NFDI4Cat

NFDI for Catalysis-Related Sciences

TDCC-NES

Natural & Engineering Sciences

Enhancing Research Data Management in Chemistry:

A Collaborative Approach for Catalysing Innovation in Germany and the Netherlands

Funded by

DFG Deutsche Forschungsgemeinschaft German Research Foundation

Project number
441926934

NWO

Project Code
ICT.TDCC.002.004

www.nfdi4cat.org
www.tdcc.nl/nes

Executive summary

This report presents a snapshot of research data management (RDM) initiatives within the chemistry domain in Germany and the Netherlands and areas where potential improvements could be made, as presented and discussed during the FAIR4ChemNL workshop, held on 4-5 June 2024 in Utrecht, the Netherlands. Not only will the results of this discussion be interesting to those wanting to enhance data-driven science, but readers are encouraged to replicate the format of this collaboration process so as to guide or contribute to the developments that are needed to transform RDM across chemistry research and other scientific fields.

The paper identifies current challenges and makes specific suggestions for technical and cultural solutions that could be implemented to catalyse the implementation of RDM practices across the chemistry discipline, as well as their adoption by the community. Some of these will be pursued as projects in Germany and the Netherlands from 2025 onwards, initially focusing on: the alignment of RDM practices and infrastructures; the collection, development and alignment of ontologies; the co-development of trainings to foster cultural change; the continuation of knowledge exchange between the two countries through regular interactions, and the organisation of joint events for the community.

The FAIR4ChemNL workshop was hosted by the Dutch Thematic Digital Competence Centre for the Natural and Engineering Sciences (TDCC-NES) and co-organised with the NFDI for Catalysis-Related Sciences (NFDI4Cat). Participants came from across chemistry and the supporting research environments.

Cover photo by L. Varat

DOI 10.5281/zenodo.15050550

Table of Contents

		Executive Summary	2
1		**Introduction**	4
2		**Current Landscape of RDM in the Netherlands, Germany, and Beyond**	5
3		**Digitalising Chemistry Research**	7
	3.1.	FAIR Data and AI: Transforming Scientific Research	7
	3.2.	Catalysis Research in the Era of Emerging Technologies	7
	3.3.	Use Cases: Data Preparation and Sharing in Catalysis Research	7
4		**Workshop Findings: Current Challenges**	9
	4.1.	Data Collection & Generation	9
	4.2.	Processing & Analysis	10
	4.3.	Publishing & Archiving	10
	4.4.	Metadata & Ontologies	11
5		**Workshop Findings: Uniting Efforts Across Borders**	12
	5.1.	Enhancing Cross-Border Collaboration and Knowledge Exchange	12
	5.2.	Building Blocks for Collaborative Improvement	13
6		**Infrastructure Designed by Researchers, for Researchers**	15
	6.1.	Centralized Data Repository for Storing Data and Primary Metadata	16
	6.2.	Personal Repositories in Catalysis Research: Addressing Privacy and Flexibility	16
	6.3.	PID Handle Server for Resource Identification	16
	6.4.	Web Application for Creating Extended Metadata TRIQ	16
	6.5.	Metaportal for Metadata-based Dataset Search	17
	6.6.	Vocabulary and Ontology Services	17
7		**Conclusion**	18
8		**Appendix A: Contributors to the White Paper**	19
9		**Appendix B: Workshop Details and Participants**	20

1. Introduction

Chemistry and catalysis-related research is rapidly becoming a data-intensive field. Despite some data making its way to scientific publications and open repositories, a significant portion remains unused and is ultimately lost. This is due to a multitude of factors, such as the lack of or limited access to robust e-infrastructure for data analysis, storage, and sharing; the frequently unstructured and proprietary nature of data; inadequate metadata; and a lack of awareness and willingness to share data. All of these barriers underscore the pressing need for Research Data Management (RDM) to unlock the full potential of chemistry research data.

RDM encompasses all activities related to the handling of research data throughout its lifecycle, from its creation to its preservation, publication, and reuse. Implementing effective RDM practices ensures the long-term value and impact of research data, as well as its wide accessibility and reusability by the broader scientific community. The overarching framework for guiding the transition to a more robust and sustainable data-intensive chemistry research landscape is provided by the FAIR (Findable, Accessible, Interoperable, Reusable) data principles, which aim to retain and maximise the value of all research data by ensuring that it is well-documented, accessible to authorised users, and suitable for reuse, both concerning technical aspects and usage rights. The transition to a FAIR data ecosystem necessitates a multi-faceted approach addressing both technical and cultural aspects of research. These span from adhering to the principles of scientific integrity from citing data sources transparently, to applying common data standards and unique persistent identifiers, developing comprehensive and uniform descriptive metadata, and providing clear and easy access to authorised users through appropriate infrastructure and software tools.

This white paper is the outcome of the two-day FAIR4ChemNL workshop, organised through a collaborative effort by the Thematic Digital Competence Centre for Natural and Engineering Sciences (TDCC-NES), the Dutch Research Council (NWO) Fundamentals and Methods of Chemistry Committee (FMC), and the German initiative NFDI for Catalysis-related Sciences (NFDI4Cat). The workshop, held on June 4–5, 2024, at Utrecht University brought together around 40 experts with diverse profiles (e.g., chemistry researchers from academia and industry, data stewards, e-infrastructure providers), affiliated with institutions across the Netherlands and Germany. The workshop aimed to facilitate the exchange of knowledge and experiences regarding various RDM initiatives and efforts within the chemistry domain in the two countries, identify persistent challenges, and explore collaborative strategies to address these issues across national boundaries. The first day of the workshop featured presentations highlighting current best practices and available solutions for RDM in catalysis-related sciences and computational chemistry, addressing all phases of the RDM cycle, from planning to storage and backup. The second day focused on identifying next steps and creating a collaboration plan, with participants divided into break-out groups to discuss topics related to 'data collection and generation', 'processing and analysis', 'publishing and archiving', and 'ontology and metadata development'. The workshop concluded with discussions on immediate actions, such as organising similar future events, co-developing RDM training materials, and advancing efforts to align ontologies at the EU level.

The key take-aways from this workshop are summarised, and a strategy outlined for future collaboration between Germany and the Netherlands, aimed at catalysing the implementation of FAIR data principles across chemistry disciplines, as well as their adoption by the chemistry research community.

***Figure 1**: Impressions from the event (Day 1). Photos by: L. Varat*

2. Current Landscape of RDM in the Netherlands, Germany, and Beyond

The Netherlands and Germany both feature well-established RDM ecosystems, characterised by a constellation of organisations and initiatives active on topics related to navigating the growing complexity of research data, reproducibility of research, and open science.

In the Netherlands, the National Coordination Point Research Data Management (LCRDM, Dutch region of the Research Data Alliance) is central to this landscape, bringing together experts from different sectors to collectively address RDM challenges and by acting as a point for collaboration, coordination, and knowledge sharing on RDM policy at national level. Organisations such as SURF (cooperative organisation for ICT in Dutch education and research), DANS (national centre of expertise and repository for research data), 4TU.ResearchData (research infrastructure, community, and training provider; collaboration of the four Dutch technical universities), and RDNL (Research Data Netherlands – a coalition of key data service providers including 4TU.ResearchData, DANS, SURF, and Health-RI) all offer research services and expertise relevant to research data management and archiving with proper support and data curation, as well as training activities and workshops on topics related to RDM. The Netherlands eScience Centre is an important actor in supporting researchers with software engineering, as well as its long-term sustainability and maintenance, enabling its proper reuse by research communities. Particularly relevant for the Dutch chemistry RDM landscape is the Big Chemistry consortium, funded by the Dutch National Growth Fund, aiming at integrating automated experiments with artificial intelligence (AI) to accelerate discoveries in molecular systems. The successful collaboration among academic and industrial partners in this consortium puts a particular emphasis on the unified RDM practices. Nation-wide networks such as the TDCCs, set up collaboratively by the Dutch academic community and NWO bring researchers, research supporters, and e-infrastructure providers together to collectively identify and tackle bottlenecks related to (among other topics) the adoption of RDM and FAIR principles, in line with their roadmaps and through dedicated calls funded by NWO[1].

Most research institutions in the Netherlands have established Digital Competence Centers (DCCs) or equivalent support structures (e.g., TUDelft-DCC, UDCC, WDCC, and others) to assist researchers with RDM, providing expertise in data stewardship, FAIR data practices, and digital infrastructure. They offer training, tools, and services to enhance data management across diverse disciplines. However, the RDM support is often not embedded within research groups and are not all focused on a specific domain or application. In general (and in contrast with Germany, see below), RDM efforts in the Netherlands tend to be broad in scope and therefore do not cater to the needs and challenges of a specific community or discipline such as chemistry. Partnering with efforts in Germany is therefore seen as crucial.

