David Linke

NFDI4Cat: Data infrastructure for catalysis research – from molecules to chemical processes

Why catalysis data is different

Managing research data in catalysis is exceptionally challenging. The field spans from molecular spectroscopy and surface-science simulations to time-series from pilot-plant reactors, routinely involves confidential industry partnerships, and – until recently – lacked any standardized vocabulary or metadata standard to describe its core concepts [1]. Unlike domains that can converge on a single data model, catalysis research generates data that vary fundamentally in structure, scale, and context.

Two additional challenges set catalysis apart. First, it covers the complete value chain from fundamental discovery through laboratory and pilot scale to industrial production. Second, catalysis research has a long tradition of close academia-industry cooperation, which brings legitimate requirements for intellectual property protection and controlled data sharing.

The NFDI4Cat vision

NFDI4Cat (DFG project 441926934) is the German National Research Data Infrastructure consortium for catalysis-related sciences, coordinated by DECHEMA e.V. [2]. Its mission is to make catalysis data FAIR – Findable, Accessible, Interoperable, and Reusable – across the entire development from early lab studies to production processes. The consortium unites 14 co-applicant institutions with complementary roles: technology providers who develop tools, domain partners who adopt them and provide feedback, and an Industrial Advisory Board that contributes guidance from an industry perspective. This advisory board, established from the outset, reflects the reality that industry collaboration is common in catalysis research, not the exception.

The guiding principle of NFDI4Cat is to support the catalysis value chain – "from molecules to chemical processes." The data infrastructure is designed to support researchers at every stage, whether they work on fundamental molecular understanding, on lab-scale catalyst screening, or on process optimization at pilot or production scale (Figure 1).

Dr. David Linke
Leibniz Institute for Catalysis (LIKAT)
Albert-Einstein-Str. 29a, 18059 Rostock, Germany
david.linke@catalysis.de

DOI-Nr.: 10.26125/d5vg-eq61

In the current second funding phase, the focus is shifting from foundational tool development to community-oriented deployment through thematic data spaces – collaborative environments organized around concrete research topics such as sustainable alcohol and olefin production, with pre-configured tools and vocabularies.

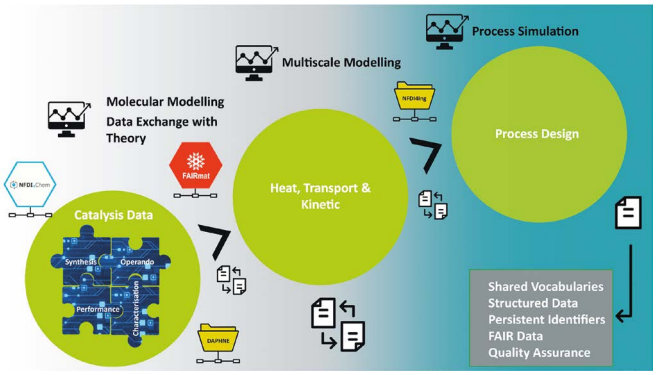

Fig. 1: The catalysis value chain spans from molecular modelling and fundamental catalysis data through reaction engineering which adds heat, transport and kinetic modelling to process design and simulation. NFDI4Cat provides data infrastructure across this entire range, with interfaces to FAIRmat, DAPHNE4-NFDI, NFDI4Chem, and NFDI4Ing at the boundaries of their respective domains.

Tools for researchers

Within NFDI4Cat, several specialized local tools have been developed independently and are connected through common standards: LARAsuite, an open-source research data management suite and electronic lab notebook designed to support high-throughput robotic platforms for biocatalysis research [3]; ADACTA, which creates digital twins of catalyst test stands to archive experimental data together with the full context of reactor configurations [4]; CaRMeN for microkinetic modeling; and the BasCat toolbox at TU Berlin, which combines a local Dataverse instance with analysis tools for high-throughput catalysis data. These tools address very different research needs. Yet they can exchange data and metadata because they build on the same standards.

The shared infrastructure that connects these local tools includes the following components:

Voc4Cat is the first standardized vocabulary for catalysis research [1, 5]. Built on SKOS (Simple Knowledge Organization

System, a W3C standard for controlled vocabularies), it provides machine-readable concepts with persistent URIs. Researchers contribute new terms via Excel templates; contributions pass through automated quality checks and peer review in a continuous integration pipeline, and new versions are published automatically. The toolchain can be adopted by any community to build their own SKOS vocabulary. Voc4Cat is operational and in production use.

