates that were already identified in FP1 include CSD and ICSD. Bilateral meetings led to the plan to integrate them into our search service. We will further
5. develop (a) a formal onboarding process, (b) classes of embedding degree, (c) assessment of infrastructure readiness, (d) onboarding documentation and (e) integration testing, risk management and reviews.

T3.5.3 Knowledge transfer to onboarded repositories with integration and FAIRification support

In FP2, we will

1. initiate a regular exchange forum for onboarded and federated repositories. These engagements will centre around (a) FAIR technologies and (b) adherence to our established standards, as well as (c) connection to our services (T3.5.2). We will offer
2. comprehensive technical support throughout the integration phase.

Table 4

Table 4.

Potential risks of the TA3 work programme and measures for their mitigation.

Description of management risks	Proposed risk-mitigation measures
RM1: Hiring and keeping qualified staff (software developers, DevOp- and IT-service operation experts). **Likelihood**: medium.	Use multiple channels to reach candidates. **TA(s) involved**: all
RM2: Continuous support for components developers. **Likelihood**: medium.	Embedding in frameworks such as deRSE, RSE4NFDI (see LoS). **TA(s) involved**: TA2, TA3, TA6
RM3: Reduction in funding with the effect that reliable infrastructure operation and maintenance cannot be guaranteed. **Likelihood**: medium.	Apply for additional funding. **TA(s) involved**: TA2, TA3, TA6
RM4: Risks in the case of budget cuts. **Likelihood**: high	See general description of this risk in TA1 **TA(s) involved**: all task areas

5.4 TA4 Metadata, Data Standards, and Publication Standards

TA4 defines and implements standards for metadata and data formats to ensure FAIRness, together with reference implementations and data validation, as well as standards for publishing workflows relevant to authors, academic publishers and developers of data and metadata infrastructures. Ontologies are used in metadata standards where possible, and missing terminological artefacts are added (in collaboration with M6.1). TA4 develops minimum information standards, harmonises metadata practices, and collaborates with international standardisation bodies like IUPAC to adopt and foster global standards. Moreover, TA4 cooperates with academic publishers to streamline the publishing processes of datasets along with corresponding scientific articles.

The development and maintenance of standards for both data and metadata are important to ensure the FAIRness (Wilkinson et al. 2016) and RIPEness (Reliable, Interpretable, Processable and Exchangeable, (McEwen and Bruno 2023)) of research data. With these goals, TA4 contributes to key objectives 2 and 5. Specifically, the objectives of TA4, designed to support FAIRness in NFDI4Chem, are:

O4.1: Create a set of **modular**, **machine-actionable** 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** ready for integration into semantically rich **linked open data**

O4.3: Increased **adoption** and **integration** of standards through generic **reference software** implementations

O4.4: Integration with the **scientific publishing processes**

O4.5: Contribution of standards and identifiers on (bio-)chemical entities **to the entire NFDI**

These objectives will be pursued through the following measures:

M4.1 Development and harmonisation of Minimum Information Metadata Standards

Goals: M4.1 contributes to the objectives O4.1, O4.2, and O4.4.

Description: Rich and harmonised metadata are required to make entities and datasets in chemistry FAIR. We will develop and maintain a set of modular, machine-actionable and interoperable metadata standards in key areas of chemistry. To be modular, we will develop requirements for Minimum Information about a Chemical Investigation (MIChIs) standards, which form the basis for chemistry-specific metadata schemas and mappings with and for the Metadata Schema Service (MSS, see M6.3). By including marginality in the MIChIs, we will achieve the key goal of identifying minimal common denominators. We will use a combination of semantically rich information following the linked open data principles (Bizer et al. 2009), using persistent identifiers and concepts obtained from terminology services.

There is still a real need for a thorough description of research objects, methods, tools and variables, which at present are partially covered by authoritative bodies like the IUPAC, e.g. via the Gold or Purple Book (Chalk and McEwen 2017, Jones et al. 2009), or the Joint Committee for Guides in Metrology (Joint Committee for Guides in Metrology 2012, Joint Committee for Guides in Metrology, Working Group 1 2008) though our partner PTB, or grassroots activities also in NFDI consortia (e.g. Metadata4Ing).

Our international network (IUPAC, RDA, PSDI), community engagement (in collaboration with M5.1), and interaction with neighbouring NFDI consortia will be used to discuss the MIChIs from the subdisciplines for all relevant chemical methods, e.g. through workshops. The harmonisation will be achieved through iterative refinement and feedback when implemented by researchers, infrastructures, journals, and participants from TA2, TA3, TA5, and TA6.

T4.1.1 Development and harmonisation of Minimum Information standards about Chemical Investigations (MIChI)

During FP1, several international workshops were conducted to bring domain experts together, to develop MIChIs for diverse methods and subdisciplines of chemistry. These typically result in guidelines and checklists, identification of (existing or missing) controlled vocabulary-derived term definitions and providing human and machine-readable examples.

In FP2, we will continue the discussion with domain experts in additional subdisciplines and for further methods used in chemical sciences to extend and broaden the coverage of MIChIs, including, but not limited to, organic chemistry, inorganic chemistry, polymer

chemistry, physical chemistry, medicinal and pharmaceutical chemistry, biochemistry, analytical chemistry, photo- and electrochemistry, environmental chemistry and theoretical chemistry This also requires vocabulary and minimum requirements for the simple semantic description of analytical tools, methods and results. We will increase the efforts in the area of computational and theoretical chemistry towards machine-readable and potentially automated computational protocols to allow for generic descriptions beyond individual simulation software solutions and workflows.

With a growing set of MIChIs in hand, we can identify a common chemistry core (CCC) of concepts that are applicable and re-usable in multiple subdisciplines. Harmonising their descriptions will lead to an interoperable CCC to be the basis of our modular and interoperable metadata standards These discussions will also include the Data and Service providers in NFDI4Chem to moderate and push harmonisation efforts across our federation of services (see M3.3 and M6.3).

T4.1.2 Modelling and Mapping metadata, linking to terminologies and existing standard models

Based on the description of CCC and subdiscipline-specific concepts resulting from MIChIs developed in T4.1.1, we will model and map the concepts into metadata schemas for the annotation of datasets that can be pushed and hosted in the MSS (see M6.3) We will ensure that generic domain-independent metadata in schemas of DataCite and Crossref can be linked with domain-specific chemistry specific metadata schemas and that the metadata is suitable to annotate and package datasets using e.g., RO-crate in our federation of repositories (see M3.3). Under the umbrella of computational chemistry community actions (e.g., euroSAMPL, TA5), the standardisation of computational chemistry protocols will be fostered.

The modular modelling of metadata schemas ensures that information is harmonised and can be mapped via cross-walks to various widely adopted and established metadata standards and will follow the recently published recommendations on the Cross-Domain Interoperability Framework (CDIF) developed as part of the WorldFAIR project (Gregory et al. 2024). The specifications of metadata schemas will aim at

1. improving the findability of data in both our generic and chemistry-specific repositories, also through the NFDI4Chem Search Service,
2. ensuring the searchability and accessibility via MSS (both in collaboration with M6.3), and
3. providing a rich description of chemistry data to improve interoperability and reusability aimed at knowledge graphs developed in M6.3, while using broadly applicable languages for its representation.

In all cases, we will make use of persistent identifiers and work with the PID4NFDI base service under development. We will document the use of the schemas, e.g. in the NFDI4Chem Knowledge Base and the MSS documentation, how to properly write, read, validate, and harvest metadata, considering both low-technology and high-level

solutions. Through the partners PTB and BAM, we expect to drive the adoption of the vocabularies and metadata schemas in norms and convey them to the industry, e.g., via regulatory requirements.

We will explore and recommend APIs to access and harvest metadata, including the well-established OAI-PMH protocol, where the FIZ OAI provider also supports cross-walks (see T6.3.4) between metadata schemes (e.g., DataCite, Dublin Core, RADAR). Newer approaches may integrate better with today's repository architectures, including Event Notifications following the push principle or the ResourceSync specification for Sitemap-based access.

M4.2 Development and maintenance of standards for data exchange and archival

Goals: Develop and specify standards for specific data types in chemical research, contributing to O4.1, O4.2, and O4.5.

Description: Data exchange and LTA require open formats for analytical data, as well as open formats to describe molecules and characterise chemical reactions. Existing open data formats will be examined, and areas of potential development will be determined, along with packaging methods with source vendor formats. The key goal is the implementation of metadata schemes, based on the MIChI standard developed in M4.1, for related data types/groups for a broad user base to maximise general acceptance and allow for a systematic extension/differentiation according to specific sub-community needs. Additionally, we will accommodate the experience of partners working on machine learning (ML) methods as key requirements for standards to improve the level of AI-readiness of standards-compliant data.

T4.2.1 Molecule and chemical reaction standards

In this task, we will further develop specifications of data standards for chemical data on molecules and chemical reactions also, including characterisation of enzymatic reactions, leveraging the experience of our partners in standardisation initiatives and data exchange formats while working with chemists, software and ontology developers to ensure a balanced focus on the chemical representation and the suitable use of the file formats in software and repositories. Our partner BI (in-kind) and USt will extend the EnzymeML data model to document analytics- and process-related metadata. Through partner PTB, we will include chemical safety as an integral part of our specifications. We will ensure that our models cover the extended concept of chemical reactions, including thermochemical conversion with its kinetics while implementing and contributing to the development of identifier systems for chemical mixtures and molecules of unknown structure through our activities in the InChI Trust. IUPAC´s International Chemical Identifier (InChI) is a unique representation of molecules, which is the basis for linking the data in ELNs and repositories. In FP1, we focused on the implementation of molecular inorganic chemistry into the InChI, which had been ignored for more than 20 years. This work is now almost accomplished (Herres-Pawlis et al. 2024) by the team of our co-applicant S. Herres-Pawlis (RWTH) in the Board of the InChI Trust. In the next phase,

inorganic stereochemistry and polymer chemistry will be integrated. Polymer chemistry in particular, represents a highly interesting industrial use case (see M5.5) because an international standard is lacking and urgently needed. BigSMILES (Lin et al. 2019) can deal with simple polymers but not complex architectures. The extension of InChI will be implemented into other services such as Chemotion ELN and repository (see T2.4.1). For the ML community, the new identifier MolBar also gains importance and will be regarded here (van Staalduinen and Bannwarth 2024).

T4.2.2 Analytical data standards

In this task, we will work with software developers, instrumentation companies, industry collaborators, publishers, repository operators and our international network (IUPAC, PSDI) on the specification of data standards for spectral data such as mzML, Allotrope Simple/Data Model (ASM/ADM), NMReData, nmrML and JCAMP-DX. This will include the enrichment of metadata based on the metadata schemes derived from MIChIs (see M4.1), thus improving the FAIRness of data format standards. Moreover, the community-established STAR format type CIF and NMR-STAR will be combined and augmented by further analytical and computational data for which standards and tags have to be defined. This results in the already described Molecular Information File (MIF), which will be revived, adapted and harnessed for synthetic and analytic molecular data.

An important aspect is automated metadata extraction, which improves the documentation of analytical data formats, along with ensuring unique and persistent instrument identifiers via the PIDINST initiative (Stocker et al. 2020, Krahl et al. 2021). Data sources include, but are not limited to, measurement devices such as NMR, mass spectrometry, X-ray diffraction, UV-Vis, IR, Raman, EPR, photoluminescence spectroscopy, and electrochemistry. The proper packaging of chemical data with provenance, usage policies, and contribution references will boost efficient data exchange between researchers, ELNs and repositories, where initiatives such as RO-Crate (see M2.3, M2.6, M3.2) will be explored.

M4.3 Implementation and support of software components for creation, validation, and consumption of standardised data formats

Goals: M4.3 contributes to O4.3, and O4.5.

