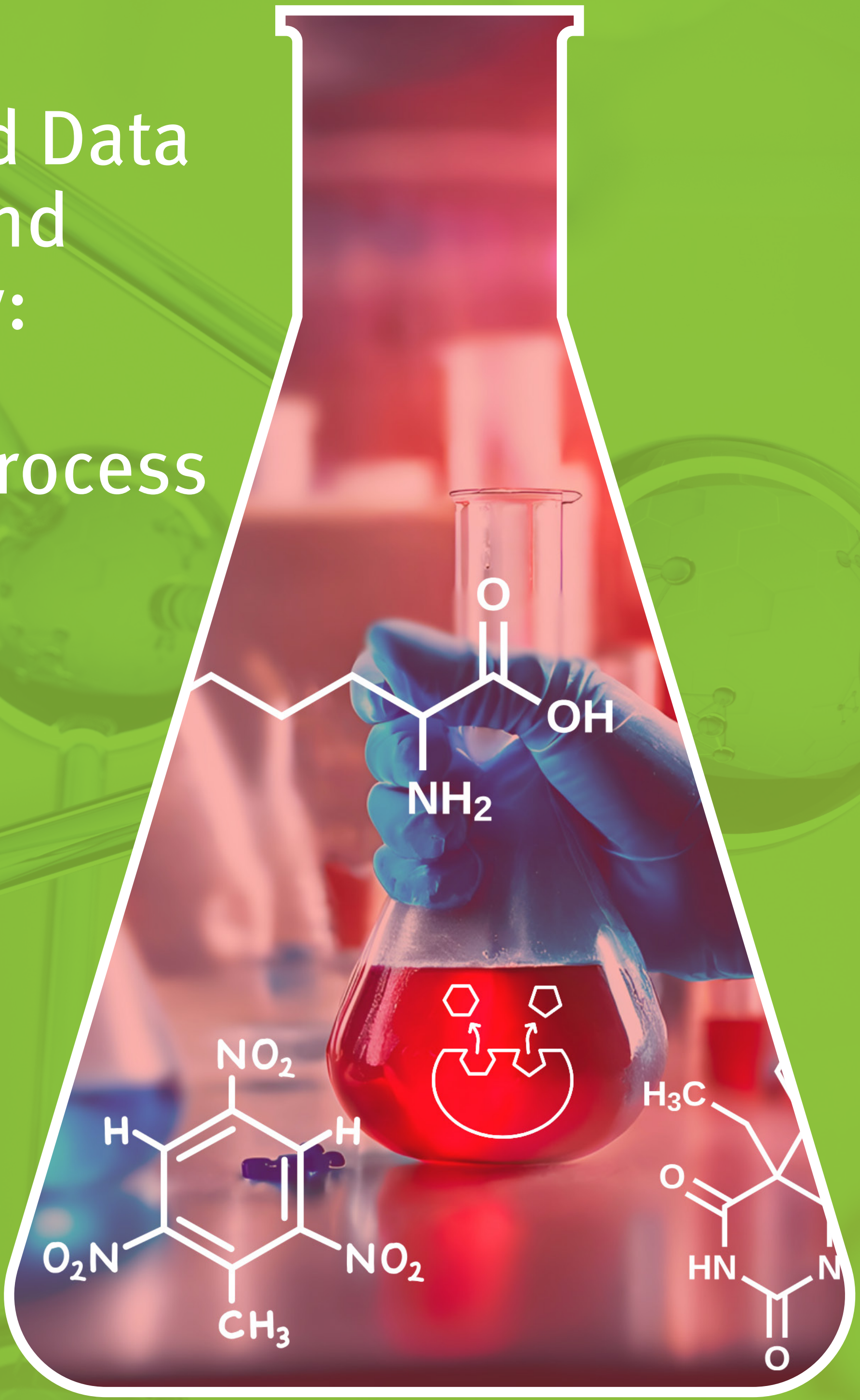
NFDI4Cat

NFDI for Catalysis-Related Sciences

White Paper

Ontology-based Data Management and Interoperability: Workflow for Catalysis and Process Research Data



Funded by

DFG Deutsche Forschungsgemeinschaft

Project number 441926934

www.nfdi4cat.org

Executive Summary

The rapidly evolving field of catalysis research generates a vast spectrum of data, necessitating innovative approaches to data management, interoperability, and utilization. This White Paper, “Ontology Mapping and Interoperability: Insights from Catalysis Research Data,” emerges from comprehensive discussions and expert contributions during the dedicated workshop in November 2023 hosted by NFDI4Cat. It encapsulates collective insights aimed at harnessing the power and applicability of ontologies and related tools to navigate the complex landscape of catalysis and catalysis related research data and the method transfer to other domains. Together with scientific, domain-specific vocabulary and standard metadata, knowledge graphs can be generated allowing to handle and store data in a structured and FAIR manner (Findable, Accessible, Interoperable, Reusable). This does not only enable consistent data handling in research data management but also leaves the data in a format ready for data analysis via advanced methods. Exemplary, the data management is shown for literature search, experimental and simulation data entry, analysis, and storage as well as research planning.

To enhance accessibility for individuals with varying levels of data literacy, the examples provided cater to different skill sets and interests. While the document extensively elucidates the overarching semantic framework, the application instances are tailored towards researchers with a practical orientation. For those seeking an introduction to this evolving landscape of Research Data Management (RDM), platforms like the Lara Suite, ELN FURTHRmind, or local Digital Labs offer valuable insights into data infrastructure and automation solutions. Advanced users can delve deeper into platforms such as LARAsuite or Nomad, while experts can actively integrate specialized applications like ADACTA, NOMAD, or the Cross Domain Meta Ontology into their projects.

Table of Contents

1. Introduction

In an era characterized by the widespread adoption of machine learning tools, including Large Language Models (LLM) such as ChatGPT, the imperative for trust in the methods and results of science and research has significantly intensified, in particular to the data content and handling process. Ensuring the reliability of information and navigating through a growing volume of publications and research data are current challenges, where organizations such as Germany's National Research Data Initiative (NFDI) aim to support researchers in producing qualitative and trusted publications. The focus is on making data FAIR (Findable, Accessible, Interoperable, Reusable) [1] and to the core aligned to open science and its principles by providing shared vocabularies.

To increase trust in data, self-describing and FAIR datasets can be created using semantics. Semantics, indicating the internal relations and meaning of content, facilitates human understanding through definitions, descriptions, and comments. Complementing semantics with logical rules and mathematical informatics establishes semantic webs that enable machines to read and "comprehend" the knowledge stored within the semantic web. These representations, which enable the semantic web, can be designed and constructed in different ways and with different (re-)usability. Starting from a simple list, e.g. a recipe, which is already self-describing thanks to its title, author, date, and structure, the semantic description improves through the use of thesauri, leading to complex description logic such as ontologies.

The semantic spectrum of knowledge organization systems is depicted in Figure 1, starting on the lowest complexity with a list, going to an ontology. Figure 1 also gives an overview on the commonly used syntaxes and formats for given semantic structures.

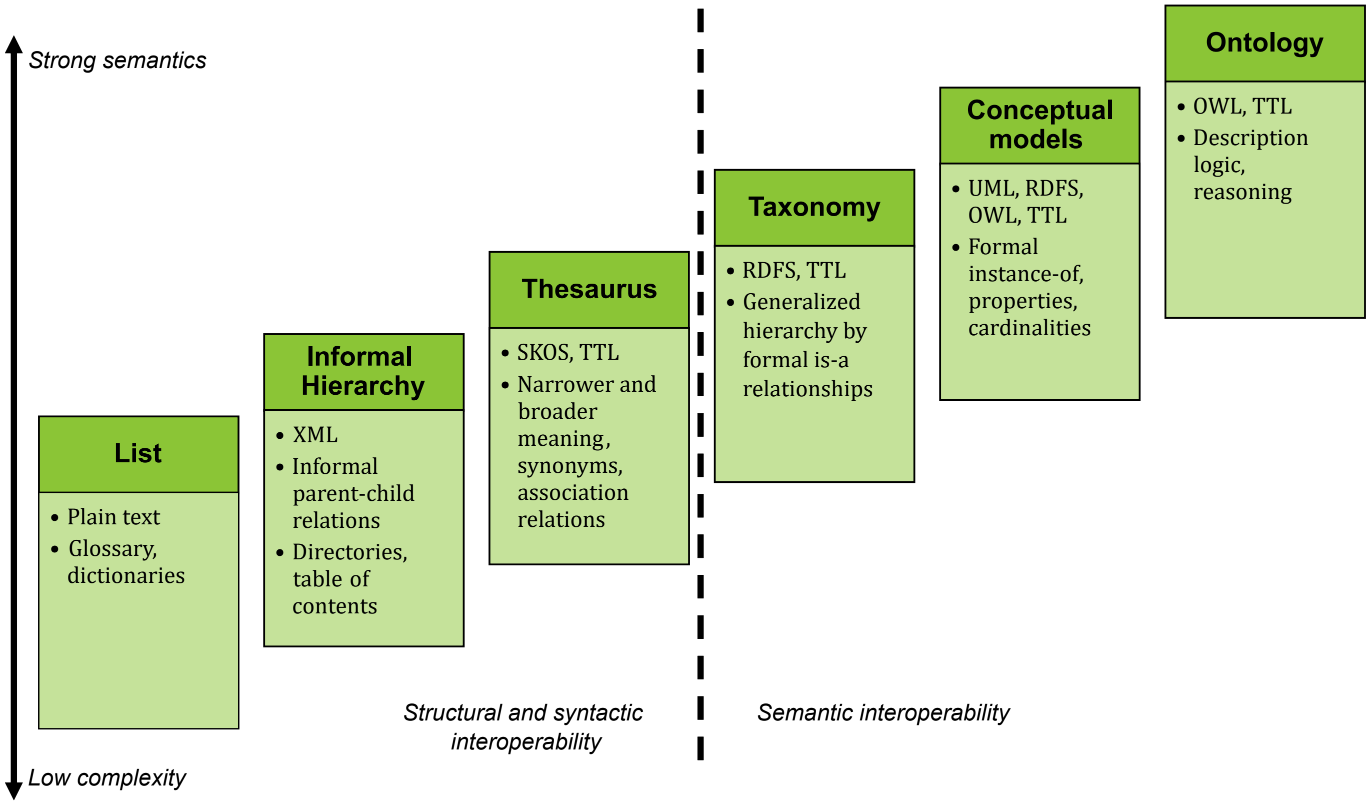

Strong semantics
Low complexity
List
• Plain text
• Glossary, dictionaries
Informal Hierarchy
• XML
• Informal parent-child relations
• Directories, table of contents
Thesaurus
• SKOS, TTL
• Narrower and broader meaning, synonyms, association relations
Taxonomy
• RDFS, TTL
• Generalized hierarchy by formal is-a relationships
Conceptual models
• UML, RDFS, OWL, TTL
• Formal instance-of, properties, cardinalities
Ontology
• OWL, TTL
• Description logic, reasoning
Structural and syntactic interoperability
Semantic interoperability

Figure 1: *Semantic spectrum from simple lists over annotated thesauri and semantically described conceptual models to ontologies with full semantic descriptions and connections [2, 3].*

Developing semantic methods for research data requires shared, community-based vocabularies and terms. To utilize these concepts, persistent identifiers (PID) for building references to data and terminologies such as ontologies, are deployed in a research data management framework and are used throughout the semantic tools to develop a semantic web. Ontologies represent advanced forms of semantic representations as they facilitate description logic capable of specifying real-world concepts using set theory and aforementioned PIDs to model domains. This allows connecting the bits and pieces of data and knowledge in a graph structure set up by information triplets (subject-predicate-object). The nodes in this graph represent classes or individuals, which can be related by predicates. An ontology adds semantics to this graph by formally defined class axioms (e.g. SubClassOf or DisjointClasses) and property axioms (like SubPropertyOf or TransitiveProperty). Individuals are related to this conceptualization by Individual axioms (like InstanceOf or SameIndividual). Based on these formally defined axioms, automated reasoning and inference over the knowledge graphs is enabled. As the complexity of class and predicate definitions increases, so does the mathematical complexity of possible inferences.

As described above, the logic of an ontology can be used to check the logic and data of a knowledge graph for correctness and thus strengthen confidence in a data set. Inference engines are the tools that are able to draw logical conclusions and thus perform this check. A brief description of how this works can be found in Figure 2.

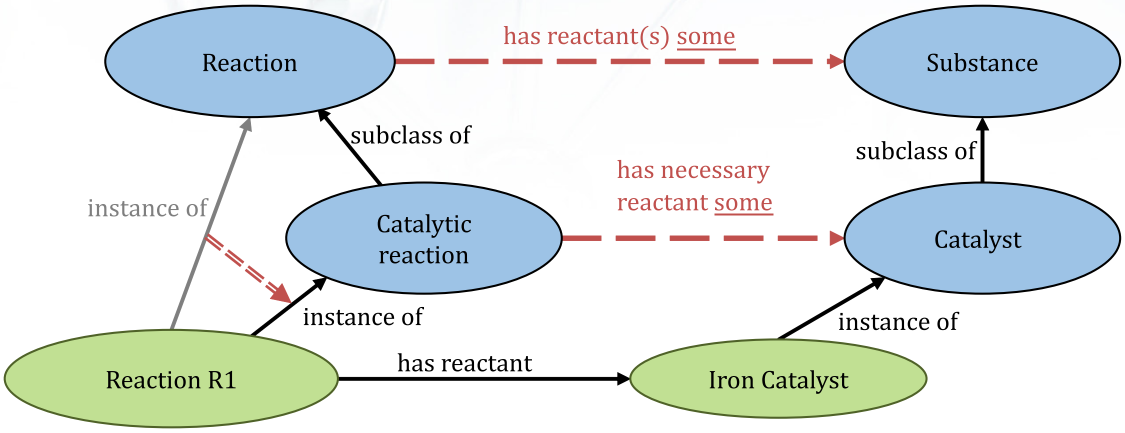


Figure 2: *Illustration of an example ontology and the function of reasoning engine, in which an Individual (green ellipse) is reclassified under a different class (blue ellipse) due to its relations (gray, black, and red directed arrows).*

In Figure 2, a simplified ontology example is depicted featuring several key classes and individuals. These include Reaction, Catalytic Reaction, Substance, and Catalyst, represented by the entities Reaction_R1 and Iron Catalyst. The connections between these entities are depicted using directional arrows in black and gray. The red dashed lines signify rules governing relationships between classes, such as the requirement that a reaction must involve reactants from the Substance class. Additionally, thanks to the rule stipulating that a catalytic reaction must involve a substance classified as a catalyst, an inference engine can deduce that Reaction_R1 is not merely a standard reaction but a catalytic one.
This inference is indicated in Figure 2 by both the gray arrow representing the relation and the red double-dashed arrow symbolizing the subclassification as a catalytic reaction.

