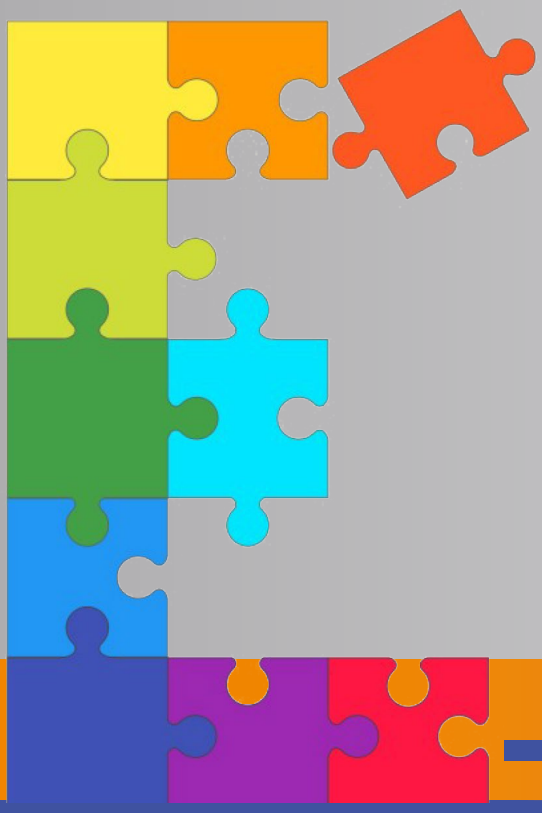


Vincent Heuveline
Nina Bisheh
Philipp Kling
(Hrsg.)

-Science-Tage 2023

Empower Your Research – Preserve Your Data

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

E-Science-Tage 2023

E-Science-Tage 2023

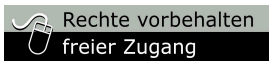
Empower Your Research – Preserve Your Data

Herausgegeben von

Vincent Heuveline, Nina Bisheh und Philipp Kling

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.



Publiziert bei heiBOOKS, 2023

Universität Heidelberg / Universitätsbibliothek
heiBOOKS
Grabengasse 1, 69117 Heidelberg
https://books.ub.uni-heidelberg.de/heibooks

Die Online-Version dieser Publikation ist auf heiBOOKS,
der E-Book-Plattform der Universitätsbibliothek Heidelberg,
https://books.ub.uni-heidelberg.de/heibooks, dauerhaft frei verfügbar
(Open Access).
urn: urn:nbn:de:bsz:16-heibooks-book-1288-4
doi: https://doi.org/10.11588/heibooks.1288

ISBN 978-3-948083-91-5 (Softcover)
ISBN 978-3-948083-90-8 (PDF)

Inhaltsverzeichnis

I Grußworte 9

Grußwort der Ministerin für Wissenschaft, Forschung und Kunst Baden-Württemberg
Petra Olschowski 11

Grußwort des Prorektors der Universität Heidelberg
Matthias Weidemüller 13

Vorwort der Herausgeber
Vincent Heuveline, Nina Bisheh, Philipp Kling 15

II Wissenschaftliche Beiträge 17

A Reproducible Machine Learning Workflow to Characterize the Solid Electrolyte Interphase
Deepalaxmi Rajagopal, Arnd Koeppe, Meysam Esmaeilpour, Michael Selzer, Helge Stein, Britta Nestler 19

Linking Domain-specific RDM to Institutional and Generic Approaches – the Case of NFDI4Biodiversity
Jimena Linares, Barbara Ebert, Judith Sophie Engel 32

Breaking Down Hurdles of Current Data Citation Practices. Use Cases and Benefits of Persistent Identifiers for Dataset Elements
Janete Saldanha Bach, Claus-Peter Klas, Peter Mutschke 41

Datenmanagementplan und Publikation von Forschungsdaten im Projekt „Emissionsminderung Nutztierhaltung“ EmiMin: Planung und Realität – Umsetzbarkeit von Forschungsdatenmanagement
Ewald Grimm, Birte Lindstädt, Katrin Wagner, Roman Riedel 56

Automating DOI Registration with DataCite API
Giuditta Parolini, Falko Glöckler 60

Research Data Policies in Scientific Journals – a Case Study
Gertraud Novotny, Thomas Seyffertitz 73

DataPLANT – Harnessing the Power of Ontologies for FAIR Research Data Management
Kathryn Dumschott, Hannah Dörpholz, Kevin Frey, Marcel Tschöpe, Heinrich Lukas Weil, Timo Mühlhaus, Dirk von Suchodoletz, Björn Usadel, Angela Kranz 89

Data Repositories 4Culture – Bedarfsorientierte Forschungsdatenrepositorien für den Kulturbereich
Alexandra Büttner, Sandra Göller, Peggy Große, Kerstin Soltau 101

Herausforderungen beim Aufbau eines föderierten Datenrepositoriums auf Basis von InvenioRDM
Dirk von Suchodoletz, Jonathan Bauer, Marcel Tschöpe, Holger Gauza, Michael Derntl, Steve Kaminski 116

Datensammlung in der Romanistik – Eine Analyse von Normierung und Standardisierung in E-Mails
Laura Bothe, Sybille Große 132

Carrots and Sticks: Motivating with Storage for Good RDM – Science Led Allocation of Research Data Storage Resources within an Integrated RDM System
Ilona Lang, Marcel Nellesen, Lukas C. Bossert, Marius Politze 140

Cat4KIT: A Cross-institutional Data Catalog Framework for the FAIRification of Environmental Research Data
Mostafa Hadizadeh, Christof Lorenz, Sabine Barthlott, Romy Fösig, Uğur Çayoğlu, Robert Ulrich, Felix Bach 149

bwVisu: A Scalable Remote Service for Interactive Data Processing and Training for Scientists
Erik Schnetter, Carlo Antonio Beretta, Martin Baumann, Sabine Richling, Florian Heuschkel, Thomas Kuner 161

Wege aus der Verantwortungsdiffusion – Vermittelnde Angebote des Forschungsdatenmanagements zwischen Top-Down und Bottom-Up
Jan Leendertse, Dirk von Suchodoletz, Saher Semaan 174

How to Choose a Research Data Repository Software? Experience Report
Nina Buck,Volodymyr Kushnarenko, Björn Schembera, Mona Ulrich, Heinz Werner Kramski, Andreas Ganzenmüller, Jan Hess, Alexander Holz, André Blessing, Pascal Hein, Kerstin Jung, Nicolas Schenk, Claus-Michael Schlesinger, Thomas Bönisch, Roland S. Kamzelak, Jonas Kuhn, Gabriel Viehhauser 188

Quo venis? Metadata for Common Scientific ASCII Files
Muhammed Bayram, Frank Tristram 197

A Machine-actionable Workflow for the Publication of Climate Impact Research Data from the ISIMIP Project
Jochen Klar, Matthias Mengel 201

Empowering Data at Leeds Beckett University: Understanding Institutional Needs and Applying Best Practice
Amy Campbell 208

NFDI4DS – NFDI for Data Science and Artificial Intelligence
Sonja Schimmler 215

Mit maßgeschneiderten Metadatenprofilen zu validierten und nachhaltigen Forschungsdaten
Matthias Grönewald, Nils Preuß 220

Automated Software Metadata Conversion and Publication Based on CodeMeta
Marie Houillon, Jochen Klar, Tomas Stary, Axel Loewe 228

Reifegradmodell für die Verwaltung des Datenzugriffs
Max Leo Wawer, Roland Lachmayer 235

Das Data Science Center an der Universität Bremen: Interdisziplinärer Knotenpunkt und Service-Infrastruktur für die datenintensive Forschung
Lena Steinmann, Heike Thöricht, Sandra Zänkert, Rolf Drechsler 246

Leibniz Data Manager – Data Management Across Various Research Data Repositories
Angelina Kraft, Anna Beer, Mauricio Brunet, Ahmad Sakor, Maria-Esther Vidal 253

Schöne neue Laborwelt – Elektronische Laborbücher digitalisieren die Labordokumentation
Bert Zulauf, Nina Knipprath 258

An Interdisciplinary Approach to Manage Materials Data with Kadi4Mat and Chemotion
Patrick Altschuh, Stefan Bräse, Thomas Hartmann, Doris Jaeger, Nicole Jung, Arnd Koeppe, Peter Krauss, Carolin Leister, Britta Nestler, Gunther Schiefer, Clemens Schreiber, Michael Selzer, Martin Starman, Giovanna Tosato 264

Stärkung von FDM-Services im Verbund – Ergebnisse einer Bedarfserhebung
Angela Ariza de Schellenberger, Evgeny Bobrov, Kerstin Helbig, Denise Jäckel, Monika Kuberek, Lea-Sophie Orozco Prado, Elisabeth Maria Schlagberger, Sibylle Söring, Britta Steinke 270

Ein Werkzeug zur XSD-basierten Metadatenannotation
Olaf Brandt, Holger Gauza, Jan Kaltenbach, Maximilian E. Müller, Gabriel Schneider, Claus Zinn 276

Standardized Metadata Collection to Reinforce Collaboration in Collaborative Research Centers
Manuel Watter, Laura Kahle, Birger Brunswiek, Urs A. Fichtner, Michelle Pfaffenlehner, Frank Werner, Denis Gebele, Harald Binder, Jochen Knaus . 282

Bringing FAIR Bioimage Data Management into Practice: the Information Infrastructure for BioImage Data (I3D:bio) Project – bottom-up Community Support for Microscopy Data Sharing and Preservation.
Christian Schmidt, Michele Bortolomeazzi, Tom Boissonnet, Julia Dohle, Tobias Wernet, Janina Hanne, Roland Nitschke, Susanne Kunis, Karen Bernhardt, Stefanie Weidtkamp-Peters, Elisa Ferrando-May 289

Implementation of an InfraStructure for dAta-BasEd Learning in environmental sciences (ISABEL)
Marcus Strobl, Elnaz Azmi, Balazs Bischof, Alexander Dolich, Sibylle K. Hassler, Mirko Mälicke, Ashish Manoj Jaseetha, Jörg Meyer, Achim Streit, Erwin Zehe 295

Data Competence for Photonic Nanotechnologies
Jörg Meyer, Nigar Asadova, Dominik Beutel, Uğur Çayoğlu, Carsten Rockstuhl, Frank Tristram 301

Bayesian Optimization Framework for Data-driven Materials Design
Giovanna Tosato, Arnd Koeppe, Bai-Xiang Xu, Michael Selzer, Britta Nestler 306

Veranstalter
. 313

Teil I

Grußworte

Grußwort der Ministerin für Wissenschaft, Forschung und Kunst Baden-Württemberg

Petra Olschowski

Auf beeindruckende und sehr unmittelbare Weise erleben wir tagtäglich in unterschiedlichen Lebensumfeldern, welche Rolle Daten in modernen Gesellschaften spielen. Ein Beispiel, das uns allen noch präsent ist, war die Pandemie. Hier haben wir die Bedeutung der Datenerhebung für die wissenschaftliche Erkenntnis wie für die tagtägliche Umsetzung von Schutzmaßnahmen gesehen – aber auch die Lücken und die Probleme bei der Interpretation von Daten. Die Nutzung von Gesundheitsdaten für medizinische Innovation und Public Health ist insgesamt eine unabweisbare Aufgabe auch hier im Land – und das ist nur ein Bereich von vielen.

Das Motto Ihrer Tagung „Empower Your Research - Preserve Your Data“ beinhaltet viele Aspekte. Es geht darum, wie die Speicherung und Sicherung von Forschungsdaten unterstützt und vorangebracht werden kann, und es geht um Forschungsdateninfrastrukturen, die Ihnen bei Ihrer Arbeit und Forschung nützlich sind, denn: Der systematische Zugang zu digitalen Datenbeständen wird für neue wissenschaftliche Erkenntnisse und damit für Innovationen und Technologietransfer immer wichtiger.

Zukunftsfelder wie Maschinelles Lernen oder Künstliche Intelligenz sind auf entsprechende Datengrundlagen angewiesen; die systematische Datensicherung und der kompetente Umgang mit riesigen Datenmengen spielt in der Forschung eine immer wichtigere Rolle. Die an unterschiedlichen Stellen auf verschiedenste Weise gesammelten Daten müssen so zugänglich gemacht werden, dass sie auch für Dritte unmittelbar und geordnet auffindbar sind. Ebenso müssen die Daten über die Grenzen einzelner Datenbanken, Fachdisziplinen und Länder hinweg analysiert und verbunden werden können.

