NFDI 4 CS & CS 4 NFDI & Partners

Non-binding Letter of Intent (2020)

Please address the following aspects in your letter of intent

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

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

2 Formal details

[this will be filled in / completed till the August, 16th 2019]

- Planned name of the consortium
- *NFDI for Computer Science & Computer Science for NFDI & Partners*
- Acronym of the planned consortium
- *NFDI4CS4NDFI**

- Applicant institution
- *University of Duisburg-Essen*
- *Prof. Dr. Ulrich Radtke, Rektor*
- Spokesperson
- *Prof. Dr. Michael Goedicke, paluno – The Ruhr Institute for Software Technology, michael.goedicke@paluno.uni-due.de*

3 Objectives, work programme and research environment

- Research area of the proposed consortium
- *Area 409 / 44 (Computer Science)*
- *Concise summary of the planned consortium's main objectives and task areas*

Please note that this LOI proposes two related areas of work: CS – related research data management and work on cross cutting topics. Currently we leave it open, whether this will be covered in two consortia or both areas are proposed to be covered in one consortium

The main goal of the consortium is to identify, define and finally deploy services to store complex domain specific data objects from the specific variety of domains from Computer Science (CS) and its applications and to realize the FAIR principles across the board. This consortium will focus on various domains where empirical data will emerge naturally. Examples are "software", "eLearning", "HCI/media informatics" which the consortium likes to start with. However, during the course of action it will be extended to all relevant areas of CS. The related findings and results in terms of methods, processes & way of communication and service will be offered to other domains beyond CS as well. Thus, this area of the work definitely reaches out to other – primarily – national but also international consortia and partners as well. A general tool becoming more and more accessible in the area of Computer Science which needs special attention is HPC (High Performance Computing). Here a specific way to create, maintain and analyze huge amount of data is provided and, of course, not only for the area of Computer Science. In particular HPC simulations produce monitoring and telemetry information collected while running a specific job on an HPC system. This type of information – generalized HPC performance data –is instrumental to provide general provenance information and to facilitate the reproducibility of results.

A main overarching task within this consortium will be the definition of CS related data types. This means related processes and methods will be defined, implemented and deployed for realizing interoperable research data management schemes for all relevant CS areas.

As a short summary there will be the following areas of concerns covered in related subprojects:

- *CS related data types*: this means related processes and methods will be defined, implemented and deployed for realizing interoperable research data management schemes for all relevant areas in Computer Science
- *Cross Cutting Services*: a set of cross cutting services will emerge from the work in the various areas defined here and the consortium will potentially collaborate with. From the start there are a number of candidates which cover important areas addressed in the FAIR principles. E.g in order to enable (R) reusability of data it is necessary to freeze the context in which this data has been collected / created, processed, viewed etc. This context consists of a range of

components e.g. a complete execution environment for the aforementioned aspects. In summary the key to realize the FAIR -principles are allocated here.

- *Infrastructure*: here the area is addressed to cover a global interoperable management scheme which enables the robust and persistent ID-management for individuals, publications and data, safe and reliable long-term storage and retrieval mechanisms across organizational borders of single institutional research data management facilities. Related blueprint, prefabricated solutions and software libraries can also be created and provided to all institutions wanting to collaborate later on.
- *HPC*: due to its specific profile the HPC – area is important to being addressed in a special way as well. The powerful tool these machines or even network of machines provides needs to be handled in a way which addresses the FAIR-principles for this area of computing as well.
- *Data Management:* In addition to the necessary computing infrastructure, general data management topics are of equal importance. The database community has developed a broad spectrum of data management solutions – both general purpose as well as specific highly tailored database management systems – most suitable for a particular research project is a non-trivial task, which is highly relevant for research data management, especially for large amounts of data and concerning trade-offs between performance, consistency and precision.
- *Data Quality:* Data Management Systems already provide a spectrum of techniques that are aimed at maintaining and improving data quality (e.g. database transactions, integrity constraints etc.). Recent research results will support the actual needs of data quality management, which are aimed at highly sophisticated domain specific quality criteria.
- *Knowledge representation and reasoning:* this area of CS, is increasingly important for all kinds of research projects. Research results about gaining knowledge in a specific domain based on a general way of representing knowledge, the evolution of knowledge and methods for reasoning on a given corpus are of paramount importance for the integration of knowledge from different origins.
- *Machine Learning:* these techniques are increasingly applied in all areas of research. This offers chances but also incorporates risks. The consortium will develop a set of guidelines to address the challenges of using machine learning techniques especially to avoid misinterpretation and bias.

