# A Unified Research Data Infrastructure for Catalysis Research – Challenges and Concepts

Christoph Wulf,[a] Matthias Beller,[a] Thomas Boenisch,[b] Olaf Deutschmann,[c] Schirin Hanf,[d] Norbert Kockmann,[e] Ralph Kraehnert,[f] Mehtap Oezaslan,[g] Stefan Palkovits,[h] Sonja Schimmler,[i] Stephan A. Schunk,[j, k] Kurt Wagemann,[l] and David Linke*[a]

Modern research methods produce large amounts of scientifically valuable data. Tools to process and analyze such data have advanced rapidly. Yet, access to large amounts of high-quality data remains limited in many fields, including catalysis research. Implementing the concept of FAIR data (Findable, Accessible, Interoperable, Reusable) in the catalysis community would improve this situation dramatically. The German NFDI initiative (National Research Data Infrastructure) aims to create a unique research data infrastructure covering all scientific disciplines. One of the consortia, NFDI4Cat, proposes a concept that serves all aspects and fields of catalysis research. We present a perspective on the challenging path ahead. Starting out from the current state, research needs are identified. A vision for a integrating all research data along the catalysis value chain, from molecule to chemical process, is developed. Respective core development topics are discussed, including ontologies, metadata, required infrastructure, IP, and the embedding into research community. This Concept paper aims to inspire not only researchers in the catalysis field, but to spark similar efforts also in other disciplines and on an international level.

## 1. Introduction

Catalysis is a key technology field for solving the challenges related to climate change and a sustainable supply of energy and materials. To tackle the challenges in reasonable time, improving the efficiency of developing new catalytic processes is of great value. Catalysis is highly interdisciplinary in its breadth of fields covering heterogeneous, homogeneous, bio-, electro- or photo-catalysis. All sub-disciplines share some common characteristics. Progress is driven by both experimental and computational methods which are often carried out in isolation by different specialists. Another aspect is that catalysis covers broad length and time scales. While ideal conditions can be often realized on small scale, this is no longer possible at larger scale. It is therefore vital to consider reaction and process engineering aspects in the early state of catalyst development. Due to the tight link between catalyst performance and optimal

[a] C. Wulf, Prof. M. Beller, Dr. D. Linke
Leibniz-Institute for Catalysis Rostock
Albert-Einstein-Str. 29a
D-18059 Rostock (Germany)
E-mail: david.linke@catalyis.de
[b] Dr.-Ing. T. Boenisch
High Performance Computing Center Stuttgart (HLRS)
University of Stuttgart
Nobelstr. 19
D-70569 Stuttgart (Germany)
[c] Prof. O. Deutschmann
Karlsruher Institut für Technologie (KIT)
Kaiserstraße 12
D-76131 Karlsruhe (Germany)
[d] Dr. S. Hanf
Karlsruher Institut für Technologie (KIT)
Engesserstr. 15
D-76131 Karlsruhe (Germany)
[e] Prof. N. Kockmann
Biochemical and Chemical Engineering, Equipment Design
TU Dortmund University
D-44221 Dortmund (Germany)
[f] Dr.-Ing. R. Kraehnert
BasCat – UniCat BASF JointLab
Technische Universität Berlin
Hardenbergstraße 36
D-10623 Berlin (Germany)
[g] Prof. M. Oezaslan
Institute of Technical Chemistry
TU Braunschweig
D-38106 Braunschweig (Germany)
[h] Dr. S. Palkovits
Institute for Technical and Macromolecular Chemistry
RWTH Aachen University
Worringerweg 2
D-52074 Aachen (Germany)
[i] Dr. S. Schimmler
Fraunhofer Institute for Open Communication Systems (FOKUS)
Kaiserin-Augusta-Allee 31
D-10589 Berlin (Germany)
[j] Dr. S. A. Schunk
the high throughput experimentation company
Kurpfalzring 104
D-69123 Heidelberg (Germany)
[k] Dr. S. A. Schunk
BASF SE
Carl-Bosch Str. 38
D-67056 Ludwigshafen (Germany)
[l] Prof. K. Wagemann
DECHEMA e.V.
Theodor-Heuss-Allee 25
D-60486 Frankfurt (Germany)

This publication is part of a Special Collection on "Data Science in Catalysis". Please check the ChemCatChem homepage for more articles in the collection.

process design, innovations may result from both the catalyst and the related process technologies.

According to a recent GeCats whitepaper a key to improve the general understanding and the development workflows in catalysis is building a bridge between theory and experiments.[1] This covers the challenge of understanding which material properties determine catalyst performance also described as the quest for the "catalyst genome".[2] One cornerstone of addressing this challenge is to boost the available amount of material, adsorption and reaction data via high throughput computation and to apply machine learning to gather further insights and to make predicting new materials more efficient.[3] However, this approach still suffers from a so called materials gap, that is that application in industry requires other data than what is currently stored in the available materials data platforms.[4] The lack of such data is also reflected in a recent mini-review on open data in catalysis which concludes that "small data" were neglected so far but are important in the big picture. With the term "small data" the authors refer to experimental data on the catalytic action i.e. kinetic data, which are believed to enable new insights at active site and mechanism levels when coupled with knowledge extraction tools.[5]

Another challenge is to develop the catalyst taking into account chemical engineering constraints, that is, to integrate catalyst and process development workflows.[1] Both challenges require better interdisciplinary collaboration between mathematical and theoretical sciences as well as experimental chemistry, chemical engineering and materials science.

Up to today, research data are hardly ever disseminated in the catalysis disciplines.[5] Although computers have become ubiquitous and are perfectly connected, research data are often not computer readable, not transferable between labs and therefore rarely re-used. Conventional means to transport research results, such as textual publications or verbal communication still dominate. Compared to other disciplines like astronomy,[6] oceanography[7] or climate research[8] sharing of data is hardly established except for some sub-disciplines such as computational material science[4] or crystallography.[9]

There are several factors that contribute to the current state. Most important, catalysis suffers from its complexity as a discipline that bridges chemistry, material science, chemical engineering, and physics. In order to make data widely useful, rather advanced, and well-coordinated approaches are needed that are beyond what a single group or institution can develop and sustain. Moreover, work in the catalysis lab often involves manual steps e.g., for catalyst preparation that are difficult and cumbersome to record in a digital format. Lab work often implies one-off setups which also change often or use tools that typically do not record data (heater, stirrer, oven). This complicates digital recording of experiments further.

While solutions exist to collect lab work in digital form in electronic lab notebooks (ELNs),[10] this is not standard in academic research labs where work and people change often, and short-living setups are used. Moreover, ELN are often tailored to local environments and exchanging data between or with ELNs is hampered by a lack of standardization. Consequently, such locally deployed ELNs have not stimulated a culture of sharing data. This may change with recent developments like the Chemotion ELN that provides a standardized interface for sharing data.[11]

The catalysis discipline suffers from this lack of data and tools, e.g.: Experiments are repeated unnecessarily. New results are not compared to existing and not put into an overall context. Information contained in the data is not extracted fully. Micro-kinetic analysis of reaction data is rarely performed. Data science developments cannot be applied to their full potential. Reproducibility and quality checks are hampered due to individual procedures and setups which are not described sufficiently in publications.[12] This slows down progress in catalysis but on the other hand opens a great opportunity to improve.[3b]

We propose applying the principles of digitalization to catalysis to enable efficient data-driven interdisciplinary development of catalysts and catalytic processes. Key requirements are (i) the use of open and well-defined data formats and (ii) using sufficient metadata to provide sufficient information on the context of the data. The latter is challenging but essential so that e.g., data from a theoretician can be reused by experimentalists, data from a chemist's lab experiment can be reused by chemical engineers or data from large experimental and computational series can be analyzed by machine-learning experts.

The above shortcomings, which also exist in other scientific communities, have motivated the German government to initiate a 10-year long cross-disciplinary initiative to coordinate research data management and stimulate data sharing and re-use in research, called NFDI (Nationale Forschungsdaten-infrastruktur).[13] The first consortia for funding were selected in June 2020. This includes an initiative from the catalysis discipline NFDI4Cat (NFDI for Catalysis-Related Sciences) which we report upon here.

NFDI4Cat has formed in a bottom-up approach base on community interests and needs.[1] NFDI4Cat addresses the needs of the catalysis community and seeks to enable the exchange of data following FAIR principles (FAIR = Findable, Accessible, Interoperable, Reusable).[14] In addition to IT specialists NFDI4Cat comprises of partners from all catalysis sub-disciplines and from chemical engineering to foster a common coordinated approach. This integrated approach is essential to realize the envisioned cross-disciplinary (re-)use of data (Figure 1).[15]

The essentials of the NFDI4Cat approach are based on four core principles:

- Open and Sustainable Data

A large part of research data created today is still produced for momentary and local use. NFDI4Cat seeks to foster a more open and sustainable approach to data where data can be found, understood, and re-used by other researchers.

- Cloudification

Currently, most data are hidden behind institutional boundaries. To maximize re-use and to enable collaboration across institutions, bringing/integrating local data to the cloud,

i.e., making them findable and re-usable on a global scale is a key goal for NFDI4Cat.

- Information Transparency

The FAIR principles will be followed. All standards and conventions related to data, metadata or interfaces will be shared with the community and community feedback will be integrated. The measures to verify data quality will be always transparent for users.

- Community Acceptance

The most important long-term goal is high acceptance in the community. NFDI4Cat will therefore provide training and tools to ease the production of sustainable data. Moreover, reward models will be developed to motivate sharing of data. In this context the protection of intellectual property and confidentiality are major challenges, which have to be tackled as part of the initiative. Hereby, the needs of academic institutes and the chemical industry require a differentiated contemplation in terms of the data sharing decision, competitiveness, and reward models.