Starke Forschung bei gleichzeitiger Datensicherheit: In den vergangenen Jahren haben die Landesregierung und das Wissenschaftsministerium in den Auf- und Ausbau gemeinsamer Forschungsdateninfrastrukturen an den Universitäten und Forschungseinrichtungen des Landes investiert. Bereits seit 2019 fördert das Wissenschaftsministerium im Rahmen der Landesdigitalisierungsstrategie den Aufbau von vier leistungsstarken Forschungsdatenzentren – die Science Data Center – mit insgesamt acht Millionen Euro. Für den Technologiestandort Baden-Württemberg sind leistungsstarke Forschungsdatenzentren von herausragender Bedeutung: Mit dem Cyber Valley in Stuttgart und Tübingen konnte Europas größtes Forschungskonsortium im Bereich Künstliche Intelligenz aufgebaut werden.

Das baden-württembergische Begleit- und Weiterentwicklungsprojekt für Forschungsdatenmanagement (bw2FDM) unterstützt nicht zuletzt mit der Organisation der E-Science-Tage den bundesweiten Wissens- und Erfahrungsaustausch.

Aktuell entwickeln wir die Landesstrategie Forschungsdaten, um diesen Bereich insgesamt auf ein Niveau zu heben, wie wir es im Bereich des High Performance Computing bereits gewohnt sind. Dabei macht Wissenschaft jedoch nicht an Ländergrenzen halt. Ein wichtiger Meilenstein ist der Aufbau der Nationalen Forschungsdateninfrastruktur (NFDI), die gemeinsam von Bund und Ländern finanziert wird. Die vier baden-württembergischen Science Data Center sind Teil dieser Infrastruktur. Mit der NFDI sorgen wir aber auch für europäische und internationale Anschlussfähigkeit, beispielsweise an die European Open Science Cloud.

Das gemeinsame Ziel muss es sein, die verschiedenen Initiativen auf Landes-, nationaler und internationaler Ebene zusammen zu denken und zu harmonisieren, denn in einem derart dynamischen Wachstumsbereich wie den Forschungsdaten bedeutet Stillstand mittelfristig einen quantitativen und qualitativen Rückschritt. Wir wollen in Baden-Württemberg gemeinsam mit den Partnern auf bundes-, europäischer-, und internationaler Ebene eine Dateninfrastruktur aufbauen und weiterentwickeln, die dem Bedarf der Wissenschaft heute und in Zukunft entspricht.

Ich freue mich daher, dass bei dieser Tagung sowohl das KIT, die Universität Konstanz als auch die Universität Heidelberg engagiert dabei sind und dass wir als Wissenschaftsministerium diese Konferenz unterstützen können! Mein Dank geht dabei insbesondere an das Organisationsteam der diesjährigen E-Science-Tage für Ihr besonderes Engagement.

Vielleicht lässt sich nicht auf alle Fragen im Rahmen dieser Tagung eine Antwort finden, zumindest erhoffe ich mir aber, dass Sie für Ihre Arbeit an den Universitäten und Instituten neue Impulse gewinnen und die Vernetzung ein Stück weit vorantreiben können.

Petra Olschowski MdL
Ministerin für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg

Grußwort des Prorektors der Universität Heidelberg

Matthias Weidemüller

Liebe Teilnehmer der E-Science-Tage 2023, lieber Vincent,

es ist mir eine ganz besondere Freude, Sie alle zu den diesjährigen E-Science-Tagen in der Neuen Aula der Universität Heidelberg zu begrüßen. Und dies nicht nur in meiner Funktion als Prorektor für Innovation und Transfer der Universität, sondern auch als Quantenphysiker, der um die Bedeutung von Datenmanagement und der adäquaten Weiterverwendung von Primärdaten nur allzu gut weiß.

Mein Dank geht an alle an der Organisation dieser E-Science-Tage Beteiligte, insbesondere an Nina Bisheh und Vincent Heuveline.

Die Universität Heidelberg hat dem Wissens- und Technologietransfer als „Dritte Mission" neben Forschung und Lehre einen besonderen Stellenwert in ihrer strategischen Ausrichtung zugewiesen. Dies manifestiert sich nicht nur an der Einrichtung eines eigenen Prorektorats für dieses wichtige Thema, sondern auch in der Schaffung ganz neuer Strukturen innerhalb der Universität. Da ist zum einen die Stabsstelle heiINNOVATION, divers und multidisziplinär besetzt mit mittlerweile 13 Mitarbeiterinnen und Mitarbeitern. Diese Stabsstelle ist dezentral organisiert mit Shared Offices, um dort zu sein, wo wir gebraucht werden. Hier kümmert man sich um alle Belange der Förderung von Innovationsprojekten, um die Beratung von Gründungsinteressierten oder der Förderung von Ideen mit gesellschaftlichem Impact. Wir scouten aktiv auf dem Campus, um Hot Spots für den Wissens- und Technologietransfer zu identifizieren, oder um Akteure auf diesen Gebieten miteinander zu vernetzen. Und wir bieten Weiterbildungsprogramme für alle Mitglieder unserer akademischen Gemeinschaft an, u.a. im Rahmen des in diesem Wintersemester erfolgreich angelaufenen Zertifikats zum Thema „Entrepreneurial Skills". Parallel hierzu haben wir eine Verwertungsgesellschaft gegründet, die Science Value Heidelberg GmbH. Hier werden alle Patente und Lizensierungsverträge der Universität verhandelt und vorbereitet.

heiINNOVATION und die SVH GmbH arbeiten Hand in Hand, und diese Aktivitäten tragen Früchte. Wir erwirtschaften Einnahmen im zweistelligen Millionenbereich durch die Verwertung von Patenten und Lizenzen, auch durch zwei große Abschlüsse aus den Lebenswissenschaften und der Medizin in den vergangenen zwei Jahren. Die Anzahl der Erfindungsmeldungen steigt. Und, und dies erfreut mich besonders: Wir unterstützen zunehmend Transferprojekte aus den Sozial- und Geisteswissenschaften, die große gesellschaftliche Wirkung entfalten.

Erlauben Sie mir noch einige finale Bemerkungen zur Bedeutung von Daten und ihrer Nutzung für die Transferaktivitäten in die Gesellschaft. Die Universität Heidelberg besitzt einen wahren Datenschatz, der sich über alle Disziplinen und Fachgebiete erstreckt. Diese Daten werden - aus meiner Sicht - derzeit noch völlig unzureichend genutzt, um damit einen über die Wissenschaft hinausgehenden gesellschaftlichen Nutzen zu bewirken. Hierfür gibt es aus meiner Sicht zwei entscheidende Gründe:

- Zum einen sind Rohdaten ohne Wert. Sie werden erst nutzbar durch die Veredelung, entweder durch adäquate Kategorisierung und Kuratierung, oder aber durch geeignete Modelle, die den Daten und ihren tieferliegenden Korrelationen einen übergeordnete Bedeutung geben. Nehmen Sie als Beispiel die Daten der Planetenbewegungen durch Tycho Brahe. Zum einen waren diese Daten auf das Feinste annotiert, was mindestens einen ebenso großen Arbeitsaufwand bedeutete wie das Messen der Daten selbst. Zum anderen wären diese Daten nur von sehr begrenztem Wert gewesen und hätten sicherlich keinerlei Speicherung von 16. bis in das heutige Jahrhundert gerechtfertigt, hätte nicht Johannes Kepler erkannt, dass die Daten in einem heliozentrischen Bezugssystem einfache Ellipsenbahnen der Planeten beschreiben. Der enorme Aufwand, der für die Kuratierung einerseits, für die Modellbildung andererseits erforderlich ist, und durch den die Daten erst einen wirklichen Wert erlangen, wird in der gegenwärtigen öffentlichen Diskussion um die Speicherung und Verwendung von Daten in meiner Wahrnehmung weit unterschätzt, vor allem auch in Bezug auf die Ressourcen, die hierfür notwendig sind und finanziert werden müssen.
- Zum anderen benötigen wir rechtliche Experimentierräume jenseits vom reinen „Datenschutz", um „veredelte" Daten so zu aufzuarbeiten, dass sie größtmöglichen Nutzen entfalten können. In solchen gesicherten Räumen muss es möglich sein, Neues auszuprobieren, ohne dass ein negatives Ergebnis gleich als Scheitern angesehen wird (hier spricht der Experimentalphysik), und dies jenseits der Datennutzungspraktiken im Silicon Valley und in Shenzhen. Ich denke hierbei z.B. an die Nutzung von medizinischen Daten, wie sich in großem Umfang an der Universität Heidelberg vorliegen, die aber nicht in ausreichendem Umfang zugänglich gemacht werden können für die Prävention, die Diagnostik oder die Therapie.

Wenn ich auf das Programm schaue, bin ich sicher, dass auch diese Themen auf den E-Science-Tage eine wichtige Rolle spielen werden. Von daher möchte ich dem weiteren wissenschaftlichen Austausch nicht länger im Wege stehen und wünschen eine erfolgreiche Tagung.

Prof. Dr. Matthias Weidemüller
Prorektor für Innovation und Transfer der Universität Heidelberg

Vorwort der Herausgeber

Vincent Heuveline, Nina Bisheh, Philipp Kling

Liebe Leserinnen und Leser des Tagungsbandes der E-Science-Tage 2023,

die Digitalisierung durchdringt inzwischen alle Bereiche wissenschaftlicher Aktivitäten. Dies gilt insbesondere für den Umgang mit Forschungsdaten: Forscherinnen und Forscher sehen sich unweigerlich mit der Fragestellung konfrontiert, wie die gewonnenen Datensätze gewinnbringend verfügbar gemacht und nachhaltig bewahrt werden können.

Dieser und anderen Fragen rund um den gesamten Lebenszyklus von Forschungsdaten widmeten sich die E-Science-Tage 2023. Unter dem Motto „Empower Your Research – Preserve Your Data“ versammelten sich internationale Wissenschaftlerinnen und Wissenschaftler verschiedener Disziplinen in Heidelberg und boten Einblicke in aktuellen Fragestellungen und Erkenntnisse. Nachdem die letzten E-Science-Tage 2021 online stattfinden mussten, begrüßten die Teilnehmenden die diesjährige Gelegenheit für persönliche Begegnungen und den direkten Dialog über Ländergrenzen und Fachdisziplinen hinweg.

Wir freuen uns sehr, in dem vorliegenden Tagungsband Beiträge aus so vielfältigen Bereichen wie Materialforschung und Biodiversität, den Sozial-, Lebens- und Sprachwissenschaften, der Physik oder der Klimafolgenforschung präsentieren zu können. Welche Systeme und Anwendungen stehen heute schon für das Forschungsdatenmanagement zur Verfügung? Welche Vorteile bieten dabei Verbundlösungen? Und welche gesellschaftlichen Chancen liegen in einer etablierten Bereitstellung von Rohdaten? Die thematische Vielfalt und die rege Beteiligung belegen die Schlüsselrolle, die dem Forschungsdatenmanagement heute zukommt.

Die E-Science-Tage 2023 konnten nur durch die Unterstützung des Ministeriums für Wissenschaft, Forschung und Kunst Baden-Württemberg und durch das Engagement all derer ermöglicht werden, die bei der Organisation oder durch ihre Teilnahme aktiv mitgewirkt haben. Hierfür möchten wir uns herzlich bedanken. Unser besonderer Dank gilt außerdem den Autorinnen und Autoren, die mit ihren zahlreichen gelungenen Beiträgen an der Entstehung dieses Tagungsbands beteiligt waren. Auch in Zukunft soll das gewinnbringende Format der E-Science-Tage fortgesetzt werden, um den interdisziplinären Austausch zu

fördern und die wissenschaftliche Gemeinschaft rund um das Forschungsdatenmanagement in Deutschland zu stärken.