Thus, it will be possible for these types of data objects and beyond to provide the necessary standards based on the FAIR principles to create persisted, sustainable, reproducible and distributable versions of related data objects. These data objects can be of any size, structure and quality.

Key is here to create an organizational and technical, cooperative / interoperable infrastructure to join forces of actors from computer science and beyond to create the competencies for the

responsible treatment of research data. Based on the existing structure of the GI – comprising of related specialist groups –, NFDI4RSE and the other co-applicants mentioned above, it will be reached out to related communities to maintain and where necessary create new standards for research data objects and their services. The specialist groups provide an excellent basis for providing help in specific problems / domains when needed.

These domain specific communities select respective steering committees for defining and pursuing the agendas to build the specific research data infrastructure.

- *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 infrastructure will be based on the services DFN already provides for secure authentication and authorization (aka eduroam, eduGAIN). Both services federate identity management systems in R&E organizations at an international level. A substantial part of these services is needed in many NFDI contexts – the required "NFDI-readiness" can only be realized in close consultation with the researchers. In addition, basic protocols and services to find and access the data being looked for can be realized and operated on the existing infrastructures of DFN. As a result, the extended services contribute significantly [but not limited] to the implementation of the F (Findable) and A (Accessible) principles.

Also, the experience is available to adopt the basic abstract version of these services which will be defined in the domains involved here and other consortia. The principles of creating and maintain webservices (from the established field of service-oriented computing) can be used and transformed into the substrate the consortium needs to build and operate the specific services to find and access the desired data based on meta-data related specification of a query.

Metadata catalogues to support the F (Findable) and I (Interoperable) principles can be, e.g., based on the GeRDI project. Leveraging the expertise of partners being involved in GeRDI, we will build as part of the NFDI activities policies and procedures on metadata standardization as well as on making data actually findable by research data search engines (e.g., Google Dataset Search or GeRDI). Minimum requirements will include the availability of core metadata properties defined by DataCite, the assignment of persistent identifiers to datasets, and an open standard interface (e.g., OAI-PMH) for metadata harvesting by search engines.

On the data curation and storage side, in many cases the size of the raw data indicates that the collected data can is unsuitable to be downloaded by end users but only a catalogue of all metadata should be centrally hosted, while the data itself has to reside within the participating organizations yet be accessible. The storage infrastructure should be provided by partner institutions from the Gauss Center for Supercomputing and the Gauss Alliance as well as other computer centers and research data management facilities available locally at the collaborating partners. It also becomes necessary to offer compute resources to support R (Re-usablity) of bulk

data within each of these centers, which are dedicated to the storage, curation and ultimately analysis of monitoring data.

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

As a follow-up to initial discussions at the NFDI conference in May members of NFDI4CS4NDFI have established contacts to other consortia addressing cross-cutting objectives, namely NFDI4RSE and 2linkNFDI. All partners agreed to continue consultations over the next months to investigate in synergies and common areas of interest before proposal submissions in 2021 or 2022 respectively.

Nation-wide (NFDI)

Based on discussions which took place at the NFDI conference, we plan to build closer collaborations with, e.g., the following consortia until the proposal submission in 2020. Please note this list in no way exhaustive or stable in any way and we take the freedom establish other or new relationships to the t:

- NFDI4ING
- MaRDI
- NFDI Web
- PAHN-PaN
- Astro-NFDI
- NFDI4Chem
- NFDI 4 MobilTech

Intensive discussions will be organized with these partners over the few next months to structure the domains and areas. It is important to define the related boundaries, interfaces and processes to discuss and agree on common ontologies and interoperation procedures as well. The expertise in Requirements Engineering, Software Architecture and Service oriented computing will be instrumental for the success of these discussion. That will lead to establish the related proposals in such a way that the overall goal of the NFDI will be addressed successfully.

