1 Binding Letter of Intent for NFDIxCS

[x] as advance notification of the full proposal in 2021

2 Formal details

- **Planned name of the consortium:**

National Research Data Infrastructure for and with Computer Science

- **Acronym of the planned consortium:**

NFDIxCS

- **Applicant institution:**

University of Duisburg-Essen
Universitätsstraße 2
45141 Essen
Rektor Prof. Dr. Ulrich Radtke

- **Spokesperson:**

Prof. Dr. Michael Goedicke
michael.goedicke@paluno.uni-due.de
paluno – The Ruhr Institute for Software Technology

- **Co-applicant institution:**

Gesellschaft für Informatik e.V. (GI)
Wissenschaftszentrum
Ahrstraße 45
53175 Bonn
President Prof. Dr. Hannes Federrath

- **Co-spokesperson:**

Daniel Krupka
daniel.krupka@gi.de
Geschäftsführer

- **Co-applicant institution:**

Universität Potsdam
Am Neuen Palais 10

14469 Potsdam
President Prof. Oliver Günther, Ph.D.

- **Co-spokesperson:**

Prof. Dr. Ulrike Lucke
ulrike.lucke@uni-potsdam.de
Komplexe Multimediale Anwendungsarchitekturen

- **Co-applicant institution:**

Karlsruher Institut für Technologie (KIT)
Kaiserstraße 12
76131 Karlsruhe
President Prof. Dr.-Ing. Holger Hanselka

- **Co-spokespersons:**

Prof. Dr.-Ing. Anne Koziolek
anne.koziolek@kit.edu
Modelling for Continuous Software Engineering
Prof. Dr. Ralf Reussner
ralf.reussner@kit.edu
Dependable Software-intensive Systems – Design, Modelling and Analysis

- **Co-applicant institution:**

Universität Hamburg
Mittelweg 177
20148 Hamburg
President Prof. Dr. Dieter Lenzen

- **Co-spokesperson:**

Prof. Dr. Hannes Federrath
federrath@informatik.uni-hamburg.de
Security in Distributed Systems

- **Co-applicant institution:**

Kiel University / Christian-Albrechts-Universität zu Kiel
Christian-Albrechts-Platz 4
24118 Kiel

President Prof. Dr. med Simone Fulda

- **Co-spokesperson:**

Prof. Dr. Agnes Koschmider
ak@informatik.uni-kiel.de
Process Analystics & Information Systems

- **Co-applicant institution:**

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
Am Faßberg 11
37077 Göttingen
Geschäftsführer Prof. Dr. Ramin Yahyapour

- **Co-spokesperson:**

Prof. Dr. Ramin Yahyapour
ramin.yahyapour@gwdg.de
GWDG Geschäftsführer und
Georg-August-Universität Göttingen, Institute of Computer Sciences,
Speaker Campus-Institute Data Science Göttingen

- **Co-applicant institution:**

Ludwig-Maximilians Universität München
Geschwister-Scholl-Platz 1
80539 München
President Prof. Dr. Bernd Huber

- **Co-spokesperson:**

Prof. Dr. Albrecht Schmidt
albrecht.schmidt@um.ifi.lmu.de
Fakultät für Mathematik, Informatik und Statistik, Human-Centered Ubiquitous Media

- **Co-applicant institution:**

Technische Universität München
Arcisstr. 21
80333 München
President Prof. Dr. Thomas F. Hofmann

- **Co-spokespersons:**

Prof. Dr. Tobias Nipkow
nipkow@in.tum.de
Fakultät für Informatik, Logik und Verifikation
Prof. Dr. Martin Schulz
schulzm@in.tum.de
Fakultät für Informatik, Computer Architecture and Parallel Systems

- **Co-applicant institution:**

Johannes Gutenberg University Mainz
Saarstr. 21
55122 Mainz
President Prof. Dr. Georg Krausch

- **Co-spokesperson:**

Prof. Dr.-Ing. André Brinkmann
brinkman@uni-mainz.de
Zentrum für Datenverarbeitung