Ontologies thereby serve as a structured set of rules, which define the structure of knowledge graphs, providing both constraints and opportunities for discovering and establishing additional relationships based on these rules.
Beyond the rules embedded within the description logic, the ontology level also encompasses semantic information aimed at enhancing the comprehensibility of data and its relationships through descriptive narratives and explanations. While in practical data description, the boundaries between data, metadata, and ontology may be blurred, functionally, these concepts merge to form a cohesive structure. When utilizing and creating new semantic structures, it is essential to adhere to these conceptual ideas to ensure that functionalities can be seamlessly built upon each other.
Achieving the expressiveness, an ontology yields, might increase methodical complexity but also improves interoperability of data. The gained expressiveness helps in understanding and trusting data, while interoperability is important for the findability and reuse of data in other applications. Good semantic representations, such as having the idea of a reaction modeled and related in an ontology, not only allow a machine to access all resources more easily but also enable cross-referenceable data. Therefore, steps such as harvesting, data cleaning, and programming as well as the corresponding data analysis can be supported in the future, for example by using ontologies for consistency checking.

Ontologies also contribute to enhancing the consistency of metadata standards, which in return enable structured handling and storage of data in suitable repositories. Ontology-based metadata enriched with data from experiments, simulations, or linked databases from the web lead to knowledge graphs, which are the base for consistent research data management. The focus lies on the integration and application of semantic aspects and tools to everyday research data workflows in the frame of NFDI4Cat (NFDI for Catalysis-Related Sciences).
To be able to design ontologies more efficiently, they are classified according to their function into three levels: on a very fundamental top level, on a generic domain and task level, and on the specific application level. On the top level ontologies aid modeling by guiding the general setup up of other ontologies and enabling the coexistence and interoperability of domains. On domain level, ontologies are used for covering the domain specific requirements and expert language, similar to the generic task level, which both are interoperable. In certain cases and for specific needs, application-level ontologies are derived, which allow for even more detailed semantic specifications tailored to a specific use case. Application ontologies are not necessarily fully compatible with other application ontologies, but by embedding them in common domain ontologies, they become at least semantically compatible

at the generic level that the domain ontology provides.

Ontologies serve as a means to establish consistent and standardized data structures known as knowledge graphs. These graphs, governed by the principles and guidelines outlined in an ontology, encapsulate details about data sources, their organization, and distinguish metadata from primary data points. They function as a comprehensive repository for data within a particular domain. By leveraging knowledge graphs, the contents of data structures can be articulated and interconnected, facilitating the linking of various information entities. For instance, one might connect a broad-scale measurement dataset with a research paper, which characterizes specific observations within that measurement.

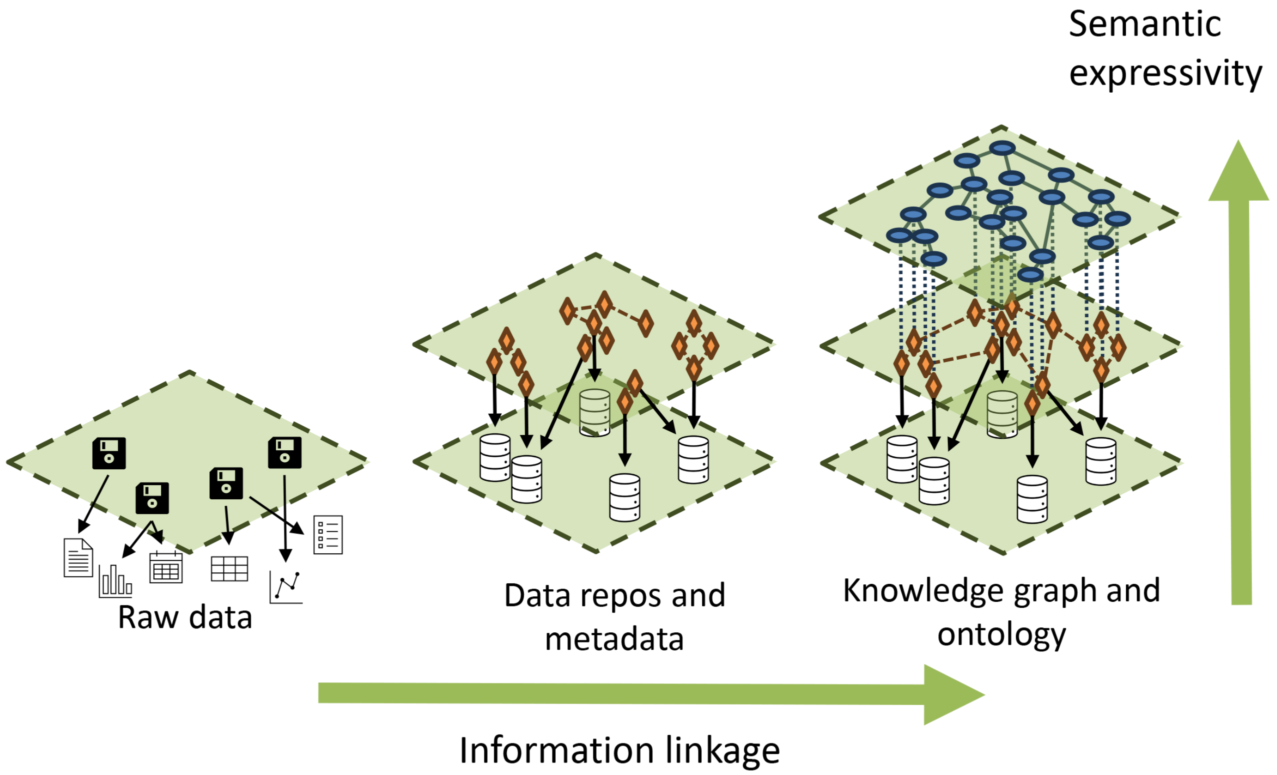


***Figure 3**: Different layers of data representation and structuring in an ontology (top layer), metadata structure (mid layer) and data containers (bottom layer).*

Figure 3 illustrates the conceptual framework for linking and describing data. Starting from left to right, a way of how the interconnectivity of data can be increased, is described. To easily present the increased semantic required to have a coherent linkage, the semantics are modeled from the bottom to the top with increasing semantic expressivity.

At the foundational layer, data generated directly from research activities can be found, including measurements by researchers, outputs from AI networks using the same data, (automatic) numerical simulations, and their resulting data sets, depicted as a data store. This data store can manifest in various forms, from local file storage to decentralized cloud storage. It also can be stored in various forms ranging from a simple file-based storage, over traditional relational databases to object-oriented databases, graph databases, and triple stores, which can represent knowledge graphs themselves. As data can be stored and induced into all of the above mentioned storage methods, a problem of uniformity often occurs. Different databases for example apply different rules for storing data, from a schema-based approach to unique SQL rules in relational databases.

To overcome these communication barriers generated by different rules, knowledge graphs and the respective metadata serve as an interconnection and translational layer. They link the data via aforementioned PIDs directly from a storage space through the modelled relations to other data points, which have their own respective PIDs. Metadata hereby serves as a method to capture not only how information is interlinked, but also the content it represents. Moving up, the next ontological level is reached, where a standardized vocabulary is established, and rules are defined regarding the linkage and permissible linkage of metadata.

In summary, these tools contribute to FAIR data management and bolster confidence in datasets, while also enabling context-based analysis of the contained data.

2. Semantic Starting Point of Data Management in Catalysis

The different consortia of the NFDI deal with their specific domains, resulting in different wishes and needs to address the research data of their respective community. While many of the domains, such as the chemistry domain (addressed, e.g. by NFDI4Chem consortium), can reuse established semantic artifacts, other domains need to start with more pioneering work. The domain of catalysis research is multi-disciplinary and very heterogeneous with regard to the type of research data. These data range from the operando measurements of, e.g., reactions on solid state or molecular catalysts, over molecular simulations of reactions on surfaces of heterogeneously catalytic reactions, to multi-step-biocatalysis, and electrochemistry. This demonstrates the large domain of catalysis research with many niche areas and resulting (meta) data to cover. Through its multi-disciplinary nature, many links to other knowledge domains, which are represented by different NFDI consortia, exist. So are, e.g., process descriptions important in NFDI4Ing or surface analysis of materials within FAIRmat.

Therefore, NFDI4Cat has been searching for established semantic artifacts since its early days and has already recorded these in an initial report [4]. As depicted in Figure 4, the domain of catalysis was divided into three topics that build on each other, spanning from catalyst data, heat and mass transfer and the associated kinetics, to process description and simulation.

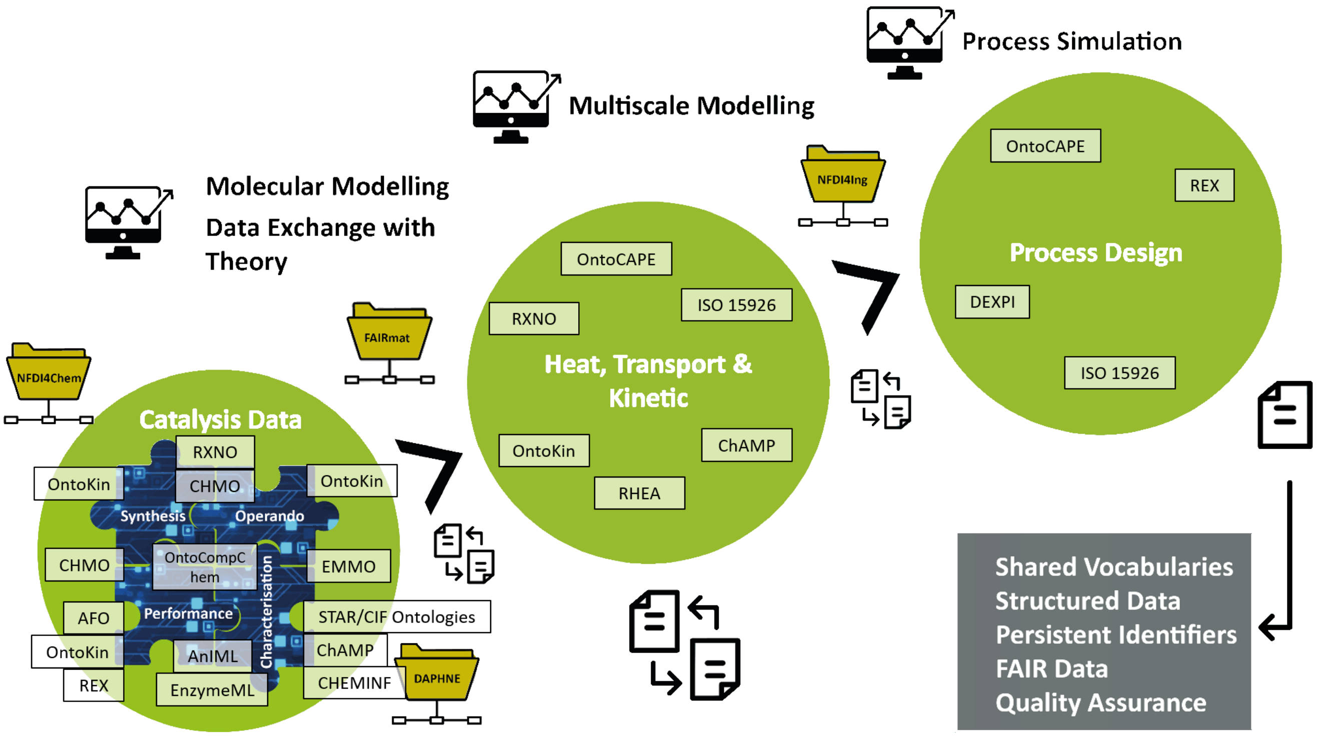


Figure 4: *Preliminary landscape of ontologies and vocabulary relevant to the data value chain of catalysis research [4]. This represented the first overview on semantic artifacts relevant to catalysis research.*

As the project work in the consortium progressed, however, some of the initially considered vocabularies and ontologies turned out to be either outdated, incomplete or designed with a strong bias, rendering them useless for most of the envisioned tasks. Thus, a second screening round was conducted for the ontologies that should cover the wide knowledge domains of catalysis research. In this second, more structured, screening, stringent criteria for the selection of ontologies have been applied which lead to a classification of ontologies in 13 subdomains of knowledge important to catalysis research.

With this approach, 30 ontologies were investigated in the second screening, resulting in not only an actively maintained ontology collection and its documentation, but also in a workflow that is agnostic to its knowledge domain. [5] Figure 5 shows an overview of the domains of knowledge and the number of ontologies that cover these domains, showing also the cross-domain nature of catalysis research.

Here, the green line indicates the number of ontologies, which contain the respective domain of knowledge of catalysis research. This shows the poor coverage of the whole domain of catalysis research by existing ontologies and with them by shared vocabularies.

Another major challenge arises in the linking and mapping of ontology classes between multiple ontologies.

This is necessary to be able to map the rather complex knowledge represented in ontologies on their heterogeneous real-world data representations, without having to compromise between one domain or another.
NFDI4Cat's aim is to build on existing ontologies and infrastructure. Therefore, it appears to be more efficient and effective to establish a comprehensive semantic architecture. This architecture should encompass the diverse array of ontologies and research data in the catalysis domain, along with their interconnections in contrast to a selective approach that focuses solely on specific applications and semantics.

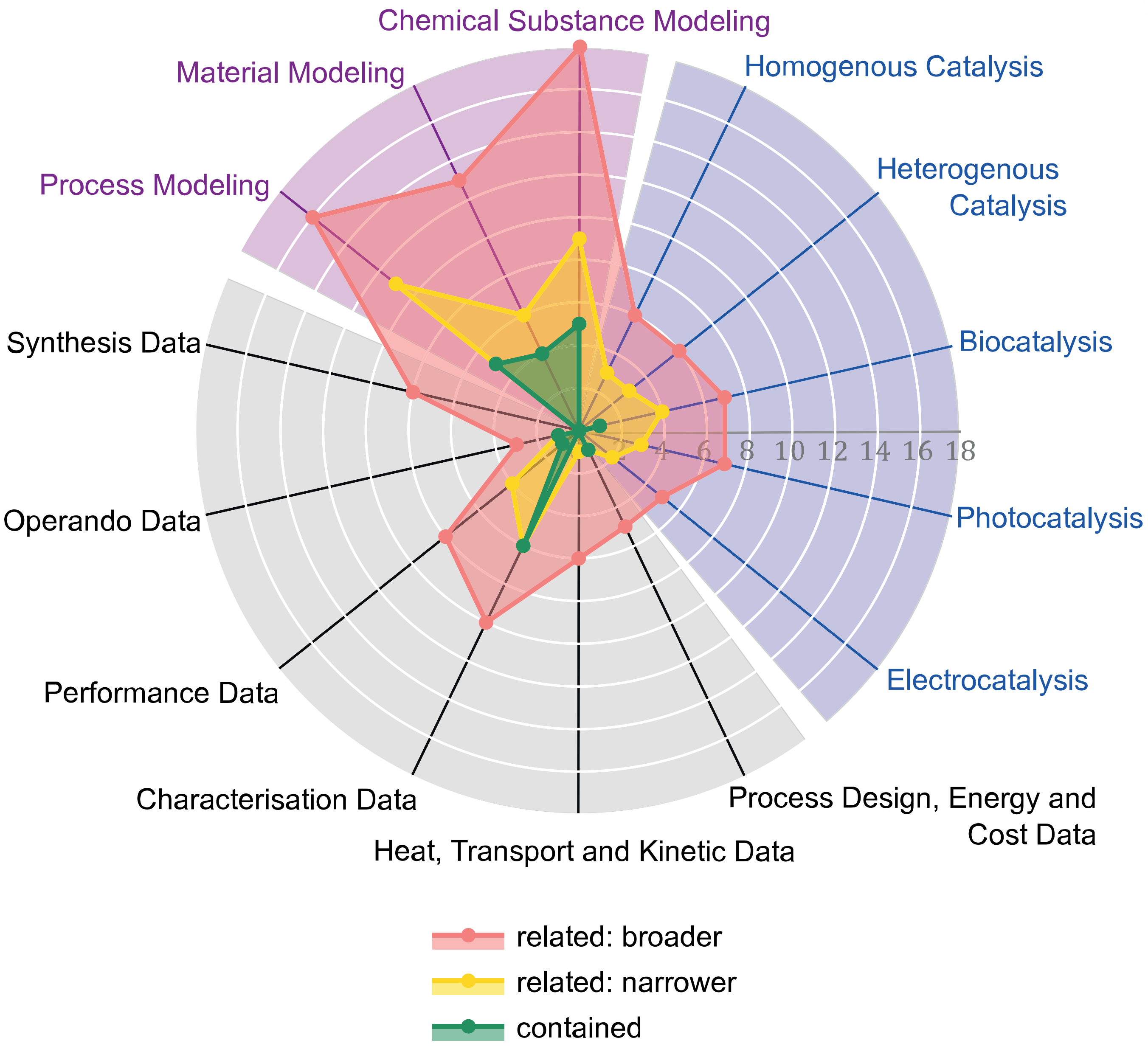


***Figure 5:** Radar plot for the number of ontologies that address the respective domains of catalysis research. The specific fields of catalysis are denoted in blue, while the fields more directed to modeling are colored purple. Fields regarding general catalytic data are written in black. Graph taken from [5].*

3. Criteria that any Functional Solution Must Meet: Academia and Industry

The architecture of any database structure should adopt a (meta) data-focused approach following the simple guideline:

"A rooted network system is wanted and open space to grow in soft malleable soil, where the developing roots have easy access to nutrient-rich data spots, and services should not be like a plant seed on rocky terrain, having trouble finding cracks to soil in the rocks."

