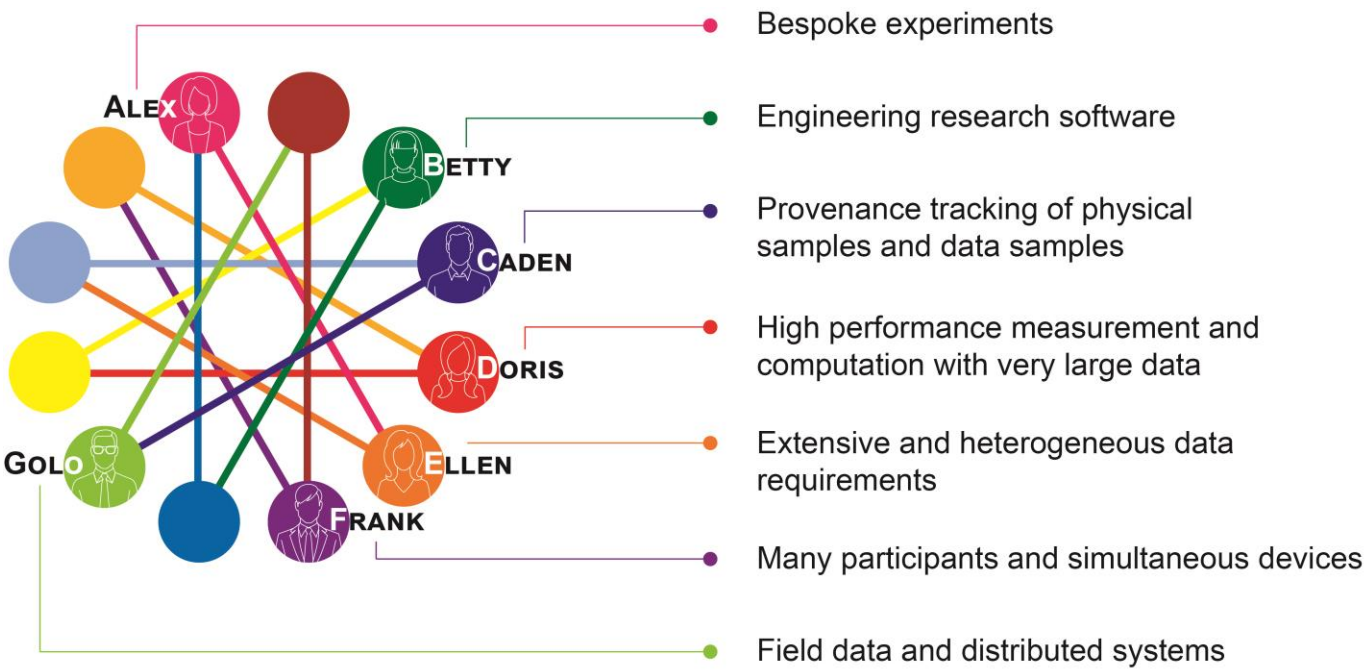
NFDI4Ing – the National Research Data Infrastructure for Engineering Sciences

Excerpt from the Funding Proposal submitted in October, 2019, to the German Research Foundation (DFG)



NFDI4ing

Table of contents

Summary

NFDI4Ing brings together the engineering communities and fosters the management of engineering research data. The consortium represents engineers from all walks of the profession. It offers a unique method-oriented and user-centred approach in order to make engineering research data FAIR – findable, accessible, interoperable, and re-usable.

NFDI4Ing has been founded in 2017. The consortium has actively engaged engineers across all five engineering research areas of the DFG classification. Leading figures have teamed up with experienced infrastructure providers. As one important step, NFDI4Ing has taken on the task of structuring the wealth of concrete needs in research data management. A broad consensus on typical methods and workflows in engineering research has been established: The archetypes. So far, seven archetypes are harmonising the methodological needs:

- ALEX: bespoke experiments with high variability of setups,
- BETTY: engineering research software,
- CADEN: provenance tracking of physical samples & data samples,
- DORIS: high performance measurement & computation,
- ELLEN: extensive and heterogeneous data requirements,
- FRANK: many participants & simultaneous devices,
- GOLO: field data & distributed systems.

A survey of the entire engineering research landscape in Germany confirms that the concept of engineering archetypes has been very well received. 95% of the research groups identify themselves with at least one of the NFDI4Ing archetypes.

NFDI4Ing plans to further coordinate its engagement along the gateways provided by the DFG classification of engineering research areas. Consequently, NFDI4Ing will support five community clusters. In addition, an overarching task area will provide seven base services to be accessed by both the community clusters and the archetype task areas. Base services address quality assurance & metrics, research software development, terminologies & metadata, repositories & storage, data security & sovereignty, training, and data & knowledge discovery. With the archetype approach, NFDI4Ing's work programme is modular and distinctly method-oriented. With the community clusters and base services, NFDI4Ing's work programme remains firmly user-centred and highly integrated.

NFDI4Ing has set in place an internal organisational structure that ensures viability, operational efficiency, and openness to new partners during the course of the consortium's development. NFDI4Ing's management team brings in the experience from two applicant institutions and from two years of actively engaging with the engineering communities. Eleven applicant institutions and over fifty participants have committed to carrying out NFDI4Ing's work programme. Moreover, NFDI4Ing's connectedness with consortia from nearby disciplinary fields is strong. Collaboration

on cross-cutting topics is well prepared and foreseen. As a result, NFDI4Ing is ready to join the National Research Data Infrastructure.

1 Consortium

1.1 Research domains or research methods addressed by the consortium, objectives

Engineering sciences play a key role in developing solutions for the technical, environmental, and economic challenges imposed by the demands of our modern society. The associated research processes as well as the solutions themselves will only be sustainable if being accompanied by a proper research data management (RDM) that implements the FAIR data principles. The proposed consortium NFDI4Ing will put engineers in research and development in the position to enhance their daily creative output with correspondingly required and useful RDM measures.

The purpose of engineering is manifold in developing methods to describe selected aspects of the world and also finding and providing means and artefacts to change it. Engineers (i) analyse technical, environmental, economic, and social systems (and interaction within). The creative core is the methodical design, (ii) the synthesis: Engineers use the knowledge gained through analysis to design technical functionalities. Setting the framework for engineering sciences, the technical synthesis of systems is initiated, firstly, by (iii) social needs. Secondly, the synthesis of technical systems methodically pursues the goal of designing technical functionalities (iv) in an efficient and sustainable way.

Research in engineering ranges from basic research (analysis), i.e. the production, use and processing of data, to applied research and technology development (synthesis). Especially for the latter, availability of data, including descriptive documentation and corresponding software, is crucial for success in this often highly interdisciplinary environment. As such, engineers face a vast variety and a wide heterogeneity of day to day tasks; they have a strong focus on software and must be able to provide and employ cross-discipline data, possibly at high speed and high volume. Consequently, the engineering community shows a high affinity and ability to develop and adapt IT systems, familiarity with standardisation (e.g. industry standards), know-how in quality management, and last but not least an established practice of systematic approaches in big projects with many stakeholders. However, the complexity of information in the focus of research (sub-)communities has resulted in highly specialised solution approaches regarding research data management and diverse engineering research profiles.

Methodically, engineering sciences rely on the understanding of fundamental research (theories) and their application on and validation in particular domains (experiments), enriched by additional information derived from domain-specific models (simulation). These sources provide a wide variety of types, objects, amount, and quality of data. Despite this diversity in research, engineers

share common needs and approaches on the methodological level. NFDI4Ing leverages those similarities in order to consolidate the key objectives and to design task areas that benefit the engineering community at large. NFDI4Ing has identified seven typical research profiles by the most commonly recurring needs, methods, and workflows classifying corresponding challenges for RDM in engineering sciences. We call these research profiles “archetypes” and use them throughout this proposal to structure our planned developments.

Task areas addressing the requirements of these archetypes are flanked by task areas that cross-link them with research-area-centred communities as well as cross-cutting topics. Together, they contribute to the following key objectives of NFDI4Ing from the perspective of engineers as well as infrastructure providers:

1. Scientists of all disciplines are able to retrace or reproduce all steps of engineering research processes. This ensures the trustworthiness of published results, prevents redundancies, and contributes to social acceptance.
2. Engineers are enabled to develop validated quality-assured engineering research software. They treat software as research data that possibly connects the different stages of stored data.
3. Recording and linking of auxiliary information and provenance is automated and optimised to reduce the manual data handling tasks as much as possible. This ensures the interpretability of data in the context of a specific project and for a hitherto unknown repurposing.
4. Sharing and integration of possibly large amounts of data is facilitated and employed by engineers across single studies, projects, institutions, or disciplines via networked technical infrastructure (repositories), open metadata standards, and cultural change.
5. Collaborative research is unhindered, while preventing unauthorised access to confidential data. Because engineering sciences are close to industry, this calls for sophisticated means in authentication, intellectual property, and license management.
6. Engineers are able to generate machine-processable representations of auxiliary information based on open standards by means of easily accessible tools. This paves the way to further reuse by data-driven analysis methods such as machine learning and artificial intelligence approaches.
7. Engineers profit from an improved data- and software-related education (data literacy) and available domain and application specific best practices.
8. Publication of data is standardised and acknowledged by the engineering community in the same way as publication of scientific documents, including peer-review measures and effects on the scientific reputation.

1.2 Composition of the consortium and its embedding in the community of interest

NFDI4Ing opted for a research method-oriented approach to meet the requirements of our community of interest: the engineering sciences (according to DFG classification of research areas: 41-45). In this heterogeneous community, the vast variety of engineering problems focused on specific research (sub-)areas has resulted in highly specialised, individualised solution tool approaches. These individual approaches can hardly be reused by engineering research groups with similar yet slightly different requirements due to their lacking modularity. Additionally, the tool approaches are often not sustainable and easily outdated from the point of IT progress or new scientific demands. In NFDI4Ing, we lay ground for systematic solutions of this challenges by following a new solution design: the archetype concept, which is described in the following section.

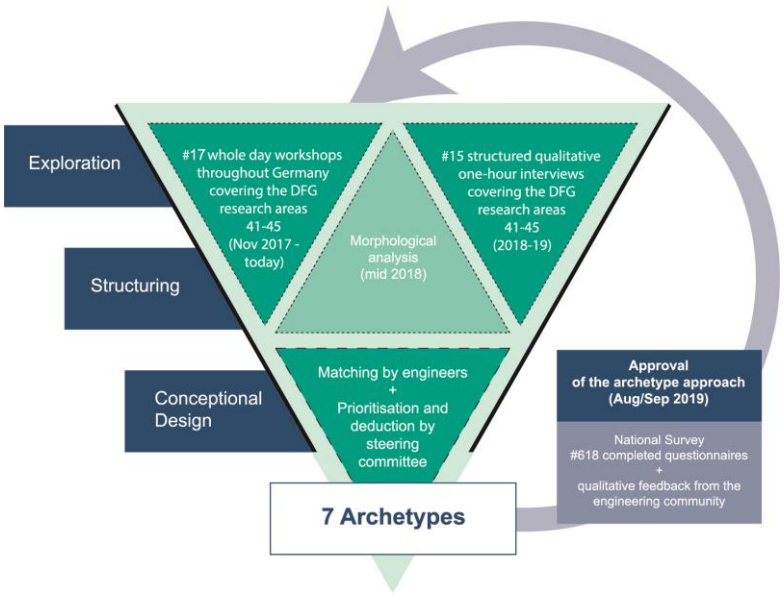


Figure 1.2.1 Exploration, structuring and conceptual design of the archetype system

For the identification of the specific needs of the engineering science community, we used a mixed-method approach including the collection of qualitative and quantitative data. Figure 1.2.1 provides an overview of our identification and prioritisation process.

The explorative phase had the objective to provide a broad overview of the current state and needs regarding RDM in engineering science: In 2018, we conducted 15 semi-standardised face-to-face interviews with representatives from the five DFG research areas of interest and have invited, since 2017, to 17 workshops with different foci each reaching from 12 to 50 attendees. Based on these means of communication and data collection, we synthesised 24 key dimensions for describing engineering science as morphological box, a heuristic problem solving method e.g. used for the development of product innovations. Then, this morphological box was used in the structuring process to categorise individual researchers according to their respective characteristics and to identify our most important target groups in terms of prototypical engineering scientists or methodological archetypes. In total, we derived seven archetypes being

representative for the majority of the heterogeneity in engineering sciences and derived method-oriented task areas from each of these archetypes:

- ALEX: bespoke experiments with high variability of setups,
- BETTY: engineering research software,
- CADEN: provenance tracking of physical samples & data samples,
- DORIS: high performance measurement & computation,
- ELLEN: extensive and heterogeneous data requirements,
- FRANK: many participants & simultaneous devices,
- GOLO: field data & distributed systems.

These archetypes define typical research methods and workflows classifying corresponding challenges for research data management. So, engineers will be reached via their identification with the methodological archetypes. NFDI4Ing has and will continue to put emphasis on the identification and harmonisation of engineering research archetypes.

Therefore, we are aiming to approve whether our community of interest feels represented by the deduced archetypes. To this end, in mid-2019, we conducted a first online survey study that allowed approaching and involving a significant number of engineering scientists. The survey consisted of two parts: (1) RDM-related questions and (2) evaluation of the archetypes. We used an extensive mailing list including engineering research associations and other potential disseminators to achieve a reasonable sample size. Thus, we reached engineering research groups at all German universities and universities of applied sciences, as well as at all non-university engineering research institutions (e.g. Fraunhofer institutes, Helmholtz centres, governmental research institutes, etc.). In total, 618 engineers completed the survey (each representing one research group from all fields of engineering), providing a solid basis for evaluating our approach. Analysis of the survey data [1] confirmed the representativeness and relevance of our seven archetypes (cf. Figure 1.2.2): 95% of all research groups identify themselves with at least one archetype. The typical engineering research group combines elements of three to four archetypes and considers on average two archetypes as very relevant, showing a good division of demands by the archetypes. These findings are mostly independent of the engineering sub-discipline. These conclusions are further supported by the aforementioned interviews and workshops giving valuable qualitative feedback regarding our method-oriented and user-centred approach.

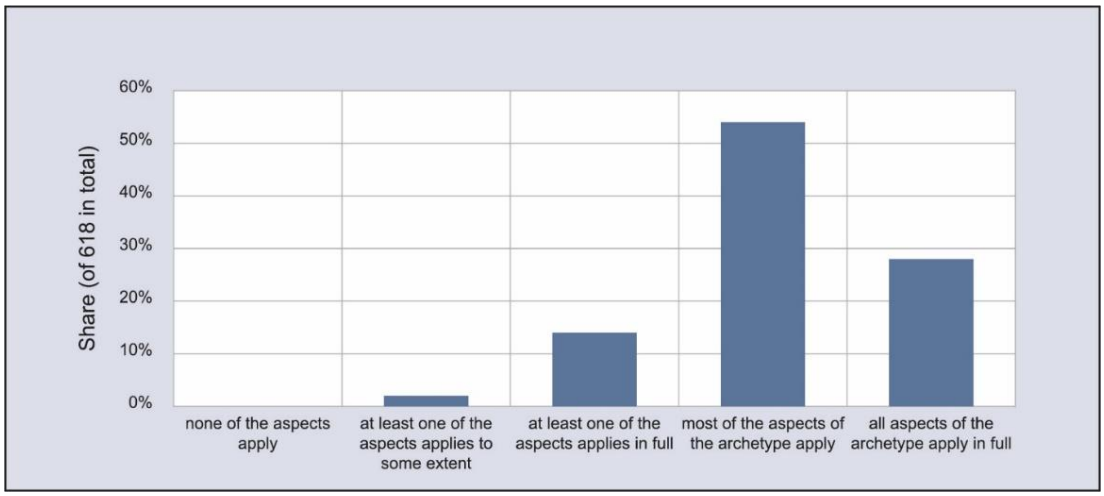


Figure 1.2.2 Survey result: Degree of identification of the respondent research groups with the archetypes

Archetypes do not only serve as a classification system helping to bundle resources, they can also act as a communication gateway and support an engineer to phrase his individual demands towards the infrastructures. To further consolidate the needs and solution approaches motivated via the different archetypes and to translate those into services that integrate well with the NFDI, NFDI4Ing dedicates the task area 3.9 Base Services to inter- and intra-consortial service topics regarding research data management. Those services within NFDI4Ing are offered by established RDM infrastructure providers across TU9 universities and well-known research associations such as Helmholtz, Leibniz, and Fraunhofer. All infrastructure providers stand out with their experience regarding engineering specific collaborations and RDM projects. Examples for such established RDM services are specified in chapter 2.2. The user-centred approach in the development of services safeguards that the interests of all stakeholders are well balanced.

Continued discussion, involvement, and engagement with the targeted research communities in the engineering sciences ensure that the community needs remain properly represented and addressed. Therefore, the task area 3.10 Community Clusters will engage in outreach activities and communicate the needs of the community to the consortium members on a regular basis. Taking account of the broad range of the engineering community, it is covering the DFG research areas from 41 to 45:

- 41 Mechanical and Industrial Engineering,
- 42 Thermal Engineering/Process Engineering,
- 43 Materials Science and Engineering,
- 44 Computer Science, Systems and Electrical Engineering,
- 45 Construction Engineering and Architecture.

Figure 1.2.3 Structure of the consortium's work programme illustrates the close interaction of the research method-oriented archetypes with both the community clusters and base services. Each of the archetype task areas bridges the five community clusters. The archetype task areas and

the community clusters are further supported by the task area Base Services focusing on general service topics in research data management.

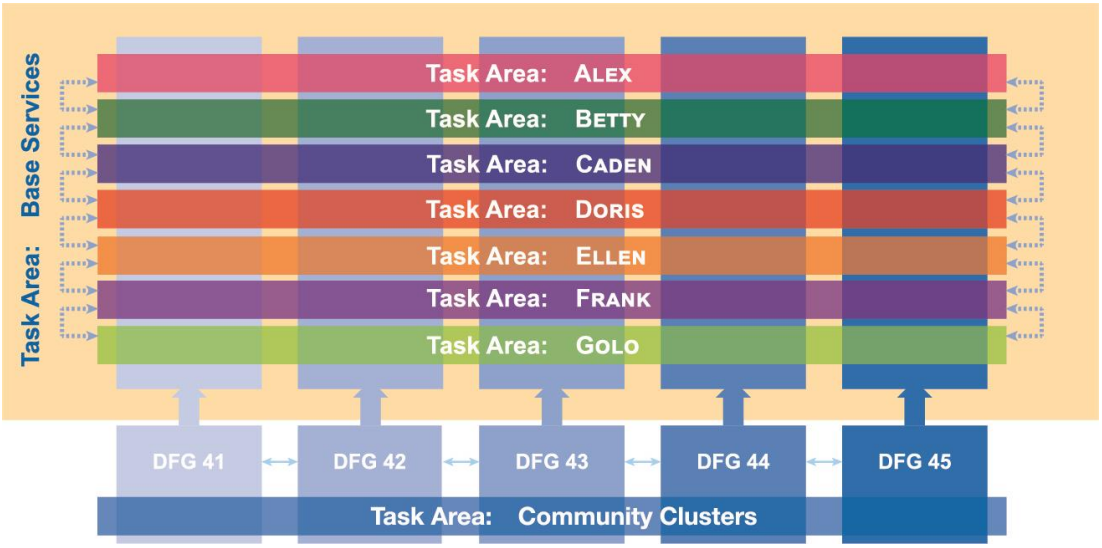


Figure 1.2.3 Structure of the consortium's work programme

NFDI4Ing aims for a maximal participation of engineers with minimal entry restrictions. Therefore, a grassroots democratic structure will be established to create a most participatory membership model (cf. chapter 1.5). Every engineer is welcome to participate in the consortium and take part in one of the archetype activities or community clusters, or use the data related services.

Composition of the consortium by partners

The consortium NFDI4Ing consists of eleven member institutions, each introducing its specific expertise and experience regarding our work programme as it is outlined in the following:

Engineering science at **RWTH Aachen University (RWTH)** is leading in worldwide rankings and takes first place in the DFG Förderatlas [2]. The University holds the status of Excellence. Two Clusters of Excellence with engineering focus are located at RWTH and integrated in this consortium. RWTH is actively developing RDM governance structures and technical services since 2014. The close collaboration between IT Center, University Library, and research communities results in joint infrastructure projects. RWTH´s RDM strategy includes data stewards in all collaborative projects. RWTH is active in a number of initiatives to improve the digital support for research (IdM.nrw, HPC.nrw, AcademicGroupware.nrw) and actively cooperates with other RDM experts from TU9, CESAER, DH-NRW, DINI/nestor, and RDA.

TU Braunschweig (TUBS) has initiated RDM support activities in 2016. TU Braunschweig is an early adopter of the DMP software RDMO. Additional RDM know-how is generated and distributed within the university by members of the HPC-user-council (responsible for administration and usage of TU HPC-Cluster PHOENIX) and the MUSEN centre (Mechanics, Uncertainty and Simulation in Engineering). TU Braunschweig's civil engineering department will expand in 2020 with two incoming tenure-track professorships granted in the context of data-driven modelling.

TU Darmstadt (TUDA) is a pioneer in developing RDM governance and services, both at university level, (e.g. guidelines, model for sustainable funding of research data infrastructures) and in distributed networks of services (major participation in the five-year joint project HeFDI). TU Darmstadt's RDM strategy sets its main effort in delivering RDM services to engineersng efforts are taken to and integrate RDM in the engineering curriculum. With respect to software development, several engineering groups have gained broad expertise in the design of respective software solutions for both heterogeneous measurements and simulations, also on HPC platforms. Its infrastructure institutions are experienced service providers (e.g. RDMO for HeFDI, TDM, Hessian Competence Center for HPC, metadata handling, search engine technology, and repository infrastructures such as DSpace).

TU Dresden (TUD) also holds the status of Excellence and offers a broad variety of disciplines in engineering and experience in all aspects of Research Data Management. For example, the TU Dresden has developed its own research data repository (OpARA) and introduced a research data policy to support its scientists in RDM. TU Dresden is also partner in several RDM-related initiatives and projects, such as the RDA, the Saxony RDM-Initiative or DataJus, and RADAR. For NFDI4Ing, the various engineering communities at TU Dresden are of special relevance, particularly the Faculty of Transport and Traffic Sciences "Friedrich List" as a unique institution for activating community cluster 44 as described in chapter 3.10.

At **Leibniz University Hannover (LUH)** research data management was officially tackled and elaborated in 2014 by the presidential board of the university with the objective of a concept development for the handling of research data. The concept includes an institutional guideline, consulting, training and information services, and a research data repository. The repository facilitates free publishing, long term archiving, the availability of research-related data, and the quotability of datasets via DOI. With experience in joint research data management for large collaborative projects, LUH together with TIB provide the consortium expertise in the development of research data and knowledge management systems supported with machine-readable, domain-specific semantic vocabularies.

TIB Hannover (TIB) - as a pioneer in the field of DataFAIRness - has been coordinating the registration process of the Digital Object Identifier (DOI) since 2005 and provides DOI services for more than 200 data centres as a DataCite founding member, e. g. all TU9. Since 2017, the TIB has been developing a data management system (Leibniz Data Manager) for the semantic indexing and visualisation of heterogeneous data collections, including a semantic description and networking of research data. TIB also has extensive experience in the field of vocabulary development. This includes the development of the ontology STO, which describes standards for the realisation of industry 4.0. The BMBF-funded project STREAM addresses the challenges of the curation, quality, and interlinking of research data in the field of materials science and materials technology.

The **Forschungszentrum Jülich (FZJ)** has a number of engineering science focal points and is closely linked to the RWTH Aachen in these areas via JARA (Jülich Aachen Research Alliance). The materials science focus is particularly pronounced for the development of new energy materials. The highly experimental work at the FZJ requires efficient data management. Therefore, a number of procedures have been developed and tested at the FZJ in the past. Particular mention should be made here of data management tools (RDMO), software development tools (GitLab, Kubernetes), the creation of repositories (Invenio, Dataverse, b2share), PIDs for research data (Datacite membership), electronic laboratory books (JuliaBase, IFFSamples), and universal data analysis methods (JupyterHub).

Karlsruhe Institute of Technology (KIT) is the research university in the Helmholtz association, a member of the TU9, and one of the leading universities for engineering in Europe. The digitisation of engineering plays an important role in the excellence strategy of the KIT. RDM is an acknowledged part of two material science related CoE. Baden-Württemberg funds the Science Data Center MoMaF, a collaboration of material scientists, chemists, and the SCC to build a digital research infrastructure. R&D in data science is visible through the GridKa, contributions to the Helmholtz Information and Data Science research platforms, organisations like RDA, and in several international projects. KIT is major partner in the joint efforts of establishing a EOSC, in particular responsible for the IT-service management.

TU Munich (TUM) will start the TUM Institute for Data Science in 2020 to support the ‘data to knowledge’ effort across the diciplines as a result of the successful excellence initiative. TUM is part of the TU9 and the EUROTECH intiatives. TUM has developed and provided platforms in the past for RDM (MediaTUM – with DOIs, WorkBench@TUM, Git repository) as well as central support through the university library (UB - ERiC-project). TUM for decades has been a strong and effective large-scale user of HPC systems at LRZ as well as nationwide. Several groups have won the prestigious Gordon-Bell award for the world’s largest and fastest HPC simulation. Large amounts of computing time is awarded to TUM on the national TIER-1 and the European TIER-0 machines after a rigorous evaluation system.

The **University of Stuttgart (US)** is a leading technical German university with a strong research focus, expressed by hosting currently two Clusters of Excellence and five Collaborative Research Centres, most of them driven by engineering sciences. Endorsed by its data policy, it expressly promotes the free access to research data and supports its researchers through the research data competence centre FoKUS and the data platform DaRUS. It will contribute the results of the Dipl-Ing project (BMBF) in which scientists and infrastructure have jointly developed concepts and solutions for RDM in the engineering sciences, including the metadata schema EngMeta. A further focus is the handling and provision of research software, reflected by the DFG projects SuSI and ReSUS. The High-Performance Computing Center Stuttgart (HLRS) of the University of Stuttgart supports researchers by providing HPC platforms, technologies, services, and support.

The **German Aerospace Center (DLR)** is the national aeronautics and space research centre of the Federal Republic of Germany. Its extensive research and development work in aeronautics, space, energy, transport, digitalisation, and security is integrated into national and international cooperative ventures. DLR is involved in different interconnected and interdisciplinary data-generating projects and initiatives. It has built up corresponding data expertise. Projects and initiatives currently underway include Digital Twin, CUBE, Factory of the Future, Big Data Platform, CAMS, EMP-E, SciGRID, ngTDP, and Digital Node 4.0.

TU Clausthal (TUC) is active in the fields of energy, materials, environmental, and information technology, as is expressed in several dedicated engineering research centres. TU Clausthal contributes expertise in HPC, simulation science, software and systems engineering, and cybersecurity. The focus will be on handling a large number of experimental results, standardisation and exchange of information between industrial and academic project partners.

Additionally to the joined forces of the eleven members of NFDI4Ing, the consortium has a great number of participants from all over the German engineering research landscape. Furthermore, all members and participants are part of wider national and international networks such as CRCs and further collaborative research projects, RDM specific networks (cf. chapter 1.4), and learned societies, etc., of which many signed letters of support. Our seed funding concept and the open governance model (cf. chapter 1.5) ensure that these wider networks can be activated for and integrated in NFDI4Ing's work programme neatly and quickly, in order to gain special expertise and pilot users needed for our user-centred appoach.

1.3 The consortium within the NFDI

NFDI4Ing represents the engineering community. Engineers typically work in an environment close to other research disciplines, on the one side, and industry, on the other side. As stated in chapter 1.1, key objective 5, being close to industry calls for sophisticated means in authentication, intellectual property, and license management. Research disciplines closely related to the engineering sciences are, for example, natural sciences, mathematics, economic sciences, medicine, and geology. This is reflected by the analysis of all the letters of intent submitted to the DFG by July 4th, 2019, regarding the planned collaborations between consortia. The results of this analysis are presented in Figure 1.3.1, where the size of a node is proportional to the Betweenness Centrality of a consortium. As can be seen from the figure, NFDI4Ing is featured prominently with the biggest Betweenness Centrality node. Therefore, co-operation with other consortia in nearby disciplinary fields is well prepared and foreseen.

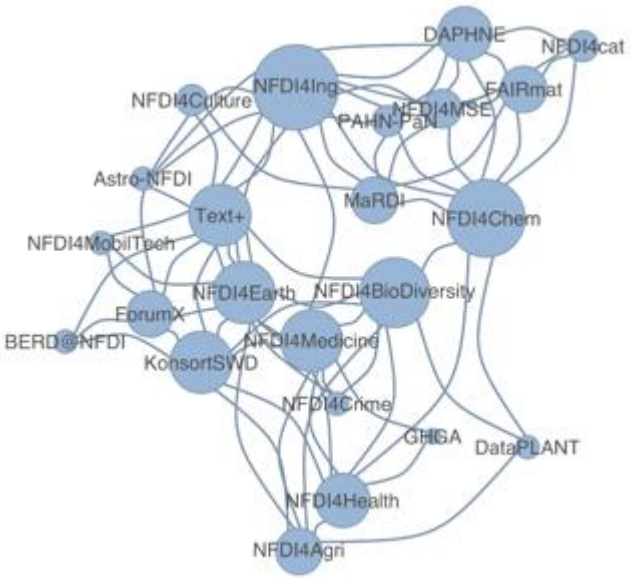


Figure 1.3.1 Collaborations between NFDI consortia. Analysis based on submitted letters of intent. [3]

Some members of NFDI4Ing will also be participating in the following consortia: DAPHNE, FAIRmat, NFDI4Chem, NFDI4Culture, NFDI4Earth, NFDI4MobileTech, NFDI4MSE, NFDI4Phys, MaRDI, PAHN-PaN, and Text+. The former discrete consortium OD-Rex accepted the invitation to integrate the domain of intelligent robotics research into NFDI4Ing (cf. chapter 3.8 and chapter 3.10, Cluster 44). For reasons of organisational size and internal community organisation, intensive previous discussions showed the best way of direct collaboration with other consortia in the engineering sciences like NFDI4MSE, FAIRmat, and NFDI4MobilTech is an agreement on shared tasks regarding cross-cutting topics and community outreach. The agreement with NFDI4MobilTech includes annual workshops, joint working groups, and coordination of community activities in the subject area of transport and mobility, with the DLR acting as a link. Already in preparation of the NFDI call, members of NFDI4Ing participated in FAIRmat workshops.

Besides the common research and disciplinary interest, cooperation is also focused on cross-cutting topics. NFDI4Ing and Text+ will collaborate in scoping out and leveraging the potential of text mining techniques for the extraction of research data out of pertinent digital scholarly literature and documents. In the area of training and qualification, NFDI4Ing and NFDI4Culture have identified Data Literacy, Code Literacy, and the provision of Open Educational Resources as cross-cutting topics for close collaboration. Additionally, both consortia aim to co-operate in the area of standardisation and curation of 3D data types (like CAAD models and other forms of 3D digital representations).

NFDI4Ing aims to support a broad spectrum of disciplines, including all different domains of engineering sciences. Depending on the subject, collaboration is also planned with consortia from different research areas. One focus of NFDI4Ing will be on developing, deploying and sustainably operating a metadata and terminology service (cf. S-3). This includes the technical platform and

the generic tools, while the subject-specific requirements remain part of the respective NFDI consortia. To this end, a number of topically related consortia (NFDI4Chem, NFDI4cat, MaRDI, NFDI4Culture, NFDI4Ing, and NFDI4MSE) agreed to collaborate on this semantical harmonisation. In the field of materials science, for example, the consortium NFDI4MSE is planning to develop sustainable and comprehensive ontological descriptions for its core domain, i.e., the materials science and engineering. With NFDI4MSE's subject-specific approach and respective networks, this profound focus onto microstructural materials representation is an appropriate supplement to the terminological approaches towards harmonisation of NFDI4Ing. The two consortia are looking forward to possible cooperation due to the close fit of their consistent views that promise to complement one another throughout the NFDI project. Among other interests we agree with FAIRmat, that we are discussing a common metadata schema and ontology. For example, NFDI4Ing's expertise in materials processing and aging combined with FAIRmat's expertise and focus on electronic and atomic structure modeling, together enable a complete digital representation of materials under realistic conditions. NFDI4Ing would benefit from the planed materials encyclopedia and analytics toolkit from FAIRmat to identify advanced materials needs in engineering, e.g. materials that exhibit tailored mechanical and material fatigue properties as well as properties of materials in the manufacturing processes. In order to leverage the full potential of this interconnecting approach, NFDI4Ing has started to consolidate a common strategy within NFDI and formed an open working group focused on metadata and ontologies. A workshop dedicated to metadata and ontology handling within NFDI is planned within the first quarter of 2020.