Description: The availability of reference implementations is important to foster the broad adoption of data standards developed in M4.1 and M4.2. This requires readers, writers, validators, and converters, either as standalone software libraries or integrated into widely used chemistry software and libraries for commonly used programming languages, such as Python or Java. We will provide extensions to existing software libraries for standardised data formats such as EnzymeML (USt, Bi in-kind), mzML, ASM/ADM, NMReData, nmrML and JCAMP-DX, or where needed, develop tools supporting the data standards. Existing OS converters require broad test data. For use within workflow systems, tools like basic conversion and validation need to be described, e.g., the Common Workflow Language (CWL) (Crusoe et al. 2022).

T4.3.1 Reference implementation of conversion, extraction and validation libraries

Data standards defined in M4.2 need a software ecosystem that supports the input and output functionality to be available for software developers processing such data. The software needs to be usable both standalone and by users and allow batch execution in connection with Chemotion ELN and repositories in NFDI4Chem. In conjunction with the Semantic Data Hub (see M6.3), we will develop the extraction of metadata from data formats through the software libraries. The tool description of the Common Workflow Language (CWL) (Crusoe et al. 2022) allows encoding rich software metadata from authors and contributors to ontology-backed information about data formats for inputs and outputs. The CWL descriptions we develop will be deposited to public repositories like the common-workflow-library or WorkflowHub (Goble et al. 2021).

T4.3.2 Workflows for processing of standards-compliant data

Extensions for seamless data processing greatly enhance the value of standardised data formats: ingesting analytical and process-related data and metadata, enabling modelling of analytical data, including time course data, relating parameters with metadata, and publishing datasets in repositories. Such implementations should form an ecosystem that can be used in non-trivial workflows with little to no human intervention.

M4.4 Improving re-usability of scientific data through data standards

Goals: M4.1 contributes to O4.4, O4.1, O4.2.

Description: In this measure, we will collect, adapt, and demonstrate approaches to standardising the access to and re-use of scientific chemistry data. This involves specifically standards and standardised APIs to integrate between ELNs, repositories, and data processing and analytics infrastructures.

T4.4.1: Adopt and improve approaches for machine-actionable data and metadata

This includes efforts to establish prototypes of ML-ready data pools based on existing data. These data pools will be augmented and annotated and also published in NFDI4Chem repositories. The suitability for AI tasks will be directly assessed via the MLOps workflows (see next task), and described in user and developer-facing documentation and articles. Additionally, work in this task improves the FAIRness scores obtained through automated FAIR assessment tools. We will make sure that all improvements actually simplify data reuse, rather than just play games to game FAIR assessment tools.

T4.4.2: Enable machine learning on chemistry datasets

In recent years, ML has been established in chemistry for the analysis of data as well as the development of predictive and generative models. Nevertheless, the research of the application of ML methods and algorithms is a complex task that requires not only knowledge of ML but sufficient, unbiased data, computing resources, and tools. Thus, we will develop and provide workflows (MLOps) for chemists to consume the standards-

compliant data specified in this TA from our services. We will integrate existing tools, e.g. ML libraries, container orchestration (Kubernetes), ML development environments (collaborating with the JupyterHub-based service), and version control (GitLab) to create an easy-to-use ML development, testing, and execution infrastructure.

This will include, among other things, interfaces to data repositories, improving standardised access to data (in cooperation with M4.2 and M4.3), integrating Cloud and High-Performance Computing resources, encapsulating standard ML development procedures, automatic documentation for reproducibility of the development steps, and a user interface specific for chemists.

The development of this service by the partner TUDr will be in close cooperation with the chemists (e.g., requirements analysis, agile development with fast feedback cycles, joint testing, and evaluation) and include AI tools like MEGAN (Teufel et al. 2022, Sacha et al. 2020), and application of the EnzymeML platform to demonstrate the reuse of data by reanalysing previously published datasets which have been stored in standardised data formats.

T4.4.3: Approaches for the development and integration of hybrid experimental/ computational datasets

This task ensures that the data standards are applicable for integrating experimental and simulated (atomistic) data. Expanding the chemical space or covered properties in datasets has become an increasing trend by padding the latter with simulation results (the most prominent examples coming from material sciences). This is not yet a widespread practice in chemistry (particularly solution chemistry), but it would hold large potential for the streamlining of data acquisition and requires that data standards between them are interoperable. Some of the biggest challenges are cross-validation and normalisation, ensuring that the resulting hybrid datasets are internally consistent. There are also theoretical datasets across different areas with different provenance and, therefore, not interoperable (due to different levels of theory, numerical thresholds, etc.).

In this task, we will pursue the development of validation tools for repetitive measurements and closely related simulation experiments using multiple models. Data outliers can be identified through learning models applied on rather diverse chemical datasets, including but not limited to kinetic (Proppe and Kircher 2022), reaction thermodynamics (Tielker et al. 2020) and spectral data (Fischer et al. 2022). Another strength of the approach is that it is not only domain-independent but can also be applied to experimental and simulated datasets (and subsets thereof). The tools developed will provide standards for testing compatibility between different methods of data acquisition, enable aggregated datasets, and provide or improve uncertainty quantification in existing datasets.

M4.5 Integration with scholarly publishing

Goals: M4.5 contributes to O4.4.

Description: Academic publishing is the primary mode of communication for scientists, and publishing is an integral part of research on all levels. Many institutions and funders are beginning to mandate that research data be open, findable, and measurable. Publishers of academic journals have a keen interest in publishing high-quality articles to avoid jeopardising their reputations. Given the integral role of data as the foundation of any scholarly publication, academic publishers also recognise the importance of the publication, curation, and review of research data. Almost all have a data availability policy. However, researchers and publishers lack the tools and workflows necessary for the seamless curation, publication, and dissemination of research data. This is especially true in the area of the chemical sciences.

In this measure, we pursue tasks to improve data publishing processes, quality control mechanisms for research data, and simplify the reuse of machine-readable research data for scientists. This will require interaction between authors, editors of journals, reviewers, and developers of data and metadata infrastructures.

This measure also has great synergy potential with other NFDI consortia due to its domain-independent aspects, including processes to encourage, validate, review, and disseminate deposited and associated data and efficient retrieval and monitoring.

T4.5.1 Co-development of academic data publishing with publishers

As an integral part of academic publishing, data publishing is handled at different points along the publishing process. This may include the manuscript and data submission step, review processes (also see T3.3.5), data processing and formatting for machine-readability, metadata registration and linking, and presentation of the repository metadata in the published article and its DOI landing page, in addition to the delivery of the information to relevant databases and repositories.

We will work within our existing Editors4Chem format with publishers, where we can connect them with developers and repositories (see M5.6) to initiate pilot projects that would allow us to determine the feasibility, propose workflows to streamline data publishing in conjunction with research articles, define and recommend exchange standards and APIs, and identify required developments that will be summarised as recommendations (see also M4.1) to publishers and chemistry repositories.

T4.5.2 The Editors4Chem workshop

While each pilot project from T4.5.1 will be conducted with an individual or a smaller group of academic publishers under the umbrella of Editors4Chem, we will use the Editors4Chem Workshop to communicate the lessons learnt, standards, and best practices developed to streamline data publishing workflows and recommend APIs for (meta)data exchange to all academic publishers active in the field of chemistry. The Editors4Chem Workshop is a biennial workshop series that brings together journal editors of diverse academic publishers, developers of standards and infrastructures and enthusiastic chemists to work on measures to increase the deposition and appreciation of FAIR chemistry data associated with scientific publications. The workshop organisation is

a joint effort with the Committee on Publications and Cheminformatics Data Standards of the IUPAC (see M6.5).

Table 5

Table 5.
Potential risks of the TA4 work programme and measures for their mitigation.

Description of management risks	Proposed risk-mitigation measures
RM1: Inability by the community to **agree on MIChI standards** due to incompatible assumptions. **Likelihood**: Medium	We aim for a modular set of recommendations for the subdisciplines, avoiding that one standard has to fit all subdisciplines. Acceptance is promoted through community actions. **TA(s) involved**: TA2, TA3, TA6 and TA4
RM2: Unwillingness of instrument manufacturers to implement open data format standards or updates to already implemented formats to their software. **Likelihood**: Medium	We will engage with instrument manufacturers, include them in the MIChI-based metadata schema development as well as the subsequent specification of data format standards and support the implementation.
RM3: 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
RM4: **Objection by publishers** to adopt stringent author guidelines ensuring good scientific practice, especially those with a business model based on APC. **Likelihood**: Medium	Existing Editors4Chem workshop series allows Editors to listen to the chemistry community and adopt realistic author guidelines. **TA(s) involved**: TA3, TA2, TA4, TA5
RM5: Risks in the case of budget cuts. **Likelihood**: high	See general description of this risk in TA1 **TA(s) involved**: all task areas

5.5 TA5 Community and Training

Research data management (RDM) has received increasing attention in the scientific community in recent years. Major funding bodies such as the DFG and the EC have introduced requirements for research data, such as DMPs and RDM strategies. While the issue can no longer be ignored and is being embraced by some, it still faces much confusion and sometimes rejection in the community.

Facilitating the cultural change towards FAIR research data remains a long-term challenge. In our TA, it is crucial to be aware of the community's needs while supporting researchers in their challenges in producing and publishing FAIR research data. The cornerstones of our strategy are a professional communications strategy (Fig. 6), providing guidance through intensive training activities, and providing a wide range of best practice examples and use cases. To ensure the sustainability of our activities, we will focus on developing modular and multidimensional training elements, with a strong emphasis on train-the-trainer approaches.

We will also provide concepts for embedding RDM into curricula, ideally by integrating it into existing courses without requiring formal, time-consuming changes. We will make our open educational resources available in our NFDI4Chem knowledge base, YouTube, Zenodo, etc., and they will be documented via the BMBF-funded DALIA platform. DALIA is the result of cross-consortium work in the Edutrain section. To best meet the community's needs, we will work closely with all other TAs to ensure that these needs are met.

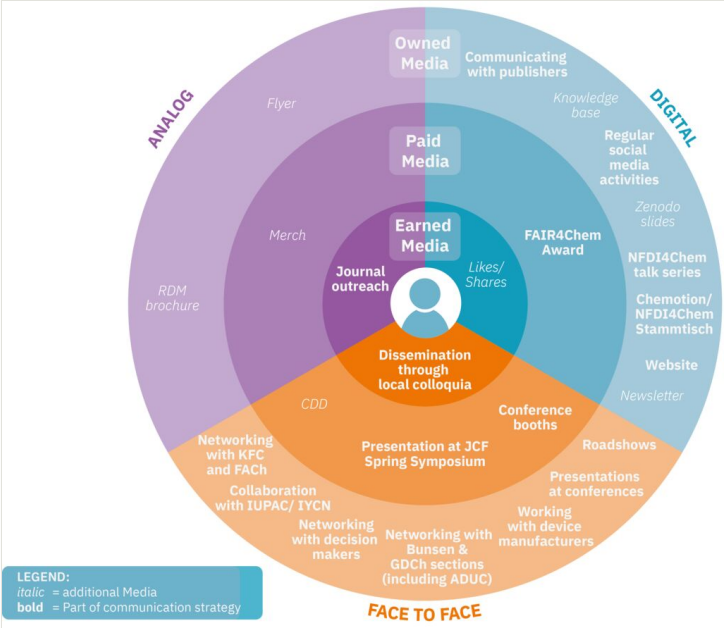

Figure 6. doi

Touchpoint analysis for the communication strategy of NFDI4Chem.

TA5 will contribute to the key objectives KO3, KO4, KO5, and KO6. More specifically, the **objectives** of TA5 are:

O5.1: Continuously 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

O5.4: Highlight best practices and FAIR-compliant machine learning applications

These objectives will be pursued through the following measures:

M5.1: Community requirements

Goals: Continuously collect community requirements and transform them into actionable tasks.

Description: This measure aims to continuously gather community requirements through surveys and events to understand and promote adopting FAIR data practices. These requirements will then be analysed, prioritised, and translated into actionable tasks to improve RDM tools and practices in different task areas.