The strategy involves molding the service architecture around knowledge graphs, allowing flexible connections to a variety of databases (differing in content or even type). Additionally, the complete and inherent integration of the FAIR principles, simplicity of use paired with an acceptance of expressivity, complexity, extendibility, and scalability are emphasized. Furthermore, the limitation to designing all workflows and demands towards serving a single database or data provider are rejected, opting for a more versatile and democratic approach. Moreover, a monolithic approach based on a single ontology is avoided based on the previous experiences made and discussed above. Instead, it appears to be a viable strategy to go forward based on a common top level ontology. Subsequently the creation of a set of domain and task specific ontologies appears important, which refer to a set of (meta) data structures and related databases.

The database structure‘s architecture is designed to satisfy key criteria, including its impact on science and technology. It is tailored to be relevant to catalytic science and process technology, encompassing areas such as reaction engineering, separations, product purification, and formulation. Emphasis is placed on usability, ensuring a user-friendly experience for researchers engaged in experimental, analytical, and numerical work. Additionally, the design prioritizes consistency and coherence across various disciplines. The sustainability of the database structure is emphasized, focusing on long-lived solutions and robust connectivity.

While developing solutions for the catalysis-related sciences community, several critical aspects must be considered to ensure their effectiveness and impact. Firstly, solutions must be relevant, meeting the demands of both individuals and the community while addressing scientific challenges. This necessitates a focus on societal needs and the practical application of tools in real-world scenarios. Additionally, the developed solutions should adhere to the FAIR principles, promoting the Findability, Accessibility, Interoperability, and Reusability of data, thus enhancing their usability and sharing capabilities. Furthermore, it is imperative to support the Open Science Principles, fostering transparency, collaboration, and accessibility in research endeavors. While simplicity of use is important, it must be balanced with an acceptance of the inherent complexity of the subject matter. Solutions should be user-friendly while retaining the expressivity required for handling intricate scientific data and processes. Moreover, they should be extendable and scalable to accommodate future advancements and growing user bases without requiring significant redevelopment efforts. Finally, solutions should not be confined to a single technology provider but should be open-source and available for public use under a creative commons license. This encourages innovation, collaboration, and accessibility, ultimately contributing to the broader advancement of catalysis-related sciences. By integrating these principles and considerations into solution development, the community can ensure the sustainability, impact, and longevity of tools and resources for present and future generation of researchers.

4. Constructing the Research Data Engine: Components, Pipelines, and User Interface

Having outlined the foundational techniques for a FAIR data approach and depicted the essential criteria to be recognized, an insight into an implementation of a structure that fulfills these criteria will be given. To start of the basic components, methodologies with which most researchers are already quite familiar with, will be presented. Subsequently, an overview of the internal pipelines, which serve as an in-depth description for those interested in building on these is given. Finally, the achieved interaction methods and the anticipated benefits resulting from this approach are presented.

4.1. Where to Start: Bits and Pieces of the Research Data Machine

After a first overview, it is presented how a vocabulary with terms, definitions, and relations helps us to link data, and how this vocabulary can be used in electronic lab notebooks to make data recording with direct linking simple and convenient. The needs of individual research groups as well as aspects of the datasets that need to be considered are then briefly discussed. Data sets in catalysis research pose some interesting challenges that are often overlooked in other research areas, not only because of the need for confidentiality but also because of the wide range of studies that can be represented in them. This leads to different requirements for the databases in which data is deposited and thus requires flexible ways to interconnect data via knowledge graphs.

4.1.1. Vocabularies and Guidelines

Catalysis is at its core a multidisciplinary field of research. Before the NFDI4Cat initiative, no catalysis-specific vocabulary existed. The development of such a vocabulary, Voc4Cat [6], aims to fill this gap and eventually lead to the creation of the respective ontologies. Vocabularies are “living organisms”, evolving alongside the field they represent, and are crucial for fostering interoperability and data sharing. They provide a standardized foundation for data exchange and collaboration, enhancing findability and reusability. Voc4Cat encompasses a three-part framework: the vocabulary itself, the guidelines on how the community can contribute, and a developed Continuous Integration (CI) pipeline. These three parts, work together to enhance the usage of the vocabulary to “real-life” research applications. The Voc4Cat vocabulary includes concepts from various subfields of catalysis: heterogeneous, homogeneous, biocatalysis, electrocatalysis, photocatalysis, reaction engineering, etc. Voc4Cat provides the specification of a Preferred Label (standardized name for a concept), explicit definitions, alternate labels (e.g., synonyms or alternative spellings), and hierarchical (parent/children) relationships between concepts. In addition, it allows the formation of collections to be used in more specific cases. URIs (via w3id.org) accompany each concept and are configured for both machine (turtle [7] file format) and human (HTML documentation) using content negotiation.

The Voc4Cat guidelines [8] follow closely the ANSI/NISO Z39.19 (R2010) standard [9]. These guidelines provide instructions for maintaining coherence and consistency across various aspects, including spelling, hyphenation, punctuation, and abbreviations, thus promoting a standardized approach to the growth of VoC4Cat.

The developed CI pipeline is discussed in more detail in chapter 4.2.4.

4.1.2. Electronic Lab Notebooks

In the Research Data machine room, ELNs log the researcher’s actions in planning, executing, and evaluating scientific experiments. Just the same as conventionally used handwritten or printed laboratory notebooks, the purpose of an ELN is to document research work to make it comprehensible and reproducible for others in a digital format. However, ELNs exceed the potential of traditional laboratory notebooks as they assure a more collaborative and easily accessible work environment as well as a sustainable longtime storage of information. By creating a metadata mask in advance all important experimental data is documented completely and the risk of transfer errors to the computer and information loss is eliminated. In addition, raw data from the experiments can be directly imported and linked with the metadata. Moreover, ELNs allow digital linking of information and data, editability, and search functions. With this the daily work of researchers is simplified and reproducibility is greatly increased. Even by students generated experimental data can be tracked across hierarchies and information can be forwarded between generations of researchers [10–12]. Ontology-based ELNs enable seamless integration of heterogeneous data sources, such as experimental protocols, materials databases, and analytical instrument outputs. By mapping experimental data to ontology terms, researchers can link related information and uncover hidden relationships between experimental variables [13].

Besides the numerous benefits of ELNs, some of their downsides should be mentioned. ELNs are more costly than hardcover notebooks and not user-friendly, making the hurdle to using ELNs much higher. Another disadvantage of ELNs is that different researchers often use different wording and metadata structure, especially among different institutions. Therefore, publications are often difficult to reproduce due to the lack of completeness and comparison of experimental results is not possible [10, 14]. Additionally, lack of terminology standardization and interoperability between different ELN systems may hinder data sharing and collaboration among research groups, especially in multidisciplinary research environments [15].

Usually, ELNs act as interfaces into Laboratory Information Management Systems (LIMS), which store the data and offer further capabilities such as planning of workflows, versioning, data analysis or data export. To further increase the capabilities of the ELN, LIMS can offer semantically structured data storage solutions and, therefore, exposing the benefits of ontologies and semantics to ELNs users. Metadata, pipelines, and data management can then be backed by semantically rich information.

To enable such integrations, ELNs must be considered as part of the framework of a RDM ontology, just as in the machine room, where a gearwheel alone has no function if it´s not connected to the whole system. Additionally, widespread usage of ELNs requires an intuitive user interface with little complexity, which still enables flexible handling of data and thus similar functionalities as the paper notebooks [12].

4.1.3. Inherently Heterogeneous Research Data

Managing the quite heterogeneous data landscape of catalysis research poses some challenges for data providers. Usually, a broad variety of tools for data storage is employed in the industry. Historically, these tools also change over time, based on change of providers or technology reasons. Long-term data storage solutions often relate to document information systems for the storage of written, language-based reports. These may lack machine readability and interoperability, as the documents are at least in part not machine-readable.

Typically, toolsets tend to be harmonized within one organization - but not within a whole industry or academia branch. Thus, an overview of the toolsets used is hardly or not available as information is seen as “internal”, hindering efforts of cross-company standardization. Furthermore, typical industry demands encompass easy usability, maximum cost-benefit, and effectiveness. Open-source and customizable, clear and well described interfaces are key technology for a wider user-acceptance of tools for data providers. Data solutions should contain all the data relevant to the respective industry and provide the data in data spaces, ideally free to access or at least with low use fees, in order to achieve the envisioned democratization of research data.

Although academic research must consider these requirements, they often have lower priority compared to other requirements. This also results in more heterogeneous toolsets, elevating the complexity of the problem.

Scientific User Groups/Data Providers

At this stage of digitalization development, scientists are not inclined to get involved with semantic modelling. Experimental science involves a multitude of complex facets, from device interoperability to manual lab work. Particularly in catalysis research, numerous manual steps such as sample preparation and configuring reactor components significantly influence experimental outcomes. In an effort to capture scientists‘ perspectives on describing their work, a thesis study within the laboratory was conducted with the aim to extract a formalized schema and, equally important, a metadata schema. Through an iteratively developed questionnaire, it was discovered that the experimental scientists conceptualized their experiments as dynamic workflows rather than fixed objects. Leveraging this insight, a skeleton for the methanol synthesis process was built as a case study, subsequently interrogating the necessary metadata that is associated with produced data. Recognizing the complexity of directly implementing an ontology encompassing all relevant entities, a more accessible approach is taken by designing a less formalized semantic model using a SKOS vocabulary, e.g. with Voc4Cat [6] (s. chapter 4.1.1) or VocPopuli [16]. The goal is to have a stable description that serves scientists in a non-intrusive way while lowering the barrier to integrate data with established semantic models from NFDI4Cat.

Chemical Process Engineering

It is the very nature of research to generate a heterogeneous collection of data to decipher the multiple aspects of a research question. Especially in catalysis, the thorough investigation of catalysts across their entire lifecycle, from manufacturing to operation and end-of-life treatment, requires various methods and techniques that provide information about the catalyst’s physicochemical properties and process-relevant performance indicators. The number of possible characterization and analysis methods commonly applied in catalysis emphasizes the heterogeneity and amount of data and data formats (.xlsx, .csv/.txt, .pdf, .log, .tif, .png, .mp4, .stl) that researchers have to analyze and evaluate - a puzzle of findings that must be put together to understand the complexity of catalysis and optimize it in a targeted manner. In the academic world, the research data is often collected in self-created folder structures and processed manually in Excel sheets or semi-automated with stand-alone applications implemented in Python or MatLab. Moreover, at universities, research work is oftentimes conducted by different students, each with its own individual “RDM strategy”. Hence, the status quo makes data handling and analysis time-consuming and unsustainable for publishing and reuse. Recent developments at universities [17] tackle the challenge of making heterogeneous data more accessible by integrating data collection and processing tools and thus automatizing research routines. Besides more FAIR data handling approaches, embracing open science principles is paramount in today‘s rapidly advancing world of research and discovery. By adhering to transparent methodologies, sharing data openly, and fostering collaborative environments, academic research efforts should contribute to the ethos of open science and amplify its value attribution. Through this commitment, researchers advance knowledge collectively and promote inclusivity, reproducibility, and innovation, driving progress for faster and more sustainable research.

4.2. Putting Everything Together: Pipelines of the Engine Room

To be able to fulfill the goals set by the criteria for an acceptable solution, not only the needs and basics must be understood. In addition, an “engine room“ needs to be build, that can be adapted to the new basics and future tasks over the years. Since NFDI4Cat aims to develop reusable methods and tools, the „engine room“ should not remain unique. Thus, it is envisioned that the tools can be adapted and modified by different stakeholders, making it even more important to draw on the principles of good design at an early stage. Here, too, the developed structures are divided into two primary functional principles that must be fulfilled. The first function is the extensibility of established semantic artifacts, coupled with the ability to incorporate new ones. The second functionality is the realization and application of these semantic artifacts to data records to obtain linked and enriched data. An overview of the structures can be seen in Figure 6, which shows the path from different data sources and interfaces, through the processing level, to the services and applications that can be realized with them. The sub-functions are described in more detail in the following individual chapters.

