Submission of Letters of Intent

NFDI4Ing - Nationale Forschungsdateninfrastruktur für die Ingenieurwissenschaften

Deutsche Forschungsgemeinschaft
Kennedyallee 40 · 53175 Bonn, Germany · Postal address: 53170 Bonn, Germany
Tel.: + 49 228 885-1 · Fax: + 49 228 885-2777 · postmaster@dfg.de · www.dfg.de

1. Binding letter of intent as advance notification or non-binding letter of intent

X	Binding letter of intent (required as advance notification for proposals in 2019)
☐	Non-binding letter of intent (anticipated submission in 2020)
☐	Non-binding letter of intent (anticipated submission in 2021)

2. Formal details

Planned name of the consortium:

Nationale Forschungsdateninfrastruktur für die Ingenieurwissenschaften

Acronym of the planned consortium:

NFDI4Ing

Applicant institution:

RWTH Aachen University

Templergraben 55

52062 Aachen

Head of the Institution: Prof. Dr. Ulrich Rüdiger, Rector

Spokesperson:

Prof. Dr. Robert Schmitt, RWTH Aachen University

contact@nfdi4ing.de

3. Objectives, work programme and research environment

Research area of the proposed consortium (according to the DFG classification system)

- 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

Concise summary of the planned consortium's main objectives and task areas

To contribute to the overarching NFDI goals, NFDI4Ing serves as a platform for scientific networks, open to all researchers in engineering science. For identifying key objectives and specific needs regarding research data management in engineering science, NFDI4Ing considers the two main aspects of the self-conception of engineering science: What is its core – i) Analysis

and ii) Synthesis of technical-economic-social systems. And, what gives it orientation – iii) consideration of social needs and iv) sustainability of technical system functionalities:

i.) Analysis – reuse data to trace or reproduce results

- Treat research software (from analysis scripts and implementations of physical models to algorithms and simulation code) as part of research data that connects the different stages of data.
- Enable researchers to easily verify trustworthiness of results by tracing them through the whole research process or by conducting intentionally redundant studies.
- Facilitate identifying starting points for further research, avoiding unwanted redundancies.

ii.) Synthesis – repurpose data as basis for (or validation of) otherwise unrelated research

- Automate recording and linking of auxiliary information and provenance – in the context of a specific project, but as much as possible beyond that – for a hitherto unknown re-use.
- Facilitate sharing and integrating of data across single studies, projects, or institutions via technical infrastructure, open metadata standards, etc.
- Establish validation of computational models by integration of simulation and experimental or observational data as a standard procedure.

iii.) Consider social needs – society in general & sub-society of the research environment

- Reduce effort for data-related tasks when conducting field studies. This allows verifying system behaviour in the actual (social) environment to ensure satisfying current social needs.
- Allow unhindered collaborative research, while preventing unauthorised access to confidential data. Because the engineering research environment is in close proximity to the industry, this calls for sophisticated authentication means and role management.
- Facilitate a change of culture towards sustainable handling of data. The engineering research environment needs an improvement of data- and software-related education (data literacy) as well as a reputation system for data publications.

iv.) Design technical functionalities to be sustainable

- Optimize the research process by automating data handling as much as possible, enabled by establishing open standards for machine-processable auxiliary information.
- Contribute to social acceptance with availability, traceability, etc. of research data as a basis for discussion of emerging technologies (renewable energy, self-driving cars, etc.).
- Provide low-effort solutions to generate machine-actionable representations of auxiliary information, supporting both availability and demand of data-driven analysis methods.

In the vast variety of engineering problems, the focus on research (sub-)communities in past efforts has resulted in highly specialized solution approaches. Common aspects on a method or workflow level must be found for the definition of task areas that benefit the engineering community at large. Therefore, NFDI4Ing will put emphasis on the identification and

harmonization of engineering research archetypes – typical research methods and workflows classifying corresponding challenges for research data management.

NFDI4Ing task areas are derived from these archetypes. The specific needs of a single research project or facility can then be represented by a combination of task areas, or intersection of task area and research community. NFDI4Ing so far identified seven archetypes (= task areas), and the corresponding challenges to be solved:

1. High variability of hardware and research software setups: Solving the combinatory explosion when composing systems of hardware and software components; unifying and automating documentation (metadata creation) for such systems; harnessing insufficiently documented legacy data.

2. High throughput of data (experiment and simulation): Integrating large decentralized data volumes and transferring large datasets to the location of computation, or code to storage location; processing data online at high rates (no storage), while triggering storage of valuable data.

3. Mature, sophisticated research software, incl. high-performance computing: Documenting, managing (versioning), and distributing source code; automating jobs and testing (unit tests and result validation); distributing executable software rather than data and ensuring reproducibility.

4. Assessable data: Covering highly heterogeneous and alternating information needs; facilitating interaction with known data sources and identification of new data sources; providing means to evaluate and certify data (sources) regarding quality, and to maintain or increase that quality.