Repo4Cat is the central data repository, hosted at HLRS Stuttgart and based on the Dataverse software [6]. Its key differentiator is a flexible access model with four working areas: project-level spaces for sensitive collaborations, organizational and personal spaces, and a publication area for open datasets. This design directly addresses the challenge of industry-academia cooperation through a "cool-off" model: data can remain private during competitive phases – patent filings, grant proposals, pre-publication embargoes – and be opened gradually on the researcher's terms. Rather than forcing a binary choice between "open" and "closed," the model lets researchers define a timeline for progressive disclosure, protecting legitimate interests while ensuring that data eventually become available to the community.

pid4cat provides lightweight persistent identifiers for samples, devices, and models – entities for which DOIs are too heavyweight (DOIs target published outputs, require membership in a registration agency, and incur per-identifier costs). Built on the same Handle system technology as DOIs, pid4cat identifiers carry rich metadata modeled in LinkML [7]. Partners can manage their own sub-namespaces, for example, to connect their ELN, allowing to upgrade local IDs to full PIDs and track identifiers from first creation through to publication. pid4cat identifiers are in production use within Repo4Cat; the identifier service for samples and devices is in beta.

Metaportal is a federated search portal for discovering catalysis data across distributed repositories. It harvests metadata from Repo4Cat, NOMAD (FAIRmat), NFDI4Chem, NFDI4Ing, and further sources, with quality metrics for assessing metadata completeness. Because the Metaportal builds on DCAT-AP, any compliant repository can be included (Figure 3) – a direct benefit of the standards-based approach.

Building on standards – a deliberate choice

The tools described above share a common design: every Voc4Cat concept, pid4cat identifier, and Repo4Cat dataset has a stable web address that machines can follow to retrieve structured information. These resources link to each other, creating a connected web of research data rather than isolated files. This Linked Data approach addresses the hardest part of the FAIR principles: genuine interoperability and reusability. This connected web of catalysis resources can form the basis for knowledge graphs supporting automated reasoning and AI-driven discovery across the domain.

To work across institutions, the infrastructure builds on broadly adopted international domain-neutral standards rather than a single-platform ecosystem. For dataset discovery and exchange, it follows the European DCAT-AP standard used by open data portals across Europe and beyond. NFDI4Cat and NFDI4Chem jointly extended it for the NFDI (DCAT-AP-plus [8]) and for chemistry-specific metadata (ChemDCAT-AP [9]). TRIQ, a web application currently in beta, provides a visual interface for creating and exploring these connections without requiring expertise in the underlying standards.

This approach stands in contrast to strategies pursued by sister consortia. FAIRmat has built an impressive ecosystem around the NOMAD platform, with its own data model and customisable at local level [10], providing a coherent single platform. NFDI4Cat, by contrast, specifies shared standards that independent tools connect through. Both approaches have merit; notably, FAIRmat already uses Voc4Cat vocabulary terms in its own platform, demonstrating that standards-based resources facilitate adoption across communities. DAPHNE4NFDI benefits from converging on the NeXus standard for its well-defined domain of photon and neutron experiments [11].

For NFDI4Cat, the payoff of this standards-based approach is tool plurality: catalysis spans far more sub-disciplines than any single platform can serve. The standards-based framework lets diverse specialized tools coexist in a common ecosystem (Figure 2) – a requirement of the domain's breadth.

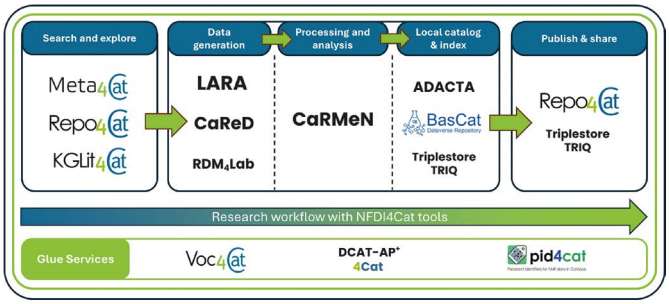

Fig. 2: Research workflow with NFDI4Cat tools, from data generation through processing to publication. Shared "glue services" – Voc4Cat, DCAT-AP+, and pid4cat – connect the specialized tools into a common ecosystem. An overview of all services is available at https://nfdi4cat.org/services.

Collaboration and community

NFDI4Cat is designed as an integrator. Cross-consortium collaboration is a defining feature and has produced concrete results.

The joint work with NFDI4Chem on DCAT-AP-plus [8] created a shared metadata foundation that benefits the entire NFDI. The chemistry-specific extension ChemDCAT-AP [9] was developed on top of this shared foundation, led by NFDI4Chem with close NFDI4Cat involvement. With FAIRmat, the Metaportal harvests NOMAD metadata for federated search. With DAPHNE4NFDI, NFDI4Cat collaborates through the "Physical Sciences in NFDI" working group and through joint research, for example in the TrackAct Collaborative Research Centre on operando X-ray absorption spectroscopy.

In the second funding phase, thematic data spaces will bring together research groups working on industry-relevant topics.

Domain partners follow a structured onboarding curriculum – internally called the "hero's journey" – that takes them from first contact with research data management to competent, independent practitioners who become multipliers in their own communities. At the same time, domain partners' experiences will feed back to tool providers, anchoring further development on real research needs. The RDM School of Catalysis provides hands-on training and is being expanded into an online learning platform.

