Datenbank-Spektrum (2024) 24:97–105
https://doi.org/10.1007/s13222-024-00475-4

SCHWERPUNKTBEITRAG

Research Data Management in Simulation Science: Infrastructure, Tools, and Applications

Bernd Flemisch[1] · Sibylle Hermann[2] · Melanie Herschel[3] · Dirk Pflüger[3] · Jürgen Pleiss[4] · Jan Range[2] · Sarbani Roy[2] · Makoto Takamoto[5] · Benjamin Uekermann[3]

Received: 1 February 2024 / Accepted: 26 April 2024 / Published online: 4 June 2024

Abstract
Research Data Management (RDM) has gained significant traction in recent years, being essential to allowing research data to be, e.g., findable, accessible, interoperable, and reproducible (FAIR), thereby fostering collaboration or accelerating scientific findings. We present solutions for RDM developed within the DFG-Funded Cluster of Excellence EXC2075 Data-Integrated Simulation Science (SimTech). After an introduction to the scientific context and challenges faced by simulation scientists, we outline the general data management infrastructure and present tools that address these challenges. Exemplary domain applications demonstrate the use and benefits of the proposed data management software solutions. These are complemented by additional measures for enablement and dissemination to foster the adoption of these techniques.

Keywords Research Data Management, Simulation Science, Software, Applications

1 Introduction

Research Data Management (RDM) has received increasing attention in recent years, e.g., to maintain research quality and reliability, promote collaboration among scientists, or accelerate impact of scientific findings. For instance, today, sharing and publishing research data is highly encouraged or even mandatory in conjunction with publishing scientific results or as part of research projects funded by agencies. The importance of research data management across a wide range of scientific disciplines is further reflected by large initiatives such as the National Research Data Infrastructure in Germany (NFDI)[1], where domain-specific consortia devise RDM solutions dedicated to the particular requirements of their field.

This paper presents data management solutions for research data infrastructures developed within the DFG-

[1] https://www.nfdi.de/.

Bernd Flemisch
bernd.flemisch@iws.uni-stuttgart.de

Sibylle Hermann
sibylle.hermann@simtech.uni-stuttgart.de

✉ Melanie Herschel
melanie.herschel@ipvs.uni-stuttgart.de

Dirk Pflüger
dirk.pflueger@ipvs.uni-stuttgart.de

Jürgen Pleiss
juergen.pleiss@itb.uni-stuttgart.de

Jan Range
jan.range@simtech.uni-stuttgart.de

Sarbani Roy
sarbani.roy@simtech.uni-stuttgart.de

Makoto Takamoto
Makoto.Takamoto@neclab.eu

Benjamin Uekermann
benjamin.uekermann@ipvs.uni-stuttgart.de

[1] Institute for Modelling Hydraulic and Environmental Systems, University of Stuttgart, Stuttgart, Germany

[2] Stuttgart Center for Simulation Science, University of Stuttgart, Stuttgart, Germany

[3] Institute for Parallel and Distributed Systems, University of Stuttgart, Stuttgart, Germany

[4] Institute of Biochemistry and Technical Biochemistry, University of Stuttgart, Stuttgart, Germany

[5] NEC Laboratories Europe, Heidelberg, Germany

funded Cluster of Excellence (EXC) EXC2075 *Data-Integrated Simulation Science (SimTech)*[2] at the University of Stuttgart. The cluster's research targets a new class of modeling and computational methods based on data available from various sources. This research both heavily relies on and contributes to novel data and software. Therefore, the dedicated EXC supporting structure for *Research Data and Software Management (RDSM)* has been established. It is substantiated by a coordinating committee formed by professors in the EXC and an operational team that consists of data stewards and research software engineers.

One central RDSM goal is that data and software originating from the EXC follow the FAIR principles [2, 14]. Going beyond that, we aim for reproducibility of all computational results published by researchers in the EXC. This requires publishing all data and code as well as the computational environment used to produce results. To achieve these goals, we identified four measures: (i) Provision and administration of infrastructure, tools, and services, (ii) Support of research data and software management activities, (iii) Enablement and exchange, and (iv) Dissemination and participation in overarching RDM activities.

This paper focuses on technical aspects, outlining the research data management infrastructure (Sect. 2) and tools we have developed to support simulation scientists (Sect. 3) as showcased in several exemplary applications (Sect. 4). Thereby, the focus of this paper lies on measures (i) and (ii), and we only provide a brief summary of highlight activities for the remaining measures.

