Letter of Intent - Bridge4NFDI

(1) Non-binding letter of intent

- anticipated submission in 2020

(2) Formal details

Planned name of the consortium

- Bridging boundaries among national research data infrastructures

Acronym of the planned consortium

- Bridge4NFDI

Applicant institution

- **Fraunhofer FOKUS** (Fraunhofer Institute for Open Communication Systems)
    - Kaiserin-Augusta-Allee 31, 10589 Berlin
    - Prof. Dr. Manfred Hauswirth & Prof. Dr.-Ing. Ina Schieferdecker

Spokesperson

- **Dr. Sonja Schimmler**, sonja.schimmler@fokus.fraunhofer.de
    - **Fraunhofer FOKUS** (Fraunhofer Institute for Open Communication Systems)
    - **Weizenbaum Institute for the Networked Society**

(3) Objectives, work programme and research environment

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

- Cross-sectional, does not apply

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

The NFDI initiative will fund consortia from various scientific disciplines. It is anticipated that most consortia will target the specific requirements of their disciplines, which is necessary. However, this sole focus on individual disciplines, e.g., according to the DFG classification system, faces the obvious **risk of leading to even further fragmentation and more data silos**, i.e., less integration. For example, NASA estimates that some of its data funded through a couple of billion US dollars is no longer accessible due to heterogeneous data storage in silos and lack of documentation.[1] In fact, a major challenge for the formation of a common, shared NFDI for Germany and beyond that in Europe and worldwide, is the **establishment of shared infrastructures and principles**, which tackle also generic requirements, which will lead to a distributed yet cooperating network of scientific data management systems. Finding of a "common ground" regarding technical, social, cultural and economic aspects of research data management needs to be addressed from the very start of the NFDI funding.

The establishment of shared infrastructure and principles **enables multidisciplinary[2] research**, which helps scientists to solve complex problems, and which may lead to innovations by providing new angles of looking at known problems. Cross-domain interoperability is the key element for multidisciplinary collaboration. Although high quality disciplinary data infrastructures are essential for the advancement of any discipline, the ability to use this data for multidisciplinary research enables scientists to address complex research questions and facilitates innovation and may even lead to novel research questions.

Multidisciplinary scientific knowledge graphs are the key building blocks for machine-processable knowledge. 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 and Facebook. Multidisciplinary scientific knowledge graphs will create an added value for the scientific community in a similar way, and will be interoperable with existing knowledge graphs.

Vertical domain-specific research data infrastructures will identify domain-specific methods for capturing and publishing metadata, for accessing data resources, for ensuring interoperability and for identifying protocols concerning data reuse. **Bridge4NFDI** will embrace this heterogeneity in the technical, semantic, and organizational layers of the infrastructure provided by the NFDI consortia. Complementing this, **a horizontal multidisciplinary meta-infrastructure** will increase discoverability and effective use of domain-specific research data, with a perspective on supporting multidisciplinary research and data-driven innovation.

The **Bridge4NFDI** consortium will build a **semantic layer** on top of them, for bridging the self-contained data infrastructures, which will enable **harmonized interfaces for humans and machines** with the overall vision of building a **general access point** for the (meta-)data of the NFDI consortia and other existing infrastructures. The semantic layer will support services, applications, metadata and data interoperability for distributed, federated resources. Our infrastructure will be developed in the spirit of **Semantic Web and Linked Open Data (LOD)**, essentially meaning that we have vocabularies and mappings among data sets. **The infrastructure will be built bottom-up**, i.e., building on standards and practices that exist in the scientific communities, which are open and extensible. Currently, the open data world (DCAT, CKAN, LOD, schema.org, etc.) and the research data world (RDA, DDI, DataCite, etc.) are

