NFDI4Cat: Local and overarching data infrastructures

Sonja Schimmler[1], Thomas Bönisch[2], Martin Thomas Horsch[2], Taras Petrenko[2], Björn Schembera[2], Volodymyr Kushnarenko[2], Bianca Wentzel[1], Fabian Kirstein[1], Harald Viemann[3], Martin Holeňa[3] and David Linke[3]

[1]Fraunhofer Institute for Open Communication Systems, FOKUS
[2]High Performance Computing Center Stuttgart, HLRS
[3]Leibniz Institute for Catalysis, LIKAT

The NFDI is a German national initiative that aims to develop repositories, tools, standards, and best practices for research data management across all scientific disciplines. Until 2022, approximately 30 consortia will be formed under the umbrella of the NFDI e.V. association. NFDI for Catalysis-Related Sciences (NFDI4Cat) is one of these consortia, which targets research data management for catalysis-related sciences, a field that is of strategic importance for the economy and the society as a whole. In this paper, we give a brief overview of the consortium and introduce its planned local and overarching data infrastructures. We further describe our approach for requirements analysis, and provide some first insights on our findings.

1 Introduction

Catalysis is one of the key technologies for tackling challenges related to climate change. This research field is investigating the acceleration of chemical transformation by using a catalyst to increase the reactions efficiency and minimize unwanted side products at the same time. Each advancement in this catalytic research is an essential foundation for addressing problems like attaining CO_2 neutrality, finding a sustainable way to feed the worlds population or improving the valorization of plastic waste. The field of catalytic research is highly interdisciplinary covering bio-, electro-, photo-, heterogeneous and homogeneous catalysis as well as disciplines like reactor design and process engineering.

Catalysis-related sciences are currently facing some problems resulting in a slowdown of research advancement. There are many different companies and institutes working on catalysis research but most of the simulations and experiments take place in isolation, resulting in the repetition of experiments and simulations. There is a lack of standardization regarding the documentation of experiments, simulations and its data and metadata. There is also a lack of exchange. These problems can be mitigated by standardization and by setting up local and overarching data infrastructures that are specifically designed for catalysis-related sciences.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13738 (CC BY-SA 4.0)

As part of the NFDI (National Research Data Infrastructure), the consortium NFDI4Cat (NFDI for Catalysis-Related Sciences) was formed to tackle these challenges in catalysis-related sciences. The core objective of NFDI4Cat is to facilitate a fundamentally improved understanding of catalysis by building a bridge between theory, simulation, and experimental studies by addressing all aspects in the catalysis value chain from the catalyst design over characterization and kinetics to engineering aspects [13].

The consortium aims for a more standardized way of handling data throughout the research data lifecycle. It will develop ontologies for catalysis-related sciences to fully describe data and processes and build local and overarching data infrastructures for the community to enable storage and exchange of semantic rich data. The local and overarching research data infrastructures will support the whole research data lifecycle and will serve as an e-science solution for the field.

One challenge is to identify and serve the real needs of the NFDI4Cat community. Therefore, we will involve different stakeholders in the whole process, including a user-centered requirements analysis and setting up a pilot system to get early feedback from the users. Another challenge is to avoid fragmentation and data silos. Therefore, we will proceed with a coordinated approach.

The remainder of this paper is structured as follows: In Section 2, we will give a brief overview of the consortium. In Section 3, we will introduce the local and overarching data infrastructures that are planned within the consortium. In Section 4, we will describe our approach for requirements analysis, and provide some first insights on our findings. In Section 5, we will give an outlook and conclude the paper.

2 Consortial structure

The NFDI4Cat consortium assembles experts from homogeneous, heterogeneous, photo-, bio-, and electrocatalysis on the one side, and from process engineering and data technology on the other side. It gathers a total of 16 dedicated partners, experts from process engineering and data technology (High Performance Computing Center Stuttgart (HLRS), Fraunhofer Institute for Open Communication Systems (FOKUS), Max Planck Institute for Dynamics of Complex Technical Systems (MPI-DCTS), Karlsruher Institute for Technology (KIT)) and from catalysis and data driven catalysis research (Leibniz Institute for Catalysis (LIKAT), Max Planck Institute for Chemical Energy Conversion (MPI-CEC), RWTH Aachen, TU Berlin, TU Braunschweig, TU Darmstadt, TU Dortmund, Friedrich-Alexander-University Erlangen-Nürnberg, University of Greifswald, University of Leipzig, TU München, University of Rostock) coordinated by the DECHEMA Gesellschaft für Chemische Technik und Biotechnologie e.V. The project is supported by an advisory board of industrial partners, including BASF SE, Clariant Produkte GmbH, Covestro Deutschland AG, Evonik Industries AG, hte GmbH, Linde AG and Thyssenkrupp Industrial Solutions AG.