Prof. Dr. Vincent Heuveline
CIO Universität Heidelberg und Geschäftsführender Direktor des Rechenzentrums der Universität Heidelberg

Nina Bisheh, M.Sc.
Organisatorin der E-Science-Tage 2023 Konferenz

Dr. Philipp Kling
Mitglied des Organisationsteams der E-Science-Tage 2023 Konferenz

Teil II

Wissenschaftliche Beiträge

A Reproducible Machine Learning Workflow to Characterize the Solid Electrolyte Interphase

Deepalaxmi Rajagopal[1a,1b], Arnd Koeppe[1a,1b], Meysam Esmaeilpour[1c], Michael Selzer[1a,1b,2], Helge Stein[3], Britta Nestler[1a,1b,2]

[1a]Institute for Applied Materials – Microstructure Modelling and Simulation (IAM-MMS);
[1b]Institute for Nanotechnology (INT) - Mikrostruktursimulation (INT-MSS);
[1c]Institute of Nanotechnology (INT);
[1]Karlsruhe Institute of Technology (KIT);
[2]Institute for Digital Materials Science (IDM), Karlsruhe University of Applied Sciences
[3]Helmholtz Institute Ulm

Research data management tools structure the data life cycle and expedite the scientific process. Applied research data management still needs to be incorporated into daily research operations at the institutional level that allows access to the entire data life cycle. The generated warm data along the research operations enable automatic knowledge base generation and interpretation using robust integrated data analysis methods. The open-source research data infrastructure Kadi4Mat provides a generic framework for FAIR data management, efficient scientific workflows, and integrated data analysis. In this use case from virtual material design, we demonstrate how to implement machine learning as a workflow to characterize the virtual Solid Electrolyte Interphase (SEI) formation in Lithium-ion batteries. A better understanding of SEI formation helps to adjust batteries for optimum performance and safety. The workflow combines data and model definition, preprocessing, training, generation, and data analysis. We utilize kinetic Monte Carlo simulations and deep learning (Variational Auto Encoders with a parallel regressor) to structure the complex, high-dimensional, and non-convex design space and demonstrate how integrated data analysis can characterize materials and predict material properties.

1 Introduction

In recent years, there has been a strong push toward digitalizing the materials science field, with data-driven methods being developed to accelerate the development of new materials. For the generation and development of such high-performing material informatics, many syntheses, experiments, and simulations must be performed; this accelerated development

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.c18061 (Freier Zugang – alle Rechte vorbehalten)

of material results in multifaceted datasets. These datasets have to be combined and analyzed to discover new knowledge. The material data produced by experiments other than simulation lacks standard formats and corresponding metadata, which makes it difficult to share, visualize, and analyze these data (Ludwig 2019; Hey and Trefethen 2003; Draxl and Scheffler 2020). Incorporating FAIR data principles into material dataset generation methods helps make the data more accessible and easier for the research community to manage, share, and reuse.

A flexible research data management tool is essential for achieving the requirements of the FAIR (Findable, Accessible, Interoperable, and Reusable) principles by providing a wide variety of tasks that can be performed, such as retrieving data from an experimental device, processing data, sharing processed data more efficiently, automating the data handling process, data mining, and visualization. Utilizing research data management tools that provide access to unprocessed information sources from which published data is derived can result in manageable research process chains. Various research data software, such as Zenodo, Dataverse (Crosas 2011), Dspace (Smith et al. 2003), and Nomad, focus specifically on published data. The other important component of research data platforms is electronic lab notebooks to facilitate the digitization of the data from experiments and simulations while offering access to a wide range of data analysis and visualization tools required for the research.

Various Electronic Lab Notebooks (ELNs), including Jupyter Notebooks (Kluyver et al. 2016), Galaxy (Jalili et al. 2020), Fireworks (Jain et al. 2015), ElabFTW (CARPi, Minges, and Piel 2017), and Aiida (Pizzi et al. 2016) exist for documenting research processes. However, their domain-specific nature often prevents researchers from utilizing them fully. Moreover, most of the mentioned ELNs lack interdisciplinary capabilities and necessitate programmatic expertise for establishing workflows that can automate research operations. Hence, it is necessary to implement a research data management system that caters to the requirements of managing interdisciplinary research processes. This system should provide access to "warm data" which refers to unpublished data yet to be analyzed. Incorporating an ELN into the repository-based Research Data Management (RDM) tools with provisions for accessing user interface-based and script-based research process workflow implementations helps to minimize the effort required for daily research activities like data retrieval from the experimental devices, data sharing, data analysis, and visualization.

To meet the aforementioned requirements, Kadi4Mat (Karlsruhe Data Infrastructure for Materials Science; Team 2022; Brandt et al. 2021), an open-source data platform that functions as a communal repository, and the ELN is being developed at the Karlsruhe Institute of Technology. The Kadi4Mat platform's repository component facilitates effectively organizing data from various sources. It expedites data sharing among fellow researchers or research project collaborators. In contrast, their ELN component enables the logging of meaningful information about the research process, the visualization, and the data analysis stored in the repository component. It uses a user and programmatic interface to construct reproducible research workflows. In addition, the Kadi4Mat ecosystem provides access to KadiAI and CIDS (Computational Intelligence and Data Science

tools; Koeppe and CIDS Team 2023) to allow the use of interactive dashboards for machine learning process definition and execution, corresponding workflow nodes to define the process of data-driven study, and programmatic interfaces for data preparation into a machine-readable format, data engineering, data-driven model architecture construction, tuning, and training. Integrating Artificial intelligence (AI) toolset and advanced material simulations with research data infrastructure accelerates the discovery of new material configurations and develops traceable research process chains for further study (Koeppe et al. 2022; Mundt et al. 2020; Koeppe et al. 2018).

This article explores Kadi4Mat ecosystem functionalities and their potential for implementing reproducible integrated data analysis. Specifically, we use a deep generative model to examine the data-driven use case based on characterizing solid electrolyte interphase in batteries.

2 Tools and Methods

2.1 The Kadi4Mat ecosystem

Kadi4Mat represents a comprehensive platform incorporating a community repository to organize and share data from various sources efficiently, enriched by an ecosystem of applications and interfaces that facilitate efficient and automatized RDM. The ELN of the Kadi4Mat ecosystem consists of both web-based and desktop-based workflow editors to streamline scientific workflows. Kadi4Mat aims to digitally document the scientific workflow in daily research, which facilitates researchers in reproducing and utilizing identical data and research workflows more efficiently. The generic architecture of Kadi4Mat enables the replication of nearly all stages involved in the research data lifecycle, except planning and publishing. However, it is possible to effectively execute these two procedures by integrating established frameworks into the Kadi4Mat. For instance, RDMO (Klar et al. 2017) may be utilized for managing research data plans, while Zenodo can serve as a platform for publishing data. In the following sections, some components of the Kadi4Mat ecosystem are discussed to provide an overview of their role in implementing reproducible integrated data analysis.

2.2 Kadi's core components: KadiWeb and Kadistudio

KadiWeb is a web-based interface that provides a user-friendly platform to access the repository and ELN components of Kadi4Mat. The resources stored in the Kadi4Mat can be accessed via the web and programmatic interfaces called KadiAPY (Schoof and Brandt 2020) based on the command line interface and Python library. Upon accessing Kadi4Mat, the web interface provides access to create and structure data with the help of different components available in the interface. The essential features of the Kadi4Mat are as follows:

Records: instances to structure resources, including data and associated metadata. The metadata includes general information such as title, description, identifier, record type, tags, and other additional metadatum denoted by key/value pairs specific to the stored data, which can be customized based on user needs.

Collections: Collect and arrange records of relevance together.

Templates: defines the record's metadata beforehand. It can contain the information required to create a record or the extra metadata specific to an individual record.

Users: displays registered users of logged-in Kadi4Mat instance.

Group: organizes users into workspaces based on their roles or research projects and facilitates transparent and efficient access management.

Kadistudio is a desktop-based workflow editor enabling the seamless implementation of scientific workflows. The software platform provides access to a wide range of pre-installed tools conveniently tailored to create and implement heterogeneous scientific workflows comprising diverse data types and sources (Griem et al. 2022; Zschumme 2021b, 2021a). To add new tools in the workflow environment, user can define their corresponding XML tool settings (Zschumme et al. 2020). The functionalities of KadiAPY (Schoof and Brandt 2020), workflow nodes that mimic user interface commands, and additional tools such as CIDS and KadiAI for data-based analysis are also available in the form of workflow nodes accessible through the Kadistudio integrated workflow editor.

2.3 KadiAI and CIDS

As an interface between RDM and machine learning applications, KadiAI aims to standardize and integrate AI projects, work packages, and workflows into the Kadi ecosystem. The interface streamlines data-driven research through structuring and automatization and offers interactive dashboards to implement data-driven studies and provide feedback to the user. CIDS is a Python-based framework to develop, implement, and standardize learning algorithms in AI workflows. The framework is integrated within the Kadi4Mat ecosystem to aid in preprocessing and converting the data, as well as develop data-driven models that learn from the defined datasets stored within the Kadi4Mat repository. Together, KadiAI and CIDS enable data-integrated AI within the Kadi4Mat ecosystem to enable a quantitative and qualitative analysis of the collected data.

The generic machine learning workflow concept in Kadi4Mat consists of four essential process elements: source and data preparation, model development, and share model for prediction. There are two different ways to implement this machine-learning workflow concept within Kadi4Mat. One method uses KadiAI user interactive dashboards, which require no programming background to establish and perform machine learning process steps. In contrast, the other method utilizes tailored Python scripts within the CIDS framework for optimized automation of node execution. Figure 1 shows the tools required to implement the machine learning process steps using the Kadi4Mat ecosystem.

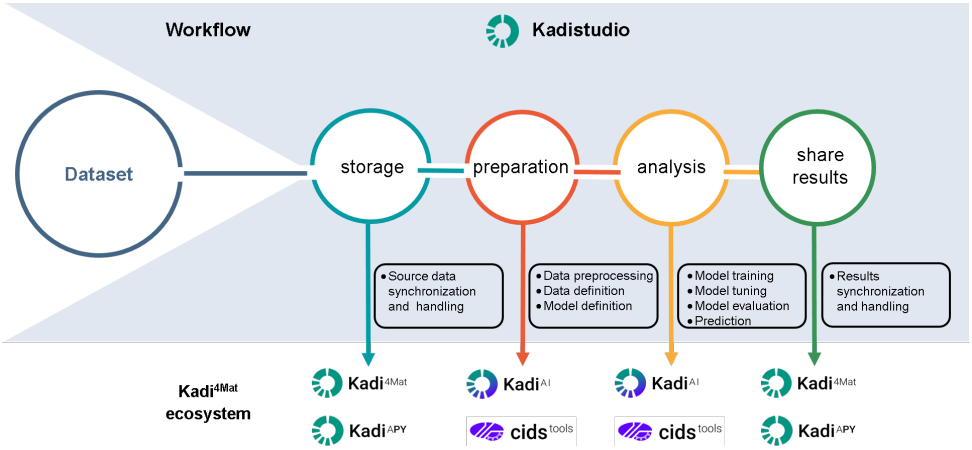


Figure 1: Concept of generic machine learning workflow in Kadi4Mat ecosystem.

3 Use case and workflow

In subsequent sections, we will address the use case research problem based on the characterization of solid electrolyte interphase using a data-driven strategy and its implementation using the Kadi4Mat ecosystem. Specifically, we used a collection of CIDS tools incorporated within Kadistudio to define and execute each processing step of machine learning workflows. Figure 2 shows the machine workflow layout using the Kadi4Mat ecosystem. Some CIDS workflow nodes can communicate with the Kadi4Mat repository to retrieve original data, update transformed data, or upload analysis results. This feature helps with tracking and ensuring the reproducibility of data provenance.