T5.1.1: Continuous collection of requirements and needs from the community

As in FP1 (s. 3.1, 4.1), we will continue to gather the needs of our community and its subdisciplines in FP2. The next surveys are planned for 2026 and 2029. In addition to the surveys, we have found that collecting requirements at the NFDI4Chem conference booths and in workshops (Chemistry Data Days, RDM and Chemotion workshops) is also very efficient in obtaining high quality detailed requirements.Pharmacy-specific requirements for RDM are being collected via the DPhG and TUBr-UB (FID Pharmacy).

T5.1.2: Transformation of requirements into feasible tasks for other Task Areas

From the surveys and the directly communicated requirements, TA5 distils the requirements and discusses the prioritisation with the other TAs. Technically, many requirements for ELNs and repositories are correlated and joint solutions need to be found between TA2 and TA3. Here, the desired features of the RDM tools, new ideas for workshops and pages in our knowledge base are transformed into feasible tasks for TA2-TA6.

M5.2: Awareness

Goal: Raise awareness by communicating with the community in a variety of ways

Description: This measure uses the website, newsletters, and social media and societies to reach diverse stakeholders. Direct engagement with the chemical community will be sought through conference stands, presentations, and workshops.

T5.2.1: Website contents, newsletters, and social media in cooperation with societies

The communication strategy developed in FP1 includes the maintenance of the NFDI4Chem website as the first point of contact for news, events, and general information and as an entry point to the knowledge base (see T5.3.1). The touchpoint analysis (Fig. 6) shows that the full range of analogue and digital media (including social media such as LinkedIn, X, Instagram, Bluesky, Mastodon, etc.) is needed to reach all stakeholder groups ranging from pre-/post-graduate students, senior researchers to PIs, data stewards, librarians, publishers and many more (cite communication strategy when published). The societies play an important role as multipliers through their newsletters, but also through their social media accounts. The frequency of the NFDI4Chem newsletter will be increased to a bi-monthly basis as more results and news need to be communicated. A special feature of NFDI4Chem is the regular monthly "Stammtisch" (all-digital), where researchers report on their RDM solutions but also on topics such as machine learning in chemistry, news from the InChI development and new repos or ELN solutions. Besides the talk, which is recorded and posted on YouTube, the "Stammtisch"

features a longer and lively discussion between a mixed audience of chemoinformatics and RDM experts and benchtop chemists.

T5.2.2: Conference booths and presentations

Direct contact with the chemical community can be obtained easily via booths at general chemical conferences and rather specific conferences for chemical sub-disciplines. In FP1, NFDI4Chem was present at more than 30 conferences, which we plan to continue in FP2 to bring the new tools into the community. To present our results, we follow two strategies: a) Direct results are presented in talks at conferences, and we also invite talks to colloquia and webinars. These presentations are more interesting to the RDM and chemoinformatics audience and serve the (inter)national networking. b) To inform the chemical community, we present best practices on how to use the ELNs, repos and other new RDM tools at all important conferences in chemistry (many GDCh, DPhG, EuChemS, IUPAC, and selected ACS conferences). The Chemiedozententagung and the Spring Symposium of the GDCh Jungchemikerforum play an important role in special multiplier conferences. We have been present here with booths and workshops for the last few years and will continue this. We will also be present at RDM community conferences such as RDA-DE, CoRDI or E-Science Days to share developments and adapt new ideas to the needs of chemists.

M5.3: Training and Support

Goal: Provide training and support resources to enable the community in RDM

Description: This measure is about enhancing the NFDI4Chem Knowledge Base, developing tailored RDM training modules, conducting extensive training sessions, operating a help desk for support, and creating train-the-trainer resources for local RDM multipliers.

T5.3.1: Improve and extend existing information resources

During FP1, we established the NFDI4Chem Knowledge Base as a community effort based on GitHub and hosted at JGU,, which offers a repository guide and an interactive best practice section with examples of actual data publications. During FP2, we will continuously improve and extend the NFDI4Chem Knowledge Base. In particular, we will work on making it visually more appealing and reducing the length of the article's text without omitting important information. New participants are asked to add specific information about domains not currently represented in the knowledgebase (e.g. electrochemistry).

T5.3.2: Development of granular target- and phase-specific training modules

Comprehensive training is essential for a successful cultural change towards RDM. This training must be tailored to different audiences at different stages of their scientific careers. This requires modular training elements that can be interchanged to meet different needs. In particular, these modules should enable data stewards and local RDM

multipliers to develop chemistry-specific training themselves (see T5.3.5). In this respect, these modules are key to establishing a sustainable training infrastructure.

T5.3.3: Extensive on-site training activities on general RDM topics and ELNs

Training courses in general chemistry RDM and Chemotion ELN will become more important during FP2, and we will allocate more human resources to them. The general RDM in chemistry training comprises the basics of FAIR data, ELNs, repos, metadata, and DMPs. DMPs for chemists are a special focus of TA5 since the community requests them for numerous funded projects as well as collaborative projects. While these courses serve an important purpose, they also serve as a testing ground for optimising the training modules developed in T5.3.2.

T5.3.4: Operate a helpdesk unit

We have seen that the demand for direct support increases as new tools and workflows are implemented in the community. During FP1, we successfully implemented the NFDI4 Chem Helpdesk at TIB (707 successfully closed tickets from 384 requesters since 04/2021), which is the first point of contact for all kinds of questions or doubts. The helpdesk is operated by a first-level team that classifies incoming requests and forwards them to the appropriate second-level experts while maintaining communication with the requesters.

This activity will continue with a focus on improving efficiency. This will include the creation of SOPs from the cases to be disseminated through the knowledge base to facilitate self-service and to enable second-level support from more local stakeholders after FP2 to ensure improved sustainability.

T5.3.5: Develop comprehensive train-the-trainer resources to empower local RDM multipliers

During FP1, we have gained extensive experience in implementing RDM and ELN activities in research groups in CH, CZ, DE, DK, ES, Estonia, FR, IN, IT, NL, PT, SK, Slovenia, Thailand, TR, UK, and USA. We are, therefore, well aware of the challenges faced not only by researchers but also by local RDM stakeholders. We will identify suitable regional stakeholders to collaborate intensively in the creation of train-the-trainer resources. Local RDM teams at the universities are the ideal partners as multipliers. We strive for the sustainable integration of our materials into the RDM structures of universities. We will use the synergies arising from the activities in T5.3.3 to facilitate sustainable and long-term decentralised training activities, ultimately independent of NFDI4Chem.

M5.4: Curricular teaching

Goal: Implement chemistry-specific RDM early in academic education

Description: This measure will develop and promote curricular recommendations for integrating RDM content into existing chemistry curricula through practical exercises and refine these models for wider implementation in chemistry departments.

T5.4.1: Develop and promote curricular recommendations

In FP1 in 2021, the GDCh published the recommendations for the new chemistry curricula (Gesellschaft Deutscher Chemiker e.V. (GDCh) 2021). This was an important step with a strong signalling effect on the chemistry community, but curricula may only be changed very slowly. For FP2, we will work on the curricular recommendations and promote them further in the community, as well as continuously survey progress from direct contacts with chemistry departments. Besides the targeted inclusion of RDM content into chemistry curricula, we observed in FP1 that the sub-curricular implementation was easier and, in many cases, straightforward. Without changing credit points or administering lecture hours between chemical subdisciplines, RDM content can be successfully implemented in lab stages (Fink et al. 2023) and selected lectures.

T5.4.2: Develop and implement material for existing curricula

The implementation of RDM content in existing curricula is guided by practical experience: laboratory phases, which are an important part of chemistry studies, are enriched by RDM seminars and the use of ELNs by students. In FP1, we have gained several years of experience with the implementation of RDM content in Master's lectures (e.g. in the Inorganic Master's lecture in Aachen and an RDM lecture at the TU Darmstadt) and the use of ELNs in bachelor's laboratory courses (e.g. Inorganic Chemistry in Aachen and Organic Chemistry in Kaiserslautern). A student survey helped to monitor the implementation of the ELN and the integration of RDM and to improve the teaching materials and concepts (Fink et al. 2023). The three-year follow-up showed a growing awareness of RDM and increased acceptance of ELNs among students. The survey results highlight the need for and progress made in teaching RDM early in chemistry studies. Moreover, the possibility of integrating the topic as a practical experience is an easy way to integrate it quickly into teaching without changing the whole curriculum. Students like the hands-on approach and connect the theoretical knowledge of RDM directly to the synthesis of compounds in the laboratory. They can easily experience the benefits of RDM tools and gain meta-competence as digital skills for their future careers. For FP2, we plan to refine this model and roll it out to other chemistry departments. Additional materials for students will be prepared. In addition, RDM content will be implemented in the chemistry teaching degrees.

M5.5: Best practice

Goal: Highlighting best practice and machine learning applications with FAIR data

Summary: This measure highlights best practices across chemical disciplines and monitors evolving RDM tools, fostering industrial collaboration on standards and solutions. A particular focus will be on FAIR-compliant machine learning applications,

such as blind prediction challenges, to promote reproducible workflows and enable machine learning applications in FP2.

T5.5.1: Assemble use cases, visions, success stories

RDM poses varying challenges for different chemistry sub-disciplines, especially ELN interoperability when exchanging data. We will highlight best practices in our Knowledge Base and training workshops, attracting partners from all subdisciplines to provide diverse visions and success stories (see T2.4.5, T4.4.2). Given the dynamic nature of RDM tool development, continuous monitoring of new developments and their acceptance is necessary.We also aim for closer collaboration with industry, including joint development of standards (see TA4) and identifying best practices. A focus on patents and embargoes will enable new RDM solutions. For example, we are collaborating with BASF to develop a community-wide standard for polymers, such as the PolymerInChI (see TA4). Identifying homopolymers, cross-linked polymers, copolymers, and blends is crucial for recycling polymeric materials in the chemical industry.

T5.5.2: Maximising reproducibility by blind prediction challenge use cases

For FP2, we are planning next-generation blind prediction challenges, extending the scope of previous SAMPL and Fe-MAN (Rahrt et al. 2024) challenges, where the community will peer-evaluate submissions, in particular for FAIR+Rness ("fairerness", extending the FAIR principles by adding a reproducibility requirement), combining prediction accuracy with a FAIR score (Fig. 7). In this way, we aim at the "gold standard" of scientific publications at the interface of experimental and computational work, as described in Ref. (Heil et al. 2021), which will allow full reproduction of workflows "at the push of a button", thus enabling future continuous ML model training and deployment. We will develop an infrastructure for a "containerised" challenge, integrating experimental and theoretical data and ensuring low-threshold data transfer between local ELNs and federated infrastructures. We will extend the target observables to include predictions for compounds with high experimental uncertainty, thereby improving the reliability of ML models.

T5.5.3: Enabling machine learning applications as use cases derived from NFDI4Chem repositories

ML methods will require standardisation efforts (see T4.4.2) and community outreach. The use cases where data from repositories are reused for ML analyses are also highly valuable for the synthetic chemistry community. In particular, the lack of 'negative' data (i.e. plausibly expected but failed experimental approaches or measurements), which biases ML predictions, will be exploited by considering 'real world' data). To this end, standardisation efforts for computational chemistry will be essential to estimate experimental error from theoretical predictions (see T4.4.2).

T5.5.4: Select and present flagship labs with exemplary RDM implementations

Flagship labs inspire chemists to work more efficiently with RDM tools. In FP1, we developed a concept for using the Chemotion ELN in research groups with members ranging from interns to the PI (Fink et al. 2022). This addresses concerns about loss of laboratory time when ELNs are introduced. In the long term, groups will be more productive with a group-wide ELN, allowing members to share data more efficiently and quickly find older information.