Regarding measure (iii), i.e., enablement and exchange, we have established the Special Interest Group Data Infrastructure (SIGDIUS), which offers a regular forum to interested researchers who want to improve their RDSM activities on an individual, working group, or institute level. Further, we have conducted multiple Software Carpentry workshops, SimTech lecture series, graduate school seminars, and hands-on RDSM workshops to educate researchers and students (in SimTech and beyond) on best practices in RDM and research software engineering (RSE).

Concerning dissemination and participation in overarching RDM activities as part of measure (iv), we coordinated the compilation of a commitment to active RDSM, signed by spokespersons of more than 20 Clusters of Excellence [3]. Moreover, we have been involved in the opinion article on "An environment for sustainable research software..." [1]. In a position paper, we propose measures to improve research software engineering at the University of Stuttgart [6]. The SimTech contributions to RDM extend beyond the EXC, e.g., through involvement of SimTech researchers in the NFDI consortia MaRDI[3], NFDI4Cat[4], NFDI4Chem[5] and NFDI4Ing[6]. We also actively participate in, for example, sections and working groups of the overall NFDI organization, a CESAER task force[7], and two EOSC task forces[8].

2 Infrastructure Overview

The contributions we describe in this paper tackle several RDM challenges that we have been facing in the interdisciplinary simulation science community. We first summarize challenges before we outline our data management infrastructure that addresses these.

2.1 Challenges

Metadata management: Given the interdisciplinary nature of SimTech, the definition of an appropriate metadata schema is difficult, as not all disciplines require the same metadata to model the relevant information to support FAIR principles. Therefore, the metadata management solution for SimTech needs to integrate and support a high degree of flexibility in modeling, provisioning, and disseminating both research data and its descriptive metadata.

Difficult data entry: Having to manually create and fill in metadata when publishing a dataset represents a significant obstacle in RDM, e.g., because it is perceived as "distraction" from the more pressing research activities or represents a tedious, complex, and time-consuming process. Clearly, there is a need to better integrate and streamline this process within researchers' workflows in a (semi-)automatic way.

Quality control: Quality control plays a crucial role in ensuring the trustworthiness of published data and software. Just like articles undergo a review process to ensure their reliability and credibility, the same principle applies to other forms of information and resources. By implementing a quality control mechanism, organizations can establish a systematic approach to assess, monitor, and enhance the quality of their outputs. This helps to build confidence in the accuracy, integrity, and usability of the published data and software, making them more suitable for reuse. Without quality control, there is a higher risk of errors, inconsisten-

2 https://www.simtech.uni-stuttgart.de/exc/.

3 https://www.mardi4nfdi.de/.

4 https://nfdi4cat.org/.

5 https://www.nfdi4chem.de/.

6 https://nfdi4ing.de/.

7 https://www.cesaer.org/task-forces/task-force/?id=11825.

8 https://eosc.eu/advisory-groups/researcher-engagement-adoption, https://eosc.eu/advisory-groups/infrastructures-quality-research-software.

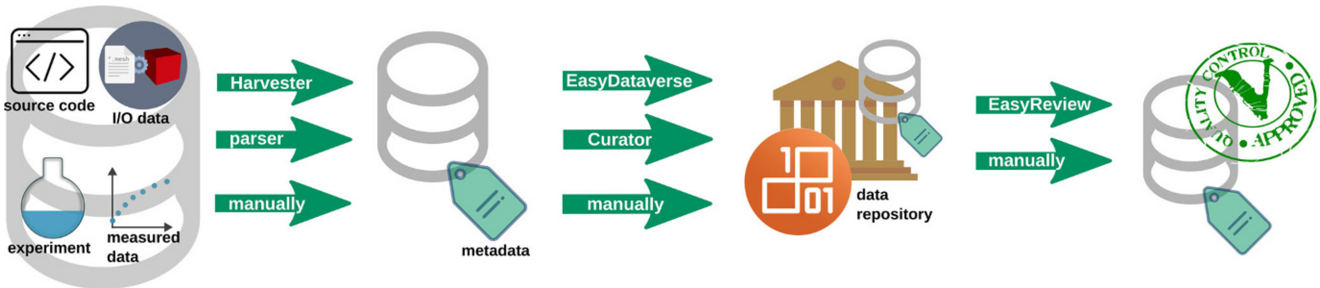

Fig. 1 Typical RDM workflow in simulation science. A locally produced dataset consisting of, e.g., source code, simulation input/output or experimental data is annotated with metadata and uploaded to a data repository. Before publication, the dataset undergoes a review process. Different possibilities for the required steps are depicted in the arrows, including the tools presented in this article

cies, and inaccuracies, which can undermine the trust and reliability of the published materials.