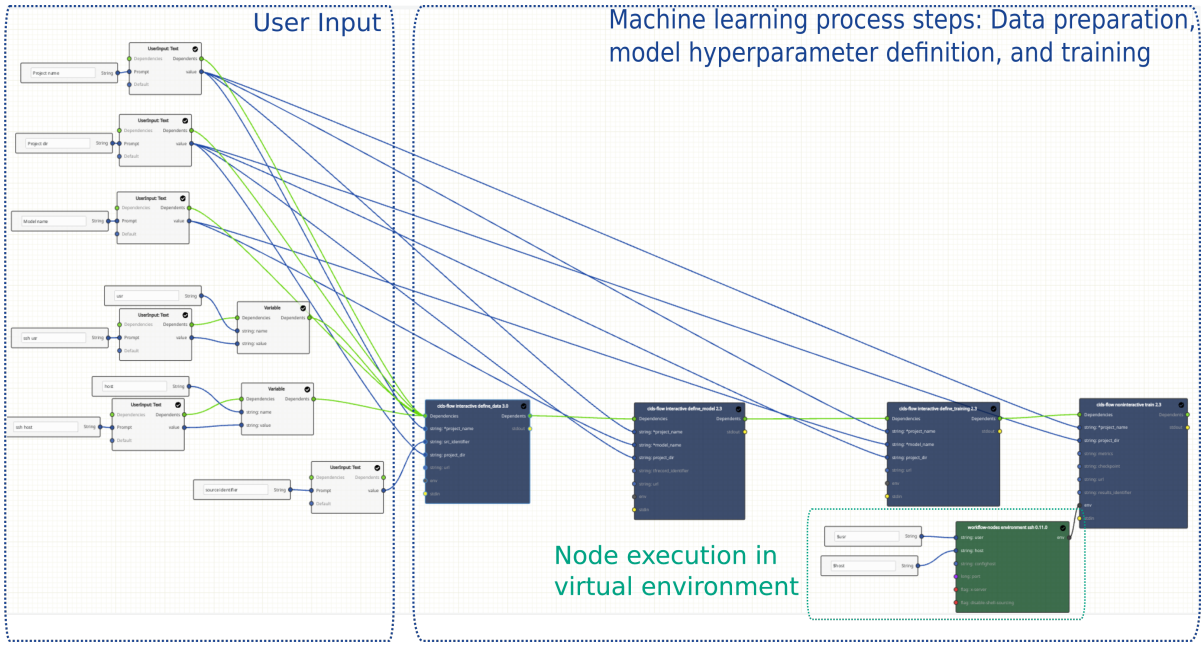


Figure 2: Machine learning workflow layout visualization in Kadistudio.

3.1 Background

The current use case involves a comprehensive investigation of the dataset acquired from kinetic Monte Carlo simulation of solid electrolyte interphase growth in batteries.

The solid electrolyte interphase is formed on the electrode of a battery as a reduction product of the electrolyte. These interphases play a vital role in performance, cycling, and even the safety of the battery (Dunn, Kamath, and Tarascon 2011). The complex interdependent relationship between the composition and morphology of the evolving solid electrolyte interphase (SEI) makes the traditional modeling approaches struggle to capture essential information to predict the behavior of SEI accurately. The recently developed Kinetic Monte Carlo protocol based on the multiscale approach with corresponding reaction kinetics as input can capture the growth of SEI and can determine the essential physical properties of the evolving SEI (Esmaeilpour et al. 2023). However, the above-proposed protocol still takes a long time to complete each simulation and lacks structure-property linkage to interpret and understand the underlying mechanism of SEI growth.

3.2 Objective and approach for characterization using deep generative models

This work aims to accelerate the design of solid electrolyte interphase according to target physical properties and to understand the underlying mechanisms that dominate its growth. We implemented a data-driven approach to expedite the Kinetic Monte-Carlo simulation to characterize the SEI configuration concerning existing physical properties for further optimization. For this approach, we used a variational auto-encoder model and a regressor to understand the underlying representation of higher dimensional SEI configurations and predict the key physical properties of SEI formation. The proposed approach demonstrates the ability to accurately predict SEI properties and structure, which can be used to optimize battery performance and safety.

3.3 Source data and preprocessing

The study utilizes input data derived from Kinetic Monte Carlo simulations that model SEI growth. The obtained SEI dataset contains 50000 samples of the final configuration of SEI growth. Each sample consists of spatial features and non-spatial features. The spatial features of an SEI configuration represent the electrolyte reduction species distinguished by colors according to their reaction product type. The non-spatial features of the SEI configuration refer to physical properties such as thickness, density, volume fraction, and porosity.

The foremost step in promoting data-oriented research using machine learning models is gathering information from experimental observations or simulations. The repository component of Kadi4Mat facilitates efficient sourcing and collection of datasets from the

collaborators. Once collected, it is essential to preprocess and clean the data to eliminate any inconsistencies or errors that may compromise the accuracy of the analysis. For instance, we received SEI configuration data generated with Kinetic Monte Carlo simulations from a collaborator via Kadi4Mat (ibid.). To proceed with analysis using this information, we need to define attributes of the obtained dataset such as feature name, data format, data type, data shape, and decoding used during data conversion. Additionally, it is necessary to preprocess the source data to achieve an efficient learning process using machine learning algorithms. The preprocessed data of each sample are then stored in the form of TF records (TensorFlow records) for efficient data serialization and storage. We utilized the CIDS framework-based interactive data definition node to streamline this process to facilitate user input requirements and functionalities for defining the data attributes. Figure 3 visualizes the interactive data definition process using CIDS interactive nodes and KadiAI dashboards.

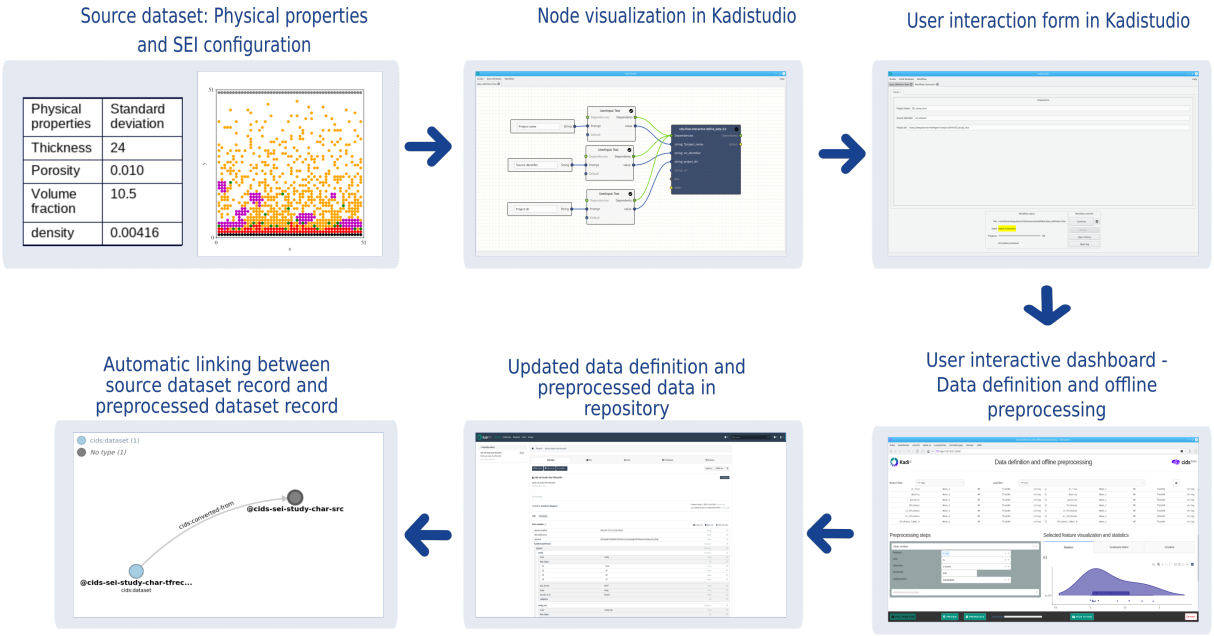


Figure 3: Data preparation process for machine learning study using CIDS interactive nodes and KadiAI dashboards.

3.4 Deep generative models

Variational autoencoder (VAE; Kingma and Welling 2013) is a deep generative model consisting of two main parts: encoder and decoder. The function of the encoder is to compress the higher dimensional input into a distribution over the lower dimensional latent space. The lower dimensional latent space defines the bottleneck of the VAE model. Thus, the encoder produces new feature space from the old feature space through extraction or selection of essential features. On the other hand, the decoder part takes a point in the lower dimensional latent space and decompresses it back to the original higher dimensional input space. An additional part called regressor at the bottleneck of VAE

aids in incorporating the information of physical properties into the learned latent space. Training the regressor component involves predicting the desired physical properties acquired latent space representation and backpropagation to automatically arrange VAE's bottleneck based on physical attributes (Gómez-Bombarelli et al. 2018). The acquired knowledge of the representations assists in predicting desired physical properties. Here, we used the CIDS interactive model definition node to aid in defining the architecture of a model. Figure 4 visualizes the model definition and hyperparameter selection using the CIDS interactive model definition node. This node aims to enable efficient input feature and output feature definition, facilitate the selection of appropriate model architectures, and specify ranges for corresponding hyperparameters. The hyperparameters defined here determine the aspects of model architecture, such as ranges for the numbers of convolution layers within the encoder and decoder, ranges and choices of latent dimensionality, and choice of activation function for the layers.

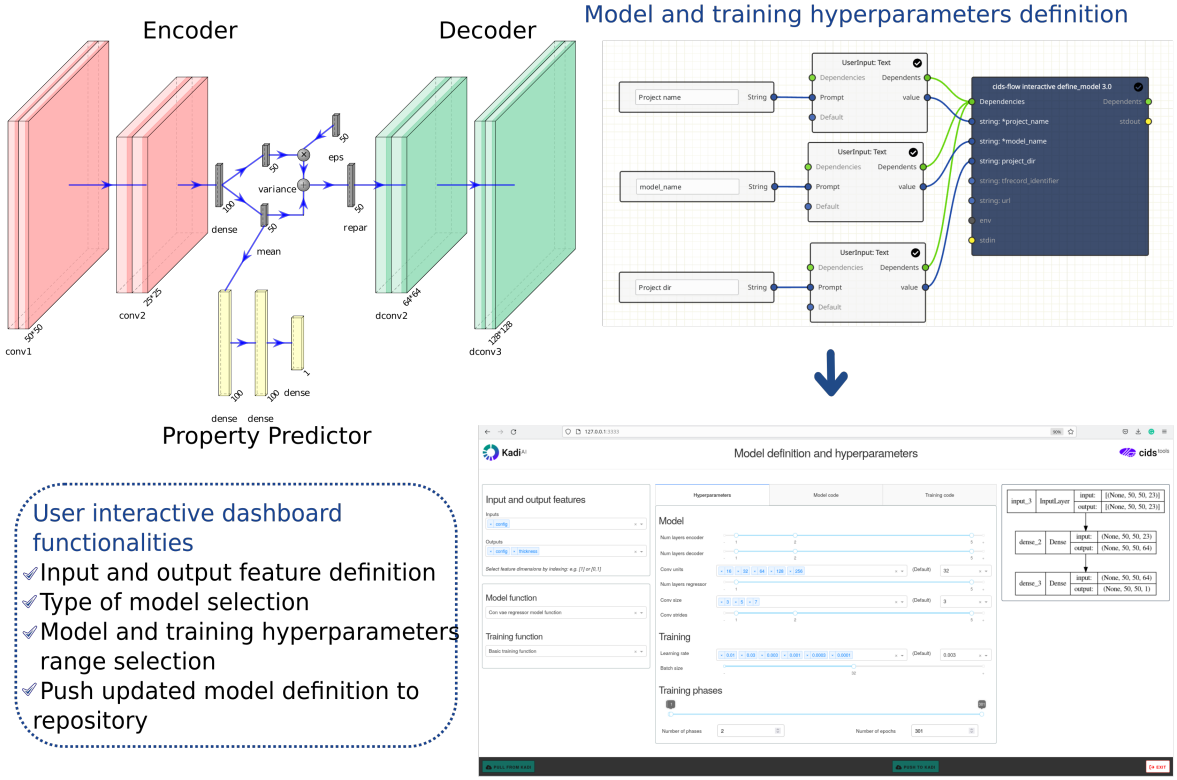


Figure 4: Model and training hyperparameter selection using CIDS interactive nodes and KadiAI dashboards.

In a machine learning workflow, the next step is to define the hyperparameters of the training, such as batch size, number of epochs, learning rate, optimizer choice, and number of training phases. These settings should be carefully selected and fine-tuned to ensure optimal model performance during training. A user-interactive training definition node assists in the definition of these training hyperparameters and their corresponding search

range. On defining the required hyperparameters for training, the next step is to start the model training with the help of a non-interactive training node which automates the training process by utilizing the hyperparameters and model architecture defined in the previous steps.

3.5 Data provenance tracking in executed machine learning workflows

The KadiAI workflow nodes developed to define and execute machine learning processes using the CIDS framework record all the essential artifacts obtained throughout the workflow execution. Each CIDS workflow node integrated with the KadiAI interface has its machine-learning metadata library. The metadata library captures and retrieves metadata containing essential information about the various workflow process steps, their executions, and the artifacts produced during the machine learning process. Logging metadata includes everything from setting up KadiAI projects to model training and evaluation. KadiAI nodes generate the corresponding record type in the Kadi4Mat repository for each machine learning process result and metadata storage.