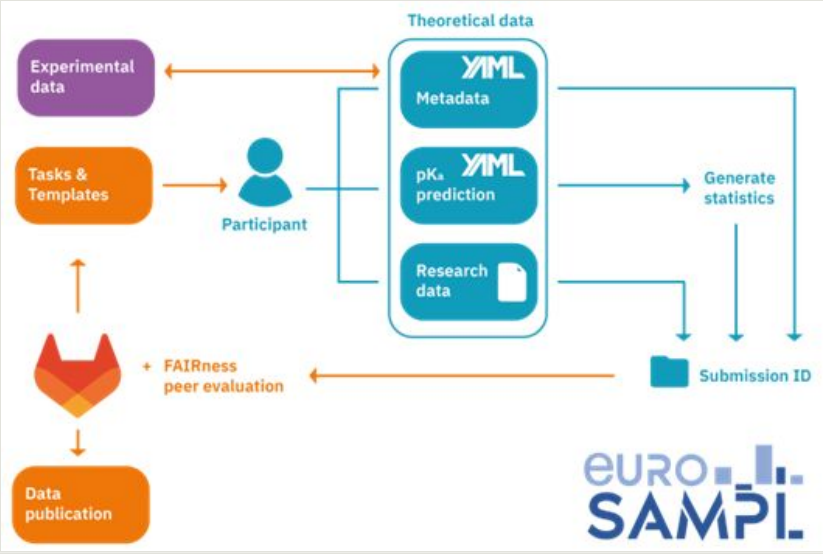

Figure 7. doi

Outline of the 1st euroSAMPL challenge. In line with the NFDI4Chem strategy, this blind prediction challenge serves as a best practice example of combined experimental and theoretical workflows with FAIR and reproducible data and code.

During FP1, we initiated the FAIR4Chem award to highlight excellent RDM practices from all corners of the community. The awardees presented their data-sharing solutions at the JCF conference, the Stammtisch and in a video. We will continue to run the FAIR4Chem Award to highlight further excellent FAIR data contributions from the community for the community. Flagship labs from all chemistry subdisciplines will be showcased in the Stammtisch (T5.2.1) and other formats.

M5.6: Community stakeholders

Goal: Networking between a multitude of stakeholders and contributing to standardisation

Description: This measure maintains ongoing stakeholder engagement to implement NFDI4Chem recommendations. It includes active participation in international

standardisation bodies and communication of results to promote community involvement in standardisation efforts.

T5.6.1: Exchange between community stakeholders continuously to ensure the implementation of NFDI4Chem recommendations

Ongoing interaction between the various stakeholders in the community is an important ongoing process. These stakeholders include, for example, learned societies, chemistry faculties, publishers (in collaboration with M4.5), industry or instrument manufacturers. We will continue to facilitate a continuous exchange between these groups, particularly on the work programme of the other actions in TA5 and other NFDI4Chem activities in general, such as standardisation or ontology development.

T5.6.2: Involvement in international standardisation bodies and communication of the process to the community

A particular aspect of NFDI4Chem's networking is its collaboration with international standards bodies such as IUPAC or the InChI trust (see T.4.3.1). In close coordination with TA4 (especially M4.1 to M4.3) and TA6 (M6.1, M6.5), we actively collaborate with these bodies at many levels and will continuously disseminate the results of this work to the community. This work has always been done on a largely voluntary basis. Therefore, parties willing to make an effort to work with standards bodies and working groups will have a tremendous impact on the outcome. We will encourage interested parties from the community to contribute to standardisation.

Table 6

5.6 TA6 Synergies and Cross-Cutting Topics

A hallmark of NFDI is the collaborative effort of consortia to create an overarching infrastructure of interconnected and interoperable services. NFDI4Chem has the same objectives in the field of chemistry. TA6 ensures the harmonisation and integration of services and data into NFDI4Chem and the NFDI-wide infrastructure, providing the Terminology and Search Service. These services will be extended to strengthen the semantic basis of chemistry research data and NFDI4Chem services. The main initiative of TA6 in FP2 is the Semantic Data Hub (SDH) (Fig. 8), which provides machine-readable FAIR chemistry research data and introduces the Metadata Schema Service and the Chemistry Knowledge Graph. TA6 will continue to formalise and coordinate the development of terminologies and contribute to data and metadata standards. TA6 also coordinates NFDI4Chem's contributions to NFDI-wide initiatives and working groups, including Base4NFDI and TS4NFDI, while integrating existing NFDI services into NFDI4Chem (e.g. IAM4NFDI). An overview of the contributions of NFDI4Chem is described in section 3.2. In addition, collaboration with international bodies such as RDA, IUPAC and projects such as PSDI in the UK will be strengthened. The overall goal for FP2 is to improve the interoperability, AI-readiness and machine-actionability of chemical research data, making it even more FAIR.

Table 6.

Potential risks of the TA5 work programme and measures for their mitigation.

Description of risks	Proposed risk-mitigation measures
RM1: 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. **Likelihood**: Medium	Results must be reviewed and critically discussed by the survey panel.Research communities must get involved in targeting the surveys. **TA(s) involved**: TA1, TA5
RM2: Developed tools, standards and materials are not adopted by the community, either by a mismatch of expectations or by ignorance. **Likelihood**: High	Development must constantly be monitored and critically discussed by the Steering Committee.Development activities must include early-stage hands-on testing by PhD students.External expertise and review must be acquired. **TA(s) involved**: TA1-5
RM3: Premature frustration of the community. The lack of usability or practicality of developed products may propagate reluctance to face challenges involved with RDM in everyday research. **Likelihood**: Medium	Careful testing of all products with representative user groups.User-friendliness by design. **TA(s) involved**: TA2, TA5
RM4: Risks in the case of budget cuts. **Likelihood**: high	See general description of this risk in TA1 **TA(s) involved**: all task areas

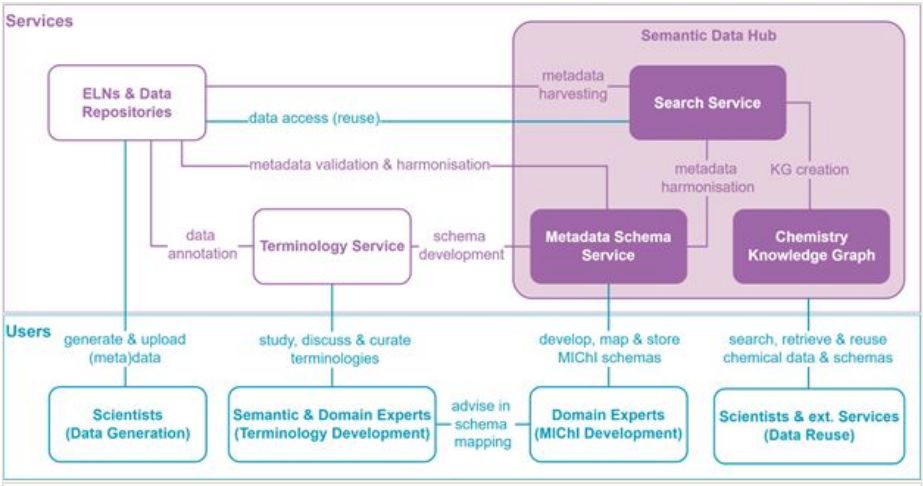

Figure 8. doi

NFDI4Chem services and data evolving into a Semantic Data Hub.

To create synergies, TA6 will contribute to individual aspects of most of the key objectives.

- Key objective 1: seamless integration of distributed data sources and uniform access to data;

- key objectives 2: minimum information standards and semantically rich linked data;
- key objective 3: seamless digital data workflows,
- key objective 4: technically reliable RDM infrastructure;
- key objective 5: collaboration with other consortia and promote cross-cutting developments and
- key objective 6: data re-use and enabling AI in chemistry.

The following objectives are derived for TA6:

O6.1 Users can develop, curate and apply terminologies to encode their domain knowledge into machine-actionable research data.

O6.2 Users can find and re-use research data across the NFDI4Chem federation and harness the full potential of semantically rich data.

O6.3 Users can use and access NFDI4Chem's services and data based on shared standards in a uniform and user-friendly manner.

O6.4 The services and data of NFDI4Chem are well-integrated and available within the NFDI and beyond.

M6.1: Ontology development, curation and harmonisation

Goal: Continue improving existing and developing new ontologies needed for FAIR chemical research data.

Description: To enable a circular use of research data, it must be annotated with rich and machine-actionable metadata mapped to FAIR terminologies. During FP1, we initiated community-driven actions to improve the availability and access to terminologies relevant to NFDI4Chem (Strömert et al. 2022b). Collaborating with data providers, curators, ontology engineers and researchers, we established workflows for curating ontologies like RXNO and CHMO. Addressing subdisciplines lacking coverage, we are developing the Vibrational Spectroscopy Ontology (VIBSO, Strömert et al. 2024c, Strömert et al. 2024b) with domain experts. With the annual Ontologies4Chem Workshop, we bring together all major stakeholders of the worldwide chemistry ontology community to elaborate on harmonising standards and pipelines for ontology management. In FP2, we will continue and increase our efforts to improve, develop and harmonise ontologies needed by the chemistry community by working on standardised pipelines for curation and deployment. Tasks will be supported by tools and features developed in the TS (M6.2) and the SDH (M6.3).

T6.1.1 Standardisation of ontology development workflows

This task focuses on establishing and promoting efficient workflows for developing and curating chemical ontologies, leveraging common tools and methodologies such as the Ontology Development Kit (ODK) (Matentzoglu et al. 2022) or shared design principles and patterns to enhance efficiency and interoperability. Embracing collaborative

platforms like GitHub for ontology maintenance and development calls for using the same tools, workflows and best practices to allow fast, iterative and coherent development by diverse contributors and to foster modular interoperability between ontologies. In FP1, we gained knowledge from developing VIBSO and contributing to existing ontologies in this regard (see 6.1.2.), which we aim to refine in FP2 to derive mature methodologies and best practices for the chemistry community, such as migrating RXNO, MOP and CHMO to an ODK-based setup. We will also evaluate the methodology behind the Voc4Cat SKOS (Simple Knowledge Organization System) thesaurus (Linke and Moustakas 2023) to streamline the ontology development process by creating a SKOS thesaurus for urgently needed common chemistry concepts. These efforts aim to improve the standardisation of ontology development, ensuring alignment with community needs and international standards.

T6.1.2 Domain ontology development

This task involves continuously contributing to community ontologies like RXNO, CHMO, CHEMINF, ChEBI, MOP, OBI, IAO and nmrCV, and developing new ontologies within NFDI4Chem, like VIBSO. We have identified the need to develop or extend ontologies for electrochemistry to enable the semantic data annotation of new data resources (Clark et al. 2021, Pingarrón et al. 2020), complementing TA4's MIChIs development (see M4.1) and TA2's semantic data annotation within ELNs. We follow a collaborative approach for ontology development, pairing domain experts (e.g. ETOS Cluster, UUlm) with ontology experts, ensuring robust and domain-informed ontologies. Building on T6.1.1's standardisation efforts and TS features (see M6.2), we will apply, evaluate and provide feedback on the workflows established in T6.1.1. in real-world scenarios to ensure their efficiency. We will also collaborate closely with NFDI consortia related to chemistry, like NFDI4Cat, NFDI4ING, NFDI4Health, NFDI4MatWerk or Dataplant, to ensure the interoperability between the used ontologies and to share the workload.

T6.1.3. Community engagement - workshop series on ontology development

We will continue the Ontologies4Chem workshop series in this task to develop and disseminate guidelines and best practices for ontology development and curation. These outcomes will be accessible to the broader chemistry community to promote adoption and establish ongoing collaboration and support. We will engage the chemistry ontology community, including members of the OBO Foundry, PSDI, and NFDI consortia related to chemistry and industry stakeholders like the Pistoia Alliance and Allotrope Foundation. We will deepen our collaboration with RDA working and interest groups (Research Data Alliance 2024a, Research Data Alliance 2024b, Research Data Alliance 2024c, Research Data Alliance 2024d, Research Data Alliance 2024e, Research Data Alliance 2024f, Research Data Alliance 2024g) and IUPAC (IUPAC 2024) to enhance international harmonisation. Additionally, we will participate in NFDI-related events and working groups to foster cross-domain ontology interoperability, continuing our leading role in the NFDI Section (Meta)data WG Ontology Harmonization and Mapping.

