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 Transfer and Application

Ekaterina Borisova[1] Raia Abu Ahmad[1] Georg Rehm[1] Ricardo Usbeck[2] Jennifer D'Souza[3] Markus Stocker[3] Sören Auer[3] Judith Gilsbach[4] Anastasia Wolschewski[5] Johannes Keller[5] Daniel Schneider[5] Thomas Neumuth[5] Sonja Schimmler[6]

Abstract: Due to the ever increasing importance of Data Science and Artificial Intelligence methods for a wide range of scientific disciplines, ensuring *transparency* and *reproducibility* of DS and AI methods and research findings have become essential. The NFDI4DS project promotes the *findability*, *accessibility*, *interoperability*, and *reusability* in DS and AI by developing an open integrated research data infrastructure in which all artefacts (e. g., papers, code, models, datasets) will be interlinked in a FAIR and transparent way. One of the key aspects is to build a bridge between NFDI4DS and other research communities which actively apply DS and AI methods. This paper describes the main actions taken to engage with the relevant (sub)communities.

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

1 Introduction

Most research artefacts nowadays are spread across repositories, digital libraries, and institutional databases. Such unsystematic and decentralised storage complicates the *findability*, *accessibility*, *interoperability*, and *reusability* (FAIR) [WDA16] of scientific data. The German National Research Data Infrastructure (NFDI)[7] initiative aims to interconnect and preserve interdisciplinary scholarly data in a FAIR way. NFDI for Data Science and Artificial Intelligence (NFDI4DS)[8] is one of the 26 NFDI consortia which promotes the idea of transparency, reproducibility and fairness in DS and AI. NFDI4DS's goal is to support all phases of the research data life cycle, including collecting/creating, processing, analysing, publishing, archiving, and reusing resources in DS and AI through innovative tools and services.

[1] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany
ekaterina.borisova@dfki.de, raia.abu_ahmad@dfki.de, georg.rehm@dfki.de
[2] Universität Hamburg, Hamburg, Germany, ricardo.usbeck@uni-hamburg.de
[3] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
jennifer.dsouza@tib.eu, markus.stocker@tib.eu, auer@tib.eu
[4] GESIS Leibniz Institute for the Social Sciences, Cologne, Germany, judith.gilsbach@gesis.org
[5] Innovation Center Computer Assisted Surgery (ICCAS), Leipzig University, Germany
anastasia.wolschewski@iccas.de, johannes.keller@iccas.de, daniel.schneider@iccas.de,
thomas.neumuth@iccas.de
[6] Fraunhofer FOKUS, sonja.schimmler@fokus.fraunhofer.de
[7] https://www.nfdi.de
[8] https://www.nfdi4datascience.de

To achieve this objective, it is important to engage with the relevant communities and to transform NFDI4DS's vision into concrete applications. We collaborate with four scientific fields, i. e., Natural Language Processing (NLP) and Language Technology (LT) including Semantic Web, Biomedical Research, Information Sciences and Social Sciences.

2 Applications in Natural Language Processing

Like the other domains in scope of the project, researchers from the LT, NLP, and Semantic Web communities apply a wide variety of Research Data Management (RDM) methodologies. Although these communities follow similar research avenues, they utilise heterogeneous RDM techniques. This leads to inconsistency in documentation and sharing of digital artifacts (e. g., data, code, models) contributing to low transparency and reusability. Our goal is to bridge the gap between LT, NLP and Semantic Web by hosting different events (e.g., workshops, shared tasks, challenges) and by spreading knowledge about RDM best practices. We currently work on the following activities:

- Three shared tasks[9] which address problems in the area of scholarly information processing, i. e., software mention detection, leaderboard mining [KDA21, KDA23a, KDA23b], and research field classification. By organising these shared tasks we enable and encourage the sharing of RDM techniques, digital artifacts and evaluation measures across LT, NLP and Semantic Web communities.
- The ISWC 2023 Scholarly Question Answering challenge[10] allows researchers to compare their Knowledge Graph Question Answering Systems.
- We also work on compiling Open Science Best Practices in DS and AI, i. e., recommendations for ensuring a FAIR life cycle of digital artifacts. The guidelines will be published on the NFDI4DS website.

3 Applications in Biomedical Research and Clinical Decision Making

In the biomedical domain, Data Science faces obstacles due to 1. the safety-critical and ethically relevant nature of clinical applications, 2. complex heterogeneous and incomplete datasets, and 3. a lack of standardisation and data-privacy regulations. We plan to leverage the NFDI4DS infrastructure to foster data and knowledge transfer between the biomedical research community and DS. This ensures that the requirements of biomedical stakeholders can be adequately met within DS research. Our effort until 2026 includes:

- Host challenges dealing with the requirements of biomedical applications, starting with a competition related to reliable uncertainty estimation in 2024. The competitions will raise awareness for open research questions relevant to biomedical DS, flatten

[9] https://www.nfdi4datascience.de/community/shared-tasks/
[10] https://kgqa.github.io/scholarly-QALD-challenge/

learning curves through tutorials and workshops, present overviews of the state of the art, provide benchmarks, and obtain novel solutions driving AI adoption in medicine.
- Promote the NFDI4DS infrastructure in the biomedical engineering, medical informatics, and DS research communities by giving talks and hosting focus sessions at relevant conferences (e. g., BMT[11], GMDS[12], and within ScaDS.AI[13]). These events will also help gather community feedback and requirements.
- Design tutorials and best practices for use cases from clinical prognosis and decision support using heterogeneous medical datasets. The tutorials will be deployed using the binder notebook service and involve topics including data preparation, synthetic data generation, dealing with missing data, and uncertainty estimation.