Withing NFDI two more consortia related to aspects of catalysis have been starting along with NFDI4Cat: NFDI4Chem[16] which deals with research data management in chemistry, and NFDI4Ing,[17] which serves engineering sciences. All consortia work closely together to realize the vision of NFDI.

In this contribution we start with examples of research data management from a few subdisciplines and present examples for innovative data re-use. We start with an example from computational catalysis which is representing the most advanced sub-discipline in catalysis regarding data handling. As an example, for linking catalysis with chemical engineering we present a tool that supports the researcher in the development of kinetic and/or reactor models. It integrates management of models, simulation and experimental data and visual model assessment and offers a web-based user interface. Third we show how publishing a research data set, in the selected example one with historical data from the oxidative coupling of methane, can stimulate creative data science work to gain additional insights and to identify paths to improved catalysts.

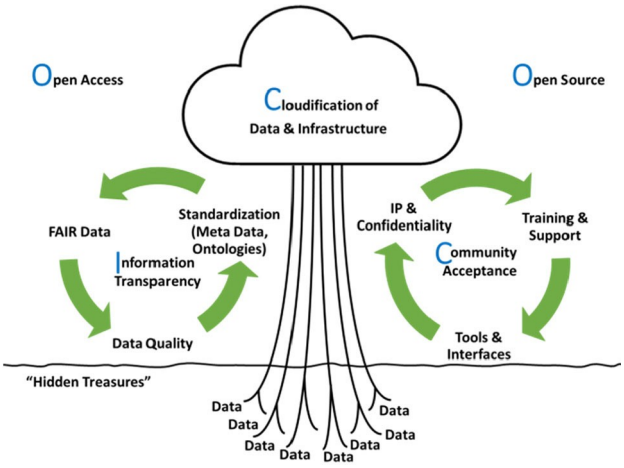


**Figure 1.** Illustration of the core principles and elements of NFDI4Cat.

Last, we present a current data management solution from industry representing state-of-the-art as an inspiration for similar solutions in academia. We conclude our example section with a critical view on the shortcomings of these solutions and identify the remaining key challenges. Finally, we present the approach of NFDI4Cat to these challenges in detail.

## 2. Examples of Existing Activities

### 2.1. Nomad – Uniting Interfaces in Computational Material Science

The management of data for individuals and in organizations plays a substantial role in an environment where communication will in large parts rely on transfer of information in the form of data. Therefore, overarching systems in which data can be stored, accessed and collaborative scientific work is fostered are of major importance for a digital catalysis community. Front runners in the field of catalysis following this approach in the context of the FAIR principles[14] are a range of initiatives driven by the community of scientists active in the field of theoretical chemistry and modelling. Many of the initiatives have been funded by public agencies in Europe and the US; an example of how politics can positively influence a culture of data sharing and progress in the field is for sure the initiation of the European Open Science Cloud[18] which is strongly supported by the national initiatives in the European Union.[19] It has to be noted that the computational scientist community in the field of catalysis have meanwhile advanced the field with respect to a proper storage of their respective data. Several databases exist to store and access especially the results of DFT calculation on solid materials.[20] The Materials Genome Initiative (MGI)[20e] is the oldest of these publicly funded projects and has in a lot of aspects acted as role model.[21] Central target of all of the efforts of MGI has always been the enhancement of the development speed of new materials and fostering of a paradigm shift in the community via digitalization. The two repositories in the MGI with largest relevance to catalysis are AFLOW (automatic flow for materials discovery) and the Materials Project.[20d]

One has to keep in mind that there is only a limited variety of DFT codes for solids available and most of the time especially the input files of the respective codes are interchangeable if suitable converters are at hand. This is the core of the project NOMAD (Novel Materials Discovery Laboratory) which is the host of the world's largest repository for input and output files of computational materials sciences codes.[22] Funding of NOMAD was provided on a basis of an EU-project under the CORDIS (Community Research and Development Information Service) framework.[23] Among the major achievements of NOMAD are the development of routine parsers which allow storing of input and output files leading to a reproducible workflow in which information like the surface/molecule geometries are retained, and version control tools (git, subversion) are used. NOMAD as most other DFT databases are searchable by a programming interface (API) making it possible to re-use/re-purpose the data in other fields of application to seek

correlations with tools from the field of artificial intelligence. One important strategy that NOMAD has since its start followed is to also host data which are publicly available in alternate databases and to convert these calculations, which are generally only available in different computer codes into a common, code-independent format. Following this strategy NOMAD hosts at present several million high-quality calculations. At the core of the mission, NOMAD programmers have developed parsers which automatically convert data sets available from open-access databases and archive the calculations in the code-independent format in the respective NOMAD archives. NOMAD is currently pursuing the following workstreams: (i) The NOMAD Encyclopedia, (ii) The NOMAD Big-Data Analytics Toolkit, (iii) a workstream for visualization tools, and (iv) High-Performance Computing Expertise and Hardware, available for purposes of the NOMAD project. Recent examples of work of NOMAD researchers in the field of catalysis include work on carbon dioxide conversion to fuels and chemicals[24] and work on structure property relationships over vast datasets.[25] Unfortunately, up-to date, similar data bases, tools and interfaces do not exist yet for experimental catalysis research – a gap which NFDI4Cat seeks to close. At a later stage, all these databases will be linked together to generate more insights.

### 2.2. CaRMeN – A Tool for Rapid Analysis and Development of Kinetic Models

The development process from finding new catalytic materials to their technological use is still a slow process. One tedious task is handling the evolving reaction engineering models along with the updated in experimental data. The recently developed software tool CaRMeN (CAtalytic Reaction MEchanisms Network) addresses the challenge of handling experimental information, model assumptions, model parameters, and equations including all metadata for the area of kinetics and reactor simulation in catalysis.[26]

The tool is designed for the rapid analysis of physical and chemical models against experimental data. It integrates tools to archive and package various forms of data along with simulation codes under a coomon graphical user interface. It improves the manual workflow of testing various models against experimental data by automating time-consuming and error-prone tasks such as setting up numerical simulations and post-processing the resulting data. Within the user interface, experimental data can be conveniently compared with the results of any simulation code under the matching experimental conditions in a plug-and-play fashion (Figure 2).[26]

CaRMeN can also be used to assess the quality of physical models such as transport models for porous media and different flow models (laminar/plug flow). False measurements in experimental data can be recognized more easily. Critical computer software issues resulting from wrongly implemented or inadequately used sub models become more obvious even for users not-so familiar with computing. CaRMeN has also been used in the areas of homogeneous gas-phase reactions

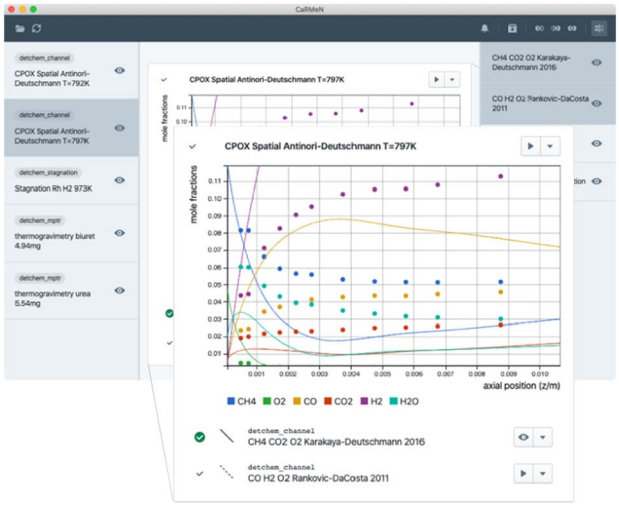


**Figure 2.** Screenshot of a comparison of experimentally measured and numerically predicted axial profiles of syngas production over Rh catalysts in a tubular flow reactor.

(combustion, pyrolysis, engines), chemical and steel industry as well as fuel and electrolysis cells.[26a,27] Hence, the tool serves as link between kinetics, reactor engineering and process engineering and can be easily extended to work with any simulation code. Extensions of the toolbox to establish direct links to DFT data and catalyst characterization data from microscopy and spectroscopy would be highly desirable.

In the CaRMeN toolbox, all raw data are accompanied by metadata of the experimental measurements as well as the processing chain of these data and associated results. The metadata is needed to generate input files for the numerical simulation of the reactors, in which kinetic data have been measured. Drivers use these metadata to combine the experiment with the specific reactor/process simulation software (e.g. CFD simulation)[28] and to set-up the input files for the numerical simulation. For instance, information of catalyst material and loading, porosity of the support structure, volumetric flow rates, temperature profiles, inlet mixtures etc. are automatically linked to the models. The user can directly access these metadata from the user interface to retrieve specific information needed. The format of the original and metadata is rather flexible; new formats, types of reactors and processes just require a specific driver, which can be written by the user. In combination with an accessible, intuitive user interface and a comprehensive search function, this approach achieves a high level of reusability. Several levels of IP rights on data and models are supported reaching from full open-access academic research and teaching to completely non-disclosed commercial use on customers' servers. Within NFDI, these approaches will be leveraged for broader application to provide new insights through their combination.

### 2.3. Meta-Analysis – Progress through Re-Use of Data

A vast amount of data is available in experimental catalysis research but hardly usable for digital processing. For some reactions, such as the oxidative coupling of methane, the data from several thousand literature reports were compiled and made available in digital form.[29] This shared data set has stipulated various groups to explore the application of data science methods to gain further insights, such as principle-component analysis,[30] artificial neural networks[31] and other machine learning tools.[32] Applying such methods proves to be promising, but faces numerous challenges including a large heterogeneity in the way the catalysts were synthesized, tested, and reported.