The KadiAI projects are locally executed at the user level, which is analogous to classic research processes. Through a structured upload, KadiAI projects synchronize the results with the repository, where each processing step is saved as a record with corresponding metadata and data. As the admin of the uploaded project, the user can grant access to other members with specific roles such as member, editor, admin, or collaborator. KadiAI projects also record the tool versions used for each processing step, allowing users to go forward or backward with tool versions or infrastructures as required. For each new model within a KadiAI, the ML model-creating function as Python code and tunable model settings are recorded in separate records, which are linked to the corresponding project record. Execute in Kadi workflows, users can skip process nodes and reuse existing derived datasets for new models if no reprocessing is required, ultimately streamlining the workflow. The created records of the KadiAI project are linked using unidirectional record links to define the context of the relationship between them. The knowledge graphs of the executed machine learning workflow aid in traceability and provenance tracking of the data used, models developed, and results obtained during the machine learning project.

Figure 5 visualizes the data provenance for the characterization of solid electrolyte interphase using the deep generative model in KadiAI as knowledge graphs. Knowledge graphs turn data collected along your machine-learning process into machine-understandable knowledge. These graphs help you to define the context of the relationship between the two methods and adapt situational changes. On the other hand, it allows you to incorporate real-world knowledge into your study, which data-driven lacks.

3.6 Data and code availability

The code for Kadi4Mat is readily available on the public Gitlab repository (Kadi4Mat Team and Contributors 2023) for seamless sharing and collaboration within the community. Additionally, we publish tool versions on Zenodo with a Digital Object Identifier

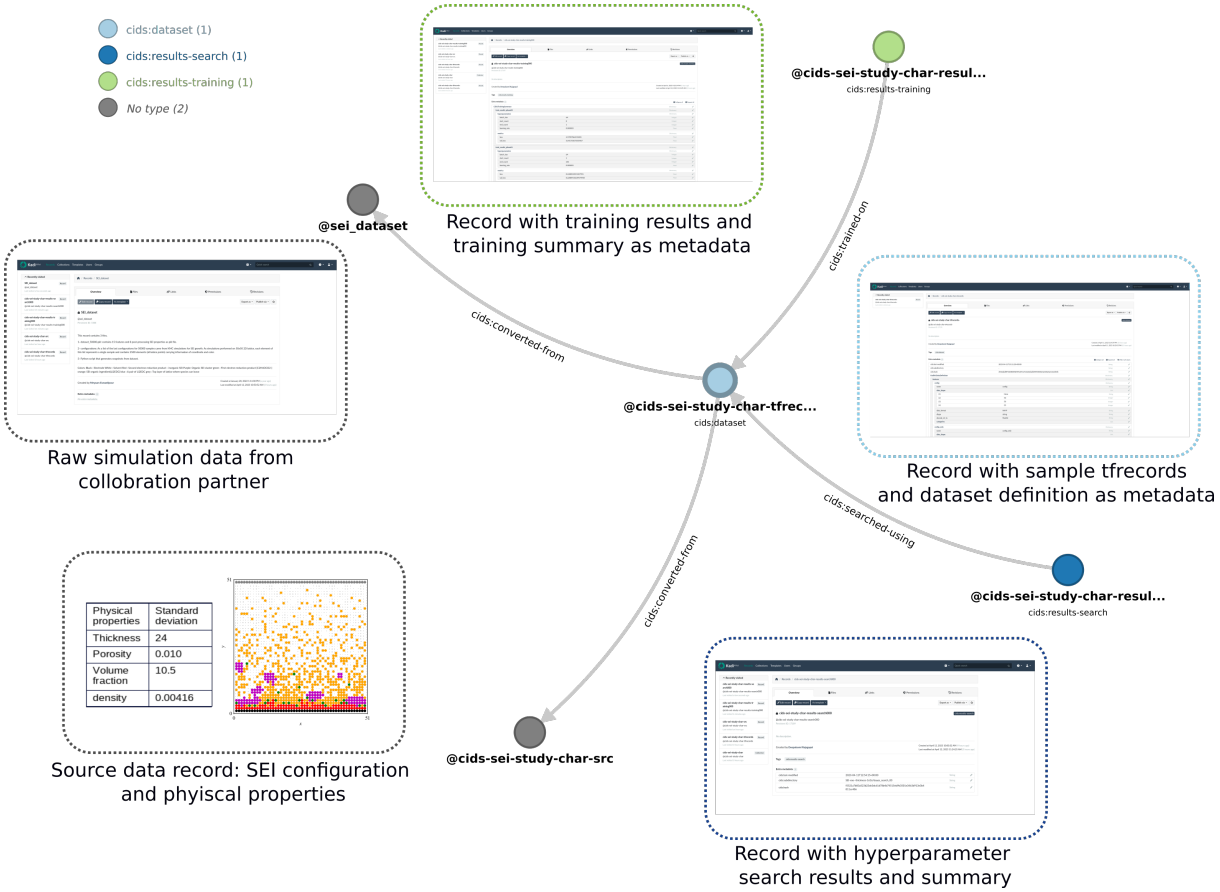


Figure 5: Visualization of tracked data provenance during machine learning process steps as knowledge graphs.

(DOI; Kadi4Mat Team and Contributors 2023) for enduring accessibility while prioritizing backward compatibility of both tools and infrastructure through the versioning tools on Zenodo. Implementing the FAIR principles made our code easily discoverable and accessible through our public Gitlab repository. Furthermore, we ensure that the code is interoperable through the commonly used Python language, and our Zenodo DOI persistence guarantees the reproducibility of tool versions.

4 Conclusion

In this article, we discussed the importance of research data management in daily scientific research activities and the role of existing research data infrastructure in tackling the challenges related to the organization, storage, and sharing of research data. Then the infrastructure and functionalities of the Kadi4Mat ecosystem, a generic research data management system to handle heterogeneous data in interdisciplinary material science, and the development of virtual research environments through scientific workflows are described. To understand the functionalities of the Kadi4Mat ecosystem, a research problem focused on the data-driven study to characterize and predict the properties of solid elec-

trolyte configuration of interest is addressed. The demonstrated machine learning workflow using the CIDS framework and KadiAI user interactive dashboards interfaced with the Kadi4Mat repository showed how data management like Kadi4Mat can be integrated with machine learning processes to ensure reproducibility, transparency, and traceability of artifacts generated during the execution of the workflow.

Acknowledgements

The authors gratefully acknowledge the financial support for the research provided by the Deutsche Forschungsgemeinschaft's "Cluster of Excellence" POLiS (project number 390874152), the BMBF's "FestBatt" competence cluster (project number 03XP0174E), the Ministry of Science, Research, and Art Baden-Württemberg (MWK-BW) in the project MoMaF-Science Data Center, with funds from the state digitization strategy digital@bw (project number 57). We would also like to thank the German Federal Ministry of Education and Research (BMBF) for its financial support of the AQuaBP project under grant number 03XP0315B. The authors acknowledge support by the Helmholtz association through the program MTET, no: 38.02.01.

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.

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.

Crosas, Mercè. 2011. "The Dataverse Network: An Open-source Application for Sharing, Discovering and Preserving Data". *D-Lib Magazine* 17 (1/2). DOI: https://doi.org/10.1045/january2011-crosas.

Draxl, Claudia, and Matthias Scheffler. 2020. "Big data-driven materials science and its FAIR data infrastructure". In *Handbook of Materials Modeling,* 49–73. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-44677-6_104.

Dunn, Bruce, Haresh Kamath, and Jean-Marie Tarascon. 2011. "Electrical energy storage for the grid: a battery of choices". *Science* 334 (6058): 928–935. DOI: https://doi.org/10.1126/science.1212741.

Esmaeilpour, Meysam, Saibal Jana, Hongjiao Li, Mohammad Soleymanibrojeni, and Wolfgang Wenzel. 2023. "A Solution-Mediated Pathway for the Growth of the Solid Electrolyte Interphase in Lithium-Ion Batteries". *Advanced Energy Materials* 13 (14): 2203966. DOI: https://doi.org/10.1002/aenm.202203966.

Gómez-Bombarelli, Rafael, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. 2018. "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules". *ACS Central Science* 4 (2): 268–276. DOI: https://doi.org/10.1021/acscentsci.7b00572.

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.

Hey, Anthony, and Anne Trefethen. 2003. *The Data Deluge: An e-Science Perspective.* DOI: https://doi.org/10.1002/0470867167.CH36.

Jain, Anubhav, Shyue Ping Ong, Wei Chen, Bharat Medasani, Xiaohui Qu, Michael Kocher, Miriam Brafman, Guido Petretto, Gian-Marco Rignanese, Geoffroy Hautier, et al. 2015. "FireWorks: a dynamic workflow system designed for high-throughput applications". *Concurrency and Computation: Practice and Experience* 27 (17): 5037–5059. DOI: https://doi.org/10.1002/cpe.3505.

Jalili, Vahid, Enis Afgan, Qiang Gu, Dave Clements, Daniel Blankenberg, Jeremy Goecks, James Taylor, and Anton Nekrutenko. 2020. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update". *Nucleic Acids Research* 48 (W1): W395–W402. DOI: https://doi.org/10.1093/nar/gkaa434.

Kadi4Mat Team and Contributors. 2023. *Kadi-Karlsruhe Data Infrastructure for Materials Science.* https://gitlab.com/iam-cms/kadi.

Kadi4Mat Team and Contributors. 2023. *kadi: 0.39.3.* Version 0.39.3. DOI: https://doi.org/10.5281/zenodo.8233992.

Kingma, Diederik P, and Max Welling. 2013. "Auto-encoding variational bayes". *arXiv preprint arXiv:1312.6114.*

Klar, Jochen, Claudia Engelhardt, Heike Neuroth, Harry Enke, and Jens Ludwig. 2017. "RDMO-Research Data Management Organiser". In *EGU General Assembly Conference Abstracts,* 15760.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. 2016. "Jupyter Notebooks – a publishing format for reproducible computational workflows." In *Positioning and Power in Academic Publishing: Players, Agents and Agendas,* 87–90. IOS Press. DOI: https://doi.org/10.3233/978-1-61499-649-1-87.

Koeppe, Arnd, Franz Bamer, Michael Selzer, Britta Nestler, and Bernd Markert. 2022. "Explainable artificial intelligence for mechanics: physics-explaining neural networks for constitutive models". *Frontiers in Materials* 8:636. DOI: https://doi.org/10.48550/arXiv.2104.10683.

Koeppe, Arnd, and CIDS Team. 2023. "CIDS and KadiAI GitLab Repository". Visited on September 5, 2023. https://gitlab.com/intelligent-analysis/cids.

Koeppe, Arnd, Carlos Alberto Hernandez Padilla, Maximilian Voshage, Johannes Henrich Schleifenbaum, and Bernd Markert. 2018. "Efficient numerical modeling of 3D-printed lattice-cell structures using neural networks". *Manufacturing Letters* 15:147–150. DOI: https://doi.org/10.1016/j.mfglet.2018.01.002.

Ludwig, Alfred. 2019. "Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods". *NPJ Computational Materials* 5 (1): 70. DOI: https://doi.org/10.1038/s41524-019-0205-0.

Mundt, Marion, Arnd Koeppe, Sina David, Tom Witter, Franz Bamer, Wolfgang Potthast, and Bernd Markert. 2020. "Estimation of Gait Mechanics Based on Simulated and Measured IMU Data Using an Artificial Neural Network". *Frontiers in Bioengineering and Biotechnology* 8 (41): 1–16. DOI: https://doi.org/10.3389%2Ffbioe.2020.00041.

Pizzi, Giovanni, Andrea Cepellotti, Riccardo Sabatini, Nicola Marzari, and Boris Kozinsky. 2016. "AiiDA: automated interactive infrastructure and database for computational science". *Computational Materials Science* 111:218–230. DOI: https://doi.org/10.1016/j.commatsci.2015.09.013.

Schoof, Ephraim, and Nico Brandt. 2020. *IAM-CMS/kadi-apy: Kadi4Mat API Library.* Version 0.2.1. DOI: https://doi.org/10.5281/zenodo.4088276.