- **Co-applicant institution:**

Technische Universität Dresden
01062 Dresden
Rector Prof. Dr. Ursula M. Staudinger

- **Co-spokesperson:**

Prof. Dr. Wolfgang Nagel
wolfgang.nagel@tu-dresden.de
Center for Information Services and High Performance Computing (ZIH)

- **Co-applicant institution:**

Forschungszentrum Jülich
Wilhelm-Johnen-Straße
52428 Jülich
Director Prof. Dr.-Ing. Wolfgang Marquardt

- **Co-spokesperson:**

Prof. Dr. Dr. Thomas Lippert

th.lippert@fz-juelich.de
Institute for Advanced Simulation (IAS)

- **Co-applicant institution:**

Universität Paderborn
Warburger Str. 100
33098 Paderborn
President Prof. Dr. Birgitt Riegraf

- **Co-spokesperson:**

Prof. Dr. Christian Plessl
christian.plessl@uni-paderborn.de
Paderborn Center for Parallel Computing

- **Co-applicant institution:**

RWTH Aachen
Templergraben 55
52062 Aachen
Rector Prof. Dr. Ulrich Rüdiger

- **Co-spokesperson:**

Prof. Dr. Matthias Müller
mueller@itc.rwth-aachen.de
IT Center

- **Co-applicant institution:**

TU Darmstadt
Karolinenplatz 5
64289 Darmstadt
President Prof. Dr. Tanja Brühl

- **Co-spokesperson:**

Prof. Dr. Christian Bischof
christian.bischof@tu-darmstadt.de
Hochschulrechenzentrum

- **Co-applicant institution:**

Schloss Dagstuhl - Leibniz Center for Informatics
Oktavie-Allee
66687 Wadern
Prof. Raimund Seidel, Ph.D. (Scientific Director)
Heike Meißner (Technical Administrative Director)

- **Co-spokespersons:**

Prof. Raimund Seidel, Ph.D.
raimund.seidel@dagstuhl.de
Schloss Dagstuhl LZI / Saarland University
Dr. Marcel R. Ackermann
marcel.r.ackermann@dagstuhl.de
dblp computer science bibliography
Dr. Michael Wagner
michael.wagner@dagstuhl.de
Dagstuhl Publishing

- **Co-applicant institution:**

FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
Direktorin Sabine Brünger-Weilandt

- **Co-spokesperson:**

Prof. Dr. Franziska Boehm
franziska.boehm@fiz-karlsruhe.de
Bereichsleiterin Immaterialgüterrechte in verteilten Informationsinfrastrukturen (IGR) und Zentrum für angewandte Rechtswissenschaft (ZAR) am KIT

- **Participants:**

Prof. Dr. Florian Alt	Universität der Bundeswehr München, Forschungsinstitute CODE
Prof. Dr. Christoph Benzmüller	Freie Universität Berlin, Dep. of Math and CS Confederation of Laboratories for Artificial Intelligence Research in Europa (CLAIRE)