Meta-analysis is a powerful statistical tool to aggregate individual studies and estimate effects across heterogeneous data sets. Applied to heterogeneous catalysis, it can identify chemically meaningful and statistically significant correlations between physicochemical catalyst properties and their performance in a particular reaction.[33] The method combines physicochemical properties inferred from catalyst composition and well-known elemental reference data to formulate a working hypothesis that divides the dataset into subsets. Differences in the catalytic performance between these subsets are then tested for statistical significance against the pooled literature data. An iterative hypothesis refinement yields a statistical model that represent probable property-performance relationships. Figure 3 illustrates exemplarily how the method is used to structure the data into meaningful subsets.[33] The method was applied to the most comprehensive data sets of OCM data.[29] In the final model four simple hypotheses suffice to sort 1802 complex multi-component catalysts into 10 groups of distinct OCM performance.[33] Catalyst properties identified to be relevant are the ability of the contained elements to form (1) carbonates and (2) thermally stable oxides, (3) the carbonate's thermal stability under the respective experimental conditions, and (4) the properties (1-3) in combination with the respective amount of oxides and / or carbonate.

The results imply general correlations between a material's physicochemical properties and its OCM performance. Good catalysts comprise at least two elements, with one element being able to form a thermodynamically stable carbonate at the temperatures of OCM reaction, and a second element forming a thermally stable (non-sintering) oxide under OCM conditions. Hence, good catalysts apparently require a support that provides a high surface area at OCM temperatures, and carbonate(s) that either contribute directly to C2 formation and/or prevent subsequent unselective oxidation of the C2 products. The results directly guide dedicated experiments to understand the specific role of CO2 and carbonates in OCM, i.e., operando Raman under OCM conditions, experiments that relate the thermal stability of a series of supported carbonates and their OCM performance, as well as DFT to understand carbonate properties.

The derived correlations and interpretations can serve as a general guide to the design of new experiments, spectroscopic studies, and quantum chemical calculations. However, creating

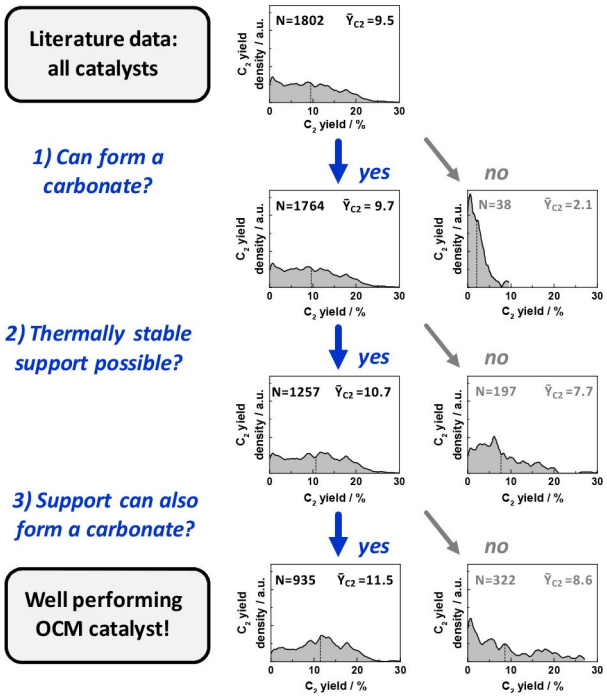


**Figure 3.** Illustration of the method output of the meta-analysis applied to OCM data. A dataset of 1802 catalysts is divided into subsets using three simple physicochemical criteria. The respective graphs report for each subset the number of catalysts, the average C2 yield in OCM as well as the resulting C2-yield density distribution. The full model and respective data are available in the paper of Schmack and co-workers.

such models would immensely benefit from the availability and accessibility of sets of data that contain large numbers of experimental results, are measured with consistent experimental procedures and well documented with the respective metadata.[34]

### 2.4. myHTE – Data Warehouse and Information Hub

The steady increase of the amount of data generated in modern laboratory environments and the subsequent storage over long periods of time, creates significant challenges in terms of the data management. Up to now, in many organizations fragmented data storage approaches are followed resulting from a lack of data governance. This bears significant disadvantages in terms of data consistency and administration. In order to enhance the overall data accessibility, consistency and short- and long-term value, reduce the data administration costs and enable smarter decision making, an integrated data approach is a vital foundation (Figure 4).[35] The main part of integrated data management approach is the central data warehouse, which connects all data storage infrastructures (hardware and cloud) for the user to provide all necessary information for data analysis and decision making. This integrated data warehouse should be administered centrally to control the process of data acquisition, management and distribution efficiently.[36] Based on this integrated data management philosophy, hte GmbH[37]

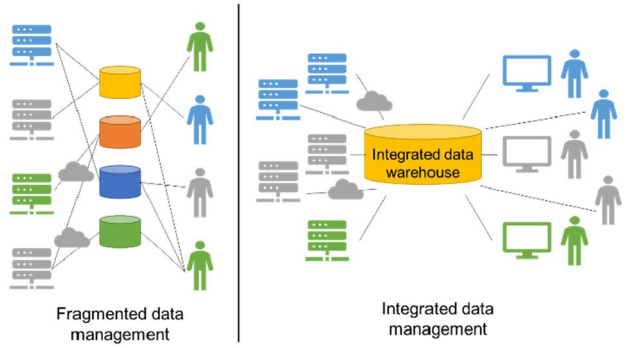


**Figure 4.** Fragmented and integrated data management strategies.

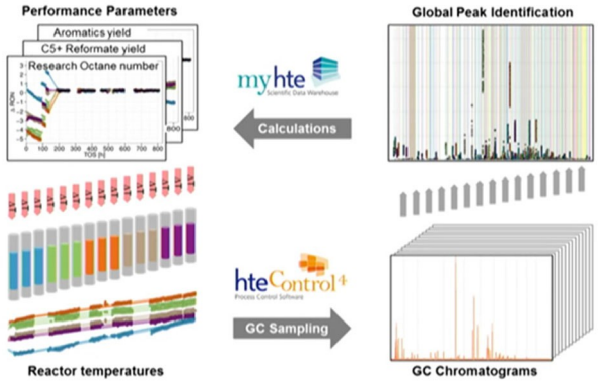


**Figure 5.** Symbiosis of hteControl™ and myhte™ in the iso-RON operation.

developed two software platforms, namely hteControl™ and myhte™, for the data collection and analyzes in the context of process catalysis related applications and high throughput experimentation.[35] hteControl™ is an advanced process control system, which allows control of experimental parameters, fully automated experiment execution and subsequent reliable data acquisition. With this process control system, parameters can be controlled and adjusted using flexible experimental sequencing in a graphical flow diagram editor. Moreover, it is possible to gain access to fast system diagnostic in a 24/7 operation. The data sets acquired can subsequently be stored and managed in the integrated data warehouse (myhte™). This data management software integrates, stores, analyzes data, and allows visualization.

It is possible to analyze large amounts of online and offline analytical data in relation to process parameters and experimental details, such as data related to catalyst synthesis, catalyst characterization and details of reactor loading. Therefore, a robust automatic quality control can be ensured through programming of automated routines. An example for this, is the automated evaluation of on- and offline analytic results from gas chromatography which includes peak assignment and automatic quantifications.

Through the interaction of the process control system and the integrated data warehouse, new modes for running experiment becomes possible, e.g., the so-called iso-run modes. In these iso-run modes, complex product features are selected as response factors, which will change dynamically over time due to an alteration of the catalyst characteristics. The objective is to keep the response factor constant via an automatic adjustment of the process parameters. This dynamic back-coupling of the response factor and the process parameters can be achieved via an automated analysis of the experimental results (Figure 5).[38] For this self-optimization process the integrated data management is crucial. It furthermore lays the basis for further data mining, statistical evaluation of the experimental data and kinetic studies.

The before described tooling serves as an example for a "high end" industrial solution for data generation and data management. Such solutions will also be of major importance to a broader research community since the fundamental challenges in obtaining and storing good data are essentially the same in an academic lab.

To sum up, the examples mentioned above demonstrate that there are already very promising approaches to manage, use and re-use research data in catalysis. These approaches are, however, still addressing only specific aspects in the respective discipline. Moreover, the data stores are rather isolated silos without much interlinking or cross-tool functionality. For example, CaRMeN cannot directly use DFT data from NOMAD and experimental data from myHTE. While the problem of linking data from different data stores is not new and has led to the invention of the semantic web,[39] the available standards and technologies for inter linking data, have (if at all) only been rudimentary applied. Only, recently the application of the full semantic web stack has been suggested.[40]

Providing user-friendly access to data science tools along with the data, is another challenge. CADS which aims to provide a multi-functional environment for assisting researchers in designing catalysts using catalyst informatics is an endeavor in this direction.[41]

Besides the above challenges, it is even more important to increase the amount of shared data which is currently very low in catalysis. Therefore, thinking the bigger picture is needed. In the next sections we propose an overall concept to address this and solve some of the mentioned problems.