2.2 RDM in SimTech

The challenges summarized above are addressed both by the central RDM infrastructure at the University of Stuttgart and tools developed to cater the specific needs of simulation scientists. Fig. 1 illustrates the positioning of the corresponding components within a typical RDM workflow from locally produced research data to a published, quality-controlled and annotated dataset.

The central organisational unit for RDM at the University of Stuttgart is the Research Data Competence Center FoKUS, combining the expertise of the university's library and its Technical Information and Communication Services. Implemented by the FoKUS team, the institutional data repository DaRUS[9] has officially gone into production in September 2019. The implementation behind DaRUS is based on the open-source software Dataverse[10], which allows to maintain and publish research data as "datasets" within hierarchical collections called "dataverses" [5]. A dataverse for SimTech has been established early after, where in January 2024 more than 100 datasets have been published[11]. Different metadata schemes suitable for SimTech have been integrated into DaRUS, as for example the EngMeta scheme [10], which is specifically designed for engineering applications.

While the data repository described above offers suitable metadata management capabilities, having data and metadata enter the data repository remains cumbersome. To alleviate this tedious (meta-)data entry, we have developed a harvester tool described in Sect. 3.2 which is capable of parsing and curating metadata from user-specified folders and subsequently populating the relevant metadata fields in a dataset.

Various application programming interfaces (APIs), allowing the automated upload of research data, are offered by Dataverse to the researchers and selected depending on the use case. The Python API is employed by our EasyDataverse tool, which further facilitates the upload process and will be introduced in Sect. 3.1.

Concerning quality control, we have established a workflow to curate the formal description of the data publication process together with the FoKUS team. This process of publishing data involves several steps. First, the dataset is checked by one of the SimTech Stewards and then by the FoKUS team. The author of the dataset receives feedback and can adjust the dataset accordingly. To facilitate the process, we have developed quality criteria[12]. Additionally, we created the tool EasyReview, serving as review management system for datasets (see Sect. 3.3).

3 Tooling

As highlighted in the previous section, besides the manual workflow to publish high-quality data and robust metadata for efficient data categorization, indexing, and retrieval, we have developed tailored metadata management solutions, resulting in a variety of tools[13]. These solutions seamlessly integrate with the DaRUS repository (see Sect. 2.2), effectively addressing the challenge of varying metadata requirements in the interdisciplinary field of simulation science. The tools described below tackle the other two challenges described in Sect. 2.1, i.e., simplifying metadata entry and supporting quality control processes.

3.1 EasyDataverse

Dataverse offers a native application programming interface (API) that allows users to interact with and configure their Dataverse instance. This API includes the ability to upload

[9] https://darus.uni-stuttgart.de.

[10] https://dataverse.org/.

[11] https://darus.uni-stuttgart.de/dataverse/simtech.

[12] https://www.izus.uni-stuttgart.de/en/fokus/darus/qualitycontrol/.

[13] https://github.com/SimTech-Research-Data-Management.

and download metadata and files for datasets. However, creating a Dataverse JSON file that follows the metadata block configurations of the Dataverse instance can be a challenging task for users without programming knowledge. This can make it difficult to programmatically upload datasets, decreasing the usability of the Dataverse repository for automated dataset uploads.

To address this issue, the Python library EasyDataverse provides a simplified metadata report interface with key-value operations that adhere to the required Dataverse JSON structure. EasyDataverse offers metadata blocks as objects that can be populated like any other Python object through attribute assignment. Compound data is handled as compositions and attached to metadata block objects. These metadata blocks are dynamically served upon connecting to a Dataverse installation, allowing users to access any Dataverse installation, even with changes made to the installation or if operations are being carried out across different installations.

Once users have created their metadata reports, they can attach files and directories to their dataset using the Python-DVUploader library[14]. By specifying the target Dataverse collection, users can upload their dataset, including metadata and files, to a Dataverse instance. EasyDataverse also enables users to download and update their dataset, making it easy to automate any step within a simulation workflow. Both EasyDataverse and the Python-DVUploader are officially endorsed by the Global Dataverse Community Consortium (GDCC)[15] and the source codes are hosted under the GDCC GitHub organization[16].

