Bridge4NFDI - Extended Abstract

Formal details

Planned title of the consortium
- Bridging boundaries among national research data infrastructures

Acronym of the planned consortium
- BRIDGE4NFDI

Lead institution or facility
- Fraunhofer FOKUS (Fraunhofer Institute for Open Communication Systems)

Name and work address of a contact person (including email address and institutional affiliation)
- Dr. Sonja Schimmler, Fraunhofer FOKUS, Weizenbaum Institute for the Networked Society
  Kaiserin-Augusta-Allee 31, 10589 Berlin
  sonja.schimmler@fokus.fraunhofer.de

Members of the planned consortium (including institutional affiliation, without address)
- Prof. Dr. Ziaawasch Abedjan, abedjan@tu-berlin.de
  - TU Berlin
- Prof. Dr. Sören Auer, auer@tib.eu
  - Leibniz University of Hannover
  - TIB (Leibniz Information Centre for Science and Technology)
- PD Dr. Carsten Baldauf, baldauf@fhi-berlin.mpg.de
  - Universität Leipzig
  - FHI (Fritz Haber Institute of the Max Planck Society)
- Dr. Oya Beyan, beyan@dbis.rwth-aachen.de
  - RWTH Aachen
  - Fraunhofer FIT (Fraunhofer Institute for Applied Information Technology)
- Prof. Dr. Stefan Dietze, stefan.dietze@hhu.de
  - Heinrich-Heine-University Düsseldorf
  - L3S Research Center
- Prof. Dr. Frank Oliver Glöckner, frank.oliver.gloeckner@awi.de
  - Jacobs University Bremen gGmbH
  - AWI (Alfred-Wegener-Institut, Helmholtz Zentrum für Polar- und Meeresforschung)
- Prof. Dr. Manfred Hauswirth, manfred.hauswirth@fokus.fraunhofer.de
  - TU Berlin
  - Fraunhofer FOKUS (Fraunhofer Institute for Open Communication Systems)
  - Weizenbaum Institute for the Networked Society
- Franziska Heine, franziska.heine@wikimedia.de
  - Wikimedia Deutschland e.V.
- Prof. Dr. Volker Markl, volker.markl@tu-berlin.de
  - TU Berlin
  - BBDC (Berlin Big Data Center)
  - BZML (Berlin Center for Machine Learning)
- Prof. Dr. Matthias Scheffler, scheffler@fhi-berlin.mpg.de
  - FHI (Fritz Haber Institute of the Max Planck Society)
- Dr. Sonja Schimmler, sonja.schimmler@fokus.fraunhofer.de
  - Fraunhofer FOKUS (Fraunhofer Institute for Open Communication Systems)
  - Weizenbaum Institute for the Networked Society
- Prof. Dr. Dietrich Rebholz-Schuhmann, rebholz@zbmed.de
  - University of Cologne
- Prof. Dr. Klaus Tochtermann, k.tochtermann@zbw.eu
  - Kiel University
  - ZBW (Leibniz Information Centre for Economics)

Participants in the NFDI conference (names, institutional affiliation and email address; max. 3 persons)
- Prof. Dr. Sören Auer, Leibniz University of Hannover, TIB
  auer@tib.eu
- Prof. Dr. Manfred Hauswirth, TU Berlin, Fraunhofer FOKUS, Weizenbaum-Institut
  manfred.hauswirth@fokus.fraunhofer.de
- Prof. Dr. Klaus Tochtermann, Kiel University, ZBW
  k.tochtermann@zbw.eu

Research area of the planned consortium (research area according to the DFG classification system [not subject areas]):
- Does not apply

Participating research institutions (without address)
- AWI, FHI, Fraunhofer FIT, Fraunhofer FOKUS, RWTH Aachen, TU Berlin, Jacobs University Bremen, University of Cologne, Heinrich Heine University Düsseldorf, Leibniz University Hannover, Kiel University, Universität Leipzig, Weizenbaum Institute

Participating infrastructure facilities and/or potential information service providers (without address)
- TIB, Wikimedia Deutschland e.V., ZB Med, ZBW

Planned proposal submission date (2019, 2020, 2021)
- 2019 intended, to accompany the other consortia from the very beginning
Overview diagram or organisational chart for the planned consortium

A Horizontal Meta-Infrastructure to Complement the Vertical NFDI Initiatives
Subject-specific and infrastructural focus of the planned consortium

Key questions/objectives of the consortium

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 pure 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, such as vocabulary and metadata standards, reference models, data usage patterns, data protection policies, licencing, and administrative domains, 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. An excellent example for such a breakthrough is the discovery of magnetic resonance imaging by Paul Lauterbur (a chemist) and Peter Mansfield (a physicist). Any independent researcher designing and conducting their own separate experiments within the borders of their disciplines would not have been able to achieve this groundbreaking discovery resulting in a rapid development in medicine.\(^3\) Cross-domain interoperability is the key element for a 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.

BRIDGE4NFDI will connect experts from different disciplines and from many domain communities, who have addressed generic research data management challenges for years. One very successful example of this past collaboration is the foundation of the DOI registration agency for research data, DataCite e.V.\(^4\) Since 2009, the German members (ZBW, ZB MED, TIB, GESIS and SUB Göttingen) provide thousands of DOIs to institutes and universities all across Germany, Europe and even globally every year, thus enabling the persistent referencing of research data sets, which is a necessary requirement for data discovery.

