NFDI4DataScience

NFDI for Data Science and Artificial Intelligence

Progress Report – Part 1

1 General Information

Name of the consortium

NFDI4DataScience (NFDI4DS): NFDI for Data Science and Artificial Intelligence

Research domains or research methods addressed by the consortium

4.43 – Computer Science, (3.31 – Mathematics)

The consortium addresses researchers in the **field** of data science and artificial intelligence. This comprises not only researchers rooted in Computer Science and Mathematics but also in Humanities and Social Sciences, in Life Sciences, in Natural Sciences as well as in Engineering Sciences.

The consortium focuses on data science and artificial intelligence **methods**, where scholarly information processing serves as a broader vision for the whole project. It utilizes knowledge graphs for unifying metadata and enabling trustworthy tools and services. Not only research data is targeted by the consortium, but also articles, data, models and scripts/code.

URL of the consortium website and repositories used for publishing output

https://www.nfdi4datascience.de

2 Summary

Main results and notable success stories

When the NFDI4DS project was planned in 2018, and also soon after its start, the pace of the proliferation of data science (DS) and artificial intelligence (AI) in academia and industry could not have been foreseen. DS and AI have rapidly become popular research fields with groundbreaking disruptions in machine learning and deep learning, fostering fast-pace changes and challenges for science, industry and society, while also creating unprecedented opportunities for start-ups, companies and public organizations.

One area that is still underestimated concerns ethical, legal and social aspects, where NFDI4DS has successfully increased awareness and improved literacy in the DS and AI community. Especially because of the broad scope spanning DS and AI in different fields, the NFDI4DS community is rather diverse compared to other NFDI consortia. Skill sets and prior knowledge vary a lot. A crucial aspect has been therefore to get a thorough understanding of who the community consists of and what their needs are.

To this aim and as part of Task Area 'Community and Training' (TA1) we were able to elicit, gather and analyze requirements in a set of elaborate surveys and interviews via a variety of communication channels and modalities. Based on those, we carefully developed personas and user stories that helped us guide not only our development, but also our community activities. We organized and participated in events on a regular basis (more than 200 events, more than 200 research artifacts) targeting the different stakeholder groups – from more 'traditional' formats like our yearly NFDI4DS Conference and our regular NFDI4DS Lecture Series to more dynamic approaches like our newly established NFDI4DS School, and our regular workshops and hackathons.

Building on the analysis done by TA1, Task Area 'Transfer and Application' (TA4) uses thirteen shared tasks on scholarly information processing, question answering, etc. to strengthen the connections between our consortium and several DS and AI sub-communities including domain-specific ones such as research software in healthcare as well as method-specific ones such as entity recognition for building knowledge graphs. Furthermore, there is a need to deepen the skills among researchers regarding the deployment of FAIR Open Science practices.

The existence of strong infrastructures within the consortium, demonstrated with work carried out in Task Area 'Research Knowledge Graphs' (TA2), enabled linking research artifacts (articles, data, workflows, models, scripts/code) from different partners, also in terms of interoperability (e.g., through the support of cross-KG linking such as the dblp KG and the ORKG KG).

As part of activities of Task Area 'Infrastructure and Services' (TA3), the consortium was able to improve existing services and to come up with new services (more than 40 services in total) such as the NFDI4DS Gateway or the NFDI4DS Portal, which allow researchers to search and explore research artifacts from within NFDI and beyond. Both services are in the prototype and testing phase and will transition into production until the end of the funding period. Joint efforts between TA2 and TA3, guided by the community work in TA1 and TA4, will improve the connectivity across individual but complementary infrastructure and services contributed by different partners, being one of the main tasks until the end of the funding period.

Finally, as far as Task Area ‘Interoperability and Cooperation’ (TA5) is concerned, our project had a rather outstanding and high-profile leading position in terms of presence and participation within NFDI, like formative involvement in the NFDI Sections (e.g., lead of Section ‘Common Infrastructures’ and co-lead of Section ‘Metadata, Terminologies, Provenance’) and in Base4NFDI (e.g., co-lead of Base4NFDI, co-development of several basic services), and also on a national and international level, like community engagement via hackathons (e.g., on metadata for research software, metadata for ML models, and FAIRification support) and workshops (e.g., on knowledge graphs, entity recognition, research management for Open Science), to name a few.

Challenges

Since the proposal of NFDI4DS, **large language models (LLMs)**, such as ChatGPT, have received unprecedented attention and are regarded as game-changers in the field of natural language processing (NLP) and AI, mainly due to their generalizability. However, they are black-box models, which often fall short of capturing and accessing factual knowledge. In contrast, Knowledge Graphs (KGs) are based on structured knowledge models and can explicitly store rich factual knowledge, although these high-quality KGs are typically rather small and fragmented because the data and knowledge they contain requires manual curation.

What we consider a challenge yet to be addressed, concerns a broadening of NFDI4DS’s research endeavors to study and **explore synergies between LLMs and KGs**, with our project providing answers and also coming up with tools and services that demonstrate the potential of unifying LLMs and KGs, customized for the use case of applying such technologies for research purposes.

In promoting Open Science practices, we strive to adopt the FAIR principles for the different artifacts considered in the project. Thus, FAIR and quality measurement represent another important aspect of ultimately ensuring higher reuse of the artifacts in DS and AI. Also, in raising awareness for ethical, legal and social aspects, we solicit a further improvement of these artifacts. The overall goal is to better cover **transparency, reproducibility and fairness of research projects**.