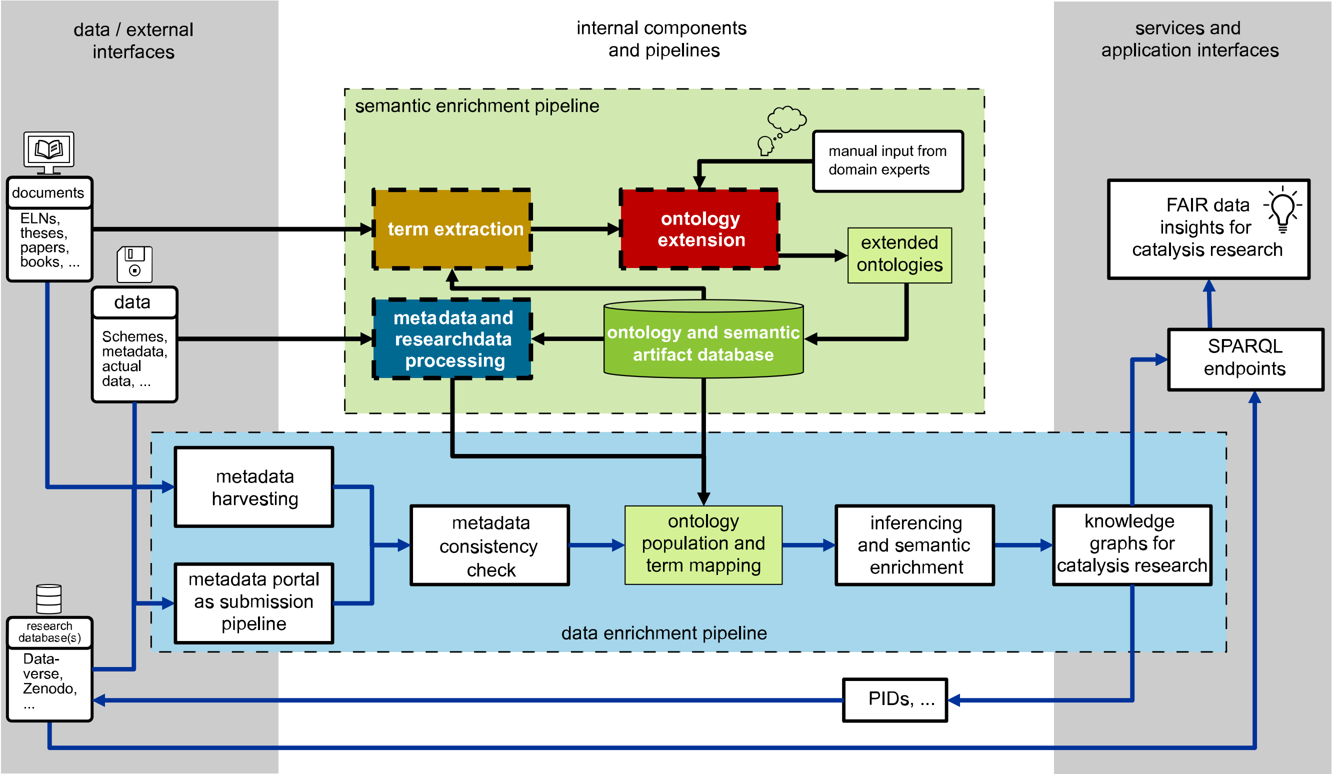

data / external interfaces
internal components and pipelines
services and application interfaces
semantic enrichment pipeline
manual input from domain experts
documents
ELNs, theses, papers, books, ...
term extraction
ontology extension
extended ontologies
data
Schemes, metadata, actual data, ...
metadata and researchdata processing
ontology and semantic artifact database
FAIR data insights for catalysis research
SPARQL endpoints
metadata harvesting
metadata consistency check
ontology population and term mapping
inferencing and semantic enrichment
knowledge graphs for catalysis research
metadata portal as submission pipeline
data enrichment pipeline
research database(s)
Dataverse, Zenodo, ...
PIDs, ...

Figure 6: Structure of the “engine room” of semantic research data management, focusing on the general interplay of internal components and pipelines.

4.2.1. Persistent Identifiers

Managing research objects like catalysts, datasets, models etc. can be challenging if they should be shared with others. This is where Persistent Identifiers (PIDs) come into play, which provide a globally unique resolvable persistent identifier for entities. Well-known examples for PIDs are the Digital Object Identifier (DOI), the Open Researcher and Contributor IDentifier (ORCID) for uniquely identifying researchers or the ROR identifier for Research ORrganizations. While DOIs were initially created as PID for publications, they can now be also used for data sets, software and since recently also for samples and devices [18]. For DOIs certain metadata are mandatory (see DataCite schema [19]). The use of DOIs also comes with certain costs per DOI. Due to the metadata obligations but also due the costs, alternative, more "lightweight" solutions such as ePIC exist [20]. For resolving a PID, ePIC uses the same underlying technology as is used for DOIs: the handle-system [21]. In contrast to DOIs, the metadata of ePIC handles are stored in the handle record itself which simplifies the system. For some application, even ePIC PIDs cause too much overhead (e.g. each PID requires a registration). Due to this, handle-based PIDs are hardly ever used for terms or classes in ontologies and vocabularies. Storing metadata about the PID itself or the distributed resolver of the handle system are of little benefit for this use case. Instead, redirect-services are used that simply offer a redirect from a stable URL to the URL describing the resource (which may change) but store no information about the individual PID. Examples for such services, which are managed by communities and typically backed-up by larger organizations, are w3id.org [22], purl.org [23] or pida.org [24].
To reflect the different use cases NFDI4Cat has been using and developing the following PID services:

- **Simple redirect service offered by w3id.org**
  - This is used for URIs of terms in ontologies or vocabularies, e.g. for the concepts and collections defined in Voc4Cat (a vocabulary for the catalysis disciplines):
    Example: "photocatalyst" has URI https://w3id.org/nfdi4cat/voc4cat_0000002
    (note, the concept has a numeric ID since it is independent of the language, the German "Photokatalysator" has the same IRI)

- **Digital Object Identifier (DOI)**
  - Zenodo-provided DOIs are used for software or linked-data artefact developed on GitHub. Examples can be found in Voc4Cat (SKOS-vocabulary): https://doi.org/10.5281/zenodo.8313341
  - Datacite-provided DOIs are used for data sets published in Repo4Cat the NFDI4Cat data portal.

- **NFDI4Cat-provided handle-based identifiers**
  - PIDs for data sharing before data publishing in NFDI4Cat´s Repo4Cat data portal or similar local data portals

- **NFDI4Cat-provided handle-based PID4Cat identifiers**
  - PID4Cat identifiers are similar to ePIC. They also store metadata in the handle record itself. In contrast to ePIC a more extensive metadata schema is suggested [https://github.com/nfdi4cat/pid4cat-model] which is available as LinkML model [25].
  - PID4Cat identifiers can be used for a variety of entities, e.g. samples, devices, or models. Partners can own and control sub-namespaces of identifiers. This allows to create PIDs locally (in their own sub-namespace) and register them later in the PID4Cat service. Existing globally non-unique IDs e.g. from ELNs can be converted to full PID by registering them as PID4Cat identifier with a namespace for the source ELN.

In addition, ORCID and ROR are used in various RDM services to identify researchers and organizations, respectively.

4.2.2. Semantic Enrichment Pipeline

At the core of the internal components and pipelines lies the semantic enrichment pipeline. This pipeline aims to accelerate the creation and extension of ontologies, leading to fast ways to transform a dataset into a knowledge graph and to ease the mapping of ontological terms to the dataset. As Figure 7 shows, documents like ELNs, scientific theses, or books give in general a good overview of the terms, which describe a specific domain. Applying Natural Language Processing (NLP)-based tools, these terms can be extracted and already clustered or set in relation to each other. With this vocabulary for a description of a domain, the terms can be searched for in existing ontologies and other semantic artefacts. This helps domain experts to choose a proper ontology for extension by the concepts not yet present in the ontology. Furthermore, this yields a database of ontologies and other semantic artefacts for catalysis research. Finally, the extended ontologies can be brought to life by the data they are describing in a last step of metadata and research data processing, generating knowledge graphs of the research data and extended ontologies for further (re)use.

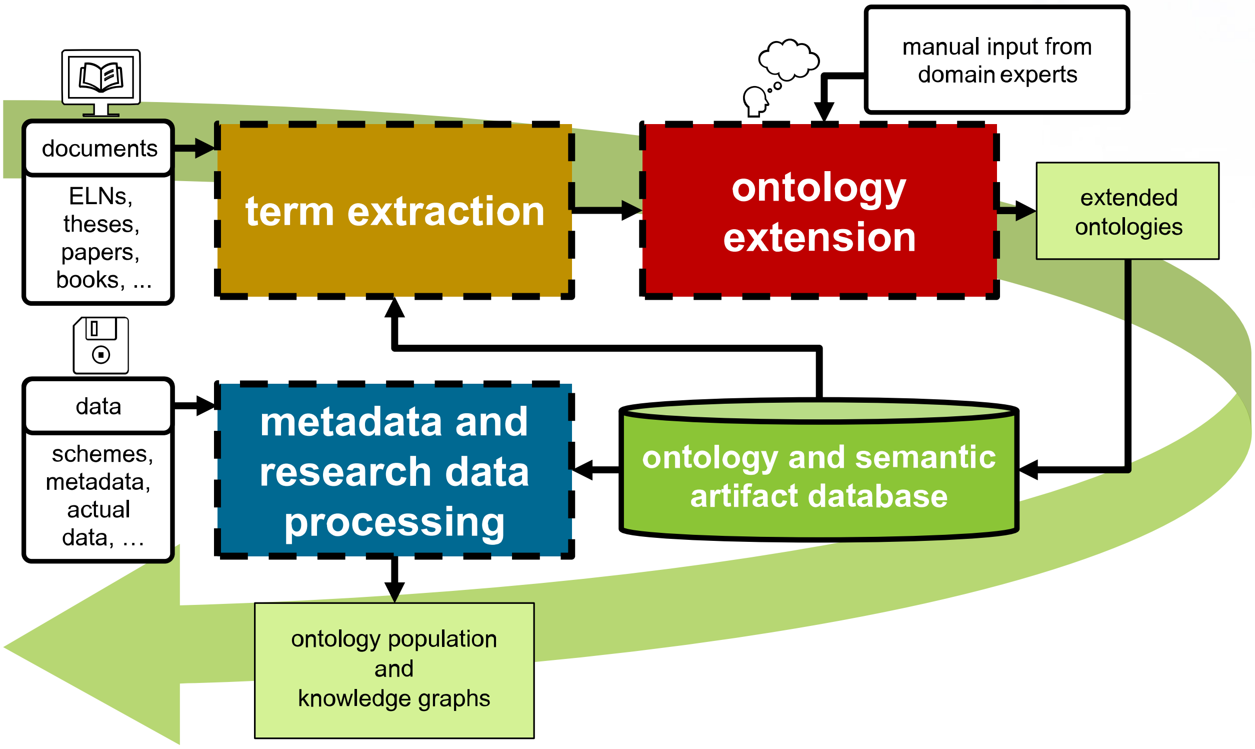

documents
ELNs, theses, papers, books, ...
term extraction
manual input from domain experts
ontology extension
extended ontologies
data
schemes, metadata, actual data, ...
metadata and research data processing
ontology and semantic artifact database
ontology population and knowledge graphs

Figure 7: *Overview on the overall semantic enrichment pipeline, modified from [4].*

4.2.3. Metadata Validation Pipeline

Moving towards linked FAIR data from simple data entails not solely relying on optimal storage practices or placing the burden entirely on researchers. Especially when dealing with intricate ontological structures, ensuring accurate (meta) data becomes paramount. For instance, an error might occur where a catalyst, along with its relations like „hasMolecularComposition“, is mistakenly categorized as a member rather than being identified as a catalyst examined by the NFDI due to a typographical mistake. Identifying such errors can be challenging within the intricacies of an ontology, and individually inspecting each data record is impractical. While tools like ELNs may incorporate validation checks, there is a necessity for complementary methods within a semantic infrastructure that operate beyond specific input interfaces.

Consider simulation software and its outcomes as an illustration. As depicted in Figure 8, the current plan outlines three primary interface options aligned with the semantic structure. Firstly, a fully automated interface ensures efficiency but maintains flexibility by accommodating potential adaptations throughout its development stages. Secondly, a semi-automated input interface caters to researchers with advanced modeling skills or those requiring a higher level of detail, aiding them in aligning their data with desired ontologies. Yet, recognizing that even this might lack adequate flexibility, provision is made for researchers to engage with semantic experts who will tailor support to their specific needs, integrating any enhancements into the validation processes of both semi-automatic and automatic metadata checks.

Focusing on the most utilized tool, the automatic interface, a more comprehensive overview reveals its capabilities. Beyond assessing the structural integrity of a knowledge graph, it is imperative to evaluate the FAIRness of data intended for storage. One approach involves utilizing databases structured using the dataverse architecture, which can be harvested with tools such as Piveau Metrics [26], for validating the quality of the given metadata. Piveau Metrics can utilize the terminology defined via DCAT and DCAT-AP, a widely adopted metadata schema in the EU, that is used in the dataverse, and evaluates it against the Data quality Vocabulary [27] to assess its quality, a process already in use within EU data platforms [28].

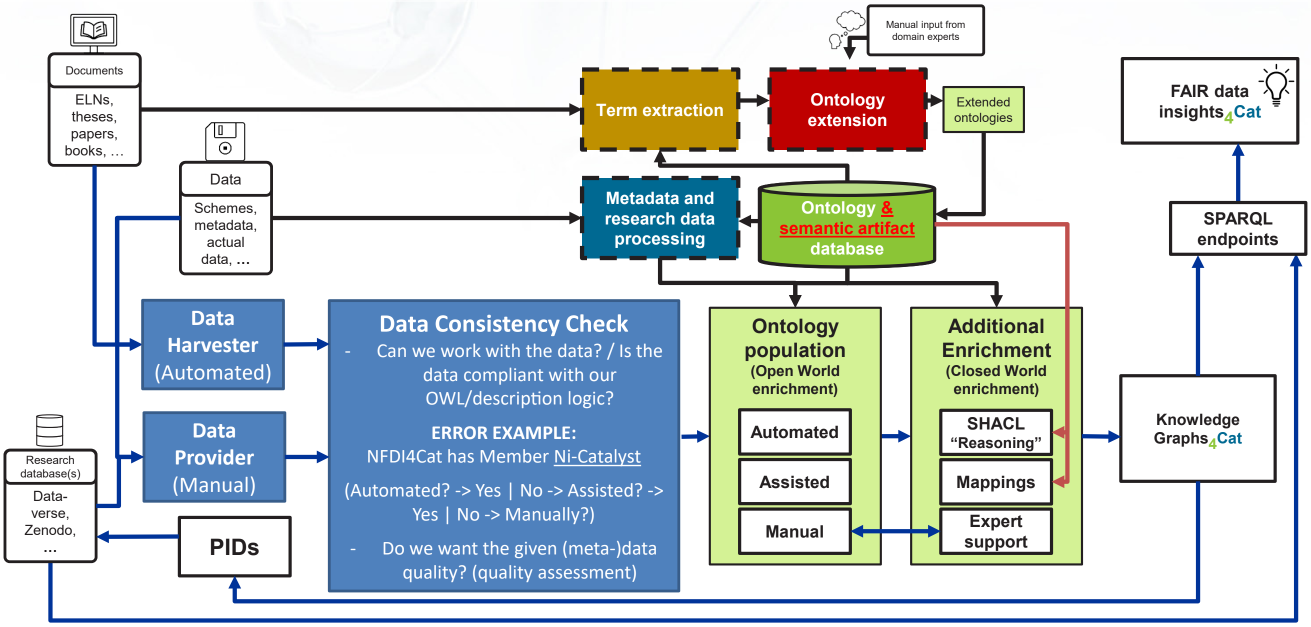

Documents
ELNs, theses, papers, books, …
Data
Schemes, metadata, actual data, …
Manual input from domain experts
Term extraction
Ontology extension
Extended ontologies
FAIR data insights4Cat
Metadata and research data processing
Ontology & semantic artifact database
SPARQL endpoints
Data Harvester (Automated)
Data Provider (Manual)
Research database(s)
Data-verse, Zenodo, …
PIDs
Data Consistency Check
- Can we work with the data? / Is the data compliant with our OWL/description logic?
ERROR EXAMPLE:
NFDI4Cat has Member Ni-Catalyst
(Automated? -> Yes | No -> Assisted? -> Yes | No -> Manually?)
- Do we want the given (meta-)data quality? (quality assessment)
Ontology population
(Open World enrichment)
Automated
Assisted
Manual
Additional Enrichment
(Closed World enrichment)
SHACL "Reasoning"
Mappings
Expert support
Knowledge Graphs4Cat

Figure 8: *Overview on the metadata validation pipeline.*

4.2.4. Semantic Development Pipelines

A user-friendly vocabulary contribution workflow (Continuous Integration -CI- pipeline) has been developed and extensively tested, facilitating community contributions to the vocabulary coupled with an effective vocabulary curation procedure and the automatic creation of the relevant resulting documentation, as shown in Figure 9.