3.2 Harvester-Curator

Despite the effort of simplifying the creation of Dataverse JSON files through EasyDataverse, researchers encounter challenges in extracting and inputting metadata into EasyDataverse. Collecting metadata, whether concurrently with ongoing research or retrospectively revisiting the entire research process, poses significant challenges — either disrupting the workflow or being cumbersome. This underscores the compelling need for developing an automated metadata collection process to streamline this intricate facet of research data management. The Python tool Harvester-Curator[17] represents a significant advancement in metadata provisioning, facilitating the seamless transition from a user's local folder to a dataset within a designated data and/or software repository. Executing the tool is straightforward, requiring a single command line for metadata harvesting and another for populating metadata fields of a dataset. Rigorous testing, conducted with demoDaRUS[18], a test server for DaRUS, improves its robust functionality. Ongoing efforts are focused on extending this functionality to other Dataverse installations and prominent repositories such as Zenodo. As an ongoing development, readers are encouraged to explore the tool's GitHub repository for the latest updates and contributions. For an overview, see the schematic diagram in Fig. 2.

The initial stage, termed Harvester, operates as a scanner, navigating through user-specified folders, categorizing files based on types, and identifying appropriate parsers for diverse file formats. These parsers are tailored to extract metadata from various structured file formats, methodically retrieving and consolidating information into a well-organized JSON file. The current version incorporates parsers for several structured file formats, including VTK, HDF5, CFF, BibTeX, JSON, and YAML. The tool's inherent extensibility allows seamless integration of new parsers, rendering it versatile and adaptable to the evolving requirements of research. Operating locally, this phase can collect metadata from large files often omitted from version control, offering researchers a streamlined and automated approach to comprehensive metadata collection.

In the subsequent stage, Harvester-Curator transforms into a Curator, leveraging gathered metadata to populate metadata fields within a designated data and/or software repository. The objective is to identify and assign (if available) a relevant metadata field of the target repository to each harvested metadata information, compiling it into a compatible metadata file for that specific repository. During this process, the aim is to minimize metadata loss and ensure the proper arrangement of metadata fields within the appropriate parent field and metadata schema of the target repository. The tool also offers options for plug-in mappings, establishing connections between the attributes of harvested metadata and the metadata fields of the target repository. This phase is currently tailored specifically for DaRUS, and we are utilizing EasyDataverse to create a new dataset and populate its metadata fields with compatible metadata file.

Through this automated process, this tool not only streamlines the process of metadata collection but also significantly contributes to enhancing data accessibility and interoperability within repositories, providing a robust and scalable solution for research data management.

[14] https://github.com/gdcc/python-dvuploader.

[15] https://www.gdcc.io/.

[16] https://github.com/gdcc.

[17] https://github.com/SimTech-Research-Data-Management/Harvester-Curator/tree/master.

[18] https://demodarus.izus.uni-stuttgart.de/.