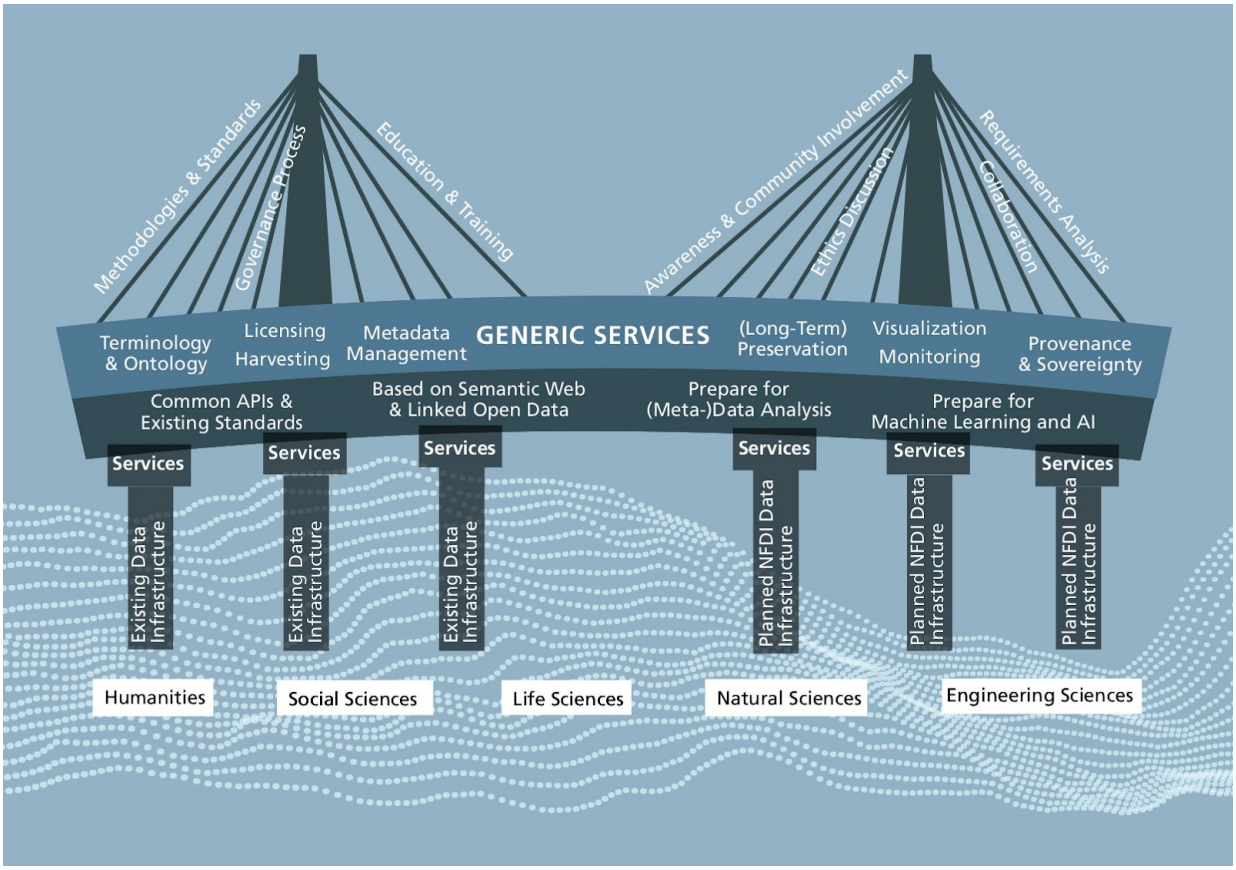
[1] https://www.economist.com/leaders/2012/04/28/bit-rot/
[2] Multidisciplinary, crossdisciplinary, interdisciplinary, or even transdisciplinary

still disparate. It is our goal to close this gap, and use open data standards to facilitate FAIR research data infrastructures.

The involvement of discipline-specific NFDI consortia in this shared infrastructure is key to achieve our goals. A **bottom-up, integrative and non-prescriptive approach** seems most promising here to build a national research data service. We plan **to reuse tools from different disciplines, to generalize them where possible and to transfer them from one discipline to another**. This will avoid that each community (re-)invents similar techniques, and instead join these efforts and facilitate reuse.

Our **main idea** is to provide a **(meta-)infrastructure for research data, which comprises a variety of generic services that can be used by the other NFDI consortia**.

The following diagram provides an overview of the planned functionalities. Technical functionality (non-exhaustive list): **harvesting service, metadata management service, visualization service, (long-term) preservation service, monitoring service, licensing service, provenance & sovereignty service,** and **terminology & ontology service**. Technical goals: **common APIs & existing standards**, **prepare for (meta-)data analysis**, and **prepare for machine learning & AI**. Measures to support user participation and involvement: **methodologies & standards**, **governance process**, **education & training**, **awareness & community involvement**, **ethics discussion**, **collaboration**, and **requirements analysis**.