Dr. Hendryk Bockelmann	DKRZ, Hamburg
Dr. Daniel Demmler	Universität Hamburg, Security in Distributed Systems
Prof. Dr. Gregor Engels	Universität Paderborn
Dr. Bernhard Fechner	Fernuniversität Hagen, Zentrum für Medien und IT
Prof. Dr. Martin Frank	Steinbuch Centre for Computing Karlsruher Institut für Technologie
Dr. Christian Grimm	Verein zur Förderung eines Deutschen Forschungsnetzes (DFN-Verein)
Dr. Stephan Hachinger	Leibniz-Rechenzentrum (LRZ)
Prof. Dr. Wilhelm Hasselbring	AG Software Engineering, Christian-Albrecht-Universität zu Kiel
Dr. Jens Heidrich	Fraunhofer IESE, Kaiserslautern
Prof. Dr. Timo Kehrer	Modellgetriebene Software Entwicklung, Humboldt Universität zu Berlin
Prof. Dr. Antonio Krüger	Director DFKI und Ubiquitous Media Technology Lab, DFKI und Universität des Saarlandes, Saarbrücken
Prof. Dr. Ulf Leser	Knowledge Management in Bioinformatics, Institut für Informatik, Humboldt Universität zu Berlin
Dr. Jan Linxweiler	TU Braunschweig
apl. Prof. Dr. Thomas Mandl	Institut für Informationswissenschaft & Sprachtechnologie, Universität Hildesheim
Prof. Dr. Vera Meister	TH Brandenburg Fachbereich Wirtschaft
Prof. Dr. Andreas Oberweis	Betriebliche Informationssysteme, Institut für Angewandte Informatik und Formale Beschreibungsverfahren, KIT, Karlsruhe
Prof. Dr. Klaus Pohl	Software Systems Engineering, Paluno, Universität Duisburg-Essen

Prof. Dr. Tillmann Rabl	Data Engineering Systems, Hasso Plattner Institute, Universität Potsdam
Prof. Dr. Michael Resch	HLRS, Universität Stuttgart
Prof. Dr. Enrico Rukzio	Media Informatics, Universität Ulm
Prof. Dr. Bernhard Rumpe	Software Engineering, Fachgruppe Informatik, RWTH Aachen
Prof. Dr. Philipp Schaer	Technische Hochschule Köln, Institut für Informations-wissenschaft
Dr. Gunther Schiefer	Betriebliche Informationssysteme, Institut für Ange-wandte Informatik und Formale Beschreibungsverfah-ren, KIT, Karlsruhe
Prof. Dr. Bernt Schiele	Max-Planck-Institut für Informatik, Saarbrücken
Dr.-Ing. Horst Schirmeier	TU Dortmund, Embedded System Software Group
Prof. Dr. Ute Schmid	Universität Bamberg, Information Systems and Applied Computer Sciences / Confederation of Laboratories for Artificial Intelligence Research in Europa (CLAIRE)
Prof. Dr. Nicole Schweikart	Theoretische Informatik, Institut für Informatik, Humboldt Universität zu Berlin
Prof. Dr. Peter Sobe	HTW Dresden, Faculty of Informatics/Mathematics, Computer Science Fundamentals and Programming
Dr. Alexander Steen	Uni Luxembourg, Faculty of Science, Technology and Medicine
Prof. Dr. Sven Strickroth	LMU München
Dr. Michael Striewe	Uni Duisburg-Essen, paluno
Hannes Thiemann	DKRZ, Hamburg
Prof. Dr. Matthias Weidlich	Humboldt Universität zu Berlin, Institut für Informatik

3 Objectives, work programme and research environment

3.1 Research area of the proposed consortium: *Area 409 / 44 (Computer Science)*

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

NFDIxCS provides a forum for discussion of computer science (CS) research data formats, metadata formats, and semantics and the implementation of the related research data management infrastructure. It works towards generally accepted standards, especially for the sustainable storage, retrieval, and delivery of CS research data. The three main objectives of NFDIxCS are:

a) to promote the implementation of the FAIR Data Principles in the CS community for research data as well as software artifacts,
b) to simplify the citability of software and CS data, and thus
c) to modernize the publication processes and culture in both CS and its applications.

NFDIxCS aims to support CS research in handling increasingly large amounts of data. This applies not only to areas such as Big Data or Artificial Intelligence and Machine Learning, but also to the operation of e.g. High Performance Computing (HPC) systems, computer architecture, and human media interaction. Additionally, numerous methods and processes that provide new insights in other disciplines would not be possible without CS. Working with huge amounts of data from those disciplines will therefore support the further development of genuine CS (research) methods. Thus, NFDIxCS will share the experience and knowledge of the CS community on system architectures, processes, standards for interoperability, data-oriented scientific publishing, and communication systems, first within NFDIxCS and consequently with all interested scientific communities and consortia represented in the NFDI. This bidirectional goal is represented by the "x" in the acronym NFDIxCS.