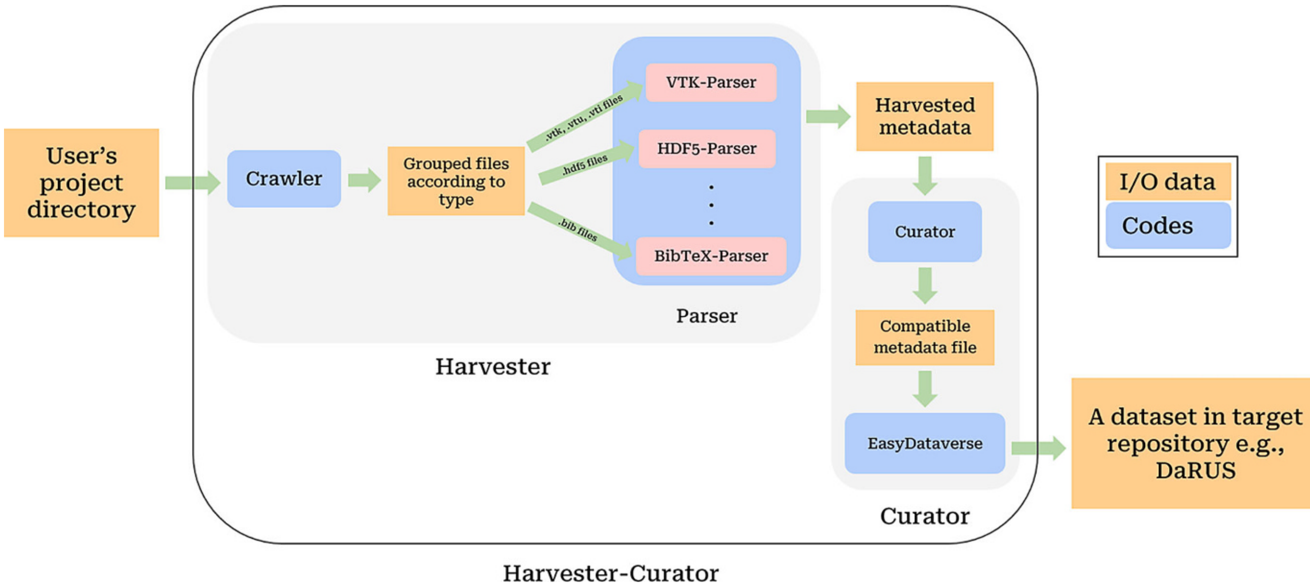

Fig. 2 Harvester-Curator, a tool to elevate metadata provision in data and/or software repositories

3.3 EasyReview

The final step in a publication workflow involves reviewing datasets to ensure that quality measures are achieved. This step is typically carried out by curators and experts of a specific field. As of April 2024, our team comprises just 7 curators who manage the data for 111 users. A manual report is compiled to summarize the review and provide suggestions for changes or additional fields needed to comply with SimTech's dataset publishing requirements. These reports are then sent to the depositor and authors of the dataset for feedback implementation. This iterative process continues until quality measures are met, ultimately leading to publication of the dataset. While this process is straightforward in theory, it presents several challenges in practice.

Firstly, it is important to ensure that the review is verbose enough for the user to understand. Secondly, the effectiveness of the review must be evaluated to ensure that all metadata fields and files have been addressed. Thirdly, it is essential to determine whether an additional expert review is required. Finally, it is important to establish transparency in the review process for both the user and repository team.

The challenges are heavily dependent on the curator's experience, capacity, and knowledge. For instance, if a curator has an insufficient background in a particular topic, the review can only be conducted superficially, and proper content review is impossible. Additionally, manually written reports may vary in quality, leaving users unsure of which actions to take. These challenges can be resolved with technical implementations that streamline, secure, and support the review process.

To address these issues, EasyReview is a web-based tool designed to provide a streamlined approach to reviewing datasets (see Fig. 3). EasyReview boasts a user-friendly graphical interface that enables reviewers to load datasets, review all fields, and accept, decline, or comment on specific parts. Reviewers can suggest improvements and even extend datasets themselves, which can then be accepted or declined by the user. EasyReview allows for dynamic communication between reviewers and users, with findings shared on a per-field basis. Furthermore, the tool allows for sharing reviews with multiple reviewers, ensuring that field experts can review the complete or partial dataset. Additionally, EasyReview provides an API and Python package to automate reviews that are capable of pre-filling and validating fields, leaving reviewers to only review fields that require human control. Ultimately, EasyReview's goal is to simplify and secure the review process on both ends such that users and reviewers are guided through the process of quality control and to maintain high-quality data using a simple graphical user interface.

4 Domain Applications

To showcase how the RDM infrastructure and the developed tools influence research in SimTech, we describe three applications as examples: The EnzymeML Platform, which uses EasyDataverse, a typical SimTech publication using simulation and data science, whose metadata is extracted by Harvester, and the benchmark platform PDEBench, which serves pre-calculated datasets and models that utilize EasyDataverse for provision and DVUploader for curation.