In Germany, multiple initiatives have emerged in recent years aiming to establish robust RDM frameworks. This includes local initiatives as collaboration between research institutions and universities, e.g., HeFDI in Hessen or SaxFDM in Saxonia, and local data competence centres that provide counselling, services and workshops in RDM. Most of those initiatives focus on a broad approach to RDM, with no specific orientation to individual scientific domains. However, a major step towards RDM in specific disciplines was taken in 2020 with the foundation of the National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur, NFDI) as a DFG-coordinated project. Within 3 years, a total of 26 consortia and the Base4NFDI alliance aimed at different scientific domains were funded within the project. Especially NFDI4Cat for catalysis-related sciences, NFDI4Chem for chemistry, FAIRmat for material sciences, NFDI4Ing for engineering, and DAPHNE4NFDI for data from photon and neutron experiments provide contributions to the development of infrastructure, best practices, and standards relevant for or adjacent to chemistry research. Examples of activities in the mentioned consortia include the development of a minimum information standard (MICHI) for data in chemistry within NFDI4Chem or the creation of standardised vocabulary for linked data (Voc4Cat) and collection of suitable ontologies in catalysis-related sciences within NFDI4Cat.

Many institutions across Germany maintain local repositories to store and manage research data. In addition, several domain-specific repositories have been developed to address the unique needs of different scientific disciplines. One notable example is Repo4Cat, a recently launched repository designed to serve as a comprehensive platform for data generated across all sub-disciplines of catalysis research. For experimental chemistry, repositories such as RADAR4Chem and the Chemotion Repository were created and are used for publication and archiving of data from experimental chemistry. Another prominent repository is NOMAD. Initially established in 2014 as part of a collaborative project between the Fritz Haber Institute (FHI) in Berlin and the Einstein Foundation Berlin, NOMAD was created to store and share computational material data. In recent years, through its integration into the NFDI consortium FAIRmat, NOMAD has expanded its scope to include experimental data from materials science and heterogeneous catalysis.

Despite the ongoing initiatives and the increasing awareness of RDM, the FAIR data principles, and open data within the

1 Brown, A., Groep, D., Teperek, M., Dedoussi, I., Sarkol, V., Som de Cerff, W., van Aert, E., & Blank, P. (2022). Roadmap Thematic Digital Competence Center - v1.0: Domain Natural and Engineering Sciences.

German scientific community in recent years, concrete practices for implementing RDM are only now beginning to take shape. This is also due to the fact that RDM is not provided in chemistry education and is only slowly being incorporated into curricula. Although universities and research institutions across Germany have increasingly centralised support structures for RDM and actively promote open science, the appointment of data stewards dedicated to supporting RDM within specific scientific disciplines, such as chemistry, is still sporadic. Comprehensive coverage across institutions has not yet been achieved. Addressing this gap is essential, as discipline-specific expertise is critical for tailoring RDM practices to meet the unique needs of chemistry research. The establishment of tools and services for RDM in chemistry research practices will be a focus of the upcoming years. This effort is further supported by the increasing awareness fostered through the NFDI initiative, which spans all scientific disciplines. Funding agencies are also progressively incorporating requirements for the implementation of FAIR (Findable, Accessible, Interoperable, and Reusable) and sustainable RDM practices as prerequisites for grant approval.

Further international initiatives, both within and beyond the EU (e.g., EU-MACE: European Materials Acceleration Center for Energy, EU-MINe: European Materials Informatics Network, the Allotrope Foundation, or DAEMON: Data-driven Applications towards the Engineering of functional Materials: an Open Network) aim to avoid duplication of standards and efforts, by promoting and democratising open science practices in chemistry and materials science, making them accessible to researchers and practitioners across diverse backgrounds and expertise. In this context, it is worth recalling the recent shift at the European level from an idealistic fully open-science concept, to a more pragmatic ''as open as possible, as closed as necessary'' approach, adopted to level the playing field with other scientific giants[2,3]. Each of these initiatives focuses on one or a few specific open science dimensions, yet, the overall activities span from the development of ontologies and repositories to automated and reproducible workflows, from training to co-design of effective policies. Among these initiatives, the DAEMON COST Action features numerous Dutch and German scientists in chemistry in its membership. It is active in open science training and dissemination activities to meet the multidisciplinary needs of the next generation of researchers in materials science. The initiative delivers robust, practical training in informatics and data management directly applicable to their specific chemistry research projects. Furthermore, DAEMON advocates for top-down policies to harmonise data management and sharing in the EU. The currently fragmented landscape is an obstacle to the circulation of skills and knowledge which underlies a pillar of innovation in Europe[4]. Finally, DAEMON advocates for stricter research data policies in journals, moving away from the "data available upon reasonable request" picture and closer to the high reproducibility standards enforced by journals championing a "data-first" mentality.

Despite the widely acknowledged importance of RDM and the numerous emerging initiatives in the Netherlands and Germany, as well as across Europe, its comprehensive integration continues to pose significant challenges across research domains. Discussions at the FAIR4ChemNL workshop highlighted several gaps and obstacles impeding the widespread adoption of RDM practices throughout the entire data lifecycle. The next section explores current issues and current challenges in greater detail.

2 Salazar, A., Wentzel, B., Schimmler, S., Gläser, R., Hanf, S., & Schunk, S. A. (2023). Chem. Eur. J., 29, e202202720. https://doi.org/10.1002/chem.202202720

3 European Research Executive Agency. (visited 2024). Open science. Retrieved December 3, 2024, from https://rea.ec.europa.eu/open-science_en#additional-resources

4 Council of the European Union. (2024). Much more than a market: Report by Enrico Letta. Retrieved from https://www.consilium.europa.eu/media/ny3j24sm/much-more-than-a-market-report-by-enrico-letta.pdf

3. Digitalising Chemistry Research

3. 1. FAIR DATA AND AI: TRANSFORMING SCIENTIFIC RESEARCH

The implementation of FAIR data practices enables the exchange and utilisation of shared knowledge and ultimately leads to the advancement of science. This section first describes how emerging technologies, such as artificial intelligence (AI), are impacting practices in chemistry research, more specifically in the discovery of novel catalysts. With use cases from the Netherlands and Germany, this section then illustrates the potential and some successful implementations of digital approaches and RDM practices in catalysis research.

3.2. CATALYSIS RESEARCH IN THE ERA OF EMERGING TECHNOLOGIES

The search for new catalysts or the improvement of established catalysts is facilitated by knowledge of the descriptors that determine catalytic properties. However, the functionality of catalytic systems is governed by interfacial and kinetic phenomena, with the relationships between catalyst structure and catalytic properties often being highly intricate. Large sets of reusable and accessible data provide the possibility to apply big data analysis which can unveil previously unseen relations between the properties of catalysts and the activity for given reactions. The knowledge obtained can then be utilised to make predictions for catalyst and process design, speeding up research.

Over the last few years, AI has emerged as a powerful tool in materials science and catalysis, capable of identifying nonlinear correlations and uncovering complex patterns in data. However, the inherent complexity of catalysis also introduces inconsistencies in experimental data, which can limit the effective application of AI in this field[5]. In addition, machine-readable and well-managed datasets (often beyond the scope of single institutions) are essential to train AI, but the lack thereof has impeded the full exploitation of the potential of these approaches in chemistry. To achieve the necessary pool of data for AI and predictive approaches, shared FAIR data is unavoidable and requires FAIR workflows from the beginning. The four case studies described below illustrate how adhering to open access, high-quality datasets has facilitated significant advancements in catalysis research.

3.3. USE CASES: DATA PREPARATION AND SHARING IN CATALYSIS RESEARCH

The section above stressed the importance of machine readability of catalysis data for it to be AI-compatible. In a pragmatic approach, experimental data can be prepared for digital analysis by adhering to standard operating procedures (SOPs)[6]. An important component of such workflows is the implementation of certified standards (e.g., industrial benchmark catalysts), which allow for the direct evaluation of measured catalysis data when published alongside the corresponding standard results. The most effective approach to making experimental data AI-compatible is the use of machine-readable SOPs in automated experiments[7]. This ensures the generation of standardised, comprehensive data and metadata sets that can be shared after publication in repositories. Recently, Bellini et al. investigated the CO oxidation catalysed by perovskites through a combination of rigorous experiments performed according to SOPs and AI[8]. A series of 13 ABO_3 (A = La, Pr, Nd, Sm; B = Cr, Mn, Fe, Co) perovskites was synthesised, analysed by advanced characterisation techniques and studied in catalysis. The resulting data set was modelled and evaluated using data science methods. In this way, a descriptor for the activity of perovskites in CO oxidation was found that depends both on parameters reflecting structural distortions and on the elementary properties of A and B. The descriptor was thus more general than previously found relationships and enabled the prediction of more active materials that contain chemical elements that were not part of the training set.