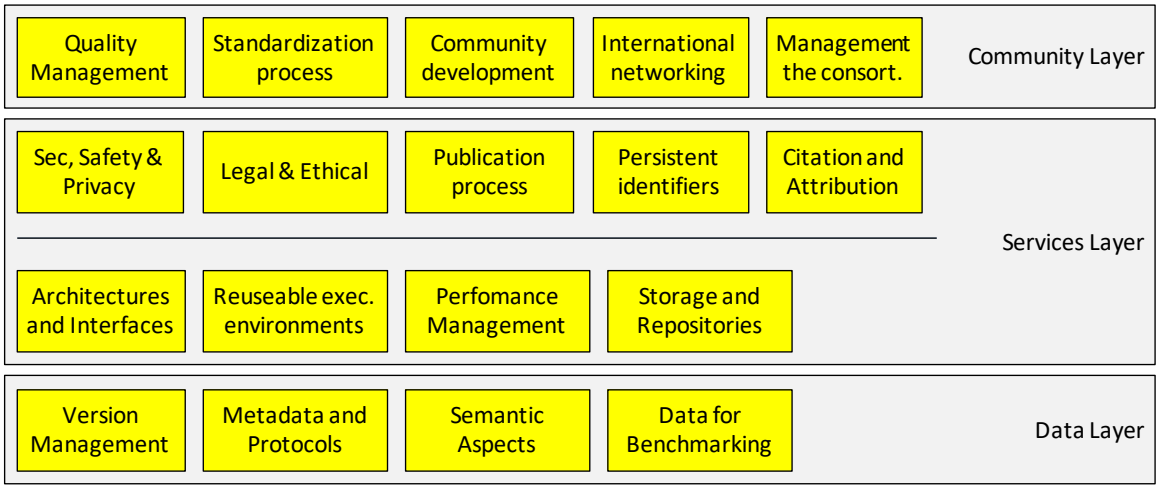
The work program follows the organizational structure of the DFG research area 409/44. For each of its sub-disciplines related types, methods, processes, tools, and concerns in the management of research data have been identified. According to the degree of maturity of these sub-disciplines the definition of formats, metadata and services have been considered. NFDIxCS comprises the following sub-disciplines of CS:

- Theoretical Computer Science,
- Software Technology & Programming,
- Security & Dependability,
- Operating Systems, Communication Systems, Databases, Distributed Systems,
- Visual Computing,
- Business Information Systems,
- Computer Architecture and Embedded Systems,
- Massive Parallel & Data Intensive Systems,
- Artificial Intelligence and Machine Learning,
- Interactive Systems.

From these sub-disciplines – with their respective needs and solutions (i.e. tools, repositories, etc.) – we derive overarching task areas which drive the work in the consortium. The organizational structure of NFDIxCS is designed to accommodate future trends and new topics in CS, which could in the future be integrated as new sub-communities.

From a **technical point of view** the overall goal of NFDIxCS is to produce re-usable data objects specific to the various types of CS data including the data itself, the related metadata as well as the corresponding context and execution information in a standardized way. Thus, we realize the **FAIR principles** to create persisted, sustainable, reproducible, and distributable versions of CS research data objects. Those data objects can be of any size, structure and quality and packaged in a container as a unit.

Based on the analysis of the sub-disciplines the entire **work of NFDIxCS is organized** in 18 task areas (c.f. figure) according to the identification of their respective properties and commonalities.



The basis is the **data layer** covering fundamental task areas like metadata definition, protocols for interoperation, and the versioning of data and software. The semantic aspects of the research data – particularly important for more complex types of data like software or formal proofs – are covered here as well. The **service layer** builds upon this and is divided into a sublayer addressing operational aspects such as the definition of architecture and interoperation related interfaces, performance management, and the management of storage and repositories. A user-centric sublayer will address services like security, safety & privacy, legal & ethical issues/properties, persistent identifiers, as well as citation & attribution – all related to the specific needs of CS. The **community layer** encompasses task areas which not only address the NFDIxCS users but *involve* them. Especially standardization, community building and international networking reach out to the wide CS community within NFDIxCS and beyond – nationally and internationally. The two task areas quality management and the project management will measure and control the success of the work and the use of funds to create and operate the infrastructure and processes for FAIR CS research data management (RDM).