The developed tool is hosted in the NFDI4Cat GitHub repository (https://github.com/nfdi4cat/voc4cat) and uses the SKOS standard. Leveraging GitHub and Excel, the CI pipeline facilitates community-driven contribution and curation. This approach combines the efficiency of GitHub features with the familiarity of Excel, making it accessible to a wide audience. The enabling Python package is a universal xlsx to SKOS converter [6]. A GitHub-repository template is offered to be re-use by other communities for their vocabularies [29]. Voc4Cat is built on this repository template.

Contributions can be submitted as an Excel file (through the developed template: https://github.com/nfdi4cat/voc4cat/tree/main/templates) or as a SKOS/turtle file. Eventually, the submitted xlsx file is removed and converted to a SKOS/turtle file, but the tool is able to convert on request between .xlsx and SKOS/turtle formats (in both directions). After submission, the CI pipeline starts automatically (typically runs for less than a minute) and initially checks for errors and suggests corrections. When all issues (if any) are addressed, and through an iterative improvement loop containing peer review and discussion with the maintenance team, the suggested changes reach an expert editor. The editor is charged with approving or rejecting the proposed request leading to the publication of an updated vocabulary.

To track the contributors, each interested community member requests an ID range and the respective contributions will be attributed via the researcher's ORCID number. This also allows independent work and avoids using the same ID repeatedly.

When a new version of the vocabulary is released, it is automatically published in Zenodo (latest release: https://doi.org/10.5281/zenodo.8313340). Voc4cat is available via the TIB Terminology Service [30], appears in FairSharing.org [31], and the European Open Science Cloud (EOSC) – portal [32].

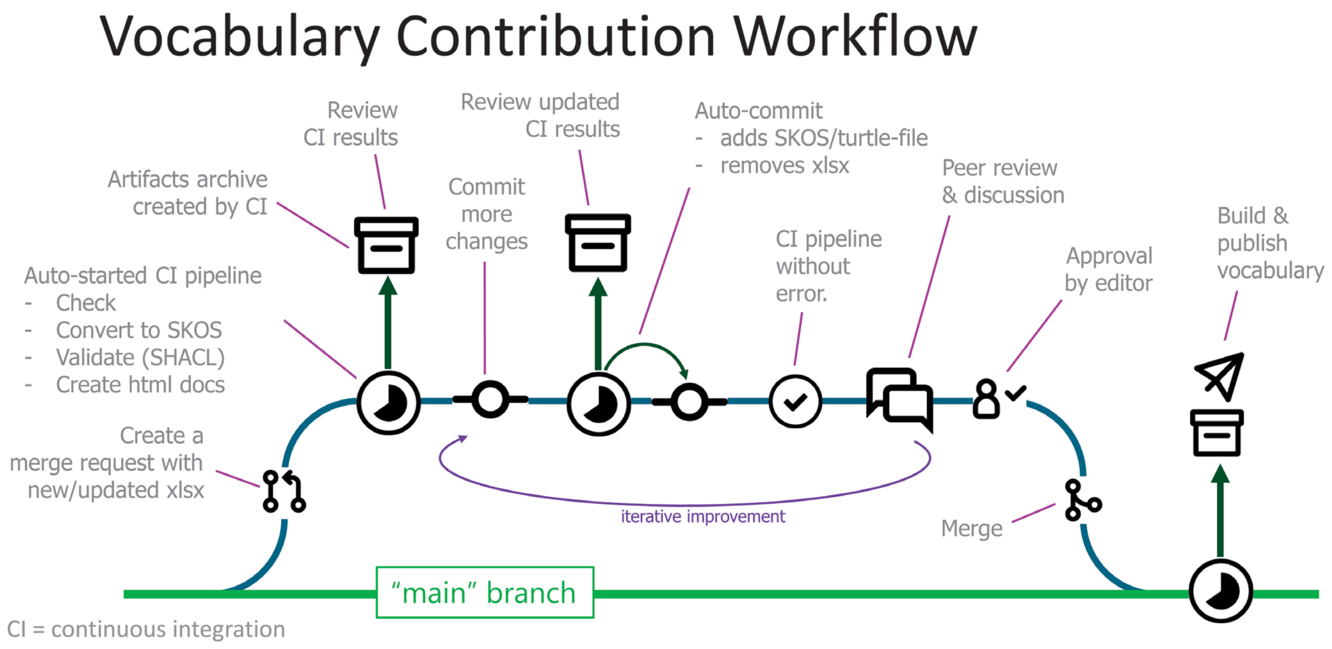


Figure 9: The continuous integration (CI) pipeline.

4.2.5. Knowledge Graphs and Triple Stores for Querying

In the context of NFDI4Cat, knowledge graphs and triple stores play a crucial role in organizing and querying data. Knowledge graphs are semantic networks that describe objects and their interrelationships, aiding in improving data search and understanding of context. As the foundation of knowledge graphs, triple stores manage large data sets in the form of triples (subject, predicate, object).

As noted in the Architecture Document [33], metadata describing the respective experiments are encapsulated into metadata sets, transformed into RDF (Resource Description Framework) format, and integrated into triple stores. This process includes data transformation, maintaining semantic consistency through ontologies, and selecting appropriate triple store technology for scalability and efficient query processing. Moreover, such organization of meta-sets facilitates compatibility and interaction among all project participants.

Using triple stores allows each metadata set to be identified by its unique URI, ensuring direct accessibility to research results and easy navigation between metadata sets. Effective storage and navigation of data in the form of RDF triples in triple stores enable SPARQL to perform complex queries and data analysis. This combination allows for high-level queries for data extraction, updating, and manipulation, which is important within the NFDI4Cat project, where data are often interconnected and multilayered.

However, selecting the appropriate knowledge graph storage solution is a complex task that requires thorough testing. The testing procedure involves analyzing administrative requirements, documentation, licenses, and ensuring user-friendliness. Additionally, the integration of the selected storage solution into the overall project structure shown in Figure 6 is a critical factor. As part of the NFDI4Cat project, testing the triple store‘s suitability and applicability is conducted. The results of this evaluation will be published following a comprehensive analysis of all requirements and community feedback review.

Thus, knowledge graphs and triple stores become tools that assist in efficiently managing user data in catalysis, providing a deeper understanding of complex chemical processes and contributing to the advancement of scientific research in this area.

4.3. Bringing the Machine to Life: Interfaces for Researchers

Besides the theoretical implications of research data management and semantic workflows, researchers need interfaces to implement the solutions in a practicable way. Thus, this chapter highlights contact points with the presented workflows.

4.3.1. Electronic Laboratory Notebooks

Most ELNs offer somehow solutions that can be aligned with semantic techniques. The interface of the ELN is an inseparable element of the researchers' everyday workflow, therefore, needs to provide the flexibility to formulate the entries in an intuitive manner. At the same time the interface should include a large number of functionalities to support the variety of aspects of the experimental work even of a single researcher, e.g. catalyst synthesis steps in comparison to a catalyst performance test, let alone within a research group or an institution. Alternatively, a research group leader could create customizable interfaces or forms for each type of experiment that fits to the group which would provide to junior scientists, which, however, has the drawback of increasing the effort and workload.

In order to avoid the researchers being hindered by the complexity of an interface, it is often preferred by institutions to use multiple solutions that are specialized in a field such as catalyst synthesis or catalytic performance tests. Holistic integration of ELNs, RDM software and even experimental process control units by collaboration between providers of such tools could not only reduce complexity in the user's daily interaction with them but it also allows for more automated research routines. The automatization of the data handling process from collection to storage can significantly reduce the initial hurdle for researchers to use ELNs and RDM tools and thus facilitate their contribution to the FAIR research data principles. These holistically integrated approaches are currently under development and will provide a platform for further applications in catalysis and beyond [17]. In this case, the interconnection between the RDM tools and automation scripts as well as the use of PIDs assigned to each catalyst become of high importance [34].

4.3.2. Querying Knowledge Graphs

Since knowledge graphs can represent complex, not even explicitly formulated, relations between semantic entities, querying these highly structured graphs is on the one hand highly desired and on the other hand challenging. The challenges are on the one side, that a questioner needs to know parts of the structure (like terms, relations) of the knowledge graph to submit a query and on the other side that the current implementations cannot cope with natural language (English) queries - this is still a subject of research. To highly structure and define queries, many attempts have been undertaken, e.g. GraphQL [35], Cypher [36], GQL [37] and SPARQL [38]. All of them have their own syntax and range of applications, but require a high degree of expertise. For querying RDF triples based knowledge graphs, stored in triple stores, the W3C consortium specified [38].

SPARQL is a powerful query (and also knowledge graph manipulation and generation) language, syntactically rooted in SQL. In contrast to other query languages, SPARQL allows federated queries, combining different data resources across the web.

Listing 1 shows a (working) simple example that queries solubilities of substances from wikidata [39]. It illustrates the complexity of the language and the necessity of prior knowledge of the structure of the triple store: e.g., all the terms, starting with wikibase as prefix need to be known at query time.

Similar to the advances made for databases, similar advances can be assumed to happen in the field of SPARQL queries. While large search engines such as Google, Yahoo, and Bing have an eye on using knowledge graphs in their search mechanics [40], they still have the problem of writing precise queries which is still a little bit unwieldy and thus use their more refined search techniques. However, developments in NLP show promising results for translating simple questions into pre-formulated SPARQL queries.

Since knowledge graphs, which are currently the primary focus of NFDI4Cat, are just a very specific way of constructing a graph database, several other methods of querying graphs are also applicable. While SPARQL has a large set of features that are specific to RDF graphs, translating a knowledge graph into another graph format allows for querying with tools such as GraphQL and Cypher. This is currently an applicable way to circumvent some of the syntactic as well as the computational complexity of SPARQL querying.

```
SELECT DISTINCT ?chemicalLabel ?value ?unitsLabel ?solventLabel WHERE {
 ?chemical ?propp ?statement .
 ?statement a wikibase:BestRank ;
  ?proppsv [
   wikibase:quantityAmount ?value ;
   wikibase:quantityUnit ?units
  ] .
 OPTIONAL {
  ?statement prov:wasDerivedFrom/pr:P248 ?source .
  OPTIONAL { ?source wdt:P356 ?doi . }
 }
 ?property wikibase:claim ?propp ;
    wikibase:statementValue ?proppsv ;
    wdt:P1629 wd:Q170731 ;
    wdt:P31 wd:Q21077852 .
  OPTIONAL {
  wd:P2178 wikibase:qualifier ?qualifierS .
  ?qualifierS a owl:ObjectProperty .
  ?statement ?qualifierS ?solvent .
 }
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
} LIMIT 10
```

Listing 1: *Exemplary SPARQL-query for querying solubilities of substances from wikidata, illustrating the complexity of SPARQL and the necessity of prior knowledge of the structure of the knowledge graph.*

4.3.3. Metadata for Domain Specific Search Engines

A pertinent consideration lies in the extraction of metadata from scholarly text providers, such as Scopus. Using the content presented in such vast repositories offers a beneficial base for enrichment of the knowledge base. Utilizing structured querying techniques via API endpoints allows for automated and systematic retrieval of the articles, texts, and metadata. Through programmatic access to the data records, relevant metadata (such as which reactants were used or which reaction was investigated in the respective publications) can be analyzed and extracted from these texts. This can then be used to enrich ontologies with contextual information. Employing NLP methodologies and pre-trained information extraction (IE) models, such as catalysisIE [41], metadata can be extracted from textual publications. Here, information such as catalysts, reactions, and products can be extracted from scholarly works, subsequently organizing this data into an ontology by the help of NLP and IE. Making the resulting knowledge graph accessible for example via SPARQL queries or Protégé‘s graphical user interface can then give deeper and faster insights in current state of the art of scientific literature. While this approach works in principle, more data labelling and further training of the models is needed to obtain more powerful IE models, focusing on more and more concepts that are also described in ontologies. However, with this metadata for domain specific search engines can be retrieved and stored in a knowledge graph automatically.

4.3.5. Metadata Quality Assessment for Increased Research Publications

As described in previous chapters, such as metadata for domain-specific- and large search engines, metadata deposited together with a publication does not only increase the visibility of a research publication for other researchers and the public, but it also demonstrates how viable information extraction is for a specific publication. In this context, measures and awards from funding agencies, publishers and other institutions and initiatives have been proposed and partially established in order to promote and guarantee a liberal and comprehensive data sharing approach among scientists. Publishers, for example, are now implementing the open science badges [https://www.cos.io/initiatives/badges], which show directly how accessible a certain dataset is. Although these badges do not quantify the metadata quality directly, they still have the potential to act as an inherent quality control measure. Also, institutions and organizations, such as the EU, already measure the metadata quality of contributors [28] against a set of criteria using metadata specifications like DCAT-AP [43]. While this is currently more of an incentive, it might become mandatory for some publications, especially for publicly funded research projects. Within the German NFDI landscape also field specific quality control measures are implemented to ensure and assess metadata quality. Within NFDI4Cat, for example, a tool based on the FAIR principles (Piveau metrics) is currently being integrated into the NFDI4Cat Metaportal. Predictive annotation with metadata can thereby help to ensure that data remains relevant in the future. Additionally, when it comes to trust in data and trustworthy data, quality assurance is essential, too, which can be enhanced or made measurable through ontologies.

4.3.6. Automating (Meta) Datastreams

Most scientists cannot spend a lot of time on enriching their data with proper metadata, therefore,
the success of the FAIR principle will only be achieved, when the "boring repetitive" work will be automated. A very powerful technology to automatically generate knowledge content, like individuals, offer open source, royalty free and vendor agnostic lab-automation communication protocols, like SiLA [44] and OPC-UA/LADS [45]. They enable an abstract and generalized control of lab devices and services, like machine learning services or lab workflow orchestration. They further help to channel data in a laboratory network, interact with ELNs and LIMS systems, and assist to enrich data with metadata in an automated fashion. This reduces the burden of the scientist to track all the data and automate many repetitive steps. Well combined and structured they have the potential to form the data transport backbone of semantically enriched data streams and a data exploration infrastructure.

4.3.7. Software-assisted Generation and Inspection of RDF Metadata

In general, the translation of the ontology-based model into valid RDF metadata for particular research steps is associated with significant challenges. Despite a seeming simplicity of RDF, manually creating RDF metadata can be akin to a programming-like experience, demanding a good understanding of both the ontology model and the syntactic rules of the chosen RDF serialization format. In addition, assigning RDF resources to ontology classes, describing relationships between resources using ontology properties and translating them into RDF triples involve meticulous attention to detail. After all, the generated RDF metadata have to be syntactically checked, semantically validated against the ontology, as well as checked for completeness. The adoption of linked semantic technologies by the community faces significant hurdles due to these challenges.

