NFDI4DataScience - Letter of Intent

1 Binding letter of intent as advance notification of a full renewal proposal

- **Binding letter of intent**[1]

2 Formal details

- **Name of the consortium**
  NFDI for Data Science and Artificial Intelligence

- **Acronym of the consortium**
  NFDI4DataScience

- **Applicant institution**
  Fraunhofer
  Hansastraße 27c, 80686 München
  Prof. Dr.-Ing. Holger Hanselka
  Fraunhofer FOKUS (FOKUS)
  Kaiserin-Augusta-Allee 31, 10589 Berlin
  Prof. Dr. Manfred Hauswirth
- **Spokesperson**
  Prof. Dr. Sonja Schimmler, sonja.schimmler@fokus.fraunhofer.de

- **Co-applicant institution**
  Bauhaus-Universität Weimar (BUW)[2]
  Geschwister-Scholl-Straße 8, 99423 Weimar
  Prof. Peter Benz
- **Co-Spokesperson**
  Prof. Dr. Benno Stein, benno.stein@uni-weimar.de

- **Co-applicant institution**
  Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

[1] Required as advance notification for renewal proposals in 2025
[2] New partners are highlighted

Trippstadter Str. 122, 67663 Kaiserslautern
Prof. Dr. Antonio Krüger

- **Co-spokesperson**
  Prof. Dr. Georg Rehm, georg.rehm@dfki.de

- **Co-applicant institution**
  Fraunhofer FIT (FIT)
  Schloss Birlinghoven, 53757 Sankt Augustin
  Prof. Dr. Stefan Decker
  Co-spokesperson
  Dr. Zeyd Boukhers, zeyd.boukhers@fit.fraunhofer.de

- **Co-applicant institution**
  FIZ Karlsruhe - Leibniz Institute for Information Infrastructure (FIZ)
  Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen
  Prof. Dr. Wolfram Horstmann
- **Co-spokespersons**
  Prof. Dr. Harald Sack, harald.sack@fiz-karlsruhe.de
  Prof. Dr. Franziska Boehm, franziska.boehm@fiz-karlsruhe.de

- **Co-applicant institution**
  GESIS - Leibniz Institute for the Social Sciences (GESIS)
  Unter Sachsenhausen 6-8, 50667 Köln
  Prof. Dr. Christof Wolf
- **Co-spokespersons**
  Prof. Dr. Stefan Dietze, stefan.dietze@gesis.org
  Dr. Arnim Bleier, arnim.bleier@gesis.org
  Dr. Philipp Mayr, philipp.mayr@gesis.org
  Dr. Brigitte Mathiak, brigitte.mathiak@gesis.org

- **Co-applicant institution**
  Hochschule Wismar (HW)
  Postfach 1210, 23952 Wismar
  Prof. Dr. jur. Bodo Wiegand-Hoffmeister
- **Co-spokesperson**
  Prof. Dr.-Ing. Frank Krüger, frank.krueger@hs-wismar.de

- **Co-applicant institution**
  Leuphana University Lüneburg (LEU)
  Universitätsallee 1, 21335 Lüneburg
  Prof. Dr. Sascha Spoun
- **Co-spokesperson**
  Prof. Dr. Ricardo Usbeck, ricardo.usbeck@leuphana.de

- **Co-applicant institution**
  Leipzig University (LU)
  Ritterstraße 26, 04109 Leipzig

Prof. Eva Inés Obergfell

- **Co-spokesperson**
  Prof. Dr. Thomas Neumuth, thomas.neumuth@uni-leipzig.de

- **Co-applicant institution**
  Schloss Dagstuhl - Leibniz Center for Informatics (LZI)
  Oktavie-Allee, 66687 Wadern
  Prof. Dr.-Ing. Holger Hermanns
- **Co-spokespersons**
  Dr. Marcel R. Ackermann, marcel.r.ackermann@dagstuhl.de
  Dr. Michael Wagner, michael.wagner@dagstuhl.de

- **Co-applicant institution**
  RWTH Aachen University (RWTH)
  Templergraben 55, 52056 Aachen
  Prof. Dr. Dr. Ulrich Rüdiger
- **Co-spokesperson**
  Dr. Christoph Lange-Bever, lange@cs.rwth-aachen.de

- **Co-applicant institution**
  TIB Leibniz Information Centre for Science and Technology (TIB)
  Welfengarten 1 B, 30167 Hannover
  Prof. Dr. Sören Auer
- **Co-spokespersons**
  Prof. Dr. Sören Auer, auer@tib.eu
  Dr. Markus Stocker, markus.stocker@tib.eu