## 3. Vision

Central to our concept for sustainable research data management in catalysis are FAIR digital objects.[19b,42] A FAIR digital object is a stable actionable unit that bundles sufficient information to enable reliable interpretation and processing of the data contained in it. It is composed of the data itself and accompanying information that provides context to the data, including persistent identifiers and metadata. Persistent identifiers are world-wide unique identifiers that allow reliably finding and citing such data objects. Metadata is "data about data"[43] that describe the context of the data. The quality of metadata determines the reusability of the digital object. It is obvious

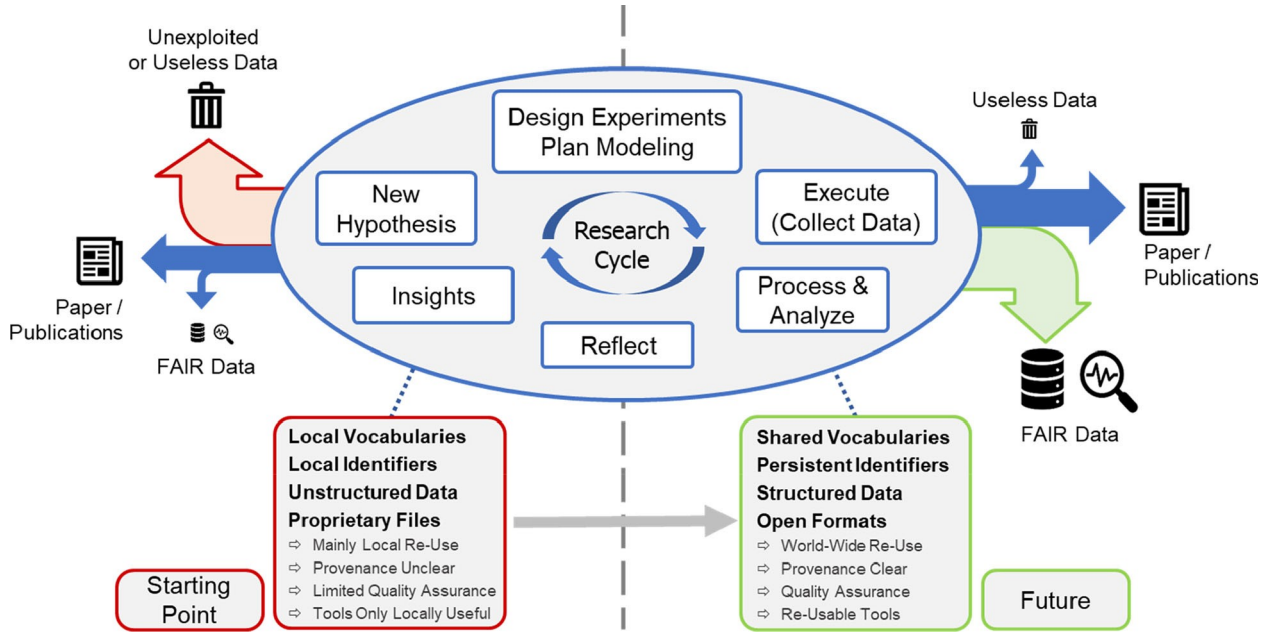


**Figure 6.** Problematic elements of current data production in the research cycle and elements for boosting the production of FAIR data in the future.

that a discipline should agree on and use standardized metadata schemes and vocabularies. Such "agreement" can be encoded in form of shared ontologies.[44] Moreover, both the data and the accompanying metadata should be re-presented in common open data formats to make them accessible and re-useable.[45] This idea of digital objects extends beyond pure data. Source code or other research outputs can and should be handled applying the same principles, too. Figure 6 shows the key elements that NFDI4Cat seeks to change in research data production. The core change is that any ambiguity related to data will be avoided from the beginning. This has tremendous benefits: All data will only be present in FAIR form. Thus, sharing is inherently possible. This replaces the data annotation in hindsight which is time consuming and adds little value for the researcher itself. Moreover, tools for ingesting such FAIR data will be re-usable by others. This will stimulate joint development of tools leading to better quality and less work for the individual researcher.

In order to enable the re-use of FAIR digital object along the complete catalysis value chain from molecules to chemical processes (Figure 7),[15] the development of metadata schemes and vocabularies should be coordinated over all catalysis sub-disciplines and related disciplines like chemical engineering.

By integrating feedback loops at every stage of the displayed stages of the data value chain the information and knowledge gained can have valuable influence in further experiments. An iterative design-of-experiments is envisioned to be an integral part of the workflow of data-driven catalysis research. Part of this approach will be the building of quantitative models to predict other regions of interest and highest potential information gain. Respective models will be modular and will be based on statistics, machine learning, theoretical calculation as well as combinations thereof.[3b] Since

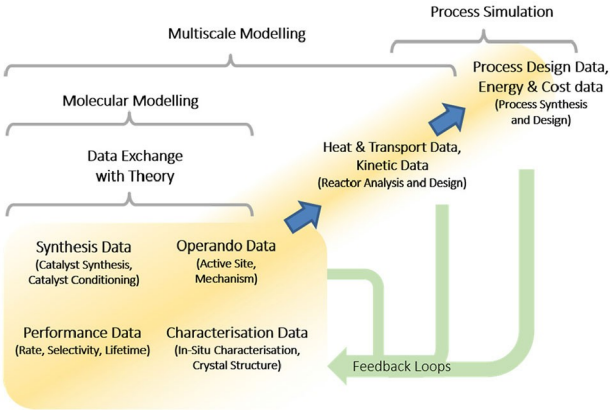


**Figure 7.** Data value chain for catalysis sciences.

hardly any of the involved sub-disciplines produces FAIR digital objects right now, there is an open window for NFDI4Cat to elaborate these and address the needs of the various sub-communities together in order to establish universally usable metadata schemes and vocabularies. Digital catalysis objects using standardized catalysis metadata will form the backbone of the digital catalysis value chain. In such a digital value chain more efficient feed-back loops are possible because data exchange and re-use is tremendously improved. The development of new processes or the adaptation of improvements to existing processes will be fostered by enabling interdisciplinarity between mathematical and theoretical sciences and experimental chemistry, chemical engineering, and materials science. FAIR digital catalysis objects will boost data-driven approaches in catalysis research.

The solution that NFDI4Cat plans for managing the digital catalysis objects is a hierarchical system with local data repositories and an overarching infrastructure for bringing (selected) local data to the cloud (see Figure 10). The local parts of the system allow keeping data private and can be tuned to the need of the sub-community that the user works in. The overarching infrastructure will index all local digital catalysis objects marked for sharing and will bring these local data to the cloud as FAIR digital catalysis objects. Moreover, the overarching infrastructure will provide a unified view on data across the catalysis disciplines to facilitate inter- and cross-disciplinary data re-use.

## 4. Core Development Topics in NFDI4Cat

### 4.1. Data Collection

In catalysis labs experiment data are stored and generated on different levels of complexity and aggregation

- raw data directly obtained from instruments or software programs during an experiment,
- processed and aggregated synthesis, property, and performance data,
- metadata that describes experimental procedures, conditions, and setups,
- metadata that describes the data processing.

For every step, data and metadata are generated and have to be processed and stored. Many measurements also alter the catalyst material; hence a history of the treatment of a catalyst is often essential for profound understanding of its properties and behavior.

While the overall workflow and fundamental concepts are similar in heterogeneous, homogeneous, electro and bio catalysis, each of the disciplines uses slightly different approaches, different nomenclatures, experimental methods as well as property and performance descriptors.

In heterogeneous catalysis research data are produced in a sequence of steps. In a typical workflow, catalysts are synthesized (often from molecular compounds called precursors) and subsequently treated (calcined, reduced, pressed, sieved…) in order to produce a solid material suited for performance testing. For catalytic tests, the materials are mounted in a reactor and then exposed to the reactants (gases, liquids) in a sequence of reaction conditions (temperature, pressure, flow rates…). The effluent product streams are then analyzed with respect to formed products and their quantity using e.g., GC, MS, or other analysis methods. The obtained data is processed to calculate or estimate aggregated numbers as a measure of catalyst performance (conversion, yield, rate, activation energy etc.). These numbers serve as an input for kinetic modelling and reactor simulations. Based on such simulations catalytic reactors and processes can be designed. To understand the respective catalytic materials better, their physicochemical properties (composition, structure, spectroscopic information…) are assessed experimentally or via quantum chemical calculations (bulk and surface structure, adsorption sites, transition states, energy barriers etc.).

A hierarchical scheme can be derived to organize such data according to the respective abstraction level. However, each of the experimental steps can modify the catalyst material and its properties. Thus, implementing a timeline or "biography" for each catalyst will be one of the crucial aspects for success. Further challenges include data collected in proprietary formats, a lack of standardized nomenclature and ontology. Furthermore, these also enclose a lack of open software tools and repositories, ways of linking publications, data, and potentially other digital objects consistently and permanently as well as paths to retrieve published data for re-use. Catalysis-specific ontologies and metadata standards will be critical in making the data accessible and retrievable.

### 4.2. Ontologies and Metadata

One of the pressing questions of the research data handling is, how can the context of data and ultimately knowledge be shared within and outside of a community? A core role in the solution play ontologies. An ontology is an explicit, formal specifications of a shared conceptualization. By using ontologies defined in a machine-readable language like OWL the concepts behind data can be represented. The formal conceptualization determines which additional information, i.e., which conceptual data are required to provide context to data.