M6.2: Terminology service and mapping service

Goal: Enable researchers and services within NFDI4Chem and the broader NFDI community to access, curate, and update terminologies pertinent to chemistry and related domains.

Description: The NFDI4Chem Terminology Service (TS) is pivotal in promoting the interoperability, discovery and use of semantic data within NFDI4Chem and beyond. Built on EBI's Ontology Lookup Service (OLS) framework, we are currently upgrading the TS to the OLS4 codebase for enhanced functionality and performance. Following the analysis of suitable chemistry ontologies, we established the NFDI4Chem ontology collection and continuously integrated user requirements to support ontology development and curation during FP1 (see also M6.1). This systematic expansion aims to transform the TS into a comprehensive platform for ontology management, allowing users to perform curation tasks while monitoring dependencies with other chemical ontologies. Consequently, the TS has become a reference source for discovering and interacting with chemistry ontologies. In FP2, we aim to integrate more features derived from user requirements, including a mapping service in collaboration with the TS4NFDI Mapping Service.

T6.2.1 Collaborative ontology management integration

In this task, we focus on enhancing collaborative ontology management in the TS by integrating ontology update pipelines and providing the latest ontology version based on commits in the original repository. We will also archive TS-defined versions every six months, making them available via the API for downstream services. We will extend and integrate the prototypical NFDI4Ing TS Semantic Diff Tool to facilitate version comparison. Additionally, collaborative curation features developed in FP1 will be improved within the TS frontend by enhancing discussion mechanisms that leverage the existing notes and GitHub integration feature and by prototyping an extension of the existing term request feature with the SKOS thesaurus solution evaluated in T6.1.1.

T6.2.2 Mapping service

Mapping services are essential for integrating or transforming data from various repositories or research communities. Despite efforts to harmonise ontology development, mapping between overlapping or conflicting terminologies remains crucial due to the inevitable emergence of subdisciplines and projects with their own sets of terms and ontologies. In this task, we will implement the Mapping Service developed and provided by the TS4NFDI basic service. Mappings will be stored using the Simple Standard for Sharing Ontology Mappings (SSSOM). This implementation includes integrating widgets into the NFDI4Chem Terminology Service frontend to create, curate, and analyse mappings between ontologies and connecting to the TS4NFDI API. These mappings will also be utilised in the Metadata Schema Service (MSS) described in T6.3.3. We will also explore how complex mappings can be created and stored, either within NFDI4Chem or in TS4NFDI, to ensure comprehensive and coherent data integration.

T6.2.3 Integration of TS in basic terminology service

The backend of the NFDI4Chem Terminology Service is a component of the TS4NFDI basic service backend architecture, providing NFDI-wide access to a federation of terminology services. This task involves coordinating our contributions to the TS4NFDI project and addressing required developments for integration into the NFDI4Chem Terminology Service not covered by the TS4NFDI project. Development efforts will focus on extending the NFDI4Chem API and underlying backend functionalities to provide necessary API endpoints for the basic service API Gateway.

M6.3 Semantic Data Hub

Goals: Provide a central point of access to harmonised and integrated FAIR data from NFDI4Chem, culminating in a chemistry knowledge graph.

Description: The digitisation of chemistry data workflows and the integration of all developed services enable us to fully leverage the benefits of machine-actionable data. This measure aims to converge and link activities related to MIChI development, metadata schemas (see M4.1), and semantic data annotation (see M4.2), addressing shortcomings regarding semantic interoperability. (Vogt et al. 2024) FAIR Metadata generation happens early in the Smart Lab or later in repositories. It relies on terminologies and MIChI, provided by the Terminology Service (TS) and the Metadata Schema Service (MSS), which will be developed in this measure. The Search Service harvests this semantically harmonised metadata from repositories, with the ability to index new metadata by loading blueprints for corresponding new metadata schemas from the MSS. This approach enables seamless metadata ingestion into a knowledge graph from the Search Service later, facilitating sophisticated querying across all resources. The MSS for MIChI-derived metadata schemas will improve the development and application of schemas for FAIR data annotations. Combining it with the TS Mapping Service will allow schema crosswalks and, thus, data transformations.

T6.3.1 Search Service

The NFDI4Chem Search Service provides central, semantically harmonised access to the subdiscipline-specific repositories of the NFDI4Chem federation. During FP1, in collaboration with TA2, TA3 and TA4, we identified shared principles and requirements for metadata and semantic descriptions, empowering repositories to apply MIChIs for more detailed data descriptions. The harvesting and indexing by the Search Service, which now provides access to over 140,000 datasets from five repositories, are optimised and extended continuously for enhanced search options (i.e. by adding a molecule index or using the TS term look-up API). In FP2, we will extend the Search Service to apply MIChI schemas, allowing for the continuous integration of more repositories and subdiscipline MIChIs. We aim to extend searchable entities to include datasets, reactions, samples, and molecules. The user interface and APIs will also be extended for improved cross-repository data exploration and reuse.

T6.3.2 Development and implementation of a Chemistry Knowledge Graph

In this task, we will develop a Chemistry Knowledge Graph (KG) fed by the Search Service data stream (T6.3.1). We will provide a SPARQL endpoint for expert or programmatic access to the KG. To make this SPARQL endpoint more accessible, we will create an intuitive UI with graphical query builders and a library of customisable query templates, reducing the need for users to write SPARQL code directly. Additionally, we will explore natural language processing (NLP) capabilities to allow entering queries in natural language. We will collaborate with the KG4NFDI project to align our work with NFDI's basic service activities.

T6.3.3 Metadata Schema Service (MSS) for MIChI

We will develop a Metadata Schema Service (MSS) as a repository and registry for MIChI-based metadata schemas (see M4.1). The MSS will provide a GUI for TA4 to develop, store and serve metadata schemas, one for browsing, understanding, and relating these and other metadata schemas, and an API for machine-actionable access and operations such as metadata validation and schema crosswalks. This centralised service will save resources for chemistry data repositories and data scientists while ensuring standardised data outputs. It will partially depend on the TS4NFDI base service's mapping service, referencing mapping sets as metadata and using them for schema crosswalks. Additionally, the MSS aims to enable the creation and storage of complex mappings needed for schema crosswalks, extending the simple entity-to-entity mapping approach planned for the TS4NFDI mapping service.

M6.4 NFDI4Chem within the NFDI and basic services

Goals: Engage in NFDI sections, working groups, and task forces and contribute to cross-cutting topics relevant to the consortium, shaping processes in our best interest. Contribute to NFDI basic services and successfully integrate these into NFDI4Chem services.

Description: NFDI4Chem identifies itself as an integral part of NFDI. We are involved in the NFDI bodies and embody the NFDI vision and strategy to provide data as a common good

for excellent research organised by the scientific community in Germany. Members of NFDI4Chem contribute to the long-term development of the overarching architecture, the research data commons. We contribute our knowledge and expertise to the NFDI sections to facilitate overall developments and processes. We see the NFDI basic services, based on cross-community requirements and rooted in the NFDI sections and Base4NFDI, as integral components of the NFDI4Chem infrastructure. Following the objectives set out in FP1, we will continue to participate actively in developing and integrating such services.

T6.4.1 NFDI4Chem contributions to sections, working groups and task forces

Coordinating NFDI4Chem's contributions to NFDI sections is crucial. We ensure active participation in sections, working groups, and task forces established in FP1 or to be created in FP2. This task aligns NFDI4Chem's expertise with other consortia and feeds results into the consortium via workshops, guidelines, position papers, and shared use cases. NFDI4Chem members are represented in all sections and work groups, such as WG Ontology Harmonization and Mapping, WG Terminology Services, and Task Force Metadata within the section metadata; WG Data Management Planning, WG Electronic Lab Notebooks, and WG Identity and Access Management within the section common infrastructure; section ELSA covering legal aspects, and WP8 Error Culture in Science within the section Edutrain. Future work in the industry engagement section will be done with our industry advisory board.

T6.4.2 Coordination of basic services integration into NFDI4Chem infrastructure

Base4NFDI services are integral to NFDI4Chem. During the initialisation phase, we contributed to IAM4NFDI, TS4NFDI, KG4NFDI, and PID4NFDI by completing surveys, participating in workshops or incubator projects, i.e. testing the AAI test instances with the NFDI4Chem TS and Chemotion ELN. These efforts will continue in the integration phase by connecting the authorisation layer and using the IAM4NFDI infrastructure proxy, integrating our TS backend into the TS4NFDI infrastructure and contributing to its API Gateway, as well as collaborating on a mapping service and terminology curation. For PID4NFDI, we will contribute to developing PIDs for samples and instruments. For future basic services, we will introduce generic methodologies and processes for NFDI4Chem that will coordinate and actively support integrating basic services into NFDI4Chem’s infrastructure and services, including incubator projects with basic service projects. We also envision collaborating with KG4NFDI for the Chemistry Knowledge Graph in the Semantic Data Hub.

M6.5 Integration into the landscape of existing infrastructure and initiatives

Goal: Ensure NFDI4Chem's integration within the international infrastructure landscape to enhance interoperability, discoverability, and sustainability of chemistry data and services.

Description: Integrating existing infrastructures and initiatives maximises NFDI4Chem's visibility and impact. This involves strategic collaborations with international projects, monitoring relevant groups, and developing sustainable standards-based business models. NFDI4Chem is already involved in initiatives promoting interoperability, standardisation, and visibility of Smart Lab components. Expanding these efforts in FP2 will improve the FAIRness of NFDI4Chem data in chemistry portals and general search engines, making our data and services FAIR for the global research community.

T6.5.1 Identify and collaborate with relevant international infrastructure projects

This task focuses on identifying and establishing collaborations with relevant international infrastructure projects. We will actively monitor and engage with RDA, EOSC, and PSDI groups that share common goals. We aim to strengthen existing

strategic collaborations with organisations like IUPAC and CODATA, contributing to sustainable business models for standard development and curation. With our engagement within this global network, we will work on long-term solutions for MIChI (see M4.1), metadata schema and terminology curation, as well as on fostering the interoperability and standardisation of Smart Lab components, related workflows and formats, like the *.eln format and RO-Crate.Additionally, we contribute to initiatives like the ELN-finder, making international solutions more visible and understandable. With these diverse efforts, we aim to position NFDI4Chem as a pivotal player in the global chemistry data ecosystem.

T6.5.2 Improvement of findability of chemistry data resources in general search engines

For scholarly literature, numerous search, aggregation, and monitoring services exist and are widely used by researchers, including the NFDI4Chem Search Service. This task aims to improve the visibility of chemistry data resources in general search engines like OpenAIRE Explorer, Google Dataset Search, and DataCite Commons. By mapping our data to (Bio)schemas.org, we enhance its representation in general search results while preserving chemistry-specific metadata. We will highlight use cases such as finding data by individuals or institutions, data linked to publications, and publications citing datasets. We will engage with repositories and search services to identify bottlenecks and provide improved metadata. Using our international network, we will develop common approaches and data transformation guidelines for NFDI4Chem repositories to enhance their visibility.

Table 7

Table 7.

Potential risks of the TA6 work programme and measures for their mitigation.

