Trusted Chemistry Data for the Digital Age: Inside NFDI4Chem's Federated Infrastructure

Felix Bach[1*] and Christian Bonatto Minella[1]

[1]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Germany
[*]Corresponding author: Felix.Bach@fiz-Karlsruhe.de

Introduction

The National Research Data Infrastructure for Chemistry (NFDI4Chem) [1] is building a comprehensive and sustainable framework for the management, publication, and reuse of chemical research data. Embedded in an international context, NFDI4Chem collaborates with global initiatives that promote and standardise FAIR (Findable, Accessible, Interoperable, Reusable) [2] implementations and interoperable infrastructures, including Research Data Alliance (RDA), International Union of Pure and Applied Chemistry (IUPAC) , European Open Science Cloud (EOSC), FAIRsharing and GO FAIR. This ensures alignment with global standards and enables accessibility and relevance of data across the international and interdisciplinary scientific community. NFDI4Chem concentrates on data describing molecules, their properties and reactions, covering a wide range of data types, formats, and structures. These data are curated, published, and preserved within a federation of dedicated repositories such as RADAR4Chem, Chemotion Repository, nmrXiv, and STRENDA DB. [3]

Objective

NFDI4Chem aims to deliver high-quality chemical research data through seamless and trusted digital workflows. The overarching goal is to ensure that data produced by chemists are FAIR throughout the entire research lifecycle. The initiative further strives to make FAIR chemistry data publication a new standard in scientific research.

Results and Discussion

In order to identify those chemistry repositories that could form the nucleus of the envisaged virtual federation, Task Area 3 (Repositories) team, led by FIZ Karlsruhe, developed a list of criteria for selection. [1] A number of relevant repositories were identified. [1] With the aim of identifying gaps in coverage across data types and disciplines, the TA3 team conducted a gap analysis of existing relevant repositories. [3] This process involved individual interviews with repository leaders that covered a broad range of topics, including general information, metadata standards and ontologies, data content, and technical aspects such as the Authorisation and Authentication Infrastructure (AAI), Application Programming Interfaces (APIs), available services and functionalities, operating environments, software architectures, and workflows. The information collected was used to assess the current maturity and operational readiness of the repositories and to develop recommendations for addressing identified gaps through adaptation or, where necessary, new development. In 2023, a list of selection criteria for new repositories was created. [4] This list serves as a benchmark for extending the current federation to other chemistry repositories. In the same year, an analysis of the chemistry repository landscape in re3data revealed that, of the 12 resulting chemistry repositories that fulfilled the selection requirements, seven are part of the NFDI4Chem federation. [5] Although a detailed analysis of the impact of this infrastructure has not yet been published since evaluating all metrics on e.g. data reuse and interdisciplinary exchange requires time and longitudinal monitoring, some positive trends are already observable. The number of users and published datasets has steadily increased since the beginning of the consortium, suggesting a growing adoption and awareness within the scientific community. Furthermore, two new repositories, RADAR4Chem and nmrXiv, have been specifically developed to support chemists, a

community that previously lacked dedicated platforms for sharing multidisciplinary and nuclear magnetic resonance data. These initiatives demonstrate the infrastructure's capacity to address domain-specific needs and foster broader data sharing. In parallel, the consortium is strongly committed to promoting a cultural shift towards open data practices and to encouraging the evaluation of scientists based not only on textual publications but also on the quality and availability of their published datasets. However, this represents a significant change within the established culture of chemistry, where, with a few notable exceptions, only text-based research outputs are typically made public, while underlying datasets are often stored locally on institutional servers without associated metadata. As a result, these data are difficult to contextualize or reuse. Encouragingly, some repositories within the federation are now explicitly recommended by Wiley in their author guidelines, providing chemists with sustainable and trusted options for data publication that align with FAIR principles. An illustrative example of innovation within this federated infrastructure is RADAR4Chem, which integrates the Terminology Service (TS4NFDI) and a Large Language Model to automatically recognize relevant keywords and ontologies directly from user-provided metadata. This integration enhances metadata quality and semantic consistency, supports automated curation workflows, and significantly improves data discoverability and interoperability across repositories. It also illustrates how AI-driven methods can be effectively embedded into repository services to facilitate data publication and reuse in a sustainable and scalable manner. At the same time, several challenges remain: the federated nature of the infrastructure implies a high degree of heterogeneity among repositories, which can complicate automated curation processes and the consistent application of AI-based methods. Harmonising metadata standards, ensuring interoperability across scientific domains, and fostering a long-term cultural transformation towards open data will be crucial to fully realise the potential of this distributed ecosystem. Curation models vary across repositories, with RADAR4Chem supporting decentralised expert curation in dedicated workspaces and Chemotion employing manual curation enhanced by automated quality control and ongoing LLM-supported workflow development. The NFDI4Chem infrastructure represents a trusted, sustainable, and inclusive environment that enables chemists to publish high-quality data of all types and formats, reflecting the full diversity of chemical research. Within this infrastructure, researchers can ensure the long-term preservation and accessibility of their data while complying with the FAIR principles with minimal effort and without the need for contractual obligations. Its federated structure and adoption of semantic standards foster improved data reuse, interdisciplinary integration, and the development of machine-actionable knowledge. Moreover, AI-supported enhancements streamline curation processes, improve metadata completeness, and enhance data quality and compliance with community standards. Finally, NFDI4Chem strengthens collaboration with global initiatives and publishers, thereby reinforcing transparency, reproducibility, and open science practices across the entire discipline.

References

[1] C. Steinbeck, O. Koepler, F. Bach *et al.*, "Nfdi4chem - towards a national research data infrastructure for chemistry in germany," *Research Ideas and Outcomes*, 2020. doi: 10.3897/rio.6.e55852 .

[2] M. Wilkinson, M. Dumontier, I. J. Aalbersberg *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, 2016. doi: 10.1038/sdata.2016.18 .

[3] F. Bach, K. Binder, C. Bonatto Minella *et al.*, "Nfdi4chem - deliverable d3.3.1: Gap analysis report for selected repositories," *Zenodo*, 2023. doi: 10.5281/zenodo.7602102 .

[4] C. Bonatto Minella, F. Bach, J. D. Jolliffe *et al.*, "Repos4chem - criteria for acquisition - for suggestion by nfdi4chem for data providers," *Zenodo*, 2023. doi: 10.5281/zenodo.8199755 .

[5] C. Bonatto Minella, T. G. Fischer, and J. D. Jolliffe, "Analysis of the landscape of repositories for chemistry in re3data," *Zenodo*, 2023. doi: 10.5281/zenodo.8347993 .