In the last decades, various disciplines have been developing ontologies and metadata standards for using, sharing, and annotating information between domain experts. In chemistry some well-established ontologies exist like IUPAC's International Chemical Identifier (InChI)[46] for describing chemicals or the Crystallographic Information Framework (CIF).[47] for describing crystals. For other parts of chemistry ontologies are still subject to current research, e.g. for chemical reactions.[48]

In process engineering the development of ontologies and data standards has a long tradition, particularly driven by the process system engineering activities.[49] Data standards and data exchange are very important in automation and control of chemical plants. These activities include the transfer of data from modelling to actual representation of a plant state and its influence on the control strategy (model-predictive control). In process engineering, data exchange is important in the development of chemical processes, from early process design, laboratory experiments, and equipment design to plant construction and commissioning.[50] These aspects are partially treated in the DEXPI initiative for the German chemical industry,[51] in DIN/ISO15926, or CFIHOS[52] activities in the oil and gas industry. The most elaborated ontology in process engineering is probably OntoCAPE developed at RWTH Aachen.[53]

OntoCAPE seeks to cover the description from molecules to the whole plant. Figure 8 gives an impression of the ontologies in OntoCAPE.[54] An example for the representation of a molecule and its properties as a pure substance is given in Figure 9.[56] Catalyst representation in OntoCAPE includes mainly

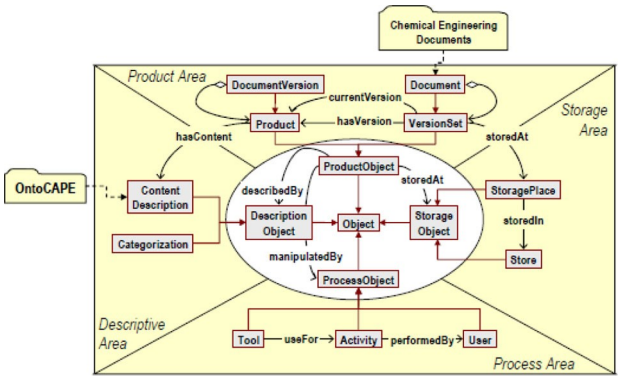


**Figure 8.** Simplified view of the OntoCAPE Core Ontology and some Peripheral Ontologies.[54]

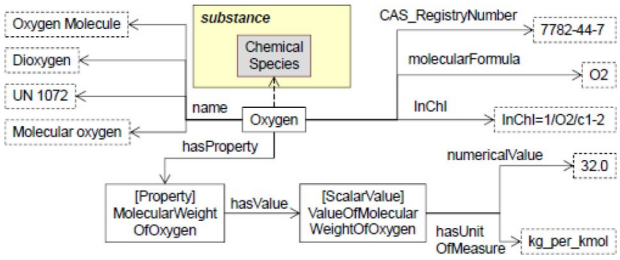


**Figure 9.** Representation of Oxygen in OntoCAPE with the physical properties of molecular weight, triple point temperature and pressure, and critical properties of temperature, pressure, and molar volume.[56]

the cost aspect of the precious metals which are often used. The multitude of typical reactors applied for catalytic reactions, particularly for the different phases and their contact mechanism as well as the heat integration is not entirely covered. Although the OntoCAPE ontology is elaborated and ready to use, only few applications have been known to date.[55] Some of the succeeding activities are bound to the DEXPI standardization activities in the planning process of chemical plants.

An ontology covering all aspects of the catalysis data value chain from Figure 7 does not exist. However, ontologies covering various parts are already available and provide a foundation for NFDI4Cat to build upon.

While the ontologies organize metadata, guidelines have to be developed which metadata needs to be supplied with the data. Presently, there are few guidelines available, e.g., from STRENDA (Standards for Reporting Enzymology Data)[57] or ESAB (European Federation of Biotechnology Section of Applied Biocatalysis)[58] that cover enzymology and biocatalysis data.

Whenever possible metadata should be added automatically without user involvement for consistency and to achieve a low error rate. One of the successful models of automated metadata descriptions have recently been achieved by EngMeta at High Performance Computing Center Stuttgart (HRLS) and Stuttgart University Library.[59] EngMeta is developed for the use case of computational engineering and enables the documentation of the entire research process in terms of descriptive, technical, process and domain-specific metadata. The most powerful tool implemented in EngMeta is the automatic extraction to collect metadata from different sources. Metadata for laboratory processes involving manual steps come ideally from ELNs. The development of interoperable ELNs that provide semantically rich data will be a focus in the NFDI4Chem consortium.[16a] NFDI4Cat will cooperate with NFDI4Chem on ELNs but does not plan to develop a separate ELN system on its own.

### 4.3. Local and Overarching Data Infrastructures

One main goal of NFDI4Cat is to set up and establish local and overarching data infrastructures. This includes a distributed repository infrastructure and other services that are needed by the NFDI4Cat community, in order to put forward a national environment for catalysis-related research data.

One challenge is to identify and serve the real needs of the NFDI4Cat community. Therefore, we will involve different stakeholders in the whole process, including a requirements analysis and user acceptance tests. Another challenge is to avoid fragmentation and data silos. Therefore, we will proceed with a coordinated approach. Existing solutions will be integrated, where reasonable, and new solutions will be pushed ahead, where necessary.

To put forward an overarching data infrastructure, a layered architecture is planned, which includes a distributed storage layer, a repository layer, and a presentation layer, see Figure 10. The distributed storage layer enables the local storage at different sites. The repository layer will provide one new general repository at HLRS and new repositories at sites with special requirements. It will also integrate already existing repository systems. For instance, data that is under intellectual property regulations can be stored safely, without being published. The presentation layer will provide a general access point to the (meta)data that is openly available in the different repositories and will offer other services that were identified of being useful for the NFDI4Cat community.

To put forward local data infrastructures, pilots will be set up in labs working in different catalysis disciplines. These data systems will be locally administered. The local researcher and/or institution decides about access rights and what to share. The

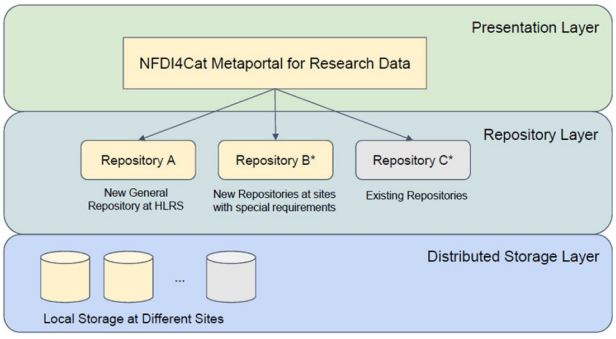


**Figure 10.** Linked extensible infrastructure for NFDI4Cat.

idea is to enable using the same system for open as well as for confidential research data. The aim in the beginning is to experiment in real-world scenarios, gain experience in the daily use, and identify challenges as early as possible. Here, the whole research data lifecycle – collect/create, process, analyze, preserve, access and reuse – will be considered. Other groups will benefit from these pilots, either by reusing some of the services established, or by learning from the setup of the pilots. Long-term goal of this effort is to include these services in a general toolbox. To ensure future viability, we will build on existing standards and principles e.g. use established vocabularies such as schema.org[60] or W3C DCAT,[61] and will synchronize with other consortia and other communities. We will favor open source solutions, will rely on modern technologies, and will develop in the spirit of Semantic Web[39] and Linked Open Data.[45,62]

One tool that is planned being used to set up local and overarching data infrastructures is Piveau,[63] a fully-fledged Open Data management solution, based on Semantic Web technologies. It forms, for instance, the technical foundation of the European Data Portal,[3c,64] a central access point for metadata of Open Data published by public authorities in Europe that acquires data from more than 70 national data providers.

### 4.4. Data Analysis and Quality Management

Data-driven catalytic science aims to identify relationships between the different data of the described workflow. However, the mentioned parameters along the whole workflow are highly interconnected, and all measured and computed values are subject to errors and error propagation. High quality data and known error margins are therefore essential to enable reliable modelling and correlation analysis. Thus, quality assurance should be an integral part of catalysis research.

In order to assure high-quality data, two main aspects have to be addressed – reliable and reproducible measurements, and equally important, the quality of documentation.[65] Common experimental pitfalls can be overcome by including in the design-of-experiment tests for catalyst stability, the assessment of mass and heat transport limitation, the calculation of mass balances and error estimation via repeated measurements at different levels (repeated analytical runs, repeated testing, repeated synthesis,…).[66] Moreover, standardized reference catalysts and common benchmarking procedures that assess catalyst performance and stability could become an integral part of the research workflow.[12a] Excellent examples from the field of electro-catalysis can be found for the hydrogen[67] and oxygen[68] evolution reaction.

The other essential aspect is the documentation of each step and parameter in a catalysts life.[34] Such documentation should be in a digital form, use open and standardized formats, be highly automatized and – most important – community accepted. This requires not only a change in research culture, but also the respective technological tools and organizational measures. These tools should facilitate quality assurance along the whole workflow of catalysis research, including experiment planning, synthesis, testing, data processing, visualization, evaluation, and modelling. Easy to use tools and low entry barriers will be key to a wide-spread adoption. Moreover, educating catalysis researchers in quality assurance via easy access to examples, tutorial, standard procedures, and reference materials will be vital.

### 4.5. IP & Confidentiality, Licenses & Reward Models

The sharing of data for the benefit of the scientific community and science in general is one of the central cornerstones of the NFDI and current movements within the scientific community. However, although the values of data sharing are self-evident, these values must be balanced with the interests of individuals and groups who intend to exploit the value of data generated within publicly funded projects of any kind. A work package in NFDI4Cat addresses the sensitive points around data sharing procedures and the resulting consequences and tries to find a balance through an open dialogue between academia and industry; from the viewpoint of NFDI4Cat a very differentiated contemplation and approach is required. The interests of all stakeholders involved need to be balanced: the views and needs of academic research groups and industrial companies might differ substantially and an approach based on modus of consensus must be found.