3 Composition of the consortium

Applicant institution

Applicant institution	Location	Duration
Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.	Munich	10/2021 – today

Spokesperson

Spokesperson	Institution, location	Duration
Prof. Dr. Sonja Schimmler	Fraunhofer FOKUS, Berlin	10/2021 – today

Co-applicant institutions

Co-applicant institutions	Location	Duration
DFKI Deutsches Forschungszentrum für Künstliche Intelligenz GmbH	Berlin	10/2021 – today
FIT Fraunhofer FIT	Sankt Augustin	10/2021 – today
FIZ FIZ Karlsruhe	Eggenstein-Leopoldshafen	10/2021 – today
FOKUS Fraunhofer FOKUS	Berlin	10/2021 – today
GESIS GESIS – Leibniz Institute for the Social Sciences	Köln	10/2021 – today
LZI Schloss Dagstuhl Leibniz Center for Informatics	Wadern	10/2021 – today
RWTH RWTH Aachen University	Aachen	10/2021 – today
TIB TIB Leibniz Information Centre for Science and Technology	Hannover	10/2021 – today
TUB Technische Universität Berlin LUH Leibniz University Hannover	Berlin Hannover	10/2021 – today 10/2021 – 12/2023
TUD Technische Universität Dresden	Dresden	10/2021 – today
UK Universität zu Köln	Köln	10/2021 – today
ULei Universität Leipzig	Leipzig	10/2021 – today
LU Leuphana University UH Universität Hamburg	Lüneburg Hamburg	01/2024 – today 10/2021 – 12/2023
ZB Med ZB MED Information Centre for Life Sciences	Köln	10/2021 – today
ZBW ZBW Leibniz Information Centre for Economics	Kiel	10/2021 – today

Co-spokespersons

Co-spokespersons	Institution, location	Task area(s)	Duration
Prof. Dr. Georg Rehm 0000-0002-7800-1893	DFKI, Berlin	TA3, TA4, TA5	10/2021 – today
Dr. Zeyd Boukhers 0000-0001-9778-9164	FIT, Sankt Augustin	TA1, TA3, TA4, TA5	01/2024 – today
Prof. Dr. Franziska Boehm --- Prof. Harald Sack 0000-0001-7069-9804	FIZ, Eggenstein-Leopoldshafen	TA1, TA2	10/2021 – today
Prof. Dr. Stefan Dietze 0009-0001-4364-9243 Prof. Dr. Claudia Wagner 0000-0002-0640-8221	GESIS, Köln	TA2, TA3, TA4	10/2021 – today 10/2021 – today
Prof. Raimund Seidel, Ph.D. 0000-0003-2349-785X Dr. Marcel R. Ackermann 0000-0001-7644-2495 Dr. Michael Wagner 0000-0002-4682-4019	LZI, Wadern	TA2, TA3 TA2 TA3	10/2021 – today 10/2021 – today 10/2021 – today
Dr. Christoph Lange-Bever 0000-0001-9879-3827	RWTH, Aachen	TA2	10/2021 – today
Prof. Dr. Sören Auer 0000-0002-0698-2864 Dr. Markus Stocker 0000-0001-5492-3212	TIB, Hannover	TA1, TA4, TA5 TA1, TA4, TA5	10/2021 – today 10/2021 – today
Prof. Dr. Ziawesch Abedjan 0000-0002-2846-1373 Prof. Dr. Manfred Hauswirth 0000-0002-1839-0372 Prof. Dr. Volker Markl 0009-0009-0964-026X Prof. Dr. Sebastian Möller 0000-0003-3057-0760	TUB, Berlin	TA2, TA3 TA3, TA6 TA3, TA6 TA3, TA6	10/2021 – today 10/2021 – today 10/2021 – today 10/2021 – today
Prof. Dr. Ricardo Usbeck 0000-0002-0191-7211	LU, Lüneburg	TA1, TA2, TA3	10/2021 – today
Prof. Dr. Wolfgang E. Nagel ---	TUD, Dresden	TA3	10/2021 – today
Prof. Oya Beyan 0000-0001-7611-3501	UK, Köln	TA4	10/2021 – today
Prof. Dr. Thomas Neumuth 0000-0001-6999-5024	ULei Leipzig	TA4	10/2021 – today
Prof. Dr. Ricardo Usbeck 0000-0002-0191-7211	UL, Lüneburg	TA3, TA4	10/2021 – today
Prof. Dr. Dietrich Rebholz-Schuhmann 0000-0002-1018-0370	ZB Med, Köln	TA2, TA3	10/2021 – today
Prof. Dr. Klaus Tochtermann 0000-0003-2471-2697	ZBW, Kiel	TA2	10/2021 – today

Participants

Participating institutions	Location	Duration
University of Bremen (Alfred Wegener Institute – Helmholtz Center for Polar- and Marine Research) Prof. Dr. Frank Oliver Glöckner	Bremen	10/2021 – today
Fritz-Haber-Institut der Max-Planck-Gesellschaft Prof. Dr. Matthias Scheffler PD Dr. Carsten Baldauf	Berlin	10/2021 – today
Wikimedia Deutschland e.V. Franziska Heine	Berlin	10/2021 – today