A collaboration with NFDI4Chem already started, focusing on the development of digitisation modules for scientific data in chemistry and material science. Further exchange of the development of data standards and open formats is planned. The same is true for NFDI4Earth, considering the importance of geographical data for various engineering fields like traffic or energy. MaRDI and NFDI4Ing will closely collaborate, including but not limited to reproducible science, the sharing of mathematical models, the generation and description of input data sets from experiments and measurements, and the simulation software developed for analysis and quantifiable predictions. Furthermore, the interface between medical and engineering sciences is rapidly gaining momentum on both sides. Common research topics like micro technology, medical device technology, simulation assisted surgery, sensor technology, or ergonomics require an intensive exchange of data of patients and test persons, which needs to be addressed. Besides common data standards, data privacy is a central aspect of the common usage of personal data at this interface. Here, NFDI4Ing brings in its know-how on data privacy previously gained from collaborations with industrial partners. We will deal with these issues in a cooperation with NFDI4Medicine and NFDI4Neuro.

Topics for NFDI-wide co-ordination

Metadata and ontologies are one of the cross cutting topics that are central to most NFDI consortia. At the same time, metadata, knowledge graphs, and ontologies have the potential to link different consortia and communities during the duration of the NFDI project and beyond. We acknowledge this as one of the highly desirable goals of the NFDI. NFDI4Ing has signed the Berlin Declaration [4] on NFDI cross-cutting topics and joins the other signatories in the effort of addressing cross-cutting topics in a coordinated fashion. NFDI4Ing also joins the signatories in the understanding that the framework for collaboration, as suggested by the DFG, is not comprehensive and remains open for specifications. At any rate, NFDI4Ing is acknowledging the common interest expressed in the Berlin Declaration and will help build a national research data infrastructure that transcends the confines of the engineering communities.

1.4 International networking

The involvement of internationally leading engineering communities and RDM organisations is important to guide the evolution of the services and standards and therefore crucial for the success of NFDI4Ing and NFDI as a whole:

Research Data Alliance 'Research Data Management in Engineering Interest Group' (RDA RDM4Eng IG)

As an important step to expand internationalisation, members of the NFDI4Ing initiative together with international colleagues successfully submitted a proposal to the RDA network to set up an interest group for the engineering community. The RDA IG RDM4Eng [5] aims to change the culture of handling data within the engineering sector, creating awareness and bridging communities and existing initiatives. The identified challenges included to bring together scientific and industrial stakeholders from all relevant sectors to discuss their legal and technological challenges around RDM practices, and the provision of a forum for exchanging knowledge, options, and experiences on a national and international level. NFDI4Ing will thus continue to reach out to the international engineering sector using the IG RDM4Eng, foster contacts to other engineering societies such as NIST [6], as well as use this framework to discuss and evaluate the international scalability of the planned NFDI4Ing services.

International Data Spaces e.V. Association (IDSA)

The IDSA network [7] aims to develop reliable solutions for digitisation in industrial production and business processes. NFDI4Ing will support this process e.g. with the definition of user requirements for the architecture of future international data marketplaces and associated data services.

Conference of European Schools for Advanced Engineering Education and Research (CESAER)

The Task Force Open Science (TFOS) [8] aims at advancing the understanding and implementation of open science within and outside the CESAER network, with particular focus on open access topics, scientific publications and RDM. Several workshops were performed representing major engineering communities such as computational engineering, mechanical engineering, construction, and thermodynamics. The results of the workshops are also reflected in the NFDI4Ing proposal.

Authentication and Authorisation for Research Collaborations (AARC2)

One of the most discussed and crucial points in the context of data management in the engineering sciences is data security and the sovereignty to regulate access to possibly sensitive data from industrial partners. The European project AARC2 offers a Blueprint Architecture (BPA) for access management solutions for international research collaborations. This can be used as a starting point for implementing the AAI (Authentication, Authorisation Infrastructure) for federated data access in the NFDI4Ing (FZJ and KIT are partners in the AARC2 project).

DataCite e.V.

By providing reliable community-owned infrastructure to register DOIs and metadata, DataCite plays a key role in making data FAIR. The DataCite service Re3data is a global registry of more than 2000 research data repositories from a diverse range of academic disciplines. DataCite was also involved in the drafting of Counter (The Code of Practice for Research Data Usage Metrics release 1) which provides a framework for comparable usage metrics by standardising the generation and distribution of usage metrics for research data. DataCite will contribute as a participant in the task area Base Services, focusing on improving and providing PID services tailored to the requirements of the engineering community.

European Open Science Cloud (EOSC)

Cooperation with EOSC is established via NFDI4Ing members who are partners in EOSC-hub and EOSC-secretariat (FZJ, KIT), as well as in EOSC-Pillar and EOSC-synergy (KIT). Input for NFDI4Ing is expected by the work on authentication and authorisation tools and components (see also AARC2-project), the services and portfolio management, and the integration of tools related to managing the federated infrastructure

FAIR working groups and initiatives (Force11, EOSC FAIR, FAIRsFAIR, GO FAIR)

The FAIR-principles are the result of international networking and the discussion about the implementation of the FAIR-principles is continued internationally in various initiatives and working groups. Force11, RDA, EOSC, GO FAIR, and FAIRsFAIR are probably the most active platforms in this context. Solutions developed by NFDI4Ing can be reviewed and discussed in these forums.

EUDAT Collaborative Data Infrastructure (EUDAT CDI)

As a network of more than 25 leading European research organisations, data-, and computing-centres in 15 countries, EUDAT CDI provides services and support as well as best practices for data management and training for various research communities. As a result, concrete services for data hosting, registration, management, and sharing [9] can be reused and adapted for the requirements of NFDI4Ing (members of the consortium: FZJ and KIT are part of EUDAT CDI).

1.5 Organisational structure and viability

From a user's perspective, NFDI4Ing offers different points of access. On the subject level, the task area Community Clusters links up with the existing communities (cf. chapter 3.10). For instance, each cluster convenes a community meeting, starting this year with the materials science and engineering cluster in Karlsruhe on November 19th-20th, 2019. On the methodological level, the archetype task areas address needs in research data management that are very specific to an engineer's day-to-day activities. Engineers from all walks of the profession thus can find their methodological match in one of the archetypes. On the membership level, NFDI4Ing's general meeting of members provides a forum. Whatever the point of access – subject, methodological, or membership – users have several ways of taking part in the work programme. One possible way will be by advancing the state of the art as pilot users or experts. Another possible way will be the scientific exchange during NFDI4Ing's workshops and events. The consortium remains open for new entrants at all times, and joining without active engagement is also an option.

From a provider's perspective, NFDI4Ing's matrix structure of task areas (cf. Figure 1.2.3 Structure of the consortium's work programme) leads to a high degree of integration and allows to pool strengths and resources. Integration is achieved by managing the links between the different task areas' work programmes. To this end, the board of co-spokespersons (cf. Figure 1.5.1) coordinates and prioritises topics across task areas.

NFDI4Ing's internal organisation is set up such as to meet the requirements of a registered voluntary association in Germany – eingetragener Verein (e. V.). This provides the option to register as soon as the wider NFDI network structure suits the need for such a status. In case a NFDI-wide voluntary association will be founded, as discussed on the NFDI Governance Workshop held on August 30th, 2019 [10], the current structure can be adapted to NFDI's organisational framework. Moreover, aligning our internal organisation with the requirements of a registered voluntary association brings transparency, warrants the participation of all members, and keeps NFDI4Ing open to new members. By this way, we ensure NFDI4Ing's viability during the course of the consortium's development. Member status is granted exclusively to legal entities. While natural persons may be admitted as members to the consortium, only legal entities will have voting rights.

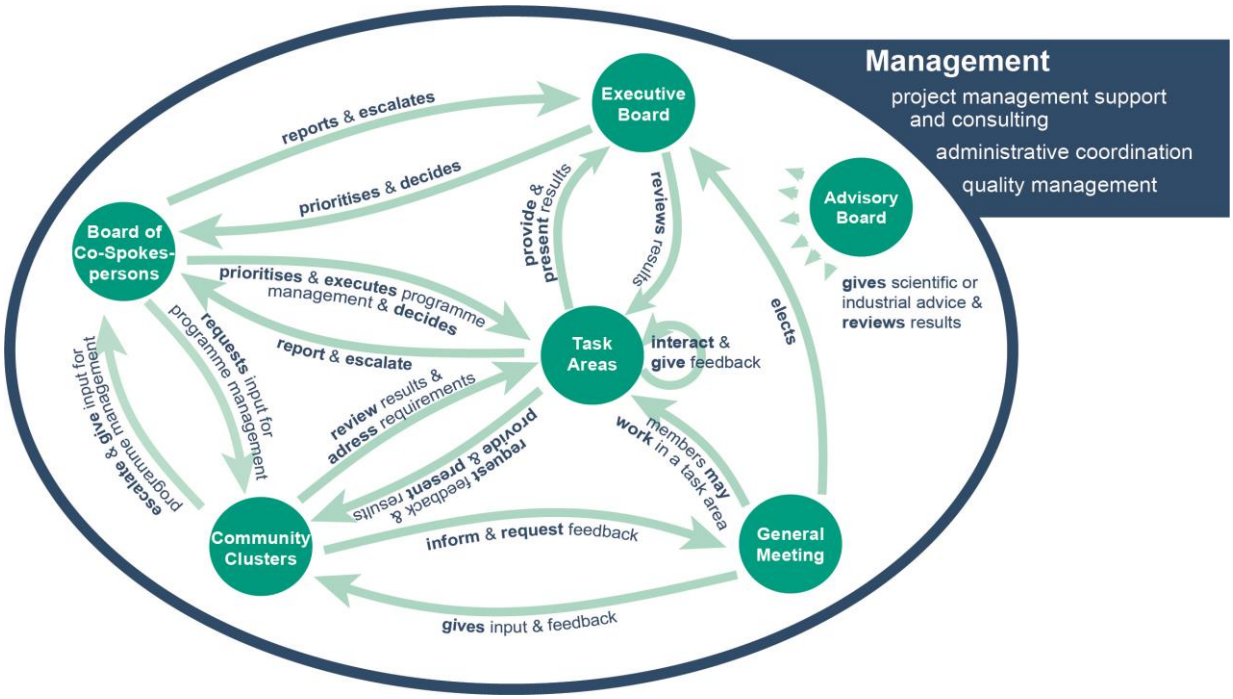


Figure 1.5.1 NFDI4Ing's structures of communication and interaction

NFDI4Ing's internal bodies are displayed in Figure 1.5.1. They comprise three boards: an executive board, an advisory board, and a board of co-spokespersons. The executive board is headed by chair Robert Schmitt from RWTH Aachen University. Schmitt is represented, among others, by deputy chair Peter Pelz from TU Darmstadt. The management is likewise located at the aforementioned two universities, supporting the character of the shared responsibility for NFDI4Ing as a whole. NFDI4Ing's responsibilities, reporting, and decision-making will be established as follows:

- The **general meeting of members** takes all decisions as stated in the envisaged charter and elects the boards accordingly. It has power to make amendments to the charter. The general meeting of members takes place once a year.
- The **executive board** includes the chair, the treasurer, and the reporter. The executive board assumes responsibility for all decisions besides the ones taken by the general meeting of members. It supports and advises the co-applicant institutions responsible for the task area work programmes. Meetings of the executive board with the management are accomplished by regular conference calls. In addition, the executive board and management meet at least two times a year in person.
- The **advisory board** supports the other boards by giving advice regarding NFDI4Ing's objectives from the points of view of other scientific disciplines, of industry, and of politics and governance. Two times a year, the advisory board meets with representatives of the executive board, of the management, and of the board of co-spokespersons.
- The **board of co-spokespersons** is responsible for the functional and scientific programme management. It takes functional and scientific decisions according to the agreed objectives

and work programme. It reports to the executive board. Board meetings are accomplished by regular conference calls. In addition, the board of co-spokespersons meets at least two times a year in person with representatives of the executive board and management.

- The **management** is responsible for the administrative coordination and quality management of NFDI4Ing. It reports to the executive board. The management is already up and running and it will play an important role in ensuring the continuity and quality of NFDI4Ing's work.

A crucial condition of the National Research Data Infrastructure is the ability to adapt to future needs in research data management. NFDI4Ing addresses this issue with agile methods for supporting NFDI4Ing's work programme (cf. chapter 3.11). Furthermore, NFDI4Ing plans to satisfy upcoming operational needs relating to the work of participants, many of which have already committed to joining NFDI4Ing, by operating a seed fund. Our rationale behind the establishment of a seed fund is the following: (i) Seeds can be used in order to adjust the work programme to user needs. User needs, in turn, can only be addressed, if testing or expert involvement is advancing the management of research data. (ii) Most task areas are already planning to involve pilot users and/or will develop prototypes. (iii) The identification and harmonisation of engineering research archetypes achieved, so far, may further evolve during the course of the consortium's development. A new archtype may be called into action. (iv) Finally, operational needs relating to points i-iii, above, should be managed centrally.

We are looking forward to the involvement of primarily, albeit not exclusively, the participants, which already have committed (by a LoC or LoI) to joining NFDI4Ing. In view of their involvement as pilot users, experts or, possibly, in another future role, we suggest that the decisions regarding the disbursement of the seed fund follow an agreed-upon procedure with three levels. At the first level, the management, alone, will approve of the funding of short-term operational needs. At the second level, mid-term operational needs require a majority decision of the board of co-spokespersons. At the third level, long-term operational needs require the consent of the executive board. The seed fund will cover solely staffing costs and applicants will have to match seeds with an equity ratio of one third. Only (co-)spokespersons can apply, meaning that participants need to use the official channel of a (co-)applicant institution. Finally, yet importantly, the approval procedure will be documented and made available for external review (cf. chapter 3.11).

1.6 Operating model

Infrastructure services for the engineering communities are widely decentralised and hence several infrastructure providers and research groups operate individual, mostly unconnected, service "islands". Initiatives aiming for centralisation and standardisation of processes or

workflows and supporting services mostly target specific machinery like HPC systems or work along industry standards. Together with scalable IT services, the alignment of existing infrastructures with best practices for research data management like the FAIR principles is a central challenge for NFDI4Ing.

NFDI4Ing brings together researchers and infrastructure providers. On their own initiative, the partners within NFDI4Ing have already set up several scalable services like GitLab or simpleArchive (cf. chapter 2, NFDI4Ing's expertise and experience) that support specific RDM workflows in engineering sciences and can already be used by researchers from several participating institutions. These existing services are prototypes to build a more extensive RDM infrastructure for engineering sciences in the future.

Other services provided by partners in the consortium or third parties will be required in the future RDM infrastructure. A comprehensive analysis of the services currently provided is in the service portfolio. These services are often only available to engineers of the local institution and need to be extended for inter-organisational use. To allow this scalability in the future, the current infrastructure partners have agreed to follow a set of standards, best practices and requirements towards services offered within NFDI4Ing. For example, as basis for authorisation and access management, all users should be part of the DFN using their infrastructure and SSO service for authorisation. These standards are formulated as "Integration Readiness Levels" (IRL) (cf. chapter 2.3) and give an indication for the development of future components to be included in the consortium's infrastructure. Infrastrcuture porviders in NFDI4Ing already operate as scalable infrastructure as shown in Table 1.6.1 that can support current and future RDM services.

Table 1.6.1 Availible infrastructure of NFDI4Ing

Member	Virtualisation	HPC	File storage
Karlsruhe Institute of Technology (KIT)	VMware Academic Program (VMAP)	bwUniCluster ForHLR	Tape Backup and Archive, LSDF (Large Scale Data Facility)
FZ Jülich	VMWare Cluster	JEWELS, JURECA, QPACE3	Tape Backup and Archive (IBM Spectrum), two OpenStack deployments with object store
RWTH Aachen	VMWare Cluster (loc. Redundant)	CLAiX	Tape Backup and Archive, Blockdevice and Object Storage
TIB Hannover	VMWare Cluster		Scaleable Virtual Machine environment for new services
TU Braunschweig	VMWare Cluster		Tape Backup and Archive (IBM Spectre Protect)
TU Darmstadt	VMWare Cluster	Lichtenberg Cluster	large-inactive-data-Storage based on IBM Spectrum Scale / Protect
Univ. Stuttgart	VMWare Cluster		Tape Backup and Archive (IBM Spectre Protect), Blockdevice and Object Storage (Netapp)

Regarding financial aspects, the operating model of the consortium will be non-profit. Charges for individual services will only be collected on a market level and to compensate for the current costs of the service. In general, if possible, we strive to open services available at local institutions for other consortium members. A billing of individual researchers for services should only be used for strategic reasons, i.e. to encourage quality consciousness in the use of NFDI4Ing services. In general, we expect that an approach, where everybody brings services to the table that all participants can share, is the appropriate mechanism to encourage information sharing and to reduce administrative overhead.

2 Research data management strategy

In the whole of this chapter, we outline the NFDI4Ing research data management (RDM) strategy consisting of the following components:

- An assessment of the current state of established RDM practices and established RDM infrastructures provided at institutions with a high affiliation towards the engineering sciences.
- The vision of future RDM within NFDI4Ing, based on our research-method oriented approach encompassing the establishment of archetypes, base services and community clusters.

A core concern of research data management in engineering is to unite the different dimensions of institutional, subject-specific and international RDM solutions in such a way that an engineering-specific approach to address the FAIR principles is provided. This is also in detail outlined in NFDI4Ing's overarching key objectives (cf. chapter 1.1).

Current state of research data management in engineering

Availability of data, including descriptive documentation and corresponding software, is crucial for engineering research. While engineers use and produce large amounts of research data, the corresponding data handling rarely follows standardised RDM strategies. There is neither a central or cross-linked infrastructure or service nor widely used guidelines or an accepted committee to propose those. In engineering research, RDM is usually handled on the basis of simple file systems. It relies most often on the manual organisation of directories, files, and metadata. Data and metadata are often created on a case-by-case basis and stored separately, inconsistently, and untraceably. In many cases, the created metadata are not even really metadata in the sense of being machine-processable information about distinct datasets. These circumstances diminish the information value of research data and hinder the development of automated workflows, relying on metadata. This may be because appreciation of maintaining and publishing research data is low, and it is not established in the scientific reputation system. This leads to the fear that implementing and following RDM workflows is an additional burden rather than an advantage for the researchers and their daily work overshadowing the long-term benefits.

A major challenge in establishing standardised RDM procedures is the vast variety of engineering problems that often result in highly specialised, bottom-up solution approaches, cf. Task Force Open Science (TFOS), American Association of Engineering Societies (AAES) and CESAER (Conference of European Schools for Advanced Engineering Education and Research). For example, experimental measurement devices as well as simulation software are typically custom-built and coded, or engineering research may involve investigating products with proprietary interfaces, lacking common standards.

As a result of insufficient data handling, research work and results suffer in several respects. New students and employees on all levels in academia and industry, from bachelor candidates to postdocs, often need long preparation times for finding and utilising data and software codes produced by their predecessors. This usually comes along with a loss of information and efficiency, and quite often results in creating new data and rewriting code, instead of reusing existing one.

NFDI4Ing's expertise and experience

The most important part of NFDI4Ing's expertise and experience in managing engineering research data lies with the members and participants from active engineering research as it is outlined in detail in the respective task areas 0– 3.8 and 3.10.

From the infrastructure provider's perspective, there are quite few existing RDM infrastructures that are (1) fitting engineering demands and (2) available to a broader engineering public. Thus, most engineers can only rely on general RDM services. None of these are tailor-made for the engineering communities' requirements. E.g., re3data lists approximately 1.900 discipline-specific data repositories, of which only 132 are intended for engineering sciences. Further analysis shows that only very few of those are actually specialised in engineering data, in terms of metadata schemes or supported data types [11]. The same holds for the RDA Metadata Standards Directory, listing not a single standard or tool for the engineering sciences, except for crystallography [12].

On the other hand, there are several examples for engineering data standards that might be used for RDM as well. There is the ISO 8000 family on data quality, rapidly being adopted by industry (mainly in manufacturing), but still in development and hardly applied to engineering research. In the field of civil engineering, the building SMART portfolio, including the Industry Foundation Classes, is a well-established example for the transfer of data within and between research and industry. The International Data Spaces Association (IDSA) will be an important part of NFDI4Ing's link to industry. IDSA has designed a reference architecture model for a virtual data space, leveraging existing standards and technologies, as well as accepted governance models for the data economy, to facilitate the secure and standardised exchange and easy linkage of data. IDSA is also closely connected to the "Plattform Industrie 4.0" initiative, bringing together expertise in digitisation of industrial production in Germany.

While building the consortium, NFDI4Ing has already identified several services that the engineering community is lacking. In independent approaches, different members of NFDI4Ing have opened locally existing services for each other in order to enhance RDM workflows for the engineering community. Examples for RDM services currently available to engineers via one or more consortium members are (1) tools for data management planning, different platforms for (2) data storage and (3) for software development, as well as (4) repositories and PID services including (5) a prototype for the exchange of metadata between repositories.

We now elaborate on the examples 1-5 given, above. (1) In NFDI4Ing, several members are running an RDMO instance, the de-facto standard tool that supports researchers in creating data management plans in Germany, some of which are jointly operated. There is preliminary work done at Aachen and Darmstadt for joint curation of DMP templates, which will be further developed by NFDI4Ing. (2) In order to help researchers who completed a project, RWTH Aachen built an easily accessible workflow for archiving of research data, “simpleArchive”. The workflow and its implementation were later generalised and transferred to TU Darmstadt. Both instances of “simpleArchive” are now jointly operated and serve researchers at both universities for preserving research data that is no longer actively needed or that needs to be kept in its original state. (3) A GitLab instance, operated in Aachen, was opened up for cooperative use and extension for RDM workflows nationwide. Even though most universities also operate their own infrastructures, members of 31 German universities and research institutions are collaborating on the service. Six are among the TU9 members showing the high demand for a common infrastructure for working with source code. (4) All members of NFDI4Ing operate institutional research data repositories based on individual software stacks. Some are providers of generic software solutions to build own disciplinary repositories (e.g., KIT Data Manager). In almost all of these repositories, DOIs for research data publications are used, for which the TIB serves as DOI provider and offers PID support. (5) To connect the existing infrastructures and to facilitate findability and reuse, a CKAN prototype at TU Darmstadt uses the DCAT standard to harvest the institutional research data repositories at LU Hannover, TU Darmstadt and RWTH Aachen. Due to the well-established DCAT vocabulary, an extension to the other members’ repositories is possible.

Relating to teaching skills within the consortium there is already considerable experience in conveying RDM contents and data literacy by means of various teaching formats. Regular RDM trainings and workshops, partly in cooperation with graduate schools, are well established with all partners. Additionally, specific formats such as, e.g., participation with the data and software carpentries, blended learning concepts, and eLearning formats including webinars or online tutorials are also provided. The TU9 members have already developed an extensive pool of modular RDM training material, which is directly targeted at the engineering community and of which some parts have already been published under open license [13]. Furthermore,

collaboration with the platform forschungsdaten.info [14] as well as with the EU framework FAIR4S [15] is taking place.

The general consortium-specific expertise and experience of the co-applying infrastructure providers in building RDM services for engineering is listed in chapter 1.2. A full, detailed portfolio of all services already existing within NFDI4Ing is presented in the service portfolio.

Envisaged state of RDM and user involvement

As an outcome of the assessment of the current state above, the NFDI4Ing strategy is characterised by a user centred and method-oriented approach. The integration of data management tools and services into the engineering research process and in particular the best possible automation of the handling and description of data is crucial for the acceptance of existing and planned services. This strategic approach consists of, first, a base layer of development and enhancement of suitable RDM services by renowned infrastructure providers (cf. chapter 3.9). Second, the seven archetypes derived from engineering-specific workflows and needs (as outlined in chapter 1.2) devise custom-build engineering RDM services based on the provided service modules. Participating experts and especially pilot users will be evaluating and disseminating the different archetypes' services. Finally, our envisaged state of RDM in engineering is the seamless integration and combination of different services (cf. chapter 2.3) from the NFDI4Ing archetypes into every single engineers' daily workflow. It is estimated that each engineering research group needs services from three or four archetypes (cf. chapter 1.2).

In the task areas (cf. chapter 3), the specific method-oriented approach of each archetype is outlined in detail. According to the RDM strategy the archetype task areas solve the problem on a researcher's level representing a specific part of the whole engineering community. In the following, we map these approaches to the overall NFDI4Ing objectives:

1. The objective to enable all engineers to retrace or reproduce all steps of engineering research processes is taken on by the archetypes ALEX (cf. chapter 3.2), CADEN (cf. chapter 3.4) and GOLO (cf. chapter 3.8). There, NFDI4Ing focuses on the development of suitable workflows for one-of-a-kind experiments (ALEX), complex sample processing (CADEN) and field data (GOLO).
2. The objective that research software is treated as research data is taken on by the archetype BETTY (cf. chapter 3.3) and in the task area Base Services (cf. chapter 3.9). There, NFDI4Ing focuses on the challenges emerging from the fact that software usually is developed by domain specialists rather than software engineers. These challenges include e.g. the need for validation and reproducibility as well as the complex dependencies on other software and possibly hardware architectures.
3. The third objective of NFDI4Ing is to enable the recording and linking of auxiliary information and provenance within RDM workflows, focusing on automatisation and consequently the

reduction of the manual data handling tasks. This objective is taken on by the archetype CADEN (cf. chapter 3.4), whose research deals with complex sequences of processing and analysing steps, applied to samples and/or data sets, and also partly reflected in the archetype ALEX (cf. chapter 3.2).

4. The objective to enable the sharing and integration of enormous amounts of data is taken on by the archetypes DORIS (cf. chapter 3.5), whose focus is on conducting and post-processing high-resolution and high-performance measurements and simulations on HPC systems, and ELLEN (cf. chapter 3.6), whose focus is on the analyses of complex systems comprising a large set of multidisciplinary interdependencies.
5. The fifth objective seeks to allow unhindered collaborative research, while at the same time preventing unauthorised access to confidential data, considering the close proximity to the industry sector. This objective is taken on in two ways: One, by the archetype ELLEN (cf. chapter 3.6), focusing on the support of engineers in their search for data by increasing the number of potential data sources. Two, by the archetype FRANK (cf. chapter 3.7), whose research deals with the management of different data types and dimensions from many participants.
6. The sixth objective focuses on efficient solutions to generate machine-processable representations of auxiliary information based on open standards, considering data-driven analysis methods such as AI and ML. This objective is taken on by the archetypes CADEN (cf. chapter 3.4), ELLEN (cf. chapter 3.6), FRANK (cf. chapter 3.7) and in the task area Base Services (cf. chapter 3.9).
7. The seventh NFDI4Ing objective aims at an improved data- and software-related education (data literacy) and availablility of domain and application specific best practices. This is mostly taken on by the community clusters (cf. chapter 3.10) and the task area Base Services (cf. chapter 3.9).
8. The final objective focuses on making research data publications a recognised and valuable research output among the engineering community. Such a change of perspective must be driven by a cultural change and is mostly addressed by the community clusters (cf. chapter 3.10).

Finally, archetypes and community clusters will ensure that the planned service portfolio of NFDI4Ing will fit and enhance their methodological approaches. NFDI4Ing will start with a focus on the engineering research areas according to the DFG classification (areas 41-45) in the first five years of the project, while at the same time searching for additional pilot users and participants from disciplines and research-methods which are not yet represented by the existing project partners.

Monitoring of user needs and derived change process

User involvement is the basis for all decisions regarding design and implementation of the services based on the eight key objectives of the NFDI4Ing consortium (cf. chapter 1.1). We will implement several measures to constantly monitor the user needs, e.g. via target group specific community events, ticket and polling systems, and surveys. To this end, the five community clusters (oriented to the DFG classification of research areas) serve as the fora for exchange on specific user needs, reach out in all engineering subdisciplines, and enhance the particular change processes leading to new RDM cultures. In particular, they mediate engineers from certain sub-disciplines and their needs to the convenient archetypes and, thus, monitor the needs for new or newly aligned archetypes. All details concerning respective measures and tasks in the community clusters are outlined in chapter 3.10.

New ideas by engineers from within and outside of the consortium can be rapidly prototyped with support through seed funding (cf. chapter 1.5). Already existing services from other providers following NFDI4Ing's guidelines can be technically integrated into the service portfolio, if desired, cf. chapter 2.3. Last but not least, NFDI4Ing's transparent and flexible governance allows for readjustments where needed. The advisory board (cf. chapter 1.5) opens up the discussion from the academic circle and aligns the engineering community with RDM discussions in other disciplines and the demands of non-academic stake-holders (e.g. from industry).

Data selection and quality management

The approach to data quality management adopted by NFDI4Ing is based on three elements: data quality assurance processes, data quality management tools, and data quality metrics development. Data quality assurance processes include process groups such as data operations, continuous data quality control processes, and data quality improvement processes. Within the framework of NFDI4Ing the data quality assurance processes carried out are both general and specific for the task area's data quality characteristics, which emphasises the importance of choosing the right tools for managing data quality (e.g. DMP).

Metrics will be used throughout the consortium to objectively evaluate data quality and thus enable engineers to select data for further curation. Metrics for monitoring and evaluating data quality have to be both standardised and specific, focused on special data types and research methods in the archetypes. Such metrics include integrity, authenticity, measurability, manageability, objectivity, traceability, provenance, accessibility, confidentiality, reusability and interoperability as well as relevance for different stake-holders (one's own, other academics, industry, and society). Most of the engineering community is familiar with quality management and its methods. In NFDI4Ing, we work on the topic of quality management incl. metrics in a dedicated measure in the task area Base Services, specific requirements and actions are outlined there (cf. chapter 3.9 S-1).

2.1 Metadata standards

Necessary minimum set of metadata for engineering data

In NFDI4Ing, we aim not only to use metadata for documentation and indexing of research data stored in repositories and to fulfil the prerequisites for citation. We also want to leverage the potential of standardised metadata during active research by using it for facilitation of tasks like (automated) retrieval, analysis or combination of complex research data as well as to enable the full scope of FAIR compliant usage afterwards. To accomplish this, the metadata must be specifically tailored to the individual research data and intended application and provide information on all parameters relevant for finding and interpretation of the research data. Typical elements for engineering are IDs of technical components and samples, technical data of sensors (e.g. characteristic curves) and samples (e.g. materials data), detailed documentation of experimental setups and test beds, input parameters of simulation runs or data processing routines, log data, software versions used and access permissions and licenses. As a result, the minimum set of required metadata will be rather large. In addition, since the field of engineering encompasses diverse subject areas and working styles (as reflected by the community clusters and archetypes contained in this proposal), we must support multiple minimum sets reflecting the individual needs of the research area and archetype.

This makes it necessary to ensure that the sets of metadata used at the active research stage are interoperable. To this end, we will implement a hierarchical metadata model, where specific metadata standards are derived as “children” of more general standards. In this model, like in object-oriented programming, each child inherits all elements contained in its parent, and extends these with new elements. This way, we will start from widely used metadata standards (e.g. DataCite MDS [16] or DCAT [17] or, if available for the area at hand, more specific standards like CodeMeta [18]) and successively derive more detailed discipline-specific children. The design will be a multi-level hierarchy, in which the scope of the derived children gradually becomes more narrow and precise, e.g. refining the subject area along a chain like engineering ➔ mechanical engineering ➔ mechanics ➔ fluid dynamics ➔ etc. The resulting hierarchical tree of related standards maximises their interoperability, since the standards are always downwards-compatible up to the level of their most specific common parent.