Description of management risks	Proposed risk-mitigation measures
RM1: Unwillingness of the ontology curators to join collaborative ontology development efforts. **Likelihood**: low	Transparent community strategy about motivation, roadmaps, guidelines for contributions and collaborations. **TA(s) involved**: TA4, TA5
RM2: Increased personal resources required to cooperate with future basic services and their implementation. **Likelihood**: medium	Close communication NFDI sections and Base4NFDI. Prioritising implementations based on community needs. **TA(s) involved**: TA4, TA2, TA3, TA5
RM3: Required APIs and functionality of TS4NFDI are not provided in time or not at all. **Likelihood**: medium	Close coordination with the TS4NFDI project to identify rising time delays and initiate counter measures. Mapping may be stored in MSS meanwhile. **TA(s) involved**: TA4, TA6 and TS4NFDI project
RM4: Risks in the case of budget cuts. **Likelihood**: high	See general description of this risk in TA1 **TA(s) involved**: all task areas

6 Additional Aspects.

6.1 Equal opportunity and diversity

Equal opportunity and diversity are key aspects of staff and team development and will be ensured as outlined in M1.4. As part of this, NFDI4Chem strongly supports the Code of Conduct adopted by the NFDI and has published an addendum with additional points of particular importance regarding equal opportunities. Other activities to promote equal opportunity and diversity are integrated in the paragraphs above but will be highlighted here:

To promote equal opportunities in our target community, NFDI4Chem has participated in two IUPAC Global women's breakfasts, as well as in a panel discussion on critical reflections on RDM at the SDG Graduate Schools Alliance midterm conference, focusing on equal opportunities for researchers in the global south. We strongly believe that free access to our services promotes equal opportunity in science, as it supports researchers regardless of their institution or background. With free training programs and multilingual instruction materials, we enable scientists of all genders and ethnicities at any stage of their careers to digitalise their workflows, store their data in a FAIR way, and hence prepare for future digital developments both in academia and industry. Our helpdesk provides additional support for those with limited resources or experience. In surveys and personal consultations, we collect feedback from users, particularly those from underrepresented groups, to improve our services. In several online and in-person events, such as the Chemistry Data Days or the Stammtisch series, we create a platform for exchange and networking possibilities, especially for early career scientists, to help them succeed in their research careers. Our services are developed with a focus on user-friendly design and accessibility to ensure usability by a wide range of scientists.

Acknowledgements

This work was done by the participants of the NFDI4Chem consortium, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI4/1 – **Project number 441958208** (NFDI4Chem). Further, we want to thank Sabrina Krohm for her intenisve support in redacting and formatting this manuscript.

Disclosure: During writing, large language models (LLMs) were used to translate the summary and for editorial polishing, *i.e.* English language improvements and moderate shortening of human-written paragraphs. All LLM output was vetted by human authors before inclusion in the document.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Amelung L, Anthofer V, Danabalan R, Demandt É, Ebert B, Elschner E, Espinoza S, Eufinger J, Fuchsloch S, Götz B, Henzen C, Hunold J, Idda T, Jansen L, Krieger U, Rodrigues CM, Meister M, Miller B, Pitroff S, Popp C, Pushkina A, Schatlowski N, Schimmler S, Schneide C, Schörner T, Schwetje T, Seegert J, Trippel T, Weber J, Wössner E, Zinke W (2023) White Paper: Interim report reference. Zenodo. https://doi.org/10.5281/ZENODO.7688729
- Bach F, Binder K, Bonatto Minella C, Lutz B, Razum M (2023) NFDI4Chem - Deliverable D3.3.1: Gap analysis report for selected repositories. Zenodo. https://doi.org/10.5281/ZENODO.7602101
- Beilstein-Institut zur Förderung der Chemischen Wissenschaften (2024) STRENDA DB. https://www.beilstein-strenda-db.org/. Accessed on: 2024-8-04.
- Bender T (2024) Successful Workshop on FAIR Terminology Services. https://www.nfdi4chem.de/successful-workshop-on-fair-terminology-services/. Accessed on: 2024-7-23.
- Bertelmann R, Holzer AC, Horstmann W, Kressin L, Lange R (2024) Schwerpunkt "Digitalität in der Wissenschaft" der Allianz der deutschen Wissenschaftsorganisationen. Digitalität gestalten – Offenheit und Souveränität vorantreiben. 2023 - 2028. Zenodo https://doi.org/10.5281/zenodo.11189484
- Biedermann Labs (2024) SupraBank. https://suprabank.org/. Accessed on: 2024-8-04.
- Biernacka K, Bierwirth M, Buchholz P, Dolzycka D, Helbig K, Neumann J, Odebrecht C, Wiljes C, Wuttke U (2020) Train-the-Trainer Concept on Research Data Management (3.0). Zenodo https://doi.org/10.34657/5179
- Bierwirth M, Glöckner FO, Grimm C, Schimmler S, Boehm F, Busse C, Degkwitz A, Koepler O, Neuroth H (2020) Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung. Zenodo. https://doi.org/10.5281/ZENODO.3895208
- Bizer C, Heath T, Berners-Lee T (2009) Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5 (3): 1-22. https://doi.org/10.4018/jswis.2009081901
- Boehm F, Buchner B, Kipker D, Kuntz A, Petri G, Sax U, Schaar K, von Suchodoletz D, Vettermann O (2021) Sektionskonzept "Ethical, Legal & Social Aspects" (section-ELSA). Zenodo. https://doi.org/10.5281/ZENODO.5646929
- Bonatto Minella C, Fischer TG, Jolliffe JD (2023a) Analysis of the Landscape of Repositories for Chemistry in re3data. Zenodo https://doi.org/10.5281/zenodo.8347993
- Bonatto Minella C, Bach F, Jolliffe JD, Fischer TG (2023b) Repos4Chem - criteria for acquisition - for suggestion by NFDI4Chem for data providers. Zenodo. https://doi.org/10.5281/ZENODO.8199754
- Castro LJ, Fluck J, Arend D, Lange M, Martini D, Neumann S, Schimmler S, Rebholz-Schuhmann D (2023) Schema.org as a Lightweight Harmonization Approach for NFDI. Proceedings of the Conference on Research Data Infrastructure 1 https://doi.org/10.52825/cordi.v1i.280
- Chalk S, McEwen L (2017) The IUPAC Gold Book. Chemistry International 39 (3): 25-30. https://doi.org/10.1515/ci-2017-0307

- Chalk S, Munday S, Kroenlein K, McEwen L, Mustafa F (2024) WorldFAIR (D3.2) Training Package: FAIR Chemistry Cookbook. Zenodo. https://doi.org/10.5281/ZENODO.10711949
- Chemotion ELN contributors (2021) Chemotion ELN and repository. https://www.youtube.com/watch?v=tZHaP6DW-Dw. Accessed on: 2024-7-29.
- Chemotion ELN contributors (2022) Chemotion Electronic Lab Notebook (ELN). v1.3.0. Zenodo. URL: https://doi.org/10.5281/zenodo.6797527
- Chemotion ELN contributors (n.d.) A CLI tool to create and manage instances of Chemotion ELN. Github. URL: https://github.com/Chemotion/ChemOrc
- Clark S, Bleken FL, Stier S, Flores E, Andersen CW, Marcinek M, Szczesna-Chrzan A, Gaberscek M, Palacin MR, Uhrin M, Friis J (2021) Toward a Unified Description of Battery Data. Advanced Energy Materials 12 (17). https://doi.org/10.1002/aenm.202102702
- ComPlat (2024) Repository-Tracker. https://github.com/ComPlat/Repository-Tracker. Accessed on: 2024-8-03.
- Crusoe MR, Abeln S, Iosup A, Amstutz P, Chilton J, Tijanić N, Ménager H, Soiland-Reyes S, Gavrilović B, Goble C, The CWL Community (2022) Methods included: standardizing computational reuse and portability with the Common Workflow Language. Communications of the ACM 65 (6): 54-63. https://doi.org/10.1145/3486897
- Davies A, Patiny L (2021) NMRium browser-based nuclear magnetic resonance data processing. Spectroscopy Europe https://doi.org/10.1255/sew.2021.a18
- Dolcet P, Schulte M, Maurer F, Jung N, Chacko R, Deutschmann O, Grunwaldt J (2003) LablMotion Electronic Lab Notebook as Research Data Management Tool in Catalysis. Proceedings of the Conference on Research Data Infrastructure 1 https://doi.org/10.52825/cordi.v1i.334
- Eberl F, Jansen L, Miller B, Amelung L, Danabalan R, Demandt É, Deschler K, Eggert M, Engel J, Espinoza S, Ferenz S, Fritzsche F, Goedicke M, Götz B, Hennig C, Hofmann A, Hoffmann C, Hunold J, Idda T, Jojo T, Jutz R, Krieger U, Meister M, Nüst D, Pittroff S, Sauerland K, Schmidt C, Schneide C, Schwarz A, Schwetje T, Seegert J, Trippel T (2024) Collaborative work in NFDI. Zenodo https://doi.org/10.5281/zenodo.15880071
- Ebert B, Fluck J, Glöckner FO, Koepler O, Miller B, Schmitt R, Schrade T, Stegle O, Steinbeck C, von Suchodoletz D, Wagemann K, Knebes J, Kraft S, Seitz-Moskaliuk H, Sure-Vetter Y, Wössner E (2021) NFDI Cross-cutting Topics Workshop Report. Zendo. https://doi.org/10.5281/ZENODO.4593769
- Fink F, Hüppe HM, Jung N, Hoffmann A, Herres-Pawlis S (2022) Sharing is Caring: Guidelines for Sharing in the Electronic Laboratory Notebook (ELN) Chemotion as applied by a Synthesis-oriented Working Group. Chemistry–Methods 2 (10). https://doi.org/10.1002/cmtd.202200026
- Fink F, Hoffmann A, Herres-Pawlis S (2023) Results of a Three-Year Survey on the Implementation of Research Data Management and the Electronic Laboratory Notebook (ELN) Chemotion in an Advanced Inorganic Lab Course. ChemRxiv https://doi.org/10.26434/chemrxiv-2023-09ljg
- Fischer TG, Ditfe T, Rauh D, Parks NA, Liermann J, Schulze T, Neumann S (2023) NFDI4Chem - Deliverable D4.4.1: 50 Experimental processes and data publications using NFDI4Chem infrastructure and beyond. Zenodo. https://doi.org/10.5281/ZENODO.8137599