The following steps will be to structure and actually write the proposal in a small number of steps following the guidelines of the DFG.

International Partners

Initial contacts have been established with the Alan Turing Institute in the UK and the Software Heritage by INRIA in France.

4 Cross-cutting topics

- *Please identify cross-cutting topics that are relevant for your consortium and that need to be designed and developed by several or all NFDI consortia.*

Below a preliminary list of aspects, functions, service and processes is given. This is only a first and non-exhaustive list of items which are relevant as cross cutting topics in research data management. Please note, that these topics are derived from two perspectives: one is at the general cross cutting level which applies virtually to all domains while the second one is derived from the specific domains CS as a branch of Science working with empirical data. Thus, at an initial stage the areas briefly described above are meant here. But all areas of CS, of course, with some obvious exceptions like theoretical foundations are to be included in the long run.

The related findings and results in terms of methods, processes and way of communication and service will be offered to other domains as well. Thus, this area of the work definitely reaches out to other – primarily – national but also international consortia and partners as well.

A set of *cross cutting services* will emerge from the work in the various areas defined here and the consortium will potentially collaborate with. From the start there are a number of candidates which cover important areas addressed in the FAIR principles. E.g., in order to enable (R) reusability of data it is necessary to freeze the context in which this data has been collected / created, processed, viewed etc. This context consists of a range of components e.g. a complete execution environment for the aforementioned aspects. In summary the key to realize the FAIR -principles are allocated here. In terms of infrastructure a number of topics are addressed to cover a global interoperable management scheme which enables the robust and persistent ID-management, safe and reliable long-term storage and retrieval mechanisms across organizational borders of single institutional research data management facilities. Related blueprint, prefabricated solutions and software libraries can also be created and provided to all institutions wanting to collaborate later on.

Here the list of cross cutting topics:

- Identity management, persistent identifiers, …
- Persistency of data and context, execution context
- Security against unauthorized access / unauthorized change
- Safety against loss of data and context
- Cost efficient storage and retrieval of context
- Anonymization / Pseudonymization of personal information
- Basic mechanisms to define domain ontologies and appropriate metadata structures, search engines based on such formalized knowledge (e.g. knowledge graphs).
- Data quality: Methods to define data quality and data consistency requirements. Methods to measure data quality.

- Methods to find appropriate data management solutions for a given research project.
- Services to support the groups of experts in collaboratively define and evolve domain ontologies and corresponding metadata structures, automate procedures to define, agree and publish related standards, generate [new versions of] related software components to handle the related meta-data
- Services to support the groups of experts in collaboratively define and evolve meta data structures, automate procedures to define, agree and publish related standards, generate [new versions of] related software components to handle the related meta-data
- Services to support groups of experts to collaboratively define interoperability protocols based on the meta-data structures to enable automated cross-site search and data [including context]-exchange

- *Please indicate which of these cross-cutting topics your consortium could contribute to and how.*

All these aforementioned cross-cutting topics as well as the infrastructure related ones are relevant to this consortium and it is necessary to connect to other consortia in order to obtain the desired level of interoperability to find, access and reproduce the data where needed – at best without further transformation.

On top of these more technical and automatic means for interoperability the consortium will develop practically usable guidelines and standard procedures for research data management, which can be used for research data counselling centers at universities. In order to achieve this, a close cooperation with other domain specific consortia is required. We are already in close contact to a growing number of consortia, which also recognize our proposal as a core element for consolidation. We intend to achieve this consolidating effect by using the German Informatics Society (GI) as a core multiplier, that reaches out to the various areas of computer science and the related application areas as well.

The various topics listed above will be clustered in a way to create related subprojects which elicit requirements, develop and apply appropriate design patters and implement them accordingly as frameworks and/or libraries.

The specific structure of the consortia’s projects and subprojects will be defined during the second half of 2019 and planned in detail during the early 2020. This will result in a number of milestones leading to a streamlined [set of] consortia which address the topics and CS related domains as well.