NFDI4DS – NFDI for Data Science and Artificial Intelligence

Sonja Schimmler

Fraunhofer FOKUS

NFDI4DataScience (NFDI4DS) supports researchers along all stages of the research data lifecycle to conduct their research in line with the FAIR principles. An infrastructure is developed targeting researchers from a wide range of disciplines working in data science and artificial intelligence.

By regularly conducting interviews and surveys, NFDI4DS identifies the needs and challenges of researchers from various disciplines regarding data science and artificial intelligence, keeping ethical, legal, and social aspects in mind. Those identified needs and challenges are continuously addressed by picking up existing services, developing new ones and integrating them into the NFDI4DS infrastructure. By systematically adding digital objects (articles, data, models, workflows, scripts/code, etc.) to the NFDI4DS research knowledge graph within the infrastructure, transparency, reproducibility, and fairness are steadily improved. Support structures, including interactive learning materials and community events, accompany the whole process.

This short paper gives an overview of NFDI4DS and its work programme. It provides details about its approach to address the current challenges. It also gives an overview of the services planned, and how they are meant to interact.

1 Introduction

The past years have seen a paradigm shift, with computational methods increasingly relying on data-driven and often deep learning-based approaches, leading to the establishment of data science as a discipline driven by advances in the field of computer science. Transparency, reproducibility and fairness have become crucial challenges for data science (DS) and artificial intelligence (AI) due to the complexity of contemporary DS methods, often relying on a combination of scripts/code, workflows, models, and data.

The *vision* of NFDI4DS is to support all steps of the complex and interdisciplinary research data lifecycle, including collecting/creating, processing, analyzing, preserving, accessing, and reusing resources in DS and AI. The *overarching objective* of NFDI4DS is the

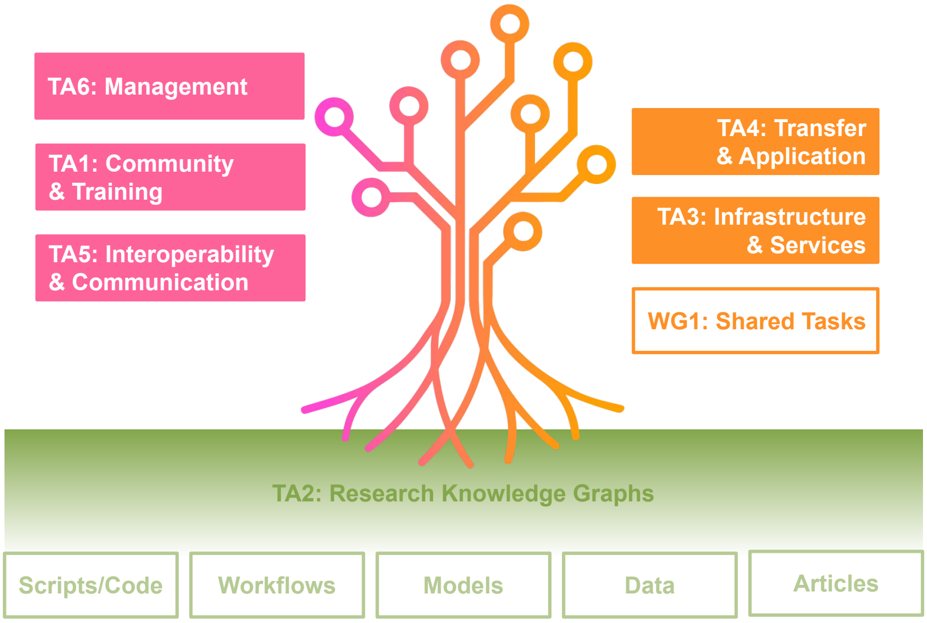
Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18079 (CC BY-SA 4.0)

development, establishment, and sustainment of a national research data infrastructure (NFDI) for the DS and AI community in Germany. This will also deliver benefits for a wider community requiring data analytics solutions, within the NFDI and beyond. The *key idea* is to work towards increasing the transparency, reproducibility and fairness of DS and AI projects, by making all digital objects available, interlinking them, and offering innovative tools and services.