Building on this approach, we will also implement a modular metadata design in which different realms of metadata (e.g. description of subject area, method, sample and result as well as components like sensors, machinery, etc.) are separated, resulting in metadata consisting of multiple smaller datasets governed by different standards. This drastically increases the standards’ applicability and maximises their reusability across disciplines. In addition, it facilitates the hierarchical design described in the previous paragraphs, since it enables us to exploit hierarchical relations independently, designing one branch for description of subject areas, one for methods, etc. By this way metadata can be highly interoperable on one level (e.g. when

possessing a specific common parent for the metadata module describing the method), even if the relation on other levels (e.g. the subject area) is much less close.

Engineering specific metadata standards

There are no metadata standards specific for engineering that have achieved a sufficiently wide community acceptance to be included in the RDA Metadata Directory [19]. The consortium is also not aware of any metadata standards that reach the level of specificity required to be useful during active research (cf. previous subsection). However, there are current developments of engineering-focused metadata standards among the participants of NFDI4Ing, e.g. EngMeta [20] or D-SI, [21] that will be used as part of the proposed strategy for metadata (cf. next subsection). In addition, there are several controlled vocabularies relevant to engineering (e.g. QUDT, [22] VIM3 international vocabulary of metrology, [23] Allotrope Taxonomies [24]) as well as commercial terminologies (e.g. Thesaurus Technology and Management, [25] DKF thesaurus for automotive technology, [26] or DIN, ISO and other norms), that need to be considered. Furthermore, there are device- or manufacturer-specific formats, in which metadata are generated by instruments or commercial software. These formats usually lack the interoperability of a true standard, but can be leveraged as sources of structured metadata available for automated extraction of metadata elements.

Coordination within NFDI regarding cross-disciplinary metadata standards

The same strategies that help allow NFDI4Ing to assure interoperability between multiple metadata standards required within the consortium (cf. Necessary minimum set of metadata for engineering data) also help on the cross-disciplinary level. NFDI4Ing's hierarchical approach to modelling metadata introduces a tree-like structure of relations between metadata standards, where increasingly more specific standards are derived as children inheriting the elements of their more general parents (cf. Necessary minimum set of metadata for engineering data). In this way, all standards are interoperable on the level of their most specific common parents. In addition, reuse and adaption of standards are facilitated and can be realised at the most suitable level of hierarchy, seamlessly enhancing the interoperability between standards.

In a similar fashion, NFDI4Ing's modular metadata design that creates separate metadata standards for different realms of metadata like subject area, method, sample, etc. (cf. Necessary minimum set of metadata for engineering data) maximises the standards' reusability across disciplines by avoiding unnecessary layers of specificity. This way, a standard specific to a certain method can be used in any scientific discipline employing that method, and a standard describing a sample can be reused without having to consider the method. To further increase interoperability, NFDI4Ing will define metadata standards using terms from controlled vocabularies as building blocks. This RDF-based [27] approach is known as application profiles [28] and is considered to represent the highest tier of interoperability for metadata available at present [29]. It significantly reduces the complexity of mapping between standards while at the

same time allowing for the degree of flexibility required for defining application specific standards. In order to leverage the full potential of this approach, we have already started to consolidate our strategy within NFDI (cf. chapter 1.3).

International metadata coordination

NFDI4Ing will actively participate within the community-specific RDA IG "Research Data Management in Engineering" and consolidate its approach to metadata handling with groups dedicated to this topic (e.g. "Metadata IG", "Research Metadata Schemas WG", "Vocabulary Services IG", "Data Discovery Paradigms IG", "Data Foundations and Terminology IG", etc.). Furthermore, the established metadata services will include international standards for referencing, such as the provision of a Digital Object Identifier (DOI). Generic metadata standards such as the DataCite or DCAT metadata standard are used in research communities of all disciplines. Their use facilitates easier interoperability between systems and disciplines, enables standard-based quality management and in this way provides reliability for re-use of data. Subject-specific metadata, on the other hand, are tailored to the special features of the data and standards used and provide very specific information that is difficult to integrate in a generic schema. Therefore the integration of more extended metadata into the metadata provided by PIDs is important for FAIR data. As a participant of NFDI4Ing, DataCite will aim for a generic integration that works with multiple metadata standards, and update the DataCite metadata schemata where needed.

In addition, our strategy for modelling metadata is chosen to facilitate compatibility between standards and promote reuse. This is meant in two ways: 1) reusing existing standards and term definitions wherever possible and 2) making standards and vocabularies newly created within NFDI4Ing available for reuse. The former is achieved by modelling metadata standards as application profiles and implementing a hierarchical inheritance concept in which new standards are defined as derived children of existing standards (cf. Necessary minimum set of metadata for engineering data). The latter is achieved by hosting and indexing newly created standards in a public repository (cf. S-3-1) and by implementing a modular design for metadata which maximises the applicability of standards across disciplines (cf. Necessary minimum set of metadata for engineering data and Coordination within NFDI regarding cross-disciplinary metadata standards).

Offering the generated standards and vocabularies for reuse is a central effort of NFDI4Ing's metadata and terminology services. All standards and vocabularies generated within NFDI4Ing will be made available for reuse alongside tools and best practices developed for their integration into typical workflows. Our findings and the generated standards will be discussed and advertised within suitable channels like the RDA IGs mentioned, above. For a detailed description of our outreach and dissemination concept, cf. section 3.10.

2.2 Implementation of the FAIR principles and data quality assurance

Implementing the FAIR principles [30] for the engineering sciences is a central goal in all of NFDI4Ing's task areas. All the consortium's overarching key objectives (cf. chapter 1.1) are strongly committed to making FAIR data a common practice in every engineer's day-to-day research. Despite their importance, the FAIR principles are far from being well-known or properly implemented in our target communities. There are some engineering specific challenges in fostering FAIR data: 1) open and accessible data is not always desired or possible in applied engineering research due to confidentiality demands; 2) the re-use of data is not common in some engineering disciplines (creating solely new data while in the first place developing the underlying technologies); 3) the idea of interoperable networked data systems meets rather scattered (self-engineered) infrastructures in engineering research institutions. Therefore, NFDI4Ing will build services following our key objectives (cf. chapter 1.1) that meet the engineer's demands as well as the FAIR principles:

Following NFDI4Ing's key objectives 4 and 8 the findability of data and software will be reached by the provision of suitable data repositories that are fitted to engineering data types, lab equipment, simulation procedures and existing research workflows. Key is the utilisation of rich, subject-specific metadata, including its automated logging, and intuitive interfaces for effort-minimised use by humans.

Accessibility to metadata and data for authorised humans and machines will be reached by a networked technical infrastructure based on standardised protocols and open standards that offers the retrieval of high data volumes and rates as automated as possible as it is expressed by the key objectives 4 and 6.

Interoperability of engineering data will be reached in fulfilment of key objectives 3 and 6 by the development of services and ontologies that integrate existing industrial (meta)data standards and enable (meta)data exchange. Thus, semantic information processing based on standardised, yet sub-discipline and method-specific metadata information becomes possible.

Reusability in the sense of conformity to community norms and standardised licensing might be the biggest challenge within NFDI4Ing. Therefore, fostering reusability on all levels stands in the focus of the key objectives 1, 2, 3, 5, 6, and 8. We will (1) foster the coordinated curation of data and engineering specific data management planning, (2) develop recommendations for suitable open and interoperable file formats for engineering data, (3) provide suitable authentication and authorisation mechanisms, (4) develop standard usage licenses and provenance marks for sensitive data, and (5) provide repositories for software, for experimental data processing and analysis, as well as facilities for systematic testing of software.

Training in FAIR data literacy for engineers is a key objective of the consortium for itself (objective 7). Actions on promotion of the FAIR principles in engineering will be central in the community

clusters (cf. chapter 3.10). One further measure to generally support the FAIR principles in engineering is the foundation of the RDA Interest Group "Research Data Management in Engineering" (IG RDM4Eng, cf. chapter 1.5). Data quality assurance is a central topic in NFDI4Ing (cf. chapter 2) and will be addressed by measures S-1 and CC-5, the latter developing engineering specific quality issues.

Picking the right combination of underlying technologies is very important in the early stages for the development of reliable services. At NFDI4Ing, the partners have agreed to follow a set of standards, best practices and requirements towards the development of FAIR infrastructures and services. These standards are formulated as “Integration Readiness Levels” (IRL) in chapter 2.3 and give an indication for the development of future components to be included in the infrastructure. Furthermore, the continuous feedback loop provided by a close interaction between NFDI4Ing archetypes, community clusters, and service providers will ensure that a community-driven development of services takes place (cf. chapter 1.6).

2.3 Services provided by the consortium

Planned NFDI4Ing service structure and minimum services

As part of this initiative, the existing RDM services of all consortium members and relevant participants have been recorded within a NFDI4Ing service portfolio document and described according to their role and assignment in the life cycle of research data. The portfolio is presented after the work programme, below. The NFDI4Ing consortium aims to form a coordinated RDM team, bringing together components of existing local services as identified in chapter 2 to form higher-level cross-cutting services which will be further adapted to the specific requirements of the engineering community. This structure is outlined in Figure 2.3.1, with the lower pyramid level providing local, institutional RDM services, the second level providing adapted NFDI4Ing cross-cutting services which are, e.g., used and modified in NFDI4Ing archetypes. Both service levels work towards the establishment of a cultural change on how RDM is treated within and outside of the engineering community.

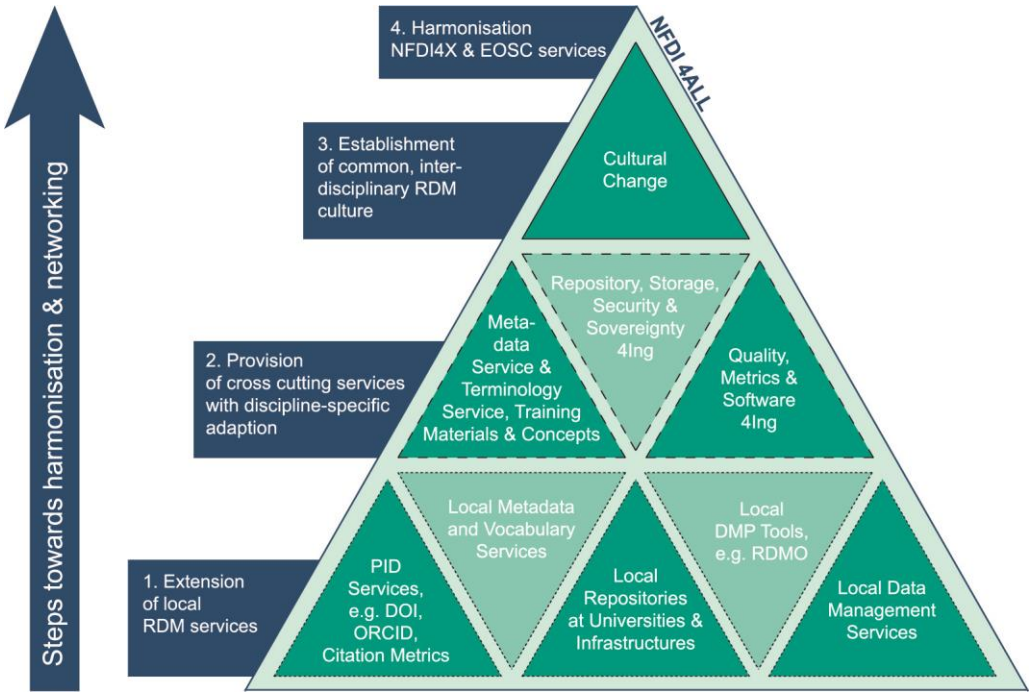


Figure 2.3.1 The envisaged NFDI4Ing service pyramid

Local RDM services (lower level) of all partners are recorded and compared with requirements from user stories. From this, joint NFDI4Ing cross-cutting services with discipline-specific adaptation (second level) are developed, which in turn can be integrated into a higher-level NFDI service structure (third level) which is driven by an overall cultural change regarding RDM in the engineering community.

In the service development, the community principle prevails, meaning that national services jointly hosted by institutions develop a higher impact than services driven by individual or only a few institutions. Through the multi-level service approach, each NFDI4Ing member and participant can continue to decide freely on the scope and design of the local services. NFDI4Ing will set up and provide the following minimum discipline-specific services within the first five years. These will be adopted and continuously expanded by the archetypes:

- Data quality assurance and metrics for FAIR data: A service to provide a commonly accepted set of quality criteria and approaches to achieve a quality standard which is accepted by the engineering community. This service will, e.g., be adopted in the archetype FRANK (cf. chapter 3.7), where tools for the treatment of corrupted and missing research data sets will be developed.
- Research software development: A service which provides support regarding the development of sustainable research software for engineering sciences (cf. chapter 3.3).
- As the provision of harmonised metadata standards and tools for the semantic description of research workflows are required in most NFDI4Ing archetypes, a central service for the development, evaluation, and maintenance of metadata and specialist vocabularies and

ontologies in the engineering sciences will be provided (cf. chapter 3.2, chapter 3.4, chapter 3.5, chapter 3.6, chapter 3.7, chapter 3.10).

- Data storage, archives and repositories: Services to support best practices and tools suitable for storage, exchange, publication, and long-term archival of data of varying quality and volume will be provided.
- Data security and sovereignty: Services to support best practices for securely providing and accessing (meta)data in a distributed architecture operated by different scientific communities will be provided. This service will, e.g., be adopted by the archetype DORIS (cf. chapter 3.5), where access concepts (accessibility, access rights, data sovereignity) at HPC level will be developed.
- Community-based training on enabling data-driven science and FAIR data: A service to support common practices to be adapted across engineering disciplines. It will provide discipline-specific trainings tailored towards specific challenges like handling large amounts of data or transitions between empirical and simulation-based research environments.
- Text and data mining: As engineering research data is still often 'hidden' in scientific literature such as articles and technical reports, a service to support data science in the context of enrichment of unstructured data with structured elements and of enabling text and data mining will be provided (cf. chapter 3.2, chapter 3.4, chapter 3.6).

From the archetypes and community clusters, requirements are formulated to drive the development of new, improved and harmonised RDM infrastructures, which are in turn used, modified and evaluated by the archetypes and in the respective engineering workflows. On their own initiative, the partners within NFDI4Ing have already set up several scalable services like GitLab or simpleArchive that support specific RDM workflows in engineering sciences and that can already be used by researchers from several participating institutions. These existing services form a nucleus to build a more extensive RDM infrastructure for engineering sciences in the future. A detailed description of the planned services can be found in the task area descriptions of the NFDI4Ing archetypes (cf. chapter 3.1) and in the task area Base Services (cf. chapter 3.9).

Contingency measures: the service life cycle in NFDI4Ing and participatory processes

The following graph shows the different phases of a service life cycle (Figure 2.3.2):

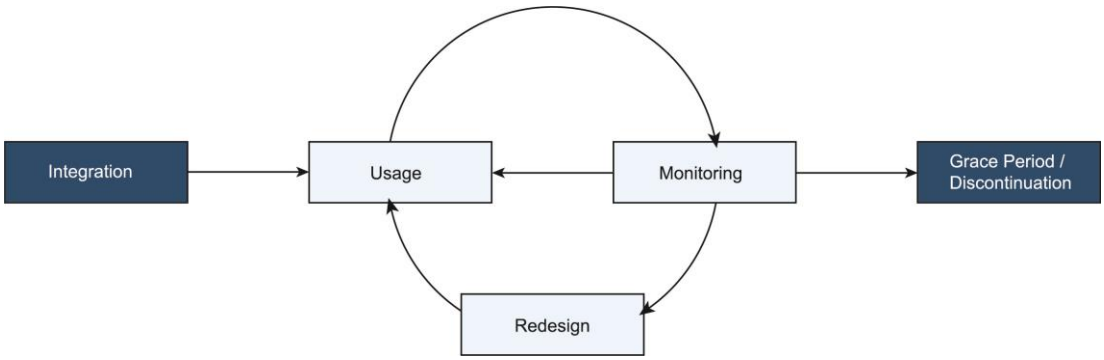


Figure 2.3.2 Service life cycle within NFDI4Ing

In the first phase of NFDI4Ing, the integration of a new service requires the providing institution to be part of the NFDI4Ing consortium. Researchers from participating institutions are, in general, entitled to the use of NFDI4Ing services, other researchers need to sign an agreement with the consortium. Best practices such as ITIL are suggested for processing problems, changes, and planning processes. New services will be integrated if the following requirements are fulfilled:

- The service addresses a specific purpose that is not yet covered by existing services. It is possible to operate distinct services with the same purpose serving different requirements of archetypes.
- The service is establishing a standard or is employing a standardised interface.
- In general, the service should be provided by a central provider. A decentralised service should only be used in case of requirements which cannot be handled by a central service. In this case, an exchange of data between the different decentralised services should be possible.
- For a new service, contracts, e.g. covering data privacy issues, between the service provider and the users of the service are required. The goal is to establish a general service agreement for NFDI4Ing services, which can be easily adapted to new services and institutes or researchers. Additionally, every service provider has to define key performance indicators (KPIs) for their service, including number of active users, size of the contents, uptime, recovery time of the service, usage of service components, number of tickets and requests to the service provider. On a regular basis, these KPIs should be evaluated against projected goals and published within the consortium.

Usage of a service: Any member can use the services of the consortium when he accepts the service level agreement. The service level agreement serves as a basis for decision of any further actions taken to redesign or extend the service.

Redesign of a service: If a service does not fulfil the defined purpose anymore, is not used, its technology is outdated, or its functionality should be extended, then it can be redesigned. The redesign should follow the same procedures like the integration of a new service and should be co-ordinated by the service provider.

Longevity, grace period and discontinuation of a service: In general, a service will only be discontinued if it is not used anymore in the consortium. If a service is discontinued, the provider offers a certain amount of time for migration and supports the migration with export tools or interfaces. After a service is discontinued, the data and programmes of the service will be archived for 10 years. Individual terms of usage might differ, but only if this is already a known fact when the service level agreement is signed. Also, software services should be up-to-date at all times following the actual development in this area.

Financial aspects of RDM services: Charges for individual services will only be collected on a market level and to compensate for the current costs of the service.

To allow the scalability of services in the future, the NFDI4Ing partners have agreed to follow a set of standards, best practices and requirements towards the planned services. These standards are formulated as “Integration Readiness Levels” (IRL) and give an indication for the development of future components to be included in the infrastructure. The "Integration Readiness Level (IRL)" will record the status of the central and decentralised components used to set up and maintain NFDI4Ing. The IRL describes the a) current and b) potential status of new services to be (in the future) included in the service portfolio (e.g. from new or associated partners and from other NFDI consortia close to the business) and comprises 5 harmonisation levels:

Level 0 - No specifications: Everybody does what he/she wants; the service is not networked, there are no or only narrowly open interfaces, and no standards; this corresponds to the status quo of many institutes and departments within the engineering sector.

Level 1 - The service includes documented best practices, training materials, standards or other community specific features (e.g. metadata standards, specific DMP templates).

Level 2 - The service includes harmonised components and open interfaces.

Level 3 - The service is part of the discipline-specific services of NFDI4Ing or contributes essential components to it and supports theming, branding, and multi-client capability within NFDI4Ing.

Level 4 - The service is part of the NFDI vision of integrated services & data within the entire German research data landscape (NFDI4X, Figure 2.3.1) and within the EOSC.

At the latest in the final expansion stage (after 10 years), the services are to be connected to a central NFDI instance (NFDI4All) according to the vision of RfII (Figure 2.3.1).

3 Work programme

3.1 Overview of task areas

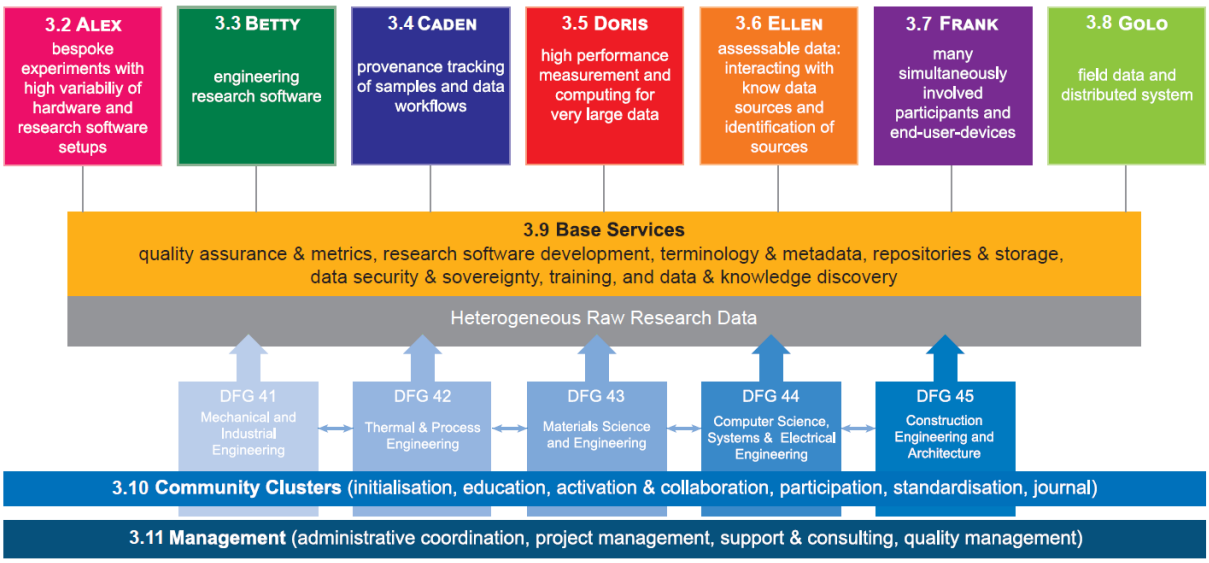


Figure 3.1.1 Overview of task areas

The work programme of NFDI4Ing is structured in eleven task areas. The task areas 3.2 – 3.8 are dedicated to the seven method-oriented engineering archetypes, while task area 3.9 works on the development and delivery of basic RDM services, and task area 3.10 on subject-related tasks specific to sub-communities within engineering. Finally, task area 3.11 organises the overall management of the consortium. The structure of NFDI4Ing's work programme can be seen at a glance in Figure 3.1.1 Overview of task areas.

Thus, the whole structure of NFDI4Ing's work programe is user-centred and reflects our engineering research method-oriented approach.

Task area	Measures	Responsible Co-Spokesperson(s)
ALEX "Bespoke experiments"	A-1 "Transmitting data between bespoke partial solutions"	Peter Pelz
	A-2 "Modular approach to reusable bespoke solutions"	
	A-3 "Persistant storage of medium to high data volumes with fine grained access"	
	A-4 "Core metadata model"	
	A-5 "Retroactive metadata generation for legacy data"	
BETTY "Engineering research software"	B-1 "Integrated toolchain for validated engineering research software"	Bernd Flemisch
	B-2 "Best practice guides and recommendations"	
	B-3 "Containerisation and generation of web frontends"	
	B-4 "Standardisation and automated extraction of research software metadata"	
	B-5 "Catalogue of engineering research software and validation data"	
CADEN "Provenance tracking of physical samples and data samples"	C-1 "Research best practices and IT tools landscape"	Thorsten Bronger, Britta Nestler
	C-2 "Compile and disseminate the task area output"	
	C-3 "Create tool prototypes for graph API, linked graphs, and shared access"	
	C-4 "Design and implement the graph API in existing tools suitable for archetype CADEN"	
DORIS "High performance measurement and computation with very large data (HPMC)"	D-1 "Accessibility and access rights, data security and sovereignty"	Christian Stemmer
	D-2 "Support for third-party users & community-based training, provision of post-processing algorithms and modules"	
	D-3 "Metadata definitions & terminologies, support to data-generating groups"	
	D-4 "Storage & archive for very large data"	
	D-5 "Reproducibility on large-scale high-performance systems"	
ELLEN "Extensive and heterogeneous data requirements"	E-1 "Semantic mapping of methodological knowledge"	Sören Auer, Detlef Stolten
	E-2 "Use of methodological knowledge to facilitate data exploration"	
	E-3 "Connection of data exploration and data generation processes"	
	E-4 "Community-based validation of data concepts and services"	
FRANK "Many participants and simultaneous devices"	F-1 "Target process specification"	Robert Schmitt
	F-2 "Technological feasibilty and decision-making"	
	F-3 "Design concept of an application program interface"	
	F-4 "Incentivation of an active and interdisciplinary RDM use"	
GOLO "Field data and distributed systems"	G-1 "Conception of a digital twin for the organisation and processing of research field data"	Regine Gerike, Roland Lachmayer
	G-2 "Creating a digital master of a technical system"	
	G-3 "Recommendations for creating and monitoring digital shadows"	
	G-4 "Conversion of the extended concept of the digital twin into a ready to use process"	
Base Services	S-1 "Quality assurance in RDM processes and metrics for FAIR data"	Matthias Müller, Irina Sens, Thomas

	S-2 “Research software development”	Stäcker, Achim Streit
	S-3 “Metadata and terminology services”	
	S-4 “Repositories and storage”	
	S-5 “Overall NFDI software architecture – data security and sovereignty”	
	S-6 “Community-based training on enabling data-driven science and FAIR data”	
	S-7 “Automated data and knowledge discovery in engineering literature”	
Community Clusters	CC-1 “Initialisation of community services”	Regine Gerike, Manfred Krafczyk, Christian Langenbach, Britta Nestler, Peter Pelz, Robert Schmitt
	CC-2 “Education”	
	CC-3 “Activation and Collaboration”	
	CC-4 “Partizipation”	
	CC-5 “Standardisation”	
	CC-6 “Journal”	
Management	M-1 “Effective internal and external communication”	Verena Anthofer, Annett Schwarz
	M-2 “Organising events”	
	M-3 “Interaction with other consortia and the NFDI bodies”	
	M-4 “Quality management”	
	M-5 “Fostering homogeneous project management”	
	M-6 “Financial administration and disbursement of funds”	
	M-7 “Traveling, purchasing, and contracting”	

Table 3.1.1 Overview of task areas and measures.

3.2 ALEX: Bespoke experiments (TUDA, SU, TUC)

3.2.1 Introducing the archetype ALEX

“Hello, I’m ALEX. I’m an engineer that, to investigate a technical system in regard to its process variables, develops and carries out a bespoke, one-of-a-kind experiment. The experiment may be real or virtual, using a custom tailored hardware or software system.

Examples of such technical systems are process plants, buildings, prototypes, components, control loops, mechatronic systems, algorithms, mass transport, interaction interfaces, etc.

My professional background may be based in production enginering, constructive mechanical engineering, thermofluids, energy systems, systems engineering or construction engineering.”

3.2.2 ALEX's key challenges with respect to research data management

Because ALEX pursues highly specialised, bleeding edge research objectives, she has to leverage individual, bespoke one-of-a-kind-setups of equipment, methods and interfaces. This means: Since she typically is solely responsible for a project, ALEX typically wants to know exactly how her system and experiment function. She also commonly needs to adapt the system and experiment to meet her research objective. ALEX needs highly flexible software or data models to represent and connect her experiment components, configurations and results.

Because ALEX often rightfully chooses an iterative or agile approach to reach her research objective, ALEX's need for flexibility has to consider the necessity of compatibility or reusability. She must be able to trace the data-flow and configuration of her experiment, as well as adapt and repeat it. ALEX mostly has to deal with at least medium volume of data, with no upper limit (gigabyte to terrabyte per year and project or more), therefore: ALEX's data rates are in almost all cases too high to manage the generated data manually, even when using document management or version control systems. ALEX's volumes of data are often too large to store the entirety of a generated dataset.

3.2.3 State of the art in research data management

The applicability of existing standards or toolchains for data management is often constrained to proprietary environments or specific domains outside of engineering and poorly transmissible [31], [32]. Currently most engineers of this archetype exploit their affinity and ability to develop their own approaches and solutions using programming languages. While Information describing internal workings of such programs are often well documented in the form of text, the use of existing managing solutions like version control systems could be improved. The generated data is stored in customised data schemas [33], [34]. Providing consistent documention via metadata, i.e. information about the investigated system, instrumentation or modelling setup, its calibration, and configuration, is a challenging task. Due to the absence of standardised formats and machine-actionable metadata, research software is usually developed for highly specific purposes and frequently needs to be modified or rewritten [35].

3.2.4 Key objectives

The key objectives of this task area aim to provide engineers with the means to address the key challenges formulated above, without constricting them into a monolithic service, contributing to the NFDI4Ing key objectives formulated in chapter 1.1:

- Facilitated integration of partial solutions (e.g. unified data transfer, increased compatibility and reusability) to improve the tracebility of system functions, data flow and configuration and lower the barrier to manipulate and adapt experiment setups (cf. objective 1).

- Self documenting modular software setups (including representations of hardware setups), thatu automatically generate machine actionable metadata (cf. objectives 3, 6).
- Providing means to store and retrieve specific sets of (meta-)data in medium to large sized collections in an intuitive and performant way to encourage publishing and re-using data collections and facilitate decoupling of code and data (cf. objectives 4, 8).

3.2.5 Measures

Measure A-1 "Transmitting data between bespoke partial solutions" (TUDA)

During the execution of an experiment, data needs to be transmitted between different parts of a system, each responsible for different functionality, possibly connected to hardware components. Those parts, which may be bespoke implementations based on varying programming languages or environments, need to be integrated and synchronised. Therefore, the aim of this measure is to establish a framework, interface or protocol to connect user-written custom modules or system parts.

Task A-1-1: Review existing technologies for transmitting data between separate software modules in regard to platforms, programming paradigms, language features, third party libraries and rate them according to ease of integration.

Task A-1-2: Publish guidelines on applicability of transmitting technologies per software environment, along with recommendations for selected use-case examples.

Provided with the means to transfer data between separate system parts, engineers need an intuitive interface to specify which modules they want to use, how they should be configured and connected. This also opens up the possibility to automate the generation of metadata for the whole system.

Task A-1-3: Boost the ease of use for cross-language (or -environment) module systems. To this end, exploit the interface solution developed in measure B-3 for steering and connecting modules provided as containers that include all required dependencies.

Task A-1-4: Provide interface suitable for workflows typical for ALEX in cooperation with pilot users (assisted by measure S-2) and standardise transmitted data (cf. A-4).

In addition, we will provide means to interchangeably use software modules within processing chains operating on volatile streamed as well as on stored persistent data by specifying suitable syntax on the technical (i.e. addresses and identifiers, cf. A-3) and semantic level (cf. A-4).

Measure A-2 "Modular approach to reusable bespoke solutions" (TUDA, SU, TUC)

This measure aims to provide engineers with the means and background knowledge to design partial solutions for system functionalities in a modular fashion, so that they are as interoperable and reusable as possible, without sacrificing flexibility and adaptibility. To achieve this goal,

training of engineers in required skills and suitable tools is crucial. The tasks of this measure are therefore carried out in cooperation with measures B-1 and S-2.

Task A-2-1: Evaluate popular design approaches for interoperable software regarding their degree of coupling, extensibility and complexity and publish the results as a guide with recommendations based on ease of use and limitations regarding performance or features.

Task A-2-2: Provide reference implementations, templates and standard modules for typical use-cases in cooperation with pilot users.

Task A-2-3: Implement a test suite that enables engineers to validate the compatibility of their custom-written modules automatically and reliably via unit testing.

As a prerequisite for reusability and reproducibility, it is also essential to separate code and data (i.e. configuration and parameters). We will therefore provide a lightweight model for representing configurations of hard- and software setups together with an interface for its integration into developed code, accompanied by guidelines for implementation and adaptation.