For circumventing these difficulties, the HLRS team is actively developing a software suite to assist researchers in creating RDF metadata. To achieve this objective, the reference ontology needs augmentation with suitable logical constructs, facilitating efficient decision-tree support through a reasoner-based questionnaire. This program is designed to integrate seamlessly with a user-friendly web interface for enhanced interaction. Another crucial concern involves the representation of RDF data, which, in most cases beyond the trivial, are hardly comprehensible in raw form. To address this, automatically transforming RDF data into accessible HTML or Markdown formats would greatly enhance the usability of knowledge graphs in the community. The above-mentioned program suite will incorporate such a tool. This task is particularly appropriate for RDF data, considering that all resources within them are assigned URIs. In this way, the generation, processing and visualization of RDF metadata can be performed without demanding advanced skills from researchers. This would consequently enhance the usability of semantic technologies within the community.

4.3.8. Mechanisms for Alignment: The Role of Domain Expert Input

The involvement of domain experts is an integral organ in the workflow of ontology-based research data management. The formal arrangement can be achieved through committees for survey of activities and corrective input to the RDM system, similarly to a mechanism that has proven useful in context of the Wikipedia system. The workflows should encompass detailed processes tailored to be nurtured by different expert inputs. The input should include mechanisms for vocabulary curation utilizing general user input derived from literature, experimental, or simulation data sources, involving text processing to identify terms and relations. Automated methods are currently developed and should be further fostered and employed to map terms to classes and individuals, with alignment activities conducted with neighboring NFDI consortia in chemistry, engineering, and materials, as well as general semantic initiatives like Wikidata. Ontology extension and curation are carried out in collaboration with NFDI base services, ensuring alignment with top-level ontology and employing reasoning mechanisms to prevent inconsistencies.

The development of the NFDI4Cat toolbox will give more information on the concrete user input and far-reaching automation of the process.

The automated workflow integrates expert input and provides regular reporting of key results, managing conflicts effectively. PID management and rights management are incorporated, with reporting occurring on a monthly, quarterly, or annual basis, alongside alignment activities with other NFDI consortia and base services. Ontology-based metadata functions as a template, seamlessly integrated into the RDM architecture. This covers data storage, curation, quality control, data analysis, and augmentation to scientific knowledge. The workflow supports data sharing and publication, enabling the reuse of data for research planning in experiments and simulations, facilitating analysis, and connecting with other data sources and information. The overarching goal is to close feedback loops and continually enhance data management practices.

5. Application of the RDM-Engine: Example Solutions and Future Implementations

This chapter aims to describe success stories of NFDI4Cat-related applications with focus on APIs and important interfaces for the workflow. As potential users might have different levels of “semantic readiness”, different usability is addressed in the following subchapters. Basic users might be at the start of using semantic technology and more prone to use solutions that offer a low hurdle but also low semantic expressivity. Experienced users might be able to use semantically more enhanced applications, such as querying a semantic data base with SPARQL. Finally, expert users might see themselves fit enough to develop their own semantic RDM workflows based on the presented ones and tweak them to their needs.

5.1. LARAsuite

The LARAsuite [46] is an open source and very modular research data management system and integrated lab automation environment, designed to cover the full scientific workflow in natural sciences (from biosciences, over earth sciences to chemical- and physical sciences). For that purpose, it consists of modules that interlink data sources of common experimentation, e.g. linking materials, substances, organisms, processes and processes to performed experiments, structuring experiments in projects, etc. It assist in structuring core information of experimentation and adding automatically a semantic metadata layer to the experiment description. Through its modular design and microservice architecture, all modules can be expanded to the individual needs of a particular experimentation. LARAsuite is designed with workflow automation and semantic metadata enrichment from the ground up, see Figure 10. This means that it generically supports and orchestrates SiLA (s. chapter 4.3.6) lab automation servers, collects the produced data into an object storage, adds the metadata to a triple store (virtuoso) and allows data evaluation and visualization with ELT pipelines for machine learning and AI (e.g., prefect). The triple store can be queried through SPARQL (s. chapter 2.3.4). For simple data access, a convenience python library is avail, which makes the large gRPC API of the LARA database extremely simple to use. Several instances of LARA can share data through a synchronization protocol. Connectors to public data repositories, like dataverse, Zenodo etc. are under development.

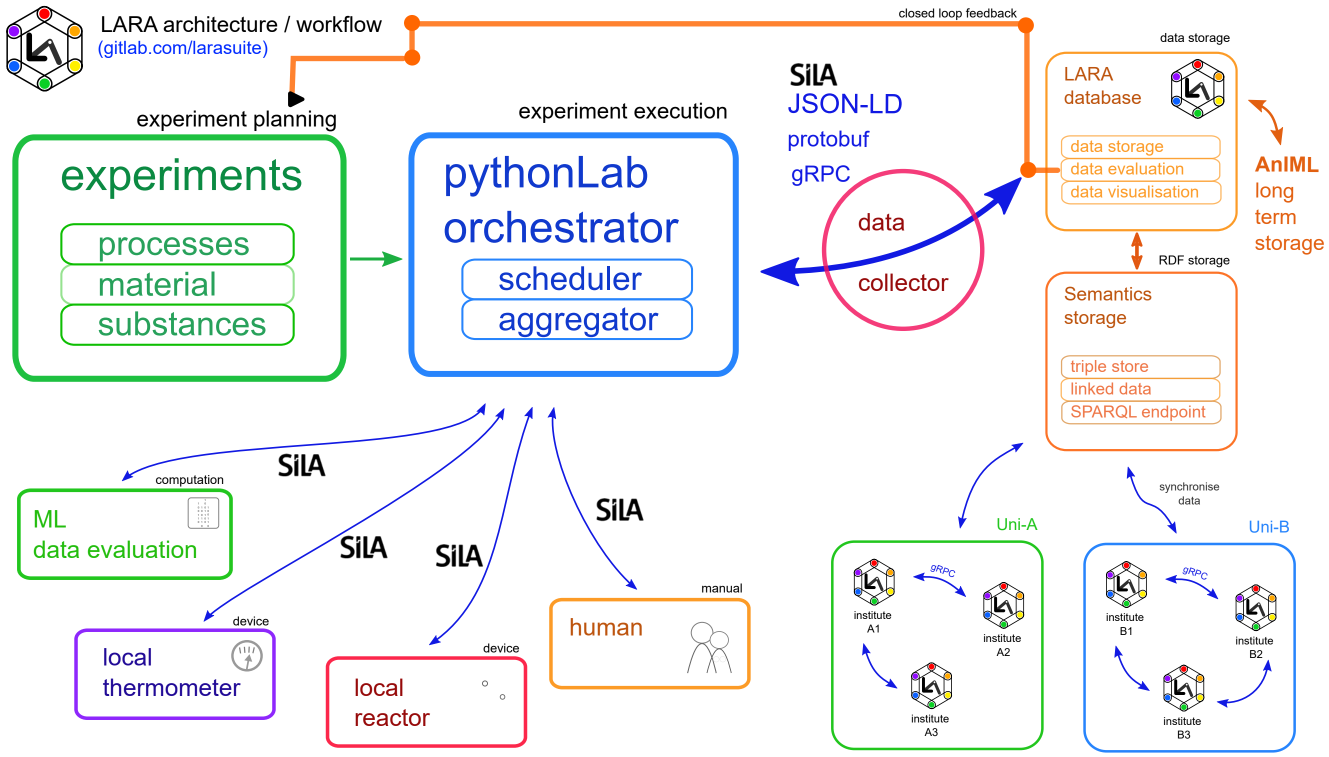


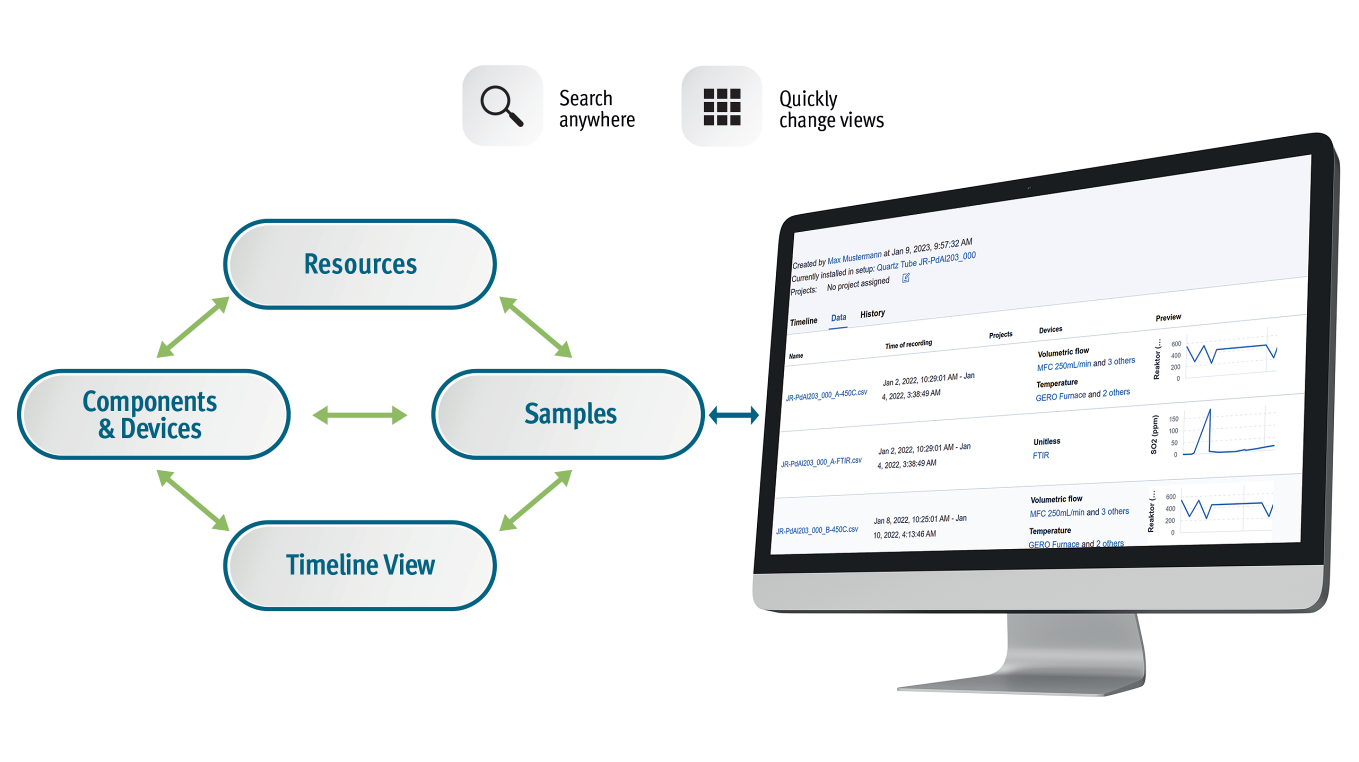
Figure 10: *LARA architecture*

5.2. ADACTA

In the context of a specialized solution for management of catalytic performance experimental data, the software Adacta was developed within NFDI4Cat. The basic concept behind Adacta is the efficient data archiving of time-stamped experimental measurements associated by the complex metadata which arise because of the reaction parameters and the used devices [17]. Adacta is highly flexible in that it does not rely on rigid predefined "forms" to input accompanying metadata which lead to high maintenance efforts due to the many customized forms needed to support a wide variety of experimental setups that can evolve with time.

In Adacta, each experimental reactor setup within a lab has its own digital twin, including all of its components. Mass flow controllers, pressure valves, reactor tubes etc. are specified along with their technical information such as range of mass flow or dimensions. The experimental setup configuration is assigned to a point in a time line and is easily maintained in case of future changes in the physical lab. This provides the possibility to keep track of the exact status of the experimental setup at the time of each experimental measurement.

The experimental raw data are also organized based on the time-stamp which assure its uniqueness for a specific reactor setup. Raw data are also associated with the specific device that produces them as for example a thermocouple measures the reactor temperature or a FTIR gives a signal of gas composition. In this way, it is assured that no information will be lost and that it can be easily traced back at any time. A schematic representation of the user interface that enables the quick transition between Resources (datasets), Device (experimental setups) and Samples is depicted in Figure 11.



Created by Max Mustermann at Jan 9, 2023, 9:57:32 AM
Currently installed in setup: Quartz Tube JR-PdAl203_000
Projects: No project assigned

Name	Time of recording	Projects	Devices	Preview
JR-PdAl203_000_A-450C.csv	Jan 2, 2022, 10:29:01 AM - Jan 4, 2022, 3:38:49 AM		Volumetric flow MFC 250mL/min and 3 others Temperature GERO Furnace and 2 others	
JR-PdAl203_000_A-FTIR.csv	Jan 2, 2022, 10:29:01 AM - Jan 4, 2022, 3:38:49 AM		Unitless FTIR	
JR-PdAl203_000_B-450C.csv	Jan 8, 2022, 10:25:01 AM - Jan 10, 2022, 4:13:46 AM		Volumetric flow MFC 250mL/min and 3 others Temperature GERO Furnace and 2 others	

Figure 11: *Schematic representation of the concept of the user interface in Adacta.*

Data repositories control the physical location where the experimental raw data and related metadata are stored. Therefore, depending on user requirements, data can be stored locally on a single computer, or remotely on a company or department server. If data are stored on remote servers, then a user can have access to the whole database from multiple computers.

Advanced searching options can filter the datasets based on the measurements with a specific sample or the measurements within a certain period in time. One more option with practical importance is the search of the raw data measured by an experimental setup composed of a specific device that was discovered at a later time that it was malfunctioning. This gives the advantage to trace which measurements were actually affected by the technical problem.

Extended visualization features allow comparison of signals from different experiments or combined view of signals within a dataset, filters etc. Additional features are the access control, i.e. a mechanism that gives the creator of the data the possibility to control which other users have access (including read-only or read/write permissions) complying with non-disclosure agreements and controlling intellectual property.

As the database of information expands, tracing the history of catalysts across reactor setups will become feasible, to identify the smallest details that impact experimental performance, even far into the future when the experimental test stand is decommissioned, and key personnel have moved on to other roles. By linking with ELNs and other resource media via common ontology and terminology, all documents, reports, and data generated for a particular experiment can be stored in a single repository.

5.3. Linking Local and Overarching Data Infrastructures in NOMAD

Complete and machine readable data sets are an important requirement for the application of artificial intelligence methods to find new catalyst materials or to better understand relationships between material properties and performance [47–49]. The basis for the generation of AI-compatible data sets is the development of local data infrastructures that are adapted to the needs of specific research tasks. The connection of initial data collection, converting it into open formats, data analysis and publishing FAIR data completes the full lifecycle of research data management (RDM).