The second study, A cost-effective strategy of enhancing machine learning potentials by transfer learning from a multicomponent dataset on ænet-PyTorch, by Aisnadaa, Artrith et al., integrates ML with first-principles approaches[9]. The study addresses the challenge of developing machine lear-

5 Marshall, C. P., Schumann, J., & Trunschke, A. (2023). Achieving digital catalysis: Strategies for data acquisition, storage, and use. Angewandte Chemie International Edition, 62(30), e202302971. https://doi.org/10.1002/anie.202302971

6 Trunschke, A., Bellini, G., Boniface, M., Carey, S. J., Dong, J., Erdem, E., Foppa, L., Frandsen, W., Geske, M., Ghiringhelli, L. M., Girgsdies, F., Hanna, R., Hashagen, M., Hävecker, M., Huff, G., Knop-Gericke, A., Koch, G., Kraus, P., Kröhnert, J., Kube, P., Lohr, S., Lunkenbein, T., Masliuk, L., Naumann d'Alnoncourt, R., Omojola, T., Pratsch, C., Richter, S., Rohner, C., Rosowski, F., Rüther, F., Scheffler, M., Schlögl, R., Tarasov, A., Teschner, D., Timpe, O., Trunschke, P., Wang, Y., & Wrabetz, S. (2020). Towards experimental handbooks in catalysis. Topics in Catalysis, 63(13-14), 1683–1699. https://doi.org/10.1007/s11244-020-01380-2

7 Moshantaf, A., Wesemann, M., Beinlich, S., Junkes, H., Schumann, J., Alkan, B., Kube, P., Marshall, C. P., Pfister, N., & Trunschke, A. (2024). Advancing catalysis research through FAIR data principles implemented in a local data infrastructure: A case study of an automated test reactor. Catalysis Science & Technology, 14(21), 6186–6197. https://doi.org/10.1039/D4CY00693C

8 Bellini, G., Koch, G., Girgsdies, F., Dong, J., Carey, S. J., Timpe, O., Auffermann, G., Scheffler, M., Schlögl, R., Foppa, L., & Trunschke, A. (2025). CO oxidation catalysed by perovskites: The role of crystallographic distortions highlighted by systematic experiments and artificial intelligence. Angewandte Chemie International Edition, 64(6), e202417812 . https://doi.org/10.1002/anie.202417812

9 El Aisnadaa, A. N., Boonpalit, K., van der Kruit, R., Draijer, K. M., Lopez-Zorrilla, J., Miyauchi, M., Yamaguchi, A., & Artrith, N. (2025). A cost-effective strategy of enhancing machine learning potentials by transfer learning from a multicomponent dataset on ænet-PyTorch. The Journal of Physical Chemistry C, 129 (1), 658–669. https://doi.org/10.1021/acs.jpcc.4c06235

ning potentials (MLPs) with limited reference first-principles data. By leveraging the Open Catalyst Project 2020 dataset, a publicly available and comprehensive resource, Aisnadaa et al., pre-trained MLP models on subsets of the dataset and fine-tuned them using minimal ab initio data. This approach significantly improved the accuracy and generalisability of MLPs, particularly for simulating complex catalytic systems such as CuAu/6H2O. The study underscores how open data resources can reduce computational costs while enhancing simulation performance, enabling breakthroughs otherwise unattainable.

The third study, Predicting the Activity and Selectivity of Bimetallic Metal Catalysts for Ethanol Reforming Using Machine Learning, by Artrith et al.[10], highlights the transformative role of shared computational resources in catalyst design. By training ML models on publicly available transition-state energy datasets, the researchers identified key reaction pathways for ethanol reforming and screened a range of bimetallic catalysts computationally. Optimal catalyst configurations with high activity and selectivity were identified and experimentally validated, demonstrating how open datasets can bridge computational predictions with experimental outcomes. In the third study, Machine learning prediction and experimental verification of Pt-modified nitride catalysts for ethanol reforming with reduced precious metal loading, ML insights combined with experimental data guided the design of Pt-modified molybdenum nitride catalysts[11]. Computational models built on public datasets enabled the discovery of efficient, economically viable catalysts for renewable hydrogen generation while minimising precious metal usage. Together, these studies exemplify the critical role of data sharing in advancing scientific innovation and accelerating the discovery in catalysis.

An additional demonstration of effective implementation and exploitation of rigorous data management practices has been exemplified for the Quantitative Description of Metal Center Organization and Interactions in Single Atom Catalysts by Kevin Rossi and coworkers[12]. This study leveraged microscopy measurement conducted by two other labs (NUS and Leeds), previously reported in other publications. The availability of data and metadata about microscopy characterisation of single atom catalysts enabled the efficient development and rigorous validation and benchmarking of a machine learning pipeline to detect single atom metal centres for diverse resolutions, imaging conditions, and elements. The possibility of re-utilising data from collaborators fast-tracked this outcome and allowed us to explore a variety of systems and scenarios otherwise difficult to access without incurring large cost and time expenditures. In agreement with this spirit, all the code and scripts utilised to obtain results and plot figures have been made available open access. In this regard, we note how different characterisation communities are at different levels of maturity in open-science practices. Crystallographic data have a strong and established set of standards, while such a standardisation in the context of microscopy is instead lacking, or currently emerging, and urgent.

10 Artrith, N., Lin, Z., & Chen, J. G. (2020). Predicting the activity and selectivity of bimetallic metal catalysts for ethanol reforming using machine learning. ACS Catalysis, 10(16), 9438–9444. https://doi.org/10.1021/acscatal.0c02089

11 Denny, S. R., Lin, Z., Porter, W. N., Artrith, N., & Chen, J. G. (2023). Machine learning prediction and experimental verification of Pt-modified nitride catalysts for ethanol reforming with reduced precious metal loading. Applied Catalysis B: Environmental, 312, 121380. https://doi.org/10.1016/j.apcatb.2022.121380

12 Rossi, K., Ruiz-Ferrando, A., Akl, D. F., Abalos, V. G., Heras-Domingo, J., Graux, R., Hai, X., Lu, J., Garcia-Gasulla, D., López, N., Pérez-Ramírez, J., & Mitchell, S. (2024). Quantitative Description of Metal Center Organization and Interactions in Single-Atom Catalysts. Adv. Mater., 36, 2307991. https://doi.org/10.1002/adma.202307991

4. Workshop Findings: Current Challenges

On the second day of the workshop, discussion groups were organised to address distinct phases of the research data cycle (see Figure 2) to identify challenges in the workflow.

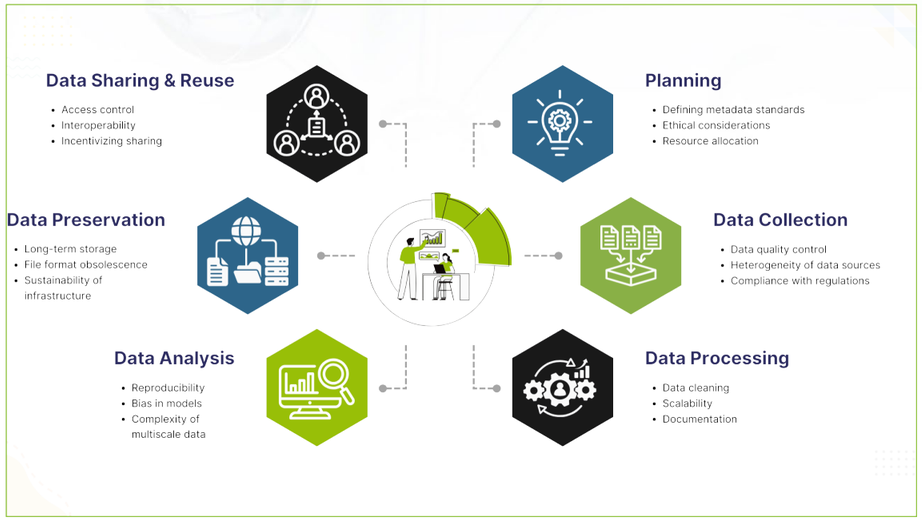

Figure 2: *The research data cycle and main challenges*

Each group focused on specific aspects of the research data cycle, exploring the current state of the art, envisioning future directions over the next five years, and developing a two-year actionable plan to achieve these goals.

