NFDI4DataScience - Letter of Intent

1. Binding letter of intent as advance notification

Binding letter of intent (submission in 2020)

2. Formal details

Name of the planned consortium

NFDI for Data Science and Artificial Intelligence

Acronym of the planned consortium

NFDI4DataScience

Applicant institution & spokesperson

Applicant institution	Spokesperson
Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. Prof. Dr.-Ing. Reimund Neugebauer Hansastraße 27 c 80686 München	**Dr. Sonja Schimmler** sonja.schimmler@fokus.fraunhofer.de Fraunhofer FOKUS, Weizenbaum Institute

Co-applicant institutions & Co-spokespersons

Co-applicant institution	Co-spokesperson
DFKI Alt-Moabit 91c 10559 Berlin	**Prof. Dr.-Ing. Sebastian Möller** sebastian.moeller@tu-berlin.de **Dr. Georg Rehm** georg.rehm@dfki.de
FIZ Karlsruhe Hermann-von-Helmholtz-Platz 1 76344 Eggenstein-Leopoldshafen	**Prof. Dr. Harald Sack** harald.sack@fiz-karlsruhe.de Karlsruhe Institute of Technology, FIZ Karlsruhe
Fraunhofer FIT Schloss Birlinghoven Konrad-Adenauer-Straße 53754 Sankt Augustin	**Dr. Oya Beyan** beyan@fit.fraunhofer.de RWTH Aachen, Fraunhofer FIT **Dr. Christoph Lange-Bever** lange@informatik.rwth-aachen.de RWTH Aachen, Fraunhofer FIT

GESIS - Leibniz Institute for the Social Sciences Unter Sachsenhausen 6-8 50667 Köln	**Prof. Dr. Stefan Dietze** stefan.dietze@gesis.org GESIS - Leibniz Institute for the Social Sciences, Heinrich-Heine-University Düsseldorf, L3S Research Center **Prof. Dr. Claudia Wagner** claudia.wagner@gesis.org GESIS - Leibniz Institute for the Social Sciences, University of Koblenz Landau
Hamburger Informatik Technologie-Center e. V. Vogt-Kölln-Str. 30 22527 Hamburg	**Dr. Lothar Hotz** hotz@informatik.uni-hamburg.de HITeC e.V. Hamburg
Leibniz University Hannover Welfengarten 1 30167 Hannover	**Prof. Dr. Ziawasch Abedjan** abedjan@dbs.uni-hannover.de Leibniz University of Hannover, BIFOLD
RWTH Aachen University Templergraben 55 52056 Aachen	**Dr. Christoph Lange-Bever** lange@informatik.rwth-aachen.de RWTH Aachen, Fraunhofer FIT **Dr. Oya Beyan** beyan@fit.fraunhofer.de RWTH Aachen, Fraunhofer FIT
Schloss Dagstuhl - Leibniz Center for Informatics Oktavie-Allee 66687 Wadern	**Prof. Raimund Seidel, Ph.D.** raimund.seidel@dagstuhl.de Schloss Dagstuhl LZI, Saarland University **Dr. Marcel R. Ackermann** marcel.r.ackermann@dagstuhl.de Schloss Dagstuhl LZI **Dr. Michael Wagner** michael.wagner@dagstuhl.de Schloss Dagstuhl LZI
Technische Universität Dresden Helmholtzstr. 10 01069 Dresden	**Prof. Dr. Wolfgang E. Nagel** wolfgang.nagel@tu-dresden.de Technische Universität Dresden, Center for Information Services and High Performance Computing, ScaDS.AI Dresden/Leipzig
TIB Leibniz Information Centre for Science and Technology Welfengarten 1b 30167 Hannover	**Prof. Dr. Sören Auer** auer@tib.eu Leibniz University of Hannover, TIB **Dr. Markus Stocker** markus.stocker@tib.eu TIB
TU Berlin Straße des 17. Juni 135 10623 Berlin	**Prof. Dr. Manfred Hauswirth** manfred.hauswirth@tu-berlin.de TU Berlin, Fraunhofer FOKUS, Weizenbaum Institute