4 Application for Information Sciences

The Open Research Knowledge Graph (ORKG)[14] [St23, Au20, Ja19] is an NFDI4DS service for semantically describing research contributions published in scientific articles in a knowledge graph. Research contributions are semantic (i. e., machine-actionable) descriptions of published research findings together with the employed materials and methods and the addressed research problem. Semantic descriptions are crowd-sourced from authors and researchers. The ORKG also leverages NLP information extraction from articles to automate knowledge graph construction. ORKG supports numerous downstream services, including comparisons of research contributions addressing a common research problem, visualisations of compared information, leaderboards, etc. The ORKG is central for the NFDI4DS Application for Information Sciences, which has three goals:

- Deploy the most promising NLP models for information extraction from scholarly articles developed in NFDI4DS Shared Tasks in an infrastructure to further advance automation of scholarly knowledge graph construction and curation.
- Catalyse the adoption of the ORKG and, more generally, Scholarly Knowledge Graph infrastructure, in relevant conference series and journals. First experiments have been conducted in 2022 with ISWC and SEMANTiCS by actively guiding authors on how to include ORKG Comparisons in their submitted papers.
- Develop approaches to ensure expressions of research findings are produced machine actionable, and automatically flow into digital scholarly communication infrastructure such as ORKG. We currently test approaches for model performance evaluation conducted using Python or R computing environments to ensure that TDMS (Task, Dataset, Metric, Score) data published in articles automatically flows into ORKG benchmarks and leaderboards.

[11] https://bmt2023.de
[12] https://www.gmds2023.de
[13] https://scads.ai
[14] https://orkg.org

5 Applications in Social Sciences

Researchers from the Social Sciences handle very heterogeneous types of data, documented and shared in very diverse ways. The same applies to code written for data wrangling and analysis. Our goal is to encourage Social Science researchers to practice FAIR science and provide them with the knowledge and infrastructure. We currently compile the status quo regarding documentation, quality assessment and sharing behaviour of data and code. Subsequently, we will contribute to the implementation of our goals in the Social Sciences with an emphasis on making data and code available to the research community in a reusable way. Until 2026 we will be engaged in the following activities:

- Conduct a survey among researchers, investigating current data and code-sharing practices as well as quality assessment of digital research objects with a focus on relational social network data. For sharing behaviour a similar survey exists in the field of computational biology [CH22].
- Define use cases from an online access web-tracking panel under development at GESIS[15] to define requirements for data quality and sharing practices.
- Host a workshop for social scientists on open science practices, present NFDI4DS services and gather feedback from the Social Science community.

6 Call for Speedboat Projects

In addition to the four scientific areas mentioned in the last paragraphs, we are issuing a call for speed boat projects. These projects are meant to kick off collaboration with further domains, and to complement our overall service portfolio.

7 Conclusion

This paper presents NFDI4DS's main strategies to promote FAIR research data management across four scientific areas: 1. NLP and LT as well as Semantic Web, 2. Biomedical Research and Clinical Decision Making, 3. Information Sciences, and 4. Social Sciences. The current community engagement tools range from conducting various surveys and proposing best practices to organising shared tasks.

Acknowledgements

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

[15] https://www.gesis.org

Bibliography

[Au20] Auer, Sören; Oelen, Allard; Haris, Muhammad; Stocker, Markus; D'Souza, Jennifer; Farfar, Kheir Eddine; Vogt, Lars; Prinz, Manuel; Wiens, Vitalis; Jaradeh, Mohamad Yaser: Improving Access to Scientific Literature with Knowledge Graphs. Bibliothek Forschung und Praxis, 44(3):516–529, 2020.

[CH22] Cadwallader, Lauren; Hrynaszkiewicz, Iain: A Survey of Researchers' Code Sharing and Code Reuse Practices, and Assessment of Interactive Notebook Prototypes. PeerJ, 10:e13933, 2022.

[Ja19] Jaradeh, Mohamad Yaser; Oelen, Allard; Farfar, Kheir Eddine; Prinz, Manuel; D'Souza, Jennifer; Kismihók, Gábor; Stocker, Markus; Auer, Sören: Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. K-CAP '19, Association for Computing Machinery, New York, NY, USA, pp. 243–246, 2019.

[KDA21] Kabongo, Salomon; D'Souza, Jennifer; Auer, Sören: Automated Mining of Leaderboards for Empirical AI Research. In: Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23. Springer, pp. 453–470, 2021.

[KDA23a] Kabongo, Salomon; D'Souza, Jennifer; Auer, Sören: ORKG-Leaderboards: A Systematic Workflow for Mining Leaderboards as a Knowledge Graph. arXiv preprint arXiv:2305.11068, 2023.

[KDA23b] Kabongo, Salomon; D'Souza, Jennifer; Auer, Sören: Zero-shot Entailment of Leaderboards for Empirical AI Research. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2023. 2023.

[St23] Stocker, Markus; Oelen, Allard; Jaradeh, Mohamad Yaser; Haris, Muhammad; Oghli, Omar Arab; Heidari, Golsa; Hussein, Hassan; Lorenz, Anna-Lena; Kabenamualu, Salomon; Farfar, Kheir Eddine; Prinz, Manuel; Karras, Oliver; D'Souza, Jennifer; Vogt, Lars; Auer, Sören: FAIR Scientific Information with the Open Research Knowledge Graph. FAIR Connect, 1(1):19–21, Jan 2023.

[WDA16] Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, et al.: The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data, 3(160018), 3 2016.