The groups were structured as follows:

1. **Data Collection and Generation** – Representing the initial steps of the research data cycle, this group focused on tools such as Electronic Lab Notebooks (ELN).
2. **Processing and Analysis** – Addressing steps three and four of the cycle, the discussion centred on processing and analysis workflows and pipelines.
3. **Publishing and Archiving** – Combining the final steps of the cycle, this group examined strategies for effective data dissemination and long-term preservation.
4. **Ontologies and Metadata Development** – Operating as a cross-cutting topic, this group explored foundational elements crucial for ensuring interoperability and data reusability.

The following sections will highlight the key issues and challenges identified by these groups, providing insights into the barriers and opportunities across the research data cycle.

4.1. DATA COLLECTION & GENERATION

Electronic Lab Notebooks (ELNs) were introduced to advance beyond traditional notebooks and serve as the foundation for data collection and management in modern research workflows. They offer an opportunity to streamline the diverse information flows and integrate them with the research documentation within the predominantly digital research infrastructure of today. ELNs facilitate RDM by aligning with data generated from scientific instruments and analytical tools and make it easier to share the research data across teams and institutions. Moreover, such features as automated report generation can streamline the preparation of publications, particularly experimental sections, saving time and ensuring consistency. However, the landscape of ELNs is rather scattered, with over 85 active ELNs on the market, and support for standard protocols and data formats missing. Due to the heterogeneous nature of research in chemistry and catalysis-related sciences, ranging from organic synthesis and catalyst development up to process engineering or simulation-based approaches, no single available ELN solution can cover all necessary applications.

While a universal exchange format for interoperability between different ELNs is in development with the ELN exchange format, solutions for a lossless import and export between ELNs and/or laboratory equipment are still missing due to different internal representations and varying metadata fields. Furthermore, the broad implementation of ELNs into the research process is still in its early stages. Many institutions and universities either just recently introduced ELNs or are currently in a trial phase to find suitable solutions. Although free and open source ELNs are available, full integration into research workflows is impeded by costs for server capacity and support staff. In addition to these factors, tablets or similar electronic devices to take notes in the lab are not available at every institution. As a result, handwritten notes have to be transcribed

to the ELN, which increases the workload for researchers and thereby lowers acceptance by the broader community.

To increase user acceptance and establish FAIR RDM in the collection and generation of data, clear and practical usefulness and advantages of ELNs and similar tools have to be directly apparent for the user, without necessity for notable additional work. This includes search functionality for all collected data, automatic storage and backup of the documentation, tools for automatic interpretation and analysis of data as well as options to easily share data with selected persons, e.g., a supervisor or collaboration partner. Additionally, ELNs should provide robust mechanisms to relate and integrate entries --usually in the form of lab notes --with data stored in external locations, such as separate repositories or storage systems (often necessary due to size limitations). For example, large datasets, such as e.g. time-resolved microscopy data, are rarely stored directly within ELNs. Instead, only the metadata is kept in the ELN disconnected from the data itself. To address this, ELNs could offer dynamic linking features that remain intact even when data is moved to different locations or folders, ensuring that the connection between the experiment‘s metadata and the associated data is not lost. Such solutions would enhance usability and streamline workflows, particularly for complex or large-scale datasets common for the modern chemistry research. Tools for this purpose have to be user-friendly and simple to operate, and maintain a low barrier of entry, while facilitating robust integration between ELN entries and external data depositories.

Security is another critical requirement, particularly for controlling access permissions when sharing data. Additionally, ELNs need to be flexible to accommodate diverse research applications and compatible with a wide range of data formats. Features such as direct data import from measurement instruments with automatic annotation can further enhance their utility. The integration of technologies such as artificial intelligence, and more specifically Large Language Models (LLMs) to transcribe audio recordings into ELN entries, could also streamline notetaking, especially in challenging research environments, such as handling corrosive or toxic substances.

Beyond technical considerations, comprehensive training programmes for researchers are essential. These programmes should emphasise the benefits of ELNs through demonstrative use cases, illustrating how these tools can simplify workflows and enhance research efficiency.

4.2. PROCESSING & ANALYSIS

As with ELNs, the landscape of processing and analysis tools is notably fragmented. Workshop participants highlighted the widespread use of diverse tools, workflows, and best practices, with the selection often motivated by the specific scientific question to be tackled and the sub-discipline within chemistry. The three most commonly used tools identified by the group were JupyterHub, PyMatGen, and OriginLab. A significant challenge noted was that many tools lack user-friendliness, experience, and stability, resulting in frustration and inefficiencies for researchers. Additionally, proprietary file formats used by some lab instruments and widely-used data processing tools like OriginLab complicates effective data sharing.

A further challenge is the lack of a centralised platform providing an overview of processing and analysis tools, including associated documentation and tutorials. This makes it difficult for researchers to navigate the landscape and find suitable tools for their needs. This lack of coordination makes it challenging for researchers to navigate the available options and identify tools that best meet their needs. Tools developed for data management should prioritise stability and user-friendliness, resembling a platform akin to Kaggle but enriched with extensive metadata and descriptors.

Moreover, the efforts invested in developing and maintaining processing and analysis resources must be properly recognised. Participants emphasised the importance of valuing publications related to tool development and dataset compilation on par with traditional research papers. Direct citations of such resources were proposed as a mechanism to incentivise and reward these contributions.

Furthermore, the lifecycle of processing and analysis tools demands careful and strategic planning, encompassing decisions on support, maintenance, and eventual decommissioning. To address these challenges, participants recommended employing more dedicated research software engineers embedded within specific research groups. This approach would facilitate the development, proper maintenance, and sustained usability of these tools.

4.3. PUBLISHING & ARCHIVING

There is currently a lack of incentive for chemistry researchers to openly publish their research data, which is an important issue in a landscape where scientific publications remain the primary metric for evaluating a researcher’s performance. Data sharing is therefore not yet widespread, and its perceived value remains unclear, leading to it being regarded as a mandatory obligation for selected publications rather than a meaningful incentive. Furthermore, participants noted that approaches to integrating research data into the publishing process vary greatly across countries and journals. For instance, in the Netherlands, researchers are required to deposit the data underlying academic papers and theses, whereas in other European countries no standardised practice exists. Similarly, publishers have not enforced sufficiently stringent requirements for data sharing, accepting PDF documents or images as sufficient, limiting the accessibility and usability of research data and therefore, limiting opportunities for other researchers to build on existing work. This disparity underscores the critical need for mandating robust RDM practices at European, national, and institutional levels.

Moreover, it was pointed out that existing requirements often fail to ensure proper data openness or reusability. For instance, while researchers at many institutions in the Netherlands and Germany are required to archive their data, this process does not necessarily guarantee long-term accessibility or usability. A further barrier to data sharing is the lack of trust that data will be used responsibly, as well as the lack of the associated recognition. Furthermore, missing unclarity on how to cite datasets leads to further uncertainties and reluctance to adopt RDM principles.

A key challenge highlighted during the event was also the sharing of negative results, which is particularly critical for preventing duplication of efforts and for advancing AI-based approaches. For example, data on catalysts that failed to exhibit the desired activity for a specific reaction are rarely shared, even though such information is invaluable for helping other researchers avoid redundant work and for informing machine learning models. Furthermore, reproducibility of scientific results does not receive the necessary attention in the chemistry domain. Competitive publication practices and incomplete experimental descriptions often undermine reproducibility. Critical details, such as residence time during a reaction or the precise location of temperature measurements in a reactor system, are frequently omitted, further complicating efforts to replicate and validate research findings.

The findability of research data also presents a significant challenge, particularly in the absence of centralised repositories for data archiving. In Germany, for instance, universities typically maintain their own repositories, leading to varied and non-standardised data formats, unstructured deposition practices, and inconsistent naming conventions. While a fragmented structure with institutional repositories can work effectively (provided metadata formats are interchangeable or mappable), this requires centralised tools to connect them, such as dedicated search engines or registries. Platforms like Google Dataset Search or technologies like OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) can be used as the examples of the approaches to improve discoverability. To fully address these issues, dedicated data repositories equipped with domain-specific search tools, interoperable metadata standards, and the possibility for the seamless integration in the existing workflows are essential. These tools must be seamlessly integrated into existing workflows to minimise additional time burdens, which currently hinder the willingness to share FAIR data. Collaborative platforms with clear regulatory frameworks and proprietary restrictions for data sharing are also crucial to support effective and compliant RDM practices.

4.4. METADATA & ONTOLOGIES

Metadata and ontologies are pivotal for enabling the consistent description, organisation, and sharing of research data.

