

Beyond Silos: Next Steps in Research Data Management Report from the 2024 Spring Symposium of FG Datenbanken

Birgitta König-Ries^{1*}, Ulf Leser², Bernhard Seeger³,
Sirko Schindler⁴

^{1*}Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Leutragraben 1, 07743 Jena, Germany.

²Department for Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, Berlin, 10099, Germany.

³Department of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str. 6, Marburg, 35032, Germany.

⁴Institute of Data Science, German Aerospace Center (DLR), Mälzerstraße 3-5, Jena, 07745, Germany.

*Corresponding author(s). E-mail(s): birgitta.koenig-ries@uni-jena.de;

Contributing authors: leser@informatik.hu-berlin.de;

seeger@mathematik.uni-marburg.de; sirko.schindler@dlr.de;

Abstract

On March 11 and 12, 2024, the Spring Symposium of the GI Fachgruppe Datenbanken took place in Jena. The overarching theme of the meeting was research data management beyond isolated repositories, and it explicitly aimed to foster connections between the FG Datenbanken and the NFDI community. Talks and posters focused on incentives and hurdles for FAIR data management, interoperability and reproducibility, automated Data Engineering, and AI-based user support.

Keywords: research data management, FAIR, distributed knowledge management



Fig. 1 The opening session of the Spring Symposium.

1 Introduction

The Spring Symposium 2024 of the Fachgruppe Database Systems (FG DB) "Beyond Silos: Next Steps in Research Data Management" was dedicated to research in FAIR research data management beyond isolated repositories and how the database community can contribute to achieve this.

This topic attracted more than ninety registrations. Even though – due to a train strike at the days of the symposium – not all of those registered could make it to Jena, the meeting was intensive and lively, characterized by active participation and intensive discussions. We thus consider it as successful in achieving its goals, i.e., fostering networking and scientific exchange.

The Symposium was structured into five plenary sessions and a poster session. Speakers and posters, in part, were specifically selected to bring together representatives of relevant NFDI consortia and the database community. In the sessions, but also in many conversations over coffee or during the social event in the evening, participants investigated different aspects of our overarching theme:

- Incentives (and hurdles) for FAIR research data management: What prevents the implementation of FAIR principles? How can the provision of FAIR data be incentivized?
- Interoperability and Reproducibility: The "I", i.e., interoperability, is considered one of the most complex aspects of FAIR, as it requires homogenization of formats, APIs, and semantics across different organizations and communities. Several contributions

presented platforms that support large-scale data integration and interoperability of tools. A second focus was reproducibility in digital experiments and how computational tools can support it.

- Automated Data Engineering and AI-based user support: Large manual efforts, be it for data quality assurance, meta data provisioning, or data integration, are major obstacles on the way to FAIR data. Contributions presented modern AI approaches promising (partial) automation of some of these time-consuming and cumbersome tasks. Other tools aim to improve human-computer interfaces to lower the barrier and effort to access and analyze data.

2 Invited Talks

In five sessions, a total of eleven invited talks were presented.

The first session was dedicated to recent developments in changing the way research outputs are assessed, aiming to put a stronger weight on qualitative aspects beyond pure publication counts, e.g., datasets, innovation, or software. Marcel Meistring presented the preliminary outcomes of the Helmholtz Task Group on Quality Indicators for Data and Software Publications, while Miriam Kip introduced the infrastructure for improving research quality established at the QUEST center at the university clinic Charité in Berlin.

The second session started with two presentations related to the Semantic Web. Felix Engel discussed the potential of using standard terminologies in databases to support interoperability and introduced a terminology service developed for NFDI4Ing. Next, Judith Wodke presented work on building knowledge graphs for Biomedicine. In the final presentation, Tanja Auge proposed a solution to support schema evolution in research data.

In the third session, Björn Grüning provided an overview of the Galaxy workflow platform and how it every day supports several thousands of non-expert users in their needs for scientific data analysis. In the last talk of the day, Dirk Junghanns took the audience on a tour of Snowflake’s main areas of work and gave information on possibilities for academia to cooperate with this company.

The second day started with a talk by Bernhard Seeger on challenges in biodiversity data infrastructures and the approaches in NFDI4Biodiversity to overcome them.

After the poster session, Fabian Panse kicked off the final session of the symposium with a presentation on the creation of synthetic tabular data (and why one may want to do so). Then, two talks followed focusing on reproducibility. Sheeba Samuel presented a comprehensive (and rather sobering) evaluation of the reproducibility of Jupyter Notebooks. Maik Fröbe introduced the TIRA platform, an infrastructure for shared tasks in inter-organizational settings, and how it contributes to reproducibility.

3 Posters

Eleven posters covering a wide range of topics related to the symposium theme were presented in total and sparked interesting conversations (see [Figure 2](#) for an impression). Topics ranged from specific database systems (Snowflake and Ocident), approaches to data integration (combining data lakes and RDBMS), legal aspects of

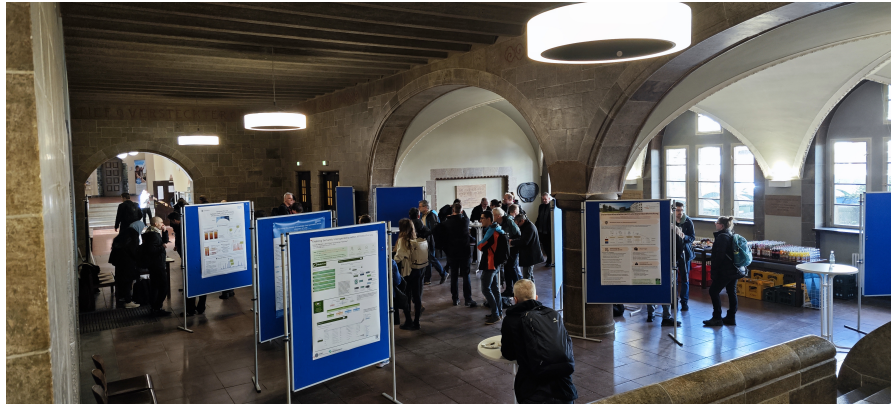


Fig. 2 Discussions at the poster session.

data management and data management for legal topics, provision and creation of semantic resources, and aspects of FAIR data (FAIR assessment, FAIR repositories, reproducibility, custom databases) to efficient data analysis in earth observation.

4 Social Event

Thanks to generous support by Snowflake, a symposium dinner could be organized in the evening. This offered plenty of opportunities for networking in a relaxed setting that invited to mingle. [Figure 3](#) offers a glimpse into the event.

5 Conclusion

Of course, this short article could only provide a small glimpse into the event. More details including abstracts, slides, posters, material from the general assembly of the Fachgruppe Datenbanken included in the symposium, and some photos can be found on the homepage of the symposium at <https://indico.rz.uni-jena.de/event/100/>. To experience networking of the database community first hand, we strongly recommend attending LWDA 2024 and BTW 2025. See you there!

Acknowledgements. The event was made possible by generous support from Snowflake, the NFDI4Biodiversity Project (DFG project number 442032008), the Gesellschaft für Informatik and Friedrich Schiller University Jena.



Fig. 3 Glimpse into the social event.