Data Management in INF Projects of Collaborative Research Centres: Building Bridges Between Research, Infrastructure and Practice

Oliver Koepler [1], Iryna Mozgova [2], Florian Nürnberger [3], Christoph Steinbeck [4], Jürgen Pleiss [5]

[1] Lab Linked Scientific Knowledge, Leibniz Informationszentrum Technik und Naturwissenschaften;
[2] Datenmanagement im Maschinenbau, Universität Paderborn;
[3] Institut für Werkstoffkunde, Leibniz Universität Hannover;
[4] Institut für Anorganische und Analytische Chemie, Friedrich-Schiller-Universität Jena;
[5] Institut für Biochemie, Universität Stuttgart

Collaborative Research Centers (CRCs), funded by the German Research Foundation (DFG), consist of a large number of research sub projects working collaboratively across research institutions, addressing complex research problems. To enhance interdisciplinary data exchange and manage research data effectively, the DFG introduced Information Infrastructure (INF) projects within CRCs. This paper presents experiences from five exemplary CRCs, including CRC 1153, CRC 1368, TRR 375, CRC 1333, and CRC 1127, spanning disciplines such as engineering sciences, materials science, chemistry, biology, and biochemistry. We analyze the role of INF projects through six key questions addressing discipline-specific challenges, collaboration with RDM local, regional, or national players, transfer of innovation into scientific communities, training, integration of AI methodologies, and sustainability of research data management (RDM) infrastructures. Common challenges identified include the integration of heterogeneous data environments, implementation of standardized metadata schemas, and ensuring interoperability and reusability of research data. Successful approaches include tailored training programs, the adoption of open-source tools, and the creation and utilization of domain-specific metadata vocabularies. Innovations from INF projects such as semantic annotation, knowledge graph generation, and AI-based data curation demonstrate the practical value of professional data management. Sustainable archiving strategies leverage local and national

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23912 (CC BY-SA 4.0).

repositories, automated metadata workflows, and open-access data publishing practices. The collaboration with local and national RDM infrastructures, including interactions with consortia of the German National Research Data Infrastructure (NFDI), significantly enhances these efforts. This article highlights effective strategies and best practices developed in INF projects, providing valuable insights into managing research data sustainably and effectively across interdisciplinary collaborations, and offering guidance to future INF initiatives within CRCs.

Keywords: Research Data Management, Collaborative Research

1 Introduction

Collaborative Research Centers (Sonderforschungsbereiche, SFBs / CRCs), funded DFG, are long-term research consortia, established for up to twelve years, in which researchers work together on interdisciplinary research questions across institutes of a university. Transregional Collaborative Research Centers (TRR), which work on interdisciplinary research problems across two or more university sites, are a special variation. CRCs usually consist of up to 20 research sub projects which contribute to an overarching research problem from various perspectives. This strong collaboration and networking inevitably results in the need to exchange data and information across projects. Motivated by these needs and the process of professionalising research data management, new methodologies, services and infrastructures, the INF projects have therefore been introduced in 2007 by the DFG funding program to support the application of Information Management and Information Infrastructure (INF) projects for CRCs and to implement sustainable and professional RDM within the CRCs. The general objectives of INF projects include the development and provision of databases or repositories for research data, the curation of data, and the development and utilization of discipline-specific metadata schema. Furthermore, the development of innovative data services for the seamless integration of research data management in research environments.

In this paper the authors report about their experience in INF projects of five exemplary CRCs:

- CRC 1127 ChemBioSys: Chemical mediators in complex biosystems, Friedrich-Schiller University, Jena
- CRC 1153: Process chain for manufacturing hybrid high-performance components through Tailored Forming, Leibniz University Hannover
- CRC 1333: Molecular heterogeneous catalysis in confined geometries, University of Stuttgart
- CRC 1368: Oxygen-free production, Leibniz University Hannover
- TRR 375 HyPo: Multifunctional high-performance components made from hybrid porous materials, Leibniz University Hannover and RPTU Kaiserlautern-Landau

The CRCs are conducting interdisciplinary research covering engineering science, material science, chemistry, catalysis, biology, and biochemistry. Out of the CRCs examined, the CRC 1127 and CRC 1153 are today in the 3rd funding period, starting 2014 and 2015. Together with the CRC 1333, starting 2018, these three research centers introduced an INF project with the beginning of the 2nd funding period. CRC 1368 and TRR 375 have started in 2020 and 2024 respectively, both with an INF project from the very beginning. We will use six key questions to describe the challenges, goals and tasks of the individual INF projects.

The paper provides answers and insights, solutions, and successful models of INF projects, highlighting how they support progress in CRCs and contribute to the advancement of RDM infrastructures. The goal is to identify common challenges and solutions and best practices to be of value for other INF and CRC projects.