Smith, Mackenzie, Mary Barton, Mick Bass, Margret Branschofsky, Greg McClellan, Dave Stuve, Robert Tansley, and Julie Walker. 2003. "DSpace: An Open Source Dynamic Digital Repository". *D-Lib Magazine* 9. DOI: https://doi.org/10.1045/january2003-smith.

Team, Kadi4Mat. 2022. *kadistudio: 0.1.0.alpha1.* DOI: https://doi.org/10.5281/zenodo.6810891.

Zschumme, Philipp. 2021a. *IAM-CMS/process-engine.* Version 0.1.0. DOI: https://doi.org/10.5281/zenodo.4442563.

———. 2021b. *IAM-CMS/process-manager.* Version 0.1.0. DOI: https://doi.org/10.5281/zenodo.4442553.

Zschumme, Philipp, Patrick Altschuh, Nico Brandt, Lars Griem, and Ephraim Schoof. 2020. *IAM-CMS/workflow-nodes.* Version 0.1.0. DOI: https://doi.org/10.5281/zenodo.4094719.

Linking Domain-specific RDM to Institutional and Generic Approaches – the Case of NFDI4Biodiversity

Jimena Linares[1], Barbara Ebert[1], Judith Sophie Engel[2,3]

[1]German Federation for Biological Data – GFBio e.V.;
[2]University of Bremen;
[3]Center for Environmental Sciences - MARUM

NFDI4Biodiversity is a consortium within the German National Research Data Infrastructure – NFDI. The consortium has been in action since October 2021 with the main goal of mobilizing data, rolling out services, and offering tools for the biodiversity community. At the moment, it has over 50 partners and a five-year work program until 2025, funded through a joint initiative of the Federal Government and its states (the Länder).

One of the partners is the German Federation for Biological Data – GFBio (e.V). GFBio has been assisting scientists and data managers for years regarding all types of inquiries about research data management in biodiversity, ecological, and environmental research, and providing a portfolio of services along the research data life cycle.

With an expanding consortium and increasing interest in research data management in universities and research-performing organizations, one challenge to overcome is how to support the scientific community. On-site (e.g. on-campus) interaction with researchers is often provided by local research data management services (RDM officers). Still, it is common that resources are limited and in-depth subject-specific advice cannot be provided. The "Front Office/Back Office" model propagated by GFBio presents a good way to serve the scientific community by linking domain-specific approaches with institutional and generic ones.

1 Introduction

Community-specific tools and expertise are certainly needed for better management of research data. NFDI4Biodiversity[1] works with different data and service providers in the domain of biodiversity, ecology, and environmental research (Weber et al. 2021). The consortium is composed of and supports a diverse community of researchers, research

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.c18062

1 https://www.nfdi4biodiversity.org/en/; *Last accessed on March 7th, 2023.*

organizations, public authorities, professional societies, and citizen science initiatives in managing their data and making them available for re-use. The community mainly generates data from field and lab work. Examples of these data are: occurrence data (observations of species); environmental data (e.g. temperature, rainfall); trait data (e.g. seed number and mass); molecular data (e.g. DNA and RNA sequences); experimental and laboratory measurements; multimedia (photographs, audio, and video of e.g., orthophotos produced using a drone, observed specimens); digital surface models; model code and statistics. Publication and distributed archiving are organized in a network of ten specialist data centers coordinated by the German Federation for Biological Data (GFBio[2]). Seven data centers are located at natural sciences collections across Germany. There, researchers can submit and get advice about the best way to submit their data according to their affiliations, specific data type (e.g. audio from birds callings), archival, and publication intentions. The other three data centers are dedicated to handling nucleotide, plant, and environmental data. The consortium's Helpdesk is the central entry point for requests from partners and researchers related to their research data management activities, including the submission of data for archiving and publication. The Helpdesk is a service run and maintained by the consortium partner GFBio e.V.[3]and is set up in the Jira project manager software[4]. The ticket workflow in the Helpdesk is a centralized dispatching system. Requests are assigned to a team of experts within NFDI4Biodiversity and cooperating partners by the Helpdesk Team.

2 Core services for biodiversity, ecological, and environmental research

The Helpdesk acts as the gateway for all requests. User requests via mail to helpdesk@nfdi4biodiversity.org and info@gfbio.org or via the contact form on nfdi4biodiversity.org generate a ticket in the Helpdesk. The Data Management Planning Tool - DMPT and the data submission service (which will be discussed later) are also connected to the Helpdesk. There are three major request types (Figure 1): 1. HELP requests for general requests around (research) data management, service support, training, event, or networking; 2. DMP-type requests (a specialized type of HELP request) generated by users that use the DMPT and require personalized support, and 3. DSUB requests generated when users are submitting data via the data submission service (aka submission system).

All communication takes place through the Helpdesk in the respective ticket(Astor et al. 2021), so the information is documented there in a transparent manner for the users, and it is internally reachable for colleagues involved and interested in the ticket. The Helpdesk workflow promotes the Front Office/Back Office model (Figure 2): The Back Office comprises the Helpdesk Team (first-level support) and the expert network in the consortium (second-level support). The Front Office is the expert network at the univer-

2 `https://www.gfbio.org/data-centers`; *Last accessed on March 22nd, 2023.*
3 `https://www.gfbio.org/gfbio_ev/`; *Last accessed on March 10th, 2023.*
4 `https://www.atlassian.com/software/jira`; *Last accessed on March 7th, 2023.*

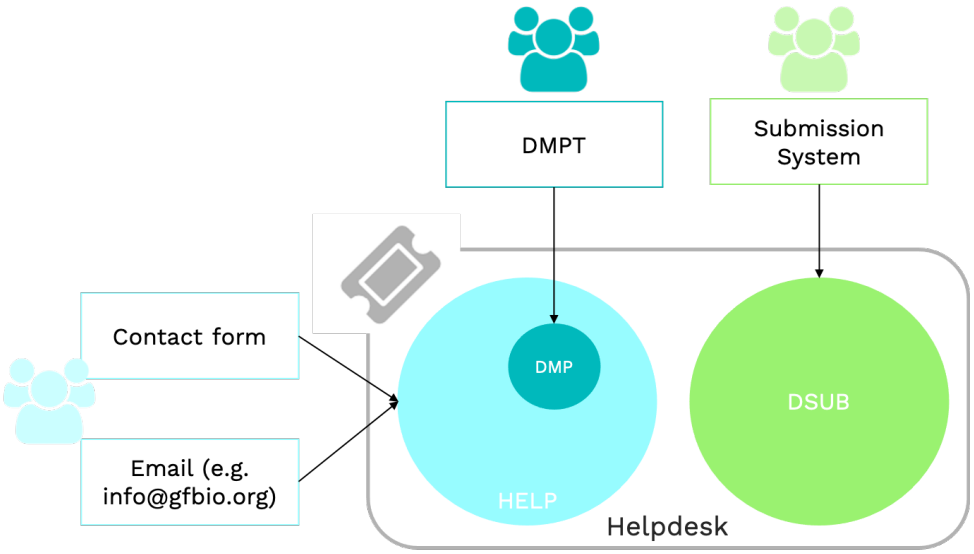


Figure 1: Helpdesk requests, including two core services in RDM.

sities and research institutions, e.g. the data stewards or data managers that offer on-site support to students and scientists of their organization.

They can address discipline-specific questions to the Helpdesk Team, or forward the specific questions directly to the Helpdesk Back Office.

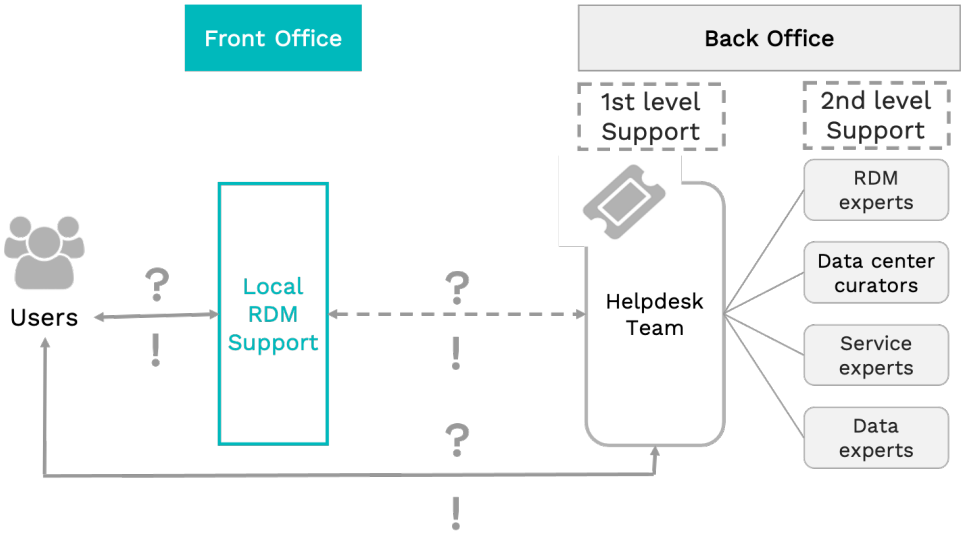


Figure 2: Front Office/Back Office support provided by the Helpdesk.

The Front Office/Back Office model has been continuously tested in two pilot projects:

a) FEdA & KonsortSWD The "BMBF Research Initiative for the Conservation of Biodiversity" (in German: Forschungsinitiative zum Erhalt der Artenvielfalt) – FEdA[5]

5 `https://www.feda.bio/en/`; *Last accessed on March 20th, 2023.*

is a long-term funding program with several cohorts of interdisciplinary projects on biodiversity and biodiversity protection. The funder expects the projects to deliver a data management plan and metadata as part of their data outputs, to feed a joint data portal for the program. The speakers of FEdA and NFDI4Biodiversity agreed to collaborate early on to link the research initiative's data management efforts to the evolving support structure in the NFDI. A focal point of the collaboration is the support of data management plans for these projects. The data manager from FEdA is included in the NFDI4Biodiversity Helpdesk, and a workflow was established to coordinate the data management plan and data archival support. The workflow has been optimized such that FEdA assumes the Front Office role, providing the initial support and dissemination of information. Then, requests are forwarded to the respective expert network in the "Back Office" of NFDI4Biodiversity for specific biological, ecological, and environmental research questions.

As FEdA's interdisciplinary projects often have social sciences components, a cooperation with KonsortSWD[6], the NFDI Consortium for social, behavioral, educational, and economic science, and specifically Qualiservice[7], was established to include additional support by social data experts. In this way, the collaboration with FEdA is also a use case for cross-consortia data management support. The Helpdesk workflows were adapted so the data experts can communicate with the researchers and the NFDI4Biodiversity colleagues as well as incorporate their input in the data management plans. When researchers decide to archive their interview data with Qualiservice, support for the necessary documentation and publication procedures are moved to the Qualiservice helpdesk, which is easier for both sides. An example is the use case project "SLInBio - Urban lifestyles and the valorization of biodiversity: dragonflies, grasshoppers, bumblebees and Co", which researches how the perception and valorization of insects can be increased in an area and what contribution cities can make to the conservation of insect diversity[8] The project works, among other data types with molecular and ecotoxicological data, interviews, and surveys. Each of them will be archived and published in the most fitted repository, which implies that diverse repositories for different disciplines (biological, environmental, and social sciences) will be used for publication.

b) UBRA/ Data Train The Helpdesk collaborates with the U Bremen Research Alliance – UBRA[9], specifically with the coordinator of the Research Data Management and Data Science Training course "Data Train"[10] for planning and organizing the courses on research data management. This serves as a notable instance of continuous collaboration in research data management with state initiatives (Landesinitiativen) and institutions, such as the University of Bremen.