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

An overall principle of NFDIxCS is to reuse, configure and assemble existing solutions and services wherever possible. An important aspect of the community work within NFDIxCS will be to balance the valid interests of researchers (to rapidly evolve infrastructure according to their very dynamic needs) and infrastructure providers (to keep installations stable, safe, affordable and maintainable). Members of the consortium contribute – along with their disciplinary expertise – versatile basic tools related to RDM from multiple perspectives. Since we consider not only data in the traditional sense as the object of our initiative, but also software as a specific artifact of our field, specific tools that support proper handling of software are of great importance to NFDIxCS:

- GitLab as a widely used platform to develop and maintain software,
- tools from rapid continuous software engineering for interlinking various branches,
- use of the Software Heritage Foundation software repository as a persistent mirror,
- evolution methods for especially research software from exploration to production level.

For the deployment of RDM services we will rely on a number of established tools and technologies in order to make use of the potential of a national infrastructure and keep the system maintainable and sustainable. For instance, we will use genuine CS technology, such as:

- database technology,
- semantic representation schemes,
- virtualization infrastructures like Docker, Proxmox, LXC Container,
- exchange protocols like OAI-PMH for metadata,
- middleware components using REST/SOAP,
- search infrastructures like Apache SOLR, and
- distributed authentication and authorization infrastructures like Shibboleth and OAuth.

Several of the co-applicants are professional infrastructure providers like academic data centers with experience in handling large amounts of data, providing high-level IT services and serving demanding scientific users. Security, privacy, and ethical & legal data handling are core principles of their daily business.

For RDM specifically, we rely on existing tools developed by the scientific community as e.g.:

- GI digital library https://dl.gi.de and dblp https://dblp.org as the established CS bibliographies, including tools for data and software evaluation, referencing (see Software Heritage Foundation) and publication,
- the online tools RDMO for supporting RDM plans in research projects,
- the Git plugin Conquaire for automated quality checks in publication processes.

This will be accompanied by an adaption of existing tools and services for community development to the specific needs of our discipline, for instance:

- the Carpentries lectures on RDM,
- the Indico service for scalable conference and publication management.

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

NFDIxCS is eager to collaborate with other NFDI initiatives on equal footing. Partnerships exist both with existing as well as planned consortia and range from closely related consortia to application-oriented initiatives in other domains with strong requirements to utilizing CS methods.

We recognize a *great disciplinary proximity* with NFDI4DataScience and intend to develop a close work relationship. Concepts and technologies will be shared where appropriate and possible. Therefore, we will design an interface to NFDI4DataScience in common areas of interest where data types and services will be developed together according to shared goals. This applies especially to the following research data types: benchmarking data, certain types of software and services.

For meta-data-based search the concept of a knowledge graph will be used commonly but in separate research domains. The domain specific data will be handled according to practices and established mechanisms of the respective research domains.

In the field of *conceptually neighboring consortia* agreements with NFDI4ING and MaRDI were sketched for engineering and mathematical RDM, as engineering, mathematics and CS share a certain overlap. This will be explored according to the NFID4ING's archetypes and our research areas and types of research data to avoid reinvention or double development of services and software.

Formalized mathematics and its artifacts are a shared interest of MaRDI and NFDIxCS. The development of meta-data standards which can express semantic information from mathematics as well as CS, is an important step for implementing the FAIR principles. Within the collaboration between NFDIxCS and MaRDI, e.g. the archive of formal proofs can be integrated into the MaRDI-Portal and the MaRDI-Knowledge Graph. Similar interests are shared in HPC and scientific computing / numerical simulation in areas like fluid dynamics.

