An Interdisciplinary Approach to Manage Materials Data with Kadi4Mat and Chemotion

Patrick Altschuh[2], Stefan Bräse[1c,1e], Thomas Hartmann[3], Doris Jaeger[1f], Nicole Jung[1c,1e], Arnd Koeppe[1b,1d], Peter Krauss[1g], Carolin Leister[1f], Britta Nestler[1b,1d,2], Gunther Schiefer[1a], Clemens Schreiber[1a], Michael Selzer[1d,2], Martin Starman[1c], Giovanna Tosato[1b]

[1a]Institute for Applied Informatics and Formal Description Methods (AIFB);
[1b]Institute for Applied Materials (IAM-MMS);
[1c]Institute of Biological and Chemical Systems (IBCS-FMS);
[1d]Institute of Nanotechnology (INT);
[1e]Institute of Organic Chemistry (IOC);
[1f]KIT Library;
[1g]Steinbuch Centre for Computing (SCC);
[1]Karlsruhe Institute of Technology (KIT);
[2]Institute of Digital Materials Science (IDM), University of Applied Sciences Karlsruhe (HKA);
[3]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure (FIZ)

Interdisciplinary approaches and diverse research data management tools are required to design and fabricate new materials with macroscopically observable properties based on changes at the molecular level. The Science Data Center MoMaF is developing strategies to enable research data management across scales using the Chemotion and Kadi4Mat RDM tools. The study presents a use-case concept showing how both tools can be used conjointly to record molecular descriptions and manage simulations of microstructures across scales. The analysis of completed projects yields a concept for future processes, emphasizing the importance of efficient and consistent research and documentation across disciplines. The conjoint use of different RDM tools bridges the gaps between research fields, such as chemistry and materials science, and pushes the frontiers of interdisciplinary research.

1 Introduction

Developing new materials with specific properties is crucial for many technological advancements. However, designing and fabricating such materials is a highly interdisciplinary task that requires diverse expertise and skills from various research backgrounds.

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18086

The creation of new materials with macroscopically observable properties is based on changes that occur at the molecular level, making the research projects highly complex and challenging. One of the major obstacles in this regard is the efficient execution and documentation of the research, which requires research data management (RDM) tools. However, as RDM tools are often specialized in specific research areas or focus on a subset of RDM tasks such as Electronic Lab Notebooks (ELNs; CARPi, Minges, and Piel 2017) or repositories for long-term data storage (European Organization For Nuclear Research and OpenAIRE 2013), they may not provide optimal solutions for interdisciplinary topics and solutions for evolving "warm" data. Therefore, using different RDM tools together is necessary to ensure consistent research and documentation across disciplines.

The Science Data Center for Molecular Materials Research (MoMaF) is developing strategies for research data management across scales, focusing on the conjoint use of RDM tools to enable work across disciplines. Chemotion (Tremouilhac et al. 2017) and Kadi4Mat (Brandt et al. 2021) are examples of research tools that cover research at the molecular, meso- and macroscopic scales. Both systems are being extended within the Science Data Center to enable the interoperable use of the systems for work across scales.

This study proposes a strategy for the RDM tools Chemotion and Kadi4Mat to conjointly record molecular descriptions, polymerization reactions, experimental outcomes, and properties. The study analyzes the procedure and documentation methods of already completed projects to propose a concept for future processes. The Chemotion ELN records necessary parameters at the molecular level, which can then be managed and transferred to the Kadi ecosystem as input for microstructure simulations that can model, e.g., time-dependent phase separation processes. Finally, the study outlines how analysis tools on time-dependent data can derive macroscopic properties as a function of the molecular composition via Kadi4Mat. This study highlights the importance of interdisciplinary approaches and the collaborative use of RDM tools for efficient and consistent research and documentation across disciplines and scales.

2 Use-case concept for interdisciplinary research data management

3D printing has revolutionized the fabrication of complex polymer objects, but limitations exist for large-scale objects with small-scale geometrical features. Porous structures at the sub-micrometer scale are essential for various applications such as sensing, separation, and biomedical applications.

Nanoporous materials have been widely studied and utilized due to their unique properties, such as high surface area, low density, and tunable porosity. For that aim, Dong et al. (2021) proposed a novel approach for 3D printing nanoporous polymers using Polymerization-Induced Phase Separation (PIPS). This approach combines digital-light-processing 3D printing with PIPS to manufacture hierarchical polymer structures that exhibit defined macroscopic geometries and tunable porosity at the micro- and nanometer scales. The produced hierarchical polymers show improved adsorption performance,

cell adhesion, and growth due to surface porosity, making them suitable for various application scales from 10 nm up to 1000 μm (ibid.).