2 Role of Information Infrastructure Projects in CRC

Introduced in 2007 into the CRC funding program, the first two INF projects started their work in 2009. Today 30 % of the CRC are supported by INF projects. In the last consultation of the DFG Senate Committee on Collaborative Research Centres nearly 50 % of the CRC proposals included an INF project. Over the years the developments vary in the individual scientific areas. The humanities and social sciences already reached a share of 50 % in 2014/2015 (CRC with INF project), and this has not changed over the years. In the life sciences, in 2014/2015 approx. 8-10 % of all CRCs had an INF project. This proportion is approx. 40 % in 2024. The increase is lowest in the natural sciences (currently below 20 %). Here numbers for 2014/2015 are missing. In contrast, we see the highest increase in engineering sciences (from under 10 % in 2014/2015 to just under 50 % currently, see Bastian et al. 2025). This reflects also the experience from the discussed examples from the CRCs in this paper. The importance of INF projects and the awareness of research data management is further fostered by the start the consortia of the National Research Data Infrastructure (NFDI) in 2020 and the introduction of the Code of Conduct "Guidelines for Safeguarding Good Research Practice" by the DFG in 2019, in which requirements towards RDM are formulated with more emphasis (Deutsche Forschungsgemeinschaft 2022).

Due to their interdisciplinary nature and the high involvement of early-career researchers in CRCs, corresponding INF projects serve as key multipliers to foster integration research data management into research workflows. Collaboration across up to 20 subprojects is an intrinsic motivation for open data exchange and reuse. CRCs are therefore particularly suitable for methodically accompanying and supporting the management of research data early on in the data life cycle. Their lifespan of 12 years is beneficial for the development

and evaluation of research data infrastructures. INF projects operate at the intersection of existing solutions at institutes, local RDM facilities or NFDI services, and the specific needs and objectives of a CRC. In addition to organizational barriers, basic technical restrictions, such as the inaccessibility of services and infrastructures at project partners' sites, often pose significant challenges. This is particularly true for Transregional Collaborative Research Centers spanning multiple locations, where isolated IT infrastructures can limit the implementation of a unified research data management approach.

Depending on the focus of an CRC, an INF project may be more service-oriented or research-driven. Key aspects addressed include development of data repositories, knowledge management systems, and collaborative data management tools, definition of data and metadata formats, design of data analysis methods, or development of AI-based approaches for research data processing.

3 Key Questions

For this paper we have defined six key questions, which we have identified as outstanding aspects of INF projects in CRCs. The selected INF projects represent different scientific disciplines with their respective requirements towards research data and their management. They also represent different project runtimes and maturity levels with regard to the collaborative work on research questions, the sharing of data and the progress of the implementation of research data services and infrastructures. Additionally with the TRR 375 we discuss the challenges of collaborative research spawning institutes from two universities.

- Q1: Which individual challenges in INF Projects arise?
- Q2: What are experiences and best practices for collaboration and the development of interfaces with local, regional, and national research data infrastructures?
- Q3: Which innovations and insights emerge from INF projects, and how can they be broadly applied within the mentioned infrastructures?
- Q4: Training and capacity building: How can the training and awareness of researchers, particularly early-career scientists, in the use of RDM tools and processes be optimized?
- Q5: How are AI and advanced data analysis methods addressed in INF?
- Q6: What strategies for the sustainable storage and archiving of research data are being pursued in INF projects, and how can they be further optimized?

With answering and discussing these key questions we aim to identify recurring challenges, approaches and methods applied.

3.1 Individual Challenges of INF Projects

This chapter comprises the overarching research questions addressed by each CRC bringing into context the aims and work of the corresponding INF projects. What discipline- and research-specific obstacles and challenges arise when implementing RDM in these interdisciplinary CRCs? These may include the development of new software tools to support data workflows, the application and extension of metadata standards to ensure interoperability, and the adoption of advanced data analysis methods to handle complex and heterogeneous datasets. Managing data access rights during the project phase is another crucial aspect, requiring a balance between openness and security.

The central objective of CRC 1153, 'Tailored Forming' is to harness the potential of hybrid solid components through an innovative tailored manufacturing process based on pre-joined semi-finished products. The development of these novel process chains involves linking multiple sequential process steps to produce hybrid components with locally customized material properties (Behrens and Uhe 2021). The INF project supports data harmonization and integration along the entire process chain, facilitates the creation of synergies between individual process steps with the aim to identify complex downstream dependencies of parameters and results across these steps. Additionally, the INF project provides training on data literacy and develops and provides workflows for FAIR data publications, in the meaning of publishing Findable, Accessible, Interoperable, and Reusable (FAIR) data.

The aim of the CRC 1368 is to gain a fundamental understanding of the processes and mechanisms in manufacturing, assembly and handling technology processes with the (technically) complete oxygen-free atmosphere which is generated by using inert gas doped with silane (SiH_4). In this way, processes with effectively oxide-free metal surfaces can be investigated (Wegewitz et al. 2023). The INF project develops an ecosystem of semantic services and RDM infrastructures to capture data generation documented with protocols and stores corresponding data in a data repository linking data about processes, devices, methods with resulting research data. The ecosystem integrates existing solutions such as the electronic lab notebook (ELN) eLabFTW and enriches its data semantically. The processes and protocols used in this research are subject to continuous further development, so that the INF project has developed semantic building blocks for digital documentation. Similar to the CRC 1153 the INF project offers a rich portfolio of training to increase the awareness of RDM in the research community.