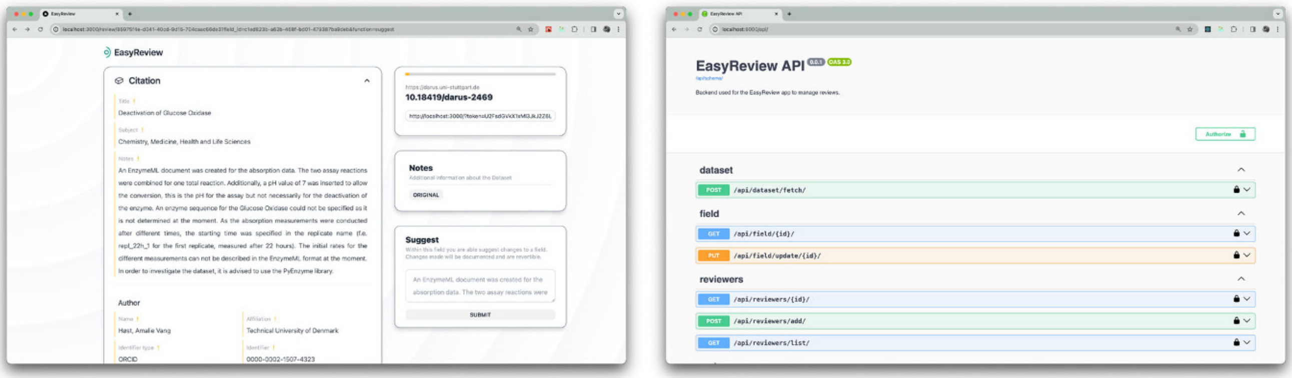

Fig. 3 Graphical user interface of EasyReview (*left*), which displays the metadata block in the *left column*. Visual indicators display the state of a field. In the *right column*, review statistics, field history and a text input for suggestions are shown. OpenAPI Swagger documentation of EasyReview's backend REST API (*right*) to automate dataset reviews

4.1 Integration of Data and Metadata by the EnzymeML Platform

Biocatalysis is a promising cornerstone of sustainable chemistry. High-throughput experimentation provides a steadily increasing stream of data, and novel methods in bioinformatics and molecular modelling enable the design of biocatalysts and bioprocesses. However, progress in development of bioprocesses is still slow, and it is not widely applied in chemistry, yet. One of the reasons is the inefficient management of data, because data acquisition is decoupled from modelling of reaction kinetics and from molecular simulations. FAIRification of the research process suggests a barrier-free access to the wealth of previously measured data, a re-analysis of data for developing new hypotheses, and a more directed design. Despite the rapidly growing number of publications in the field of biocatalysis, measured data is not published. However, training in machine learning demands large, machine-readable sets of positive and negative data. The EnzymeML platform was developed for the seamless integration of biocatalytic data with software for data acquisition, data analysis and modelling, and the automated upload to repositories [8]. The platform is based on the EnzymeML exchange format [9] and makes data management scalable, reproducible, collaborative, and modular. In the EnzymeML project, Jupyter Notebook serves as a modular workflow platform for data analysis, because it allows a rich documentation, reproducible execution of analysis steps, and automation of data management for high throughput experimentation. The EnzymeML platform uses Python as programming language, because it has a large user base and gives access to a rapidly developing toolbox for scientific computing. The EnzymeML GitHub repository[19] serves as collaborative development platform. Software is deployed to the user by installation from GitHub, preferably by containerization using Docker. Acquisition of data and metadata is achieved manually or by using the parser for reading in specific data formats generated by the analytical instruments. The raw data (time courses of light absorption or peak areas) are stored together with a standard curve that relates the measured property to the respective concentration of a reactant. Thus, postprocessing of raw data becomes reproducible, and a future re-analysis of raw data becomes feasible. In many biocatalytic experiments, a parameter study investigates the effect of cosolvent, temperature, pH, or amino acid sequence of the enzyme to the catalytic activity or selectivity of the enzyme by using microtiter plates, flow reactors, or microfluidic devices. In order to enable a systematic analysis of the measured data, the metadata of each experiment is reported and linked to the respective measured data. Because all experiments are stored and published, also negative data becomes available, and reaction conditions leading to low catalytic activity or low selectivity can be identified, which is crucial for data analysis, especially for training a learner. Because experimental data and metadata is machine-readable, existing platforms for modelling and data analysis can be seamlessly applied by integrating them into the data analysis workflow. The selected kinetic model and the estimated kinetic parameters are then added to the dataset. Thus, experimental and modelling data are interoperable and reusable. For making the dataset interoperable, reusable, findable and accessible, it is uploaded to DaRUS[20] by integrating EasyDataverse and DVUploader into the EnzymeML Python library PyEnzyme. Data and metadata are stored in the standardized data exchange format EnzymeML, a searchable metadata block is generated, and a digital object identifier (DOI) is

[19] https://github.com/EnzymeML.

[20] https://darus.uni-stuttgart.de/.

provided. In addition, the experimental results are uploaded to specialized enzyme reaction databases such as SABIO-RK[21] or STRENDA DB[22], referring to the respective DOI provided by DaRUS.

4.2 Automatic Metadata for Publication on Machine-Learned Constitutive Relations

To evaluate the Harvester-Curator tool, we study a current research article within the EXC (Coltman et al. *Data-Driven Scale Bridging Parametrizations with Metrics* [4]) and its associated supplementary data. The article describes a machine learning approach to find constitutive relations for multi-scale problems – typical research within the EXC combining simulation and data science. Accordingly, the supplementary data "suffers" from complex metadata, for instance complex software dependencies combining simulation with data science software. Without automation, chances are poor that metadata ends up being documented at all.