Task A-2-4: We will evaluate available exchange formats based on ease of use and industry adoption to decide which ones to support, taking into account the complexity of what a minimalistic module specification should include (cf. A-4).

Task A-2-5: We will enable tracking of provenance across modules by leveraging the solution for connecting modules developed in measure A-1, taking into account version control systems for reliable referencing of code and versions.

Measure A-3 "Persistant storage of medium to high data volumes with fine grained access" (TUDA)

In many use-cases typical for archetype ALEX engineers deal with medium to high data rates and arbitrarily complex auxillary information. A suitable storage solution therefore needs to combine performant storage of the primary data (transmitted data, cf. A-2) and flexible storage of complex metadata (cf. A-2). A solution that fulfils these requirements benefits engineers independent of archetype ALEX. Therefore, co-development together with measure S-4 and rapid transformation into a service is a high priority.

Task A-3-1: Evaluation of existing technology stacks in regard to the requirements outlined above and validation of acceptance together with pilot users. The results will be published as recommendations and implementation guidelines.

Task A-3-2: Development of reference implementations for the APIs included in the recommended technical stacks, enabling a) querying data for processing and analysis and b) harvesting & publishing metadata (integration of PID services).

Task A-3-3: Development of an easy-to-use user interface layer on top of the available technical access API together with pilot users, enabling queries according to the semantics developed in A-4.

Measure A-4 "Core metadata model" (TUDA, SU, TUC)

To provide an intuitive and unified basis to interact with data and metadata without sacrificing the engineers' ability to adapt, this measure will create a graph based representation of research processes with software modules and hardware as nodes (cf. A-2), and transmitted data as edges (cf. A-1). In order to maximise interoperability, the representation will be based on an abstract provenance model (cf. chapter 3.4).

Task A-4-1: Use existing engineering-focused ontologies as basis to define an upper ontology that is compatible with the abstract provenance ontology (cf. C-3).

Task A-4-2: Identify the common denominator of what must be documented in the context of archetype ALEX, regarding software and hardware configuration (of both instrumentation / modelling setup and unit under test) as well as the process variables under observation.

Task A-4-3: Integrate the identified core semantics into interfaces for transmitted data (cf. A-1) and software configuration (cf. A-2) as well as with the storage API (cf. A-3). Co-develop the interfaces together with measures S-3, S-4 and validate with pilot use-cases as well as standard modules of measure A-2.

Provided with a core data model that is unified across domain specific applications in the context of archetype ALEX, engineers can add domain specific, more detailed metadata. The ecosystem needed for this is part of the overall hierarchical and modular approach to model metadata pursued by measure S-3 and NFDI4Ing as a whole (cf. section 3.1).

Task A-4-4: Provision of guidelines for addition of domain specific metadata elements to the core model as well as for the curation and standardisation of metadata. The development will be performed in cooperation with pilot users and archetype ELLEN (long term data re-use, cf. chapter 3.6).

Measure A-5 "Retroactive metadata generation for legacy data" (TUDA)

Metadata records accompanying legacy data can be incomplete due to several reasons (e.g. based on an outdated standard, failure to record necessary parameter, redaction due to confidentiality, errors, etc.). To ensure that the corresponding legacy data remains usable, at least in the research group that produced it, this measure aims to reconstruct incomplete metadata by using information from textual sources like published articles or reports.

Task A-5-1: Evaluation of available methods and tools for data mining regarding their applicability for the objective described above (in cooperation with measure S-7).

Task A-5-2: Creation of synthetic test-cases of incomplete metadata for the evaluation process (based on use-cases of measures A-1 through A-4 where the complete metadata is known), examining the following scenarios: a) random missing information due to user error; b) migration of existing but incompatible metadata based on outdated standard; c) parts of metadata deliberately hidden because of confidentiality.

Task A-5-3: Validation of the approach against real world use cases of incomplete metadata from pilot users and dissemination of success stories within community clusters.

Possible risks of implementation

Several measures of this task area rely on close collaboration with measures of other task areas. The expertise and resources of the NFDI4Ing management will be leveraged to ensure the corresponding tasks are coordinated effectively.

3.2.6 Synergies and demarcations with regard to other task areas

The standardised and generally facilitated dissemination of software modules together with the necessary environments is (B-1, B-3) supports the modular approach to reusable bespoke solutions (A-2). This lays a foundation for the integration of standardised data transferring mechanisms (A-1) and directly supports the development of best pratices for interoperable software modules (A-2), further facilitated through training of engineers in required skills and suitable tools (S-2). By automating the documentation of setups and systems generating data (A-1), this information is easily made available leveraging flexible storage with fine grained access (A-3) through metadata harvesting (S-4), which supports web-based search and selection of research data and software (E-3). The systematisation of data quality criteria and metrics (S-1) enables identification of methodological knowledge hidden in conventional publications (E-2, S-7), which can be leveraged in the reconstruction of incomplete metadata (A-5). The creation of a core metadata model for the context of bespoke experiments (A-4) supports the development of a terminology service to access, curate and update terminologies (S-3). The development of a semantic framework (E-1) and APIs for linked graphs (C-3) will support its interoperability beyond this context, and facilitate polishing of user friendly query interfaces for storage technologies (A-3). The contributors as well as the participants and pilot users provide a strong connection to several communities (cf. DFG 41, DFG 42, and DFG 44 respectively). This ensures that the users maintain an active role over the long term, their needs are addressed appropriately and acceptance in the whole community is fostered.

3.3 BETTY: Engineering research software (US, TUC, CRC 1194, SE²A, SimTech)

3.3.1 Introducing the archetype BETTY

"Hello, I'm BETTY. I'm an engineer and self-taught programmer that develops research software. Very often, this software represents a computational model for the simulation of an engineering application. For validating such a model, I have to compare my results with data such as other simulation data or experimental observations. For this and other purposes, I also write code for analysing and converting research data. My software usually has a lot of dependencies in form of the operating system and third-party libraries. While I'm very keen on guaranteeing the reproducibility of my computational results, I can't dedicate too much working time to achieve this. My professional background can be located in any engineering discipline."

3.3.2 BETTY's key challenges with respect to research data management

As defined in [36], "*research software* (as opposed to simply *software*) is software that is developed within academia and used for the purposes of research: to generate, process, and analyse results. This includes a broad range of software, from highly developed packages with significant user bases to short (tens of lines of code) programs written by researchers for their own use." We address this whole range and in particular the fact that the software usually is developed by domain specialists rather than software engineers. While research software certainly is research data, it exhibits particular characteristics compared to more "conventional" data. This yields the following challenges:

- Code inside a research software project that is under continuous development is subject to permanent change. Employing version control is mandatory to keep track of changes and to provide means to refer to a particular code instance.
- A particular piece of research software often exhibits complex dependencies on other software such as third-party libraries and operating systems as well as possibly on the hardware architecture on which the software is supposed to be executed.
- The broad variety of engineering applications is reflected in a huge number of engineering research software projects exhibiting very different scope, size, quality etc. This demands correspondingly adaptable RDM processes and tools.
- Engineering research software is often dedicated to numerical simulation. To ensure the quality of the software, the underlying model needs to be validated, which in turn usually requires the comparison of simulation results with other data. Analogous validation requirements hold for other types of software such as software for systems control.

- Software itself generates data in form of computational results. While the results are commonly discussed in a scientific publication, a proper RDM strategy for research software should aim for ensuring the reproducibility of these results.

As described below, a broad range of tools already exists to tackle several technical challenges listed above. The fundamental challenge within this task area rather is to offer the engineering community a consistent toolchain that actually will be employed in daily engineering research.

3.3.3 State of the art in research data management

Many tools already exist to facilitate the development of research software and to foster the FAIR principles of software treated as data. In particular, findability and accessibility can be achieved in terms of publicly available software repositories. Today, usually Git is employed for the version control itself and a version control management system such as GitLab [37] is used on top. Concerning metadata, no standards for research software have been adopted or developed by the engineering communities so far. Not even the citation of research software is standardised today; several efforts are ongoing [38], [39]. As outlined above, interoper-ability and reusability might be particularly difficult to achieve due to complex dependencies. Containerisation, the act of packing an executable together with its complete software environment, is a promising measure to cope with this issue. Docker [40] and Singularity [41] are commonly employed tools for creating containers. However, the sole provision of a container doesn't guarantee reusability, as it might still be infeasible to install the container locally and run the executable inside. An attractive solution to this can be a web frontend that allows to steer the container on a corresponding backend. JupyterLab [42] and JupyterHub [43] are a prominent example for developing and hosting such frontends and backends. Several best practice guides on research software development exist [44], [45], but they usually don't target engineers and their specific requirements and skill sets, with exception of [46].

3.3.4 Key objectives

From the abovementioned key challenges and state of the art in RDM activities, we identify the following three key objectives for this task area. In particular, substantiating our common objectives 1, 2, 4, and 7 for the case of research software, we would like that *every engineer*

1. can be equipped with the tools and knowledge that are necessary and useful to develop validated quality-assured engineering research software,
2. is able to guarantee the reproducibility or at least the transparency of his computational results and to provide his peers with usable solutions for the actual reproduction,
3. can easily equip his own engineering research software and validation data with standardised metadata and find such software and data of others for his research.

3.3.5 Measures

We implement five measures to achieve the abovementioned goals. Measures B-1 and B-2 target primarily the first goal, measure B-3 the second, and measures B-4 and B-5 the third one.

Measure B-1 "Integrated toolchain for validated engineering research software" (US, TUC, CRC 1194, SE²A, SimTech)

As outlined above, several solutions for individual RDM measures in the context of software exist already. By integrating these tools, we will implement and support a consistent toolchain for developing validated engineering research software. We will perform the following tasks:

Task B-1-1: Integrate individual software-related RDM services into a toolchain that accompanies the daily engineering research work, including version-controlled code development as well as automated testing. Start with the tools at hand and adapt to the ones developed in measures S-2 and B-3 once they become available.

Task B-1-2: In the context of software and particularly model validation, facilitate the required comparison of data that is possibly distributed over different repositories with the help of measures S-4 (interoperability of repositories) and S-5 (authentication).

Task B-1-3: Follow a bottom-up approach by developing and adapting the required components in collaboration with several pilot users. Design use cases with the selected participants in terms of individual projects, each lasting for about one year. Create success stories that will attract more users via dissemination by means of the community clusters (cf. chapter 3.10).

Measure B-2 "Best practice guides and recommendations" (US)

In this measure, we aim to design best practice guides and recommendations for developing engineering research software by means of the toolchain that results from the tasks of measure B-1, starting from the first lines of code to the provision of web frontends and backends for the reproduction of computational results. The associated tasks are:

Task B-2-1: Draft the guides and recommendations by evaluating pilot user experiences and by involving the RDA Interest Group RDM4Eng (cf. chapter 1.4).

Task B-2-2: Integrate individual components such as the best practice guide for continuous integration (CI) developed in measure S-2. Incorporate the results of measure S-1 on the RDM maturity model and archetype-specific DMP templates.

Task B-2-3: Coordinate and author a white paper on developing engineering research software, describing the results achieved in this task area as well as by measure S-2.

Task B-2-4: Develop a wiki that can be extended by the community and ensure that it is adopted by a wider audience by involving the community clusters (cf. chapter 3.10).

Task B-2-5: Integrate software-related issues in subject-specific recommendations for handling research data that are to be developed for the DFG (cf. chapter 3.10 measure CC-5).

Measure B-3 "Containerisation and generation of web frontends" (US, TUC, CRC 1194, SE²A, SimTech)

Within this measure, we focus on two main ingredients to reach the second goal of enabling every engineer to guarantee and facilitate the reproducibility or at least the transparency of computational results: containerisation and generation of web frontends. With Docker and Singularity regarding containerisation as well as JupyterLab concerning web frontends, the technical tools are readily available. The individual tasks are:

Task B-3-1: Define workflows for the pilot use cases from an executable within a local software and hardware environment to a possibly public web frontend. As first step, this involves containerising the complete software stack, while the second step is the frontend development.

Task B-3-2: Automate, generalise and standardise the two steps listed above in close collaboration with measure S-2.

Task B-3-3: Develop solutions for software stacks containing proprietary components that prohibit containerisation or distribution. Achieve maximum transparency by record-ing input to and output from such components as well as all necessary metadata.

While the selection of the backend that a generated web frontend is supposed to work on should be flexible, this measure will be complemented by the provision and administration of a JupyterHub server hosted at the TIK Computing Center Stuttgart as part of measure S-2.

Measure B-4 "Standardisation and automated extraction of research software metadata" (US, TUC, CRC 1194, SE²A, SimTech)

Up to now, no standardised metadata format for engineering research software has emerged. CodeMeta [47] defines subsets of the schema.org [48] vocabulary as possible entries for the description of software. However, it is discipline- and even research-agnostic and doesn't include any engineering-related entries such as the software's application areas or details on a computational model. On the other hand, there exist recent efforts on metadata formats for engineering sciences with no particular focus on software [49]. The tasks in this measure are:

Task B-4-1: Combine the two efforts mentioned above in order to arrive at a metadata standard for the description of engineering research software. This will be undertaken in close collaboration with measure S-3.

Task B-4-2: Facilitate the annotation of research software with metadata by employing automated extraction from the underlying Git repositories. A corresponding service will be developed by measure S-3. Assist in its improvement by application to the pilot use cases and by employing it for task B-5-1 described below.

Measure B-5 "Catalogue of engineering research software and validation data" (US)

In order to help an engineer to find software that is supposed to be capable of solving his current research question, we aim to establish a catalogue of engineering research software by means of the following tasks:

Task B-5-1: Assemble the catalogue by employing the metadata harvester developed in measure S-3. Start with the projects on the GitLab instance provided by S-2 and successively widen the evaluation base with help of the community clusters (cf. chapter 3.10).

Task B-5-2: Annotate the catalogue with the code quality measures developed in measure S-2. An example of such a catalogue is given by the OpenHub [50] project which is, however, discipline-agnostic and not focusing on research software.

Task B-5-3: Especially for simulation software, also list research data that have been employed to validate a corresponding computational model. This will become particularly useful for researchers developing new models and searching for validation data.

Possible risks of implementation

In several measures of this task area, services resulting from measure S-2 are to be employed. Before their provision, we can build upon alternatives available at the University of Stuttgart which might be limited in features and usability but are sufficient for prototypical developments.

3.3.6 Synergies and demarcations with regard to other task areas

Since research software is an integral part of today's engineering research, this task area is naturally related to several other task areas. Most notably, it is strongly connected to measure S-2 "Research Software Development" of task area Base Services. While S-2 is concerned with the development, deployment and provision of services related to research software, this task area here will integrate these and other services to reach the key objectives listed in 3.3.4 with a consistent toolchain that is employed in daily engineering research work.

Another strong connection is given to task area DORIS (cf. chapter 3.5) concerning research software for HPMC. Both areas can be demarcated by the fact that task area BETTY focuses on the treatment of software as research data and the associated development process, while task area DORIS (cf. chapter 3.5) is dedicated to research data resulting from the application of software on HPC architectures. Both areas overlap when it comes to testing and reproduction. Here, challenges resulting from the fact that a test or reproduction run has to be performed on a dedicated HPC architecture will be dealt with by task area DORIS (cf. chapter 3.5) together with S-2.

Task areas ALEX (cf. chapter 3.2) and FRANK (cf. chapter 3.7) also deal explicitly with research software. Synergies can be expected by supporting these task areas with the undertaken measures and in turn receiving use cases from them. Clear connections can also be drawn to

task area ELLEN (cf. chapter 3.6), as measures B-1 and B-3 facilitate the handling of software programs combined to workflows and particularly the employment of web frontends there. As already mentioned in the measure descriptions, connections are also given to several other measures of task area Base Services, particularly S-3 on terminology and metadata, S-4 on integration and interoperability of repositories, and S-5 on authentication mechanisms. Finally, we expect to attract new users and improve the developed tools, services, and documents by close collaboration with the community clusters (cf. chapte 3.10).

3.4 CADEN: Provenance tracking of physical samples and data samples (FZJ, KIT)

3.4.1 Introducing the archetype CADEN

"Hello, I'm CADEN. I'm an engineer whose research deals with complex sequences of processing and analysing steps, applied to samples and/or data sets.

My professional background is mostly informed by materials science, building materials science, materials technology, process engineering, and technical chemistry.

For me, the output of one processing step is the input of a subsequent one. The processed objects can be physical samples (specimens) or data samples. For instance, I synthesise an alloy sample, temper, and etch it. After that, I analyse the sample in various measurement setups, creating data sets. Thus, in the context of my data management, physical and data samples are treated the same and both are referred to as *"entities"* in the following. Similarly, all kinds of processing steps, whether working on physical or data samples are referred to as *"activities"*. The resulting *graph* of entities and activities comprises my workflow."

Figure 3.4.1 illustrates such a graph. Here, the entities are boxes and activities are arrows. All entities with dashed frames are lost (i.e., samples in a state that was irreversibly changed by processing).

In this graph, a sample is processed, split into two pieces, one of which is processed further. Then, both pieces are measured and the results compared. Dataset #3 is the result of this comparison.

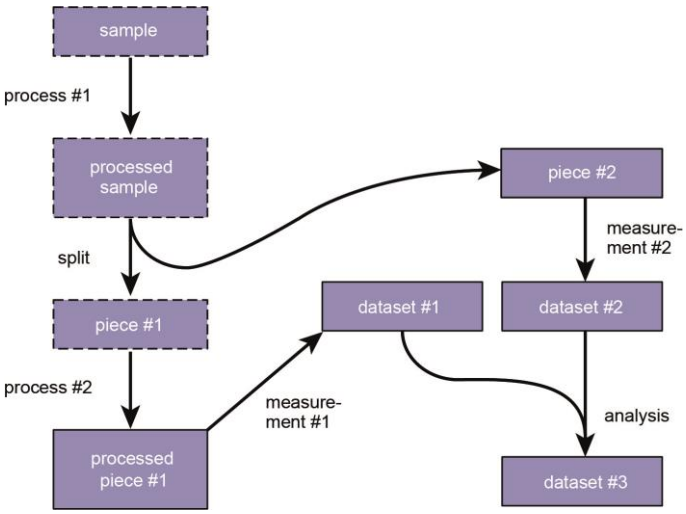


Figure 3.4.1 Provenance graph in the sample-based workflow

3.4.2 CADEN's key challenges with respect to research data management

In archetype CADEN, we need to be able to track the provenance of samples and data. The key challenge is to store the following in a well-structured way:

- all data entities; here, both data and metadata must be stored,
- all parameters of the activities (e.g. temperatures, pressures, simulation parameters),
- the graph topology.

The structure of the data and the management software needs to allow:

- querying the graph data,
- extracting the graph data for further processing,
- linking to entities and activities using globally unique persistent identifiers (PIDs).

The graph in archetype CADEN may be highly individual and non-linear (i.e., contain forks and junctions). This leads to complex data structures in the data management.

Another challenge for us in this archetype are collaborations, which usually preclude shared data management: So far, institutions store metadata isolated from each other, and a machine-actionable link between the fragments of the graph is not available.

3.4.3 State of the art in research data management

Typically, archetype CADEN addresses their challenges with idiosyncratic methods for data management, often using inadequate standard office software. Off-the-shelf ELN software used by archetype CADEN (e.g. [51], [52], [53]) scarcely fulfils any of the key challenges. Moreover, archetype CADEN does not benefit from best practices of others with similar workflows. As far as RDM is concerned, the community is poorly connected.

In 2017, the Research Centre Jülich interviewed 15 institutes of archetype CADEN (engineering and natural sciences) on its campus about their RDM. [54] Only one of them used an institute-wide tool capable of managing provenance tracking. None of the interviewed researchers made use of external data management competence (best practices, tool recommendations, trainings). However, all of them regarded such a tool as well as knowledge about best practices as beneficial. Similarly, a survey at 2019 amongst 12 archetype CADEN groups at KIT [55] revealed that only 3 used an ELN while 8 used exclusively network shares or USB sticks for internal data exchange.

3.4.4 Key objectives

In this task area, we address our common objectives 1, 3, 5, and 7 from chapter 1.1 with the following key objectives. Given the size and complexity of our community, we formulate them with the principle of help for self-help in mind.

The first key objective is promoting awareness for best practices and tooling for friction-less digital workflows. For instance, this includes propagating the various methods for getting experimental data into a graph-managing RDM tool. On the organisational side, it includes advice on effective collaboration with working group colleagues as well as with external peers on the same provenance graphs. We believe that a comprehensive RDM handbook dedicated to archetype CADEN is the most effective framework for disseminating detailed success stories and beneficial use of IT solutions. Besides, a handbook allows for informational depth which is not feasible in RDM trainings and workshops, thus complimenting cross-cutting awareness means.

The second key objective is the harmonisation of the interfaces of our tools, addressing the challenges of querying the data and of sharing data in collaborations. This comprises the possibility to access the graph data via an API so that third-party software can reuse the data, and PIDs for each part of the graph in order to enable linked data even between institutions. We will help our community to realise this goal by working out data structures and protocols for it, and by providing reference implementations.

3.4.5 Measures

Measure C-1 "Research best practices and IT tools landscape" (FZJ, KIT)

In order to fulfil the archetype CADEN desideratum for a comprehensive compilation of best practices, research effort is necessary. For this, we will screen the publications already used to determine the state of the art to extract archetype CADEN success stories, and extend this effort to non-German publications. Moreover, this screening lets us identify partners for own structured interviews with scientific engineering work groups in Germany that match archetype CADEN. Further, eligible work groups will be found amongst the partners of NFDI4Ing and the parties that declared interest in NFDI4Ing by submitting survey answers, of which 275 consider archetype CADEN relevant for their work. [C7] Close interaction with community cluster 43 (materials science and engineering) ensures that our efforts have good coverage of the community. The interviews will be followed by workshops with engineers for formulating novel best practices for archetype CADEN that are recommended to be deployed in the community.

Additional research will identify freely available (not necessarily open-source) RDM tools that are useful for archetype CADEN. This includes input from the interviews and workshops. We will compile a report that allows engineers evaluating and comparing the tools. This work depends on task C-2-1 and thus takes place after that.

Task C-1-1: Analyse published interviews with scientists, and perform and analyse own interviews.

Task C-1-2: Hold workshops for formulating best practices.

Task C-1-3: Compile a report about existing IT tools for archetype CADEN.

Measure C-2 "Compile and disseminate the task area output" (FZJ, KIT)

By reviewing the results of C-1-1 and C-1-2, we will characterise archetype C so that the reader is able to determine whether their methods (or parts of them) are represented accurately by it.

The RDM handbook for archetype CADEN will consist of the following parts:

- Description of various forms of this archetype, letting the reader identify whether their workflow (or parts thereof) falls into this category.
- Presentation of best practices for this archetype, with a clear focus on everyday tasks.
- Presentation and discussion of existing IT tools supporting this archetype, with a clear focus on free software.

By identifying success stories in the results of C-1-1 and C-1-2, we compile best practices for archetype CADEN, partly in the form of success stories, partly as guidelines. We will enrich this with well-established RDM concepts like FAIR principles and data management plans.

The outcome of C-1-3 is fleshed out into a detailed discussion which tool or combination of tools enhances productivity for the sub-profiles of archetype CADEN. Besides an annotated tabular summary of feature coverage, we demonstrate how the success stories can be supported by tools. We consider this connection of success stories and tool recommendations a substantial benefit of our approach to a compilation of best practices.

All textual output of this task area (handbook, results of the pilots, API documentation) will be created using interoperable text-based markup formats and published with a permissive licence on an NFDI4Ing GitLab instance. We will act analogously for our code. The associated issue trackers will be advertised via our community clusters as a feedback channel for comments and suggestions from the community. This way, we want to make particularly the handbook a living document with continual updates and improvements.

Writing documentation needs a lot of effort for the engineer. On the other hand, tool and process documentation is necessary for self-reflection and frictionless re-use. The current tooling landscape focuses on the documentation for developers [56], [57], [58]. By contrast, in this task, we focus on the toolchain for engineers to write handbooks that enable the reader to make use of the cutting edge features of research software and make use of this in C-2-4.

Task C-2-1: Compile the characterisation of archetype Caden.

Task C-2-2: Compile the best practices.

Task C-2-3: Compile the description and discussion of the tools.

Task C-2-4: Disseminate the task area output and establish a workflow for continual updates.

Task C-2-5: Develop toolchains to ease writing documentation.

Measure C-3 "Create tool prototypes for graph API, linked graphs, and shared access" (FZJ, KIT)

For prototyping the graph API and PID tagging, we plan to build on existing, mature, free, and open-source tools. Examples are IFFSamples [59], JuliaBase [60], Kadi4Mat [61], and repository technology (together with S-4). We will evaluate further tools from C-1-3, however, available competence for the software within NFDI4Ing members is an important criterion. Another one is that the software used later in the pilots will need to use the graph API.

We will extend the tools with PIDs for all entities and activities as well as an export of graphs (or parts thereof) through a web API. Thus, we will not create software from scratch but design a thin layer on top of mature code. We will evaluate existing API technology: Simulation science provenance APIs like that of AiiDA [62] may be a blueprint for our own. Critical is here the adaptability of a data-only API to a sample-based workflow, and the consideration of existing standards such as PROV [63], RDF DCV [64], and GQL [65]. For the underlying ontology, we plan close collaboration with the task area Base Services (measure S-3 "ontologies & metadata service") and with the consortium NFDI4MSE.

We will deploy the result of C-3-2 in multiple instances, located at different collaborating work groups that are willing to test the prototype. As one pilot we will share data between experimental working groups (IAM-ESS/KIT) and simulation-based working group (IAM-CMS/KIT). Further pilots are the multiple departments spanning battery research at IEK/FZJ, as well as the existing collaborations between the institutions. An open design process and continual user feedback allow for efficient tool optimisation. When implementing the shared access for all participants on all instances, we anticipate a project risk of considerable technical (e.g. shared login) and organisational (e.g. institutional access policy) challenges. We will build on the results of archetype FRANK (F-3) to handle these challenges. Finally, with the support of S-7, we will import legacy data into the central tools.

Task C-3-1: Evaluate candidate tools for the prototypes.

Task C-3-2: Design and implement the graph API and PID tagging.

Task C-3-3: Test, evaluate, and optimise the implementation with cooperating scientific work groups.

Measure C-4 "Design and implement the graph API in existing tools suitable for archetype CADEN" (FZJ, KIT)

We will identify those mature tools that are suitable for the extension with the graph API and PIDs. For this, we will analyse the data model in the respective software and estimate whether it can be mapped onto a graph. Some data repository software may be eligible, too.

For each tool identified in C-4-1, we will design a concept for the API extensions and send it to the respective software maintainers with the kind request to do the actual implementation.

Incentives for the maintainers are the proven benefit of the prototype while the necessary effort is manageable. This approach obviously bears a serious project risk: The probability is considerable that the software maintainers are not prepared to do the necessary code changes. We will address this risk by prioritising the tool list for archetype CADEN, and supporting external maintainers by providing a prototype implementation and wrapper scripts to access the API. We consider NFDI4Ing as a catalyst for the adoption of a common graph API, using our resources to trigger change rather than making it fully ourselves.

Task C-4-1: Compile a list of eligible existing tools.

Task C-4-2: Conceptualise an API extension to each of the tools and communicate it to the respective maintainers.

Possible risks of implementation

We identify two risks, both due to dependencies on external stakeholders: C-3-3 may be delayed or hindered due to institutional objections to a shared access for external groups. And, C-4-2 may not be applicable to all desired tools because the respective developers do not cooperate.

3.4.6 Synergies and demarcations with regard to other task areas

Archetype CADEN has partial overlap with archetype ALEX (cf. chapter 3.2), which needs to track data flows, too. By contrast, the graphs of these data flows do not vary strongly between experiments. Instead, different input parameters and variations of the setup itself are in the focus of research. Besides, provenance graphs do not span multiple institutions. Still, we will take input from the task area for archetype ALEX (cf. chapter 3.2) to optimise also the overlap domain in order to support that archetype.

Furthermore, we depend heavily on the task area Base Services. There, the ontologies and metadata services to be created (S-3) will form an integral part of our graph API. This will be a bi-directional process, as the community workshops (C-1-2) provide the underlying vocabulary to S-3, which then translates this into a formal ontology for C-3-2. As we will generate significant awareness and training material in C-2, we will provide S-6 with content about provenance tracking and sample-based research workflows.

3.5 DORIS: High-performance measurement and computation with very large data (HPMC) (TUM, RWTH, LRZ, HLRS)

3.5.1 Introducing the archetype DORIS

“Hello, I’m DORIS. I’m an engineer conducting and post-processing high-resolution and high-performance measurements and simulations on High-Performance Computing systems (HPC). The data sets I work with are extremely large (hundreds of TB or even PB) such that they are, by and large, immobile. They are too large to be copied to work stations and the (post-) processing of the experimental and computational data generally is done on HPC systems. The HPC background mandates tailored, hand-made software, which takes advantage of the high computational performance provided.

The data sets accrue in the combustion, energy generation and storage, mobility, fluid dynamics, propulsion, thermodynamics, and civil engineering communities.”

3.5.2 DORIS’s key challenges with respect to research data management

DORIS faces the key challenge that data from high-performance measurement and computation (HPMC) applications, which form the base of publications, are not accessible to DORIS as they are too large to be included in state-of-the-art repositories (objective 4). The HPMC data is currently stored and archieved at HPC centres in the personal account of the data generator. The data neither is documented nor are metadata sets available, as the semantics for HPMC in the engineering sector still needs to be adapted to HPMC applications. Due to the lack of local access models at HPC centres and the huge data size (into the PB-range), data is generally immobile. The very specific hardware and software environment at HPC systems make it time-consuming to get effective access to the data. Reproducibility of data beyond the life cycle of HPC systems is still an open issue as software container performance is much too low to be able to reproduce HPC results in an acceptable time frame.

3.5.3 State of the art in research data management for DORIS

The data sets from HPMC projects are often used as reference (benchmark) data for modellers and experimentalists. As per status quo, to gain access to the data as a third-party researcher, one has to contact the research group that generated the data and has to figure out a method to access or obtain the data. Common search algorithms and local repositories will be able to find metadata-like keywords in the publications. Data DOI are supplied to smaller data sets in

connection with the publications' DOI. The multitude of data formats, hardware and software stacks at the different HPC centres further complicate the effective usage of the huge data sets.

Version control of the constantly changing HPC and post-processing codes is currently ensured through the usage of GitLab [66] and GitHub [67]. The progress in the major code lines, as well as the experimental project progression with links to the respective data can be realised, e.g., with TUM workbench [68], which serves as an electronic lab book. The usage of software containers (e.g. a docker [40]) in the HPMC community is not feasible at the moment as HPC environments don't allow dockers due to data security issues (the docker provides root access to the machine) and poor performance on HPC systems.

3.5.4 Key objectives

Making data reusable by other research groups in order to validate their modelling approaches and their chosen methods requires heterogeneous solutions, which will be provided in task area DORIS. Engineers rely on high-fidelity, high-resolution data to compare their novel modelling approaches in order to verify and validate the proposed method. The development of specific metadata standards specific to HPMC applications in close cooperation with the task area Base Services group on metadata is an important keystone to RDM in this area. Software modules simplifying access to the data and high IO-speeds for the large data sets on the specific platforms will be developed and provided to the community. Showcase software modules for post-processing algorithms enable other research groups to speed up the time needed to get a start on reusing the single large data sets. Access to the data stored at the HPC centres for non-HPC customers will be realised. The future widespread accessibility of these unique data sets will further spur machine learning, neural network and AI research on these complex engineering problems fostering the effort to acquire knowledge from data. A workflow shift from 'code development ➔ code validation and verification ➔ data generation ➔ data post-processing' to 'new machine learning/neural network approach ➔ data reading ➔ learning from data' will open new possibilities in emerging data-to-knowledge research branches.

