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

Community and Training in NFDI4DS

Anna-Lena Lorenz[1], Maria Christoforaki[2], Christine Hennig[3], Angelie Kraft[4], Stephanie von Maltzan[5], Sonja Schimmler[3]

Abstract: Key to NFDI4DS's success is an active and vibrant community as establishing a common data culture and practice relies on the community's participation and acceptance. We address this challenge by leveraging the network of NFDI4DS partners to raise awareness for topics around FAIR data and establish international standards. By identifying requirements, we improve our services and develop new strategies for building and finding user communities.

Keywords: NFDI; NFDI4DS; Community; Training; Requirements Engineering; ELSA

1 Introduction

Due to the broad scope spanning Data Science (DS) and Artificial Intelligence (AI) in different fields, the NFDI4DS community is rather diverse compared to other NFDI consortia. Skillsets and prior knowledge vary a lot, even between people of the same career level. A crucial aspect is thus to get a thorough understanding of who the community consists of and what their needs are. Therefore we continuously elicit, gather and analyze requirements in interviews and surveys via our various communication channels and in-person events.

2 Outreach

We regularly host events targeting the different stakeholder groups - from traditional Lecture Series to more dynamic approaches like our Science Slam that already took part twice in the context of the Berlin Science Week. This event was open to the general public and invited perspectives from all NFDI consortia.

To connect to our community online, we also run several communication channels. We started with a Twitter account[1] and are now also on Mastodon[2], as part of our community moved there due to the recent developments. We can also be found on LinkedIn[3] and for

[1] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany, anna.lorenz@tib.eu
[2] University Hospital of Cologne, Germany, maria.christoforaki@uk-koeln.de
[3] Fraunhofer FOKUS, Berlin, Germany
sonja.schimmler@fokus.fraunhofer.de, christine.hennig@fokus.fraunhofer.de
[4] Universität Hamburg, Germany, angelie.kraft@uni-hamburg.de
[5] FIZ Karlsruhe, Germany, stephanie.maltzan@fiz-karlsruhe.de
[1] https://twitter.com/nfid4ds
[2] https://nfdi.social/@nfdi4ds
[3] https://www.linkedin.com/company/nfdi4ds

hosting our lecture series and training videos, we have a YouTube channel[4]. We are also on Zenodo[5], GitLab[6] and GitHub[7], since our community is active there.

While building a community is a crucial first step, training researchers and practitioners is equally important. We thus also develop interactive training material facilitating well established platforms and services within the community and curricula. We are currently in the process of creating a video series to introduce our consortium and our services. Not only do we develop material, but we also organize and participate in community events on different topics regarding data management, data ethics and data protection.

All of our created material is given a common branding. Since our beginning we have a clear corporate design consisting of an orange-pink color scheme and a stylized tree with nodes as a logo. This design emphasises the modernity of the topic via its fresh, young colors and symbolizes the interconnection between services, data and communities which is crucial for our work. Since our material is as broadly diverse as our communities, this is a unique way of creating a uniform framework, while allowing for formats, services and training material to be as manifold and specific as necessary. And of course, the design can not only be found in our digital material, but also on our merchandise that we hand out at in-person events to spread the word. Currently we have stickers, pencils, blocks, sticky notes, and t-shirts. Our slogan is "We are the missing link", as we believe that NFDI4DS truly provides a unique unified approach for all research data of our community.

3 Personas

To not think about our community as an abstract concept, but of a heterogeneous group of people with different skill sets and requirements, we developed personas - fictional individuals - with their distinct wants and needs. The first input for those personas was developed during our first in-person consortium meeting with several people from different scientific backgrounds and career levels present. In this first session, we introduced three fictional people. A computer science PhD student, a post-doc from biomedical research and a professor in social sciences. The participants were asked to take the perspective of these people based on their own experiences in the respective fields or career stages. We then collected answers to three broad questions: (1) What are this person's tasks? (2) What are their problems and obstacles in executing these tasks? (3) What do they wish for? What could help them with the aforementioned problems?

As a first step, we encouraged the participants to not only answer these questions with specific services in mind but rather generally. After this brainstorming session, we clustered

[4] https://www.youtube.com/@nfdi4ds
[5] https://zenodo.org/communities/nfdi4ds
[6] https://gitlab.com/nfdi4ds
[7] https://github.com/nfdi4ds

the answers into different categories. We excluded the points that can not be addressed by our consortium, such as the wish for more money, and identified the main issues.