Metadata refers to structured information that describes, explains, or provides context to the research data, making it findable and usable. In the context of chemistry RDM, metadata refers to the detailed information that describes experimental or computational data (e.g. reaction conditions, instrumentation parameters, catalyst compositions, or simulation parameters). Ontologies, in the context of chemistry and catalysis, are structured frameworks that define the relationships between key concepts (e.g. molecules, reactions, catalysts, and conditions), providing a shared vocabulary to describe and organize data within these domains. Metadata and ontologies not only enhance collaboration but also drive advancements in scientific discovery by standardising how data is understood and utilised across disciplines. While collections of ontologies, such as those explored in Ontologies4Cat: Investigating the Landscape of Ontologies for Catalysis Research Data Management by Behr et al., provide valuable resources, the alignment and integration of these ontologies remain a significant challenge[13]. A lack of interoperability and standardisation across ontologies often leads to fragmented and inconsistent applications in RDM. This misalignment complicates efforts to establish universally applicable standards and to achieve seamless data exchange.

To address these challenges, it is imperative to improve the visibility and adoption of existing ontologies. Despite their application in various domains, inadequate documentation and limited dissemination of their practical uses have restricted their accessibility. Addressing these gaps would not only improve adoption rates but also enhance understanding across research communities.

A closer look at some ontologies and their potential applications underscores the breadth of the challenge and the opportunity for integration:

- **ChEBI (Chemical Entities of Biological Interest):** Provides unique identifiers for small chemical substances, widely used in biochemical research.
- **RXNO/CHMO/MOP:** Industrial-led initiatives aimed at connecting reaction databases for improved data interoperability.
- **CHEMINF:** Encompasses resources such as PubChem, eNanoMapper, and CDK, focusing on chemical informatics and compound characterisation.
- **DCAT (Data Catalogue Vocabulary):** Metadata schema used for data cataloguing.
- **Allotrope Foundation Ontology (AFO):** Focuses on metadata for laboratory equipment, especially in analytical chemistry.
- **FOAF (Friend of a Friend):** Originally designed for social networking (e.g., Facebook), highlighting the versatility of ontologies in various contexts.

A recurring issue in ontology usage is the lack of transparency regarding their implementation and impact. This limits researchers’ ability to assess their utility and integrate them effectively within their workflows. Clearer documentation, better dissemination strategies, and collaborative efforts to harmonise ontologies could address this shortfall, promoting the development of interoperable metadata standards.

The next section will highlight the ideas collected by the participants during the event to address the challenges described previously. These ideas aim to offer actionable strategies and innovative approaches for improving RDM across borders.

13 Behr, A.S., Borgelt, H. & Kockmann, N. (2024). Ontologies4Cat: investigating the landscape of ontologies for catalysis research data management. J Cheminform 16, 16. https://doi.org/10.1186/s13321-024-00807-2

5. Workshop Findings: Uniting Efforts Across Borders

5.1. ENHANCING CROSS-BORDER COLLABORATION AND KNOWLEDGE EXCHANGE

The rapid advancement of scientific discovery relies on high-quality, reliable data, which in turn demands an efficient approach to RDM. Nonetheless, significant constraints often impede researchers from fully incorporating RDM practices into their daily workflows. Having explored the key aspects and challenges of RDM in chemistry and catalysis-related sciences, we outline a collaborative framework between the Netherlands and Germany designed to enhance RDM skills, promote cross-border data sharing, and improve scientific outcomes through joint initiatives. This chapter presents a vision for seamlessly embedding RDM into daily research practices.

The foundation of this vision is the automation of RDM throughout the entire research lifecycle, ensuring that data is seamlessly linked, structured, and integrated in a way that enhances its immediate utility for researchers. For RDM to be genuinely effective, it must be intuitive, automated, and capable of providing real-time insights into research activities. The ideal system would automatically annotate and store data, minimising the need for manual input and enabling smooth integration into research workflows. The key to this system's success lies in its simplicity—an accessible, low-threshold platform that ensures data can be stored and annotated automatically, freeing researchers from the burden of manual entry and increasing efficiency.

A key aspect of our vision is fostering collaboration between Germany and the Netherlands through strategies that promote cross-border data sharing, knowledge exchange, and data sharing, in alignment with the European Union's philosophy of „as open as possible, as closed as necessary." This balance between openness and protection ensures scientific rigour while safeguarding intellectual property and addressing concerns related to the misuse of publicly available data, especially in the context of machine learning models. NFDI4Cat explores this framework in further detail in the perspective publication, How Research Data Management Plans Can Help Harmonise Open Science and Digital Economy Approaches[14]. Moreover, establishing an infrastructure that supports interoperability across different research domains and both countries will ensure that tools and metadata are accessible, adaptable, and flexible enough to cater to the diverse needs of scientists working across borders. In addition, researchers must be provided with flexible sharing options, allowing them to share data across countries with minimal effort while maintaining robust security measures and preserving their autonomy over their data. Customisable access levels will give researchers control over their data, ensuring that they can manage who has access, while also fostering collaboration.

To ensure inclusivity, this system should be adaptable to a wide range of data formats, with automated tools for annotation and integration into the research ecosystem. Another critical area of focus is the development of clear standards for data sharing. Establishing rigorous but practical guidelines for metadata will ensure data reusability, further promoting the value of data sharing. By creating robust support systems and offering incentives for data reuse, we can encourage researchers to embrace a culture of openness, which is vital for achieving reproducibility in scientific research. To demonstrate the potential of interoperable data and metadata, we advocate for pilot projects that showcase existing tools and their ability to support cross-border research. These projects can serve as proof-of-concept examples for other researchers.

Improving RDM skills among researchers is crucial for realising the full potential of this framework. Training and skills development programmes must target a diverse audience, including researchers and data stewards, ensuring that all stakeholders possess a clear understanding of proper data management practices. These programmes should focus on enhancing data literacy, particularly in areas such as documentation, repository use, and software lifecycle management. By offering hands-on training on data processing and repository management, researchers can better understand how to use these tools effectively. Examples from Germany, such as the RDM School of Catalysis offered by NFDI4Cat or DALIA highlight how targeted initiatives can address specific needs of the chemical community.

In the Netherlands, several general initiatives provide RDM training, such as the trainings in RDM practices and services and tools (e.g., Yoda, built on the iRODS software), offered by SURF and EuroCC, as well as several courses and workshops on data, programming and version control offered by 4TU. ResearchData. While these programmes help build foundational RDM skills, their broader focus means they often fail to cater to the chemistry community effectively.

For example, the RDM training modules for catalysis developed by the RDM School of Catalysis (Germany) could serve as a foundation for tailored programmes. These modules, initially designed for PhD students, could be adapted to meet the needs of data stewards within the domain and piloted in the Netherlands. Furthermore, the content could be aligned with the Carpentries framework to ensure scalability and inclusivity. This adaptation has the potential to connect with projects like the TDCC-NES bottleneck project, which focuses on developing discipline-specific training through the Carpentries framework. Workshops addressing Voc4Cat, focusing on catalysis-specific vocabulary and standards, were also proposed as an avenue for fostering deeper understanding and collaboration. Additionally, exploring connections with international bodies such as IUPAC could further amplify the reach and impact of these efforts.

Despite the promising potential of cross-border data sharing, several obstacles remain. Cultural resistance, particularly the competitive nature of research, often discourages openness and collaboration. Additionally, reluctance to share unique know-how can be a significant barrier. To address these challenges, we propose fostering a shift in mindset by emphasising the internal benefits of RDM, such as improved lab efficiency and the prevention of redundant work. Incentives that reward proper RDM practices, even in situations where immediate openness is not required, could also encourage adoption. NFDI4Cat is currently addressing this need through the Digital Chemist Award, presented in collaboration with Chemistry Europe, a Wiley-VCH journal. This annual award, given at the ADCR conference, honours scientists whose research and methods significantly advance the application of FAIR principles in catalysis and catalysis-related fields. The award is open for nominations from all community members.

To support the future vision of RDM, a shift toward a robust pipeline for data publishing and archiving is necessary. We recommend establishing a clear framework that guides researchers from data collection to publication, ensuring traceability, trustworthiness, and accessibility throughout. One possible approach is adopting dual publications, where data is published separately from research findings. This model would place greater emphasis on data FAIRness, enabling more thorough validation and reuse of scientific data. For example, some journals in the Nature Research family, PLOS, and Science have implemented policies mandating data sharing, where authors must make the data supporting their findings publicly accessible, often in dedicated repositories. Some journals also encourage or require a „Data Availability Statement" to specify where the data can be accessed and how it was handled, which further complements the extended methods section.

To advance this future vision of RDM, establishing a robust pipeline for data publication and archiving is essential. A notable example is the recently launched NFDI4Cat main repository, (Repo4Cat), which provides a platform for users to store, organize, and share data while ensuring long-term accessibility through persistent identifiers. Dutch participants from the workshop intend to explore the development of a tailored solution for their local chemistry research community, aligning with the standards and methodologies of NFDI4Cat to enhance interoperability and foster collaboration.