NOMAD is a web-based application dedicated to managing and disseminating data from materials science research and serves as an open repository [50]. It serves a broad spectrum of needs within the materials science community, for example in computational and experimental research fields, including heterogeneous catalysis as an important use case. At its core, the NOMAD central service acts as a comprehensive archive and repository, housing over 13 million entries from thousands of contributors. This wealth of data is not only stored but made ready for AI applications through its structured format, rich metadata, and adherence to FAIR (Findable, Accessible, Interoperable, Reusable) data principles. Such organization ensures that the information is both machine and human-readable, facilitating seamless access and analysis through robust Application Programming Interfaces (APIs). By integrating general and example data schemas for heterogeneous catalysis, and ensuring data items such as sections and quantities are semantically enriched and interoperable (referencing community standards like ontology definitions or Voc4Cat with unique internationalized resource identifiers (IRIs)), NOMAD is positioned to facilitate advanced and interoperable RDM for catalysis-based research.

Standard operating procedures (SOPs), which are documented in machine-readable handbooks [51, 52], increase the reproducibility and comparability of experimental catalysis data. In order to comply with standardized data acquisition and storage, a local RDM concept was developed at the Department of Inorganic Chemistry at the Fritz Haber Institute of the Max Planck Society, which focuses on the automated execution of experiments. The automation of laboratory reactors, routine and advanced characterization techniques is implemented using EPICS, an open-source system that provides tools for experiment control [53–55]. A database (AC/CATLAB Archive) was established to store, visualize and retrieve the data and metadata [56]. The database functions both as a data archive and as an electronic laboratory notebook. The scientist develops measurement methods that are entered via graphical user interfaces (GUIs) and automatically stored in the database in the form of JSON and HDF5 files, where they can be re-accessed at any time. In this way, machine-readable handbooks are generated via GUIs also for manual experiments [57]. Routines developed based on EPICS carry out the experiments [58, 59]. Raw data and data evaluated with the help of Python scripts in a standardized way are automatically uploaded to an S3 server and into the database in different formats including the HDF5 format [60]. The unique feature of the present RDM solution is that the data and metadata are automatically linked to the catalyst sample and other relevant information in the database, such as the measurement method, the reactor and the chemicals used in the experiment. Catalyst IDs are also generated automatically after automated treatments that chemically or physically change the sample and thus lead to a new sample, such as a spent catalyst after a catalytic test. The development of tools using Python, with which the database entries and connections can be displayed graphically as nodes and edges, is in progress. The structured HDF5 format enables easy comparison between data sets. The API allows the data to be used directly for machine learning algorithms or to be uploaded to overarching repositories such as NOMAD.

Another local data management and storage solution developed in FAIRmat is NOMAD Oasis [50]. A high level of adaptability allows institutions to customize data schemas to fit their unique research setups and experiments with controlled vocabulary, including options for automated processing and analysis. This customization extends to faceted search interfaces, enhancing the discoverability and accessibility of research data. Furthermore, NOMAD Oasis integrates Electronic Laboratory Notebook (ELN) features, allowing for a hybrid approach to data entry - combining manual input with automatic parsing - to ensure the creation of consistent, structured datasets. A distinctive feature of NOMAD Oasis is its ability to connect with the central NOMAD service, enabling researchers to publish their data seamlessly.

5.4. Exemplary ELN - FURTHRmind

RWTH Aachen University is also working on the implementation of RDM solutions such as the integration platform for research data Coscine („Collaborative Scientific Integration Environment"). This is currently in the pilot phase and is intended to support researchers in organizing, storing and sharing their research data. Established services such as the research data repository (FDS.NRW) and GitLab are already being used and external services such as Sciebo will also be used in the future. Coscine is an open source solution of the IT Center of the RWTH and tries to implement data storage according to the FAIR principle. Above all, the software works according to the principle that no single solution can be found that fits all researchers, which is why Coscine pursues an integrative approach. As already described in chapter 4.3.1, there are many solutions for ELNs. The spin-off founded at AVT.CVT at RWTH Aachen University focuses on a different approach. FURTHRmind is a state-of-the-art software solution to manage all research data, meaning all samples, measurements, experiments, metadata, and raw data of the researchers. Additionally, data can be analyzed directly in FURTHRmind or from other software for data analysis via the REST-API. The test stands at the AVT.CVT are controlled via control boxes from the spin-off ZUMOLab. These contain a direct integration with the data management software FURTHRmind, whereby the metadata set during an experiment and the measured experimental data are stored directly in memory and uploaded for further analysis. Templates were created for a wide range of application areas that can be used for any application and reduce the workload. The template database is constantly updated, which means that a large pool of metadata, a standardized vocabulary and a standardized storage of metadata can be operated across all levels from students to professors.

5.5. Cross Domain Meta Ontology

This chapter introduces ontologies utilized to represent various aspects of research data within different NFDI NFDI-MatWerk. The focus lies on the NFDIcore ontology [61], serving as a generic framework for representing research resources across consortia. The modular approach is employed to extend core concepts based on the requirements of each consortium, enhancing consistency, clarity, and reusability of representations, and facilitating knowledge discovery across diverse domains. NFDIcore is developed to represent research resources such as data sets, providers, persons, and areas of expertise across NFDI consortia, forming the basis for further domain-specific ontologies, such as the NFDI4Culture [62] and NFDI-MatWerk [63] ontologies, which address specific domain needs for NFDI4Culture and NFDI-MatWerk respectively [64]. This modular approach establishes standard vocabularies and structures, boosting reusability and uncovering new relationships and insights across different domains. The current version of the NFDIcore ontology comprises 33 classes and 60 properties covering various aspects of research data management, guided by established standards and best practices, including the FAIR Data Principles. To ensure interoperability, NFDIcore links to around 20 external vocabularies. Additionally, extensions cater to unique requirements of different research fields, such as the NFDI4Culture and NFDI-MatWerk ontologies. Development follows a user-centered design and evaluation methodology, with incremental and iterative requirements development for different user groups. As NFDI ontologies are intended to be dynamic and evolving to adapt to changing community needs, every version is accessible and referential online.

5.6. Research Data Management Tool for Catalysis Laboratory Courses

The Department of Chemistry at the Technical University of Munich (TUM) offers the „Technische Chemie Praktikum" Laboratory Course to Chemistry (CHEM) and Chemical Engineering (CIW) undergraduate students. The laboratory coursework includes catalysis-related experiments (e.g., homogeneous catalysis, heterogeneous catalysis, biocatalysis) that are performed by students. The course typically has ~100 students enrolled each semester and each student performs at least eight different experiments. This results in more than 800 experimental datasets and reports generated each year that are evaluated by the supervisors (typically PhD students and Postdocs). A Research Data Management tool has been developed, „RDM4Lab", for systematically storing the data generated in this Laboratory Course in compliance with FAIR (Findable, Accessible, Interoperable, and Reusable) storage principles. The tool incorporates features like visualization of the current and historic data and automated data analysis to help the supervisors grade the reports. It is plan to make this tool compatible with ontologies, vocabularies, and metadata standards developed as a part of NFDI4Cat. The future is to use this tool as a real-world test case for the implementation of other tools developed at NFDI4Cat. The implementation of RDM4Lab at undergraduate level will also enable training of students with research data management. This tool will be implemented in full scale at TUM from summer terms 2024. Due to its modular nature, it can be implemented in other universities and institution for their (catalysis) laboratory related courses.

5.7. Envisioned Future Implementations

Future applications of the tools and workflows discussed and presented encompass crucial areas of cooperation within the catalysis-related sciences. One important element is the focus on creation and development of specific knowledge graphs tailored to collaboration topics of interest among collaborating entities. This is envisioned to enhance data FAIRness and also to improve workflows cost- and time-wise, as these tailored knowledge graphs serve as dynamic repositories of specialized information. Future developments will encompass the structured querying of knowledge graphs with SPARQL-queries formed by Natural-Language-Processing tools.

Moreover, the scale-up of these infrastructures involves the interlinkage of knowledge graphs, yielding organic connections, enabling complex ways of knowledge representation. On the other hand, this also makes data storage, e.g. in a dataverse, a more complicated task. Here, methodologies are to be determined to populate repositories in a meaningful and trustful fashion, to enable the full potential of the pipelines presented in this work.

In order to arrive at data spaces for catalysis related sciences and be able to link open science with a data economy, the creation and use of data spaces tailored for catalysis related sciences are to be investigated more. Further pathways must be identified and established focusing on the population of a data space via linked knowledge graphs.

In future, applications, the tools and pipelines presented in this work aim to enable the researchers for enhanced research data FAIRness with minimal or no additional workload during data uptake and recording. On top of that, the creation of data repositories and knowledge graphs will allow for faster data retrieval, also increasing efficiency for the researchers of catalysis related sciences.

6. Conclusion

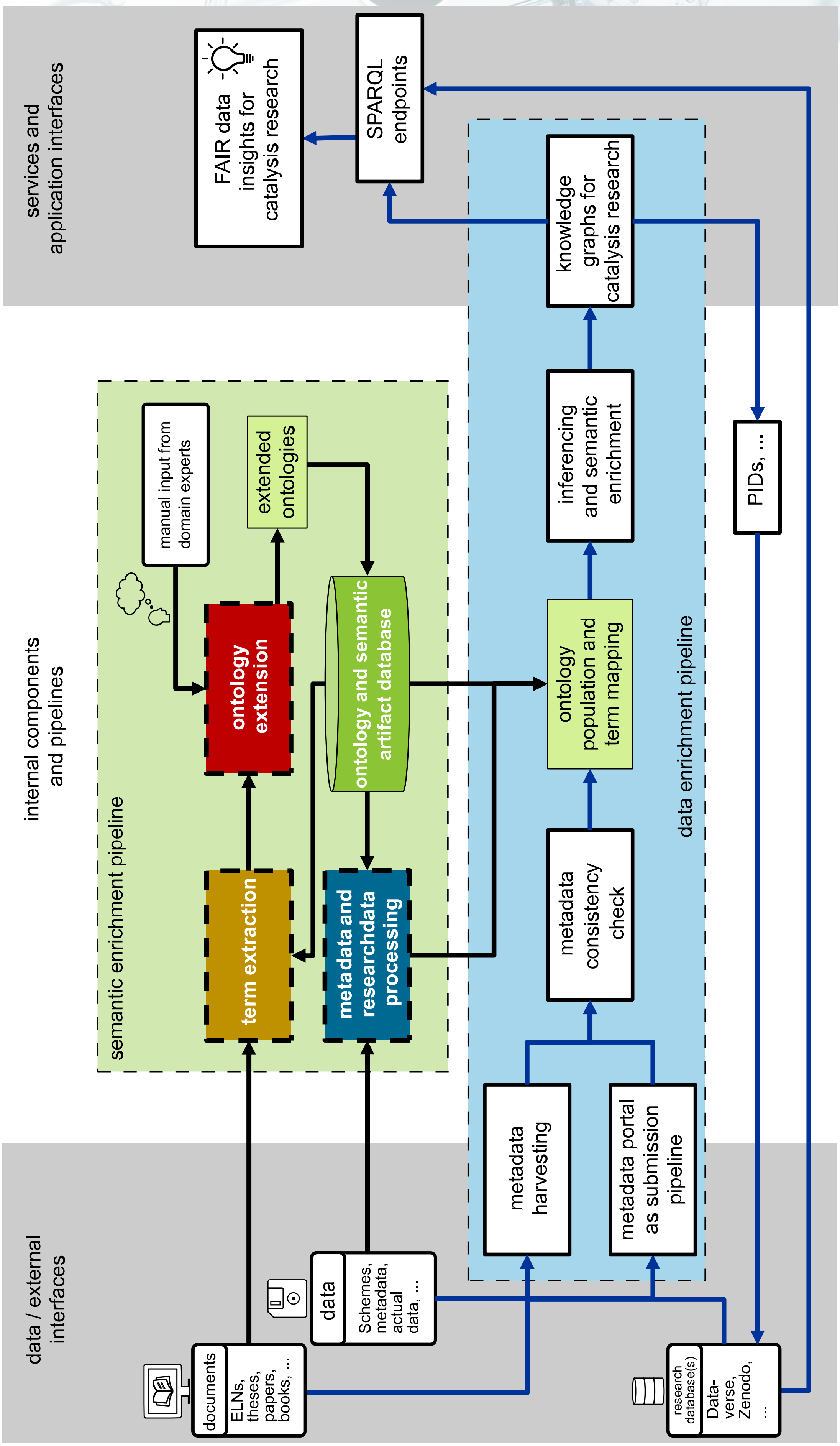
This White Paper represents a current overview on the landscape of semantic applications in the domain of catalysis research and related sciences. After the introduction, important criteria for the RDM landscape are defined and some major findings from the last three years are summarized.

Technical insights in catalysis research involve the utilization of vocabularies and Electronic Laboratory Notebooks (ELNs) to structure, record, and organize diverse research data. The recent insights include the interplay between various technical solutions such as Persistent Identifiers (PIDs), semantic enrichment pipelines, metadata validation pipelines, and knowledge graph setups. The application of these solutions extends to ELNs, querying knowledge graphs, metadata applications, and software-guided generation of knowledge graphs, enhancing research efficiency and data organization in catalysis studies. Selected application showcases are highlighted for currently already working examples of the semantic techniques. These include LARAsuite, the ELN FURTHRmind, and ADACTA as mature RDM software, thus enabling for semantic RDM with good FAIR practice. Furthermore, a way of setting up an own, local data infrastructure is shown, as well as the implementations of NOMAD and NOMAD Oasis, which provide web-interfaces for highly semantic databases. To connect existing and future developed ontologies, a cross-domain meta ontology is presented, that aims to connect the semantics of the different research domains for better alignment of FAIR data workflows. Finally, an exemplary laboratory course shows the implementation of the FAIR principles in a hands-on teaching environment.

In the near future, the following tasks are envisaged to complete the picture:

- further development of the vocabularies for catalysis research, refining the selection of ontologies by help of competency questions,
- use-case oriented design of ontologies,
- pilot applications for showcasing the benefit of ontologies, and
- enhanced querying systems for natural language - based queries on knowledge graphs will/should be developed to ease data retrieval from knowledge graphs.

As a final vision, an overall research data management workflow is necessary with user input from researchers including literature and patent survey, gathering experimental and simulation data, integrated data curation and analysis with information extraction, data storage and sharing (partly), publications and re-usage by different user groups. Different contact points are important for end users within the presented workflows and pipelines. Here, more experience with different applications is necessary in the near future.


data / external interfaces
documents
ELNs, theses, papers, books, ...
data
Schemes, metadata, actual data, ...
research database(s)
Data-verse, Zenodo, ...
internal components and pipelines
semantic enrichment pipeline
term extraction
metadata and researchdata processing
ontology extension
ontology and semantic artifact database
manual input from domain experts
extended ontologies
data enrichment pipeline
metadata harvesting
metadata portal as submission pipeline
metadata consistency check
ontology population and term mapping
inferencing and semantic enrichment
knowledge graphs for catalysis research
PIDs, ...
services and application interfaces
SPARQL endpoints
FAIR data insights for catalysis research

7. References

[1] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. ,t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, Scientific data 2016, 3, 160018. DOI: https://doi.org/10.1038/sdata.2016.18

[2] M. J. Menke, A. S. Behr, K. Rosenthal, D. Linke, N. Kockmann, U. T. Bornscheuer, M. Dörr, Chemie Ingenieur Technik 2022, 94 (11), 1827 – 1835. DOI: https://doi.org/10.1002/cite.202200066

[3] Goldbeck Gerhard, Simperler Alexandra, Report On Workshop On Interoperability In Materials Modelling, Zenodo 2018, DOI: https://doi.org/10.5281/zenodo.1240229.