To achieve the project goals, the working programme consists of eight task areas. Task area one is responsible for metadata standardization and ontology development. Task area two focuses on data standards, data collection and interfaces, and task area three on data analysis, quality management and data reusability. Task area four handles the development of linked extensible infrastructures and data access management. Task area five takes care of the dissemination, outreach and training of catalysis researchers, task area six of the networking with other NFDIs, SFBs (Collaborative Research Centres) and international initiatives. Task area seven focuses on intellectual property, licences and reward models, and task area eight on the overall management of the consortium.

3 Planned national data infrastructures

One main goal of NFDI4Cat is to set up and establish local and overarching data infrastructures, as shown in Figure 1. This includes a distributed repository infrastructure and other local and overarching services that are needed by the NFDI4Cat community, in order to put forward a national environment for catalysis-related research data. To ensure future viability, the infrastructure will be built on existing standards and principles, *e.g.*, by using established vocabularies such as schema.org or W3C DCAT, and synchronized with other consortia and other communities. Open source solutions will be favored, relying on modern technologies, and using Semantic Web technologies.

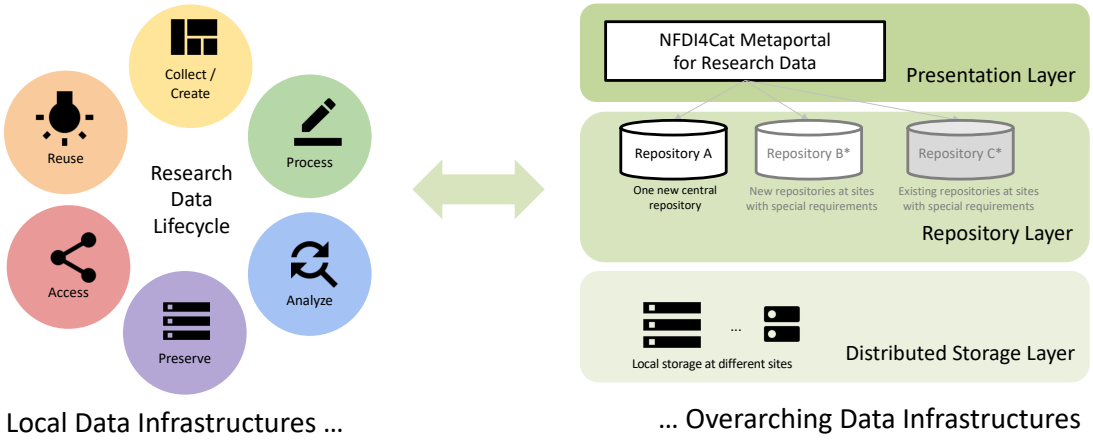

Figure 1: Local and Overarching Data Infrastructures.

3.1 Overarching data infrastructures

A distributed repository infrastructure will be set up, which will serve as an overarching data infrastructure. A layered architecture is planned, which includes a distributed storage layer, a repository layer, and a presentation layer. The distributed storage layer will enable the local storage at different sites. The repository layer will provide one new

central repository at HLRS and, where required, existing and new local repositories. The central repository as well as the new local repositories will be based on well suited existing solutions, such as Dataverse, DSpace or Invenio. The presentation layer will provide a general access point to the metadata and data that is openly available in the different repositories and will offer other overarching services that are identified of being useful for the NFDI4Cat community. This includes a graphical user interface as well as a SPARQL endpoint. The presentation layer will also provide an interface to interact with related infrastructures.

3.2 Local data infrastructures

Besides the overarching data infrastructure, local data infrastructures will be put forward, which support the whole research data lifecycle – collect/create, process, analyze, preserve, access and reuse. For this purpose, several labs working in different catalysis disciplines will setup pilots. Other groups will benefit from these pilots, either by reusing some of the local services established, or by learning from the pilots. Long-term goal of this effort is to include these services in a general toolbox.

4 Requirements analysis

Within the consortium, we follow an agile approach, meaning that the development is performed incrementally. Focusing on the overarching data infrastructure, we started with an initial requirements analysis, which will be refined later on. As a next step, a pilot system will be set up to get early feedback from the users.