Considering these close connections, we will ensure to have representatives from NFDI4DataScience, NFDI4Ing and MaRDI as active players on interfaces with other consortia in our governance structure.

Furthermore, NFDIxCS agreed to cooperate with the application-oriented consortia NFDI4Chem, Punch4NFDI, NFDI4Mobility and FAIRmat. Collaboration with further application-oriented consortia including future NFDI candidates NFDI4Energy and NFDI4Phys is envisaged.

4 Cross-cutting topics

4.1 Cross-cutting topics that are relevant for the consortium and that need to be designed and developed by several or all NFDI consortia

Cross-cutting topics will be addressed in the context of and by means from CS. The special challenge for CS as a relatively young scientific discipline is the rapid evolution of standards, processes, and procedures. NFDIxCS plans to first agree upon general directions and approaches for such cross-cutting topics and then to develop and deliver them to the community in terms of (wherever possible: web-based) infrastructure services. At a general level the principles of **ethics, security, privacy, and safety** play an important role in RDM; monitoring for **quality** and **appropriate usage of resources** belong here as well.

Based on these principles a list of cross-cutting aspects with a very close connection to primary research topics in CS is given which can also be used in other realms:

- Identity management, persistent identifiers,
- Persistency of data, software and execution context,
- Standards to tackle heterogeneous, yet scalable environments:
    - Reference architectures and web-interfaces on systems level,
    - Metadata and protocols on application level,
    - Taxonomies/ontologies at a semantic level,
- Virtualization and interfaces to HPC,
- Variants and versions management,
- Tracking provenance of data and software,
- Quality and reputational management,
- Security against unauthorized access or change,
- Safety against loss of data or context,
- Anonymization/pseudonymization of data, differential privacy,
- Cost efficient storage and retrieval of context,
- Processes for data and software publication & attribution,
- Cross-site search and data/software exchange, and
- License and digital rights (IPR) management.

These cross-cutting topics are relevant in CS research data management and could be relevant for other disciplines as well. Moreover, ethical and legal issues as well as community development and cultural change will be relevant issues with a less technical focus in all consortia where we would highly benefit from. NFDIxCS seeks a close collaboration with the other NFDI initiatives and will reach out to the international community as well. We will always use established / emerging standards and solutions and will put great effort into the evolution and maintenance of related standards and services.

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

NFDIxCS will contribute to cross-cutting topics particularly in three areas:

- **Research data and software containers**: It will be necessary to provide a distributable and portable unit for packaging a whole range of data, meta information, software, and context for the various usage forms of the available research data. Specific variants of such types of units will be provided and can be provided for general use.
- **Privacy and data reuse**: An important and particular form of research data management is reuse of the stored data. Depending on the sub-discipline and research method the lawful reuse of data might involve complex privacy issues. These will be handled using general procedures developed within NFDIxCS. Such types of reuse procedures and related methods to achieve the desired level(s) of privacy are of general interest, and can be exported into other NFDI initiatives.
- **Virtualization and interfaces to HPC**: The demand for computationally intensive methods grows quickly. This makes convenient access to various types of computational resources an important requirement. Thus, virtualization of (traditional and advanced, e.g. GPU) resources will be important research support. In addition, even higher demands will be addressed by related interfaces to proper HPC resources.

Since issues at the intersection of **IT and law** are directly addressed in our consortium, we will bring this perspective into the NFDI community and intend to embed it into a broader context of legal aspects.

Furthermore, we appreciate contributing to an NFDI board on cross-cutting issues with representatives from our consortium with efficient solutions for technical issues. We would appreciate an intense exchange with the other consortia in order to better understand their specific requirements and to consider them in our work.

For all other cross-cutting issues we intend to contribute to NFDI-wide discussions from our disciplinary perspective, to articulate our specific requirements and existing approaches, to participate in upcoming developments, and to gradually transfer emerging methods and services to our field.