5. Many simultaneously involved participants and end-user-devices: Providing mobile access; synchronization and backup of documents; integrating rights and role management across services; considering usability; concurrent access and security.

6. Field data and distributed systems: Ensuring connectivity (hardware interfaces and protocols, data transfer infrastructure); considering confidentiality of data (user data, customer-, partner-related data), providing data access which is agnostic of physical location via references and Persistent Identifiers (PIDs).

7. Sample-centric workflows: Unifying output of a heterogeneous pool of standardized instrumentation; nonlinear provenance tracking (i.e. history of manipulation and properties) of samples; ensuring adaptability of workflows and compatibility to established electronic lab notebooks (ELNs) or laboratory information management systems (LIMS).

Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfil the planned consortium's objectives

The current members of NFDI4Ing have powerful existing research data management services that address all phases of the data lifecycle. These tools and services will serve as a basis for a community-based integration and evaluation of the task areas, which are derived from the archetypes described above. These services also mark the beginning of a transition process from local to shared services as intended by the NFDI.

As an example, the NFDI4Ing partners aim for a distributed but connected landscape of their repository infrastructures, which will facilitate data findability and reuse. This has already been established in a prototype of a CKAN integration at the TU Darmstadt Computing Centre that connects to repository instances of TIB and LUH, Aachen and Darmstadt. During the project phase of NFDI4Ing, more additions are planned. These include data infrastructures at Stuttgart (DaRUS), Munich (MediaTUM), KIT (including associated EUDAT services) and FZ Jülich, the latter two as members of the Helmholtz Data Federation. As a second example, we consider RDMO a promising future standard tool for handling data management plans in the NFDI. Four of the nine current RDMO instances are operated at NFDI4Ing partners.

The International Data Spaces e.V. (IDSA) aims to develop reliable solutions for digitization 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. At the same time, the IDSA is intended to promote the industrial data sovereignty maintained on the basis of defined standards – a requirement that must be taken into account when adapting the FAIR principles by NFDI4Ing. This will create synergy effects between the academic community and industrial partners.

A common data and metadata space will be established by propagating technologies such as persistent identifiers (e.g. ePIC handles and DOIs) and data lakes for (RDF) metadata to engineering communities. These technologies will be used to ingest, access, and retrieve fragmented data artefacts while explicitly allowing for decentralized storage. Data management workflows need to be integratable into the existing, individual environments of the researchers using "scriptable interfaces" and APIs. A common identity space will be established using various existing base technologies such as DFN-AAI or ORCID.

Source code is a special kind of research data of high importance. NFDI4Ing recognizes this e.g. by operating a GitLab instance that is available to the German research community. Apart from the centralized service per se, this activity yields a set of best practices for operation of local GitLab instances.

Interfaces to other proposed NFDI consortia: brief description of existing agreements for collaboration and/or plans for future collaboration

NFDI4Ing is an open and constantly growing representation for requirements regarding research data infrastructure in the engineering sciences. Therefore, cooperation with other consortia in nearby disciplinary fields is well prepared and foreseen. The former discrete consortium OD-REx accepted the invitation to integrate the domain of intelligent robotics research into NFDI4Ing. For reasons of organizational size and internal community organization, 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. For example, the software architecture envisioned by NFDI4Ing and NFDI4MSE combines efforts of international data spaces and the FAIR Data Principles. It therefore provides researchers with best practices for collaboration and data management and thus contributes to the overall vision of NFDI. 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 joint between the two consortia to ensure a reliable flow of information. The focus of NFDI4MobilTech will be on fine-grained, decidedly domain-specific challenges of ground-based transport and its travel behaviour aspects, while NFDI4Ing will enable the integration of engineering-specific requirements in the mobility domain, including air-borne transport, with other engineering disciplines based on their common denominators. For the planned community workshops in NFDI4Ing, the neighbouring disciplinary consortia are invited to contribute, to work on joint solutions, and to disseminate research results. Already in preparation of the NFDI call, members of NFDI4Ing participated in FAIRmat workshops.

Depending on the subject, collaboration is also planned with consortia from different research areas. NFDI4Chem collaboration already started focusing on the development of digitization 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, energy, etc. 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. 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. We will deal with these issues in a cooperation with NFDI4Medicine and NFDI4Neuro. Besides the common research and disciplinary interest, cooperation is also focused with regards to the 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 cooperate in the area of standardisation and curation of 3D data types (like CAAD models and other forms of 3D digital representations).

4. Cross-cutting topics

NFDI4Ing understands cross-cutting topics as an essential vehicle for the intra- and inter-consortial collaboration with the aim to share lessons learnt and benefit from each other's expertise.

Based on data curation and consideration of the FAIR principles, NFDI4Ing identifies the cross-cutting topics below. These topics play a particularly important role in the task areas, e.g. by supporting the development of applicable solutions within and beyond NFDI4Ing.