[4] M. Horsch, T. Petrenko, V. Kushnarenko, B. Schembera, B. Wentzel, A. Behr, N. Kockmann, S. Schimmler, T. Bönisch, in Data Analytics and Management in Data Intensive Domains, Vol. 1620, Communications in Computer and Information Science (Eds: A. Pozanenko et al.), Springer International Publishing. Cham 2022.

[5] A. S. Behr, H. Borgelt, N. Kockmann, Journal of cheminformatics 2024, 16 (1), 16. DOI: https://doi.org/10.1186/s13321-024-00807-2

[6] D. Linke, N. Moustakas, nfdi4cat/voc4cat: Release 2023-09-03, Zenodo 2023, DOI: https://doi.org/10.5281/zenodo.8313341.

[7] D. Beckett, T. Berners-Lee, E. Prud'hommeaux, G. Carothers, RDF 1.1 Turtle: Terse RDF Triple Language, W3C, https://www.w3.org/TR/turtle/ (Accessed on May 14, 2022).

[8] N. Moustakas, A. S. Behr, H. Borgelt, N. Huskova, R. Khare, M. Talab, J. Köbl, V. Chandrashekhar, T. Petrenko, M. Dörr, D. Linke, Voc4cat: Vocabulary guidelines for NFDI4Cat, Zenodo 2023.

[9] American National Standards Institute/National Information Standards Organization (ANSI/NISO), ANSI/NISO Z39.19-2005 (R2010) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies 2010.

[10] E. M. Williamson, Z. Sun, L. Mora-Tamez, R. L. Brutchey, Chem. Mater. 2022, 34 (22), 9823 – 9835. DOI: https://doi.org/10.1021/acs.chemmater.2c02924

[11] H. A. Nguyen, F. Y. Dou, N. Park, S. Wu, H. Sarsito, B. Diakubama, H. Larson, E. Nishiwaki, M. Homer, M. Cash, B. M. Cossairt, Chem. Mater. 2022, 34 (14), 6296 – 6311. DOI: https://doi.org/10.1021/acs.chemmater.2c00640

[12] R. Buonsanti, Chem. Mater. 2023, 35 (3), 805 – 806. DOI: https://doi.org/10.1021/acs.chemmater.3c00019

[13] C. L. Bird, C. Willoughby, J. G. Frey, Chemical Society reviews 2013, 42 (20), 8157 – 8175. DOI: https://doi.org/10.1039/C3CS60122F

[14] K. Cruse, A. Trewartha, S. Lee, Z. Wang, H. Huo, T. He, O. Kononova, A. Jain, G. Ceder, Scientific data 2022, 9 (1), 234. DOI: https://doi.org/10.1038/s41597-022-01321-6

[15] B. Bayerlein, M. Schilling, H. Birkholz, M. Jung, J. Waitelonis, L. Mädler, H. Sack, Materials & Design 2024, 237, 112603. DOI: https://doi.org/10.1016/j.matdes.2023.112603

[16] I. Bagov, M. Meller, F. Bresser, N. Ye, N. Garabedian, VocPopuli, Zenodo 2023.

[17] H. Gossler, J. Riedel, E. Daymo, R. Chacko, S. Angeli, O. Deutschmann, Chemie Ingenieur Technik 2022, 94 (11), 1798 – 1807. DOI: https://doi.org/10.1002/cite.202200064

[18] K. Stathis, C. Ross, S. Vogt, K. Siziva, Introducing DataCite Metadata Schema 4.5, DataCite 2024.

[19] DataCite Metadata Working Group, DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs v4.5, DataCite 2024.

[20] ePIC Consortium, ePIC: A Consortium Providing PID Services for the Research Community 2009, http://www.pidconsortium.net/.

[21] Corporation for National Research Initiatives, Handle.Net Registry (HNR) and Corporation for National Research Initiatives (CNRI) 2023, https://handle.net/.

[22] W3C, w3id.org 2024, https://w3id.org/.

[23] Internet Archive, PURL Administration 2022, https://purl.archive.org/.

[24] Helmholtz Metadata Collaboration (HMC), Institute for Materials Data Science, Informatics (IAS-9), PIDA - Persistent

Identifiers for Digital Assets 2022, https://purls.helmholtz-metadaten.de/.

[25] S. Moxon, D. Unni, G. Vaidya, H. Hegde, S. Patil, K. Schafer, P. Kalita, N. Harris, T. Putman, H. Solbrig, M. Haendel, C. Mungall, LinkML, Zenodo 2024.

[26] F. Kirstein, K. Stefanidis, B. Dittwald, S. Dutkowski, S. Urbanek, M. Hauswirth, in The Semantic Web, Vol. 12123, Lecture Notes in Computer Science (Eds: A. Harth et al.), Springer International Publishing. Cham 2020.

[27] R. Albertoni, A. Isaac, Data on the Web Best Practices: Data Quality Vocabulary 2016.

[28] European Commission, European Open Data 2023, https://data.europa.eu/.

[29] D. Linke, nfdi4cat/voc4cat-template: Template release matching release 0.7.8 of nfdi4cat/voc4cat-tool, Zenodo 2023, DOI: https://doi.org/10.5281/zenodo.8306845.

[30] TIB – Leibniz Information Centre for Science, Technology, University Library, TIB Terminology Service 2023, https://terminology.tib.eu/ts/.

[31] S.-A. Sansone, P. McQuilton, P. Rocca-Serra, A. Gonzalez-Beltran, M. Izzo, A. L. Lister, M. Thurston, Nature biotechnology 2019, 37 (4), 358 – 367. DOI: https://doi.org/10.1038/s41587-019-0080-8

[32] European Open Science Cloud (EOSC), European Open Science Cloud 2023, https://eosc-portal.eu/.

[33] T. Bönisch, Y. Dikova, M. Doerr, N. Huskova, V. Kushnarenko, D. Linke, T. Petrenko, P. Rodrigues, S. Schimmler, B. Wentzel, Y. Zhang, NFDI4Cat: Architecture document, Zenodo 2023.

[34] R. Chacko, H. Gossler, S. Angeli, O. Deutschmann, ChemCatChem 2024, 16 (4). DOI: https://doi.org/10.1002/cctc.202301355

[35] A. Quiña-Mera, P. Fernandez, J. M. García, A. Ruiz-Cortés, ACM Computing Surveys 2023, 55 (10), 1 – 35. DOI: https://doi.org/10.1145/3561818

[36] N. Francis, A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault, S. Plantikow, M. Rydberg, P. Selmer, A. Taylor, in Proceedings of the 2018 International Conference on Management of Data (Eds: G. Das, C. Jermaine, P. Bernstein), ACM. New York, NY, USA 2018.

[37] ISO/IEC, Information technology - Database languages - GQL, 1st ed.

[38] W3C, SPARQL 1.1 Overview, W3C Recommendation 2013, https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/.

[39] Wikidata, Wikidata:SPARQL query service/queries/examples 2024, https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples#Solubilities_of_chemicals.

[40] R. V. Guha, D. Brickley, S. MacBeth, Queue 2015, 13 (9), 10 – 37. DOI: https://doi.org/10.1145/2857274.2857276

[41] Y. Zhang, C. Wang, M. Soukaseum, D. G. Vlachos, H. Fang, Journal of chemical information and modeling 2022, 62 (14), 3316 – 3330. DOI: https://doi.org/10.1021/acs.jcim.2c00359

[42] A. J. G. Gray, Carole Goble, R. C. Jimenez, in ISWC 2017 Posters & Demonstrations and Industry Tracks, CEUR workshop proceedings (Eds: Nadeschda Nikitina, Dezhao Song, A. Fokoue, P. Haase), RWTH Aachen University. Germany 2017.

[43] Jitse De Cock, Makx Dekkers, Pavlina Fragkou, Arthur Schiltz, Anastasia Sofou, Bert Van Nuffelen, DCAT-AP 3.0 2024, https://semiceu.github.io/DCAT-AP/releases/3.0.0.

[44] The SiLA Consortium, Standardisation in Labautomation 2021, https://sila-standard.org/.

[45] OPC Foundation, SPECTARIS, Mechanical Engineering Industry Association, LADS – Laboratory and Analytical Device Standard 2024, https://www.spectaris.de/en/association/thespectarisindustries/networked-laboratory-equipment/.

[46] M. Doerr, The LARA Suite 2023, https://github.com/LARAsuite.

[47] L. Foppa, L. M. Ghiringhelli, F. Girgsdies, M. Hashagen, P. Kube, M. Hävecker, S. J. Carey, A. Tarasov, P. Kraus, F. Rosowski, R. Schlögl, A. Trunschke, M. Scheffler, Materials genes of heterogeneous catalysis from clean experiments and artificial intelligence, arXiv 2021.

[48] L. Foppa, F. Rüther, M. Geske, G. Koch, F. Girgsdies, P. Kube, S. J. Carey, M. Hävecker, O. Timpe, A. V. Tarasov, M. Scheffler, F. Rosowski, R. Schlögl, A. Trunschke, Journal of the American Chemical Society 2023, 145 (6), 3427 – 3442. DOI: https://doi.org/10.1021/jacs.2c11117

[49] C. P. Marshall, J. Schumann, A. Trunschke, Angewandte Chemie (International ed. in English) 2023, 62 (30), e202302971. DOI: https://doi.org/10.1002/anie.202302971

[50] M. Scheidgen, L. Himanen, A. N. Ladines, D. Sikter, M. Nakhaee, Á. Fekete, T. Chang, A. Golparvar, J. A. Márquez, S. Brockhauser, S. Brückner, L. M. Ghiringhelli, F. Dietrich, D. Lehmberg, T. Denell, A. Albino, H. Näsström, S. Shabih,

F. Dobener, M. Kühbach, R. Mozumder, J. F. Rudzinski, N. Daelman, J. M. Pizarro, M. Kuban, C. Salazar, P. Ondračka, H.-J. Bungartz, C. Draxl, JOSS 2023, 8 (90), 5388. DOI: https://doi.org/10.21105/joss.05388

[51] A. Trunschke, G. Bellini, M. Boniface, S. J. Carey, J. Dong, E. Erdem, L. Foppa, W. Frandsen, M. Geske, L. M. Ghiringhelli, F. Girgsdies, R. Hanna, M. Hashagen, M. Hävecker, G. Huff, A. Knop-Gericke, G. Koch, P. Kraus, J. Kröhnert, P. Kube, S. Lohr, T. Lunkenbein, L. Masliuk, R. Naumann d'Alnoncourt, T. Omojola, C. Pratsch, S. Richter, C. Rohner, F. Rosowski, F. Rüther, M. Scheffler, R. Schlögl, A. Tarasov, D. Teschner, O. Timpe, P. Trunschke, Y. Wang, S. Wrabetz, Top Catal 2020, 63 (19-20), 1683 – 1699. DOI: https://doi.org/10.1007/s11244-020-01380-2

[52] A. Trunschke, Catal. Sci. Technol. 2022, 12 (11), 3650 – 3669. DOI: https://doi.org/10.1039/D2CY00275B

[53] The EPICS Archiver Appliance, https://slacmshankar.github.io/epicsarchiver_docs/index.html.

[54] EPICS Documentation 2024, https://epics-controls.org/.

[55] Bluesky Project, Ophyd Documentation 2014, https://blueskyproject.io/ophyd.

[56] Michael Wesemann, archive GitHub Repository 2022, https://github.com/fhimpg/archive.

[57] Abdulrhman Moshantaf, FHI-AC JSON-Scripte 2022, https://gitlab.fhi.mpg.de/fhi-ac/json-scripte.

[58] A. Moshantaf, P. Oppermann, H. Junkes, Haber - Catalytic test reactor for ammonia decomposition 2023, https://gitlab.fhi.mpg.de/fhi-ac/haber.

[59] A. Moshantaf, P. Oppermann, H. Junkes, Ertl - Catalytic test reactor for CO oxidation 2023, https://gitlab.fhi.mpg.de/fhi-ac/ertl.

[60] Abdulrhman Moshantaf, Talos 2021, https://gitlab.fhi.mpg.de/fhi-ac/talos.

[61] O. Bruns, T. Tietz, E. Posthumus, J. Waitelonis, H. Sack, NFDIcore Ontology 2024, https://nfdi.fiz-karlsruhe.de/ontology/2.0.0.

[62] H. Sack, T. Tietz, S. Bruhns, E. Posthumus, NFDI4Culture Ontology 2022, https://nfdi4culture.de/ontology.

[63] A. Azocar Guzman, A. Z. Ihsan, S. Fathalla, A. Gedsun, V. Hofmann, E. Norouzi, A. Laadhar, F. Fritzen, H. Sack, S. Sandfeld, MatWerk Ontology (MWO) 2024, http://matportal.org/ontologies/MWO (Accessed on March 19, 2024).

[64] T. Tietz, O. Bruns, L. Söhn, J. Tolksdorf, E. Posthumus, J. J. Steller, H. Fliegl, et al., in 3rd Workshop on Metadata and Research (objects) Management for Linked Open Science (DaMaLOS), co-located with ESWC 2023.

List of Authors

N. Kockmann, TU Dortmund University, Germany
A. S. Behr, TU Dortmund University, Germany
H. Borgelt, TU Dortmund University, Germany
M. Dörr, Institute of Biochemistry, University Greifswald, Germany
D. Linke, Leibniz-Institute for Catalysis, Germany
N. G. Moustakas, Leibniz-Institute for Catalysis, Germany
M. Khatamirad, BasCat, UniCat BASF JointLab, Technical University of Berlin, Germany
S. A. Schunk, hte GmbH, Germany
S. Hanf, Karlsruhe Institute of Technology, Germany
E. Norouzi, Karlsruhe Institute of Technology, Germany
E. Saraci, Karlsruhe Institute of Technology, Germany
M. Heßelmann, CVT, RWTH Aachen University, Germany
S. Zimmer, CVT, RWTH Aachen University, Germany
F. Wiesner, CVT, RWTH Aachen University, Germany
M. Wessling, CVT, RWTH Aachen University, Germany
T. Petrenko, HLRS, Stuttgart University, Germany
N. Huskova, HLRS, Stuttgart University, Germany
Y. Dikova, HLRS, Stuttgart University, Germany
R. Khare, TU Munich, Germany
A.Trunschke, Fritz Haber Institut, Germany
J. Schumann, Humboldt Universität zu Berlin, Germany
S. Angeli, Institute of Catalysis Research and Technology, Karlsruhe Institute of Technology.
H. Gossler, Institute for Chemical Technology and Polymer Chemistry, Karlsruhe Institute of Technology.
O. Deutschmann, Institute for Chemical Technology and Polymer Chemistry, Karlsruhe Institute of Technology.
R. Lenz, FAU Erlangen-Nürnberg, Germany

DECHEMA
Gesellschaft für Chemische Technik
und Biotechnologie e.V.

nfdi4cat.org

linkedin.com/company/nfdi4cat

twitter.com/nfdi4cat

youtube.com/@nfdi4cat

github.com/nfdi4cat

KONTAKT

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

Dr. Sara Espinoza
Email: sara.espinoza@dechema.de

DOI: 10.5281/zenodo.10932849