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

It is pertinent to build on experience, which is available **nationally** as well as **internationally**. To do so, **we will stay in close contact with relevant other initiatives, to integrate and extend existing solutions, wherever possible.** Some of these initiatives and solutions are mentioned below.

With the **European Open Science Cloud**[3] **(EOSC)**, a meta-infrastructure is under development, following a similar approach on the European level. **OpenAIRE**[4] is another European initiative, which is an open ecosystem for scholarly publishing. Furthermore, it is pertinent to learn from initiatives, such as **B2FIND**[5] and **DataONE**[6], which offer a meta search across EUDAT data centers and direct access to earth observational data through a distributed network, respectively.

Bridge4NFDI will connect **experts from different disciplines and from many domain communities**, who have addressed generic research data management challenges for years. There exist several **success stories of cross-domain interoperability frameworks** of the **Bridge4NFDI** partners, which demonstrate the big amount of experience we have on board:

GeRDI[7] is a running DFG project to create a research data infrastructure to store, share and re-use research data across disciplines with an emphasis on small amounts of data. Kiel University develops its architecture, whereas ZBW focuses on the metadata and operational aspects of the project.

DataCite[8] e.V. is the leading global non-profit organisation that provides persistent identifiers (DOIs) for research data, emerged from a successful DFG project and provides generic research data management services since 2009. GESIS, TIB, ZBW, and ZB MED are members of the e.V.

FAIR-DI[9] (FAIR Data Infrastructure for Physics, Chemistry, Materials Science, and Astronomy e.V.) is an association that was founded to make the treasure trove of research data from several fields available according to the FAIR principles. The **NOMAD Repository and Archive**, i.e., the computational materials science pillar of FAIR-DI, was accepted as Go-FAIR Implementation Network.

The **European Data Portal**[10] (EDP) is a central access point for metadata of heterogeneous Open Data published by public authorities in Europe with close to 900.000 datasets, 60 million RDF triples in total, from 77 data providers. The EDP is Europe's Linked Data-enabled one-stop-shop for open public sector information. Fraunhofer FOKUS is the developer of the core technical components of the EDP.

Wikidata[11] (in combination with **Wikibase**) is a free and open knowledge base that can be read and edited by humans and machines. It is a central storage for the structured data of Wikipedia, Wikivoyage, Wiktionary, Wikisource, and other projects. Wikimedia Deutschland is the developer of these very successful crowdsourcing projects.

[3] https://www.eosc-portal.eu/
[4] https://monitor.openaire.eu/
[5] https://eudat.eu/services/b2find/
[6] https://www.dataone.org/
[7] https://www.gerdi-project.eu/
[8] https://datacite.org/
[9] https://fairdi.eu/
[10] https://www.europeandataportal.eu/
[11] https://www.wikidata.org/

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

Our consortium is complementary to all other disciplinary proposals, and not in competition to other initiatives. We want to give technical as well as non-technical support, and we hope that the other future consortia will benefit from the outcomes of our work. Our aim is to build a cross-disciplinary, integrating infrastructure. Thus, it **does not target a specific discipline**. Instead, **Bridge4NFDI** focuses on the provisioning of a network of generic and domain-specific research data systems and services.

As such, the **Bridge4NFDI consortium** includes **members from different disciplines and domain communities** that have a lot of data and technical know-how already available to cover all scientific disciplines according to the DFG classification: Humanities and Social Sciences, Life Sciences, Natural Sciences and Engineering Sciences. This enables us to stay in close contact to the requirements stated by researchers representing the different disciplines.

Several members of our consortium are involved in **domain-specific consortia** (e.g., KonsortSWD, MaRDI, NFDI4BioDiversity, NFDI4Chem, NFDI4Crime, NFDI4Culture, NFDI4Earth, NFDI4Health, NFDI4Memory, NFDI4MSE, NFDI4MobilTech, NFDI4NanoSafety, NFDI4Physics, Web4NFDI), **semi-cross-sectional consortia** (e.g., NFDI4Ing, NFDI4Life) and other **cross-sectional consortia** (e.g., RSE4NFDI). This serves as the basis for deep exchange among different NFDI consortia, necessary for such a project.

It is anticipated to **include a few consortia as domain-specific use cases** in our project. These use cases will serve as a test bed for the cross-disciplinary, integrating infrastructure developed. At this point, we are already connected the domain-specific consortia mentioned above and also in contact with several other initiatives (e.g., DeBioData, NFDINeuro). In the next few months, we will concretise our plans and will reach out to other domain-specific consortia as well, if necessary.