In the following application of the interoperation strategy, we leverage the synergies of Chemotion and Kadi4Mat to digitalize and automatize data management, processing, and analysis for polymerization in 3D printing. The digital footprint of the use-case encompasses data entities with associated metadata in the Kadi4Mat and Chemotion repositories, as well as workflow entities that can execute experimental and numerical studies, record results, and communicate through APIs. Figure 1 visualizes how the two RDM infrastructures can communicate through the interoperation strategy, where RESTful APIs enable authentication, querying, and data exchange. The level of integration for each process (rounded boxes), from user execution, through scripted execution, to full integration, describes how much user interaction is required and how automatized the exchange of data and metadata functions. Initially, data at the molecular description level can be saved in the Chemotion repository. Using the interoperation strategy between the two RDM systems, users can authenticate in both systems, and requests from Kadi can be sent to Chemotion for molecular descriptions. Both processes are realizable through KadiStudio's fully-integrated workflow engine and/or short scripts that perform , e.g., API requests. From the delivered molecular data, the mesoscale simulation can be executed within the implemented KadiStudio workflow, which executes all the domain-specific computations and yields macroscopic properties recorded as structure data in the Kadi4Mat repository. Finally, the results can be mirrored back to Chemotion, where the data is structured with fixed templates based on the Chemical Methods Ontology.

At the molecular level, we focus on the chemical properties of the solvents and describe how Chemotion pushes experimental research forward. Chemotion is a powerful system for gathering and managing scientific data with several advantages. Firstly, it provides an easy and standardized way to store and share experimental data according to the FAIR (Findable, Accessible, Interoperable, Reusable) principles. Researchers can quickly access and share their data with others, increasing collaboration and transparency. Chemotion offers powerful tools for the work with data on molecular structures and facilitates gaining information on chemical entities without further databases or search engines. Data that deal with molecules and reactions is processed in a discipline-specific manner and can be represented, enriched, or used for further calculations without the need for additional software or tools. Chemotion allows researchers to track their data over time, making it easier to identify trends and patterns in their research. Finally, Chemotion offers secure and reliable data storage and management, ensuring that researchers' data is protected and can be accessed and used for future publications. The data supporting a publication can be easily published on the Chemotion repository (Bräse 2023). For more details on Chemotion, see Tremouilhac et al. (2017) and Tremouilhac et al. (2021). Chemotion excels in managing experimental chemical data, particularly at the molecular level, but the results of numerical simulation methods as in Dong et al. (2021) are also relevant to be captured. These simulations depend on experimentally collected data such as diffusivity, density, viscosity, and surface tension of the solvents, which can be made available in a structured and findable way in Chemotion via a RESTful API.