- **Co-applicant institution**
  Technische Universität Berlin (TUB)
  Straße des 17. Juni 135, 10623 Berlin
  Prof. Dr. Geraldine Rauch
- **Co-spokespersons**
  Prof. Dr. Sonja Schimmler, sonja.schimmler@fokus.fraunhofer.de
  Prof. Dr. Manfred Hauswirth, manfred.hauswirth@tu-berlin.de
  Prof. Dr. Volker Markl, volker.markl@tu-berlin.de
  Prof. Dr.-Ing. Sebastian Möller, sebastian.moeller@tu-berlin.de
  Prof. Dr. Ziawasch Abedjan, abedjan@tu-berlin.de

- **Co-applicant institution**
  Technische Universität Dresden (TUD)
  01062 Dresden
  Prof. Dr. Ursula M. Staudinger
- **Co-spokespersons**
  Prof. Dr. Wolfgang E. Nagel, wolfgang.nagel@tu-dresden.de
  Dr. Matthias Lieber, matthias.lieber@tu-dresden.de

- **Co-applicant institution**
  University of Cologne (UC)
  Albertus-Magnus-Platz, 50923 Köln
  Prof. Dr. Joybrato Mukherjee
- **Co-spokespersons**
  Prof. Dr. Oya Beyan, oya.beyan@uni-koeln.de
  Dr. Adamantios Koumpis, adamantios.koumpis@uk-koeln.de

- **Co-applicant institution**
  Weizenbaum Institute (WI)
  Hardenbergstraße 32, 10623 Berlin
  Prof. Dr. Christoph Neuberger
- **Co-spokesperson**
  Prof. Dr. Manfred Hauswirth, manfred.hauswirth@tu-berlin.de
  Prof. Dr. Sonja Schimmler, sonja.schimmler@fokus.fraunhofer.de

- **Co-applicant institution**
  ZB MED Information Centre for Life Sciences (ZB MED)
  Gleueler Straße 60, 50931 Köln
  Prof. Dr. Dietrich Rebholz-Schuhmann
- **Co-spokespersons**
  Prof. Dr. Dietrich Rebholz-Schuhmann, rebholz-schuhmann@zbmed.de
  Dr. Leyla Jael Castro, ljgarcia@zbmed.de

- **Co-applicant institution**
  ZBW - Leibniz Information Center for Economics (ZBW)
  Düsternbrooker Weg 120, 24105 Kiel
  Prof. Dr. Klaus Tochtermann
- **Co-spokesperson**
  Prof. Dr. Klaus Tochtermann, k.tochtermann@zbw.eu

3 Objectives, work programme and research environment in the second funding period

Research area of the proposed consortium[3]

4.43 Computer Science, (others as application)

[3] According to the DFG classification system: www.dfg.de/dfg_profil/gremien/fachkollegien/faecher/index.jsp

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

The importance of research data in computer and data science has steadily increased over the years, most notably for testing, evaluating, reproducing and training computational methods. In particular, within the field of AI and the rise of generative AI, data has become a key factor for advancing the state of the art in various research areas, including machine learning and its application in natural language processing and information retrieval as well as in domains such as biomedical science, physics and psychology. Data includes unstructured and (semi-)structured corpora, labelled benchmark and ground truth datasets, and experimental result and training data. Next to source code and software libraries, pretrained models and a plethora of benchmark datasets have become ubiquitous in computer and data science, where transparency about provenance, underlying data sources and model architectures have emerged as crucial challenges.

NFDI4DataScience (NFDI4DS) aims to establish a community-driven research data infrastructure for the **AI and data science community**. In this regard, we will focus on several types of data and artifacts established within the AI and data science communities: As in virtually all other disciplines, research contributions in AI and data science are conveyed via **scientific articles**. These articles are often accompanied by structured descriptions of research problems or tasks, benchmark and evaluation **datasets,** as well as **models** and **source code** (implementing the particular approach). To ensure transparency and reproducibility, it is important that all these artifacts are stored and systematically interlinked. NFDI4DS will continue to utilize recent advances in scholarly information processing to break those black boxes open and make them human- and machine-readable.

Knowledge graphs are increasingly used by industry for building large-scale data representations, e.g., by Google or Meta. For representing and managing scientific data and metadata, **research knowledge graphs** have seen wider adoption and will create added value for the computer and data science community in a similar way, and can interoperate with existing knowledge graphs. In particular, for AI and data science, knowledge graphs facilitate transparency and reproducibility.

NFDI4DS **key objectives** are to develop and maintain a **community-driven research data infrastructure** for systematically managing the complete **research data lifecycle**, including creation/collection, processing, analyzing, preserving and reusing relevant artifacts in a coherent, distributed and interoperable manner. One main goal is to overcome the replication crisis, which is currently an important challenge in this domain.