One of the key publications in the context of this discussion are the Horizon 2020 guidelines for "Open access and Data management".[69] The European Union with their research and innovation program is for sure one of the pacemakers in the context of data-sharing policies. The Horizon 2020 guidelines are fully aligned with the FAIR principles, which are, at present, the most concise summary of guiding principles in open data-sharing, emphasizing that data should be treated to be findable, accessible, interoperable, and reusable.[14] The purpose of the FAIR data governance strategy is to maximize the use and therefore the value of research data. In context of Horizon 2020, the European Commission has also launched the European Open Science Cloud (EOSC) to foster exchange of scientific data, data handling and processing and services around data processing.[18] This service is part of the Horizon 2020 program and builds on a series of demonstrator projects and accompanied by changes in regulation around EU's General Data Protection Regulation. Although open access is the default setting for Horizon 2020 and therefore within the NFDI and NFDI4Cat, it has to be acknowledged that not all data can be open. According to the current state of discussion in the European Commission, an approach is suggested that follows the view of an "as open as possible, as closed as necessary" policy; open access is therefore not required if the following facts apply:[70]

- The participation is incompatible with the obligation to protect results that can reasonably be expected to be commercially or industrially exploited.
- The participation is incompatible with the need for confidentiality in connection with security issues.

- The participation is incompatible with rules on protecting personal data.
- The participation would mean that the project's main aim might not be achieved.
- The project will not generate/collect any research data.
- There are other legitimate reasons.

From an industrial viewpoint, the obligation to protect certain data to remain competitive is obvious. According to the SusChem, a European technology platform for sustainable chemistry, industrial competitiveness in domestic (EU) and global markets is crucial to maintain an economic growth, especially for small and medium-sized companies.[71] It has to be acknowledged that academia is traditionally also not less competitive than industry, and that advantages through data realized in knowledge and know-how guarantee access to grants and collaborators, participation in excellence initiatives, as well as to excellent students. Specifically, for academia a balancing of sharing data via reward models in the context of a competitive research and grant application environment must be considered. In this context, it is important to avoid negative effects which outweigh the potential gains of a competitive research environment.[72] Independent from the industrial or academic environment, a competitive framework where the best ideas compete for funding and attention is still a dominant cultural paradigm for innovation policies with knowledge and data being the most precious goods.[73]

We are therefore entering an age where data increase in value almost in the same way in academia as well as in industry for a number of reasons, therefore one of the major objectives of NFDI4Cat is to create a culture of data-sharing where the motivation and incentives to contribute catalysis data must be fostered.

Publishing of data alongside with interpretation and explanations is state of the art in academia, therefore in principle data-sharing should not stand in contrast with goals of NFDI. It is vital to establish, as above said, new reward strategies,[74] which for example are based on the number of citable data sets published, preferably also in combination with annotations to data quality. An evident reward model could be the allocation of a digital object identifier (DOI) number, through which each deposited dataset will be a citable source of data. By associating the digital objects with their authors via persistent author identifiers like ORCID, credit can be given to data providers and in analogue way to tool providers. It can be envisaged that researchers can build their reputation in a more diverse way in the future. Citable "digital object publications" will become a new element for esteem in science and will motivate sharing of data and tools in a FAIR way.

However, it must be considered that such next-generation metrics are in theory susceptible to very similar difficulties as traditional and often quantitative measurements, such as the journals impact factor.[75,74c] Therefore, a qualitative assessment of data, based on expert judgement, should be implemented to further develop policies for rewarding open data sharing. Rewards for open science activities could be granted in the form of promotions. In addition, data sharing activities could be explicitly used as criteria in recruiting processes or funding applications. Apart from the direct rewards, the deposition of experimental and theoretical data in a digital format will lay the foundation for future collaborations and could be the starting point for the development of new business models dealing with data handling and data analyses. Obviously, an open data research management should be considered as state of the art in the future. However, this change in data handling and the not self-serving data sharing culture has to be embraced by the community. Therefore, NFDI4Cat aims to promote the open data policy as final reward strategy with the aim of bringing science to a next level in a digital format.

In this context, one of the NFDI4Cat's major interests is to develop practical measures, which ensure confidentiality, allow for measures for securing intellectual property and a high data quality, without the FAIR principles being passed over. These guidelines for industrial and academic research groups are summarized in Figure 11 and are based on a so-called "cool-off model", which could help to classify data according their critical or uncritical status and lay the foundations for a sensible process in a culture of data-sharing. The distinction of uncritical data can be made based on the "opting-out" factors given by the European Commission.[70] If data is worth to protect, it must be decided whether the results will be patented or whether the information is kept and protected internally as trade secrets without any procedural formalities.

### 4.6. Integrating the Community

Beyond technical challenges also a change in research culture and RDM literacy is important. Therefore, it will be important to educate not only a new generation of scientists and engineers towards an improved data awareness but also to provide knowledge for the catalysis community and related organizations and disciplines. Collaborations (e.g. NFDI4Chem,[16a]

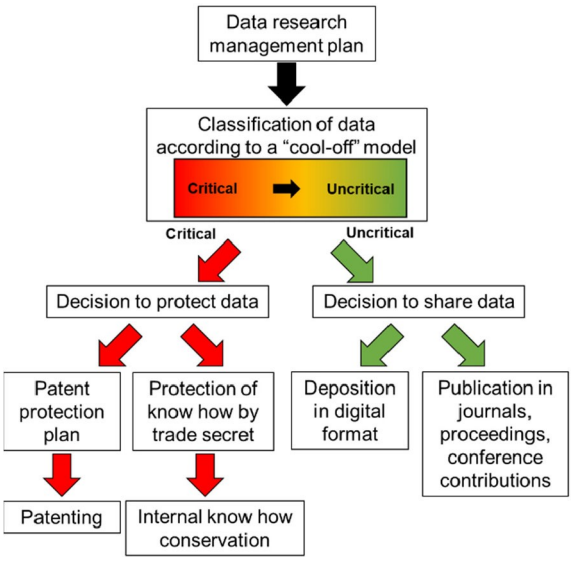


**Figure 11.** Data management for the decision process using a "cool-off" model.

NFDI4Ing,[17] IUPAC[76]) are therefore an important part of the outreach within NFDI4Cat. NFDI4Cat will take several measures to improve the Data Science education in Catalysis related sciences.

Establishing feedback loops to gather the information from the community is always important to establish the connection between the developers (the NFDI4Cat consortium) and the final users (all users in Catalysis related sciences). Therefore, NFDI4Cat will use measures at different scales to establish a stable feedback loop with the community and towards the proposed best practices. The actions will reach from simple surveys and public relations up to the organization of an annual NFDI4Cat conference with the help of DECHEMA as organizing organization. But NFDI4Cat is not only aiming at dissemination of the respective outcome on own national conferences but the consortium will also organize sessions at international conferences to establish and foster the collaboration with international stakeholders in Catalysis.

One very important measure will be the Research Data Management School of Catalysis. The aim of the Research School is to make the community and new generations of scientist more aware how data should be stored to be FAIR. Therefore, it is important that the participants get a feeling which data is important for a reproducible study and how to keep the data not only for themselves but how it should be made available for the community as a whole.

The Research Data Management School of Catalysis will be split into several parts including modules about

- Data quality and open formats,
- Data acquisition,
- Data storage and
- Publication of the respective data for a study.

Teaching Research Data Management and the related tools, skills and techniques will gain much in importance in the future. As a possible blueprint the "Data 8: The Foundations of Data Science" course of UC Berkeley can be used.[77] The course spans from basic skills to Machine Learning and covers most of the aspects needed to work with research data.

Data intensive studies show that one important skill for future researchers will be the evaluation of the increasing amount of data. This often goes well beyond the possibilities of tools like MS Excel or Origin. Therefore teaching Research Data Management will be also about teaching new tools like programming and evaluating data in programming languages like Python, Julia or R, combining Machine Learning libraries with Web techniques like JavaScript or including final algorithms in languages like Go or C. Teaching Research Data Management also means showing the next generation of researchers how to work with version control (especially git) or cloud-based computations as clearly many computational studies move away from computation on a single workstation. Therefore at least some awareness of concepts like containerization and related techniques are valuable.

The consortium plans to publish the outcome of the initiative as Best Practice Guides compiling the important outcome of the initiative how NFDI4Cat recommends working with data generated around theoretical and experimental work in Catalysis. To get started with the best practice concepts, access to data generated by NFDI4Cat will be provided. This should enable users to dive into Research Data Management without own data but by a blueprint already available. Apart from these dissemination spotlights, NFDI4Cat will actively contribute to the distribution of modern tools and techniques for Research Data Management in all its aspects for the whole Catalysis community.

## 5. Outlook

Within the German NFDI initiative the consortium NFDI4Cat embarks on the endeavor of realizing a data-oriented "digital catalysis value chain" supporting research along the development chain from molecules to chemical processes. Core motivation is a fundamentally improved understanding in catalysis sciences, the creation of workflows in catalysis that build a bridge between theory/simulation and experimental studies in design, characterization and kinetics of catalysts and the related engineering aspects. This challenge requires a unified view on all catalysis disciplines to reveal universal guiding principles common to homogenous, bio-, heterogeneous and electro-catalysis. By integrating stakeholders from all catalysis sub-disciplines in Germany, NFDI4Cat is in a unique position to realize this vision in the years ahead and inspire similar efforts on an international level and in other disciplines.

The initial focus will be on enabling the German catalysis community to exchange data following FAIR principles. To make data (re-)usable and enable collaboration across organizations and between (sub-) disciplines on a data level, catalysis specific new open standards or extensions of existing standards for storing data and the metadata are urgently needed. NFDI4Cat will work on ontologies, metadata and data standards and finally build prototypes that are built upon this foundation. All standardization efforts will be coordinated on international level. From the current point of view, it is also important to emphasize that the time scale, until a full implementation and the final goal of a fully digitalized scene in catalysis can be reached, is expected to be on the order of a decade. It is anticipated that ultimately the information architecture will become an indispensable tool of the research community in catalysis on a national and international basis.