---

\(^1\) [https://www.economist.com/leaders/2012/04/28/bit-rot/](https://www.economist.com/leaders/2012/04/28/bit-rot/)

\(^2\) Multidisciplinary, crossdisciplinary, interdisciplinary, or even transdisciplinary

\(^3\) [https://www.nobelprize.org/prizes/medicine/2003/summary/](https://www.nobelprize.org/prizes/medicine/2003/summary/)

\(^4\) [https://datacite.org/](https://datacite.org/)
**Multidisciplinary scientific knowledge graphs** are the key building blocks for machine-processable knowledge. The FAIR principles focus on enhancing the ability of machines to automatically find, access, interoperate and (re)use data towards the goal of having machine-actionable data, and enable big data analysis, data mining, large-scale machine learning and AI. 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, such as the Microsoft Academic Graph, the Google Knowledge Graph, Wikidata and the Linked Open Data Graph.

**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 infrastructures 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 them, for bridging the self-contained data infrastructures, which will enable **harmonized interfaces for human and machines** with the overall vision of building a **general access point** for accessing 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 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.
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**\(^6\) (Generic Research Data Infrastructure) 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** 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**\(^5\) (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**\(^7\) (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** (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 e.V. is the developer of these very successful crowdsourcing projects.

**Known needs/current status of research data management in the relevant discipline/subject-specific relevance of the planned consortium**

This consortium is meant to be complementary to all other disciplinary proposals, and not in competition to other initiatives. We aim 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.

---

\(^5\) [https://www.gerdi-project.eu/](https://www.gerdi-project.eu/)
\(^6\) [https://fairdi.eu/](https://fairdi.eu/)
\(^7\) [https://www.europeandataportal.eu/](https://www.europeandataportal.eu/)
Our consortium has several **research institutions** on board, which are experts in Semantic Web and Linked Open Data, Big Data analysis, and machine learning and AI. They also provide practical knowledge in building Open Data platforms as well as Research Data infrastructures.

Several members of our consortium are involved in domain-specific consortia (e.g., NFDI4ING, NFDI4EARTH), semi-generic consortia (e.g., NFDI4LIFE for Life Sciences) and other generic consortia (e.g., RSE4NFDI). This serves as the basis for deep exchange among different NFDI consortia, necessary for such a project.

**Summary of the planned research data infrastructure that is specifically intended to address the needs of research users in their respective work processes**

Our **main idea** is to provide a **(meta-)infrastructure for research data**, which provides a variety of **generic services that can be used by the other 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, we would like to enable the following:

- use and propagate **common APIs** and **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 and AI** by integrating data from different disciplines.

**Planned implementation of the FAIR principles and information about any existing policies or guidelines in the relevant discipline**

FAIR principles provide guidance and a roadmap to communities who aim at creating, managing and sharing their research data in a sustainable way. The goal is to accelerate the reuse of data and to leverage data-driven science and innovation towards creating societal and economic benefits. However, FAIR principles do not prescribe concrete solutions and technical implementations. Therefore, each community should decide on the targeted maturity levels for their context, and what methods and technologies should be applied to achieve these maturity levels.

**BRIDGE4NFDI aims at supporting discipline-specific consortia by providing tools that enable the application of state-of-the-art technologies and increase FAIRness.** This can include an introduction into FAIRmetrics\(^8\) and related methods, which are intended to help checking and enhancing the maturity level of the (meta-)data, and thus facilitate the (technical) transformation and application for the setup of a national research data infrastructure to improve FAIRness. The LOD 5 star rating system\(^9\) formulated by Tim Berners-Lee follows a similar approach, and should also serve as a basis. E.g., the planned services based on LOD and machine learning will help to increase findability of datasets across disciplines, the planned terminology & ontology service will provide a better description and (machine) accessibility of datasets, etc. In terms of assistance, ZBW, one of our partners, hosts the BMBF-funded GO FAIR initiative\(^{10}\), whose aim is to implement the FAIR data principles.

**Planned measures for user participation and involvement**

To support user participation and involvement, we plan a number of measures:

- 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.

---

\(^{8}\)http://fairmetrics.org/

\(^{9}\)https://www.w3.org/DesignIssues/LinkedData.html/

\(^{10}\)https://www.go-fair.org/go-fair-initiative/
- 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.

**Existing and intended degree of networking of the planned consortium**

Nationally, this consortium is complementary to all other disciplinary proposals, and not in competition with other initiatives. We want to give technical as well as non-technical support, and we hope that the future consortia (and, on the long run, other existing state-level initiatives) will benefit from the outcomes of our work. **We will stay in close contact with relevant other initiatives to integrate and extend existing solutions, wherever possible.**

It is pertinent to build on experience, which is available **internationally**. With the European Open Science Cloud\(^{11}\) (EOSC), a meta-infrastructure is under development, following a similar approach on the European level. OpenAIRE\(^{12}\) is another European initiative, which is an open ecosystem for scholarly publishing. Furthermore, it is pertinent to learn from initiatives, such as B2FIND\(^{13}\) and DataONE\(^{14}\), which offers a meta search across EUDAT data centers and direct access to earth observational data through a distributed network, respectively.

We plan to establish a good balance between **infrastructure facilities and research institutions**. Inside the consortium, there are TIB, Wikimedia Deutschland e.V., ZB MED and ZBW as infrastructure providers on the one side, and AWI, FHI, Fraunhofer FIT, Fraunhofer FOKUS, BBDC and BZML, RWTH Aachen, TU Berlin, Jacobs University Bremen, University of Cologne, Heinrich Heine University Düsseldorf, Leibniz University Hannover, Kiel University, Universität Leipzig, and Weizenbaum Institute 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. For instance, a collaboration with GESIS is currently being discussed. It is also planned to have a lively exchange with other research institutions, especially from the other consortia of the NFDI.

---

\(^{11}\) https://www.eosc-portal.eu/
\(^{12}\) https://monitor.openaire.eu/
\(^{13}\) https://eudat.eu/services/b2find/
\(^{14}\) https://www.dataone.org/