- Fischer TL, Bödecker M, Zehnacker-Rentien A, Mata RA, Suhm MA (2022) Setting up the HyDRA blind challenge for the microhydration of organic molecules. Physical Chemistry Chemical Physics 24 (19): 11442-11454. https://doi.org/10.1039/d2cp01119k
- FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur GmbH (2024) RADAR4Chem. https://radar.products.fiz-karlsruhe.de/de/ueber-uns/impressum-rechtliches. Accessed on: 2024-8-04.
- Gemeinsame Wissenschaftskonferenz (2018) Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018. https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/NFDI.pdf. Accessed on: 2024-7-30.
- Gesellschaft Deutscher Chemiker e.V. (GDCh) (2021) Empfehlungen der GDCh-Studienkommission zum Bachelorstudium Chemie an Universitäten. https://www.gdch.de/fileadmin/downloads/Service_und_Informationen/Downloads/Schule_Studium/PDF/2021_GDCh_Studienkommission_Druckversion.pdf. Accessed on: 2024-8-04.
- Glöckner FO, Diepenbroek M, Felden J, Overmann J, Bonn A, Gemeinholzer B, Güntsch A, König-Ries B, Seeger B, Pollex-Krüger A, Fluck J, Pigeot I, Toralf K, Mühlhaus T, Wolf C, Heinrich U, Steinbeck C, Koepler O, Stegle O, Weimann J, Schörner-Sadenius T, Gutt C, Stahl F, Wagemann K, Schrade T, Schmitt R, Eberl C, Gauterin F, Schultz M, Bernard L (2019) Berlin Declaration on NFDI Cross-Cutting Topics. Zenodo. https://doi.org/10.5281/ZENODO.3457212
- Goble C, Soiland-Reyes S, Bacall F, Owen S, Williams A, Eguinoa I, Droesbeke B, Leo S, Pireddu L, Rodríguez-Navas L, Fernández JM, Capella-Gutierrez S, Ménager H, Grüning B, Serrano-Solano B, Ewels P, Coppens F (2021) Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory. Zenodo. https://doi.org/10.5281/ZENODO.4605654
- González–Cebrián A, Bradford M, Chis AE, González–Vélez H (2024) Standardised Versioning of Datasets: a FAIR–compliant Proposal. Scientific Data 11 (1). https://doi.org/10.1038/s41597-024-03153-y
- Gregory A, Bell D, Brickley D, Buttigieg PL, Cox S, Edwards M, Doug F, Gonzalez Morales LG, Heus P, Hodson S, Kanjala C, Le Franc Y, Maxwell L, Molloy L, Richard S, Rizzolo F, Winstanley P, Wyborn L, Burton A (2024) WorldFAIR (D2.3) Cross-Domain Interoperability Framework (CDIF) (Report Synthesising Recommendations for Disciplines and Cross-Disciplinary Research Areas). Zenodo. https://doi.org/10.5281/ZENODO.11236870
- Heil BJ, Hoffman MM, Markowetz F, Lee S, Greene CS, Hicks SC (2021) Reproducibility standards for machine learning in the life sciences. Nature Methods 18 (10): 1132-1135. https://doi.org/10.1038/s41592-021-01256-7
- Herres-Pawlis S, Pelz P, Kockmann N, Gläser R, Richter M, Liermann J, Ortmeyer J, Heine I, Metzmacher A, Andres A, Münzmay A, Heuer J, Hagener M, Dierkes J, Wiljes C, Lindstädt B (2021) Sektionskonzept Training & Education zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V. Zenodo. https://doi.org/10.5281/ZENODO.5599770
- Herres-Pawlis S (2023) How to initiate the cultural change towards digital chemistry. https://blogs.rsc.org/chemspider/files/2023/11/Vortrag-Herrespawlis-RSC-071223a.pdf. Accessed on: 2024-8-03.
- Herres-Pawlis S, Bach F, Jung N, Koepler O, Liermann J, Neumann S, Razum M, Steinbeck C, Fink F, Hoffmann A, Ortmeyer J, Jollife J (2023) NFDI4Chem: from chemical

research data management to digital chemistry. Zenodo. https://doi.org/10.5281/ZENODO.8340453

- Herres-Pawlis S, Blanke G, Brammer J, Baljozovic D, Khan N, Lange F, Bänsch F (2024) Making the InChI FAIR and sustainable by moving to open-source on GitHub. ChemRxiv https://doi.org/10.26434/chemrxiv-2024-w6kws
- Herres-Pawlis S, Liermann JC, Koepler O (2020) Research Data in Chemistry – Results of the first NFDI4Chem Community Survey. Zeitschrift für anorganische und allgemeine Chemie 646 (21): 1748-1757. https://doi.org/10.1002/zaac.202000339
- Herres-Pawlis S, Bach F, Bruno I, Chalk S, Jung N, Liermann J, McEwen L, Neumann S, Steinbeck C, Razum M, Koepler O (2022) Minimum Information Standards in Chemistry: A Call for Better Research Data Management Practices. Angewandte Chemie International Edition 61 (51). https://doi.org/10.1002/anie.202203038
- Herrmann D, Hodapp P, Starman M, Huang P, Lin C, Le L, Klar J, Bizzarri C, Roese P, Oppel N, Fischer TG, Tremouilhac P, Holzhauer L, Herres-pawlis S, Hoffmann A, Seitz T, Zeitler K, Dorn A, Jung N, Bräse S (2023) Enhancing FAIRdata by providing digital workflows from data generation to the publication of data: an open source approach described for cyclic voltammetry. ChemRxiv https://doi.org/10.26434/chemrxiv-2023-95s19
- Huang P, Lin C (2024) LabIMotion. Zenodo. URL: https://zenodo.org/records/15083705
- Huang Y, Tremouilhac P, Nguyen A, Jung N, Bräse S (2021) ChemSpectra: a web-based spectra editor for analytical data. Journal of Cheminformatics 13 (1). https://doi.org/10.1186/s13321-020-00481-0
- Huang Y, Tremouilhac P, Kuhn S, Huang P, Lin C, Schlörer N, Taubert O, Götz M, Jung N, Bräse S (2024) (Semi-) Automatic Review Process for Common Compound Characterization Data in Organic Synthesis. ChemRxiv https://doi.org/10.26434/chemrxiv-2024-1r9tb
- Hunold J, Koepler O, von Maltzan S, Klix M (2023) Activity Update - Report of relevant cross-cutting topics for NFDI4Chem. Zenodo. https://doi.org/10.5281/ZENODO.8334137
- IAM4NFDI (2024) Incubator Dashboard. https://incubators.nfdi-aai.de/. Accessed on: 2024-8-02.
- IUPAC (2024) IUPAC Committees. https://iupac.org/who-we-are/committees/. Accessed on: 2024-8-04.
- Jacob C, Mück-Lichtenfeld C, Wolter M (2024) Research Data Management in Theoretical Chemistry. Institut für Physikalische und Theoretische Chemie https://doi.org/10.24355/dbbs.084-202407310658-0
- Joint Committee for Guides in Metrology (2012) International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM). 3rd edition. JGCM https://doi.org/10.59161/jcgm200-2012
- Joint Committee for Guides in Metrology, Working Group 1 (2008) Evaluation of measurement data — Guide to the expression of uncertainty in measurement. JCGM https://doi.org/10.59161/jcgm100-2008e
- Jones RG, Wilks ES, Metanomski WV, Kahovec J, Hess M, Stepto R, Kitayama T (Eds) (2009) Compendium of Polymer Terminology and Nomenclature: IUPAC Recommendations 2008. Royal Society of Chemistry [ISBN 978-1-84755-942-5] https://doi.org/10.1039/9781847559425

- Jung N, Tremouilhac P, Hodapp P, Herrmann D, Starman M, Mehzabin T, Klar J (2024) Integrating devices and converting data with Chemotion ELN. Zenodo. https://doi.org/10.5281/ZENODO.13120738
- Jupp S, Burdett T, Malone J, Leroy C, Pearce M, McMurry J, Parkinson H (2015) A new Ontology Lookup Service at EMBL-EBI. Workshop on Semantic Web Applications and Tools for Life Sciences. https://www.semanticscholar.org/paper/A-new-Ontology-Lookup-Service-at-EMBL-EBI-Jupp-Burdett/b83bfbfc1f2f08e5b88af5ef65ef2a8687ac4112. Accessed on: 2024-7-29.
- Karlsruhe Institute of Technology (KIT) (2024a) LablMotion Template Hub. https://www.chemotion-repository.net/home/genericHub. Accessed on: 2024-7-29.
- Karlsruhe Institute of Technology (KIT) (2024b) Chemotion Repository. https://chemotion-repository.net/. Accessed on: 2024-8-04.
- KIT Karlsruhe Institute of Technology (2024) Chemotion Docker repository. https://hub.docker.com/r/ptrxyz/chemotion/tags. Accessed on: 2024-7-29.
- Koepler O, Liermann J, Schön F, Herres-Pawlis S (2020) Forschungsdatenmanagement - Zeit für den Abschied vom analogen Laborbuch. Nachrichten aus der Chemie 68 (12): 20-23. https://doi.org/10.1002/nadc.20204095910
- Koepler O, Schrade T, Neumann S, Stotzka R, Wiljes C, Blümel I, Bracht C, Hamann T, Arndt S, Hunold J (2021) Sektionskonzept Meta(daten), Terminologien und Provenienz zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V. Zenodo. https://doi.org/10.5281/ZENODO.5619089
- Koepler O, Strömert P, Hunold J (2023) Activity Update - Report on 2nd Ontologies4Chem Workshop 2023. Zenodo https://doi.org/10.5281/zenodo.10260337
- Konsortialversammlung des Vereins Nationale Forschungsdateninfrastruktur (NFDI) e.V. (2022) Stellungnahme der NFDI-Konsortien zu Basisdiensten. Zenodo. https://doi.org/10.5281/zenodo.6091657
- Krahl R, Darroch L, Huber R, Devaraju A, Klump J, Habermann T, Stocker M, RDA PIDINST WG Members (2021) Metadata Schema for the Persistent Identification of Instruments. Research Data Alliance https://doi.org/10.15497/rda00070
- Linke D, Moustakas NG (2023) voc4cat - A SKOS vocabulary for the catalysis disciplines. From proof of concept to a reliable service. Zenodo. https://doi.org/10.5281/ZENODO.10200693
- Lin T, Coley CW, Mochigase H, Beech HK, Wang W, Wang Z, Woods E, Craig SL, Johnson JA, Kalow JA, Jensen KF, Olsen BD (2019) BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. ACS Central Science 5 (9): 1523-1531. https://doi.org/10.1021/acscentsci.9b00476
- MassBank consortium (2024) MassBank. https://massbank.eu/MassBank/. Accessed on: 2024-8-04.
- Matentzoglu N, Goutte-Gattat D, Tan SZK, Balhoff JP, Carbon S, Caron AR, Duncan WD, Flack JE, Haendel M, Harris NL, Hogan WR, Hoyt CT, Jackson RC, Kim H, Kir H, Larralde M, McMurry JA, Overton JA, Peters B, Pilgrim C, Stefancsik R, Robb SM, Toro S, Vasilevsky NA, Walls R, Mungall CJ, Osumi-Sutherland D (2022) Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. Database 2022 https://doi.org/10.1093/database/baac087
- McEwen L, Bruno I (2023) WorldFAIR Project (D3.1) Digital recommendations for Chemistry FAIR data policy and practice. Zenodo. https://doi.org/10.5281/ZENODO.7887283

- Neumann S, Bach F, Castro LJ, Fischer T, Hofmann S, Huang P, Jung N, Katabathuni B, Mauz F, Meier R, Nainala VCS, Rayya N, Steinbeck C, Koepler O (2023) Harmonising, Harvesting, and Searching Metadata Across a Repository Federation. Proceedings of the Conference on Research Data Infrastructure 1 https://doi.org/10.52825/cordi.v1i.202
- Neumann S, Andres A, Bach F, Bender T, Bonatto Minella C, Eberl F, Fischer T, Golub B, Harivyasi SS, Herres-Pawlis S, Huang P, Hunold J, Jollife JD, Jung N, Liermann JC, Nainala VC, Razum M, Koepler O, Steinbeck C (2024) Interim Report NFDI4Chem 2023. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e124977
- NFDI4Chem (2023) Ontologies4Chem Workshop Series. German National Library of Science and Technology (TIB). https://doi.org/10.25798/27ny-xb37
- NFDI4Chem (2024a) Helpdesk. https://www.nfdi4chem.de/helpdesk/. Accessed on: 2024-8-04.
- NFDI4Chem (2024b) Chemistry Repositories. https://www.nfdi4chem.de/repos/. Accessed on: 2024-8-04.
- NFDI4Chem (2024c) Electronic Lab Notebooks ELN. https://www.nfdi4chem.de/eln/. Accessed on: 2024-8-04.
- NFDI4Chem (2024d) NFDI4Chem Knowledge Base. https://knowledgebase.nfdi4chem.de/knowledge_base/de/. Accessed on: 2024-8-04.
- NFDI4Chem (2024e) TS4NFDI Community Hands-On Workshop. https://www.nfdi4chem.de/event/ts4nfdi-community-hands-on-workshop/. Accessed on: 2024-7-23.
- NFDI4Chem (2024f) Demo ELNs. https://demo.chemotion.scc.kit.edu/. Accessed on: 2024-7-29.
- NFDI4Chem (2024g) A workshop for institutions: FAIR Research Data Management: Basics for Chemists. https://www.nfdi4chem.de/de/fdm-grundlagen/. Accessed on: 2024-8-04.
- NFDI4Chem (2024h) Chemotion ELN Hands-on Workshop. https://www.nfdi4chem.de/event/chemotion-eln-hands-on-workshop/. Accessed on: 2024-8-04.
- NFDI4Chem (2024i) LabIMotion: Online Hands-on Workshop. https://www.nfdi4chem.de/event/labimotion-online-hands-on-workshop/. Accessed on: 2024-8-04.
- nmrXiv Project Group (2024) nmrXiv. https://nmrxiv.org/. Accessed on: 2024-8-04.
- OPC Foundation (2024) LADS - laboratory and analytical device standard. https://opcfoundation.org/markets-collaboration/lads/. Accessed on: 2024-7-29.
- Parks NA, Fischer TG, Blankenburg C, Scalfani VF, McEwen LR, Herres-Pawlis S, Neumann S (2023) The current landscape of author guidelines in chemistry through the lens of research data sharing. Pure and Applied Chemistry 95 (4): 439-450. https://doi.org/10.1515/pac-2022-1001
- Parks NA, Fischer TG, Herres-Pawlis S, Neumann S (2024) Editors4Chem: Implementing Publication Standards in Chemistry. RWTH Aachen University. https://doi.org/10.18154/RWTH-2023-10440
- Patiny L, Musallam H, Zasso M, Bolaños A, Kostro D, Wenk M, Ahkrin S, Jeannerat D, Ziegler E, Liermann J, Schloerer N, Wist J (2023) NMRium: ReactJS component to display and process NMR spectra. v0.44.0. Zenodo. URL: https://doi.org/10.5281/zenodo.8345822
- Patiny L, Musallam H, Bolaños A, Zasso M, Wist J, Karayilan M, Ziegler E, Liermann JC, Schlörer NE (2024) NMRium: Teaching nuclear magnetic resonance spectra

interpretation in an online platform. Beilstein Journal of Organic Chemistry 20: 25-31. https://doi.org/10.3762/bjoc.20.4