In both the CRC 1153 and the CRC 1368 domain-specific vocabularies and ontologies are developed to create semantically rich, machine-actionable metadata.

The scientific goal of CRC 1333 is to identify, quantify, and exploit confinement effects in catalysis by synthesis and characterization of mesoporous materials, by immobilization and characterization of organometallic and organic catalysts inside the pores, and by studying the influence of the mesopores on the performance of the catalysts (trans-

port, reaction kinetics, selectivity) by analytical and simulation tools. To support the acquisition of (meta)data and the publication of datasets according to the FAIR data principles, electronic lab notebooks (Chemotion, sciformation ELN, eLabFTW) are used, and datasets are uploaded to the university data repository. During the second funding period the INF project further extends this infrastructure by developing project-specific data models, which structure experimental data and metadata from the beginning of the project and makes it accessible to modular workflow for data acquisition, data analysis, and the automated upload to the DaRUS repository (Lauterbach et al. 2023). The data models are streamlined to existing formats and ontologies to enable the exchange of data by using standardized data exchange formats in chemistry such as ADF, DEXPI, ThermoML (Gültig et al. 2022), EnzymeML (Range et al. 2021), and nmrML (Schober et al. 2018).

The aim of TRR 375 is to establish a new class of multifunctional high-performance components (Platz et al. 2024). These consist of hybrid porous materials that are characterized by a combination of different metallic materials and the targeted incorporation of pores. The particular challenges faced by the Transregio initiative lie in establishing interdisciplinary research data management services across multiple university and institutional locations. This requires integrating heterogeneous data spaces, data tools and data workflows as well as the cross-site exchange of very large amounts of data (Computed Tomography measurements, Finite Element simulations etc.) while respecting site-specific university access rights. For the development and operation of this cross-site RDM platform existing software tools like the electronic lab book eLabFTW need to be fully integrated. A key focus is further the integration of simulation and experimental data into knowledge graphs, facilitating the generation of novel insights through combined knowledge modeling approaches.

The objective of CRC 1127 ChemBioSys is the investigation of fundamental regulatory mechanisms within complex biosystems that significantly influence everyday life. Researchers of the CRC seek to identify novel chemical mediators and their corresponding targets that contribute to shaping complex biological communities, and to elucidate the underlying mechanisms responsible for the formation and maintenance of community structures and their diversity. The long-term goal involves the targeted modulation of complex biosystems through the application of chemical mediators, including their metal complexes. The INF project addresses varying data literacy among chemists through tailored training. It establishes common metadata standards spanning disciplines while enabling the capture of discipline-specific details. Training covers introduction of electronic lab notebooks into traditional paper-based wet-lab workflows, data management with CRC's SEEK platform, and translating the abstract FAIR data concepts into practical daily laboratory procedures. The project further creates preservation strategies for diverse data types (spectra, images, code, molecular structures).

In summary, the INF projects share commitments to develop community tailored data models, metadata profiles including semantics. Tools like ELNs are frequently used to

support data management, minor differences in workflows and chosen repositories arise from the nature of the data.

3.2 Experiences and Best Practices for Collaboration with Local, Regional, National, or International Research Data Infrastructures

Data repositories are the core of RDM infrastructures. The CRCs 1127, 1153 and 1368 work on data migration pathways from private project-specific storage to public long-term archival systems at their university computing centers (see Table 1).

In cooperation with the LUIS IT-Center this approach yields key benefits, including the assignment of Digital Object Identifiers (DOIs), thus significantly enhancing the datasets' compliance with FAIR principles. Additionally, CRC 1153, CRC 1368, and TRR 375 collaborate with LUIS on the introduction of the ELN software eLabFTW as early adopters. Additionally, the Metadata4Ing ontology (Arndt et al. 2023), developed by NFDI4ING, is utilized in Semantic MediaWiki (SMW) to document data generation within research protocols (Altun et al. 2024). In Jena the CRC 1127 offers the ChemBioSys data platform with private and public dataspaces that serves as a central hub for raw and curated research data across the entire CRC. The INF project further provides COCONUT 2.0, the world's largest open natural products database, to disseminate results to the international research community. A collaboration with NFDI4Chem enables the provision of DOIs for data collections in COCONUT 2.0. Further, the CRC partners with international initiatives like LOTUS to enhance the semantic representation of natural product data in Wikidata.

Table 1: Data Management Approaches of CRCs.