6 `https://www.konsortswd.de/en/`; *Last accessed on March 19th, 2023.*
7 `https://www.qualiservice.org/en/`; *Last accessed on March 19th, 2023.*
8 `https://www.isoe.de/en/nc/research/projects/project/slinbio-1/`; *Last accessed on July 31st, 2023.*
9 `https://www.bremen-research.de/en/`; *Last accessed on March 22nd, 2023.*
10 `https://www.bremen-research.de/data-train/`; *Last accessed on March 22nd, 2023.*

As mentioned, the Helpdesk tickets also involve two special types of tickets that are produced once the users access and make use of two core services: data management plan support (through the DMPT) and the data submission service. These two are the most popular services offered by the consortium and, together with other services[11], aim to facilitate the best practices in research data management following the FAIR principles (Wilkinson et al. 2016), as well as domain-specific data standards and practices. Both services are explained below:

Data management plan support

The DMPT[12] has been the go-to tool for supporting and conceptualizing Data Management Plans – DMPs in biodiversity research for years, with a special focus on DFG requirements. In 2021, the DMPT underwent a re-implementation using the Research Data Management Organizer – RDMO (Klar et al. 2023) as a backbone, while retaining its familiar front end. This was done to optimize personalized DMPs for specific cases and to facilitate interdisciplinary cooperation, among other benefits. The revamped DMPT-RDMO base is an example of a specific RDM tool transitioning from an in-house development to a community-supported open-source tool, ensuring both sustainability and better support of research data management efforts across disciplines and institutions. The tool's questionnaire includes the common components of a DMP, and its sections follow the Guidelines of the German Research Foundation (Deutsche Forschungsgemeinschaft) – DFG[13] on Handling of Research Data[14], Handling of Research Data in Biodiversity Research[15], and for Safeguarding Good Research Practice[16]. The support of individual researchers on the making of a DMP as well as on the posterior archival and publication of the data that will be generated in the research project is a collaborative effort among experts who work with domain-specific data types. The DMP produced follows consensus among all involved parties (as defined by their roles: user, DMP Officer, DMP controller, collaborators such as contacts from data centers involved), and other experts (e.g. colleagues from initiatives such as FEdA, colleagues from other consortia). An example of a GFBio-approved (consented) DMP can be seen in the model DMP (Mau, Timmermann, and Astor 2020).

The data submission service

The data submission service[17] works with the above-mentioned data centers to offer a comprehensive curation service of the research data provided to GFBio (as seen in Figure 1).

11 `https://www.gfbio.org/services`; *Last accessed on March 29th, 2023.*
12 `http://dmp.gfbio.org`; *Last accessed on March 29th, 2023.*
13 `https://www.dfg.de/`; *Last accessed on March 29th, 2023.*
14 `https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/guidelines_research_data.pdf`; *Last accessed on March 29th, 2023.*
15 `https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/guidelines_biodiversity_research.pdf`; *Last accessed on March 29th, 2023.*
16 `GuidelinesforSafeguardingGoodResearchPractice`; *Last accessed on March 24th, 2023.*
17 `http://submissions.gfbio.org`; *Last accessed on March 20th, 2023.*

The data centers include international data archives such as the Data Publisher – PANGAEA[18] or the European Nucleotide Archive – ENA[19] as well as several specialized German Collection Data Centers. Services are organized along five major biological data types[20]. All data centers have agreed on common (meta)data standards[21] for submission and curation. There is a minimum of consensus metadata required before uploading the data. Data center staff cooperates through the Helpdesk infrastructure to organize distributed archiving. In this way, thematically and structurally heterogeneous data can be archived in different data centers (e.g. ENA and PANGAEA), and the persistent identifiers will be linked with each other. The funding obtained through the National Research Data Infrastructure is crucial to ensure this engagement.

To create a submission through GFBio, the user is required to obtain a GFBio account to access the submission service. This can be done with the individual Academic ID, or by obtaining an ID from the GFBio Identity provider service. After that, the direct support on the specific submission and curation begins. Each submission requires individual support by a Helpdesk Team member, as there can be cases where the data files and/or types do not fit within the submission criteria of the data centers, so external alternatives for archival and publication are suggested to the submitters. File sizes in the submission helpdesk are currently limited to 200 MB. This is different from the size of files accepted in each data center for archival and publication (ENA, for example, has no restrictions). The curation of datasets is supported by the curators of the data centers through communication with the submitter in the ticket, which is automatically created by the submission request made through the helpdesk. For general requests about submissions landing on the helpdesk, besides the support provided by the experts in the request, there is also a self-serve online library, the Knowledge Base[22], with information for current and future submitters[23]. Depending on the type of data, dedicated templates, and documentation are also available to facilitate standardization. For example, the molecular submission template[24] for datasets aimed to be brokered to ENA has documentation on the formats and values expected[25] for the minimum mandatory and optional fields. This documentation is continuously updated by the GFBio submissions team to comply with the latest requirements of ENA. The finished dataset is archived in the corresponding data center with a persistent identifier (usually DOI for non-molecular or an accession number for molecular data) for a long time.

18 `https://www.pangaea.de/`; *Last accessed on March 29th, 2023.*

19 `https://www.ebi.ac.uk/ena`; *Last accessed on April 4th, 2023.*

20 `https://gfbio.biowikifarm.net/wiki/Major_Types_of_Biological_Data`; *Last accessed on July 31st, 2023.*

21 `https://gfbio.biowikifarm.net/wiki/Data_exchange_standards,_protocols_and_formats_relevant_for_the_collection_data_domain_within_the_GFBio_network`; *Last accessed on July 31st, 2023.*

22 `https://kb.gfbio.org/display/KB/Knowledge+Base`; *Last accessed on July 31st, 2023.*

23 `https://kb.gfbio.org/dosearchsite.action?queryString=submission&where=KB&additional=page+excerpt&labels=faq&contentType=page`; *Last accessed on July 31st, 2023.*

24 `https://gitlab.gwdg.de/gfbio/molecular-submission-templates/-/raw/master/full_template.csv?inline=false`; *Last accessed on March 31st, 2023.*

25 `https://gitlab-pe.gwdg.de/gfbio/molecular-submission-templates/-/blob/master/Template-Description.md`; *Last accessed on March 31st, 2023.*

3 Beyond the integration of the services in a single domain - the Helpdesk as the focal point

Feedback shows that the community appreciates the support provided by the Helpdesk and the human component of the assistance (as it is possible to address a contact person as an "assignee" on the ticket) during and after the consultation. Indicators of success for the team behind the Helpdesk are the loyalty of many users who have relied on GFBio for years, the dissemination among the scientific community evidenced in new requests submitted to the Helpdesk, the growing number of cooperations with projects and institutions, and the gratitude of users once their consultation is completed.

Working on projects such as the pilot with FEdA & KonsortSWD is a good example of the usability, potential, and expansion of the Helpdesk's model to also integrate interdisciplinary exchange in one single service point. In the case of FEdA, both the support in the creation and completion of DMPs as well as support through the Helpdesk has led to an extended cooperation formalized as a 'memorandum of understanding' between the speakers of FEdA and the NFDI4Biodiversity consortium for continuous and future support on the development of (interdisciplinary) data management support[26].

Naturally, there have been challenges when working in interdisciplinary RDM support. The social sciences have carefully designed workflows to guide their users through a multi-step process for archiving person-related research data early on in the research process, including support with informed consent, formalized study reports, and data transfer agreements. These are different from the archival support for biological data types, which is mostly scheduled toward the end of the research process and based on a structured DMP. Although the difference in expectations regarding the consultation process was initially challenging, we found the early engagement with archiving services a useful concept to improve relations with our core community. There have been mutual learning benefits throughout the process of encouraging researchers in their effort towards reaching FAIRness and producing consented interdisciplinary DMPs.Working together on data management planning has made us re-think the ways of incorporating RDM practices inside the scientific research process. We are taking action by providing workshops, having 1-to-1 support encounters with the researchers, and creating self-learning materials.

The NFDI4Biodiversity Helpdesk is also committed to providing further support beyond the consortium's offers. If there are inquiries in RDM (including DMPs) in another domain that reach the Helpdesk, the team will suggest or direct users to an appropriate point of support, such as the local RDM officer or experts in the respective field. The latter also implies communication with other NFDI Consortia. Accordingly, if certain submitted data do not thematically or qualitatively belong to any of the GFBio data centers, alternatives for archiving and publication are suggested to the submitter.

26 `https://www.nfdi4biodiversity.org/en/who-we-are/cooperation-feda/`; *Last accessed on April 3rd, 2023.*

4 Conclusions

With more than 100 requests processed since the NFDI4Biodiversity consortium began, its Helpdesk has proven successful as a central entry point of access to the consortium's services, tools, and expertise. Furthermore, the Helpdesk is also facilitating transparent communication among the community working in biodiversity, ecological, and environmental research.

This is also evidenced by the Front Office/Back Office. The pilot cooperations with the BMBF Research Initiative for the Conservation of Biodiversity (FEdA) and the UBRA Data Train are examples of how to build productive working relations with institutions, universities, and research centers, e.g. by allowing local RDM officers to create and address support tickets and to participate directly in the consultations together with domain experts from the consortium. Within NFDI4Biodiversity, we aim to encourage more universities and institutions to implement this model in their RDM support on-site and to expand the number of Front Offices step by step.

The pilot cooperation with FEdA is an example of how interdisciplinary research projects can be effectively supported in the National Research Data Infrastructure NFDI. As this cooperation continues and new project cohorts start, we will need to enlarge the group of partners (e.g. from the medical domain) due to the new disciplines involved.

We think that over time Helpdesk services like ours could consolidate across consortia, especially where similar data types or research methods are concerned. This would support the transition to a "ONE NFDI" with interoperable, user-facing services across consortia. Linking domain-specific and local or generic support structures is an opportunity to create workflows that make it easy for any researcher to find the best service for their data management issues.

Acknowledgments

The GFBio project as well as the NFDI4Biodiversity consortium have been funded by the German Research Foundation -DFG under the grant agreement numbers 229241684[27] and 408180549[28] (GFBio) and 442032008[29] (NFDI4Biodiversity). NFDI4Biodiversity is part of NFDI, the National Research Data Infrastructure Program in Germany.

References

Astor, Tina, Judith Weber, Ivaylo Kostadinov, Frank Oliver Glöckner, and Jens Nieschulze. 2021. "Potential für ein starkes Netzwerk zwischen GFBio und FDM-Bera-

27 `https://gepris.dfg.de/gepris/projekt/229241684`; *Last accessed on May 8th, 2023.*
28 `https://gepris.dfg.de/gepris/projekt/408180549`; *Last accessed on May 8th, 2023.*
29 `https://gepris.dfg.de/gepris/projekt/442032008?context=projekt&task=showDetail&id=442032008&`; *Last accessed on May 8th, 2023.*

tenden an Universitäten und Forschungsinstituten". *Bausteine Forschungsdatenmanagement,* number 1: 22–31. DOI: https://doi.org/10.17192/BFDM.2021.1.8311.

Klar, Jochen, Olaf Michaelis, David Wallace, Max Schröder, Enke Harry, Giacomo Lanza, David Martínez Muñoz, and Dario Pilori. 2023. *Research Data Management Organizer (RDMO).* DOI: https://doi.org/10.5281/ZENODO.596581.

Mau, Franziska, Britta Timmermann, and Tina Astor. 2020. *GFBio Model Data Management Plan (DMP).* DOI: https://doi.org/10.25625/W3YEEQ.

Weber, Judith, Barbara Ebert, Michael Diepenbroek, Ivaylo Kostadinov, and Frank Oliver Glöckner. 2021. "NFDI4BioDiversity – NFDI-Konsortium für Biodiversitäts-, Ökologische und Umweltdaten". *Bausteine Forschungsdatenmanagement,* number 2: 98–109. DOI: https://doi.org/10.17192/bfdm.2021.2.8334.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Breaking Down Hurdles of Current Data Citation Practices. Use Cases and Benefits of Persistent Identifiers for Dataset Elements

Janete Saldanha Bach, Claus-Peter Klas, Peter Mutschke

GESIS – Leibniz Institute for the Social Sciences

The paper introduces a service to assign Persistent Identifiers (PIDs) on the level of the inline data objects of a dataset, such as survey variables in the Social Sciences, resulting from the consortium KonsortSWD of the German National Research Data Infrastructure (NFDI). This technical solution aims to make data findability and accessibility on the lower granularity level of studies more efficient. In the Social Sciences, for instance, PIDs are commonly available on the study level, which is insufficient to unambiguously identify the dataset elements used in a paper and ensure an accurate data citation. By assigning PIDs to the fine-grained level of attributes, individual dataset elements can be referenced and retrieved with the required metadata. Referencing research data and their inherited entities by PIDs supports FAIR data usage, i.e., research data can be Findable, Accessible, Interoperable and Reusable. Textual data citations without a unique identifier are non-standard practices that lead to considerable time-consuming problems in unambiguously identifying relevant elements of a dataset and reusing them. From the technical perspective, it also hinders automated access to data elements below study level. Our PID service simplifies FAIR data management and benefit both researchers and research data centres (RDCs), fostering credibility results and ensuring the sustainable reusability of data. RDCs directly benefit from PIDs as they enable citation tracking and impact measurement, linking articles using the same dataset elements. It empowers the RDC's authority by demonstrating a commitment to best practices, enhancing its reputation in the research community by adopting recommendations to support PIDs at multiple granularity levels, such as the European Open Science Cloud (EOSC) PID policy. Furthermore, it promotes digital connections among researchers, organisations, and research outputs. Explicit relations between those elements are possible and favour the formation of a network into a knowledge graph representation. Since PIDs are machine-actionable, they are the technical bridges to the FAIR principles as they increase the traceability of research results.