With the Harvester, we first prepared a CodeMeta file from the private project folder, namely, a JSON file to describe software specification following CodeMeta schema[23]. We moreover exported the bibtex citation file from the archived paper [4]. Using both as input, we managed to successfully harvest the citation and software metadata from the associated data directory[24].

However, challenges emerged during attempts to extract process and EngMeta metadata, primarily due to the utilization of the JSON file format in the data directory. JSON's inherent flexibility poses a substantial obstacle for parsing methods, despite our efforts to align JSON file keys with metadata fields of DaRUS, yielding no relevant metadata and highlighting the constraints of a general JSON parser.

Presently, our efforts are thus concentrated on two key fronts: first, the design of a specialized parser tailored to JSON files containing simulation I/O data, and second, methodologies for distinguishing such JSON files from other JSON data formats. The curation phase currently remains untested for the use-case as we first continue to refine the process of harvesting process metadata and engineering metadata.

4.3 PDEBench

PDEBench is a repository[25,26,27] that provides a diverse and comprehensive set of benchmarks for scientific machine learning based on partial differential equations (PDEs), including realistic physical phenomena [13]. It offers pre-calculated simulation data as well as code to generate own data. Additionally, PDEBench provides publicly available pre-trained machine learning models that can be tailored to specific research purposes [11, 12, 15]. These models and benchmark datasets are stored in DaRUS datasets, and to make file and metadata retrieval more efficient, users can utilize scripts that use EasyDataverse as an interface to Dataverse.

Utilizing the adaptable features of Dataverse, PDEBench continuously curates and updates its datasets and pre-trained models through ongoing communication with users and the collaborative progress of the community. Large files are uploaded efficiently using the Python DVUploader library which supports direct upload to the DaRUS S3 object storage and bypasses the default Dataverse upload interface. In adddtion, direct uploads also facilitate large data files to be transmitted efficiently by splitting them into smaller files that can be uploaded asynchronously. Overall, the popularity of PDEBench datasets and pre-trained models, with over 50,000 downloads, highlights the usefulness of Dataverse as a database for researchers.

5 Conclusion and Outlook

The Cluster of Excellence Data-Integration Simulation Science (SimTech) conducts highly interdisciplinary research, as our exemplary domain applications in chemistry, engineering, and scientific machine learning demonstrate. The SimTech support structure on RDSM is invested in supporting SimTech researchers in the FAIRification of their data, software, and computational results.

The RDM infrastructure presented in this paper comprises a central data repository. It integrates several metadata schemata that cater to the different needs of SimTech researchers in terms of documenting their digital artifacts. To overcome the hurdle of gathering the necessary metadata and providing it in the right format for (automatic) upload, we have developed several tools, e.g., EasyDataverse and Harvester-Curator. Ensuring high quality artifacts and asso-

[21] http://sabio.h-its.org/.

[22] https://www.beilstein-strenda-db.org/strenda/.

[23] https://github.com/codemeta/codemeta.

[24] https://github.com/SimTech-Research-Data-Management/Harvester-Curator/blob/feature/ml-use-case/example/harvester_output.json.

[25] https://github.com/pdebench/PDEBench.

[26] https://darus.uni-stuttgart.de/dataset.xhtml?persistentId=doi:10.18419/darus-2986.

[27] https://darus.uni-stuttgart.de/dataset.xhtml?persistentId=doi:10.18419/darus-2987.

ciated metadata requires quality control, which we support by the EasyReview tool.

While this article primarily focuses on research data, developing FAIR research *software*, and, particularly, ensuring the reproducibility of computational results exhibits unique challenges. In [7], we describe a sustainable infrastructure concept for improved accessibility, reusability, and archival of research software. We identify research software artifacts as snapshots in the software development, research, and publication cycle, aiming to enhance archival, access, and reuse through installable, configurable, and extensible research software on open-access infrastructure.