	Prof. Dr. Volker Markl volker.markl@tu-berlin.de TU Berlin, BBDC, BZML, BIFOLD **Prof. Dr.-Ing. Sebastian Möller** sebastian.moeller@tu-berlin.de TU Berlin, DFKI
Universität Leipzig Ritterstraße 26 04109 Leipzig	**Prof. Dr. Thomas Neumuth** thomas.neumuth@uni-leipzig.de Innovation Center Computer Assisted Surgery, Universität Leipzig, ScaDS.AI Dresden/Leipzig
ZB MED - Information Centre for Life Sciences Gleueler Str. 60 50931 Köln	**Prof. Dr. Dietrich Rebholz-Schuhmann** rebholz@zbmed.de University of Cologne, ZB MED
ZBW Leibniz Information Centre for Economics Düsternbrooker Weg 120 24105 Kiel	**Prof. Dr. Klaus Tochtermann** k.tochtermann@zbw.eu Kiel University, ZBW

Participants

Participant institution
AWI **Prof. Dr. Frank Oliver Glöckne**r, frank.oliver.gloeckner@awi.de Jacobs University Bremen gGmbH, AWI
FHI **Prof. Dr. Matthias Scheffler**, scheffler@fhi-berlin.mpg.de Fritz-Haber-Institut der Max-Planck-Gesellschaft **PD Dr. Carsten Baldauf**, baldauf@fhi-berlin.mpg.de Fritz-Haber-Institut der Max-Planck-Gesellschaft
FIZ Karlsruhe **Prof. Dr. Franziska Boehm,** franziska.boehm@fiz-karlsruhe.de Karlsruhe Institute of Technology, FIZ Karlsruhe **Thomas Hartmann**, tho.hartmann@fiz-karlsruhe.de FIZ Karlsruhe
Wikimedia Deutschland e.V. **Franziska Heine**, franziska.heine@wikimedia.de Wikimedia Deutschland e.V.

3. Objectives, work programme and research environment

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

409, Computer Science, (312, Mathematics; others as application)

Concise summary of the planned 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 Artificial Intelligence and the rise of deep and transfer learning, data has become a key factor for advancing the state of the art in various fields, including natural language processing, machine learning and information retrieval. Data includes unstructured, semi-structured and structured corpora, labelled benchmark and ground truth datasets, experimental result data as well as training data. Next to source code and software libraries, pretrained models have become a ubiquitous ingredient of Computer and Data Science, where transparency about provenance, underlying data sources (including aspects such as bias) and model architectures have emerged as crucial challenges.

NFDI4DataScience aims to establish a community-driven research data infrastructure for the *Data Science and Artificial Intelligence community within Computer Science*. In this regard, we will focus on several types of data and artifacts established within the Data Science and Artificial Intelligence communities:

- Scientific articles
- Research data: Structured task and benchmark definitions; benchmark and evaluation datasets (including raw data and experiment output); training data for supervised machine learning and language models
- Software: Source code and libraries (including documentation and configuration); scripts and executable notebooks; (pretrained) machine learning models; OS-level virtual containers such as docker images
- Knowledge graphs

As in virtually all other disciplines, research contributions in Data Science and Artificial Intelligence are conveyed via scientific articles. These articles are often accompanied by structured descriptions of research problems or tasks as well as source code (implementing the particular approach), and benchmark and evaluation datasets. For ensuring reproducibility, it is important that all these artefacts are stored and systematically interlinked.

Semantic Web and Linked Data are core technologies supporting the integration of machine-processable semantics. They are increasingly used by industry for building large-scale knowledge graphs, e.g., by Google, Facebook or Springer. *Scientific knowledge graphs* are also increasingly used and will create an added value for the Computer and Data Science community in a similar way, and can interoperate with existing knowledge graphs. In particular, for Data Science and Artificial Intelligence, knowledge graphs play a key role to realize trustworthiness, responsibility, reliability, explainability and transferability.

In recent years, there has been a dramatic growth in terms of breadth and depth of knowledge assets for Data Science and Artificial Intelligence, for example, including *task descriptions, datasets* and *leaderboards* on platforms such as Kaggle or Papers With Code, *bibliographic graphs* such as dblp, Microsoft Academic Graph, Springer's SciGraph, Research Graph or OpenCitations and *large knowledge graphs* such as DBpedia, Wikidata or YAGO. In order to truly realize the potential of Data Science and Artificial Intelligence we need to systematically interlink and synergistically integrate such data assets and services.

The key objectives of NFDI4DataScience are summarized in the following. These will be addressed by six main task areas: data science community & training; data science artefacts; data science infrastructure & services; data science transfer & application; data science in context; and data science management.