## *Acknowledgements*

*NFDI4Cat is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with the project number 670389-NFDI 2/1. Open access funding enabled and organized by Projekt DEAL.*

## *Conflict of Interest*

The author Stephan Andreas Schunk is Vice President and Executive Expert at BASF SE and hte GmbH. hte GmbH

commercializes and distributes the software packages hteControl™ and myhte™. All other authors have declared no conflict of interest.

**Keywords:** GeCats · NFDI · NFDI4Cat · Catalysis Community · Digitalization · Research Data Management

[1] "GeCatS Whitepaper – The Digitalization of Catalysis-Related Sciences", can be found under http://gecats.org/gecats media/Downloads/GeCatS Whitepaper + 2019 engl ezl.pdf, Accessed on 10.12.2020.
[2] J. K. Nørskov, T. Bligaard, *Angew. Chem. Int. Ed.* **2013**, *52*, 776–777; *Angew. Chem.* **2013**, *125*, 806–807.
[3] a) C. Bo, F. Maseras, N. López, *Nat. Catal.* **2018**, *1*, 809–810; b) A. J. Medford, M. R. Kunz, S. M. Ewing, T. Borders, R. Fushimi, *ACS Catal.* **2018**, *8*, 7403–7429; c) K. Takahashi, L. Takahashi, I. Miyazato, J. Fujima, Y. Tanaka, T. Uno, H. Satoh, K. Ohno, M. Nishida, K. Hirai, J. Ohyama, T. N. Nguyen, S. Nishimura, T. Taniike, *ChemCatChem* **2019**, *11*, 1146–1152.
[4] L. Himanen, A. Geurts, A. S. Foster, P. Rinke, *Adv. Sci.* **2019**, *6*, 1900808.
[5] P. S. F. Mendes, S. Siradze, L. Pirro, J. W. Thybaut, *ChemCatChem*, DOI: 10.1002/cctc.202001132.
[6] A. Zuiderwijk, H. Spiers, *Int. J. Inf. Manage.* **2019**, *49*, 228–241.
[7] a) "BCO-DMO – Biological & Chemical Oceanography Data Management Office", can be found under https://www.bco-dmo.org/, Accessed on 01.02.2021; b) "PANGAEA – Data Publisher for Earth & Environmental Science", can be found under https://www.pangaea.de/, Accessed on 01.02.2021.
[8] a) "CEDA – Centre for Environmental Data Analysis", can be found under https://www.ceda.ac.uk/, Accessed on 01.02.2021; b) "ESA climate office", can be found under https://climate.esa.int/en/, Accessed on 01.02.2021; c) "KNMI Climate Explorer", can be found under https://climexp.knmi.nl/start.cgi, Accessed on 01.02.2021.
[9] I. Bruno, *Chem. Int.* **2017**, *39*, 41–42.
[10] a) "Electronic Lab Notebooks", can be found under https://datamanagement.hms.harvard.edu/analyze/electronic-lab-notebooks, Accessed on 01.02.2021; b) S. Kanza, C. Willoughby, N. Gibbins, R. Whitby, J. G. Frey, J. Erjavec, K. Zupančič, M. Hren, K. Kovač, *J. Cheminf.* **2017**, *9*, 31.
[11] P. Tremouilhac, C.-L. Lin, P.-C. Huang, Y.-C. Huang, A. Nguyen, N. Jung, F. Bach, R. Ulrich, B. Neumair, A. Streit, S. Bräse, *Angew. Chem. Int. Ed.* **2020**, *59*, 22771–22778; *Angew. Chem.* **2020**, *132*, 22960–22968.
[12] a) T. Bligaard, R. M. Bullock, C. T. Campbell, J. G. Chen, B. C. Gates, R. J. Gorte, C. W. Jones, W. D. Jones, J. R. Kitchin, S. L. Scott, *ACS Catal.* **2016**, *6*, 2590–2602; b) F. Schüth, M. D. Ward, J. M. Buriak, *Chem. Mater.* **2018**, *30*, 3599–3600; c) *Nat. Catal.* **2018**, *1*, 229; d) *Nat. Catal.* **2020**, *3*, 471–472; e) 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*, 1683–1699.
[13] "Nationale Forschungsdateninfrastruktur (NFDI) e. V.", can be found under https://www.nfdi.de/, Accessed on 01.02.2021.
[14] M. D. Wilkinson, M. Dumontier, I. 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, *Sci. Data* **2016**, *3*, 160018.
[15] "NFDI for Catalysis-Related Sciences – NFDI4Cat", can be found under http://www.nfdi4cat.org/, Accessed on 10.12.2020.
[16] a) "NFDI4Chem – Chemistry Consortium in the NFDI", can be found under https://nfdi4chem.de, Accessed on 10.12.2020; b) C. Steinbeck, O. Koepler, F. Bach, S. Herres-Pawlis, N. Jung, J. C. Liermann, S. Neumann, M. Razum, C. Baldauf, F. Biedermann, T. W. Bocklitz, F. Boehm, F. Broda, P. Czodrowski, T. Engel, M. G. Hicks, S. M. Kast, C. Kettner, W. Koch, G. Lanza, A. Link, R. A. Mata, W. E. Nagel, A. Porzel, N. Schlörer, T. Schulze, H.-G. Weinig, W. Wenzel, L. A. Wessjohann, S. Wulle, *Res. Ideas Outcomes* **2020**, *6*, e55852.
[17] "NFDI4Ing – the National Research Data Infrastructure for Engineering Sciences", can be found under https://nfdi4ing.de/, Accessed on 10.12.2020.
[18] a) "EUROPEAN OPEN SCIENCE CLOUD", can be found under https://www.eosc-portal.eu/, Accessed on 10.12.2020; b) "European Open Science Cloud (EOSC)", can be found under https://ec.europa.eu/research/openscience/index.cfm?pg = open-science-cloud, Accessed on 10.12.2020.
[19] a) B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, M. D. Wilkinson, *Inf. Serv. Use* **2017**, *37*, 49–56; b) S. Hodson, S. Jones, S. Collins, F. Genova, N. Harrower, L. Laaksonen, D. Mietchen, R. Petrauskaité, P. Wittenburg, "Turning FAIR data into reality: interim report from the European Commission Expert Group on FAIR data", can be found under https://doi.org/10.5281/zenodo.1285272, Accessed on 01.02.2021, **2018**.
[20] a) "COMPUTATIONAL MATERIALS REPOSITORY – CMR", can be found under https://cmr.fysik.dtu.dk/, Accessed on 10.12.2020; b) "Catalysis Hub", can be found under https://www.catalysis-hub.org/, Accessed on 10.12.2020; c) "Aflow – Automatic – FLOW for Materials Discovery", can be found under http://www.aflowlib.org/, Accessed on 10.12.2020; d) "The Materials Project", can be found under https://www.materialsproject.org/, Accessed on 10.12.2020; e) "Materials Genome Initiative", can be found under https://www.mgi.gov/, Accessed on 10.12.2020; f) K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich, T. Bligaard, *Sci. Data* **2019**, *6*, 75.
[21] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.
[22] "NOMAD – Centre of Excellence", can be found under https://nomad-coe.eu/, Accessed on 10.12.2020.
[23] "The Novel Materials Discovery Laboratory", can be found under https://cordis.europa.eu/project/id/676580, Accessed on 10.12.2020.
[24] S. Posada-Pérez, P. J. Ramírez, J. Evans, F. Viñes, P. Liu, F. Illas, J. A. Rodriguez, *J. Am. Chem. Soc.* **2016**, *138*, 8269–8278.
[25] B. R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler, L. M. Ghiringhelli, New *J. Phys.* **2017**, *19*, 013031.
[26] a) H. Gossler, L. Maier, S. Angeli, S. Tischer, O. Deutschmann, *Phys. Chem. Chem. Phys.* **2018**, *20*, 10857–10876; b) H. Gossler, L. Maier, S. Angeli, S. Tischer, O. Deutschmann, *Catalysts* **2019**, *9*, 227.
[27] K. Keller, P. Lott, H. Stotz, L. Maier, O. Deutschmann, *Catalysts* **2020**, *10*, 922.
[28] a) G. D. Wehinger, *Chem. Ing. Tech.* **2019**, *91*, 583–591; b) J. Mularski, N. Modliński, *Energies* **2020**, *13*, 6467.
[29] U. Zavyalova, M. Holena, R. Schlögl, M. Baerns, *ChemCatChem* **2011**, *3*, 1935–1947.
[30] E.-J. Ras, G. Rothenberg, *RSC Adv.* **2014**, *4*, 5963–5974.
[31] E. V. Kondratenko, M. Schlüter, M. Baerns, D. Linke, M. Holena, *Catal. Sci. Technol.* **2015**, *5*, 1668–1677.
[32] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, K.-i. Shimizu, *ACS Catal.* **2020**, *10*, 2260–2297.
[33] R. Schmack, A. Friedrich, E. V. Kondratenko, J. Polte, A. Werwatz, R. Kraehnert, *Nat. Commun.* **2019**, *10*, 441.
[34] J. R. Kitchin, *ACS Catal.* **2015**, *5*, 3894–3899.
[35] a) C. Futter, L. T. A. Rupflin, N. Brem, R. Födisch, A. Haas, A. L. d. Oliveira, M. L. Lejkowski, A. Müller, A. Sundermann, S. Titlbach, S. K. Weber, S. A. Schunk, High Throughput Experimentation Applied in the Field of Technical Catalysis: Past, Present, Future, in Modern Applications of High Throughput R&D in Heterogeneous Catalysis (Eds.: A. Hagemeyer, J. Anthony F. Volpe), Bentham Science Publishers, 2014; b) A. Gordillo, S. Titlbach, C. Futter, M. L. Lejkowski, E. Prasetyo, L. T. A. Rupflin, T. Emmert, S. A. Schunk, in Ullmann's Encyclopedia of Industrial Chemistry (7th Edition), Vol. 7 (Ed.: B. Elvers), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2014.
[36] a) D. L. Myers, K. C. Burke, J. D. Burke, K. S. Culp, *Managed Care Interface* **2000**, *13*, 68–72; b) T. Töpel, B. Kormeier, A. Klassen, R. Hofestädt, *J. Integr. Bioinf.* **2008**, *5*, 343–349.
[37] "hte Company", can be found under https://www.hte-company.com/, Accessed on 10.12.2020.
[38] M. Kirchmann, A. Haas, C. Hauber, S. Vukojevic, *PTQ* **2015**, *Q4*, 119–131.
[39] "Semantic Web", can be found under https://www.w3.org/standards/semanticweb/, Accessed on 01.02.2021.