An effort in the framework of NFDI4Ing is geared towards broader usage of these very large single data sets. Therefore, a certain number of key objectives have to be addressed in the measures described below: 1) accessibility, access rights (F**A**IR), 2) support for third party users (FAI**R**), 3) metadata definitions & terminologies (**F**AI**R**), 4) reusability (FAI**R**), 5) reproducibility on unique high-performance systems, 6) storage & archive, 7) provision of post-processing algorithms and modules, 8) support to data-generating groups, 9) data security and sovereignty, and 10) community-based training.

3.5.5 Measures

Measure D-1 "Accessibility and access rights, data security and sovereignty" (TUM, RWTH, LRZ, HLRS)

The very large data sets described in the introduction are stored at the infrastructure participants LRZ, HLRS, FZ Jülich (Gauss Centre for Supercomputing - GCS) and with Gauß-Alliance (GA) computing centres in i.a. Darmstadt, Erlangen, KIT. Access to the data up to now is very restricted and the respective data 'belongs' to the user in terms of Unix access rights. In close cooperation with the HPC centres as the providers of computational time and storage space, new user models for pure post-processing projects on other people's data will be developed. The role of access rights will be devised under consideration of data security and data sovereignty on HPC systems, which provide a high level of security. Direct cooperation with task area BETTY and ALEX will be fruitful in the development of solutions in the measure. Measure S-2 on research software development will interact with this measure. A license system (like creative commons) has to be evaluated to ensure that the data is properly cited and that the results obtained and published by the third-party data-analysing groups refer to the original data set. Close cooperation with measure S-5 in the task area Base Services will accompany this measure.

Task D-1-1: Develop new user models with HPC centres to reuse very large data sets.

Task D-1-2: Devise appropriate access rights to other researcher's data for HPC.

Task D-1-3: Mechanisms to maintain data security and sovereignty on HPC systems.

Task D-1-4: Development of a front end to apply for and grant access to data, logging of access to the data, consent to a (creative commons) license system, documentation.

Measure D-2 "Support for third-party users & community-based training, provision of post-processing algorithms and modules" (TUM, RWTH)

The full benefit for the NFDI effort in the context of HPMC with very large data lies in the usage of previously inaccessible data through third-party research groups. Currently, personal interaction on a 1:1 base is necessary to introduce the data-reading group to the procedures of the data-writing algorithms as there is no common data-format standard in this community. A common data-format standard would hamper IO-speed in an unacceptable manner. Therefore, the expertise of how to effectively read the very large data with the highest IO-performance (i.e. code modules) on the various HPC-systems will be documented and provided as a service to the data-reading groups. Reading and processing the data has to be done on the HPC systems due to their sheer size. The implementation of post-processing modules depends on the precise implementation. Therefore, the exact code module used for the transformation of the highly resolved temporal and spatial data into published, condensed graphs and figures representing mathematical and physical correlations will be provided as ready-to-use modules. The same holds for algorithms extracting dynamical features on a fully time- and spatially-resolved level. The

modules and measures developed with pilot users will be brought to the HPMC community also in collaboration with the community cluster “thermal engineering and process engineering”.

Community-based training in regular biannual workshops at changing locations will disseminate the progress and measures to the communities (with S-6 and CC-2). The benefit of the measures to the user and the scientific community will be shown in examples to increase the acceptance of and the contribution to the developments created in NFDI4Ing.

Task D-2-1: Community-based training in biannual workshops updated with the progress of NFDI.

Task D-2-2: Support third-party users to effectively read the very large data sets.

Task D-2-3: Development and publication of post-processing methods with pilot users.

Measure D-3 “Metadata definitions & terminologies, support to data-generating groups” (TUM, RWTH)

For HPC users, not only the physical relevant parameters and procedures have to be included in the metadata but also information on how the data were exactly generated (exact specifications of the numerical method, boundary conditions, initial conditions, etc.). To reach maximum performance on the HPC systems, vendor-specific software on many levels (IO, optimisation directives and procedures, memory-specific access, etc.) as well as special programming techniques are used. Together with the task area Base Services (S-3), we will develop metadata standards enabling other researchers to find and interpret the data. The issue of data quality will be addressed in this measure (with S-1) as well as the provenance of the methods used. In cooperation with pilot users, best-practice guidelines on metadata and terminologies will be compiled.

Task D-3-1: Develop terminologies for high-performance numerical methods.

Task D-3-2: Define and disseminate metadata standards for HPMC environments.

Task D-3-3: Support to data-generating groups in enabling the provision of modules accelerating the effort to access the very large data sets by third-party users.

Task D-3-4: Develop and prepare best-practice guidelines with pilot users.

Measure D-4 "Storage & archive for very large data" (TUM, RWTH, LRZ, HLRS)

One of the central goals of the task area "HPMC with very large data" is the reusability of the data by research groups that have not been involved with the generation of the data. The large-scale benchmark simulations and measurements are highly valuable to the community and beyond. Together with the involved service and infrastructure providers (the Gauss Centres LRZ, HLRS, FZ Jülich, and other GA members), methods and capabilities will be developed to store and archive these large data volumes beyond a ten-year period for broader access. Front-end capabilities will be implemented providing searchable and accessible metadata information together with persistent URLs and DOIs. These fundamental means to make data accessible are considered and further developed in measure S-4 "Repositories & Storage".

Task D-4-1: Provide long-term archiving solution with a searchable and indexable front end.

Measure D-5 "Reproducibility on large-scale high-performance systems" (TUM, RWTH, LRZ, HLRS)

Reproducibility on large-scale high-performance systems is an unsolved problem. Unlike in experiments, a standard is missing that would describe the level of reproducibility, which in turn may depend on the hardware and software platforms used. This conflict will be addressed defining minimum standards for reproducibility issues. Together with the system providers, the evaluation of software containers will be evaluated for HPC environments, which at the moment is not feasible due to the low performance of the containers. Best practice guidelines for reproducibility will be developed for HPMC users (S-6 and CC-2). Direct cooperation with task area BETTY and CADEN, as well as task area S-2, is going to facilitate community-wide minimum standards.

Task D-5-1: Develop standards on reproducibility on HPC systems.

Task D-5-2: Prepare best-practise guidelines on reproducibility issues for HPMC users.

Task D-5-3: Evaluate the feasability of containers/dockers for reproducibility for HPC systems.

Possible risks of implementation

The realisation of the measures depends in parts on the cooperation in measure D-5 addressing reproducibility issues will generate technical feasibility of todays and future HPC-systems at the TIER-1 national GCS. Therefore, the HLRS and LRZ have started their participation in a LoC to show the willingness to mutually develop practical solutions to the stated challenges. The acceptance of the measures with the heterogeneous HPC-user community will be a keystone in the community. Community dialogue, information and meetings are therefore indispensable.

3.5.6 Synergies and demarcations with regard to other task areas

The metadata aspect in the task area will be discussed and developed together with the university libraries and repository providers which are engaged, i.e., in the working group metadata4ing and in the respective service area S-3. Synergies within NFDI4Ing will mainly (but not exclusively) be

on the topics of task area BETTY (cf. chapter 3.3), task area ALEX (cf. chapter 3.2) as well as task area CADEN (cf. chapter 3.4)

To effectively realise D-3, metadata services (S-3), as well as the aspect of data quality, data security and sovereignty (S-5) will be addressed in unison with the task area Base Services. The measure dealing with storage issues and repositories will be developed hand in hand with S-4. As the community-based training D-2 will share many communalities across the task areas in the engineering community, measure S-6 will be directly involved. The community clusters activities (cf. 3.10.2) especially in the thermal and processing engineering CC (DFG research area 42), will be aligned with community meetings in task area DORIS. Measure D-5 addressing reproducibility issues will generate best practice guidelines for HPMC users (S-6 and CC-2). In order to make the data traceable, algorithms developed in S-7 will be tested on HPMC publications and data.

Cooperation with other consortia as NFDI4Chem and NFDI4Phys as well as FAIRmat and Astro-NFDI are fruitful to tackle problems regarding high-performance computational and measurement methods in an interdisciplinary manner.

3.6 ELLEN: Extensive and heterogeneous data requirements (FZJ, TIB)

3.6.1 Introducing the archetype ELLEN

"Hello, I'm ELLEN. I'm an engineer who analyses complex systems comprising a large set of multidisciplinary interdependencies. Working within the computational sciences, I do not work in labs, but exclusively on computers and computing clusters. I conduct research by performing model-based simulations and optimisation calculations, whereby I often utilise algorithms coming from statistics and computer science. Inputs to my analyses are the scenarios I investigate. They typically are very data-intensive requiring information from many different disciplines, such as politics, business economy, jurisdiction, physics, chemistry, demography, geography, meteorology, etc. My professional background is typically based in electrical, chemical or energy systems engineering and is often complemented by several aspects of computer science, physics, and economics."

3.6.2 ELLEN's key challenges with respect to research data management

To gather all the information ELLEN needs from the many heterogeneous disciplines, she has to identify and investigate many diverse data sources, which are not homogeneously represented, not interconnected and not searchable over a common interface. The types of data sources are heterogeneous and range from databases containing articles over repositories providing research

data to individual websites of institutions, agencies or companies publishing surveys, statistical or historical data, or merely individual parameters. Since many of the data sets ELLEN requires quickly become obsolete, she regularly has to search for updates, whereby the best source for one distinct piece of information changes periodically. She mostly has to draw certain facts from different sources, in order to derive more reliable forecasts and estimations. In addition, the general striving for ever more precise statements with higher spatial, temporal, and content-related resolutions leads to constantly increasing data requirements. Therefore, the satisfaction of ELLEN's information needs takes up a significant portion of her daily working hours.

In case ELLEN is not able to find the specific data she needs (cf. Figure 3.6.1), she looks for similar data (B). If no suitable data is available, she has to generate the required data herself, preferably using scientifically recognised methodological concepts already implemented by colleagues from the respective research disciplines (C). If she cannot find usable and comprehensible model implementations from other scientists that serve her exact needs, she sometimes conceptualises and writes the necessary software herself (D). The more data ELLEN cannot retrieve through any of these procedures, the more she is forced to resort to inexact estimates and assumptions, which limit the reliability and legitimacy of her research outcomes.

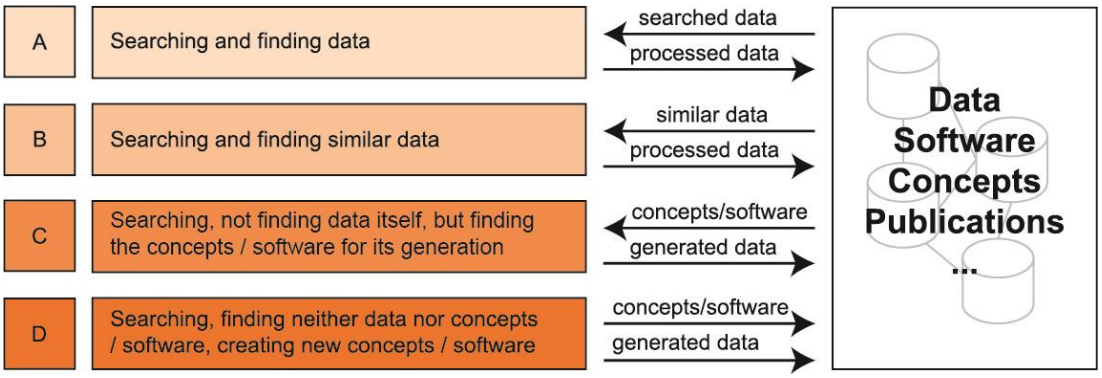


Figure 3.6.1 ELLEN's gradually extended detour to retrieve required data

3.6.3 State of the art of knowledge-based data exploration

Knowledge Graphs such as DBpedia [69], Yago [70], and WikiData [71], industrial initiatives by Google, Bing, IBM, BBC, or Thomson Reuters, and domain-specific developments like KnowLife [72] and iASiSKG [73] have demonstrated the feasibility of representing encyclopaedic and factual knowledge using RDF and Linked Data. Knowledge Graphs use semantic knowledge representation based on RDF and Description Logics, which makes them comprehensible to humans and machines. Therefore, Knowledge Graphs have been extensively applied in Question Answering, text disambiguation, categorisation, search, and retrieval, as well as in digital assistant systems. Question Answering systems provide intuitive interfaces through which users can access information in natural language and obtain direct answers [74]. Recently, first studies emerged that utilises Knowledge Graphs as background knowledge to provide extra signals for

machine learning [75]. While there has been a vast amount of work related to representing and managing bibliographic metadata, relatively few works focus on representing the information inside scientific publications semantically. The Semantic Publishing and Referencing Ontologies [76] focus primarily on metadata but also on document structure to some extent. There has been some work on enriching various document formats with semantic annotations such as Dokie.li [77], RASH [78] or MicroPublications [79] for HTML and SALT [80] for LaTeX. The representation of key findings of survey articles focussing on semantically describing research problems, approaches, implementations and evaluations has been started [81]. Further work focusses on developing ontologies for representing scholarly knowledge in specific domains, for example, mathematics [82], the RXNO ontology in chemistry or the OBO Foundry ontologies [83] in the life sciences.

3.6.4 Key objectives

The aim of this task area is to support engineers in their search for data by increasing the number of potential data sources, raising their level of integration and reducing the amount of time required for the search process (objective 4). To this end, in the case of unavailable data sets, scientifically recognised methodological concepts and their software implementations will be made available to generate the missing data (objectives 1, 2, and 6). Since neither publications nor software programs are suitable to be used as a guide to the implementation of a methodology, concepts will serve this purpose within the research data landscape. Within this task area, a semantic framework will be developed, enabling the semantically enriched and machine-interpretable representation of methodological knowledge, allowing the retracement and comprehension of each individual step of a procedure. Through the utilisation of scalable data storage backends and the development of web-based service structures, the engineering community will be enabled to store methodological knowledge within the structures of a knowledge graph, to link it with related information and to retrieve it through advanced query techniques (objective 3). Through joint developments with the community and its participation in the integration and exemplary application of knowledge-based research data management techniques, the validity and practical suitability of the project results will be guaranteed. Research findings and best practice guides will be published and disseminated within various community platforms and networks (objective 7).

3.6.5 Measures

Measure E-1 “Semantic mapping of methodological knowledge” (FZJ, TIB)

To enable the automated processing of methodological knowledge, a machine-interpretable structure for semantically enriching and unambiguously describing the individual steps of methodological approaches will be developed. For the resulting methodological concepts, a

metadata structure will be designed, particularly informing about their purpose, required input data and resulting output data. The structures of the methodological concepts will be extended to enable their combination to more complex conceptual workflows, allowing required input data to be substituted by suitable data-generation concepts. In order to increase the findability and reusability of the concepts in the context of data and information queries, the semantic framework will be extended to also map the questions a methodological concept can help to answer. The following tasks will be performed in this measure:

Task E-1-1: Development of a semantic framework for the machine-readable representation and mapping of scientific concepts.

Task E-1-2: Design of metadata structures to uniformly describe scientific concepts.

Task E-1-3: Design of semantic structures to enable concept combinations.

Task E-1-4: Extension of the framework to represent concept related information queries.

Measure E-2 "Use of methodological knowledge to facilitate data exploration" (FZJ, TIB)

In order to integrate the methodological concepts into the research data management landscape and to facilitate their public accessibility, a scalable data storage backend infrastructure is to be utilised. To ensure the accessibility of the semantic descriptions and representations, search algorithms typically employed by repositories to identify data-sets are enhanced to also present concepts capable of generating the sought data. The methodological concepts will be linked to publications, data sets, authors and other elements of the Open Research Knowledge Graph [84]. In addition, they will be tied to open-source software programs, e.g. contained in GitHub/GitLab repositories. In order to provide an information base that supports the selection of one of the offered methodological options, procedures will be developed that estimate the effort required to apply a methodology and to compile its required input data. The development of concept-related quality scales utilised by assessment procedures will further support the decision process. The developed algorithms and procedures will be integrated into a comprehensive knowledge-based research data exploration service, capable of processing complex information queries (cf. Figure 3.6.2). The framework will be provided to the community over an exemplary web-based frontend. The tasks of this measure are:

Task E-2-1: Integration of the framework in the storage structures of a knowledge graph.

Task E-2-2: Enhancement of search algorithms for the consideration of scientific concepts.

Task E-2-3: Development of procedures estimating the specific effort to apply concepts.

Task E-2-4: Personalisation of search results according to preferences and quality scales.

Task E-2-5: Implementation of a web and knowledge-based data exploration framework.

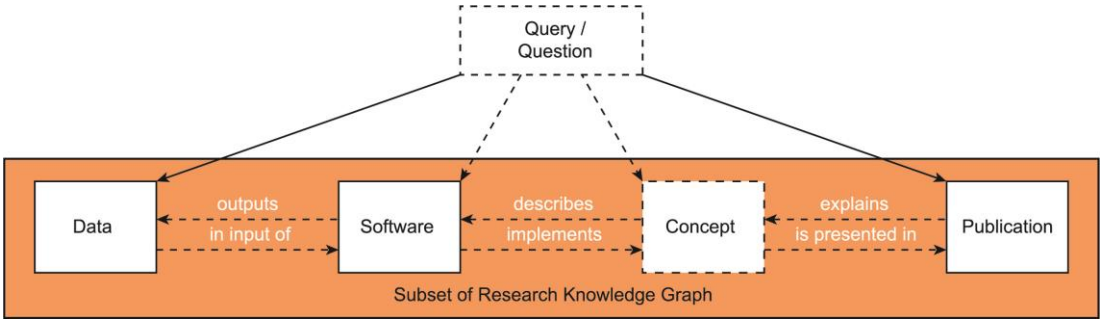


Figure 3.6.2 Subset of Research Knowledge Graph elements relevant for the querying of data generation methodologies (the focus of this task area is indicated by dotted lines)

Measure E-3 "Connection of data exploration and data generation processes" (FZJ, TIB)

In order to support engineers not only in their search for data and methodologies but also in their application, an overarching process framework is to be designed. Depending on the availability and interoperability of software components, services and input data, its web-based implementation will seamlessly link the search process with processes of software selection and parametrisation and data selection and transformation. As an interface to the following step of executing a computational workflow, a procedure will be developed to compile and provide all necessary data in a machine-interpretable format. The associated tasks are:

Task E-3-1: Development of a web frontend for the selection and parameterisation of software components necessary to implement a scientific methodology.

Task E-3-2: Development of a web frontend for the selection of data sets and data transformation functionalities necessary to apply a computational workflow.

Task E-3-3: Development of a procedure compiling all information necessary for a container-based implementation and execution of a computational workflow.

Measure E-4 "Community-based validation of data concepts and services" (FZJ, TIB)

All development work will be carried out in close collaboration with the community clusters. Workshops and surveys will be conducted at regular intervals throughout the project to collect requirements, experience reports and feedback regarding knowledge-based data exploration. To ensure the validity and practicability of the concepts, implementations and processes, various use cases will be implemented together with pilot users from the community. In this context, both the provision and the utilisation of knowledge-based storage and retrieval services will be tested. A detailed guideline will be published, which addresses the engineering community and describes the maintenance and retrieval of research data within a common data infrastructure. Also, infrastructure providers who want to offer knowledge-based services will be addressed. The participating institutions will activate their respective networks (such as the DataCite, arXiv, and ORCID communities) in order to disseminate, adapt and sustainably establish the research results and the developed services. Measure E-4 comprises the following tasks:

Task E-4-1: Collection of feedback regarding knowledge-based exploration techniques.

Task E-4-2: Implementation of engineering use cases comprising the provision of knowledge-based storage and retrieval services.

Task E-4-3: Implementation of engineering use cases comprising the utilisation of knowledge-based storage and retrieval services.

Task E-4-4: Compilation and dissemination of a best practice guide for utilizing scientific concepts, targeting the engineering community as well as service providers.

Possible risks of implementation

The embedding of methodological knowledge as well-defined scientific concepts into the research data infrastructure will fully exploit its potential, if the concepts and services of this task area have been harmonised with those of the task area services. Until this is achieved, the measures can be implemented protoypically with the resources of FZJ and TIB.

3.6.6 Synergies and demarcations with regard to other task areas

Within the task area BETTY (cf. chapter 3.3) the handling of software programmes combined into workflows is to be standardised and generally facilitated (B-1). To achieve a high degree of automation, also web frontends are to be employed (B-3). By transferring information on data generation workflows, which are to be assembled within E-3, workflow selection and execution processes can be seamlessly linked. In task area FRANK (cf. chapter 3.7), publications are analysed to obtain information on applied methodologies (F-1). The utilisation of a semantic framework capable of representing this information in a machine-interpretable form (E-1) will simplify this task. The creation of a terminology service (S-3) supports the development of a semantic framework for the representation and mapping of scientific concepts (E-1). In S-3 also a uniform utilisation of metadata is established, supporting the design of metadata structures for scientific concepts (E-1). The knowledge discovery algorithms to be developed in S-7 will automate the identification of methodological knowledge hidden in conventional publications (E-2). Due to their strong relation to the publications describing the respective methodology, the concepts will serve as a training set for the algorithms at the same time (S-7). The integration of methodological concepts into the data storage infrastructures of a knowledge graph (E-2) will benefit from the best practices concerning the utilisation of storage infrastructures and their integration with repositories (S-4). The systematisation of data quality criteria and metrics (S-1) is preliminary to the personalisation of search results according to user preferences and quality scales (E-2). The development, dissemination and establishment of concepts and services will be carried out in close interaction with the engineering community (E-4), thereby cooperating especially with the community clusters targeting computer science, systems and electrical engineering (community cluster 44).

3.7 FRANK: Many participants and simultaneous devices (RWTH, TUB, IoP)

3.7.1 Introducing the archetype FRANK

"Hello, I'm FRANK. I'm an engineer who works with a range of different and heterogeneous data sources - including the collection of data from test persons up to manufacturing networks. One of the main challenges during my research process refers to the synchronisation and access management of simultaneously and distributedly generated data.

My professional background is mostly informed by production engineering, industrial engineering, ergonomics, business engineering, product design and mechanical design, automation engineering, process engineering, civil engineering and transportation science."

For this archetype, many simultaneously involved participants, end-user-devices and machines will generate different data types and dimensions. Many participants can either refer to many researchers working together or to many research objects, e.g. test persons. The generated data can be clustered into machine and device related or human related (cf. Figure 3.7.1). Managing such different data types and dimensions from many participants is unique to FRANK.

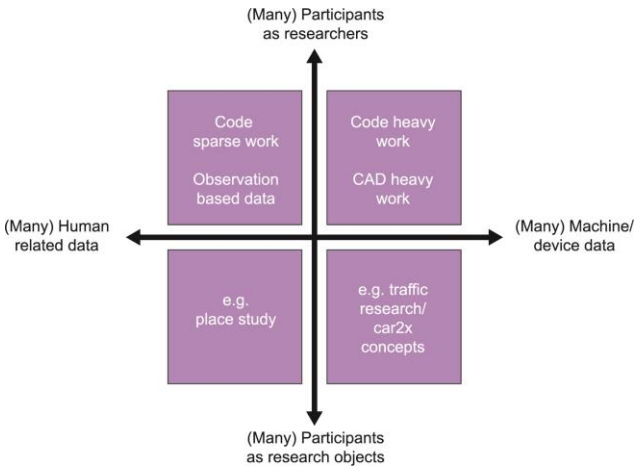


Figure 3.7.1 Typical domains and dimensions of FRANK's research activity

3.7.2 FRANK's key challenges with respect to research data management

Key challenges for the archetype FRANK arise from (a) the variety of involved engineering disciplines and (b) from the collaborative nature of working with many participants (such as common language cf. S-3).

a.1) Diversity of raw data: Data differs greatly in its level of structure, e.g. operationalised variables (measurement of empirical phenomena) in contrast to sensory data (physical entity) and software

data (e.g. CAD, ERP). FRANK has to record, store, describe (metadata and variable definition), analyse and publish (inter- or intra-organisational access) a great variety of file types, e.g. ranging from transcripts to STL-files.

a.2) Recording methodology as an integral part of a data set: Due to the aforementioned key challenge, the need for efficient documentation of recording methodologies, description of research environments and study setups is identified as an integral part of the data set.

b.1) Common language: Regarding collaboration with many participants and due to the variety of disciplines, a standardised vocabulary in terms of discipline-specific issues as well as research methodologies is required but not yet established (S-3).

b.2) Anonymisation and access: Research in engineering often involves working with corporate and confidential data. Therefore, anonymisation schemes for distributed data are required. Access to stored data has to be restricted, protected and managed in a scalable manner.

3.7.3 State of the art in research data management

Anonymisation schemes: Several methods exist for data sanitisation but most are not applicable for FRANK's use cases because of lacking efficiency and protection against re-identification (key challenge b.2). Methods comprise of pseudonymisation with artificial identifiers replacing personally identifiable information and K-anonymisation, where k-1 individuals within the data set cannot be further distinguished. Suppression, generalisation, noise addition and random permutation of information are further methods. [85], [86] Special emphasis lies within the simultaneous anonymisation of distributedly created data, concurrent data streams and heterogeneous sources, which is currently not fulfilled due to the lack of process automation in anonymisation schemes (e.g. automatic anonymisation when exporting data from ERP) and computational costs [87], [88], [89].

Role and Access Management: *Role-Based Access Control (RBAC)* regulates access to certain data by defining user groups with a set of permissions. Those roles are generally easily manageable and can be assigned by non-expert personnel. Adversely those roles must be defined upfront. *Attribute-Based Access Control (ABAC)* is the diametric framework to RBAC and grants permissions via rules based on attributes assigned to users and actions, which allows a more granular and dynamic control. This option eliminates the need to define fixed roles, which are not always certain in advance. Implementation hurdles for ABAC consist of a large number of rules and attributes. Therefore, a need for expert personnel emerges. [90], [91], [92] For the management of distributedly and simultaneously generated data, still no framework exists (key challenge b.2), which combines the advantages of both approaches: the simplicity of Role-Based and dynamics of Attribute-Based Access Control.

3.7.4 Key objectives

Overarching key objectives are the establishment of ontologies, which include research methodologies and environments, and how to integrate those into the data set as well as reducing the collaboration effort by redesigning a decision-supporting RDM framework. Furthermore, to increase the acceptance and application of RDM the following requirements are predefined:

1. No code required to actively use RDM (e.g. Graphical User Interface) but supported (e.g. using terminal prompts instead of GUI).
2. Easy access sharing of heterogeneous data regarding many engineers, that simultaneously work on shared data (cf. overarching objective 4 sharing and integration of data, objective 5 unhindered collaborative research and objective 6 cost-efficient solutions).

Overall, we emphasise retracability in the context of heterogeneous and many data sources, which also relates to the recording and linking of auxiliary information. This includes research environments and methodologies (cf. chapter 3.6). As for collaboration, we further emphasise sharing and access management to data as well as cooperation amongst many engineering disciplines. FRANK's key objectives comply with the key objectives 1, 3 - 6 of NFDI4Ing.

3.7.5 Measures

Measure F-1 "Target process specification" (RWTH)

Based on existing definitions of RDM process frameworks, we consolidate, adapt and test those in terms of FRANK's research activities. We use an iterative process development approach in order to elaborate a specific guideline and decision support regarding how to handle RDM for FRANK step-by-step.

*Results***:** Well-defined and hands-on RDM process tailored to the needs of FRANK's archetype, typical workflow processes to be taken into account within RDM (e.g. workflow in traffic science differs greatly from production engineering), identification of current implementation hurdles (e.g. deploying certain kinds of databases to institutional servers).

Synergies: Discovered implementation hurdles will be reported to the task area Base Services S-4 and S-6. Input from as many community clusters and pilot users as possible will be crucial for elaborating the target process specification. Input from the task area Base Services S-1 *Data Quality Assurance* will be taken into account to ensure the integration of quality management and DMP frameworks as well as the matching between the maturity model of S-1 and the target process specification regarding RDM.

Task F-1-1: We assess existing RDM process frameworks regarding criteria, which are directly derived from FRANK's requirements and key objectives. Such criteria comprise e.g. collaborative planning of data life cycles with many researchers (concepts regarding work

organisation) and management of conducting many simultaneous experiments and their data streams (concepts regarding operative manners).

Task F-1-2: After selecting and supplementing by further frameworks like CRISP-DM (cross-industry standard process for data mining) [93], we define concise tasks per each step in our proposed RDM process. Furthermore, best practices, risks and links to further information and literature will be provided for each step.

Task F-1-3: The concept developed above will be deployed and tested with research groups in order to identify remaining disadvantages and process weaknesses. Those optimisation potentials will be evaluated in close consultation with the researchers.

Measure F-2 "Technological feasibility and decision-making" (RWTH, TUB, IoP)

Whereas measure F-1 defines the underlying needs (as a gap between the target process and the process in reality) that arise from current implementation and practicability problems, measure F-2 examines which technologies at hand can be used.

Results: By comparing their workflow, engineers will be able to locate their planned methodological research approach within the above-mentioned framework. After the identification of relevant workflow elements, engineers will be guided in their decision-making regarding relevant technological aspects, e.g. pros, cons and implementation expenses.

Synergies: Results will be reported to the archetype ELLEN (cf. chapter 3.6) due to her methodological focus and will be discussed with the task area Base Services S-4 regarding costs and S-6 regarding potential training contents. To ensure covering enough engineering specific workflows, results will be elaborated in close conjunction with the community clusters and pilot users.

Task F-2-1: Evaluation of technologies and software implementations that meet the previously identified requirements of FRANK's archetype. Therefore, missing technologies (e.g. for applying interoperable metadata standards) are identified from measure F-1 and complemented by technologies, which already exist, but may not be in use yet.

Task F-2-2: Clustering of identified technologies aco

cording to their fit regarding different research methodologies (e.g. Design of Experiments states different requirements than behavioural location tracking of traffic).

Task F-2-3: Conduct feasibility check and gap analysis regarding typical everyday-issues (e.g. costs, implementation workload, IT security, etc.).

Task F-2-4: Identification of current collaboration barriers regarding data exchange as well as organisational matters.

Measure F-3 "Design concept of an application program interface" (RWTH, TUB)

With the previously identified implementation hurdles regarding a vivid RDM practice as well as feasible technological options at hand, a general concept of a program interface must be designed. This concept comprises how to combine and interface several technologies.

Results: Standard procedure in designing program interfaces for RDM systems, key requirements that can be handed over to third parties, e.g. IT-centres of universities, institutions and the task area Base Services.

Synergies: Consultation by the archetype BETTY (cf. chapter 3.3) and the task area Base Services S-2, S-4 and S-5 will be necessary. Acquired results can be reused within the mentioned task areas.

Task F-3-1: Deduce potential interfacing as well as segmentation points between workflows and feasible technologies.

Task F-3-2: Conceptionalise a unifying approach/framework to integrate technologies into workflows and define key requirements for e.g. third-party implementation.

Task F-3-3: Conceptionalise a role-based access framework and the application of anonymisation schemes in a preferably automated manner.

Task F-3-4: Support domain-specific visualisation of data via role-based access points.

Measure F-4 "Incentivation of an active and interdisciplinary RDM use" (RWTH, IoP)

For FRANK, it is necessary to consider organisational measures as FRANK's projects often involve many researchers in an interdisciplinary setting. Incentivation is one of those organisational measures, which ensures, that the above-developed RDM target process will be complied with.

Synergies: community clusters and pilot users will deliver great insight into their micro-universes in order to identify potential levers of incentivation. Results will be reported to the task area S-6 and S-1.

Task F-4-1: Stakeholders such as the German Academic Association for Production Technology (WGP) will be consulted. In close coordination, we will discuss potential measures, as well as plan and accompany the rollout of those measures together. Potential measures include e.g. awards for RDM, RDM as a requirement for publishing, proof that members have regularly conducted RDM-trainings at their institutions, etc.

Possible risks of implementation

Risks regarding F-1*:* It is not guaranteed that our newly proposed RDM process concept is easier to handle with less work effort. This is given when administrative tasks cannot be obviated.

