NFDI4DataScience

National Research Data Infrastructure for Data Science and AI

Spokesperson: Dr. Sonja Schimmler, Fraunhofer FOKUS
sonja.schimmler@fokus.fraunhofer.de

The importance of research data in computer 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 NLP, machine learning or information retrieval. Data includes unstructured corpora, labelled benchmark and ground truth datasets as well as structured knowledge graphs. On the other hand, next to source code and software libraries, pretrained models have become a ubiquitous ingredient of computer science, where transparency about provenance, underlying data sources and model architectures have emerged as crucial challenges.

NFDI4DataScience aims to establish a research data infrastructure for research communities with a strong focus on *computer and data science*, or on *information science*. Thereby, we will focus on several types of data and artifacts, specifically:

- Scientific articles
- Scientific knowledge graphs (and other knowledge representations)
- Research data
    - Structured task and benchmark definitions
    - Benchmark and evaluation datasets, including raw data
    - Training data for supervised machine learning and language models
    - Bibliographic metadata
- Software
    - Source code and libraries
    - Scripts and executable notebooks
    - (Pretrained) machine learning models
    - OS-level virtual containers

As in virtually all other disciplines, research contributions in computer and data science are conveyed via scientific articles. However, articles are often accompanied by structured descriptions of tasks or research problems as well as source code (implementing the particular approach), and benchmark and evaluation datasets.

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 or Facebook. *Scientific knowledge graphs* will create an added value for the computer and data science community in a similar way, and will be interoperable with existing knowledge graphs. In particular, for data science and artificial intelligence, knowledge graphs will play a key role to realize trustworthiness, responsibility, reliability, explainability and transferability.

In recent years, there has been a dramatic growth in terms of breath and depth of knowledge assets for data science, 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.

In NFDI4DataScience, we aim to develop and maintain a *comprehensive research data infrastructure* for systematically managing the complete lifecycle including maintaining, describing, extracting, publishing and reusing relevant artifacts in a coherent, but at the same time distributed and semantically rich manner. In particular, we plan to expand and synergistically integrate the following *services*:

- Establish a research data repository for computer and data science connecting the open data world (DCAT, CKAN, LOD, schema.org, etc.) and the research data world (RDA, DDI, DataCite, etc.)
- Advance and integrate community-specific open access repositories, e.g. CEUR-WS or arXiv
- Expand and integrate community-specific bibliographic metadata services, e.g. dblp
- Establish a repository for semantic descriptions of research contributions based on scientific knowledge graphs (Open Research Knowledge Graph)
- Realize an executable notebook platform to ensure the reproducibility of computer and data science research
- Maintain a source code repository and collaboration infrastructure based on Git

For each of these services, we will realize horizontal aspects, such as *persistent identification*, *provenance and sovereignty information tracking*, *licensing, 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.

To 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 an *ethics discussion*; and create standards, guidelines and supporting materials and act as a driver for establishing these *methodologies and standards*.

There exist several *success stories of the NFDI4DataScience partners*, which demonstrate the large amount of experience we have on board in data management: *Wikidata*[1] (in combination with *Wikibase*) is an open knowledge base for humans and machines. It is a

[1] https://www.wikidata.org/

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. *GeRDI*[2] is a research data infrastructure to store, share and re-use research data across disciplines with an emphasis on small amounts of data. Kiel University developed its architecture, whereas ZBW focused on the metadata and operational aspects of it. The *European Data Portal*[3] (EDP) is a central access point for metadata of heterogeneous open data published by public authorities in Europe. 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. *DataCite*[4] e.V. is the leading global non-profit organisation that provides persistent identifiers (DOIs) for research data since 2009. GESIS, TIB, ZBW, and ZB MED are members. *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.

Most members of NFDI4DataScience are also involved in other NFDI consortia. This is a good starting point for deep exchange with other relevant stakeholders. NFDI4DataScience will build on experience, which is available nationally as well as internationally. To do so, we will stay in close contact with relevant other initiatives, and integrate and extend existing solutions, wherever possible.

NFDI4DataScience has been actively involved in writing the "Berlin-Leipzig Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung"[6] (as Bridge4NFDI). We think that the paper is a good starting point for discussing cross-cutting topics that are relevant for the NFDI initiative as a whole.

NFDI4DataScience will contribute to cross-cutting topics by increasing the useability of *open research knowledge graphs*. Furthermore, as data science is of growing importance not only in computer science, other disciplines may benefit from the experiences gained by our consortium, and may reuse and adapt the research infrastructure and services developed within our consortium.

By participating in the 2nd NFDI conference, we would like to get in touch with other consortia and to discuss our goals. We further would like to address cross-cutting topics and see where synergies might arise.

[2] https://www.gerdi-project.eu/
[3] https://www.europeandataportal.eu/
[4] https://datacite.org/
[5] https://fairdi.eu/
[6] http://doi.org/10.5281/zenodo.3895208

Members of the planned consortium

Prof. Dr. Ziawasch 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 GESIS (Leibniz Institute for the Social Sciences)
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. Harald Sack harald.sack@fiz-karlsruhe.de	KIT (Karlsruher Institut für Technologie) FIZ Karlsruhe, Leibniz-Institut für Informationsinfrastruktur GmbH
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 ZB Med (Information Centre for Life Sciences)
Prof. Dr. Klaus Tochtermann k.tochtermann@zbw.eu	Kiel University ZBW (Leibniz Information Centre for Economics)