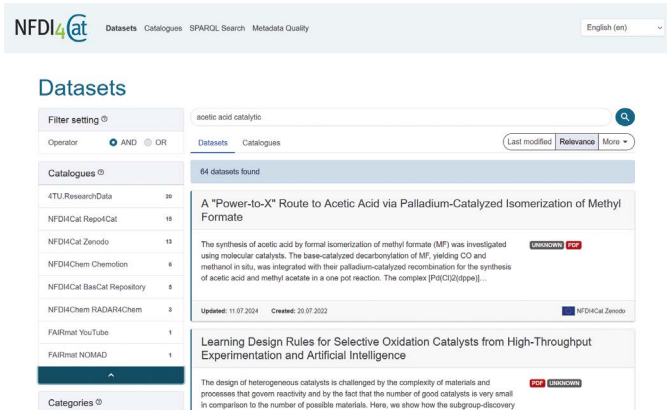

Fig. 3: The NFDI4Cat Metaportal provides federated search across catalysis data from distributed sources including Repo4Cat, NOMAD, NFDI4Chem and other NFDI repositories.

Outlook

NFDI4Cat's standards-based foundation is bearing fruit. Core tools – Voc4Cat, pid4cat, Repo4Cat, the Metaportal – are operational, and the majority of the software developed within the consortium is open source [2b]. In the second phase, the focus shifts to community growth: as researchers populate data spaces with FAIR data, they create demonstrators for AI-ready catalysis datasets and establish workflows that others can adopt.

The broader vision is catalysis data that flows across boundaries – from lab bench to production plant, from academia to industry and back – connected through international standards. Tools built on W3C standards and European frameworks are designed to outlive any single funded project, and integration with the European Open Science Cloud (EOSC) ensures relevance beyond Germany.

Researchers interested in catalysis data management are invited to explore the NFDI4Cat tools: use Repo4Cat for data storage and sharing, contribute terms to Voc4Cat, adopt pid4cat identifiers for samples and devices, and discover existing catalysis data through the Metaportal via the NFDI4Cat homepage https://nfdi4cat.org/ or directly at https://meta4cat.fokus.fraunhofer.de.

References

[1] C. Wulf et al., "A Unified Research Data Infrastructure for Catalysis Research – Challenges and Concepts," *ChemCatChem* 2021, **13**, 3223. DOI: 10.1002/cctc.202001974

[2] (a) NFDI4Cat homepage: https://nfdi4cat.org/ ; (b) NFDI4Cat open-source tools: https://github.com/nfdi4cat

[3] M. Doerr, The LARA Suite. https://lara.uni-greifswald.de/larasuite/

[4] N. Kockmann et al., NFDI4Cat Whitepaper: Ontology-based Data Management and Interoperability: Workflow for Catalysis and Process Research Data. Zenodo, 2024. https://doi.org/10.5281/zenodo.11082928

[5] (a) Voc4Cat vocabulary: https://nfdi4cat.github.io/voc4cat/ ; (b) D. Linke et al., Voc4Cat – A SKOS Vocabulary for the Catalysis Disciplines, Release 2025-10-14. Zenodo, 2025. https://doi.org/10.5281/zenodo.17351357

[6] V. Kushnarenko et al., "Repo4Cat: How to Setup, Configure and Operate a Data Repository for Catalysis-Related Sciences," *IDAACS 2025*, 2026. https://doi.org/10.1109/IDAACS68557.2025.11322400

[7] S. A. T. Moxon et al., "LinkML: An Open Data Modeling Framework," *GigaScience* 2025, giaf152. https://doi.org/10.1093/gigascience/giaf152

[8] Strömert et al., DCAT-AP-plus: A metadata profile for NFDI. https://doi.org/10.5281/zenodo.17702369

[9] Strömert et al., ChemDCAT-AP: Enabling Semantic Interoperability with a Contextual Extension of DCAT-AP, 2026. https://doi.org/10.48550/arXiv.2602.01822

[10] M. Scheffler et al., "FAIR data enabling new horizons for materials research," *Nature* 2022, **604**, 635. https://doi.org/10.1038/s41586-022-04501-x

[11] DAPHNE4NFDI: https://www.daphne4nfdi.de

Dr. David Linke

David Linke studied chemistry at the University of Cologne and received his PhD from the Technical University of Berlin. Since 2006, he heads the department of catalyst development and reaction engineering at the Leibniz Institute for Catalysis (LIKAT) in Rostock. His research spans high-throughput technologies for heterogeneous catalysis, upscaling of catalytic processes, and research software engineering. After more than twenty years of academia-industry collaboration, he helped create NFDI4Cat in 2019 and serves as one of its co-spokespersons. His group developed Voc4Cat and pid4cat and co-authored the DCAT-AP extensions.