Alternatively, expanding the methods section in traditional journal articles could serve a similar purpose by including detailed descriptions of the data and metadata used in the research. Standardised metadata requirements across disciplines would further enhance the discoverability and reuse of research outputs, ensuring that the data is well-documented and accessible.

Ontologies play a critical role in making research data FAIR. By integrating ontologies into electronic lab notebooks (ELNs) and text mining efforts, we can enhance the discoverability and reuse of scientific data across disciplines. A recognisable system, like a "badging" system, could be introduced to indicate when ontologies are used in research outputs. This would signal to other researchers that the data is structured in a way that facilitates reuse.

Moreover, ontologies could be leveraged for automated annotation in ELNs and scientific literature, enabling consistent use of predefined metadata across research projects. This consistency is essential for achieving interoperability and ensuring that data can be easily shared and reused across borders.

Our proposed framework for cross-border collaboration in RDM sets the stage for seamless, integrated research in the digital age. However, achieving this vision will require continuous adaptation, innovation, and collaboration. By leveraging automation, fostering a culture of data sharing, and enhancing training and standards for data management, we can address current challenges and unlock the full potential of cross-border research. Furthermore, participants have committed to replicating this meeting format in 2025 and, if feasible, annually thereafter, to ensure continued progress and alignment on joint initiatives.

5.2. BUILDING BLOCKS FOR COLLABORATIVE IMPROVEMENT

To achieve seamless collaboration in digital catalysis and materials chemistry, the German and Dutch catalysis communities are prioritising the development of interoperable data management practices and infrastructure. Rather than constructing a joint system, the partnership aims to align existing infrastructures in both countries to facilitate efficient data sharing, collaborative workflows, and access to computational resources. This approach respects the unique structures and technological investments already established within each community while fostering a cross-border research environment.

A core component of this interoperability effort is the establishment of a standardised data-sharing protocol/procedure. This protocol will enable researchers to exchange datasets and metadata across institutional and national boundaries, ensuring that data from German and Dutch platforms can be readily interpreted and utilised in collaborative projects. To support this goal, both communities are working to define common data formats and metadata standards tailored to the specific needs of catalysis. This alignment will enhance data compatibility and allow researchers to retrieve, analyse, and cross-reference data from both systems without extensive reformatting or conversion processes.

14 Salazar, A., Wentzel, B., Schimmler, S., Gläser, R., Hanf, S., & Schunk, S. A. (2022). How research data management plans can help in harmonizing open science and approaches in the digital economy. Chemistry – A European Journal, 28(70), e202202720. https://doi.org/10.1002/chem.202202720

A crucial part of achieving interoperability is the collection and alignment of ontologies for chemistry and catalysis-related sciences. To ensure widespread adoption, metadata must be treated as the connection point to ontologies and FAIR data. Achieving this goal will require the development and widespread implementation of pre-defined metadata standards that bridge raw data with semantic frameworks. Consistent ontology integration must be prioritised, supported by hands-on examples and use cases tailored to the needs of specific research communities to incentivise the experimental and computational chemists to incorporate ontologies into their workflows as a fundamental part of their research practice.

In addition to ontology alignments and standardised data formats, the interoperability initiative includes the development of interfaces and RDM tools that facilitate data exchange between the German and Dutch infrastructures. These tools should be designed for non-experts to deposit and retrieve data easily and accompanied by documentation, tutorials, and validation through software/data papers. By establishing secure application programming interfaces (APIs), researchers from each community will be able to access data repositories, computational workflows, and shared software libraries hosted on either infrastructure. Clear lifecycle planning for these tools is essential to ensure their sustainability and relevance.

Training in digital technologies can help reduce reluctance to use new tools, while collaborative grants that require the exchange of tools and data can drive this transition. These measures will foster a research culture that values and evaluates openness and data sharing alongside traditional measures of academic excellence. Publications of common RDM tools and datasets should be valued equally to research papers, incentivised by direct citation to encourage adoption and reward contributions.

Finally, fostering interoperability also requires ongoing communication and feedback loops between German and Dutch researchers and infrastructure providers. Regular workshops, collaborative hackathons, and joint research forums are integral to this process, allowing teams to continuously assess the effectiveness of interoperability strategies and developed tools, and make adjustments as needed. This collaborative approach ensures that the infrastructure remains responsive to the evolving needs of researchers and is capable of integrating emerging technologies, such as AI or advanced simulation tools, as they become available.

6. Infrastructure Designed by Researchers, for Researchers

Chemistry researchers in the Netherlands and Germany already make use of various solutions supporting their research and data management endeavours across the data lifecycle. Mirroring the national landscape described in Section 2, RDM solutions offered in the Netherlands tend to cater to communities from multiple, diverse disciplines and domains, therefore featuring generic functionalities relevant to a variety of scientific applications. Conversely, the German NFDI consortium, NFDI4Cat, develops and subsequently provides a variety of services specifically tailored to researchers active in chemistry and catalysis-related disciplines.

The tables in section 6.1 provide an overview a few key services and tools offered by national e-infrastructure providers in the Netherlands. Some of the German services are then described below the tables in sections 6.2 – 6.7.

6.1. CENTRALIZED DATA REPOSITORY FOR STORING DATA AND PRIMARY METADATA

The three tables in this section contain information relevant to some key RDM services offered in the Netherlands by SURF, 4TU.ResearchData, and DANS, respectively. For more information on the specific services described here, and to know more about additional ones and/or upcoming plans, readers are encouraged to get in touch with these organisations.

Table 1: *A selection of RDM services offered by SURF, the cooperative organisation for IT in Dutch education and research.*

Service name	Functionalities	Description
Yoda Hosting	Relevant for entire data lifecycle	Yoda is an RDM software providing a web portal and WebDAV interface for users to securely store, share, archive, and publish their data throughout the research lifecycle in line with FAIR and Open Science principles. Its underlying infrastructure is iRODS, an open-source RDM platform that virtualises data storage resources, allowing users to manage their data regardless of its physical location by providing a common API for accessing data across diverse infrastructures. Yoda is developed by the University of Utrecht with financial support from the Yoda consortium. Through the Yoda Hosting service, SURF runs a Yoda instance for research and education institutions in the Netherlands, offering the power of iRODS and the benefits of an easy user experience.
Data Archive	Data Preservation	The Data Archive provides secure and long-term storage of research data, including very large (reaching tens of PBs) datasets which are not actively used. The iRODS data management system uses this service for the long-term preservation of data.
Object Store	Data Preservation	Object Store is an online storage service allowing users to store large amounts of research data made of diverse data types. It enables indefinite storage space expansion, while maintaining quick data accessibility. It is the default storage environment for Research Drive.
SURF Data Repository	Data Sharing, Re-use	SURF Data Repository offers a web interface allowing users to securely store and publish large (›1 TB) and diverse research datasets to ensure their long-term preservation and availability through the SURF Data Archive. The published data is stored on tape and is enriched with metadata and unique identifiers to ensure appropriate citation in publications. SURF Data Repository is a collaborative effort between DANS, 4TU and SURF.
ResearchDrive	Data Sharing, Re-use	Research Drive is a platform for storing and sharing research data, and synchronising with various devices. It is especially suitable when collaborating with other institutions, both at national and international level.
Persistent Identifiers	Data Sharing, Re-use	The persistent identifiers service provides researchers with persistent identifiers for their data, in cooperation with the Persistent Identifiers for eResearch (ePIC) consortium. It provides a stable reference to research data, ensuring continued findability even if its location or the underlying infrastructure changes.

Table 2: *A description of the 4TU.ResearchData repository offered by 4TU.ResearchData, a collaboration of the four universities of technology in the Netherlands (Delft University of Technology, Eindhoven University of Technology, University of Twente, Wageningen University & Research) providing research infrastructure and trainings relevant to research data and software management.*

Service name	Functionalities	Description
4TU.ResearchData repository	Data Preservation, Sharing, Re-use	The 4TU.ResearchData repository offers a trusted digital environment aimed at the permanent and sustainable preservation of technical-scientific research data, in line with FAIR principles. The repository provides persistent identifiers to make the data citable and ensures its readability by humans and machines. It also offers a variety of licenses to guarantee open access and reusability. The infrastructure on which the repository runs is available as Free and Open Source software, and the service is managed by the Library of Delft University of Technology, with data stored in two locations - in Delft, and a backup in Leiden.