Risks regarding F-2*:* Implementation hurdles largely depend on the level of IT support and resources of the research institution. For this measure, the risk exists of not finding a common criteria set and therefore no generally feasible technologies.

Risks regarding F-3*:* The main risk of this measure lies within its theoretical character. Close collaboration with software experts will ensure the applicability of the concept.

3.7.6 Synergies and demarcations with regard to other task areas

Task area BETTY (cf. chapter 3.3) will be consulted for designing a generalised API concept in order to exchange requirements and adhere to best practices. Task area's S-2 expertise in software quality will further support F-2 and F-3 to ensure a quality-oriented API concept. Task area CADEN's (cf. chapter 3.4) provenance tracking will be utilised to enable many researchers to simultaneously augment data, which opens up new opportunities in collaborative data analysis and will be considered during designing a target RDM process specification. Task area ELLEN's (cf. chapter 3.6) focus on methodologies is used for integrating methodologies into ontologies, as FRANK has to record and store the research environment and methodology in use. This approach will be further supported by task area Base Services S-3 regarding the use of a common language amongst various engineering disciplines, in which FRANK can deliver valuable input. Further cooperation is planned with the task area GOLO (cf. chapter 3.8), as GOLO has to cope with simultaneous data from many different sensors. Results from specifying a target RDM process will be used in the task area S-6 as input for new training contents. FRANK will also work in close cooperation with task area S-5 to harmonise approaches for the role and access management. Recommendations and catalogues from task area S-4 will be used as input in F-2, which in turn can give new insights for S-4 regarding costs and implementation challenges.

3.8 Golo: Field data and distributed systems (LUH, TUD, DFKI)

3.8.1 Introducing the archetype GOLO

"Hello, I am GOLO. I'm an engineer involved in the planning, recording and subsequent analysis of field data for:

- examination of the operating conditions of a technical system and/or subsequent adaptation of a system to the environment and real operating conditions;
- use of the results to adapt the model and improve the methodology and parameters of the modeling of the technical system;
- preparation of data sets and parameters for learning processes and testing system models.

My professional background is formed by production technology, constructive mechanical engineering, systems engineering, robotics and information technology. My interests include methods and tools for collection, quality assurance and quality control, analysis and reuse of field data. Field data is generated from the actual or experimental operation of cyberphysical systems."

3.8.2 GOLO's key challenges with respect to research data management

Effective use of field data requires not only storaging, archiving and subsequent access to the data itself, but also the description of methods and techniques for data collection and analysis. Actual RDM environments and platforms support the processes of data storage, archiving, finding, and access, but do not provide ready-made tools that methodologically support both: the principles of data storage and access, and the methods and tools for field data processing and analysis.

The reusability of field data is often severely restricted, as the operation of a technical system in a real environment is influenced by many factors, which are almost impossible to fully cover in a test environment. In addition, information on data collection methods and data analysis methods is an important, yet currently widely neglected aspect. This information and its standardisation is necessary to meet the interoperability and reusability criteria of the FAIR data principles. Furthermore, no unified interfaces and protocols for accessing and reusing field data are available, yet. This also holds for a standardised information feedback for improving subsequent generations of technical systems.

3.8.3 State of the art in research data management

Traditionally, Product Data Management (PDM) and Product Lifecycle Management (PLM) systems are used to organise the storage and use of field data in the industry. PDM systems are used to manage computer-generated models, technical drawings and other development and product documents. By extending functionality to the lifecycle view of a product, PDM is often identified as an important part of PLM [94]. Basic functions of PDM systems [95], [96], [97] do not fully support FAIR data principles. In addition to PDM and PLM systems a concept of field data structuring using the digital twin approach [98] provides more opportunities for networking and interpretation of field data. Current definitions focus primarily on the mapping of physical systems into the digital world [99], [100], so the concept need an adaptation for RDM.

3.8.4 Key objectives

The main objective of this task area is to increase the possibilities of (re-)using field data and support their research data management as a whole. Therefore, a methodological approach will be developed and implemented in order to use a digital twin of a physical system as data

representation framework. This may also include communication of the system with other objects connected to it, and software responsible for management, operation, maintenance monitoring. A digital twin is of particular value when it most accurately displays the actual state and performance of its physical twin while also providing a history of states and performances under different conditions.

The digital twin enables the structured representation of research field data in terms of (i) storage, (ii) processing, and (iii) (re-)use of field data. The digital twin will contain information on the system model, methods of collection and analysis of the system's field data, as well as the field data itself. Automated classification of field data will be facilitated by the approach. Generated metadata and analyses as well as derived information from raw data will be attached to the digital twin concept. These two objectives relate to NFDI4Ing's overarching key objectives 1, 3, and 6. The concept will be expanded over the whole process of capturing, providing, using, and storing field data over the lifecycle of the technical system and beyond. Hence, information about the system model as well as about the methods of collection and analysis of the system data and the research field data are considered. This enables the correct interpretation of the collected research data during later reuse. This objective relates to NFDI4Ing's overarching key objectives 1, 2, and 3.

3.8.5 Measures

The general approach to the concept of a digital twin of a technical system adopted in the task area GOLO is shown in Figure 3.8.1.

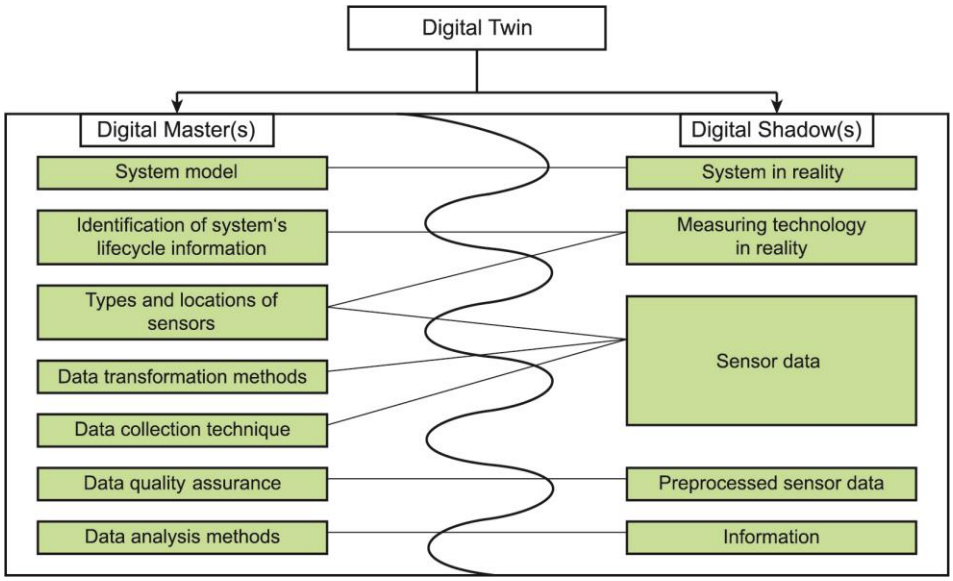


Figure 3.8.1 General idea of the concept of a digital twin for a technical system

The concept assumes that the twin has two parts: a digital master and a digital shadow. The digital master is responsible for interpreting and processing research field data. The digital shadow contains research data and the information obtained from its processing.

All measures will be developed and implemented with the involvement of the community clusters, in particular the clusters CC41: Mechanical and Industrial Engineering and CC44: Computer Science, Systems and Electrical Engineering.

Measure G-1 "Conception of a digital twin for the organisation and processing of research field data" (LUH, TUD, DFKI)

We will classify heterogeneous field data for metadata generation, automatic generation of data set summaries and for label generation. Both, human users and machines or machine-learning algorithms can make use of the generated metadata and labels. Furthermore, we will include additional data such as process parameters and environmental data into the digital twin concept.

Task G-1-1: Definition and establishment of structure and data flows in the digital twin framework: This task will be iteratively conducted from beginning to end of this measure.

Task G-1-2: Classification of research field data: Classes for data clustering will be established. This facilitates automated processes and machine-learning on the field data. In an iterative process, the set of generated classes will be extended throughout the runtime of the measure.

Task G-1-3: Description of the phases and processes of working with field data: Typical phases in a product lifecycle and in the field testing will be identified and described. Several iterations of the task's results will be conducted throughout the measure.

Task G-1-4: Adaptation and development of technical systems and models based on information from the research field data. This task represents the information feedback for system evolution.

Measure G-2 "Creating a digital master of a technical system" (LUH, DFKI)

We will create the digital master for a technical system and its field data. The digital master facilitates the correct collection and interpretation of field data based on data processing and analysis methods. With regard to copy instances of the same generation of a technical system, the digital master can be identical.

Task G-2-1: Identification of types of necessary and sufficient information about a technical system: Typical systems are identified and a minimal set of information required is defined.

Task G-2-2: Classification of data collection methods: Methods for data collection are classified, for example sensor data of a system vs. reference (ground truth) data.

Task G-2-3: Classification of data analysis techniques: Methods for analysis and representation of quantitative and qualitative data to facilitate the reproduction of results.

Measure G-3 "Recommendations for creating and monitoring digital shadows" (LUH, TUD)

This measure is concerned with the establishment of a digital shadow of a technical system as representation of a specific instance of a physical system. The combination of digital shadows of different instances will be used to get a better understanding of the general characteristics of research objects, fleets or generations of technical systems. Here, the principles of big data are applied to the specific technical system under study.

The digital twin approach will also be adapted to the organisation of qualitative data collection. In this case, the measuring technique and the sensor data do not participate in the twin profile neither in the master part nor in the shadow part. However, part of the master should contain instructions on how to collect and analyse quality data.

Task G-3-1: Analysis and quality control metrics for different types of field data.

Task G-3-2: Step-by-step technique for analysing digital shadows: collecting, understanding and interpreting field data using the models and methods contained in the digital master.

Task G-3-3: Development of a concept for transferring the approach of digital shadows to socio-spatial structures (e.g. cities) and prototypical applications.

Measure G-4 "Conversion of the extended concept of the digital twin into a ready to use process" (LUH, DFKI, TUD)

We will consolidate the efforts of the previous measures into a process ready to use for an exemplary group of pilot users. This requires to identify the necessary interfaces and adaptation of existing tools from RDM. The prototype digital twin from the previous measures will be extended and application as well as transfer will be validated.

Task G-4-1: Identification of the relevant interfaces and workflows over the life cycle of field data.

Task G-4-2: Formulation of adaptation requirements for existing tools and standards in research data management.

Task G-4-3: Prototype extension of the field of application.

Task G-4-4: Validation of a concept for transferring the approach of digital twins to socio-spatial structures (e.g. cities) and prototypical applications.

Possible risks of implementation

To develop the concept of digital twins regarding field data, support from measures S-3 and S-4 task area Base Service is required. In case of limited capacities, additional support by the IT Service LUIS of the Leibniz University Hannover could be provided to realise and test the prototype.

3.8.6 Synergies and demarcations with regard to other task areas

To implement the concept of digital twins, the task area GOLO includes the exchange of requirements and results with the task areas CADEN (cf. chapter 3.4) and ELLEN (cf. chapter 3.6) in developing the concept of digital masters. Since numerous sensors and devices can be used for research field data collection and analysis, there will be a cooperation with task area FRANK (cf. chapter 3.7).

Furthermore, there is a close cooperation with the task area Base Services (cf. chapter 3.9) in the process of development and implementation of a framework of digital twins for storing, processing and using research field data. It is necessary to realise the processing, monitoring and control of data quality (cooperation with S-1), methods and tools of versioning scripts for data processing (cooperation with measure S-2), development and support of machine readability and searchability of field data (cooperation with measure S-3). An important role in data storage, communication and interaction between the elements of the digital master and the digital shadow of the technical system belongs to the methods of data storage and the metadata of the system. Therefore, it is necessary to work together with measures S-3 and S-4. The results of task area GOLO can be presented in the form of trainings, for which joint work with measure S-6 is planned.

3.9 Base Services (TUDA, RWTH, KIT, TIB, LUH, US)

In this task area, all basic RDM services are bundled. It provides central services for the research archetype task areas described in task areas 0 – 3.8 in order to bring together the common aspects of the archetype-specific requirements. In the same spirit, it develops and maintains services considered as relevant by the community clusters (cf. chapter 3.10). This task area is designed very interlocked between the partners reflecting the strong interdependencies of the measures. Each measure is worked on by three or four partner institutions with defined lead. Each staff member of the task area Base Services will participate in at least two measures to foster the unified view on the service portfolio.

3.9.1 Key objectives of task area Base Services

NFDI4Ing builds all of its services in a modular and user-centred style (cf. chapter 2.3). The basic services are delivered by this task area, while the archetype-specific task areas build upon these services their differentiated services tailored to their respective methodological demands from the engineering communities. In the other direction, the RDM implementations in the archetype task areas serve as prototypes for the base services according to our user-centred approach. We have

identified seven service objectives relevant for all engineering archetypes and communities and thus for NFDI4Ing's overarching key objectives (cf. chapter 1.1):

- All archetypes and communities rely on data quality assurance processes, the respective tools, and data quality metrics to make their data FAIR and to enable engineers to appraise and select data for further curation (S-1);
- the support of research software development is by now urgent in all fields of engineering, particularly but not only in computational engineering (S-2);
- providing easy-to-use yet comprehensible metadata tools for the engineering research daily routine as well as establishing detailed terminologies for engineering are the common ground for all RDM processes (S-3);
- the safe and secure storage and long-term archiving of data as well as the possibility to share or publish data in suitable repositories is relevant to all archetypes and communities, yet in different shape (S-4);
- all archetype services participate in NFDI4Ing's overall software architecture incl. authentication, authorisation, and role management schemes, absolutely necessary for confident data from research projects close to industry (S-5);
- concepts and materials for training are needed in all task areas (S-6);
- most of the outcomes of engineering research is still hidden in human-readable publications only. Making engineering data FAIR needs advanced techniques of data extraction from and knowledge discovery in the engineering literature (S-7).

All services are designed as open, modular, and standardised as possible, in order to foster cross-consortial reusability. Furthermore, we strive for technical and structural connectivity to parallel and prospective developments in RDM on the national and international level (e.g. in the EOSC).

3.9.2 Measures

Measure S-1 "Quality assurance in RDM processes and metrics for FAIR data" (LUH, RWTH, TUDA)

The purpose of this measure is to assist and support researchers in organising and self-monitoring research data processes to ensure and control data quality. Within the scope of this measure, a framework will be created which provides standards, metrics, and guidance for the organisation of data curation based on existing best practices and the FAIR principles. As a basis for such a framework a common research data management maturity model and special models for each of the archetypes will be developed. In order to assure data quality, a commonly agreed set of criteria will be defined that is valid not only for the engineering community but for the scientific community as a whole.

Task S-1-1: Development of research data management maturity models. (LUH)

We will develop a data management maturity model, including a description of five maturity levels of the organisation research data management processes, by analogy with the literature [101], [102]: initial, managed, defined, quantitatively managed, and optimised. Each level of maturity is assigned to processes, which in turn are grouped into different process areas: data management planning (S-1-2, S-1-3), monitoring and control of data storage processes, data access management, ontology and metadata management, risk management of research data, data quality management and organisational training. These processes are assigned to generic and specific goals and define the degree to which a set of characteristics of data fulfills requirements and which will be described and achieved by several best practices, self-control criteria, checklists and metrics (S-1-4). Depending on the type of the research project, these checklists can be used for self-control in small research projects or for the management of large projects to control and motivate work with research data in and between subprojects.

Task S-1-2: Support of data management planning with RDMO. (TUDA)

One important and established tool for quality assurance in RDM is data management planning (DMP). NFDI4Ing will make use of the open source RDMO software [103] and will offer a DMP service to all engineering scientists through a central multitenant RDMO instance maintained by ULB Darmstadt [104]. It provides a highly adaptable user interface, authorisation procedures and customizable DMP templates per client, but runs on a single database, thus allowing easy collaborative data management planning in multi-institutional research projects. We will implement new features in RDMO regarding 1) the validation of entries in DMPs in order to connect RDMO with the maturity model (S-1-1); 2) the realisation of the RDA machine actionable DMP model [105], [106] and thus 3) connection to other RDM systems via the RDMO API (e.g., repositories, metadata services, ORCID, CRISs).

Task S-1-3: Fostering DMP templates in engineering. (TUDA, RWTH)

We will develop, test and evaluate archetype and sub-discipline specific DMP templates, for different stages of research projects, as well as in close collaboration with all archetypes and community clusters. Preliminary work has been done at RWTH Aachen and TU Darmstadt for mechanical engineering [107], [108]. We will also investigate how to support researchers and local RDM offices by designing a DMP review service based on the maturity levels. The establishment of an RDM certification scheme introducing an “NFDI4Ing seal for FAIR data” is our goal in the long run.

Task S-1-4: Providing and utilizing FAIR data metrics. (RWTH)

We will develop metrics and create relevant KPIs for the FAIRness of engineering data management. KPIs have to be added to general, already established research performance measurements, e.g. the number of publications. In the course of developing relevant KPIs, effective goals have to be defined first and in close collaboration with the engineering communities

to ensure a participative author-critic-cycle. Deduced goals are then operationalised, e.g. in the form of metrics such as data clearity, data completeness, etc. The operationalised goals and metrics form the basis for self-assessment at institutional level, which subsequently can be utilised to track performance, benchmark RDM performance comprehensively, and impose RDM by performance incentivation. As those steps resemble a quality forward chain, in return, a quality backward chain will be implemented in the form of a continuous improvement process. This process is triggered periodically by this measure but optimally conducted by every NFDI4Ing member whenever improvement potentials attract one's attention. Therefore, NFDI4Ing will implement corresponding feedback platforms.

Measure S-2 "Research software development" (TUDA, RWTH, KIT, US)

Many scientific workflows are governed by algorithms written in software codes. Therefore, software is an important type of research data in the engineering sciences in its own right. It is thus important to build RDM services that take the special properties of research software into account.

A particular computer experiment is usually represented by a snapshot in time, with respect to the state of the source codes and scripts employed as well as the operational environment it is executed on. Replicability guarantees that a software-driven computer experiment, repeated in the same operational environment, produces the same results. Reproducibility, on the other hand, ensures that a software-driven computer experiment can be repeated in a different context. As also outlined in the challenges for the archetype BETTY (cf. chapter 3.3.2), the situation is particularly complex in engineering, as many different programming languages are employed, often within the same computer experiment (e.g. python scripts for orchestration of simulations that are written in C++, C and FORTRAN). The actual execution environment of a piece of software is also affected by issues such as the hardware platform (e.g. accelerators of various flavors), libraries, the operating system, and compilers. The goal of this measure is to provide services and best practices that allow researchers to combine RDM and enterprise-grade software development workflows.

Task S-2-1: Infrastructure for replicable & reproducible software-based experiments. (TUDA, RWTH, US, KIT)

We will set up reusable workflows, e.g. script-based, that relate source codes, experimental setups, and data processing routines. Best-practices in continuous integration, arising, for example, from the activities of measure B-1 (cf. chapter 3.3.5) will be standardised, documented and continually assessed, in particular with respect to version control, automated testing, deployment, and linking to publication records. We will make them usable not only for highly-skilled software developers, but also for the large number of researchers without a background in software engineering. We employ and assess container solutions such as Singularity for

packaging up simulation experiments. We will closely collaborate with measure B-3 (cf. chapter 3.3.5), emphasising generality of usability as well as enhanced awareness of these platforms of HPC environments and their requirements on performance and emulation of hardware features (see also measure D-5 in chapter 3.5). Finally, we will deploy a JupyterHub server to lower further the threshold for new users. This service dovetails the part of measure B-3 which is concerned with the development and automated generation of JupyterLab frontends.

Task S-2-2: Services for assessment of the quality of software created by engineering researchers. (TUDA, RWTH)

Templates and best practice examples will be made available that allow to create quality metrics to make sustainability and reusability of source codes tangible for researchers using existing enterprise grade solutions like SonarQube [7] (cf. chapter 3.3).

Based on existing continuous integration software, we create a service that can be used by engineering researchers for continuous and automatic generation of quality metrics to enhance their software during development and for judging the quality of existing codes before they are being reused (cf. chapter 3.2, 3.3, 3.5, 3.6, 3.7). We will connect quality metrics with RDM workflows to extract additional metadata from existing source codes or documentations to make RDM tasks like publishing easier for engineers.

Measure S-3 "Metadata and terminology services" (TIB, TUDA, RWTH, KIT)

The definition of metadata and concepts, associated attributes and relations that are readable and understandable, not only to target audiences but also to machines, is key to FAIRness. To address this concern, NFDI4Ing will provide services to facilitate the creation of subject and application-specific standardised metadata and their integration into engineering workflows, as well as a Terminology Service (TS) to enable researchers and infrastructure providers to access, curate, and update terminologies.

In NFDI4Ing, it is essential that metadata is not only used for documentation and indexing of research data stored in repositories, but also to facilitate tasks such as (automated) retrieval, analysis, or combination of complex research data during active research. As such, the usage of metadata and semantic concepts of ontologies requires two components: a) the ability to quickly generate application-specific but standardised metadata without delaying science and b) the provision of a single-point-of-entry for terminology in NFDI4Ing. As tools for the generation, management, and use of vocabularies and ontologies, NFDI4Ing will evaluate, test, and deploy the most suitable tools for reliable management of subject-specific vocabularies. This requires metadata and terminologies to be at the same time a) detailed enough to allow for specific applications, b) interoperable enough to enable combination of data from different data sets (even across subject-areas and disciplines), and c) flexible enough to accommodate highly

heterogeneous workflows without hampering scientific freedom. In the assessments of the NFDI4Ing archetypes, the following requests concerning the management and provision of metadata and terminologies are stated:

- Develop metadata with formal semantics that are generic enough to be interoperable but specific enough to represent discipline-specific concepts (e.g. model predictions, HPMC data, data quality aspects of 'field data') (cf. chapter 3.5, 3.6, 3.8)
- Provide (semantic) documentation of all steps of data production (i.e. provenance tracking, cf. chapter 3.4) and outcome including metadata adjustments, raw data, and intermediate data to describe and predict the behaviour of real and virtual experiments (cf. chapter 3.2)
- Metadata and ontologies should describe the validation and quality-control processes of research software with a possible focus on simulation software (cf. chapter 3.3)
- Develop suitable vocabularies to describe the operational function of devices used to collect, analyse, and visualise data generated and/or processed in industrial production workflows (cf. chapter 3.7)

Based on these requests, we identify three specific tasks that are outlined in the following tasks.

Task S-3-1: Provision of tools for standardising metadata based on application profiles. (TUDA, RWTH, KIT)

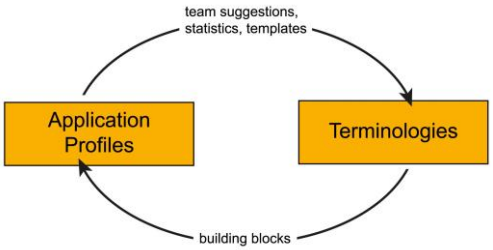


Figure 3.9.1 Synergies between application profiles and terminolgies

The variability of methods used in engineering implies a constant need to standardise application-specific metadata. NFDI4Ing therefore provides the tools needed to quickly develop flexible, interoperable, and reusable metadata standards according to the approach described in chapter 2.1. Using existing terminologies as well as those provided by S-3-2 as basis (cf. Figure 3.9.1), we will offer a smart interface that allows to find and select suitable terms and assemble them into application profiles. Term suggestions will take into consideration statistics on term usage within the defined standards and take into account settings for filtering and preference of underlying vocabularies that can be set based on community recommendations. If no fitting term is found, a custom term may be specified as provisional building block, automatically triggering a term request for the terminology service (S-3-2). To ensure that standards are shared and reused, a connected repository will archive and index the standards as well as make them available for reuse and adaptation. While the metadata standards will need to be defined within the task areas that require them, information experts from metadata services will support the scientists collaboratively. They also assist in developing tools for integration of metadata standards into scientific workflows (e.g. harvesters for extracting metadata from available sources like file-headers, log files, software repositories and tools for quality control of metadata).

Task S-3-2: Provision of a Terminology Service. (TIB)

The terminology service (TS) will develop subject-specific terminologies, simultaneously fueling S-3-1 with terms and using the application profiles created in S-3-1 as a basis for refining formal ontologies for the archetypes presented above (cf. Figure 3.9.1). The terminology service will implement technical infrastructure for access, curation, and subscription to terminologies, offering a single-point-of-entry to terminologies in NFDI4Ing. A RESTful API will be implemented to provide access to terminology in a uniform way regardless of their degree of complexity. The TS will allow the handling of requests such as custom terms for new terminologies or updates of existing terminology by stakeholder communities. For this, a ticket-based help-desk will be provided. The TS will also allow users to evaluate, test, and finally deploy the most suitable tools for reliable management of subject-specific terminologies such as the web-based working environment VoCol [109]. The service will include transformation tools from textual and tabular documents into semantic formats, a linked data interface, and terminology integrity checks and validation. Within the TS, a tool to enable semantic terminology subscription and notification will be developed. The subscription tool will recognise new matching terminologies according to the specifications of a user, activate defined processing of the terminology if requested, and inform the subscriber via email about the availability of new terminologies with a link to access them. The tool will also notify users when new terminologies of interest are available or have been changed recently.

Task S-3-3: Provision of a Metadata Hub. (TUDA, RWTH, KIT)

NFDI4Ing will provide a repository for publishing the (full) metadata sets describing actual research data according to the application profiles and ontologies developed in S-3-1 and S-3-2. This metadata hub will enable highly specific queries that can be used to access research data stored in repositories that do not support the full scope of the supplied metadata. The use of DOIs enables the linking between all components of the research process incl. experiments, raw data, software, subject-specific metadata sets, and the tracking of usage and citations. This task will also enable applications that require analysis of metadata. Applications include data-level metrics and provenance tracking based on published metadata sets as well as extraction of statistics on term frequencies fuelling the smart interface of the standard generator described in S-3-1.

Measure S-4: "Repositories and storage" (KIT, RWTH, TUDA)

When accessing or generating data, when analysing or publishing data, engineers are dealing with very different experimental settings and research circumstances. Besides, the number of research data management solutions is increasing. Different kinds of repositories, storage systems, and recommendations for archiving exist but are not shaped on the discipline specific requirements. Engineers are looking for flexible and adaptable repositories and storage solutions that are easily integrable in their individual workflows. For repositories, we distinguish on the one hand institutional repositories (e.g., RADAR, Zenodo, KITopen) and on the other hand community

specific repositories (e.g., NOMAD for material sciences). This is a very useful method for publishing research data and can be part of a long-term archiving plan, but the required flexibility for heterogenous data, which is still in the mode of changing and analysis, is missing. Therefore, we provide tools and develop interfaces for integration and interoperability of repositories. Building up community specific repositories, which serve as best practises and tell stories of successful research inside the engineering sciences, will contribute to the overall goal of realising the FAIR principles. Harmonisation and standardisation of protocols and interfaces is one of the tasks planned in this measure to implement the "A" in FAIR and to enable engineers to share their data easily across local storage and repositories.

Especially when it comes to experimental test benches, the means of generating data are often part of the initial research question. This leads to the common practice that engineers operate their own storage infrastructures to support their scientific workflows. Existing storage infrastructures available to the researchers hence are distributed among local, national, and international providers from academia and industry. While this wide range of offers greatly encourages sovereign scientific work, it also leaves engineers widely unsupported by infrastructure providers when it comes to their individual scenario and makes it harder to reuse data collected by other research groups. The heterogeneity of data generation workflows and storage formats poses a challenge to researchers in engineering. Within this measure, it is therefore our goal to define best practices and tools for storage, exchange, and long-term preservation for data of varying quality and volume to foster reusability of research data in this highly decentralised environment.

In order to tackle these challenges, we identify four larger tasks that are described in the following:

Task S-4-1: Establishment and maintaintenance of best practices and recommendations for community specific repositories & storage solutions. (RWTH, KIT, TUDA)
We will compile a catalogue of data formats and storage technologies for engineers considering sustainability, interoperability, and accessibility including a review of suitable technologies. First requirements will be defined by task area ALEX (cf. capter 3.2) and CADEN (cf. chapter 3.4) (experimental test benches) and task area DORIS (cf. chapter 3.5) (large-scale scalability). Furthermore, we will deliver recommendations for building new repositories based on the continuous analysis of the state of the art and a collection of existing community specific repositories (to be registered in re3data) and provide tutorials and other training material in cooperation with measure S-6.

Task S-4-2: Development of software for federated storage services. (RWTH, KIT, TUDA)
We will develop a software stack based on existing solutions that implements repositories, harmonised protocols/interfaces, and best practices for operators of federated storage services that incorporate good scientific practice and the FAIR principles. Considering data storage, the software will take into account data management workflows but also authentication, user

management, and access for users from external organisations (in cooperation with measure S-5). With the definition in place, we will reach out to existing infrastructure providers to achieve a higher level of standardisation of storage infrastructures for engineering researchers. For repositories, the software will be designed and adapted for a specific archetype (in close collaboration with task area ALEX (cf. chapter 3.2) and CADEN (cf. chapter 3.4)) to meet the requirements of the different engineering research methods. The developed tools (metadata harvester, search&find) will access the repositories and can be used locally. To enable access of ELNs (elaborated in task area CADEN (cf. chapter 3.4)) to the repository, the software will provide the necessary interfaces and adaption to the design of the ELN. To enable cross-repository search by providing a specialised meta-search, the software will connect repositories and develop a search interface.

Task S-4-3: Development of a cost and distribution model for storage. (RWTH, KIT)

We will develop a cost and distribution model that allows participating institutions to share storage infrastructures and compensate for costs inflicted by acquisition and operation of the infrastructures when research data is accessed by researchers of the community.

Task S-4-4: Defining long-term archiving processes in engineering. (RWTH, KIT)

We will define prerequisites for long-term archiving in the engineering community. With all archetypes, we will define a set of curation criteria that allows controlling long-term preservation processes in addition to bitstream preservation.

Measure S-5 "Overall NFDI software architecture – data security and sovereignty" (RWTH, KIT, TUDA, TIB)

The services established within this task area but also the activities within other task areas will eventually contribute to the overall vision of the NFDI. To achieve this goal, all offered services should be based on common grounds following consolidated practices. The goal of this measure is to establish these best practises for securely providing and accessing (meta)data in a distributed architecture operated by service providers from different scientific communities, so called data spaces, as for example implemented by the IDSA [110]. The tasks of this measure will consider the following topics: Authentication and authorisation infrastructures (AAI), role management, data space interfaces, connectors as well as data identifiers and discovery.

Task S-5-1: Harmonisation of authentication and authorisation infrastructures. (RWTH, KIT)

We will strive for harmonisation of authentication and authorisation mechanisms including comprehensive recommendations and best practice implementations that can be reused by federated services in the consortium. As such, we will harmonise existing approaches for federated AAIs like DFN-AAI on national, or EDUGAIN on international level, with initiatives such as ORCID that aim at identification of researchers throughout their career at different institutions. If necessary, we combine existing approaches using present reference architectures like AARC.

Task S-5-2: Development of a role and access management service. (RWTH)
We will develop a role and access management service that uses AAI to identify users and provisions access to other services offered by the federation. Thus, we enable researchers to form workgroups across institutional boundaries. We will implement a service for role and access management that is able to provide access rights to other services within a federation.

Task S-5-3: Development of interfaces and marketplaces towards a common federated data space. (RWTH)
We will develop a set of interfaces and a marketplace-like service infrastructure that allows single engineering researchers to include their research data in a common federated data space. Therefore, we will define a set of interfaces as a reference architecture of a federated data space for research data that is comparable, or better compatible, with existing interfaces for repository harvesting, linked data interfaces, or the International Data Space of IDSA [110]. We will harmonise interfaces of all NFDI4Ing services into an overarching service architecture, such that repositories, metadata, and storage services fit the engineers' challenges captured by the archetypes. Thus, we enable engineers to decide sovereignly, which parts of their data they share with whom, and implement applications to translate between otherwise unconnected data space implementations.