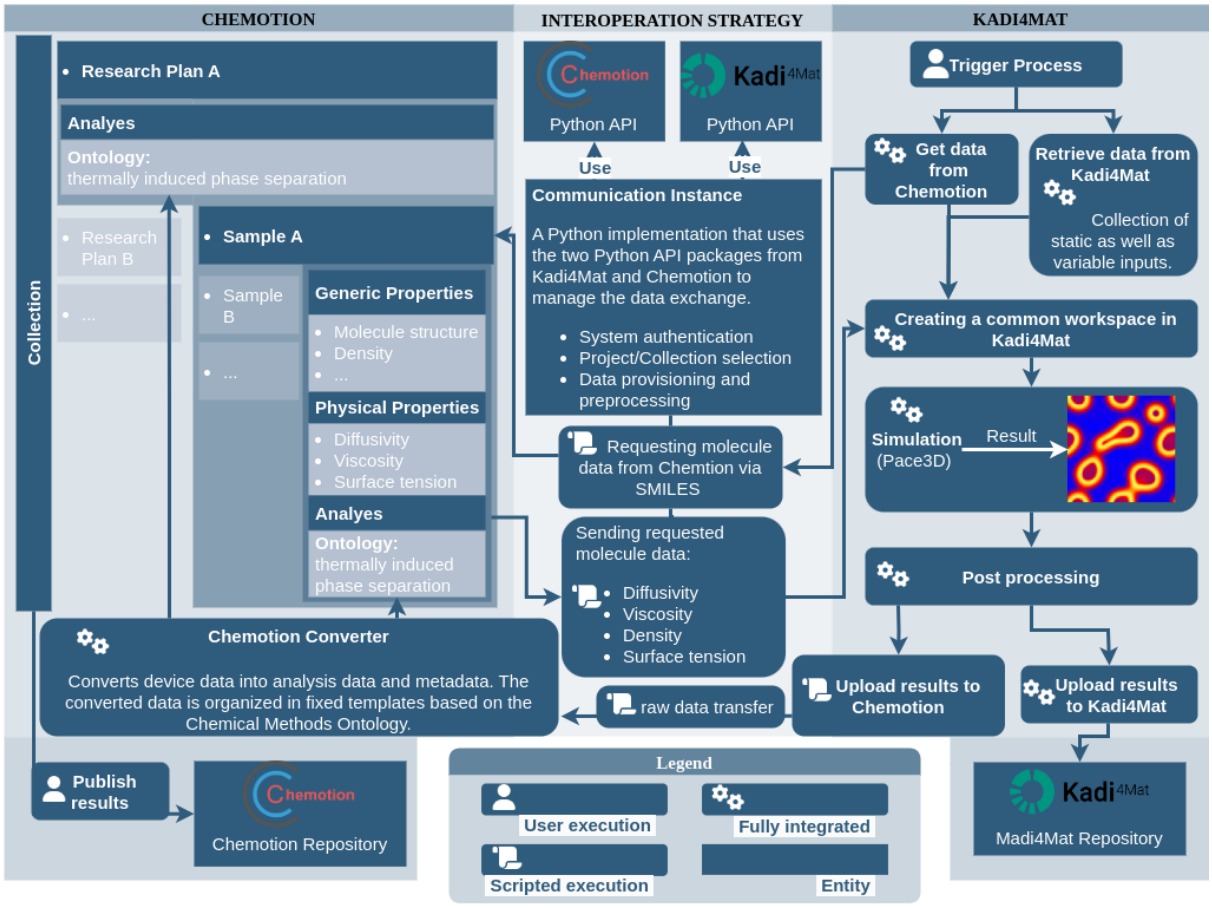

Figure 1: Interaction strategy between the Kadi4Mat and Chemotion RDM infrastructures. The legend explains the different levels of integration for each process, from user execution, via scripted execution, to full integration.

At the mesoscale, Dong et al. (2021) use Kadi4Mat to manage the extensive data accumulated from the simulations. The Kadi ecosystem provides flexible and intuitive solutions to manage data and automatize research, thereby focusing on warm and linked data that are actively used while they keep evolving. Kadi's generic implementation realizes powerful methods and tools to produce, manage, analyze, search, and share data according to the FAIR principles. For this use-case concept, the workflow engine KadiStudio (Griem et al. 2022) can automatize the microstructure simulation study, synchronize the results with internal warm-data repositories, and archive them in research data repositories. Data are accumulated from Chemotion and Kadi4Mat into a common workspace. Subsequently, the numerical solver Pace3D produces simulation results of (3D+t) microstructure evolutions, which are analyzed for their macroscopic properties during postprocessing. Finally, the results are uploaded to Kadi4Mat through its REST-like API and mirrored to Chemotion. This workflow-centric ELN 2.0 approach enables an RDM that automatizes and expedites research.

3 Conclusion and future developments

The design and manufacturing of new materials with specific properties require interdisciplinary approaches and diverse RDM tools. The Science Data Center MoMaF is developing strategies for research data management across scales. The presented study shows how Chemotion and Kadi4Mat can be used conjointly to document molecular descriptions and manage microstructure simulation processes. A concept for future joint applications for material development is proposed, and the importance of efficient and consistent research and documentation across disciplines is highlighted. The collaborative use of RDM tools enables the derivation of macroscopic properties across scales as a function of the molecular composition, which enables the development of new materials with specific properties and advancements in various technological fields.

Current trends in natural and engineering sciences strongly favor efficient data analysis methods, such as machine learning and artificial intelligence, which can analyze, describe, and enrich interdisciplinary research data. In this regard, large language models are capable of understanding and interpreting chemical notation (White et al. 2023). In the Kadi ecosystem, these methods evolve through the KadiAI interface and CIDS (Computational Intelligence and Data Science) framework (Koeppe and CIDS Team 2023). Independently from the platform these methods use, we plan to provide these methods to support researchers with the extraction and aggregation of relevant research data. For Chemotion, the goal is to become a tool that supports scientists from data acquisition to data analysis and publication for all disciplines that refer to molecular information. To achieve this goal, interfaces to access databases and further research tools are evaluated, and efforts are being made to standardize the data exchange among ELNs to enhance the interoperability of common RDM software. To train researchers and students in RDM and ELNs, we develop courses covering Kadi4Mat and Chemotion. The service team RDM@KIT guides and trains researchers from all disciplines (Serviceteam RDM@KIT 2023) and bring RDM to the broader scientific community.