[40] L. Takahashi, I. Miyazato, K. Takahashi, *J. Chem. Inf. Model.* **2018**, *58*, 17422–11754.
[41] J. Fujima, Y. Tanaka, I. Miyazato, L. Takahashi, K. Takahashi, *React. Chem. Eng.* **2020**, *5*, 903–911.
[42] a) P. Wittenburg, G. Strawn, B. Mons, L. Boninho, E. Schultes, "Digital Objects as Drivers towards Convergence in Data Infrastructures", can be found under https://doi.org/10.23728/b2share.-b605d85809ca45679b110719b6c6cb11, Accessed on 01.02.2021, 2019; b) K. De Smedt, D. Koureas, P. Wittenburg, *Publications* **2020**, *8*, 21.
[43] J. Riley, "Understanding Metadata: What is Metadata, and What is it For?: A Primer", can be found under http://www.niso.org/publications/understanding-metadata-2017, Accessed on 02.02.2021, 2017.
[44] "OWL 2 Web Ontology Language", can be found under https://www.w3.org/TR/owl2-overview/, Accessed on 01.02.2021.
[45] "Semantic Web – Linked Data", can be found under https://www.w3.org/standards/semanticweb/data, Accessed on 01.02.2021.
[46] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, *J. Cheminf.* **2015**, *7*, 23.
[47] S. R. Hall, B. McMahon, *Data Sci. J.* **2016**, *15*, 1–15.
[48] a) F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca, M. Kraft, *J. Chem. Inf. Model.* **2020**, *60*, 108–120; b) C. Pachl, N. Frank, J. Breitbart, S. Bräse, arXiv:2002.03842 2020.
[49] >W. Marquardt, J. Morbach, A. Wiesner, A. Yang, OntoCAPE: A Re-Usable Ontology for Chemical Process Engineering, Springer-Verlag, Berlin-Heidelberg, 2010.
[50] N. Kockmann, *React. Chem. Eng.* **2019**, *4*, 1522–1529.
[51] "DEXPI – Data Exchange in the Process Industry", can be found under https://dexpi.org/, Accessed on 10.12.2020.
[52] "CFIHOS - Capital Facilities Information HandOver Specification", can be found under https://uspi.nl/index.php/projects/frameworks-methodologies/cfihos-overview, Accessed on 10.12.2020.
[53] "OntoCape", can be found under https://www.avt.rwth-aachen.de/cms/AVT/Forschung/Software/~ipts/OntoCape, Accessed on 10.12.2020.
[54] A. Wiesner, J. Morbach, A. Yang, B. Bayer, W. Marquardt, Chemical Process Systems, Technical Report LPT-2008-29, 2008.
[55] A. Devanand, G. Karmakar, N. Krdzavac, R. Rigo-Mariani, Y. S. Foo Eddy, I. A. Karimi, M. Kraft, *Energy AI* **2020**, *1*, 100008.
[56] J. Morbach, A. Yang, W. Marquardt, Material, Technical Report LPT-2008-27, 2008.
[57] "STRENDA – Standards for Reporting Enzymology Data", can be found under https://www.beilstein-strenda-db.org/strenda/, Accessed on 10.12.2020.
[58] "ESAB – European Society of Applied Biocatalysis", can be found under https://esabweb.org/, Accessed on 10.12.2020.
[59] a) B. Schembera, D. Iglezakis, in Metadata and Semantic Research (Eds.: E. Garoufallou, F. Sartori, R. Siatri, M. Zervas), Springer International Publishing, Cham, 2019, pp. 127–132; b) B. S. D. Iglezakis, *Int. J. Metadata Semantics Ontologies* **2020**, *14*, 26–38.
[60] "Schema.org", can be found under https://schema.org/, Accessed on 01.02.2021.
[61] "Data Catalog Vocabulary (DCAT)", can be found under https://www.w3.org/TR/vocab-dcat-2, Accessed on 01.02.2021, **2020**.
[62] T. Berners-Lee, "Linked Data", can be found under https://www.w3.org/DesignIssues/LinkedData, Accessed on 05.01.2021, **2006**.
[63] "Piveau", can be found under https://www.piveau.de/en/, Accessed on 10.12.2020.
[64] F. Kirstein, B. Dittwald, S. Dutkowski, Y. Glikman, S. Schimmler, M. Hauswirth, Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP, in Electronic Government. EGOV 2019. Lecture Notes in Computer Science, Vol. 11685 (Eds.: I. Lindgren, M. Janssen, H. Lee, A. Polini, M. P. Rodríguez Bolívar, H. J. Scholl, E. Tambouris), Springer International Publishing, Cham, **2019**, pp. 192–204.
[65] J. R. Kitchin, *Nat. Catal.* **2018**, *1*, 230–232.
[66] "EUROKIN – Excellence in commercial reaction kinetics", can be found under https://eurokin.org/?cat=30, Accessed on 10.12.2020.
[67] C. C. L. McCrory, S. Jung, I. M. Ferrer, S. M. Chatman, J. C. Peters, T. F. Jaramillo, *J. Am. Chem. Soc.* **2015**, *137*, 4347–4357.
[68] S. Geiger, O. Kasian, A. M. Mingers, S. S. Nicley, K. Haenen, K. J. J. Mayrhofer, S. Cherevko, *ChemSusChem* **2017**, *10*, 4140–4143.
[69] "HORIZON 2020 – WORK PROGRAMME 2018–2020", can be found under https://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-intro en.pdf, Accessed on 10.12.2020.
[70] "HORIZON 2020 Online Manual – Open access & Data management ", can be found under https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access en.htm, Accessed on 10.12.2020.
[71] "SusChem 2016: Sustainable Chemistry – Innovation for Competitiveness", can be found under http://www.suschem.org/highlights/suschem-2016-sustainable-chemistry-innovation-for-competitiveness, Accessed on 10.12.2020.
[72] L. Carson, C. Bartneck, K. Voges, *Disruptive Sci. Technol.* **2013**, *1*, 183–190.
[73] P.-B. Joly, *J. Innovation Econ. Manage.* **2017**, *22*, 79–96.
[74] a) R. Benedictus, F. Miedema, M. W. J. Ferguson, *Nature* **2016**, *538*, 453; b) W. van den Akker, J. Spaapen, "Productive interactions: Societal impact of academic research in the knowledge society", can be found under https://www.leru.org/publications/productive-interactions-societal-impact-of-academic-research-in-the-knowledge-society, Accessed on 01.02.2021, **2017**; c) P. Ayris, A. L. d. S. Román, K. Maes, I. Labastida, "Open Science and its role in universities: a roadmap for cultural change", can be found under https://www.leru.org/publications/open-science-and-its-role-in-universities-a-roadmap-for-cultural-change, Accessed on 01.02.2021, **2018**.
[75] J. Wilsdon, L. Allen, E. Belfiore, P. Campbell, S. Curry, S. Hill, R. Jones, R. Kain, S. Kerridge, M. Thelwall, J. Tinkler, I. Viney, P. Wouters, B. J. J. Hill, The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management, SAGE Publications Ltd, London, **2015**.
[76] "DIGCHEM – A VISION FOR CHEMICAL DATA STANDARDS", can be found under https://iupac.org/digchem-a-vision-for-chemical-data-standards/, Accessed on 10.12.2020.
[77] "Data 8: The Foundations of Data Science", can be found under http://data8.org, Accessed on 10.12.2020.

Manuscript received: December 11, 2020
Revised manuscript received: March 3, 2021
Accepted manuscript online: March 10, 2021
Version of record online: ■■, ■■■

# CONCEPTS

Data Sharing: The German NFDI initiative (National Research Data Infrastructure) aims to create a unique research data infrastructure covering all scientific disciplines. One of the firstly funded consortia, NFDI4Cat (NFDI for Catalysis-related Sciences), proposes a concept that serves all aspects and fields of catalysis research.


*C. Wulf, Prof. M. Beller, Dr.-Ing. T. Boenisch, Prof. O. Deutschmann, Dr. S. Hanf, Prof. N. Kockmann, Dr.-Ing. R. Kraehnert, Prof. M. Oezaslan, Dr. S. Palkovits, Dr. S. Schimmler, Dr. S. A. Schunk, Prof. K. Wagemann, Dr. D. Linke**

1 – 15

**A Unified Research Data Infrastructure for Catalysis Research – Challenges and Concepts**