4.1 User-centered approach for requirements analysis

Potential benefits from novel development work usually cannot be fully appreciated *in advance* by the majority of its future users; therefore, user-centered requirements analysis is limited and cannot be used as a substitute for good design. Within these limitations, however, it can play a role as an element of an encompassing strategy, supporting developers at anticipating community concerns and at ranking multiple design objectives. Thereby, it can assist at making and justifying conceptual design choices.

Accordingly, it is common practice in major coordinated software development efforts to conduct a user-centered requirements analysis. Previous experience has corroborated that this can be beneficial in developing research data infrastructures. In related fields, *e.g.*, Chen and Wu [3] chose a similar proceeding. Within the NFDI, *e.g.*, the sister project NFDI4Chem conducted an extensive community survey, identifying requirements for data generation, processing, annotation, sharing, publication, and reuse [1]. A similar questionnaire-based approach was followed for the NFDI4Ing requirements analysis [2].

The NFDI4Cat consortium relies on extensive in-person discussions with prospective users for its initial requirements analysis. At the time of writing, NFDI4Cat is in the process of collecting and evaluating a set of typical perspectives of prospective users of the research data infrastructures. So far, 17 individual users affiliated with member institutions have been interviewed, collecting pre-existing data management practices as well as requirements for the system architecture, for data documentation and annotation, and, hence, for metadata standardization.

In accordance with agile practices [4], we are currently in the process of extracting *personas*, *epics* and *user stories* from these perspectives. The expectation is that the requirements and design objectives obtained by analysing these perspectives are a sufficient first approximation to the needs of scientific research and development in catalysis at large, and that any further refinements can be carried out at a subsequent stage on the basis of first concrete experiences in working with a pilot system. So far, 4 personas (scientist, administrator/developer, data officer, external) were identified, and approx. 50 epics and 250 user stories were gathered.

4.2 Requirements for system architecture

The majority of the interview partners have expressed a strong desire or mandatory requirement for the bulk of their data to be hosted locally, with reliable mechanisms in place to ensure that intellectual property is protected and non-disclosure agreements with industrial and other research partners are honoured.

In some exceptional cases, especially where extensive previous work on digitalization of research data can be reused, bespoke local repositories will be developed or maintained; in other exceptional cases, where a local repository solution is required, a generic local repository (installed locally, but developed centrally) is preferable, since the benefit from reusing an established architecture will outweigh any potential benefit from developing a dedicated architecture for a bespoke local repository at the respective institution; in most other cases, however, a central repository will fulfil the requirement by enabling local storage at different sites. This also includes cases, where extremely large data sets (*e.g.*, from synchrotron beamline experiments) need to be processed.

For publishable data as well as training materials, and similar research and development assets, obversely, the interview partners expect NFDI4Cat to support a wide dissemination by enhancing their findability and accessibility, which would be optimally handled by a central component of the infrastructure, in particular, through a single point of entry.

4.3 Requirements for interoperability

Intra-platform interoperability requirements are deduced from the need for local and overarching components as well as multiple tiers of the repository architecture to communicate with each other in a well-defined way. This also concerns the exchange of data

and metadata with electronic laboratory notebooks (ELNs) used by the consortial partners and other researchers in catalysis-related sciences. The user interviews indicate that solutions for interoperating should be explored for various open source ELNs like Chemotion [7], Kadi4Mat ELN component [6] or LARAsuite [9] but also for commercial ELNs like FURTHRmind [8].

Inter-platform interoperability requirements, obversely, focus on the communication with external digital infrastructures that are expected to interact closely with NFDI4Cat in the future. A cross-disciplinary integration of services with other NFDI consortia is of interest (in particular, concerning data ingest, retrieval, and extraction), where major synergies are expected from interactions with NFDI4Chem[1] (for chemistry), NFDI4Ing[2] (for engineering sciences), and FAIRmat[3] (for materials science). A coordination with similar domain-specific consortia such as the UK Catalysis Hub, working towards inter-platform interoperability, may be advisable as well. Furthermore, it is highly desirable to attain a status of affiliation with the European Open Science Cloud (EOSC), and to develop the required cross-platform standards.

5 Conclusion and outlook

In this paper, we have given an overview of NFDI4Cat and its planned local and overarching data infrastructures. We have further described our approach for requirements analysis, and provided some first insights on our findings.

The requirements analysis is accompanied by documenting research workflows within the different labs. Research workflows discussed in the user interviews are documented and annotated, yielding preliminary semantic artefacts, *e.g.*, lists of typical steps in catalysis research workflows, measurement methods and tools, observed properties and data formats, and key performance indicators associated with catalyst performance assessment.