In total we determined six personas for our core stakeholder groups from these inputs. At our next meeting we then presented the personas and aligned them with our services to ensure that we can offer added value to these fictional people and also for our services to keep in mind. During that process, we identified missing information, unclear descriptions and redundancies, which then left us with our refined version of the personas:

- **Alex** is a PhD student in computer science. He spends a lot of time searching for and cleaning datasets. For some of his tasks he wants to re-use sensitive data from another group. Sometimes he struggles to access this data and often he wishes there would be a system to help him find, access and semi-automatically clean data. Especially the last point would reduce his workload significantly. Moreover, better search options supporting different content types would help him.
- **Ben** does his PhD in biomedicine and searches for papers and methods for his task. He often gets lost in the flood of publications and does not have time to read all papers. Additionally, often methods and data have incomplete descriptions. He wishes for a better documentation as well as an overview of the current state of his research area.
- **Cassie** just started her PhD in social sciences and thus still has a steep learning curve. She wishes for better training material tailored to her needs.
- **David** is in the early postdoc phase of his career in life sciences. He wants to compare his results with some experimental data in the field, but often is missing standards for sharing data. He also wishes for more computational time and power.
- **Emma** plans to run a journal and thus has to deal with a lot of editorial tasks. She needs to find a publisher, acquire suitable reviewers and keep up with the latest community news in terms of topics, papers and people. The ongoing reviewer fatigue does not make her situation easier. A system that informs her about current developments and also suggests possible reviewers would be useful for her.
- **Finn** is a professor in social sciences and responsible for his group's data management plans. In his proposals he wants to fulfill the expectations, but sometimes doing open science and protecting sensitive data seem mutually exclusive. He needs clearer guidelines and a safe and secure storage for his group's sensitive data that should still be linked to executable code.

For these fictional people we develop pathways of how they can navigate our website and services. We keep these people's needs in mind when designing and optimizing services. Moreover, these personas are not only used for development. We also create marketing material such as videos tailored to their use cases. We have a set of secondary stakeholders as well, including e.g. developers who use our services as a back-end to their own product.

4 Ethical, Legal and Social Aspects

When working with data, it is crucial to keep ethical, legal and social aspects (ELSA) in mind. Especially when scraping data or training models on sensitive or personal data, e.g. patients' medical diagnoses or survey answers, researchers need to make sure to comply with data protection laws and copyright restrictions. Furthermore, other European regulations, particularly the forthcoming AI Act, must be taken into account. Nevertheless, open science and the use of personal data do not mutually exclude each other. We want to support open data by giving out guidelines and developing curricula to train researchers on these issues. A key step is to understand the legal and ethical challenges our community faces. Therefore, we started conducting interviews with researchers at all career levels and from all our disciplines. Based on the results of the interviews we will be able to make recommendations on how to address these challenges.

Storing and accessing personal data is a challenge faced by many NFDI consortia, but in the domain of AI, another topic arises: In times of generative AI like ChatGPT and StableDiffusion researchers have not only to think about copyright or privacy issues, but also about the societal implications of their work. Choices and jobs that have traditionally been done by people are gradually assigned to algorithms. But as these algorithms are only as fair as their training data, there is a danger of teaching the model an implicit bias and increasing the risk of discriminating people treated by the algorithms. NFDI4DS aims to identify and address these issues. For that we regularly host events such as the Weizenbaum Forum on 'Chat GPT, Stable Diffusion and Co', which recently took place.

In addition, we are currently producing a video series about important concepts and questions regarding AI ethics. This series will be a knowledge resource for practitioners in DS- and AI-related fields and will facilitate critical reflections of our present and future with AI. The videos are comprised of interviews with renown experts in the fields of ethics, sociology, computer science, design, and law. A wide range of topics are covered, such as the EU AI Act, trustworthiness, explainability, and algorithmic bias.

5 Conclusion

NFDI4DS builds and trains a community of DS and AI researchers and practitioners to establish a common culture around FAIR data, data ethics and data management. The focus of our outreach measures is on our services. By providing easy entry points and showing pathways, we simplify data management as much as possible. To train researchers, we develop curricula and participate in community events regularly.

Acknowledgements

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