NFDI4DataScience aims to develop and maintain a *community-driven research data infrastructure* for systematically managing the complete data lifecycle including maintaining, describing, extracting, publishing 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. The *core services* of the planned infrastructure are:

- Scientific articles: Expand and integrate community-specific bibliographic metadata services, e.g., dblp. Advance and integrate community-specific open access repositories, e.g., CEUR-WS, DROPS, or arXiv.
- Research data: Advance and integrate data repositories for Data Science and Artificial Intelligence.
- Software: Enrich and integrate source code repositories, e.g., Git. Realize an executable notebook platform and comparable infrastructures to ensure the reproducibility of Data Science and Artificial Intelligence research.
- Knowledge graphs: Establish a public repository for semantic descriptions of research contributions based on scientific knowledge graphs, i.e., ORKG. Collect and integrate semantic resources (such as terminologies, controlled vocabularies, ontologies) into a knowledge graph representation, and link knowledge graphs via reference ontologies.
- Interoperability: Ensure that the NFDI4DataScience core services interoperate with each other, with other NFDI services, and beyond.

For each of these services, we will realize horizontal aspects, such as *persistent identification*, *licensing*, *provenance and sovereignty information tracking, semantic integration, visualization*, *monitoring*, and *long-term archiving*. This is to ensure the compliance with the *Open Science* and *FAIR principles* as well as *Data on the Web Best Practices*.

The infrastructure will be built bottom-up, i.e., building on standards and practices that exist in the Computer and Data Science communities, which are open and extensible. We are currently performing a survey in the Data Science and Artificial Intelligence community, with the goal to better understand its needs, and to foster collaboration. We aim to accompany this dynamically growing community, where currently many new professorships are put in place, and new bachelor's and master's degrees are set up.

To further support user participation and involvement, we will perform a *requirements analysis* of the envisioned infrastructure; foster *collaboration* to keep up to date with developments in the research data community; reach out to facilitate *awareness* and *community involvement*; establish *governance processes* for joining and using the infrastructure; collect best practices, and use them for *education and training*; accompany the whole process with a discussion of *legal and ethical aspects*; and create standards, guidelines and supporting materials and act as a driver for establishing these *methodologies and standards*.

Proposed use of existing infrastructures, tools and services

NFDI4DataScience will build on experience available within the consortium as well as via national and international research projects and wider community initiatives, such as *RDA* or *EOSC*. To do so, we will actively engage with other relevant initiatives. We will integrate and extend existing technical solutions, and will ensure interoperability between infrastructures. The many *success stories of the NFDI4DataScience partners* demonstrate the large amount of experience we bring on board in data engineering and management:

Initiatives our partners are involved in: We collaborate closely with the *working group on Data Science* of the Gesellschaft für Informatik (GI), which brings together the Data Science community in Germany. DFKI hosts the German chapter of the *World Wide Web Consortium (W3C)*, which develops and standardises the technical building blocks of the World Wide Web (e.g., HTML, XML, standards related to the Semantic Web and Linked Data). Several of our partners are members of *DataCite e.V.*, the leading global non-profit organisation that provides persistent identifiers (DOIs) for research data, software and other artefacts since 2009.

Infrastructures from our partners that are planned being used for our community include: Schloss Dagstuhl LZI operates the *dblp computer science bibliography*, the world's most comprehensive, curated, open metadata collection and knowledge graph of Computer Science publications. Wikimedia Deutschland e.V. is the developer of the very successful crowdsourcing project *Wikidata*, which is an open knowledge base for humans and machines. It is a central storage for the structured data of Wikipedia, and other projects. The *European Language Grid (ELG)*, coordinated by DFKI, is developing the primary European platform for Language Technology and NLP. DFKI is also involved in the *AI4EU* consortium, which establishes the first European Artificial Intelligence on-demand platform and ecosystem. HiTEC's *GERBIL* is a FAIR, open source benchmarking platform for several NLP tasks which has already run over 100.000 experiments. In recent years, TIB and L3S Research Center of Leibniz University Hannover have intensively worked on the *Open Research Knowledge Graph (ORKG)*, a digital infrastrastructure for FAIR scholarly knowledge where research contributions are represented in machine-actionable form.

Infrastructures from our partners developed for other domains include: ZBW brings in the DFG-funded project *GeRDI*, a research data infrastructure to store, share and re-use research data across disciplines with an emphasis on small data. Fraunhofer FOKUS is the developer of the core technical components of the *European Data Portal (EDP)*, a central access point for metadata of heterogeneous open data published by public authorities in Europe. Next to hosting a number of dataset registries and portals such as *da|ra*, GESIS contributes a range of research knowledge graphs and infrastructures for creating those. ZB MED and TIB are providing specialized data repositories for microbial researchers using graph technologies and text mining in the life science domain. Our partners are involved in several *GO FAIR Implementation Networks*: GESIS coordinates *GO Inter*, which aims at fostering semantic interoperability of heterogeneous research data across domains. Fraunhofer FIT co-coordinates *Personal Health Train*, which develops an infrastructure for distributed analysis of sensitive health data. FHI co-coordinates *NOMAD*, the computational materials science pillar of FAIR-DI, an association that was founded to make research data from several fields available according to the FAIR principles.