CRC	**Repository Type**		
	local	**regional, national**	**international**
375 1153 1368	private CRC Repo	public LUIS Repo with DOIs	
1127	embargoed (private) and public Datasets in ChemBioSys Repo		COCONUT with DOIs
1333	embargoed (private) and public Datasets in DaRUS with DOIs		

The CRC 1333 utilizes the DaRUS data repository of the University of Stuttgart to publish datasets. In a collaboration with the research software engineers of the Kompetenzzentrum für Forschungsdaten (University of Stuttgart) and developers of the open-source Dataverse repository software, the INF project contributes to the automation of dataset

generation and publication by developing data processing workflows. In collaboration with the Allotrope Foundation and the national consortia NFDI4Chem and NFDI4Cat, the CRC 1333 INF project fosters the application of open standards in chemical sciences. In summary we experience a stepwise process, first on the level or private data either in separated, closed repositories or as embargoed datasets and then the final public release.

3.3 Transfer and Application of Innovations into Broader Application in Research Infrastructures

The close integration of INF projects in the cross-disciplinary research environment of CRCs usually results in demand-oriented approaches based on the researcher's specific research problems. The development of the RDM tools and infrastructures also follows the Open-Source principles, which ensures that these developments are generally easy to reuse in similar contexts.

The CRC 1368 INF project has developed a reusable software adapter for the open-source electronic lab notebook eLabFTW, which enables data migration, harmonization and mapping into a SMW. The mapping allows an enrichment of data into a semantic data structure. This adapter significantly improves interoperability across different ELN systems, allowing for seamless data exchange and collaborative research efforts. Both INF projects of CRC 1153 and CRC 1368 have also developed standardized data exchange workflows and protocols for data generation, facilitating smooth integration and efficient sharing of research data. Both INF projects develop CRC-customized data repositories based on the open-source CKAN software. Several CKAN plugin developments are shared with the open-source community including plugins for data visualization, automated metadata extraction from research data files, and simplified data uploads, thus enhancing usability and extending CKAN's application across diverse research contexts. One plugin connects the private data space of a CRC with public data repositories supporting the easy-to-use publication of FAIR data towards the end of the data life cycle by simply pushing data sets with already prepared metadata from the CRC data repository to the repository of the Leibniz University Hannover.

The TRR 375 investigates hypo-porous materials using computer tomography (CT), very similar to the activities of NFDI4Bioimage consortium, which applies CT techniques to biological materials. Despite differences in the studied materials, both initiatives face comparable challenges related to managing large volumes of imaging data, particularly regarding efficient data transfer, storage, and management. The INF project and NFDI4Bioimage share best practices and experiences to identify and implement suitable systems and workflows for optimal data transfer and storage of large amounts of data.

The work of the CRC 1127 INF projects resulted in several innovations which have a benefit to the community beyond the CRC. Its contributions include DECIMER.ai, an open

platform for chemical structure recognition, expanding its applications from academia to industry through collaboration with IBM Research Zurich. ChemBioSys also enhanced the architecture of the COCONUT 2.0 database to support international initiatives such as LOTUS (Rutz et al. 2022) and Wikidata, providing a centralized hosting environment for multiple open data repositories. Additionally, the STOUT V2 tool was developed from a research concept into a publicly accessible web tool, facilitating chemical nomenclature translation (SMILES-to-IUPAC) for the chemistry community. Utilizing Google's Tensor Processing Units (TPUs), ChemBioSys made computationally intensive models accessible to researchers globally. Further Cheminformatics Microservices provide standardized access to previously fragmented chemistry toolkits through unified APIs. Openly available training datasets and benchmarks for hand-drawn chemical structure recognition have also been created to support technological development in this area.

The CRC 1333 INF project develops reusable and adaptable workflows and data models for catalytic reactions, SAXS, NMR, IR, and Raman spectroscopy, flow chemistry, and synthesis of mesoporous materials, which are of general interest for the chemical sciences. The CRC provides use cases and comprehensive tutorials on the application of data and metadata standards like EnzymeML, ThermoML, nmrML, or DEXPI, which aim at establishing best practices for experimentation, research data management, and data analysis. The INF project further contributes to the development of the open-source software Dataverse, which is the underlying technology of the local data repository DaRUS at Stuttgart University. Publishing data and metadata of the experiment and the subsequent data analysis steps in standardized data exchange formats enhances interoperability and promotes data reuse within and beyond the chemical research community (Pleiss 2024; Behr et al. 2024).

3.4 Training and Capacity Building

This key question deals with the broad activities of INF projects supporting cultural change in the research communities towards a FAIR research data management. How can the training and awareness of researchers, particularly early-career scientists, in the use of RDM tools and processes be optimized? How well does collaboration with local, regional, and national RDM infrastructures function in this regard?