2 Challenges and Approach

Sharing scientific knowledge is not just about publishing articles. Instead, it involves documenting the entire research data lifecycle and providing a multitude of digital objects in compliance with the FAIR principles by making them findable, accessible, interoperable and reusable. As DS and AI are continuously evolving, the methods used become more complex, and it is difficult to maintain transparency, reproducibility, and fairness in research. Challenges related to ethical, legal, or social aspects further limit the willingness and/or ability of researchers to conduct, archive, or publish their research in line with the FAIR principles.

NFDI4DS[1] is part of the NFDI initiative to build a German National Research Data Infrastructure. It supports all stages of the complex and interdisciplinary research data lifecycle to enable the efficient and effective reuse of research data and other digital objects. Additionally, the consortium steadily contributes to establishing best practices in research, fostering open science to enable researchers to make full use of valuable resources.



TA6: Management
TA1: Community & Training
TA5: Interoperability & Communication
TA4: Transfer & Application
TA3: Infrastructure & Services
WG1: Shared Tasks
TA2: Research Knowledge Graphs
Scripts/Code
Workflows
Models
Data
Articles

Figure 1: NFDI4DS Task Areas.

[1] https://www.nfdi4datascience.de

NFDI4DS is organized around six task areas (see Figure 1): (1) Community and Training, (2) Research Knowledge Graphs, (3) Infrastructure and Services, (4) Transfer and Application, (5) Interoperability and Cooperation, and (6) Management. In addition, working groups are temporarily set up on further important topics such as (1) Shared Tasks.

NFDI4DS intends to represent the DS and AI community in academia, which is an interdisciplinary field rooted in computer science. The consortium currently focuses on four DS intense application areas: (1) language technology and natural language processing, (2) biomedical research and clinical decision-making, (3) information sciences and (4) social sciences. Further application areas are involved via speedboat projects later on.

By regularly conducting interviews, insights are gathered on the needs and challenges of researchers, especially about ethical, legal, and social aspects. Systematically conducted surveys identify gaps for new implementations, as well as tools and services that already exist and are useful for the NFDI4DS infrastructure. As DS and AI are important in many disciplines, with often contradicting requirements, building an infrastructure is a collaborative effort involving the scientific communities.

Support structures are accompanying the whole process, addressing the identified needs and challenges. Interactive training materials such as educational videos are provided, and community events such as challenges are organized.

By regularly providing benchmark datasets and fostering joint work on shared tasks interdisciplinary as well as domain-specific services and solutions are achieved. Each shared task focuses on a specific aspect of the research data lifecycle and has the goal to initiate a concrete service that is being integrated into the NFDI4DS infrastructure later on.

3 Core Services

The core services focus on six main areas around digital objects: collecting/creating, processing, analysing, preserving, accessing, and reusing (see Figure 2). Digital objects include artefacts beyond articles, such as data, models, workflows, and scripts/code.

The NFDI4DS infrastructure is based on a number of already existing software components and already well-established tools and services, which target different phases of the research data lifecycle. There are also new technologies, tools, and services uncovered regularly which will continuously be integrated in the NFDI4DS infrastructure. Our service integration strategy is inspired by the EOSC Interoperability Framework, which is based on: (1) persistent identification using PIDs, such as DOI, ORCID or authoritative URIs, (2) authentication and authorisation (AAI) adhering to common standards, (3) semantic interoperability using RDF, vocabularies and ontologies, and (4) API integration based on REST principles.

The NFDI4DS research knowledge graph forms the basis of the infrastructure, providing details about digital objects and their interrelation. Key elements of the infrastructure are the NFDI4DS gateway and portal as well as the NFDI4DS registries and repositories.