Task S-5-4: Building a federated linked data space for discoverability. (RWTH, KIT, TIB)
We will connect decentralised engineering-related (meta)data services and repositories with interfaces and services for data identifiers and discovery mechanisms in a federated data space. It is important to make interfaces not only human operable but also support comprehensible machine-operable interfaces to eventually become a foundation for a scientific knowledge graph. We will evaluate how existing systems like handle.net, DOI or ARK can serve as part of a data discovery service within the data space and if neccesary translate between these PID systems.

Measure S-6: "Community-based training on enabling data-driven science and FAIR data" (RWTH, TIB, KIT)

Community-based training is key to enable researchers of all engineering disciplines to work successfully in data-driven environments. NFDI4Ing will specifically address the engineering community with trainings tailored towards specific challenges encountered during an engineer's typical research, e.g. handling large amounts of data or frequent transitions between empirical and simulation-based research environments. All deliverables will be developed and evaluated in close collaboration with local RDM training and support units at engineering research institutions, e.g. the TU9 community for RDM Training Materials for the Engineering Sciences [111]. We further collaborate with the respective international working groups like CESAER and FAIR4S [112] (cf. chapter 1.4).

Task S-6-1: Defining and delivering RDM training contents for engineers. (RWTH, KIT)
Training contents such as workflow-specific RDM guidelines and concepts will originate from the archetypes' and other task areas' work and are compiled under the lead of this task. They are structured into (a) topics related to data processes and (b) topics regarding work organisation. NFDI4Ing will deliver the following training contents:

(a) Data process related topics.

- Data life cycle: from data management planning to archiving research results
- Technologies and tools: from data creation to data archiving
- Handling: from database architectures to models and preprocessing routines

(b) Work organisational topics.

- Organisational theory for research teams in engineering and data quality management
- Best practises and insights for RDM workflows gathered from community participation

In addition to these already identified training contents, an ongoing requirements analysis of the specific needs of the engineering community is crucial. This is especially important considering the importance of training to increase the awareness towards RDM and drive a cultural change. Therefore, an open and interactive communication platform for feedback and requirements will be established (cf. CC-1 and CC-4) in order to assist the demand-oriented development of training materials and courses. New materials will be published under free licences and thus be provided as Open Educational Resources (OER) for reuse and further development by the members of NFDI4Ing and the whole scientifiy community.

Task S-6-2: Defining RDM training formats for engineers. (TIB)
NFDI4Ing will offer many different formats of RDM trainings for different needs, such as: workshops, short seminars, webinars and other eLearning units. A special focus will be set on workshops carried out in cooperation with the Data and Software Carpentries, a worldwide community teaching foundational coding and data science skills to researchers [113]. A combination of online and face-to-face elements, the so-called blended learning approach, is another format that will be considered as it helps face-to-face training to become much more meaningful in practice – using eLearning material for preparation of the trainees who can keep up in their own pace. In addition, the concept of the FAIR Study Group [114], which has already been established at TIB Hannover, is a favoured alternative to classical workshop formats. Train-the-trainer measures are a further effective possibility for knowledge management and leveraging resources into the community. Such train-the-trainer workshops will be developed and conducted.

Measure S-7 "Automated data and knowledge discovery in engineering literature" (TUDA, KIT, TIB)

As outlined in the FAIR principles, the aim of research data management is not only to enable humans to find and interpret data, but also to enable automated statistical analysis of data by machines, i.e. data mining and machine learning. In engineering science, the vast majority of

available data is hidden in textual publications, e.g. in the NFDI4Ing survey three out of four research groups stated that they do not publish results in data or code repositories [1]. Additionally, open access publishing is rather uncommon in engineering [115]. Therefore, generally available engineering research data in the sense of structured representations of information such as tables or datasets is in fact not accessible for state-of-the-art scientific methods. Well-prepared and high quality (meta)data is a prerequisite for algorithmic approaches from fields like high performance data analytics, machine learning, artificial intelligence, and text mining.

Task S-7-1: Providing a service to enable text and data mining in engineering literature. (TUDA)

Based on the communities' and archetypes' requirements, we will build a large digital corpus from articles, proceedings, reports, grey literature, dissertations, monographs, and other documents relevant for engineering. On one hand, this contains open access materials that have to be harvested from different sources, on the other hand there are subscription materials that have to be licensed. To this end, all documents will be converted to machine readable text, using OCR and layout recognition where needed. Those documents already available in machine-friendly formats (e.g. JATS [116], [117]) will be harmonised in a structured XML-format as far as possible. As required by engineers (S-7-3), parts of the corpus will be further processed into this target format. Where possible, semantic facts and topics will be identified and used to enrich data with structured elements (e.g. named entities, document structures), addressing especially engineering specific challenges (like formulae, unit conversion, description of materials, references to technical standards). By using topic modelling on the data, we will help engineers to identify and cluster relevant research subjects and data from digital documents. This task is strongly linked to measure S-3 utilising the developed ontologies. We will establish general indexes and aggregate significant collections for conducting searches on full text and for leveraging the full potential of text and data mining algorithms [118]. Intelligent user interfaces and, if lawfully possible (S-7-2), a variety of download options via a single point of entry will be provided. First concept and workflow implementations for this task have already been and will be further developed at ULB Darmstadt [119].

Task S-7-2: Providing guidelines for the legal aspects of text and data mining. (TIB)

In this task, we will provide the legal basis for copyright clearing. Researchers will be able to check if and how particular publications can be made available in digital form for the sake of TDM. In particular, this tool will provide guidelines concerning publications with different licence models, including open access licences and the compliance with copyright regulations. This task will be carried out at TIB Hannover, which is specialised in licensing engineering literature.

Task S-7-3: Applying data science methods for knowledge discovery. (KIT)

In the field of data science, there is a huge need for accessing as much data as possible from any source to get new insights in the research work. From several archetypes there are

requirements regarding knowledge discovery (e.g., ALEX (cf. chapter 3.2), CADEN (cf. chapter 3.4), ELLEN (cf. chapter 3.6)). We will evaluate how the information gathered with the help of tasks S-7-1 and S-7-2 can be used to apply data science approaches to fulfil these demands. Therefore, we access data from legacy documents and combine them with newly generated research data to facilitate the application of machine learning algorithms [120]. The documentation of this evaluation will provide best-practice examples to be used for other research topics in engineering sciences. Thus, we foster the use of TDM and data science also for engineers unfamiliar to these methods.

Possible risks of implementation

The NFDI4Ing partners combine a deep knowledge of the engineering community with the experience of providing reliable service infrastructures, also with a deep understanding of RDM. The partners have applied the FAIR principles from the very start of the RDM movement. As such, the risk that services might not meet the FAIR principles can be minimised. As a central risk, the services face the possibility to be not accepted by the engineering research community. NFDI4Ing faces this risk providing a new, interactive approach by the introduction of method-related archetypes and community clusters carried out by engineers and their communities, both serve as evaluators and multipliers for the task area Base Services. The development and provisioning of the base services is highly interlocked with the evolution of the Research Data Commons [152]. This will guarantee the technical interoperability within the German NFDI and to international developments. One risk lies in the different time scales of the provisioning of the NFDI4Ing base services which are required timely and a slower consensus process for the Research Data Commons. In NFDI4Ing this will involve continuous subsequent improvement of the base services to meet common and open formats, technical protocols and interfaces.

3.9.3 Synergies and demarcations with regard to other task areas

The archetype task areas will make extensive use of and interact closely with the service measures of this task area, as it is mentioned above for each task. Maximal synergy effects arise from this modular approach. There are clear demarcations between the archetypes' and this Base Services task area, cf. chapter 2.3.

3.10 Community Clusters (RWTH, TUDA, TUD, TUBS, KIT, DLR)

The task area Community Clusters pursues two major objectives. Firstly, it serves as the gateway to the research engineering communities and their needs. For this, five community clusters are defined in NFDI4Ing following the DFG classification and positioned transverse to the other task areas as shown in Figure 1.2.3 Structure of the consortium's work programme. The matrix structure guarantees close interaction and efficient communication between the task areas for archetypes and services on the one hand and to the engineering communities on the other hand. This allows the continuous adaption to changing needs in a client-oriented manner. As this client orientation is a guiding principle of NFDI4Ing from its initiation, the archetype concept covers the vast majority of engineering communities, as the survey result shown in Figure 1.2.2 supports. Beyond this current status, the task area community clusters itself is adaptable to enable the inclusion of communities that are currently less present in NFDI4Ing, or beyond the scope of engineering alone. The measure CC-1 "Initialisation of communication services", lays the foundation to achieve this objective.

Secondly, the task area Community Clusters works on measures to disseminate the developed services and best-practices into the research world. On the one hand, this comprises RDM-related training for engineers from students and PhD students to research assistants working at higher education or non-university research institutions. The derived tasks are addressed by the measure CC-2 "Education". One the other hand, this means intensive user and community involvement in the development processes, achieved by means of an agile approach as below:

The community clusters gather cluster-specific information about RDM-related practices and requirements and direct their specific user communities to available solutions within NFDI4Ing. Developed services are tested by the research communities and disseminated via application and training. Emerging new needs trigger the development of solutions in the archetype task areas. The measures CC-3 "Activation and collaboration" and CC-4 "Participation" are designed to ensure vast community coverage and intensive engagement. Since RDM will be normative to some extend, NFDI4Ing will work on standardisation together with the engineering community in academia and industry, nationally and internationally, within the consortium and across the consortia of NFDI. This task is addressed in the measure CC-5 "Standardisation". As an appropriate academic platform to exchange ideas, the "Journal of Research Data Management in Engineering Science" will be initiated and operated in measure CC-6 "Journal".

3.10.1 Competence and expertise

"Mechanical and industrial engineering" (DFG research area 41)

This community cluster is coordinated by Prof. Robert Schmitt (RWTH). He is an active member of the *German Academic Association for Production Technology* (WGP) and the *International Academy for Production Engineering*. These research societies comprise important universities, institutions and research projects (e.g. CRC) in terms of contributions for NFDI4Ing and are strongly involved in DFG funded research. The WGP has been already consulted during the NFDI4Ing consortium's forming process. Next to WGP, there are several additional research societies such as MHI (assembly, operation, and robotics), *acatech*, and VDI/VDE, which will be approached to ensure that all the important stakeholders of the community cluster are engaged from the early project phase onwards.

"Thermal engineering and process engineering" (DFG research area 42)

This community cluster is coordinated by Prof. Peter Pelz (TUDA, speaker CRC 805). His experience in international engineering standardisation processes (representative of DIN at ISO/TC 117) will support the standardisation tasks of NFDI4Ing. To activate the community, the community cluster has been and will continue to work in several directions: Presentation of NFDI4Ing to the *Fakultätentag für Maschinenbau und Verfahrenstechnik* (FTMV); Integrating CRC 1194, CRC 805, DECHEMA/ENPRO projects as early adopters; Establishing transfer projects on research data management in industry; Jointly organising workshops on NFDI for the sub-communities of thermal turbomachinery, fluid energy machinery, fluid mechanics, also including process engineering with the support of DECHEMA; Integrating data literacy and RDM competences into the curriculum of engineering education.

"Materials science and engineering" (DFG research area 43)

This community cluster is led by Prof. Britta Nestler who was active in the *commission of engineering sciences @ BW2025* and in the expert group for PhD programmes of the MWK. She is a member of the DFG Cluster of Excellence: *POLiS* at KIT and the University of Ulm and the BMBF Cluster of Competence *FestBatt*. She is also setting up the Science Data Center *MoMaF* at KIT. Several community activities were organised together with the *Gesellschaft für Angewandte Mathematik und Mechanik* (GAMM) and *Deutsche Gesellschaft für Materialkunde* (DGM). A joint workshop of *FestBatt* and members of NFDI4Ing will take place from 19. – 20.11.2019 in Karlsruhe. In 2019 the *Karlsruhe Data Infrastructure for Material science* (kadi4Mat) will be established.

"Computer science, information technology, electrical and systems engineering" (DFG research area 44)

This community cluster is particularly heterogenous including various disciplines such as automation and robotics, biomedical systems technology, human-machine systems, electrical

engineering, transport and computer sciences. Co-spokesperson Prof. Regine Gerike is a member of various network institutions, e.g. DVWG, FGSV as well as the Scientific Board of the *Scientific Information Service for Mobility and Transport Research* (FID Move). She will be supported by co-spokesperson Christian Langenbach from DLR (Energy and Transport Systems), Prof. Detlef Stolten (Energy Systems), Prof. Frank Kirchner (Robotics), and Prof. Andreas Schütze on behalf of the *Fakultätentag für Elektrotechnik und Informationstechnik* (FTEI). The Cluster of Excellence *SE²A* committed its interest and input to this cluster.

"Construction engineering and architecture" (DFG research area 45)

This cluster is lead by Prof. Manfred Krafczyk (DFG-Fachkollegiat FK 410, member of Arbeitskreis Bauinformatik). He introduced the NFDI4Ing consortium to the *Fakultätentag für Bauingenieurwesen, Geodäsie und Umweltingenieurwesen* (FTBGU) which unanimously guaranteed support of NFDI4Ing. The *Dekane- und Abteilungsleiterkonferenz für Architektur, Raumplanung und Landschaftsarchitektur* (DARL) also agreed to cooperate with NFDI4Ing. The Vice-President of DARL, Prof. Frank Petzold (Chair for architecture informatics at TUM and member of AK Bauinformatik) will contribute to organise the cluster focusing on the research community of architects in close cooperation with Prof. Krafczyk.

3.10.2 Measures

In all of the measures described below, all community Clusters will be active, thus only the lead responsibility or, if applicable, the shared lead of the measure is indicated.

Measure CC-1 "Initialisation of communication services" (TUDA)

This measure provides the tools for interacting with the research communities. In close collaboration with the general dissemination activities in NFDI4Ing as described in chapter 3.11, we will set up, maintain and constantly update facilities and tools that are used by the community clusters in NFD4Ing for intensive engagement with their research communities and for maintaining their involvement into the development of the NFDI4Ing services.

Task CC-1-1: Set up of a community hub embedded into the general NFDI4Ing website, allowing to set up community-specific sites in order to address the heterogenous communities as close as possible to their individual needs, structures and practices.

Task CC-1-2: Integration of tools into the community hub, e.g. for setting up mailing lists, newsletters, chats, wikis, blogs, webinars, competitions, polling systems and ticket systems for user requests as well as tools for conducting user surveys, for promoting achievements in NFDI4Ing but also from the research community outside the NFDI4Ing consortium.

Task CC-1-3: Provision of further material for community engagement. This includes templates for roll-ups, posters, leaflets, presentations, newsletters and press releases.

Measure CC-2 "Education" (RWTH, TUDA, DLR)

As stated in key objective 7, engineers will profit from access to RDM-related education and available domain and application specific best practices. The aim of this measure is to provide learning materials (e.g. lecture slides, exercise materials, multimedia documents) for various training concepts (e.g. webinars, blended learning, train-the-trainer workshops, learning nuggets) and best practices for integration of data literacy and RDM competencies in the education of academics and students as our two main target groups in engineering sciences.

Task CC-2-1: Training materials and concepts will be collected and prepared (rehashed for community specific contents or edited for different learning formats).

Task CC-2-2: Access to training materials will be provided through self-developed communication networks (CC-1 and CC-3), in the wake of community specific events (CC-4) and through other established channels (e.g. social media, podcasts). The creation of new training materials (S-6-1) and new training concepts (S-6-2) will be done in close cooperation with task area Base Services S-6 (in which trainings will be conducted) and is supplemented with subject-specific application examples and tested for applicability in various disciplines and workflows.

Task CC-2-3: Promotion of established best practices for introducing students to basic RDM concepts as a recurring topic for the study programmes' curricula (e.g. managing data and code snippets via repositories, version control, attaching basic metadata, referencing the contents in subsequent work). This includes the evaluation of platforms and environments for programming exercises and test certificates as well as the assessment of options for integrating RDM in grading processes. To foster a widespread inclusion of RDM in engineering curricula, the task's outcomes will be discussed and disseminated through the association of engineering faculties.

Measure CC-3 "Activation and collaboration" (TUBS)

This measure focuses on the activation and organisation of communication between the individual community clusters and the archetypes. Building on the surveys already conducted and described in chapter 1.2, all RDM-related user requests from the clusters will be collected, harmonised and communicated to the archetypes. Their solutions are efficiently propagated back to the community clusters using the communication infrastructure provided by CC-1. The co-spokespersons identify multipliers and key representatives for each of the community clusters to establish community boards and compile current RDM practices and needs using regular polls. All tasks are adapted to the specific needs of the individual community clusters.

Task CC-3-1: Initiation and management of a community board for each community cluster.

Task CC-3-2: Implementation of regular user surveys per cluster. Respective communities are invited at least annually to submit problems or ideas for services to be developed by NFDI4Ing.

Task CC-3-3: The result of the surveys will be presented to the task area archetypes in terms of an agile backlog. The task area archetypes will decide jointly with task area community clusters

(cf. Figure 1.5.1) which issues they are going to work on. If necessary, they will apply for seed funding.

Task CC-3-4: Innovative RDM solutions from external communities or other NFDI-consortia will be evaluated and, if appropriate, be awarded, e.g. at the annual NFDI4Ing conference.

Measure CC-4 "Participation" (TUD)

This measure is concerned with events that will be organised or supported by NFDI4Ing in order to reach out to the communities, to raise awareness, identify requirements, inform about NFDI4Ing, and to attract new users.

Task CC-4-1: Initiation, hosting and organisation of community meetings (one per year and community cluster), addressing either whole clusters or dedicated sub-communities. The events may contain workshops or trainings dedicated to specific topics.

Task CC-4-2: Initiation and organisation of presentations, sessions, stands and side events at scientific conferences, fairs and workshops. These may range from traditional presentations to interactive sessions to feed into the development cycle.

Task CC-4-3: Presentations at non-scientific events such as meetings of faculty associations ("Fakultätentage") or DFG round-table discussions. This serves mostly to raise awareness for the necessity of RDM and to anchor NFDI4Ing as a stable counterpart for the communities.

Measure CC-5 "Standardisation" (KIT, DLR)

To establish successful services for the community there is a need to develop standards. These standards enable the interconnection of researchers by harmonisation of tools and services, as appealed for by the DFG [121]. The development of those standards must be in conjunction with the community as well as other consortia (cf. chapter 2), research centres (like Fraunhofer, Max-Planck-Gesellschaft (MPG) and Helmholtz Association) and industrial syndicates (like VDMA, VDI, VDE). We see a broad identification of the community with the archetypes (cf. chapter 1). Therefore, this measure interlinks the community with the archetypes within NFDI4Ing and the outside research communities to efficiently identify needs and opportunities for standardisation. User tests and feedback have a major influence on the standardised services to be developed and provided by NFDI4Ing.

Task CC-5-1: We will organise the development of community related guidelines and standards (in extension to measure CC-2 and jointly conducted with task area S-6) on the handling of research data for individual research communities in engineering. The special focus is on guidelines and standards for collaborative research within academia and joint research between academia and industry. For the latter the challenging tasks of confidentiality vs. open science will be addressed in the guidelines and standards.

Task CC-5-2: Further work on cross-cutting guidelines and standards with other consortia. For this we will trigger and organise workshops devoted to this special cross-cutting topic.

Task CC-5-3: In communication with the community and other NFDI consortia, we will identify reasonable interfaces which are worthy to be standardised. Those standards will be made visible to the communities as requests for commentssimilar to organisations like e.g. IETF.

Measure CC-6 "Journal" (TUDA)

This measure provides a much needed platform for communication of RDM-related topics within the engineering community in form of a scientific "Journal of Research Data Management in Engineering Science". By means of a peer reviewed process, both quality of content as well as reputation for the authors will be achieved. The journal will present best practice examples of RDM-related services, standards, and education. The readership is the engineering community in academia and industry, national and international. The journal will be open access and electronic. The concept of an overlay journal is considered. The editorial board will include co-spokespersons of the task area community clusters.

Task CC-6-1: Finalise the conception regarding target direction, editorial board, and review processes. Create the journal's infrastructure, and initialise, promote and operate the journal.

Task CC-6-2: Evaluate the journal's success regularly by means of established metrics and adapt the concept as necessary.

Possible risks of implementation

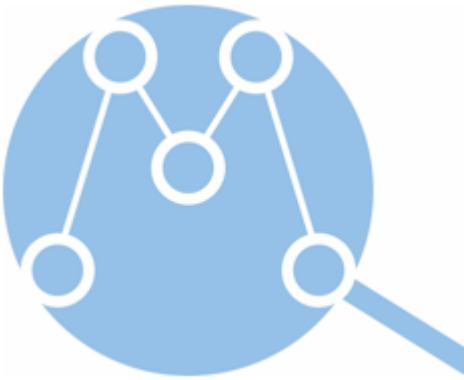
Internal project risks comprise qualitatively inferior results from individual measures, due to a lack of coordination between several task areas and measures or by insufficiently allocated resources. Those internal risks will be addressed by thorough project management and project controlling supported by close cooperation of a coordinator of the task area Community Clusters (Prof. Pelz) and the Management task area. In addition, external risks such as insufficient activation or participation of communities, caused by low acceptance of the utilised communication channels are addressed by definition of targets, regular reporting, balanced scorecards, and progress monitoring of KPIs. Furthermore, the community cluster spokespersons will distribute surveys on a yearly basis to evaluate community involvement and the communities' perception about the project. The management consolidates the evaluations across community clusters.

3.10.3 Synergies and demarcations with regard to other task areas

As introduced in chapter 3.10.1, the community clusters ensure that users maintain an active role and their needs are monitored on a regular basis. This comprises consolidating the feedback from pilot users testing prototypical solutions (cf. chapters 3.2 - 3.8), but also refining cluster specific requirements and triggering the development of new solutions in the archetype task areas (CC-1, CC-3). An especially important synergy is the provision of community tailored learning materials and concepts for data literacy and RDM (CC-2). The community clusters are supported by the

expertise and resources of the NFDI4Ing management in organising events, communication and project management (M-3 to M-6).

3.11 Management (RWTH, TUDA)



NFDI4Ing's management is already up and running. The consortium has originally been founded in 2017, and the management team has acquired two years of experience actively engaging with the engineering communities [122]. Our experience shows that it is important to keep NFDI4Ing's internal organisational structure lean enough to be manageable and, at the same time, tight and personal enough to retain the motivation and engagement of all applicants and participants at very high levels.

As layed out in section 1.5, NFDI4Ing's management will be responsible for the administration of seed funds. Likewise, the material expenses for all task areas will be centrally managed (cf. M-1, M-2, M-6, M-7).

3.11.1 Key objectives

The task area Management supports all task areas in their administrative, organisational, and communicative actions. Key objectives of the task area Management are:

- Fostering effective internal and external communication,
- Providing organisational assistance to NFDI4Ing events,
- Coordinating the interaction with other consortia and the NFDI bodies,
- Executing quality management,
- Providing project management consultancy,
- Being a reliable contact for the DFG and overseeing the disbursement of funds,
- Lending support for purchasing and contracting expenses.

3.11.2 Measures

Measure M-1 "Effective internal and external communication"

NFDI4Ing's online and offline meeting structure requires communication and reporting (cf. chapter 1.5). The management will set up the agendas for the board meetings and will ensure that the minutes are kept. It will also help with organising the regular conference calls. When it comes to communication, NFDI4Ing uses centralised resources. The management is responsible for the communication infrastructure. A number of centralised communication resources is already in place, such as the corporate design, the homepage www.nfdi4ing.de, which includes information about research data management projects and events, conference call management software,

extensive mailing lists, and a SharePoint. We will further develop these resources in close collaboration between the management and the task area Community Clusters (cf. chapter 3.10 and CC-1). This includes the dissemination of task area activities via a newsletter and target group specific social media activities. The management also supports the community-based trainings provided by the task area Base Services (cf. chapter 3.9, S-6). Finally yet importantly, the management will coordinate the publication of NFDI4Ing's work and achievements. The funds for NFDI4Ing's communications, publications, and outreach will be allocated to the task area Management.

Measure M-2 "Organising events"

One of the management's responsibilities lies with organising each year's main event, the general meeting of members. Since the general meeting of members is open to the public and shall attract interested persons from the outside, the management will reach out both within and outside of NFDI4Ing. NFDI4Ing's management will also lend its support for organising and drawing attention to the Community Clusters' events. The funds for NFDI4Ing's general meeting as well as for the Community Clusters' events will be allocated to the task area Management.

Measure M-3 "Interaction with other consortia and the NFDI bodies"

The management coordinates the interaction with other NFDI consortia and supports the executive board´s participation in NFDI's general meeting of consortia. The management also supports NFDI4Ing's work programme on cross-cutting topics in research data management that are to be addressed between consortia (cf. section 1.3). To this end, the management is tracking and reviewing collaborations and it oversees adherence to the Berlin declaration [4].

Measure M-4 "Quality management"

NFDI4Ing's management will play an important role in ensuring the quality of NFDI4Ing's work. The management does not only oversee NFDI4Ing's meeting structure, but also the escalation processes (cf. Figure 1.5.1) and overall progress. To this end, the management uses agile methods. Work results of all task areas are continuously tracked and reviewed. Review cycles are kept as short as possible and will include meetings for retrospection. Focus is put on the user needs and on the appropriate prioritisation of requirements.

Measure M-5 "Fostering homogeneous project management"

As outlined in measure M-4, above, we employ agile methods in order to track and review work results. Agile methods are greatly facilitated by a common understanding of project management. Running a solid project management office across all task areas will support knowledge creation and will bring about innovation continuously and incrementally [123]. The management team will provide project management consultancy. It will lend support with monitoring and reporting tools and will continue developing project management templates.

Measure M-6 "Financial administration and disbursement of funds"

An important responsibility of NFDI4Ing's management lies with the internal disbursement and relocation of funds. The main applicant, RWTH Aachen University, will disburse the funds for the co-applicant institutions. The management will establish an overall monitoring and control of the financial figures. It will oversee the expenses and report them to the executive board, on the one hand, and – if required – to the DFG, on the other hand. The management will also issue the figures for NFDI4Ing's general meeting of members, for project management purposes, and for seed fund proposals. Decisions regarding the disbursement of the seed fund follow an agreed-upon procedure with three levels, all of which involve the management (cf. chapter 1.5). Both the seed fund proposals and the decisions leading to disbursement or non-disbursement will be documented and made available for external review.

Measure M-7 "Traveling, purchasing, and contracting"

NFDI4Ing's management will ensure the correct handling and administration of all material expenses. These include the expenses related to traveling, purchasing, and contracting. Prior to disbursement, the funds for these expenses will be allocated to the task area Management. Disbursement will be coordinated with the board of co-spokespersons and with the administrations of the local departments of the co-applicant institutions. NFDI4Ing's management will also lend support to the hiring process, if requested by and clearly in cooperation with the local departments.

Current Service Portfolio NFDI4Ing

Name of the service / service component	Link
b2share	https://www.fz-juelich.de/ias/jsc/EN/Expertise/SciCloudServices/B2Share/_node.html
bwSync&Share	https://bwsyncandshare.kit.edu/login
bwDataArchive	https://www.rda.kit.edu/
Computer Aided Engineering Software	http://www.scc.kit.edu/dienste/4276.php
DaRUS	https://www.izus.uni-stuttgart.de/en/fokus/darus/
DataCite	https://datacite.org/
Data Science Storage (DSS)	https://doku.lrz.de/display/PUBLIC/DSS+TUM-DSS+system+architecture
DOI Service	https://www.tib.eu/en/publishing-archiving/doi-service/
DSpace	https://duraspace.org/dspace/ , https://tudatalib.ulb.tu-darmstadt.de/
ePIC	https://www.gwdg.de/application-services/persistent-identifier-pid
eRIC	https://www.ub.tum.de/forschungsdatenmanagement
F*EX	https://fex.uni-stuttgart.de/index.html
GeRDI	https://www.lrz.de/forschung/projekte/forschung-daten/GeRDI/
GigaMove	https://www.itc.rwth-aachen.de/cms/IT-Center/Services/kompletter-Servicekatalog/Anwendungen-und-Prozessunterstuetzung/~essv/Gigamove/?lidx=1
GitLab	https://about.gitlab.com/
Invenio	https://invenio-software.org/
JuliaBase	https://www.juliabase.org/
JupyterHub	https://jupyter.org/hub
KITopen	https://www.bibliothek.kit.edu/cms/kitopen.php
KIT DataManager	http://datamanager.kit.edu/index.php/kit-data-manager
Leibniz DataManager	https://datamanager.tib.eu
LSDF	http://www.scc.kit.edu/dienste/11228.php
MATLAB / R	https://www.scc.kit.edu/produkte/3841.php
Mattermost	https://www.mattermost.org/
MediaTUM	https://mediatum.ub.tum.de/
Metadata-Extractor	https://github.com/drewnoakes/metadata-extractor
MyCoRe	https://www.mycore.de/
NextCloud / OwnCloud / Powerfolder / Sync&Share services	https://nextcloud.com/ https://owncloud.com/ https://www.powerfolder.com/de/
ORCID DE	https://www.orcid-de.org/
RADAR	https://www.radar-service.eu/
RDMO	https://rdmorganiser.github.io/ , https://tudmo.ulb.tu-darmstadt.de
Redmine	https://www.redmine.org/
Rosetta, simpleArchive, Archivematica	https://www.tib.eu/en/publishing-archiving/digital-preservation/ https://www.hrz.tu-darmstadt.de/forschungsdaten_management/simplearchive_hrz/index.en.jsp
TIB AV Portal	https://av.tib.eu

Abbreviations

AAES	American Association of Engineering Societies
AAI	Authentication, Authorisation Infrastructure
AARC	Authentication and Authorisation for Research Collaborations
ABAC	Attribute Based Access Control
API	Application Programming Interface
BMBF	Bundesministerium für Bildung und Forschung / Federal Ministry of Education and Research
BPA	Blueprint Architecture
CAD	Computer Aided Design
CAMS	Copernicus Atmospheric Monitoring Service
CESAER	Conference of European Schools for Advanced Engineering Education and Research
CIRP	College International pour la Recherche en Productique / International Academy for Production Engineering
CoE	Cluster of Excellence
CRC	Collaborative Research Centre
CRIS	Current research information system
CRISP-DM	Cross-Industry Standard Process for Data-Mining
CUBE	Concurrent Certification Center
DARIAH	Digitale Forschungsinfrastruktur für die Geistes- und Kulturwissenschaften / Digital Research Infrastructure for the Arts and Humanities
DARL	Dekane- und Abteilungsleiterkonferenz für Architektur, Raumplanung und Landschaftsarchitektur in der Bundesrepublik Deutschland
DC	DataCite e.V.
DFG	Deutsche Forschungsgemeinschaft / German Research Foundation
DFKI	Deutsches Forschungszentrum für Künstliche Intelligenz / German Research Center for Artificial Intelligence
DFN	Deutsches Forschungsnetz / German National Research and Education Network
DH-NRW	Digitale Hochschule NRW
DINI	Deutsche Initiative für Netzwerkinformation
DLR	Deutsches Zentrum für Luft-& Raumfahrt e.V./ German Aerospace Center
DMP	Data Management Plan(ning)
DOI	Digital Object Identifier