Acknowledgements

This work is funded by the Ministry of Science, Research and Art Baden-Württemberg (MWK-BW) in the Science Data Center MoMaF, with funds from the state digitization strategy digital@bw (project number 57), the BMBF and MWK-BW as part of the Excellence Strategy of the German Federal and State Governments in the project Kadi4X and the support of the Karlsruhe Nano Micro Facility (KNMFi, `www.knmf.kit.edu`), a Helmholtz Research Infrastructure at Karlsruhe Institute of Technology within the program MSE, no. 43.31.01.

References

Brandt, Nico, Lars Griem, Christoph Herrmann, Ephraim Schoof, Giovanna Tosato, Yinghan Zhao, Philipp Zschumme, and Michael Selzer. 2021. *Kadi4Mat: A Research Data Infrastructure for Materials Science.* 20:8. 1. Ubiquity Press. DOI: https://doi.org/10.5334/dsj-2021-008.

Bräse, Stefan. 2023. "Chemotion Website". Visited on May 12, 2023. https://www.chemotion-repository.net/welcome.

CARPi, Nicolas, Alexander Minges, and Matthieu Piel. 2017. "eLabFTW: An open source laboratory notebook for research labs". *The Journal of Open Source Software* 2 (12): 146. ISSN: 2475-9066. DOI: https://doi.org/10.21105/joss.00146.

Dong, Zheqin, Haijun Cui, Haodong Zhang, Fei Wang, Xiang Zhan, Frederik Mayer, Britta Nestler, Martin Wegener, and Pavel A. Levkin. 2021. "3D Printing of Inherently Nanoporous Polymers via Polymerization-Induced Phase Separation". *Nature Communications* 12 (1): 247. ISSN: 2041-1723. DOI: https://doi.org/10.1038/s41467-020-20498-1.

European Organization For Nuclear Research and OpenAIRE. 2013. *Zenodo.* DOI: https://doi.org/10.25495/7GXK-RD71.

Griem, Lars, Philipp Zschumme, Matthieu Laqua, Nico Brandt, Ephraim Schoof, Patrick Altschuh, and Michael Selzer. 2022. "KadiStudio: FAIR Modelling of Scientific Research Processes". *Data Science Journal* 21 (1): 16. ISSN: 1683-1470. DOI: https://doi.org/10.5334/dsj-2022-016.

Koeppe, Arnd, and CIDS Team. 2023. *CIDS: 3.1.* Zenodo. Visited on January 11, 2023. DOI: https://doi.org/10.5281/zenodo.7524476.

Serviceteam RDM@KIT. 2023. "Train & Edu". Visited on May 10, 2023. https://www.rdm.kit.edu/train-edu.php.

Tremouilhac, Pierre, Pei-Chi Huang, Chia-Lin Lin, Yu-Chieh Huang, An Nguyen, Nicole Jung, Felix Bach, and Stefan Bräse. 2021. "Chemotion Repository, a Curated Repository for Reaction Information and Analytical Data". *Chemistry–Methods* 1 (1): 8–11. ISSN: 2628-9725. DOI: https://doi.org/10.1002/cmtd.202000034.

Tremouilhac, Pierre, An Nguyen, Yu-Chieh Huang, Serhii Kotov, Dominic Sebastian Lütjohann, Florian Hübsch, Nicole Jung, and Stefan Bräse. 2017. "Chemotion ELN: an Open Source electronic lab notebook for chemists in academia". *Journal of Cheminformatics* 9 (1): 54. DOI: https://doi.org/10.1186/s13321-017-0240-0.

White, Andrew D., Glen M. Hocky, Heta A. Gandhi, Mehrad Ansari, Sam Cox, Geemi P. Wellawatte, Subarna Sasmal, et al. 2023. "Assessment of chemistry knowledge in large language models that generate code". *Digital Discovery* 2 (2): 368–376. DOI: https://doi.org/10.1039/D2DD00087C.