Table 3: *A selection of RDM services offered by DANS, the national centre of expertise and repository for research data.*

Service name	Functionalities	Description
DataverseNL	Data Preservation, Sharing, Re-use	Dataverse is a research data repository where users can share, archive, and cite research data. DataverseNL is a repository co-provided by DANS (managing the technical infrastructure) and participating institutions, who are responsible for granting permissions to user accounts, managing and curating deposited research data.
Data Stations	Data Preservation, Sharing, Re-use	Data Stations provide domain-specific digital repositories for researchers to store datasets with detailed metadata, version control, and tools for automated information entry. These repositories use the Dataverse software, developed by Harvard University. Currently Data Stations are offered for the (i) Social Sciences and Humanities, (ii) Archaeology, (iii) Life, Health, and Medical Sciences, and (iv) Physical and Technical Sciences.
Data Vault	Data Preservation	The Data Vault is a secure, reliable, and certified long-term preservation repository that contains all DANS datasets.

6.2. CENTRALISED DATA REPOSITORY FOR STORING DATA AND PRIMARY METADATA

The NFDI4Cat infrastructure for RDM in catalysis, initially described in NFDI4Cat: Architecture document by Boenisch et. al., provides a structured and adaptable framework for managing research data across centralised and local repositories[15]. Its design emphasises centralised and local repositories, resource identification, metadata enrichment, and discoverability, addressing the diverse needs of researchers in catalysis and materials science.

The centralised data repository (Repo4Cat) is designed to streamline access to research data and ensure consistency in data storage, providing a reliable foundation for the effective management of research outputs. In this context, primary metadata refers to the essential information necessary for identifying and citing datasets, which typically includes citation metadata such as authorship, publication date, title, and a unique identifier.

Additionally, the centralised repository provides links to an extended metadata set, which connects datasets to more detailed contextual information, including methodology, and relevant experimental conditions (further described in the TRIQ section).

In support of user engagement and education, the repository team offers a range of tutorials designed to guide researchers through the data submission and metadata creation processes. These resources are aimed at ensuring that users understand the importance of proper metadata documentation and how to effectively utilise the repository's features. A comprehensive collection of tutorials is also available in the demo data repository. This platform allows researchers to explore all features in a simulated training environment before transitioning to the live repository, ensuring a smooth and confident user experience. By combining efficient storage, comprehensive metadata, and educational initiatives, the centralised repository enhances the accessibility and usability of research data.

6.3. PERSONAL REPOSITORIES IN CATALYSIS RESEARCH: ADDRESSING PRIVACY AND FLEXIBILITY

Local repositories are used to store sensitive or preliminary data, offering researchers greater privacy and control over their data. The Repo4Cat enables the creation of personalised Dataverse repositories, where users can manage their data independently or collaborate with colleagues.

- **Privacy Features:** All unpublished materials remain private and inaccessible to others unless explicitly shared by the user. Only published datasets are visible to the broader community.
- **User Control:** Researchers have complete control over who can view their data by managing permissions for individual repositories and datasets. This flexibility supports both solitary work and collaborative efforts.
- **Data Management:** New dataverses and datasets can be created within a user's personal workspace, ensuring adaptability to diverse research needs.

6.4. PID HANDLE SERVER FOR RESOURCE IDENTIFICATION

The developed PID Handle server offers a vital function by providing stable, long-term identifiers for datasets, devices, and external resources ensuring that data can be easily referenced and tracked over time. One of the key advantages of an in-house PID Handle server is that it allows for customisable identification protocols that align closely with the unique workflows and data management needs of German researchers in this field. Unlike standard DOI systems, which are typically geared towards published academic outputs, this PID system can be adapted to cover a wider array of research assets beyond traditional publications. For instance, PIDs can be assigned not only to datasets but also to research devices, software tools, or even specific experimental configurations. This adaptability makes the PID Handle server well-suited to the dynamic and diverse data landscape in catalysis, where unique identifiers for specific experimental setups or data processing workflows are often necessary for accurate data tracking and reproducibility. To fully leverage the capabilities of the PID Handle server, particular attention must be given to the underlying PID model that governs the assignment and management of persistent identifiers. This model has been carefully designed to meet the specialised needs of catalysis, chemistry, and materials science research.

6.5. WEB APPLICATION FOR CREATING EXTENDED METADATA TRIQ

The web application for generating detailed metadata, named TRIQ, represents a significant advancement toward semantic data enrichment in the NFDI4Cat infrastructure. By leveraging a structured ontology, TRIQ enhances the discoverability and contextual relevance of datasets, facilitating more effective data sharing and reuse across the research community. One of the important features is the connection of TRIQ to a centralised triple store powered by Apache Jena Fuseki. This triple store provides a robust framework for managing the semantic metadata generated through the application, facilitating complex queries and efficient data retrieval. By connecting to a triple store, TRIQ can effectively utilise linked data principles, enabling relationships between datasets. The choice of Apache Jena Fuseki, following careful selection, ensures that the infrastructure is built on a reliable and scalable platform capable of handling the intricate data relationships inherent in catalysis and materials science research.

In summary, the TRIQ web application, complemented by a centralised triple store, enhances the NFDI4Cat RDM infrastructure's capability to manage and enrich research metadata effectively. By focusing on user needs, maintaining an adaptive ontology, and leveraging advanced data management technologies, TRIQ aims to create a seamless experience for researchers, ultimately fostering a more collaborative and innovative research environment.

6.6. METAPORTAL FOR METADATA-BASED DATASET SEARCH

The NFDI4Cat Metaportal is a critical component for data discoverability, as it provides an interface for searching datasets using metadata. However, the effectiveness of the metaportal relies on the quality and consistency of the metadata. To improve search precision and user experience, the metaportal could incorporate machine learning algorithms that optimise search results based on usage patterns and user feedback. Integrating cross-referencing features that allow users to find related datasets or studies could further enhance the utility of the metaportal.

Additionally, enabling metadata search functions that span both the central and local repositories would increase the transparency of research activities and improve collaborative opportunities. A potential improvement would be to incorporate tools for visualising search results, helping users explore datasets more intuitively and assess their relevance to specific research questions.

6.7. VOCABULARY AND ONTOLOGY SERVICES

Vocabulary (Voc4Cat) and ontology services are essential for standardising terminology and ensuring consistency in metadata annotations. However, maintaining these services requires continuous coordination and alignment with international standards such as the STRENDA guidelines. Establishing a feedback mechanism for researchers to propose new terms or request modifications to the ontology could help keep it current and relevant. Additionally, collaboration with international ontology projects in catalysis and materials science could help ensure compatibility with external datasets and facilitate cross-border data integration. Such a collaboration was explored, for example, in Behr et al. where an automated workflow was developed to transfer biocatalysis data to process simulation software[16]. The transfer and structuring of the data with a tailored ontology allowed for efficient data transfer with minimal user input.

15 *Bönisch, T., Dikova, Y., Doerr, M., Huskova, N., Kushnarenko, V., Linke, D., Petrenko, T., Rodrigues, P., Schimmler, S., Wentzel, B., & Zhang, Y. (2023). NFDI4Cat: Architecture document (Version 1) [Working paper]. Zenodo. https://doi.org/10.5281/zenodo.10391091*

16 Behr, A.S., Surkamp, J., Abbaspour, E., Häußler, M., Lütz, S., Pleiss, J., Kockmann, N., & Rosenthal, K. (2024). Fluent Integration of Laboratory Data into Biocatalytic Process Simulation Using EnzymeML, DWSIM, and Ontologies. Processes 2024, 12, 597. https://doi.org/10.3390/pr12030597

7. Conclusion

The adoption of robust RDM practices is essential for advancing chemistry and catalysis research considering its growing data reliance and the increasing demands for open science. To achieve a seamless, integrated RDM ecosystem that accelerates innovation and reproducibility in chemistry, this white paper outlines a clear roadmap for action:

- **Aligning RDM efforts and infrastructure:** The Netherlands and Germany will co-develop interoperable RDM practices, protocols, and standards to enable efficient data sharing and collaborative workflows across borders, facilitated by local infrastructure, to create a cross-border knowledge exchange and research environment in chemistry.
- **Collecting and aligning ontologies:** The Netherlands and Germany will collect and align ontologies and design suitable metadata standards.
- **Offering training activities to foster cultural change:** The Netherlands and Germany will co-design and provide training activities tailored to chemistry researchers, focusing on both RDM practices and digital tools/solutions available for chemists. This will help drive a cultural shift towards collaboration and openness in the field.
- **Promoting continuous knowledge exchange across borders:** The Netherlands and Germany will maintain regular communication and organise regular workshops, hackathons, and community events to facilitate knowledge exchange and the dialogue between researchers and infrastructure providers, to collectively address arising issues, and navigate and integrate emerging technologies.