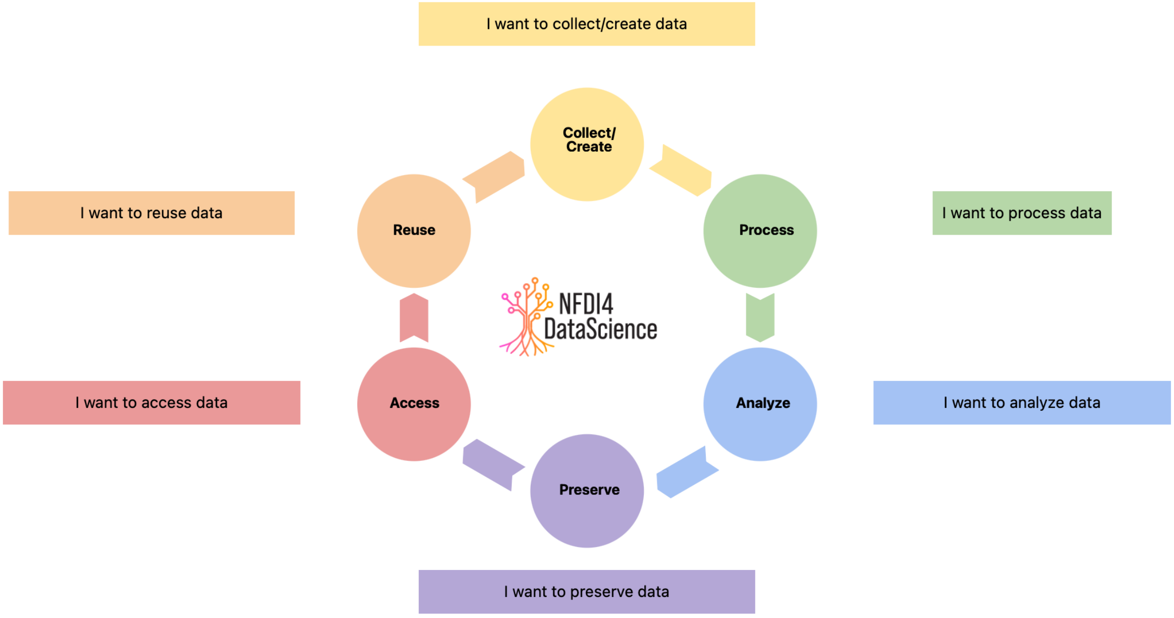


Figure 2: Core Services Dashboard.

Digital objects are harmonized, aggregated, and preserved via the repositories and exposed via the registries and the gateway and portal.

NFDI4DS Research Knowledge Graph. The research knowledge graph will entail automatically extracted metadata about resource relations. The component makes use of *the Open Research Knowledge Graph (ORKG)*, a service for semantically describing research contributions in a knowledge graph. The semantic descriptions of articles are crowd-sourced from authors and researchers leveraging NLP of articles. The component also utilizes *the GESIS Knowledge Graph Infrastructure*, which consists of tools and pipelines for constructing actual research knowledge graphs of research information, metadata and primary research data.

NFDI4DS Registries and Repositories. The consortium aims at providing registries and repositories for different digital objects. One registry being integrated is *the DBLP Computer Science Bibliography*, an open bibliographic data base, search engine, and knowledge graph on computer science publications.

NFDI4DS Gateway and Portal. Through a unified and intuitive search interface, users are enabled to query a wide range of scientific databases such as DBLP, Zenodo, and OpenAlex. While the gateway queries APIs in an ad-hoc fashion, the portal provides a harvesting-based service. The component makes use of *the Data Management Platform Piveau*, which provides services and pipelines for harvesting data and metadata from various sources saving it into a knowledge graph utilizing semantic linking.

NFDI4DS Tools and Services. The consortium aims at integrating different tools and services, including a JupyterHub instance. To facilitate the further analysis and visualisation of digital objects contributed by the participating services, they can be directly loaded into a JupyterHub instance for further programmatic processing. The component utilizes *GESIS Notebooks*, an online reproducibility service for FAIR digital objects. Its main components are a BinderHub, a JupyterHub, and a place to publish, explore, try out, and learn about DS and AI methods.

NFDI4DS Compute Infrastructure. The consortium aims at providing a compute infrastructure. While most of the NFDI4DS tools and services will be hosted in a cloud infrastructure located at various sites of the partners, some DS and AI tasks require access to specialized high-performance computing (HPC) resources such as GPU accelerators and large memory capacities.

4 Conclusions

This short paper gives an overview of NFDI4DS and its work programme. It provides details about its approach to address the current challenges. It also gives an overview of the services planned, and how they are meant to interact.

Acknowledgements

This joint project received funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number: NFDI4DataScience (460234259). We thank all project partners for their contributions.