FAIR data and data quality assurance and metrics: Standardization and quality metrics build the common ground for the transition towards a data sharing culture in the engineering sciences. Within NFDI4Ing, we will develop standards for data sharing and preservation based on existing best practices and recommendations. In order to assure data quality, a commonly agreed set of criteria needs to be defined that is reflected not only by the engineering community but by the scientific community as a whole. This is especially important in the task area "assessable data".

Research software development: Increased digital literacy in academia results in a wide range of software tools, from automatization of individual workflows to the development of entire software libraries, across all scientific disciplines. Researchers in engineering form a nucleus of early adopters within this area. Within NFDI4Ing, we therefore aim to provide professional research software development services (e.g. GitLab) and best practices to all scientific communities within NFDI. A service based on standard technologies will be enhanced and extended to promote FAIR research data management. While research software development is common across all task areas, it has an especially high priority within the task areas „mature, sophisticated research software" and "high variability of hardware and research software setups".

Vocabularies, ontologies, terminologies: An important element of NFDI4Ing is the semantic classification and description of data using vocabularies and ontologies. For all task areas, NFDI4Ing will develop a basic semantic classification that is as general as possible, has a

manageable hierarchical structure, and consists of a finite number of logical relationships that relate to the archetypes as mentioned above. Both, the generic and discipline-specific aspects of this classification have to evolve community-driven, iterative and evolutionary. This requires appropriate technical and organizational support.

Metadata services (schema registries, standard registries, collection registries, PID services such as the Data-Level Metrics): A high priority within all identified task areas is appropriate metadata, representing the basis for almost every action within FAIR research data management. Approaches to metadata handling must allow the flexibility required by the heterogeneous and dynamic scientific methods in engineering, while still providing the level of standardization and interoperability required for beneficial utilization of metadata. Citations are the heart of the academic realm. Therefore, solutions should be re-used or developed for data-level metrics.

Storage & Archive (long-term archiving): Currently the community uses a variety of storage and archiving possibilities for research data. Most of them are provided either by the local infrastructure providers (e.g. IT centres at universities) or by national and international service providers (e.g. academic and commercial repositories). A wide range of aspects like bitstream preservation, sustainable data maintenance, and operating cost options depend on the academic requirements towards the quality or volume of data. For long-term archiving, for example, decisions about which data should be archived can only be made in close cooperation of community scientists and infrastructure providers. In NFDI4Ing, this topic has special priority within the task areas “high throughput of data” and “assessable data”.

Repositories: For storing, finding, and sharing of research data, repositories are the most flexible and sustainable solution. According to the FAIR principles, repositories (necessarily in combination with metadata) are important for all task areas of NFDI4Ing, and certainly other consortia as well. Solutions for generic repositories need adaptions in order to meet the requirements of the engineering sciences. An example has already been established in a prototype of a CKAN integration at the TU Darmstadt Computing Centre that connects to repository instances of TIB and LUH, Aachen and Darmstadt.

Overall NFDI software architecture – data security and sovereignty, interoperability of repositories: Engineering research is often characterized by a close proximity to the industry. As a consequence, sophisticated means in authentication and role management are necessary. Within NFDI4Ing, we will aim at improving and extending available AAI concepts by academic and industrial service providers for authentication, authorization, and role management. As many other consortia (e.g. industrial chemistry in NFDI4Chem) will have to consider these aspects as well, a close cooperation will enable harmonized approaches in this area. This topic is important to all task areas, but it has special priority within the task areas “many simultaneously involved participants and end-user-devices” and “field data and distributed systems”.

Community-based training and cultural change on enabling data-driven science and FAIR data is required to enable researchers of all scientific disciplines to work in data-driven environments. While there are many common practices that can and should be adopted across disciplines, NFDI4Ing will specifically address the engineering community with trainings tailored towards specific challenges like handling large amounts of data or transitions between empirical and simulation-based research environments. Only if trainings fit the research environments, they can be successfully integrated to enable the cultural changes needed to spread structured data management within the community. This topic is important to all task areas.

“From Data to Knowledge”: As a result, NFDI will create an infrastructure that allows to access data across scientific fields. This will create new ways of generating knowledge from existing data sets. Research on innovative solutions for data analytics depends on well-prepared and high quality (meta)data and is a prerequisite for algorithmic approaches from fields like high performance data analytics, machine learning, artificial intelligence, and text mining. In NFDI4Ing, this topic will be particularly important for “sample-centric workflows” and “mature, sophisticated research software”.

NFDI4Ing will focus on these cross-cutting topics motivated by the community of the engineering sciences and act as a reliable partner in the NFDI. The tasks arising from the cross-cutting topics will be solved in close collaboration with other consortia. Examples are the planned cooperation with text+ and NFDI4Culture on the issue of “community-based training and cultural change on enabling data-driven science and FAIR data” and MaRDI on the issue of “research software development” (including harmonized efforts on software development as well as training). For other cross-cutting topics (e.g. legislation matters) further collaboration is needed.