- Pearman-Kanza S, Knight N, Joliffe J, Ortmeyer J (2024) Mitigating Barriers and Maximising Opportunities in Chemistry Research. Zenodo. https://doi.org/10.5281/ZENODO.12723678
- Pingarrón JM, Labuda J, Barek J, Brett CA, Camões MF, Fojta M, Hibbert DB (2020) Terminology of electrochemical methods of analysis (IUPAC Recommendations 2019). Pure and Applied Chemistry 92 (4): 641-694. https://doi.org/10.1515/pac-2018-0109
- Proppe J, Kircher J (2022) Uncertainty Quantification of Reactivity Scales. ChemPhysChem 23 (8). https://doi.org/10.1002/cphc.202200061
- Proteau PJ (2023) Journal of Natural Products – New NMR Data Requirements and Editor Changes. Journal of Natural Products 86 (4): 653-654. https://doi.org/10.1021/acs.jnatprod.3c00281
- Rahrt R, Hein-Janke B, Amarasinghe KN, Shafique M, Feldt M, Guo L, Harvey JN, Pollice R, Koszinowski K, Mata RA (2024) The Fe-MAN Challenge: Ferrates–Microkinetic Assessment of Numerical Quantum Chemistry. The Journal of Physical Chemistry A 128 (23): 4663-4673. https://doi.org/10.1021/acs.jpca.4c01361
- Rauh D, Blankenburg C, Fischer TG, Jung N, Kuhn S, Schatzschneider U, Schulze T, Neumann S (2022) Data format standards in analytical chemistry. Pure and Applied Chemistry 94 (6): 725-736. https://doi.org/10.1515/pac-2021-3101
- Research Data Alliance (2024a) Harmonised terminologies and schemas for FAIR data in materials science and related domains WG. https://www.rd-alliance.org/groups/harmonised-terminologies-and-schemas-fair-data-materials-science-and-related-domains-wg/. Accessed on: 2024-8-04.
- Research Data Alliance (2024b) Vocabulary Services IG04/08/2024. https://www.rd-alliance.org/groups/vocabulary-services-interest-group/. Accessed on: 2024-8-04.
- Research Data Alliance (2024c) InteroperAble Descriptions of Observable Property Terminology WG (I-ADOPT WG). https://www.rd-alliance.org/groups/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg/. Accessed on: 2024-8-04.
- Research Data Alliance (2024d) Chemistry Research Data IG. https://www.rd-alliance.org/groups/chemistry-research-data-interest-group/. Accessed on: 2024-8-04.
- Research Data Alliance (2024e) Metadata IG. https://www.rd-alliance.org/groups/metadata-ig/members/. Accessed on: 2024-8-04.
- Research Data Alliance (2024f) Research Metadata Schemas WG. https://www.rd-alliance.org/groups/research-metadata-schemas-wg/. Accessed on: 2024-8-04.
- Research Data Alliance (2024g) Metadata Standards Catalog WG. https://www.rd-alliance.org/groups/metadata-standards-catalog-working-group/. Accessed on: 2024-8-04.
- Sacha M, Błaż M, Byrski P, Dąbrowski-Tumański P, Chromiński M, Loska R, Włodarczyk-Pruszyński P, Jastrzębski S (2020) Molecule Edit Graph Attention Network: Modeling Chemical Reactions as Sequences of Graph Edits. arXiv. https://doi.org/10.48550/ARXIV.2006.15426
- Scheidgen M, Himanen L, Ladines AN, Sikter D, Nakhaee M, Fekete Á, Chang T, Golparvar A, Márquez JA, Brockhauser S, Brückner S, Ghiringhelli LM, Dietrich F, Lehmberg D, Denell T, Albino A, Näsström H, Shabih S, Dobener F, Kühbach M, Mozumder R, Rudzinski JF, Daelman N, Pizarro JM, Kuban M, Salazar C, Ondračka P, Bungartz H, Draxl C (2023) NOMAD: A distributed web-based platform for

managingmaterials science research data. Journal of Open Source Software 8 (90). https://doi.org/10.21105/joss.05388

- Soltau K, Hofmann S, Jasinska A, Linés C (2024) FAIRness assessment as a catalyst of chemistry research repository transformation in Germany. Zenodo https://doi.org/10.5281/zenodo.13135009
- Starman M (2023) StarmanMartin/EfwBuilder: v1.0. v1.0. Zenodo. URL: https://doi.org/10.5281/zenodo.8337273
- Starman M, Held M, Jung N, Kirchner F, Eschke C (2024) ChemotionAPI. Zenodo https://doi.org/10.5281/zenodo.12569190
- Steinbeck C, Koepler O, Bach F, Herres-Pawlis S, Jung N, Liermann J, Neumann S, Razum M, Baldauf C, Biedermann F, Bocklitz T, Boehm F, Broda F, Czodrowski P, Engel T, Hicks M, Kast S, Kettner C, Koch W, Lanza G, Link A, Mata R, Nagel W, Porzel A, Schlörer N, Schulze T, Weinig H, Wenzel W, Wessjohann L, Wulle S (2020) NFDI4Chem - Towards a National Research Data Infrastructure for Chemistry in Germany. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e55852
- Stocker M, Darroch L, Krahl R, Habermann T, Devaraju A, Schwardmann U, D'Onofrio C, Häggström I (2020) Persistent Identification of Instruments. Data Science Journal 19 https://doi.org/10.5334/dsj-2020-018
- Strömert P, Hunold J, Koepler O (2022a) 1st Ontologies4Chem Workshop – Ontologies for chemistry.German National Library of Science and Technology. https://doi.org/10.25798/frnp-sn04
- Strömert P, Hunold J, Castro A, Neumann S, Koepler O (2022b) Ontologies4Chem: the landscape of ontologies in chemistry. Pure and Applied Chemistry 94 (6): 605-622. https://doi.org/10.1515/pac-2021-2007
- Strömert P, Limbachia V, Oladazimi P, Hunold J, Koepler O (2023) Towards a Versatile Terminology Service for Empowering FAIR Research Data: Enabling Ontology Discovery, Design, Curation, and Utilization Across Scientific Communities. In: Acosta M, Peroni S, Vahdati S, Gentile A, Pellegrini T, Kalo J (Eds) Knowledge Graphs: Semantics, Machine Learning, and Languages. 56. IOS Press [ISBN 978-1-64368-425-3]. https://doi.org/10.3233/ssw230005
- Strömert P, Hunold J, Chalk S, McEwen L, Koepler O (2024a) Guidance on how to use the IUPAC Gold Book as a canonical source for textual definitions in chemical ontologies. ChemRxiv https://doi.org/10.26434/chemrxiv-2024-fvzpq
- Strömert P, Lanza G, Hunold J, Neumann S, karillio (2024b) NFDI4Chem/VibrationalSpectroscopyOntology: v2024-09-23. Zenodo https://doi.org/10.5281/zenodo.10410722
- Strömert P, Lanza G, Hunold J, Stosch R, karillo (2024c) VibrationalSpectroscopyOntology. https://github.com/NFDI4Chem/VibrationalSpectroscopyOntology. Accessed on: 2024-8-03.
- Teufel J, Torresi L, Reiser P, Friederich P (2022) MEGAN: Multi-explanation graph attention network. arXiv. https://doi.org/10.48550/ARXIV.2211.13236
- TIB-Leibniz Information Centre for Science and Technology (2024a) NFDI4Chem Search Service. https://search.nfdi4chem.de/. Accessed on: 2024-7-29.
- TIB-Leibniz Information Centre for Science and Technology (2024b) Terminology Service. https://terminology.nfdi4chem.de/ts/. Accessed on: 2024-8-04.
- TIB-Leibniz Information Centre for Science and Technology (2024c) NFDI4Chem Search Service. https://search.nfdi4chem.de/. Accessed on: 2024-8-04.

- TIB-Leibniz Information Centre for Science and Technology (2024d) TIB Terminology Service Backend Swagger API Documentation. https://service.tib.eu/ts4tib/swagger-ui.html#. Accessed on: 2024-7-29.
- Tielker N, Eberlein L, Hessler G, Schmidt KF, Güssregen S, Kast SM (2020) Quantum–mechanical property prediction of solvated drug molecules: what have we learned from a decade of SAMPL blind prediction challenges? Journal of Computer-Aided Molecular Design 35 (4): 453-472. https://doi.org/10.1007/s10822-020-00347-5
- Tremouilhac P, Lin C, Huang P, Huang Y, Nguyen A, Jung N, Bach F, Ulrich R, Neumair B, Streit A, Bräse S (2020) The Repository Chemotion: Infrastructure for Sustainable Research in Chemistry**. Angewandte Chemie International Edition 59 (50): 22771-22778. https://doi.org/10.1002/anie.202007702
- Van de Sompel H, Klein M, Jones S, Nelson ML, Warner S, Devaraju A, Huber R, Steinhoff W, Tykhonov V, Boruta L, Meijers E, Soiland-Reyes S, Wilkinson M (2023) FAIR Signposting Profile. https://signposting.org/FAIR/. Accessed on: 2024-8-03.
- van Staalduinen N, Bannwarth C (2024) MolBar: A Molecular Identifier for Inorganic and Organic Molecules with Full Support of Stereoisomerism. ChemRxiv https://doi.org/10.26434/chemrxiv-2024-k40v5-v2
- Venkata C, Sharma N, Koepler O, Neumann S, Steinbeck C (2023) Ontology Elements. Zenodo https://doi.org/10.5281/zenodo.8212133
- Vogt L, Strömert P, Matentzoglu N, Karam N, Konrad M, Prinz M, Baum R (2024) FAIR 2.0: Extending the FAIR Guiding Principles to Address Semantic Interoperability. arXiv. https://doi.org/10.48550/arXiv.2405.03345
- Wahlster W, Winterhalter C (Eds) (2022) German Standardization Roadmap on Artificial Intelligence (2nd edition). German Institute for Standardization (DIN e.V.). URL: www.din.de/go/roadmap-ai
- Wiley-VCH (2024) Notice to Authors. https://onlinelibrary.wiley.com/page/journal/15213773/homepage/notice-to-authors. Accessed on: 2024-8-04.
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, 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