Both CRC 1153 and CRC 1368 initially conducted RDM workshops in collaboration with local university RDM teams, enhancing early awareness among researchers. A partnership with the university's IT Center facilitated the early adoption of ELN software. The INF team regularly provides hands-on tutorials and workshops on specific RDM system components, tailored to researchers' practical needs. Surveys collect direct feedback from researchers, shaping the development of new features and improvements in workflows. The most prominent challenge identified during workshops and training sessions, is to teach early career scientists how to identify the appropriate data life cycle stages, select suitable

data, and to provide comprehensive descriptions for publication, either as independent data publication or as supplemental publication in combination with an article.

TRR 375, being in the early stage of the project, aims to raise awareness for professional research data management by demonstrating how RDM tools directly benefit researchers' own work from the very beginning. Training activities, including workshops, brief "coffee lectures", and video tutorials, cater to diverse learning preferences. PhD summer schools explicitly integrate RDM training by teaching FAIR data principles and emphasizing the connections between robust data management and sound good scientific practice. Researchers are provided with practical guidance on preparing data publications effectively.

Similarly, the CRC 1127 INF project emphasizes practical experience through "Bring Your Own Data" workshops, effectively bridging theoretical RDM knowledge and real research scenarios. Embedding RDM education within specific research contexts, such as cheminformatics, enhances relevance and encourages adoption. Additionally, practical Python workshops empower researchers to develop custom tools, fostering greater programming proficiency and technical confidence in daily research activities. Development of open-source tools accompanied by thorough documentation and interactive tutorials reduces barriers for early-career researchers. Collaboration with national infrastructures such as NFDI4Chem and local university computing centers ensures sustainable and long-term implementation of RDM practices. The CRC 1333 INF project demonstrates the feasibility and benefits of FAIR RDM by applying structured workflows from the beginning of a research project. The training of doctoral researchers includes Python programming courses, dedicated RDM courses and workshops, and hands-on practical experiences. Best practices are established in experimentation, data analysis, and data management through close collaboration between experimentalists and research software engineers. Additionally, successful implementations of workflows and practices are documented and published jointly, highlighting effective approaches and their advantages (Giess et al. 2023).

3.5 Integration of AI and innovative Data Analysis Methods

In this chapter we discuss how new technologies such as AI and machine learning can be integrated into INF projects to provide automated solutions for RDM. In order not to lead expectations too far in the wrong direction, we have to keep in mind the overarching objectives of INF projects in CRCs which is the creation of structured, ideally semantically annotated, machine-interpretable FAIR data. INF projects are therefore mainly enablers of AI applications in different contexts by generating and making available highly structured data.

The generation of semantically rich, machine-actionable data is consequently the focus of the CRCs 1153 and CRC 1368, which develop semantically annotated Linked Data

with the aim to enable cross-subproject data analysis. The linked data cloud of the CRCs consists of protocols maintained in SMW and datasets stored in repositories, both are systematically annotated with controlled vocabularies. This semantic rich metadata enables seamless export and ingestion of serialized RDF data, applying the DCAT standard and derived DCAT application profiles (AP), into a triplestore creating a knowledge graph.

Extending the semantic foundation of controlled vocabularies and ontologies the CRC 1153 INF project has developed a Visual Inspection Ontology, which is used to represent quality assurance measures systematically across various subprojects and different stages within the research process chain (Sheveleva et al. 2022). CRC 1368 INF employs semantic modeling to represent hardware, machinery, and tools used across its projects, managed through a machine registry linked to experimental protocols and datasets, underpinned by the MATO ontology (Altun et al. 2021). Additionally, key concepts within CRC 1368, such as specimens, variables, protocols, and datasets, are mapped to the Metadata4Ing ontology developed by NFDI4ING, ensuring broader interoperability and standardization.

TRR 375 aims to leverage AI technologies specifically for automating the extraction of metadata directly from data resource files, associated data publications and journal articles. By automating these metadata extraction processes, researchers can ensure greater consistency, reduce manual effort, and enhance data discoverability.

CRC 1127 ChemBioSys extensively applies machine learning for automated metadata extraction from scientific literature, exemplified by the DECIMER.ai platform, which mines chemical structures from textual publications. The CRC is also developing AI-powered data curation workflows, harmonising diverse data formats, and automatically identifying data quality issues before repository submission. Additionally, image recognition technologies have been developed to digitize analog research information, converting legacy hand-drawn chemical structures and texts into structured, machine-readable formats. Natural language processing (NLP) techniques are further used for the systematic extraction of structured chemical information from unstructured text in scientific articles and lab notebooks. ChemBioSys is also fine-tuning openly available multimodal models on its generated datasets, creating domain-specific AI models optimized for data extraction tasks. AI-assisted data classification tools automatically align research outputs with community-established ontologies and domain-specific standards.

The CRC 1333 also identified structured and machine-readable (meta)data as one of the most important prerequisites to the successful application of machine learning (Giess and Pleiss 2025). The INF project aims to integrate data and metadata into a comprehensive knowledge graph, facilitating advanced queries and data linkage. It leverages large language models in combination with structured data models for automated metadata extraction from textual sources. Moreover, data-driven modeling approaches such as neural ordinary differential equations are applied to kinetic modeling of chemical reactions, illustrating the application of AI-driven paradigms within experimental workflows.