In the next few months, we will also further **coordinate with the other cross-sectional consortia**, especially with 2linkNFDI, RSE4NFDI, and AI4NFDI, in order to make sure that the cross-sectional consortia complement each other well.

We plan to establish a good balance **between infrastructure facilities and research institutions**. **Inside the consortium**, there are FIZ Karlsruhe, GESIS, TIB, Wikimedia Deutschland, ZB MED, and ZBW as infrastructure providers on the one side, and AWI, BBDC, BZML, FHI, Fraunhofer FIT, Fraunhofer FOKUS, L3S Research Center, Weizenbaum Institute, RWTH Aachen, TU Berlin, Jacobs University Bremen, University of Cologne, Heinrich Heine University Düsseldorf, Leibniz University Hannover, Karlsruhe Institute of Technology, Kiel University, and Leipzig University as research institutions on the other side. **Outside the consortium**, it is planned to integrate further infrastructure facilities, especially from the other consortia of the NFDI. It is also planned to have a lively exchange with other research institutions, especially from the other consortia of the NFDI.

(4) Cross-cutting topics

All topics, the consortium intends to contribute to, are cross-cutting topics that need to be designed and developed by several or all NFDI consortia.

The following non-exhaustive list provides an overview of the planned technical functionality:

- **harvesting service**: Harvests metadata and schemas and registers them. It will fetch the metadata, and transform it into the target data format for metadata governance and interoperability insurance based on harmonized vocabularies and classification schemas.
- **metadata management service:** Includes all features, concerning the storage and management of data and metadata. It offers a query interface, enhanced search functionalities and cross-disciplinary user interfaces.
- **visualization service**: Provides tools to facilitate the creation of problem-specific visualizations and data previews.
- **(long-term) preservation service**: Offers functionality to enable (long-term) preservation of research data.
- **monitoring service**: Provides methods and tools to check the maturity level of the metadata for leveraging data governance. This includes methods and tools for providing services to generate statistics, check the quality of (meta-)data or even services to suggest quality enhancements.
- **licensing service**: Includes tools to assess and consolidate licence and usage terms.
- **provenance & sovereignty service**: Provides integrated tracking and management of provenance and ownership information to enable repeatability of research along with the possibility to assess data quality and origins.
- **terminology & ontology service:** Supports users in the development of semantic models, vocabularies, ontologies, mappings and taxonomies. Enables hosting the vocabularies and provides functionality for vocabulary publication, curation and collaboration; fosters the ongoing transition from the practice of using ambiguous words, imprecise phrases, etc. to encode data, towards common formal and semantically enriched languages for knowledge representation.

Furthermore, it is anticipated to enable the following:

- use and propagate **common APIs & existing standards** in close cooperation with existing generic research data management infrastructures (c.f. DataCite, GeRDI, etc.).
- enable multidisciplinary **(meta-)data analysis**, with a special focus on big data analytics (e.g., examine different metadata formats), data mining (e.g., find links between data sources), as well as distributed approaches.
- provide a data basis for **machine learning & AI** by integrating data from different disciplines.

To support user participation and involvement, a number of measures are planned:

- Perform **requirements analysis** of the envisioned shared infrastructure: Perform regular workshops with domain experts from the consortium and other NFDI consortia to assess their needs using an agile approach, to be able to react fast on changing requirements.
- Foster **collaboration** to keep up to date with developments in the research data community: Set up mechanisms (e.g., topic-specific working groups) to foster collaboration following a bottom-up, integrative and non-descriptive approach.

- Reach out to facilitate **awareness** and **community involvement**: Adapt methods from Wikimedia, e.g., integrate online tools for commenting.
- Establish **governance processes** for joining and using the portal: Install an advisory board and have technical staff and data stewards available who accompany the process.
- Collect best practices, and use them for **education and training** of the different stakeholders: Help them to understand how the infrastructure and the underlying technological components can be used, provide online study material and regular online events (e.g., webinars) as well as offline events (e.g., tutorials and hands-on-workshops).
- Accompany the whole process with an **ethics discussion:** Install an interdisciplinary ethics committee that covers this topic.
- Create standards, guidelines and supporting materials and act as a driver for establishing these **methodologies and standards:** We think along the very successful W3C model, i.e., install interest groups with tight monitoring: Schedule one telco per week with tools to write online minutes. Write a report after one year, and if the report is promising, go for a standard.