The collaborative efforts of German and Dutch researchers, supported by this framework, will lay the foundation for a global culture of FAIR data practices. By addressing technical, cultural, and financial barriers, we can unlock the potential of shared data to revolutionise research in chemistry and catalysis. Let us act together to make RDM a cornerstone of our scientific future.

Appendix A: Contributors to the White Paper

The following is an alphabetical list of contributors to this paper, along with their respective roles as defined by the Contributor Role Taxonomy (CRediT).

Name	ORCID	Affiliation	E-Mail	Contribution
Nongnuch Artrith	https://orcid.org/0000-0003-1153-6583	Utrecht University	n.artrith@uu.nl	Writing – original draft
Melusine Billig		DECHEMA e.V.	melusine.billig@dechema.de	Writing – original draft
Irene Bonati	https://orcid.org/0000-0002-5728-8979	SURF	irene.bonati@surf.nl	Conceptualisation, Writing – original draft Writing – review & editing
Mark Doerr	https://orcid.org/0000-0003-3270-6895	Universität Greifswald	mark.doerr@uni-greifswald.de	Writing – original draft
Sara Espinoza	https://orcid.org/0000-0002-5902-650X	DECHEMA e.V.	sara.espinoza@dechema.de	Conceptualisation, Writing – original draft Writing – review & editing
Nadiia Huskova	https://orcid.org/0000-0002-5901-4189	High Performance Computing Center Stuttgart	nadiia.huskova@hlrs.de	Conceptualisation, Writing – original draft
Rachit Khare	https://orcid.org/0000-0002-1519-5184	TU München	rachit.khare@tum.de	Writing – original draft
Norbert Kockmann	https://orcid.org/0000-0002-8852-3812	TU Dortmund University	norbert.kockmann@tu-dortmund.de	Writing – review & editing
Michael Liebau	https://orcid.org/0000-0002-1081-3051	Universität Leipzig	michael.liebau@uni-leipzig.de	Conceptualisation, Writing – original draft
David Linke	https://orcid.org/0000-0002-5898-1820	Leibniz Institute for Catalysis	david.linke@catalysis.de	Conceptualisation, Writing – original draft
Jelte P. Nimoth	https://orcid.org/0000-0003-3286-7082	Rijksuniversiteit Groningen	j.p.nimoth@rug.nl	Writing – review & editing
Evgeny A. Pidko	https://orcid.org/0000-0001-9242-9901	TU Delft	E.A.Pidko@tudelft.nl	Conceptualisation, Writing – original draft
Kevin Rossi	https://orcid.org/0000-0001-8428-5127	TU Delft	k.r.rossi@tudelft.nl	Writing – original draft
Sonja Schimmler	https://orcid.org/0000-0002-8786-7250	Fraunhofer FOKUS	sonja.schimmler@fokus.fraunhofer.de	Writing – review & editing
Mira Stanic	https://orcid.org/0009-0000-3382-8443	TDCC-NES	nes@tdcc.nl	Conceptualisation, Writing – original draft
Annette Trunschke	https://orcid.org/0000-0003-2869-0181	Fritz Haber Institute Berlin	trunschk@fhi-berlin.mpg.de	Writing – original draft
Joanne Yeomans	https://orcid.org/0000-0002-0738-7661	TDCC-NES	j.yeomans@tudelft.nl	Conceptualisation, Funding Acquisition, Writing – review & editing

Appendix B: Workshop Details and Participants

The FAIR4ChemNL workshop on 4-5 June 2024 was organised as part of the bottleneck project “Community building and project development for the TDCC-NES during 2024 – 2025” with file number ICT.TDCC.002.004 of the research programme “Implementation Plan Investments Digital Research Infrastructure” which is (partly) financed by the Dutch Research Council (NWO). We acknowledge the Dutch Research Council (NWO) in The Netherlands for awarding this bottleneck project. The presentation summarising the event‘s programme and discussions is available for download on Zenodo.

Participants (alphabetical order) in the workshop “FAIR4Chem”, 4-5 June 2024, Utrecht, Netherlands, some of whom provided corrections or other input to this paper:

TU Dortmund University	ORCID	Affiliation
Nongnuch Artrith	https://orcid.org/0000-0003-1153-6583	Utrecht University
Atul Bansode	https://orcid.org/0000-0002-1972-6704	TU Delft
Alexander Behr	https://orcid.org/0000-0003-4620-8248	TU Dortmund University
Erik Bergman		Avantium
Irene Bonati	https://orcid.org/0000-0002-5728-8979	SURF
Thomas Bönisch	https://orcid.org/0000-0003-3108-8597	High Performance Computing Center Stuttgart
Sylvestre Bonnet	https://orcid.org/0000-0002-5810-3657	Leiden University
Hendrik Borgelt	https://orcid.org/0000-0001-5886-7860	TU Dortmund University
Simone Ciarella	https://orcid.org/0000-0002-9247-139X	Netherlands eScience Center
Mark Doerr	https://orcid.org/0000-0003-3270-6895	Universität Greifswald
Sagar Dolas	https://orcid.org/0000-0001-5026-8724	SURF
Bernd Ensing	https://orcid.org/0000-0002-4913-3571	University of Amsterdam
Sara Espinoza	https://orcid.org/0000-0002-5902-650X	DECHEMA e.V.
Alessa Gambardella	https://orcid.org/0000-0002-4930-2662	Leiden University
Stan Gielen		Big Chemistry
Nadia Huskova	https://orcid.org/0000-0002-5901-4189	High Performance Computing Center Stuttgart
Cristina Izquierdo Lozano	https://orcid.org/0000-0003-0917-8860	TU Eindhoven
Roel Janssen	https://orcid.org/0000-0003-4324-5350	4TU.ResearchData
Maithili Kalamkar-Stam		SURF
Rachit Khare	https://orcid.org/0000-0002-1519-5184	TU München
Norbert Kockmann	https://orcid.org/0000-0002-2367-1988	TU Dortmund University
Michael Liebau	https://orcid.org/0000-0002-1081-3051	Universität Leipzig
David Linke	https://orcid.org/0000-0002-5898-1820	Leibniz Institute for Catalysis
Mathijs Mabesoone	https://orcid.org/0000-0003-1314-7215	Radboud University Nijmegen
Florian Meirer	https://orcid.org/0000-0001-5581-5790	Utrecht University
Naomi Messing		NWO
Jelte Nimoth	https://orcid.org/0000-0003-3286-7082	University of Groningen
Evgeny Pidko	https://orcid.org/0000-0001-9242-9901	TU Delft
Robert Pollice	https://orcid.org/0000-0001-8836-6266	University of Groningen
Gleb Popov		Fontys University of Applied Science
Nicolas Renaud	https://orcid.org/0000-0001-9589-2694	Netherlands eScience Center
William Robinson	https://orcid.org/0000-0002-7627-0192	Big Chemistry
Preston Rodrigues		High Performance Computing Center Stuttgart
Kevin Rossi	https://orcid.org/0000-0001-8428-5127	TU Delft
TU Delft	https://orcid.org/0000-0001-8270-6979	University of Amsterdam
Sonja Schimmler	https://orcid.org/0000-0002-8786-7250	Fraunhofer FOKUS
Julia Schumann	https://orcid.org/0000-0002-4041-0165	Fritz Haber Institute of the Max Planck Society Berlin/Humbolt University of Berlin
Mark Somers	https://orcid.org/0009-0004-4101-7789	Leiden University
Mira Stanic	https://orcid.org/0009-0000-3382-8443	Thematic Digital Competence Centre, Natural & Engineering Sciences (TDCC-NES)
Annette Trunschke	https://orcid.org/0000-0003-2869-0181	Fritz Haber Institute of the Max Planck Society Berlin
Bianca Wentzel	https://orcid.org/0000-0002-9218-5676	Fraunhofer FOKUS
Aleksandra Wilczynska		4TU.ResearchData
Egon Willighagen	https://orcid.org/0000-0001-7542-0286	Maastricht University
Joanne Yeomans	https://orcid.org/0000-0002-0738-7661	Thematic Digital Competence Centre, Natural & Engineering Sciences (TDCC-NES)

Notes

nfdi4cat.org
www.tdcc.nl/nes

linkedin.com/company/nfdi4cat

twitter.com/nfdi4cat

youtube.com/@nfdi4cat

github.com/nfdi4cat

CONTACT

DECHEMA e.V.
Theodor-Heuss-Allee 25
60486 Frankfurt am Main
Germany
www.dechema.de

Dr. Sara Espinoza
+49 (0)69-7564-354
sara.espinoza@dechema.de

DOI: 10.5281/zenodo.15050550

For more information about related projects in the Netherlands, TDCC staff can be reached via nes@tdcc.nl