3.6 Sustainability and Long-term Archiving

Sustainable storage and long-term archiving of research data are essential components of successful RDM. INF projects within CRCs adopt diverse yet interconnected strategies to ensure data sustainability, while also focusing on integration into existing local, regional, and national infrastructures. The final key question discusses what different or similar strategies for the sustainable storage and archiving of research data are being pursued in INF projects, and how they can be further optimized.

CRC 1153 and CRC 1368 have established dedicated CKAN-based data repositories within a closed data space, ensuring secure and private data availability among subprojects of the CRC only. Comprehensive, semantically rich documentation of data generation processes is maintained through SMW, which directly links to datasets stored in these repositories. Furthermore, a specialized "data pusher" enables automated transfer of datasets from private CRC repositories into the institutional repository at Leibniz University Hannover without additional manual efforts like metadata entry or data uploads by researchers. Data transferred this way are assigned a DOI and become publicly available under FAIR principles. CRC 1153 specifically is developing a detailed archiving strategy for protocol data in preparation for the conclusion of its third funding period.

Due to multiple institutional locations with isolated IT systems, the TRR 375 INF project employs a two-tiered storage strategy. Initially, data is housed within a CRC-internal repository to support immediate cross-site access and in an additional repository for the exchange of big data files such as CT-scans by means of a special service of the Leibniz University Hannover (High-Seas). Protocols and procedural documentation are systematically recorded in SMW, facilitating data integration from distributed storage sites. An RDM portal consolidates data and protocols from these dispersed locations. As one participating university lacks a local institutional repository, TRR 375 is also exploring national and domain-specific solutions to facilitate automated data transfer into publicly accessible repositories and is working towards integrating existing ELN systems across locations.

Research data at ChemBioSys is primarily stored on a dedicated repository[1], currently migrating to a university-managed virtual machine to enhance integration and sustainability. Long-term archiving utilizes university backup services, ensuring robust preservation. Researchers are encouraged to deposit datasets openly in platforms such as Zenodo and FigShare, significantly improving accessibility and reuse. Additionally, the INF project promotes the publication of so-called data note papers to increase the visibility and reuse potential of datasets. Data notes provide concise descriptions of well-curated research datasets, aiming to enhance their visibility, transparency, and reuse potential. They also assist researchers in meeting funding agencies' requirements for data sharing. Before submitting a data note, datasets usually must be deposited in an appropriate repository, en-

1 https://data.chembiosys.de; *Visited on September 3, 2025 by the editors.*

suring their completeness and accessibility. Data notes are exemplified by ChemBioSys's hand-drawn chemical structure dataset publications. Furthermore, the widespread adoption of open-source software like the electronic lab notebooks and the RDM technology stack of the CRC reduces reliance on proprietary solutions, supporting long-term sustainability. Integration with local, regional, and national infrastructures, combined with rigorous adherence to FAIR principles, further guarantees data interoperability and prolonged usability.

The CRC 1333 INF project has developed automated workflows for uploading datasets and generating metadata blocks specific to the DaRUS repository, enhancing the findability and accessibility of data. Datasets are stored in standardized data exchange formats to ensure interoperability and reusability. Detailed documentation of data acquisition and analysis processes accompanies stored datasets, adhering closely to the FAIR principles and ensuring the longevity and transparency of research data.

4 Discussion

Examining the key questions reveals several common challenges and solutions across the presented CRCs, likely due to their closely related scientific disciplines. All INF projects successfully integrate into local, regional research data management infrastructures and maintain active collaborations with National Research Data Infrastructure (NFDI) consortia. Regular cooperation with local RDM teams raises awareness among researchers, particularly early-career scientists, through workshops on data management plans (DMPs) and discipline-specific training sessions such as ELNs and Python courses. CRC-specific data policies and customized DMPs establish a common regulatory framework that everyone can refer to. Practical, domain-specific training sessions, often using researchers' own data, consistently receive higher acceptance compared to generic formats, significantly boosting data science literacy and openness towards robust RDM practices. The INF projects across different CRCs have generated significant innovations and insights not only for their research area, but in the creation of reusable, interoperable software and workflows that facilitate effective RDM. Common approaches include developing standardized data exchange protocols, leveraging open-source solutions for data capture and semantic data annotation to enhance interoperability. All approaches reflect the need for community-tailored solutions. The INF projects have addressed the emerging topics of AI and machine learning methods into their RDM processes with varying approaches. The generation of semantically rich, machine-actionable data, employing knowledge graphs and controlled vocabularies for interoperability is a general objective of INF projects. CRC 1153 and CRC 1368 specifically emphasize structured semantic annotations using specialized ontologies like Visual Inspection and MATO, supporting advanced cross-subproject data analyses. Conversely, CRC 1127 leverages AI extensively for metadata extraction and data curation, utilizing platforms such as DECIMER.ai, NLP, and multimodal modeling. CRC 1333 similarly integrates AI-driven techniques but focuses more heavily on

structured data modeling and data-driven approaches. With regards to sustainability and long-term archiving all CRCs initially employ private data spaces, either for embargo purposes, as private data collections within public repositories, or CRC-specific internal repositories. A common emphasis is placed on capturing comprehensive data provenance, supported by various platforms including ELNs and knowledge management systems, ultimately aiming at standardized, sustainable, and FAIR-compliant archiving practices. Private or embargoed data is later on made available in publicly available data repositories following the FAIR principles.