- Using our network, we will continue to foster deep exchange on AI and data science. We will bring together the currently rapidly growing and evolving **AI and data science community**. We will proactively identify emerging community needs and will foster exchange of expertise.

- We will further provide **training on best practices** for AI and data science research, thus educating local multiplicators which will implement these data-driven strategies, fostering a new generation of AI and data science experts.
- We will deepen our emphasis on coping with **ethical, legal and social aspects** (ELSA) of AI and data science. The recognition of these aspects will be strengthened in the community, and competencies on these topics will be developed via training activities.
- We will continue to interpret the **FAIR principles** and the future of science to what it means for our community and implement it in our context, also making use of FAIR Digital Objects (FDOs) to ensure interoperability, reusability and machine-actionability.
- We will further develop, consolidate, integrate and maintain our community-driven **research data infrastructure**, which includes several core services. All researchers with a focus on AI and data science will be invited to adapt and reuse these infrastructures and services.
- We will put forward open **research knowledge graphs**, and provide specific tools and services for benefiting from their underlying information. All consortia that plan to use research knowledge graphs as well will benefit from these activities.

The key objectives of NFDI4DS will be addressed by **five main task areas**:

- **(A) Benchmarking, Shared Tasks & Community**, focusing on methods, benchmarking and shared tasks, specifically addressing ELSA aspects and the future of science with the aim of ensuring transparency and reproducibility of AI and data science research;
- **(B) Research Knowledge Graphs**, providing knowledge graphs that allow for FAIR access to scientific knowledge and metadata in AI and enabling a federation over the growing amount of disciplinary knowledge graphs - as a nexus;
- **(C) Infrastructure & Services**, targeting consolidation and integration, and extending on artifact storage and access to foster community outreach;
- **(D) Research Co-pilots,** establishing access points and executing different use cases, to widen the applicability of the NFDI4DS resources and
- **(E) Management**, organizing the consortium.

The **research data infrastructure** will continue to be built bottom-up, i.e., building on standards and practices that exist in the AI and data science community, which are open and extensible. We are regularly performing surveys and interviews in the AI and data science community to better understand its needs, and user tests to get feedback early on. We are collaborating with the dynamically growing community, where currently many new professorships are being put in place, and new bachelor's and master's degrees are being set up.

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

In recent years, there has been a dramatic growth in terms of breadth and depth of knowledge assets and services for AI and data science, for example, including task descriptions, datasets and leaderboards on platforms such as Kaggle, or Papers With Code, and large (bibliographic) knowledge graphs like Wikidata, or SemOpenAlex. In order to truly realize the potential of AI and data science, we need to **synergistically integrate** such **knowledge assets** and **services**.

The many **success stories of the NFDI4DS partners** demonstrate the experience we bring on board: This includes infrastructures such as the European Language Data Space, coordinated by DFKI, which embodies the European language marketplace, and data.europa.eu, the European open data platform for the public sector, developed by Fraunhofer FOKUS; domain-specific search engines like GESIS Search, operated by GESIS, and repositories like PUBLISSO, operated by ZB Med; and research knowledge graphs such as the GESIS Knowledge Graph (GESIS KG), provided by GESIS, Open Research Knowledge Graph (ORKG), managed by TIB, a digital infrastructure for FAIR scholarly knowledge, and the dblp computer science bibliography, operated by LZI, the world's largest metadata collection about computer science publications.

The **core services** of the NFDI4DS infrastructure will be established and sustained in the second funding phase. Two building blocks are the **NFDI4DS Gateway**, a service hub integrating all artifact sources and services relevant for the AI and data science community, and the **NFDI4DS Portal**, a portal giving access to all artifact sources available in the NFDI. Regarding **scientific articles,** we will continue to expand and integrate community-specific bibliographic metadata services (e.g. dblp), and advance and integrate community-specific repositories (e.g. CEUR-WS, DROPS, arXiv). With respect to **research data**, we will further ensure the availability and reusability of high-quality data for AI and data science research by providing benchmarking corpora and integrating repositories for AI and data science. Targeting **models** and **code**, we will continue to ensure accessibility and reusability of computational models and basic data science methods for research purposes, and by integrating with established repositories (e.g. GitHub, Hugging Face). With respect to **knowledge graphs**, we will establish a public knowledge graph of metadata of AI and data science resources, as well as semantic descriptions of research contributions based on scientific knowledge graphs. We will further develop our **FAIR digital object**-based solution to make all artifacts fully compliant and machine-actionable. We will also continue developing our **training materials** on AI and data science best practices, contributing to the broader NFDI and EOSC activities. For each of our services, we will realize **horizontal aspects**, such as access management or persistent identification, collaborating closely with the **NFDI sections** and using **NFDI basic services** where possible. This is to ensure compliance with the open science and FAIR principles.