The development of research software management in SimTech will be tightly integrated with the RDM infrastructure to offer a coherent and "one-stop-solution" for joint RDSM in the future. Next steps towards this goal include the following. Based on our experience gained in exemplary domain applications, we first plan to generalize, further simplify, and possibly automate the processes put in place. This requires a continuation of our efforts in developing or integrating libraries and interfaces that fit simulation scientists' diverse research eco-systems. Beyond this, our work is driven by the goal of a one-click reproduction of simulation results (which combines both data and software) originating from our EXC. In terms of RDSM infrastructure, a crucial ingredient will be the extension of DaRUS to allow a proper archival of container images, a standard unit of software that packages up code and all its dependencies so that the application runs quickly and reliably from one computing environment to another. An interested party can use the containerized environment to run and modify the research codes corresponding to a dataset. Together with the FoKUS team, we plan to host a container image registry and a container metadata block in DaRUS. The container metadata block will be able to capture the run-time environment when the image was built. For the backend, a tool will be developed to run containers from the image and its metadata. Wherever possible and appropriate, we will integrate services made available through the NFDI.

Our contributions on quality control will be brought to a broader audience through a Data Overlay Journal.

Providing RDM infrastructure, tools, and services, and supporting RDSM activities of SimTech researchers covers two of the four measures we pursue to foster RDSM, as outlined in Sect. 1. Indeed, to broaden its adoption, flanking measures such as enablement and exchange, as well as dissemination and participation in overarching RDM activities are essential. We plan to pursue our activities in these areas.

We are convinced that the efforts we invest in RDSM and the solutions we contribute to the SimTech community and beyond represent an important cornerstone to not only safeguard the integrity and accessibility of relevant data and software but also exemplify a commitment to excellence in scholarly endeavors, paving the way for robust and reproducible research outcomes.

Acknowledgements Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – project number 390740016 and, as part of Germany's National Research Data Infrastructure – NFDI 9 – project number 442146713.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Anzt H, Bach F, Druskat S et al (2021) An environment for sustainable research software in germany and beyond: current state, open challenges, and call for action. F1000Res. https://doi.org/10.12688/f1000research.23224.2
2. Barker M, Chue HNP, Katz DS et al (2022) Introducing the FAIR principles for research software. Sci Data 9(1):622. https://doi.org/10.1038/s41597-022-01710-x
3. Brecher C, Buchmeiser MR, Burkert A et al (2022) Commitment zu aktivem Daten- und -softwaremanagement in großen Forschungsverbünden: Commitment to active data and software management in large research alliances. Bausteine Forschungsdatenmanage. https://doi.org/10.17192/bfdm.2022.1.8412
4. Coltman E, Schneider M, Helmig R (2023) Data-driven scale bridging parametrizations with metrics: Dispersive transport. 2311.13975
5. Crosas M (2011) The dataverse network®: an open-source application for sharing, discovering and preserving data. D Lib Mag 17(1):2. https://doi.org/10.1045/january2011-crosas
6. Flemisch B, Hermann S, Holm C et al (2020) Umgang mit Forschungssoftware an der Universität Stuttgart: Managing research software at the University of Stuttgart https://doi.org/10.18419/opus-11178
7. Koch T, Gläser D, Seeland A et al (2023) A sustainable infrastructure concept for improved accessibility, reusability, and archival of research software https://doi.org/10.48550/arXiv.2301.12830
8. Lauterbach S, Dienhart H, Range J et al (2023) Enzymeml: seamless data flow and modeling of enzymatic data. Nat Methods 20(3):400–402. https://doi.org/10.1038/S41592-022-01763-1
9. Range J, Halupczok C, Lohmann J et al (2022) Enzymeml—a data exchange format for biocatalysis and enzymology. FEBS J 289(19):5864–5874. https://doi.org/10.1111/FEBS.16318
10. Schembera B, Iglezakis D (2020) EngMeta: metadata for computational engineering. Int J Metadata Semant Ontol 14(1):26. https://doi.org/10.1504/ijmso.2020.107792

11. Sharma P, Chung WT, Akoush B et al (2023) A review of physics-informed machine learning in fluid mechanics. Energies. https://doi.org/10.3390/en16052343
12. Takamoto M, Alesiani F, Niepert M (2023) Learning neural pde solvers with parameter-guided channel attention. 2304.14118
13. Takamoto M, Praditia T, Leiteritz R et al (2023) Pdebench: an extensive benchmark for scientific machine learning. 2210.07182
14. Wilkinson MD, Dumontier M, Aalbersberg IJ et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3(1):1–9. https://doi.org/10.1038/sdata.2016.18
15. Zhang X, Wang L, Helwig J et al (2023) Artificial intelligence for science in quantum, atomistic, and continuum systems. 2307.08423