The requirements analysis is further accompanied by collecting competency questions from the users [10, 11]. This way, input for data documentation and annotation, and, hence, metadata standardization and semantic interoperability are retrieved. Competency questions are representative queries formulated by prospective users in informal language (*e.g.*, "what experimental data on catalyst material class X for reaction Y were published in year Z?"), which are expected to become formally expressible as SPARQL queries by ontology-based metadata standardization [11, 12].

[1] https://nfdi4chem.de
[2] https://nfdi4ing.de
[3] https://www.fair-di.eu/fairmat

Acknowledgments

This work was funded by the German Research Foundation (DFG) through the National Research Data Infrastructure for Catalysis-Related Sciences (NFDI4Cat), DFG project no. 441926934, within the National Research Data Infrastructure (NFDI) programme of the Joint Science Conference (GWK).

Bibliography

[1] S. Herres-Pawlis, J. C. Liermann, O. Koepler, "Research data in chemistry: Results of the first NFDI4Chem community survey," *Zeitschrift für allgemeine und anorganische Chemie* 646(21), 1748–1757, https://doi.org/10.1002/zaac.202000339,2020.

[2] G. W. Jagusch, N. Preuß, "NFDI4Ing: Rückmeldung aus den Forschungscommunities," Umfragedaten, NFDI4Ing consortium, https://doi.org/10.25534/tudatalib-104,2019.

[3] X. Chen, M. Wu, "Survey on the needs for chemistry research data management and sharing," *Journal of Academic Librarianship* 43(4), 346–353, https://doi.org/10.1016/j.acalib.2017.06.006,2017.

[4] M. Cohn, *User Stories Applied for Agile Software Development*, Boston: Pearson Education, ISBN 978-0-321-20568-1, 2004.

[5] S. Herres-Pawlis, O. Koepler, C. Steinbeck, "NFDI4Chem: Shaping a digital and cultural change in chemistry," *Angewandte Chemie International Edition* 58(32), 10766–10768, https://doi.org/10.1002/anie.201907260,2019.

[6] N. Brandt *et al.*, "Kadi4Mat: A research data infrastructure for materials science," *Data Science Journal* 20(1), 8, https://doi.org/10.5334/dsj-2021-008,2021.

[7] P. Tremouilhac *et al.*, "Chemotion ELN: An open source electronic lab notebook for chemists in academia," *Journal of Cheminformatics* 9, 54, https://doi.org/10.1186/s13321-017-0240-0,2017.

[8] H. Roth, D. Menne, J. Kamp, S. Emonds, H. Wollf, M. Wessling, "Schnell zu neuen Materialien: Effizientes Forschungsdatenmanagement an der Aachener Verfahrenstechnik," *Chemie Ingenieur Technik* 92(9), 1254–1255, https://doi.org/10.1002/cite.202055486,2020.

[9] M. Dörr, U. T. Bornscheuer, "Program-guided design of high-throughput enzyme screening experiments and automated data analysis/evaluation," pp. 269–282 in U. T. Bornscheuer *et al.* (eds.), *Protein Engineering: Methods and Protocols*, New York: Humana, ISBN 978-1-4939-7364-4, 2018.

[10] P. C. Barbosa Fernandes, R. S. S. Guizzardi, G. Guizzardi, "Using goal modeling to capture competency questions in ontology-based systems," *Journal of Information and Data Management* 2(3), 527–540, 2011.

[11] A. Fernández Izquierdo, R. García Castro, "Requirements behaviour analysis for ontology testing," pp. 114–130 in C. Faron Zucker, C. Ghidini, A. Napoli, Y. Toussaint (eds.), *Proceedings of EKAW 2018*, Cham: Springer, LNCS 11313, ISBN 978-3-030-03666-9, 2018.

[12] C. Bezerra, F. Santana, F. Freitas, "CQChecker: A tool to check ontologies in OWL-DL using competency questions written in controlled natural language," *Learning and Nonlinear Models* 12(2), 115–129, https://doi.org/10.21528/LNLM-vol12-no2-art4,2014.

[13] C. Wulf, M. Beller, T. Boenisch, O. Deutschmann, S. Hanf, N. Kockmann, R. Kraehnert, M. Oezaslan, S. Palkovits, S. Schimmler, S. A. Schunk, K. Wagemann, D. Linke, "A Unified Research Data Infrastructure for Catalysis Research - Challenges and Concepts," *ChemCatChem* 12(2), 115–129, https://doi.org/10.1002/cctc.202001974R2,2021.