DSS	Data Science Storage
e.g.	for example / exempli gratia
e.V.	eingetragener Verein / registered voluntary association
ELN	Electronic lab notebook
EMP-E	Energy Modelling Platform for Europe
EOSC	European Open Science Cloud
ERP	Enterprise Resource Planning
EUDAT CDI	EUDAT Collaborative Data Infrastructure
FAIR	Findable, Accessible, Interoperable, Reusable
FAIR4S	EOSCpilot framework of FAIR data stewardship skills for science and scholarship
FAIRmat	FAIR Data Infrastructure for Materials Science and Related Fields
FestBatt	Competence Cluster for Solid State Batteries
FID move	Fachinformationsdienst Mobilitäts- und Verkehrsforschung / Scientific Information Service for Mobility and Transport Research
FIZ Karlsruhe	Leibniz-Institut für Informationsinfrastruktur
FOKUS	Kompetenzzentrum für Forschungsdaten / Competence Center for Research Data Management
FSC	Cluster of Excellence 2186 - The Fuel Science Center
FST	Institut für Fluidsystemtechnik
FZ	Forschungszentrum
FZJ	Forschungszentrum Jülich
GA	Gauß-Allianz e.V.
GCS	Gauss Centre for Supercomputing
GridKa	Grid Computing Centre Karlsruhe
GUI	Graphical User Interface
HeFDI	Hessian Research Data Infrastructures
HLRS	High-Performance Computing Center Stuttgart
HPC	high-performance computing/computation
HPMC	high performance measurements and computations
HRZ	Hochschulrechenzentrum / Central University IT Services
IAM-CMS	Institute for Applied Materials – Computational Materials Science (KIT)
IAM-ESS	Institute for Applied Materials – Energy Storage Systems (KIT)
IDSA	International Data Spaces Association
IEK	Institute of Energy and Climate Research (FZJ)

IEK-3	Institute of Energy and Climate Research / Techno-Economic Energy Systems Analysis (FZJ)
IETF	Internet Engineering Task Force
IFC	Industry Foundation Classes
IG	Interest Group
IoP	Cluster of Excellence 2023 - Internet of Production
IRB	Fraunhofer-Informationszentrum Raum und Bau IRB
IRL	Integration Readiness Level
IT	Information Technology
ITIL	IT Infrastructure Library
ITP	Fraunhofer-Institut für Produktionstechnologie IPT/ Fraunhofer Institute for Production Technology IPT
JARA	Jülich Aachen Research Alliance
JATS	Journal Article Tag Suite
KIT	Karlsruhe Institute of Technology
KPI	Key Performance Indicator
LRZ	Leibniz Supercomputing Center of the Bavarian Academy of Sciences and Humanities
LUH	Leibniz University Hannover
MaRDI	Mathematical Research Data Initiative
MoMaF	Science Data Center for Molecular Material Research
MPG	Max-Planck Gesellschaft
NFDI	Nationale Forschungsdateninfrastruktur / National Research Data Infrastructure
NFDI4Chem	NFDI for Chemistry
NFDI4Culture	NFDI Consortium for research data on material and immaterial cultural heritage
NFDI4Earth	NFDI Consortium Earth System Science
NFDI4Medicine	NFDI Consortium of the Medical Informatics Initiative (MII) and the German Centers for Health Research (DZG)
NFDI4MobilTech	NFDI Consortium for Mobility Technology
NFDI4MSE	NFDI Consortium National Research Data Infrastructure for Materials Science and Engineering

NFDI4Neuro	NFDI Neuroscience
ngTDP	Next Generation Traffic Data Platform
NIST	National Institute of Standards and Technology
OCR	Optical characater recognition
OD-Rex	NFDI Consortium Open Database for Real World Robot Experiments
OER	open educational resources
ORCID	Open Researcher & Contributor Identification
PDM	Product Data Management
PhoenixD	Cluster of Excellence 2122 - Photonics, Optics, and Engineering – Innovation Across Disciplines
PID	Persistent identifier
PLM	Product Lifecycle Management
POLiS	Cluster of Excellence Energy storage systems beyond lithium
PTB	Physikalisch-Technische Bundesanstalt/ Technische Bundesanstalt/The National Metrology Institute of Germany
RADAR	Research Data Repository
RBAC	Role Based Access Control
RB-Nr.	Review Board Number
RDA	Research Data Alliance
RDM	Research Data Management
RDM4Eng	Research Data Management in Engineering
RDMO	Research Data Management Organiser
RfII	Rat für Informationsinfrastrukturen / Council for Scientific Information Infrastructures
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen
SE²A	Cluster of Excellence 2163 Sustainable and Energy-Efficient Aviation
SimTech	Cluster of Excellence 2075 - Data-Integrated Simulation Science
SLUB	Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden / Saxon State and University Library Dresden
SSC	Steinbuch Centre for Computing
STL	Stereolithography CAD file format / Standard Triangulation Language
STO	The Standards Ontology
STREAM	Semantische Repräsentation, Vernetzung und Kuratierung von qualitätsgesicherten Materialdaten
SU	Saarland Universität

TDM	Text and Data Mining
Text+	Language- and Text-Based Research Data Infrastructure
TFOS	Task Force Open Science
TIB	Technische Informationsbibliothek / Leibniz information centre for science and technology University library
TRR	Transregio
TU	Technische Universität / technical university
TU9	German leading Universities of Technology
TUB	Technische Universität Berlin
TUBS	Technische Universität Braunschweig
TUC	Technische Universität Clausthal
TUDA	Technische Universität Darmstadt
TUD	Technische Universität Dresden
TUM	Technische Universität München / Technical Universtity of Munic
ULB	Universitäts- und Landesbibliothek Darmstadt / University and State Library Darmstadt
US	Universität Stuttgart / University of Stuttgart
USB	Universal serial bus
VDE	Verband der Elektrotechnik, Elektronik und Informationstechnik
VDMA	Verband Deutscher Maschinen- und Anlagenbau / German Mechanical Engineering Industry Association
VDI	Verein Deutscher Ingenieure
VoCol	An Integrated Environment for Collaborative Vocabulary Development.
WGP	Wissenschaftliche Gesellschaft für Produktionstechnik / German Academic Association for Production Technology
WiGeP	Wissenschaftliche Gesellschaft für Produktentwicklung
WZL	Werkzeugmaschinenlabor / Laboratory for Machine Tools and Production Engineering
XML	Extensible Markup Language
Z-INF	Zentrales Informationsinfrastrukturprojekt / information infrastructure project
ZKI	Zentren für Kommunikationsverarbeitung in Forschung und Lehre

Bibliography and list of references

[1] G. Jagusch and N. Preuß. [Online]. Available: http://dx.doi.org/10.25534/tudatalib-104. [Accessed 30 09 2019].

[2] Deutsche Forschungsgemeinschaft, "Förderatlas 2018 - Kennzahlen zur öffentlich finanzierten Forschung in Deutschland," [Online]. Available: https://www.dfg.de/sites/foerderatlas2018/. [Accessed 11 10 2019].

[3] D. Strecker, "Daten und Skript für die Analyse von Kollaborationen zwischen potentiellen NFDI-Konsortien (Antragstellung 2019)," [Online]. Available: https://raw.githubusercontent.com/dorothearrr/NFDI_Netzwerk/master/NFDI_Netzwerk_2.png. [Accessed 11 10 2019].

[4] F. O. Glöckner, M. Diepenbroek, J. Felden, J. Overmann, A. Bonn, B. Gemeinholzer, A. Güntsch, B. König-Ries, B. Seeger, A. Pollex-Krüger, J. Fluck, I. Pigeot, T. Kirsten , T. Mühlhaus, C. Wolf, U. Heinrich, C. Steinbeck, O. Koepler, O. Stegle, J. Weimann, T. Schörner-Sadenius, C. Gutt, F. Stahl, K. Wagemann, T. Schrade, R. Schmitt, C. Eberl, F. Gauterin, M. Schultz and L. Bernard, "Berlin Declaration," [Online]. Available: https://doi.org/10.5281/zenodo.3457213. [Accessed 13 10 2019].

[5] [Online]. Available: https://www.rd-alliance.org/group/research-data-management-engineering-ig/case-statement/research-data-management-engineering-ig. [Accessed 11 10 2019].

[6] [Online]. Available: https://www.nist.gov/. [Accessed 11 10 2019].

[7] [Online]. Available: https://www.internationaldataspaces.org/. [Accessed 11 10 2019].

[8] [Online]. Available: https://www.cesaer.org/task-forces/task-force?id=34. [Accessed 11 10 2019].

[9] [Online]. Available: https://www.eudat.eu/catalogue. [Accessed 11 10 2019].

[10] Deutsche Forschungsgemeinschaft, "Protokoll: DFGNationale ForschungsdateninfrastrukturGovernance-Workshop 2019," 27 09 2019. [Online]. Available: https://www.dfg.de/download/pdf/foerderung/programme/nfdi/nfdi_governance_workshop_protokoll.pdf. [Accessed 14 10 2019].

[11] [Online]. Available: https://www.re3data.org/search?query=&subjects%5B%5D=4%20Engineering%20Sciences&types%5B%5D=disciplinary. [Accessed 27 09 2019].

[12] [Online]. Available: https://rd-alliance.github.io/metadata-directory/subjects/engineering.html. [Accessed 27 09 2019].

[13] [Online]. Available: https://zenodo.org/communities/rdm_training_engineering_sciences/?page=1&size=20. [Accessed 11 10 2019].

[14] [Online]. Available: https://www.forschungsdaten.info. [Accessed 11 10 2019].

[15] [Online]. Available: https://eosc-fair4s.github.io/. [Accessed 11 10 2019].

[16] [Online]. Available: https://schema.datacite.org/. [Accessed 11 10 2019].

[17] [Online]. Available: https://www.w3.org/TR/vocab-dcat-2/. [Accessed 11 10 2019].

[18] [Online]. Available: https://codemeta.github.io/. [Accessed 11 10 2019].

[19] [Online]. Available: https://rd-alliance.github.io/metadata-directory/subjects/engineering.html. [Accessed 11 10 2019].

[20] [Online]. Available: https://www.izus.uni-stuttgart.de/fokus/engmeta. [Accessed 11 10 2019].

[21] [Online]. Available: https://www.ptb.de/si/smartcom/d-si/v1_0_1/SI_Format.xsd. [Accessed 11 10 2019].

[22] [Online]. Available: http://www.qudt.org/. [Accessed 11 10 2019].

[23] [Online]. Available: https://www.bipm.org/en/publications/guides/vim.html. [Accessed 11 10 2019].

[24] [Online]. Available: https://www.allotrope.org/allotrope-framework. [Accessed 11 10 2019].

[25] [Online]. Available: https://www.wti-FRANKfurt.de/de/. [Accessed 11 10 2019].

[26] [Online]. Available: http://www.dkf-ev.de/dkf_thes.pdf. [Accessed 11 10 2019].

[27] [Online]. Available: https://www.w3.org/RDF/. [Accessed 11 10 2019].

[28] [Online]. Available: http://dublincore.org/documents/profile-guidelines/. [Accessed 11 10 2019].

[29] C. Harper, "Dublin Core metadata initiative: Beyond the element set," *Information Standards Quaterly,* vol. 1, no. 22, pp. 20-28, 2010.

[30] [Online]. Available: https://www.force11.org/group/fairgroup/fairprinciples. [Accessed 17 09 2019].

[31] J. H. Jones, Ed., Laboraty Informatics Institute Inc., The Complete Guide to LIMS & Laboratory Informatics, Atlanta: LiMSwiki.org, 2017.

[32] Z. Chen, D. Wu, J. Lu and Y. Chen, "Metadata-based Information Resource Integration for Research Management," *Procedia Computer Science,* vol. 17, pp. 54-61, 01 01 2013.

[33] R. R. Panko and S. Aurigemma, "Revising the Panko–Halverson taxonomy of spreadsheet errors," *Decision Support Systems,* vol. 49, no. 2, pp. 235-244, 01 05 2010.

[34] D. J. Smith, "Appendix 6 - Human Error Probabilities A2,," in *Reliability, Maintainability and Risk*, 8 ed., Oxford, 2011, pp. 295-397.

[35] J. Gray, D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt and G. Heber, "Scientific Data Management in the Coming Decade," *SIGMOND Recor,* vol. 34, no. 4, 2005.

[36] S. Hettrick, "Research Software Sustainability," in *Report on a Knowledge Exchange Workshop*, The Software Sustainability Institute, 2016.

[37] [Online]. Available: https://about.gitlab.com/. [Accessed 25 09 2019].

[38] A. M. Smith, D. S. Katz, K. E. Niemeyer and FORCE11 Software Citation Working Group, "Software citation principles," *PeerJ Computer Science,* vol. 2, p. e86, 2016.

[39] R. Di Cosmo, M. Fenner and D. Katz, "Software Source Code Identification WG," [Online]. Available: https://www.rd-alliance.org/groups/software-source-code-identification-wg. [Accessed 30 09 2019].

[40] [Online]. Available: https://www.docker.com/. [Accessed 25 09 2019].

[41] [Online]. Available: https://www.sylabs.io/docs/. [Accessed 25 09 2019].

[42] [Online]. Available: https://jupyterlab.readthedocs.io/en/stable/. [Accessed 25 09 2019].

[43] [Online]. Available: https://jupyter.org/hub. [Accessed 25 09 2019].

[44] V. Stodden and S. Miguez, "Best practices for computational science: Software infrastructure and environments for reproducible and extensible research," *Journal of Open Research Software,* vol. 2, no. 1, pp. 1-6, 2014.

[45] G. Wilson, J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt and T. K. Teal, "Good enough practices in scientific computing," *PLoS Computational Biology,* vol. 13, p. e1005510, 22 06 2017.

[46] T. Schlauch, M. Meinel and M. Haupt, "DLR software engineering guidelines," 2018. [Online]. Available: https://zenodo.org/record/1344612#.XaBA10FCSUk. [Accessed 11 10 2019].

[47] [Online]. Available: https://codemeta.github.io/. [Accessed 25 09 2019].

[48] [Online]. Available: https://schema.org/. [Accessed 25 09 2019].

[49] [Online]. Available: https://www.izus.uni-stuttgart.de/fokus/engmeta. [Accessed 25 09 2019].

[50] [Online]. Available: https://www.openhub.net/. [Accessed 25 09 2019].

[51] "Labfolder," [Online]. Available: www.labfolder.com. [Accessed 12 10 2019].

[52] "RSpace Community," [Online]. Available: www.researchspace.com. [Accessed 12 10 2019].

[53] "eCAT - The flexible electronic lab notebook that includes sample management," [Online]. Available: http://researchspace.com/electronic-lab-notebook/. [Accessed 12 10 2019].

[54] T. Bronger, "Umgang mit Forschungsdaten beim Forschungszentrum Jülich. Darstellung der Ergebnisse einer Online Befragung sowie Einezelinterviews," 2019.

[55] N. Brandt and M. Selzer, "Fragebogen zur Datennutzung in FestBatt/Questionary inside the competence cluster FestBatt," 2019.

[56] D. Van Heesch and et al, "Doxygen," [Online]. Available: www.doxygen.nl. [Accessed 12 10 2019].

[57] G. Valure and et al, "Natural Docs," [Online]. Available: www.naturaldocs.org. [Accessed 12 10 2019].

[58] Sun Microsystems, "Javadoc," [Online]. Available: docs.oracle.com/javase/9/javadoc/javadoc.htm. [Accessed 12 10 2019].

[59] F. Rhiem, "iffsamples.fz-juelich.de/documentation/," 2017–2019. [Online]. Available: https://scientific-it-systems.iffgit.fz-juelich.de/SampleDB/user_guide/citations.html. [Accessed 11 10 2019].

[60] T. Bronger, "JuliaBase," 2015-2019. [Online]. Available: https://juliabase.org/. [Accessed 12 10 2019].

[61] M. Selzer, "Kadi4Mat," 2019. [Online]. Available: www.iam.kit.edu/cms/Forschung_4519. [Accessed 12 10 2019].

[62] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari and B. Kozinsky, "AiiDA: automated interactive infrastructure and database for computational science," 2016. [Online]. Available: www.aiida.net. [Accessed 12 10 2016].

[63] L. Moreau and P. Groth, "PROV-Overview," W3C Working Group Note, 2013. [Online]. Available: www.w3.org/TR/2013/NOTE-prov-overview-20130430/. [Accessed 12 10 2019].

[64] R. Cyganiak and D. Reynolds, "The RDF Data Cube Vocabulary," W3C Recommendation, 2014. [Online]. Available: www.w3.org/TR/2014/REC-vocab-data-cube-20140116/. [Accessed 12 10 2019].

[65] S. Plantikow and et al, "ISO/IEC WD 39075 – Information Technology – Database Languages – GQL," [Online]. Available: www.iso.org/standard/76120.html. [Accessed 12 10 2019].

[66] [Online]. Available: https://gitlab.lrz.de. [Accessed 11 10 2019].

[67] [Online]. Available: https://github.com/. [Accessed 11 10 2019].

[68] [Online]. Available: https://workbench.ub.tum.de. [Accessed 11 10 2019].

[69] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. G. Ives, "DBpedia: A Nucleus for a Web of Open Data," in *In The Semantic Web - International Semantic Web Conference 2017 and Asian Semantic Web Conference 2017*, Springer, 2017, pp. 722-735.

[70] J. Hoffart, F. M. Suchanek, K. Berberich and G. Weikum, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," *Artificial Intelligence,* no. 194, pp. 28-61, 2013.

[71] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM,* no. 57, pp. 78-85, 2014.

[72] P. Ernst, A. Siu and G. Weikum, "KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences," *BMC Bioinformatics,* no. 16, p. 157, 2015.

[73] M. E. Vidal, K. Endris, S. Jazashoori, A. Sakor and A. Rivas, "Transforming Heterogeneous Data into Knowledge for Personalized Treatments - A Use Case," *Datenbank-Spektrum,* no. 19, pp. 95-106, 2019.

[74] N. Chakraborty, D. Lukovnikov, T. Maheshwari, P. Trivedi, J. Lehmann and A. Fischer, "Introduction to neural network based approaches for question answering over knowledge graphs," Computing Research Repository, 2019.

[75] J. Lehmann, G. Sejdiu, L. Bühmann, P. Westphal, C. Stadler, I. Ermilov, S. Bin, N. Chakraborty, M. Saleem, A. C. Ngonga Ngomo and H. Jabeen, "Distributed Semantic Analytics using the SANSA Stack," in *In The Semantic Web – International Semantic Web Conference 2017*, 2017, pp. 147-155.

[76] S. Peroni, "The Semantic Publishing and Referencing Ontologies”.," in *In Semantic Web Technologies and Legal Scholarly Publishing*, Springer, pp. 121-193.

[77] S. Capadisli, A. Guy, R. Verborgh, C. Lange, S. Auer and T. Berners-Lee, "Decentralised Authoring, Annotations and Notifications for a Read-Write Web with dokieli," in *In Web Engineering – 17th International Conference on Web Engineering*, Springer, 2017, pp. 469-481.

[78] S. Peroni, F. Osborne, A. Di Iorio, A. G. Nuzzolese, F. Poggi, F. Vitali and E. Motta, "Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles," *PeerJ Computer Science,* 2017.

[79] T. Clark, P. N. Ciccarese and C. A. Goble, "Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications," *Journal of Biomedical Semantics,* vol. 5, no. 1, p. 28, 2014.

[80] T. Groza, S. Handschuh, K. Möller and S. Decker, "SALT-Semantically Annotated LaTeX for Scientific Publications," in *In The Semantic Web: Research and Applications - European Semantic Web Conference*, Springer, 2007, pp. 518-532.

[81] S. Fathalla, S. Vahdati, S. Auer and C. Lange, "Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles," in *In Research and Advanced Technology for Digital Libraries - International Conference on Theory and Practice of Digital Libraries*, 2017, pp. 315-327.

[82] C. Lange, "Ontologies and languages for representing mathematical knowledge on the Semantic Web," *Semantic Web Journal,* no. 4, pp. 119-158, 2013.

[83] B. Smith, M. Ashburner, C. Rosse , J. Bard, W. Bug, A. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall and N. Leontis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnology,* no. 25, pp. 1251-1255, 2007.

[84] Technische Informationsbibliothek (TIB), "Open Research Knowledge Graph," [Online]. Available: http://orkg.org. [Accessed 17 09 2019].

[85] B. C. M. Fung, K. Wang , A. W.-C. Fu and P. S. Yu, "Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques," Chapman & Hall/CRC, 2010, pp. 15-42.

[86] B. C. Chen, D. Kifer, K. LeFevre and A. Machanavajjhala, "Privacy-Preserving Data Publishing," *Foundations and Trends in Databases,* no. 2, pp. 1-167, 2009.

[87] J. Cao, B. Carminati, E. Ferrari and K. Tan, "CASTLE: Continuously Anonymizing Data Streams," *IEEE Transactions on Dependable and Secure Computing (TDSC),* no. 8, pp. 337-352, 2011.

[88] L. Zou, L. Chen and M. Tamer, "K-Automorphism: A General Framework For Privacy Preserving Network Publication," *PVLDB,* no. 2, pp. 946-957, 2009.

[89] J. Cheng, A. Fu and J. Liu, "K-isomorphism: Privacy preserving network publication against structural attacks," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2010, pp. 459-470.

[90] R. S. Sandhu, E. J. Coyne, H. L. Feinstein and C. E. Youman, "Role-Based Access Control Models," *Computer 29,* pp. 38-47, 1996.

[91] D. F. Ferraiolo, R. Sandhu, S. Gavrila, D. R. Kuhn and R. Chandramouli, "Proposed NIST standard for role-based access control," *ACM Trans. Inf. Syst. Secur.,* no. 4, pp. 224-274, 2001.

[92] S. Verma, B. S. Kumar and M. Singh, "Comparative analysis of Role Base and Attribute Base Access Control Model in Semantic Web," *International Journal of Computer Applications,* no. 46, pp. 1-6, 2012.

[93] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, "CRISP-DM 1.0 Step-by-step data mining guide," in *The CRISP-DM consortium*^, 2000.

[94] J. Feldhusen and K.-H. Grote, Pahl/Beitz Konstruktionslehre, Berlin-Heidelberg: Springer-Verlag, 2013.

[95] J. Feldhusen and B. Gebhardt, Product Lifecycle Management for practice, London, UK: Springer-Verlag, 2008.

[96] "Entscheidungshilfe zur Einführung von PDM Systemen," VDMA-Verlag, Frankfurt am Main, 2005.

[97] W. Scheidel, I. Mozgova and R. Lachmayer, "Structuring Information in Technical Inheritance by PDM Systems.," *Proceedings of the 21st International Conference on Engineering Design (ICED17),* vol. 6, p. 217–226, 21-25 08 2017.

[98] E. Glaessgen and D. Stargel, "The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles," *Proceedings of the 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural and Material Conference,* p. 1818, 2012.

[99] F. Tao, J. Cheng, Q. Qi, M. Zhang, H. Zhang and F. Sui, "Digital twin-driven product design, manufacturing and service with big data," *The International Journal of Advanced Manufacturing Technology,* vol. 94, no. 9-12, pp. 3563-3576, 2018.

[100] E. Negri, L. Fumagalli and M. Macchi M, "A Review of the Roles of Digital Twin in CPS-Based Production Systems," in *Value Based and Intelligent Asset Management*, A. Crespo Márquez, M. Macchi and A. Parlikad, Eds., Springer International Publishing, 2020.

[101] J. Qin, K. Crowston and A. Kirkland, "Pursuing Best Performance in Research Data Management by Using the Capability Maturity Model and Rubrics," *Journal of eScience Librarianship,* vol. 6, no. 2, p. e1113, 06 10 2017.

[102] J. Oppenländer, F. Glöckler and J. Hoffmann, "Reifegradmodelle für ein integriertes Forschungsdatenmanagement in multidisziplinären Forschungsorganisationen," Heidelberg, 2017.

[103] [Online]. Available: https://rdmorganiser.github.io. [Accessed 14 10 2019].

[104] [Online]. Available: https://tudmo.ulb.tu-darmstadt.de. [Accessed 14 10 2019].

[105] [Online]. Available: https://rda-dmp-common.github.io/RDA-DMP-Common-Standard. [Accessed 14 10 2019].

[106] T. Miksa, S. Simms, D. Mietchen and S. Jones, "Ten principles for machine-actionable data management plans," *PLoS Comput Biol,* vol. 15, no. 3, p. e1006750, 03 28 2019.

[107] [Online]. Available: https://blog.rwth-aachen.de/forschungsdaten/2018/04/30/erfolgreiche-teilnahme-im-ideenwettbewerb-zur-wissenschaft-im-digitalen-wandel/.

[108] [Online]. Available: https://github.com/rdmorganiser/rdmo-catalog/tree/master/shared/nfdi4ing/rdmo_mechanical_engineering. [Accessed 14 10 2019].

[109] [Online]. Available: https://vocol.iais.fraunhofer.de/. [Accessed 20 09 2019].

[110] "International Data Spaces Association," [Online]. Available: https://www.internationaldataspaces.org/. [Accessed 13 10 2019].

[111] [Online]. Available: https://zenodo.org/communities/rdm_training_engineering_sciences/. [Accessed 20 09 2019].

[112] [Online]. Available: https://eosc-fair4s.github.io. [Accessed 20 09 2019].

[113] [Online]. Available: https://carpentries.org. [Accessed 20 09 2019].

[114] [Online]. Available: https://tibhannover.github.io/FAIR-studyGroup/. [Accessed 20 09 2019].

[115] C. Elsner, N. Rosenke, M. Weber, C. Hoppe, S. Drößler and S. Hermann, "Von Bottom up zu Top down," *O-Bib. Das Offene Bibliotheksjournal,* vol. 6, no. 2, pp. 80-91, 09 07 2019.

[116] [Online]. Available: http://jats.nlm.nih.gov/archiving/tag-library/1.1d1/index.html. [Accessed 11 09 2019].

[117] [Online]. Available: https://www.springernature.com/gp/researchers/text-and-data-mining. [Accessed 13 09 2019].

[118] [Online]. Available: https://doi.org/10.1007/978-0-387-09823-4. [Accessed 13 09 2019].

[119] [Online]. Available: http://dx.doi.org/10.25534/tudatalib-110. [Accessed 09 10 2019].

[120] A. D. Sendek, Q. Yang, E. D. Cubuk, K. A. N. Duerloo, Y. Cui and E. J. Reed, "Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials," *Energy and Environmental Science,* vol. 10, no. 1, pp. 306-320, 2017.

[121] Deutsche Forschungsgemeinschaft, "Handling of Research Data: Discipline-specific conventions," [Online]. Available: https://www.dfg.de/en/research_funding/proposal_review_decision/applicants/research_data/. [Accessed 15 10 2019].

[122] [Online]. Available: https://nfdi4ing.de/veranstaltungen/. [Accessed 13 09 2019].

[123] H. Takeuchi and I. Nonaka, in *The knowledge-creating company: how Japanese companies create the dynamics of innovation*, New York, Oxford University Press, 1995, p. 3.

[124] [Online]. Available: https://www.sfb1313.uni-stuttgart.de/. [Accessed 25 09 2019].

[125] [Online]. Available: https://isse.tu-clausthal.de/de/forschung/laufende-projekte/recycling-40/. [Accessed 25 09 2019].

[126] [Online]. Available: https://www.tu-braunschweig.de/se2a. [Accessed 25 09 2019].

[127] [Online]. Available: https://www.simtech.uni-stuttgart.de/. [Accessed 25 09 2019].

[128] B. Flemisch, M. Darcis, K. Erbertseder, B. Faigle, A. Lauser, K. Mosthaf, S. Müthing, P. Nuske, A. Tatomir, M. Wolff and R. Helmig, "DuMux: DUNE for Multi-{Phase, Component, Scale, Physics, . . .} Flow and Transport in Porous Media," *Advances in Water Recources,* vol. 34, no. 09, pp. 1102-1112, 2011.

[129] [Online]. Available: https://www.lrz.de/forschung/projekte/forschung-daten/GeRDI/. [Accessed 11 10 2019].

[130] [Online]. Available: https://mediatum.ub.tum.de/. [Accessed 11 10 2019].

[131] [Online]. Available: www.sonarqube.org. [Accessed 20 09 2019].

[132] N. Preuß and P. F. Pelz, "Integrated management of experimental research- and meta-data for fan test rigs," in *International Conference on Fan Noise, Aerodynamics, Applications and Systems 18.-20. April 2018*, Darmstadt, Germany, 2018.

[133] N. Preuß, G. Staudter, M. Weber, R. Anderl and P. F. Pelz, "Methods and Technologies for Research- and Metadata Management in Collaborative Experimental Research," *Applied Mechanics and Materials,* vol. 885, pp. 170-183, 20 11 2018.

[134] "Institut für Fluidsystemtechnik - Fluidsystemtechnik - Technische Universität Darmstadt," [Online]. Available: https://www.fst.tu-darmstadt.de/fachgebiet/index.de.jsp. [Accessed 29 09 2019].

[135] "Reaktive Strömungen und Messtechnik - Technische Universität Darmstadt," [Online]. Available: https://www.rsm.tu-darmstadt.de/home_rsm/index.de.jsp . [Accessed 29 09 2019].

[136] "Simulation of reactive Thermo-Fluid Systems - Technische Universität Darmstadt," [Online]. Available: https://www.maschinenbau.tu-darmstadt.de/stfs/stfs/welcome.en.jsp. [Accessed 29 09 2019].

[137] "CRC 1194 - Interaction between Transport and Wetting Processes - Technische Universität Darmstadt," [Online]. Available: https://www.sfb1194.tu-darmstadt.de/sfb_1194/index.en.jsp. [Accessed 29 09 2019].

[138] " SFB/TRR150: Turbulente, chemisch reagierende Mehrphasenströmungen in Wandnähe - Technische Universität Darmstadt," [Online]. Available: https://www.trr150.tu-darmstadt.de/der_sonderforschungsbereich/CRC.en.jsp. [Accessed 29 09 2019].

[139] "SFB/TRR 129 Oxyflame," [Online]. Available: http://www.oxyflame.de/ . [Accessed 29 09 2019].

[140] "SFB 805 - Technische Universität Darmstadt," [Online]. Available: https://www.sfb805.tu-darmstadt.de/sfb805/Index.de.jsp . [Accessed 29 09 2019].

[141] "Lab for Measurement Technology," [Online]. Available: http://www.lmt.uni-saarland.de/index.php/en/ . [Accessed 29 09 2019].

[142] "Forschungszentrum Energiespeichertechnologien," [Online]. Available: http://www.est.tu-clausthal.de/. [Accessed 29 09 2019].

[143] "Fraunhofer Heinrich-Hertz-Institut," [Online]. Available: https://www.hhi.fraunhofer.de/. [Accessed 29 09 2019].

[144] "Gesellschaft für Energiewissenschaft und Energiepolitik e. V.," [Online]. Available: https://www.gee.de. [Accessed 17 09 2019].

[145] "Forschungsnetzwerk Systemanalyse," Projektträger Jülich | Forschungszentrum Jülich GmbH, [Online]. Available: http://www.forschungsnetzwerke-energie.de/systemanalyse. [Accessed 17 09 2019].

[146] "iea hydrogen," [Online]. Available: http://ieahydrogen.org/. [Accessed 17 09 2019].

[147] "FZJ-IEK3-VSA," Forschungszentrum Jülich GmbH, [Online]. Available: https://github.com/FZJ-IEK3-VSA. [Accessed 17 09 2019].

[148] "METIS PLATFORM," Forschungszentrum Jülich GmbH, [Online]. Available: metis-platform.net. [Accessed 17 09 2019].

[149] "enargus," Forschungszentrum Jülich GmbH, [Online]. Available: https://www.enargus.de/pub/bscw.cgi/?op=enargus.eps2&q=lod-geoss&v=10&id=1213112. [Accessed 17 09 2019].

[150] "TIB AV-Portal," Technische Informationsbibliothek (TIB), [Online]. Available: https://av.tib.eu. [Accessed 17 09 2019].

[151] "TIB Leibniz Data Manager," Technische Informationsbibliothek (TIB), [Online]. Available: https://datamanager.tib.eu. [Accessed 17 09 2019].

[152] "Slide Wiki," Technische Informationsbibliothek (TIB), [Online]. Available: https://slidewiki.org/. [Accessed 17 09 2019].

[153] "Data Science Storage," [Online]. Available: https://doku.lrz.de/display/PUBLIC/Data+Science+Storage . [Accessed 13 10 2019].

[154] "RWTH Publications," [Online]. Available: http://publications.rwth-aachen.de/. [Accessed 13 10 2019].

[155] "DaRUS," [Online]. Available: https://darus.uni-stuttgart.de/. [Accessed 13 10 2019].