From the experiences discussed, several recommendations emerge for future INF projects. Integrating research data management, and therefore an INF project, early in the research process significantly enhances its acceptance and effectiveness. Utilizing concrete research data from researchers and projects during training sessions and tutorials helps researchers immediately recognize the practical value of RDM services, infrastructures, and tools. Additionally, early planning for connectivity to local and national RDM infrastructures facilitates the identification and application of established standards. Implementing user-friendly tools like ELNs and data annotation platforms further raises acceptance and promotes professional RDM practices. Finally, structured documentation combined with automatic metadata extraction and discipline-specific metadata standards substantially benefits interdisciplinary research collaborations.

5 Conclusion and Outlook

In conclusion, the discussion of the presented CRC INF projects demonstrates several shared approaches, but also some distinct strategies in managing interdisciplinary research data effectively. Common successful practices include developing domain-specific metadata schemas, employing open-source software for the various RDM tasks, and establishing automated workflows for data management. Tailored training programs that blend theoretical knowledge with hands-on application consistently improve researcher acceptance, particularly among early-career scientists. One difference between the examined CRC INF projects discussed in this paper is their different role as enabler of ML- and AI-applications compared to actively moving towards implementations ML- and AI-methods in RDM systems and workflows. Engineering-focused CRCs emphasize the integration of semantic annotation and structured metadata for managing complex process chains, whereas chemistry- and biology-oriented projects highlight the role of AI-driven extraction and curation of data from heterogeneous sources like literature and images. Infrastructure integration also varies, with some CRCs directly embedding into institutional repositories, while others manage private data spaces before transitioning datasets into national or domain-specific repositories. Notably, the current selection of CRC examples lacks diversity across scientific domains, limiting broader insights into RDM strategies applicable to other fields such as humanities or social sciences. Recognizing this limitation, future INF workshops should aim to incorporate a wider array of disciplines. Such expanded

participation will enhance the exchange of experiences and best practices, providing fresh perspectives and innovative approaches to research data management for all scientific communities.

Acknowledgements

The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding and support.

Work of the described projects is funded by CRC 1368 DFG Project number 394563137, CRC 1127 DFG Project number 239748522, CRC 1153 DFG Project number 252662854, CRC 1333 DFG Project number 358283783, and TRR 375 DFG Project number 511263698.

Authorship Contributions

Oliver Koepler drafted the idea, concept and initial version of the paper. He specifically contributed for CRC 1153, CRC 1368 and TRR 375. Iryna Mozgova contributed information about CRCs 1153 and 1368. Christoph Steinbeck contributed information about work and challenges in CRC 1127. Florian Nürnberger contributed information regarding TRR 375. Jürgen Pleiss contributed information about work and challenges in CRC 1333.

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

Altun, Osman, Marc Hinterthaner, Khemais Barienti, Florian Nürnberger, Roland Lachmayer, Iryna Mozgova, Oliver Koepler, and Sören Auer. 2024. “Contextualization for generating FAIR data: A dynamic model for documenting research activities”. In *Product Lifecycle Management. Leveraging Digital Twins, Circular Economy, and Knowledge Management for Sustainable Innovation,* edited by Christophe Danjou, Ramy Harik, Felix Nyffenegger, Louis Rivest, and Abdelaziz Bouras, 116–126. IFIP

advances in information and communication technology. Cham: Springer Nature Switzerland. ISBN: 9783031625787. https://doi.org/10.1007/978-3-031-62578-7_11.

Altun, Osman, Tatyana Sheveleva, André Castro, Pooya Oladazimi, Oliver Koepler, Iryna Mozgova, Roland Lachmayer, and Sören Auer. 2021. "Integration eines digitalen Maschinenparks in ein Forschungsdatenmanagementsystem". In *DS 111: Proceedings of the 32nd Symposium Design for X,* 1–10. DFX2021. The Design Society. https://doi.org/10.35199/dfx2021.23.

Arndt, Susanne, Benjamin Farnbacher, Marc Fuhrmans, Stephan Hachinger, Johanna Hickmann, Nils Hoppe, Martin Thomas Horsch, et al. 2023. *Metadata4Ing: An ontology for describing the generation of research data within a scientific activity.* Zenodo. https://doi.org/10.5281/zenodo.14982558.