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

Similar to computer science, AI and data science inhabit two roles, being a discipline itself and also acting as a 'supporting science'. One of the very first outcomes of this interchange was the Leipzig-Berlin Declaration driven by our consortium. Our main goal is to enable more effective and efficient AI and data science research, and meanwhile ensuring transparency and reproducibility thereof. In this scope, we are actively involved in all **NFDI sections**, to collaborate closely with other consortia on cross-cutting topics. At the moment, we are acting as spokesperson of the section common infrastructures and as co-spokesperson of the section (meta)data, terminologies, provenance. We are also deeply involved in the working groups overall architecture (section infra), AI and data science (section infra) and ethics (section elsa).

We have demonstrated our collaboration with literally all **other consortia** in multiple joint activities such as hackathons and will continue and broaden this in the second phase. Most members of NFDI4DS are also involved in other NFDI consortia as well: BERD@NFDI, FAIRAgro, FAIRmat, GHGA, KonsortSWD (NFDI4Society), MaRDI, NFDI4Cat, NFDI4Chem, NFDI4Culture, NFDI4Earth, NFDI4Energy, NFDI4Health, NFDI-MatWerk, NFDI4Memory, NFDI4Ing, NFDIxCS and Text+. Furthermore, we collaborate with other consortia that have a strong focus on AI and data **science**, including PUNCH4NFDI.

NFDI4DS has agreed on a close partnership with **NFDIxCS**, the other consortia in the computer science domain. Together, we will provide a well-coordinated infrastructure: NFDIxCS will cover computer science as a whole, whereas we will concentrate on AI and data science, altogether with their domain-specific applications. We have also agreed on a close collaboration with **MaRDI**, which will approach AI and data science from a mathematical perspective.

Our **speedboat projects**, which have been successfully established in the first funding phase, will be intensified. The speedboat projects focus on community engagement and the development of innovative tools and services, often in collaboration with new domains and with other consortia.

We are also actively engaged in **Base4NFDI**, putting forward basic services for our overall infrastructure to contribute to the vision of 'One NFDI'. We have been involved early on, our spokesperson acting as co-spokesperson, and our partner institutions being involved in literally all basic services. We specifically implemented an IAM incubator project, and intensely contributed to KGI4NFDI, Jupyter4NFDI and nfdi.software.

4 International and national networking

NFDI4DS builds on experience available within the consortium as well as via national and international research projects and wider community initiatives.

On a **national level**, we are cooperating closely with the **AI Competence Centers**, especially BIFOLD in Berlin and ScaDS.AI in Dresden/Leipzig, and are in close contact with the **NHR Centers**. We synergistically cooperate with the **National Data Competency Centers**, where several of our partners are directly involved. On a **European level**, we are putting an emphasis on **RDA** and **EOSC**. We are actively involved in various working groups and attend relevant events. In the second phase, we especially plan to bring in our services in the **German EOSC National Node**.

To engage with the **AI and data science community**, we collaborate closely with the **Gesellschaft für Informatik (GI)**. We also regularly organize events, such as our yearly summer school, targeting early-career researchers. We are also frequently present at diverse scientific conferences, offering formats such as workshops and tutorials to interested researchers. To engage with the **research data infrastructure community**, we established a monthly lecture series and a yearly conference. We are regularly present at conferences with workshops and shared tasks. We are also performing several hackathons a year with international participation. Targeting **society**, we are regularly hosting events such as our yearly science slam.

We are continuously contributing to **political debates** and are driving **standardisation efforts**, mainly on a national and European level. With regards to political activities, this includes our involvement with regards to the **German Research Data Act** and the **European AI Act**. In terms of standardisation, DFKI hosts the German chapter of the **World Wide Web Consortium (W3C)**, which develops and standardizes the technical building blocks of the World Wide Web. This year, for instance, we put forward a DIN spec on ‘knowledge graphs for language models and language models for knowledge graphs’.

Towards **industry**, we will continue maintaining our connections to the **Common European Data Spaces** (e.g., for language), to the Fraunhofer-coordinated **Data Spaces Support Centre** that supports their building, and to business-to-business data sharing initiatives such as Gaia-X, all of which are increasingly transitioning from mere trusted data exchange to AI services. We will also leverage our links to the **AI Service Centers** and their EU-level counterpart, the **AI Factories**.