Editors (Hrsg.): INFORMATIK 2023 – Workshop on Research Data Infrastructures in DS and AI,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023 1

NFDI4DS at a Glance

Sonja Schimmler[1] Christine Hennig[1]

Abstract: The consortium NFDI4DS supports researchers along all stages of the research data lifecycle to conduct their research in line with the FAIR principles. By conducting interviews and surveys, NFDI4DS continuously 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, machine learning models, workflows, scripts/code, etc.) to the NFDI4DS research knowledge graph within the infrastructure, transparency, reproducibility, and fairness are steadily improved. The process is continuously accompanied by providing resources such as educational videos and organizing events such as community challenges.

In this short paper, we give an overview of NFDI4DS, and provide details about our approach to address the current challenges. We will report on how we plan to utilize FAIR digital objects (FDOs) and Research Knowledge Graphs (RKGs) as a basis for the infrastructure envisioned. We will also give an overview of the services planned, and how they are meant to interact.

Keywords: NFDI; NFDI4DS; Data Science; Artificial Intelligence; Research Data Infrastructures

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, models, workflows and data used for training. NFDI4DS will promote fair and open research data infrastructures supporting all involved resources such as scripts/code, workflows, models, data, or articles through an integrated approach.

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

The overarching objective of NFDI4DS is the development, establishment, and sustainment of a national research data infrastructure (NFDI) for the DS and AI community in Germany.

[1] Fraunhofer FOKUS, Berlin, Germany
sonja.schimmler@fokus.fraunhofer.de, christine.hennig@fokus.fraunhofer.de

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 artifacts available, interlinking them, and offering innovative tools and services. Based on the reuse of these digital objects, this enables new and innovative research.

NFDI4DS intends to represent the DS and AI community in academia, which is an interdisciplinary field rooted in computer science. We aim to reuse existing solutions and to collaborate closely with other NFDI consortia and beyond. In the beginning, the consortium will focus on four DS intense application areas: language technology and natural language processing, biomedical research and clinical decision-making, information sciences and social sciences. The expertise available in NFDI4DS ensures that metadata standards are interoperable across domains and that new ways of dealing with digital objects arise.

2 Challenges and Approach to Address Challenges

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 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. These range from humanities and social sciences, life sciences, and natural sciences, to engineering sciences. The research data lifecycle includes six stages: collection/creation of data, its processing, analysis, preservation, access, and reuse. Additionally, NFDI4DS steadily contributes to establishing best practices in research, fostering Open Science to enable researchers to make full use of valuable resources.

NFDI4DS is organized around six task areas: (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 in case they are needed.

To achieve its goals, the consortium is continuously striving for an integrated approach to research data management. The infrastructure is composed of already existing tools and services as well as new ones developed within the scope of this initiative and beyond serving specific purposes to enable all stages of the research data lifecycle. These are embedded in a

framework making the individual tools freely available, and easily usable without extensive prior knowledge.

By regularly conducting interviews, insights are gathered on the needs and challenges, especially about ethical, legal, and social aspects. In addition, yearly surveys identify gaps for new implementations, as well as tools and services that already exist and are useful for the initiative. As DS and AI are important in many disciplines, with often contradicting requirements, building an infrastructure is a collaborative effort including the scientific communities. 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. Besides the development and integration of tools, another important aspect of the consortium are learning materials, best practice guidelines, tutorials, workshops, and hackathons.

The NFDI4DS core services focus on different steps within the research data lifecycle utilizing digital objects. These digital objects include artifacts such as articles, data, machine learning models, workflows, and scripts/code, building upon the emerging concept of FAIR Digital Objects. The NFDI4DS Research Knowledge Graph will form the basis of the infrastructure, providing details about digital objects and their interrelations. Key elements of the infrastructure will be the NFDI4DS gateway and portal as well as the NFDI4DS registries and repositories. Digital objects will be harmonized, aggregated, and preserved via the repositories and exposed via the registries and the gateway and portal. The gateway and portal will offer different functionalities to search and explore the knowledge base as well as provide recommendation and evaluation services e. g. for benchmarking or assessing quality, bias, harm and FAIRness. They will utilize facets allowing browsing datasets, methods, or tools complying with certain community-driven requirements.

3 Conclusion

In this short paper, we gave an overview of NFDI4DS, and provided details about our approach to address the current challenges. We reported on how we plan to utilize FAIR Digital Objects (FDOs) and Research Knowledge Graphs (RKGs) as a basis for the infrastructure envisioned. We also gave an overview of the services planned, and how they are meant to interact.

Acknowledgements

This work has received funding through the German Research Foundation (DFG) project NFDI4DS (no. 460234259).