Bastian, Alina, Ortrun Brand, Jens Dierkes, Janna Neumann, and Jürgen Windeck. 2025. *SFB-INF-Workshop 2024: Let's share! Gute Praktiken und Fallstricke im Datenmanagement in SFBs und Forschungsgruppen – Veranstaltungsbericht.* Zenodo. https://doi.org/10.5281/zenodo.15475394.

Behr, Alexander S., Julia Surkamp, Elnaz Abbaspour, Max Häußler, Stephan Lütz, Jürgen Pleiss, Norbert Kockmann, and Katrin Rosenthal. 2024. "Fluent integration of laboratory data into biocatalytic process simulation using EnzymeML, DWSIM, and ontologies". *Processes (Basel)* 12 (3): 597. https://doi.org/10.3390/pr12030597.

Behrens, Bernd.Arno, and Johanna Uhe. 2021. "Introduction to tailored forming". *Production Engineering* 15 (2): 133–136. https://doi.org/10.1007/s11740-021-01022-w.

Deutsche Forschungsgemeinschaft. 2022. *Guidelines for Safeguarding Good Research Practice. Code of Conduct.* Zenodo. https://doi.org/10.5281/ZENODO.6472827.

Giess, Torsten, Selina Itzigehl, Jan Range, Richard Schömig, Johanna R. Bruckner, and Jürgen Pleiss. 2023. "FAIR and scalable management of small-angle X-ray scattering data". *Journal of Applied Crystallography* 56 (2): 565–575. https://doi.org/10.1107/s1600576723001577.

Giess, Torsten, and Jürgen Pleiss. 2025. "Digitalization of biocatalysis: Best practices to research data management". In *Biocatalysis Identifying novel enzymes and applying them in cell-free and whole-cell biocatalysis,* edited by Dirk Tischler, 714:19–43. Methods in Enzymology. Elsevier. ISBN: 9780443317880. https://doi.org/10.1016/bs.mie.2025.01.040.

Gültig, Matthias, Jan P. Range, Benjamin Schmitz, and Jürgen Pleiss. 2022. "Integration of simulated and experimentally determined thermophysical properties of aqueous mixtures by ThermoML". *Journal of Chemical & Engineering Data* 67 (11): 3340–3350. https://doi.org/10.1021/acs.jced.2c00391.

Lauterbach, Simone, Hannah Dienhart, Jan P. Range, Stephan Malzacher, Jan-Dirk Spöring, Dörte Rother, Maria Filipa Pinto, et al. 2023. "EnzymeML: seamless data flow and modeling of enzymatic data". *Nature Methods* 20 (3): 400–402. https://doi.org/10.1038/s41592-022-01763-1.

Platz, Jacques, Johanna Steiner-Stark, Benjamin Kirsch, and Jan C. Aurich. 2024. "Fertigung funktional gradierter Materialien auf porösen Metallen durch Laserauftragschweißen". *ZWF Z. Wirtsch. Fabr.* 119 (7–8): 515–519. https://doi.org/10.1515/zwf-2024-1092.

Pleiss, Jürgen. 2024. "FAIR data and software: Improving efficiency and quality of biocatalytic science". *ACS Catal.* 14 (4): 2709–2718. https://doi.org/10.1021/acscatal.3c06337.

Range, Jan, Colin Halupczok, Jens Lohmann, Neil Swainston, Carsten Kettner, Frank T. Bergmann, Andreas Weidemann, Ulrike Wittig, Santiago Schnell, and Jürgen Pleiss. 2021. "EnzymeML-a data exchange format for biocatalysis and enzymology". *FEBS J.* 289 (19): 5864–5874. https://doi.org/https://doi.org/10.1111/febs.16318.

Rutz, Adriano, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G Graham, et al. 2022. "The LOTUS initiative for open knowledge management in natural products research". *eLife* 11. https://doi.org/10.7554/elife.70780.

Schober, Daniel, Daniel Jacob, Michael Wilson, Joseph A. Cruz, Ana Marcu, Jason R. Grant, Annick Moing, et al. 2018. "nmrML: A Community Supported Open Data Standard for the Description, Storage, and Exchange of NMR Data". *Analytical Chemistry* 90 (1): 649–656. https://doi.org/10.1021/acs.analchem.7b02795.

Sheveleva, Tatyana, Kevin Herrmann, Max Leo Wawer, Christoph Kahra, Florian Nürnberger, Oliver Koepler, Iryna Mozgova, Roland Lachmayer, and Sören Auer. 2022. "Ontology-Based Documentation of Quality Assurance Measures Using the Example of a Visual Inspection". *Lecture Notes in Networks and Systems* 546:415–424. https://doi.org/10.15488/13181.

Wegewitz, Lienhard, Wolfgang Maus-Friedrichs, René Gustus, Hans Jürgen Maier, and Sebastian Herbst. 2023. "Oxygen-free production – from vision to application". *Advanced Engineering Materials* 25 (12). https://doi.org/10.1002/adem.202201819.