Interfaces to other proposed NFDI consortia

NFDI4DataScience has agreed on a close partnership with *NFDIxCS* and *MarDI*, the two other consortia in the Computer Science and the Mathematics domain. Together, we will provide a well coordinated infrastructure: *NFDIxCS* will cover Computer Science as a whole, whereas we will concentrate on Data Science and Artificial Intelligence, altogether with their domain-specific applications. *MarDI* will cover Mathematics as a whole, and will briefly address aspects of Data Science and Artificial Intelligence from a Mathematics point of view. Our consortium will cover the topic as whole, and will approach it from other perspectives.

Similar to Computer Science, Data Science and Artificial Intelligence inhabits two roles, being a discipline itself and also acting as a “supporting science”. Considering the second role, we are in close contact with all cross-cutting consortia, including *2linkNFDI*, *Bridge4NFDI*, *CompeNDI*, *NFDI4LifeUmbrella*, *NFDI4RSE* and *NFDI Web*. One of the first outcomes of this interchange is the Leipzig-Berlin Erklärung.

Most members of NFDI4DataScience are also involved in other NFDI consortia. Amongst others, we act as co-spokespersons in the consortia *NFDI4Culture*, *KonsortSWD*, *NFDI4Health*, *NFDI4BioDiversity*, *NFDI4Cat*, *NDFI4Chem* and *NFDI4Ing* from the first round. Furthermore, we are in contact with other consortia that have a strong focus on Data Science and Artificial Intelligence, including *NFDI4Microbiota* and *NFDI-MatWerk*, and agreed on collaboration.

Using our network, we will foster deep exchange on Data Science and Artificial Intelligence, allowing all stakeholders to gain from the expertise available. Collaboration inside the NFDI is planned being implemented via working groups and other formats.

To foster exchange even more, we will include *use cases* from different disciplines, ranging from social sciences and humanities, over life sciences and natural sciences to engineering sciences. We will start off with three use cases from within Data Science and Artificial Intelligence intense disciplines, and will add further ones later on. To identify suitable use cases, we will use our network within the NFDI and beyond.

4. Cross-cutting topics

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

NFDI4DataScience has been actively involved in formulating the Leipzig-Berlin Erklärung, and has signed it as Bridge4NFDI. We think that the paper is a good starting point for collaborating on cross-cutting topics within the NFDI. NFDI4DataScience understands cross-cutting topics as an essential vehicle for collaboration with the aim to share expertise and experience gained.

For our consortium, all aspects mentioned in the paper will be relevant to a certain extent. Most prominently, coping with very specific types of data and artefacts, ranging from publications and research data to software and knowledge graphs will be central for our consortium.

Cross-cutting topics our consortium will contribute to and how

Our main goal is to enable more effective and efficient Data Science and Artificial Intelligence research, and meanwhile ensure reproducibility thereof. In this scope, our consortium will contribute to several cross-cutting topics.

We will *bring together* the currently rapidly growing and evolving Data Science and Artificial Intelligence community within Computer Science and beyond. We will pick up their needs and will foster exchange of expertise.

We will provide training on reproducible Data Science and Artificial Intelligence research, thus *educating* local multiplicators which will implement these data-driven strategies into their curricula, fostering a new generation of Data Science an Artificial Intelligence experts able to benefit from the cornucopia of existing resources in NFDI4DataScience.

We will further put an emphasis on coping with *legal and ethical aspects*. The recognition of legal and ethical aspects of Data Science and Artificial Intelligence will be strengthened in the community, and competencies on these topics will be developed via training activities.

We will interpret the *FAIR principles* to what they mean for our community and will implement them in our context. The outcomes will be beneficial for all researchers dealing with Data Science and Artificial Intelligence.

We will develop and maintain a *community-driven research data infrastructure*, which includes several core services. All researchers with a focus on Data Science and Artificial Intelligence will be invited to adapt and reuse this infrastructure and services.

We will put forward *open research knowledge graphs*, and provide specific tools and services for coping with them. All consortia that plan to use research knowledge graphs as well will benefit from these activities.