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.c18063

1 Introduction

Persistent identifiers (PIDs) are the backbone of FAIR data infrastructures as they enable a reliable data citation. FAIR stands for the Findability, Accessibility, Interoperability, and Reusability of research data (Wilkinson et al. 2016). However, in many cases PIDs are only available on study or dataset level but not on the level of the inline data objects that are usually used by researchers. In the Social Sciences, for instance, survey datasets usually contain hundreds of so-called variables but usually only a few of them are used in a research article, making it difficult to clearly identify and reference the variables used, as PIDs are only assigned on study level and the variables used are described in various, semantically often ambiguous textual forms. Moreover, researchers may cite the data provider or papers referencing the data instead of the data itself, or they place footnotes, image captions, or acknowledgments, rather than locating citations in the reference lists (Gregory et al. 2023). Moreover, researchers from various fields often do not follow any standard, such as the Data Citation Principles (Data Citation Synthesis Group 2014).

These inconsistent data citation practices create significant challenges in reliably identifying and reusing the data underlying a paper. Thus, current data citation practices often lack unique identifiers for relevant dataset elements, making it difficult for researchers to cite their data in a reliable way, for other researchers to reuse the data and for data providers to identify and annotate the important elements of datasets.

This paper introduces a PID service resulting from the consortium KonsortSWD[1] of the German National Research Data Infrastructure (NFDI[2]). The primary objective of this service is to enhance the reusability and findability of data by focusing on a more detailed level of datasets. The service assigns PIDs to specific, fine-grained dataset elements which represent the primary entities of research, such as survey variables in the Social Sciences, making it easier to reference and find relevant data objects. The PIDs are retrieved with the necessary metadata, facilitating both machine-actionable and human access. This metadata provides essential information about the data element, enabling users to understand its context and relevance. By incorporating PIDs and metadata, it ensures that users can effectively comprehend and utilize the retrieved information, in a more efficient and user-friendly way to manage and access data at a granular level. Hence, this detailed citation approach ensures data provenance, findability, and accessibility, fostering trust and promoting efficient data reuse.

The paper demonstrates how the service can simplify FAIR research data management at lower granularity levels, in section 2. Section 3 explains the PID registration service provider and the process of assigning PIDs for dataset elements. Section 4 details different social science for assigning PIDs to dataset elements below study level, and section 5

1 KonsortSWD (Consortium for the Social, Behavioural, Educational and Economic Sciences) is funded by the National Research Data Infrastructure (NFDI). KonsortSWD Homepage: `https://www.konsortswd.de`.

2 German National Research Data Infrastructure (NFDI) Homepage: `https://www.nfdi.de`.

concludes with the key contributions of the services for researchers and research data centres (RDCs).

2 PIDs simplify FAIR research data management at lower granularity levels

A PID is a persistent, unique, and globally resolvable identifier based on an openly specified PID Scheme, which allows for reliable and lasting reference to the associated research outputs (European Commission. Directorate General for Research and Innovation and Board. 2020). PIDs serve as the foundation for the long-term referencing of scientific publications, ensuring the consistent identification of digital objects. In the context of Social Sciences, for instance, research outputs have a range of granularity levels. The study and dataset represent the most common granularity levels identified with Persistent Identifiers (PIDs) when researchers publish their findings. However, datasets on survey data in the Social Sciences typically comprise questions, variables, variable values, indicator values or scales, and researchers are interested in the content of the variables. To this end, a more significant effort is necessary to understand the variable content meaning and its values. Figure 1 depicts granularity levels of research data PID are commonly used.

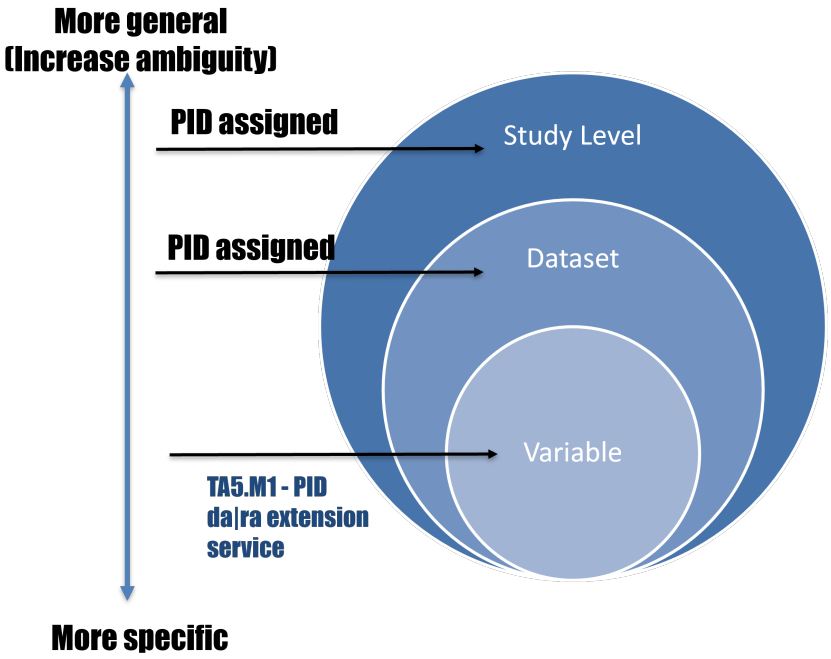


Figure 1: The Research data granularity levels (for the case of survey data in the Social Sciences).

In terms of data citation, once researchers locate and obtain a dataset of interest, they experience a long and complex process to extract the most relevant information and have to analyse data documentation exhaustively to find relevant dataset elements for their

research (Bensmann et al. 2020). In the following example we considered the case of a survey variable in the Social Science as the relevant dataset element in question. If a PID at the that level is unavailable, researchers also must:

1. Locate data citations in the paper: Researchers must identify data citations in the text of relevant studies, examining citations, quoted questions, or other hints such as websites or dataset provider institutions.
2. Identify and access the data source: After finding an interesting dataset, researchers should locate and access it on the data provider's website.
3. Verify dataset version: Researchers must ensure they are using the correct version, as some studies may cite earlier dataset versions.
4. Review data documentation: Users should examine data documentation and draw inferences.
5. Find matching dataset elements in the documentation: Users must identify variables that correspond to the referenced data in the paper.
6. Obtain dataset access: Researchers need to familiarize themselves with access rules (open, limited, or restricted/sensitive) and apply for access if necessary.
7. Learn how to open dataset files: Users must determine the appropriate software or driver for opening files, based on the file format.
8. Download the dataset: After meeting all requirements, users can download the dataset for further use.
9. Open the dataset and identify elements and their values: Users can manually locate dataset elements or use statistical software commands, depending on the file type (CSV, spreadsheets).
10. Apply statistical analysis: Users must understand dataset elements values and analyze the information accordingly.
11. Reuse variables: Users should generate new insights and alternative analyses using the same dataset.
12. Cite the data: Without a PID, users may cite the dataset name, provider name, or a report or study where the dataset elements were published, continuing the non-standard citation cycle.

Figure 2 illustrates the process of accessing and reusing dataset elements (here, survey variables) without PIDs. In contrast, assigning PIDs to identify dataset elements will simplify FAIR data management at lower level in three aspects:

1. boosting subsequent citation,
2. getting direct (meta)-data access, and
3. promoting data reuse.

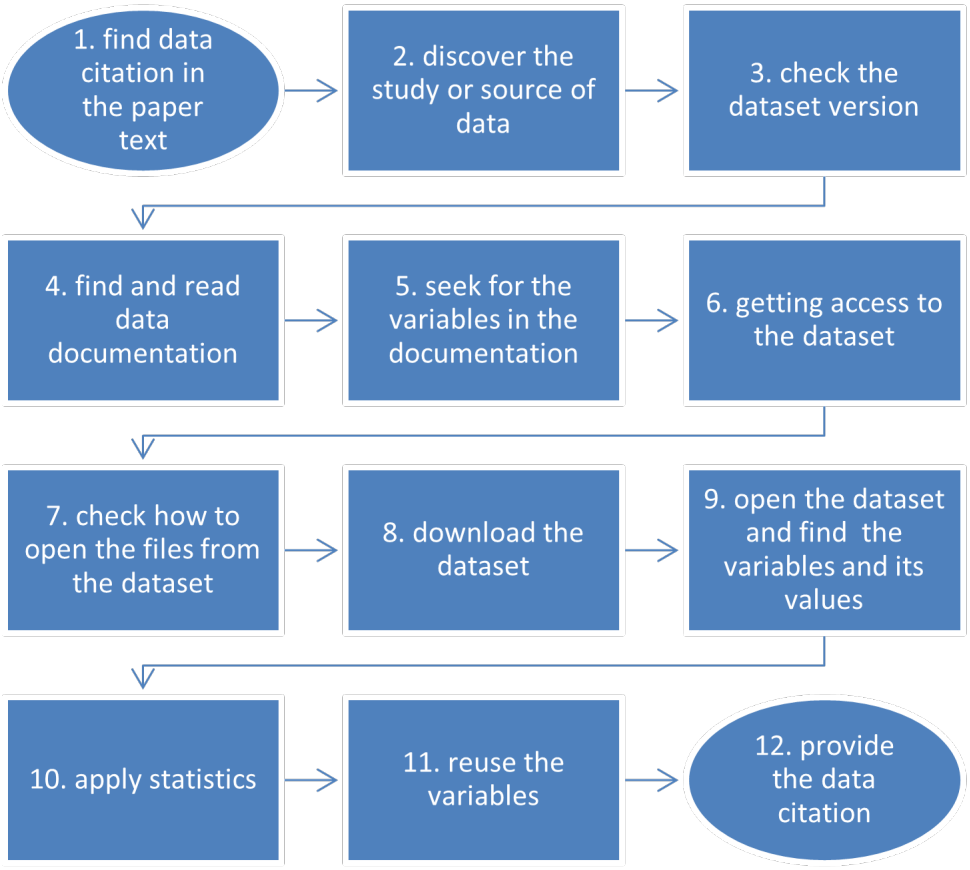


Figure 2: Steps to be taken for accessing and reusing dataset elements, here survey variables, without the availability of PIDs at dataset element level.

If one PID is registered for each element, it can (1) boost subsequent citation. In this case, automated scripts (do-files, R scripts, etc.) can be applied (2) get direct (meta)-data access, obtaining the selected element from the dataset automatically. The automatic access to the data in a dataset is enabled by just executing a script (i.e., using R, Stata, SPSS) that resolves a given PID and returns the data "behind" the PID in a proper format. One of the most common data formats in the Social Sciences are surveys and questionnaires results, a tabular dataset frequently stored in statistical packages files such as R or Stata, as well as spreadsheets or comma-separated values (.csv) files. Variables are distributed as rows (which contain objects) and columns (which contain properties) within datasets.

Some conditions are required to automatic access, which relies on the dataset's PID (typically a DOI) to retrieve the data. However, if multiple datasets are associated with the same DOI, this method will fail as the script cannot distinguish which dataset to access. And there are DOIs registered for dataset collections. For the automatic access

function to work effectively, only one dataset must be registered per PID, which is used to register the PID for a variable.

The dataset might have not restricted or closed access, and a REST API is available (Klas and Hopt 2022). Once the technical requisites are met, these automated scripts can technically give access to the dataset elements for direct usage without downloading the entire data file, either the complete dataset, but singled-out elements using PIDs (Klas, Saldanha Bach, and Mutschke 2023). The following workflow (Figure 3) depicts the process of accessing and reusing elements with PIDs. Researchers can take advantage of these machine-actionable features when a dataset element is identified with a PID. Getting data through automated access is faster and (3) promote data reuse, going through fewer steps if a PID is unavailable, compared to Figure 2.

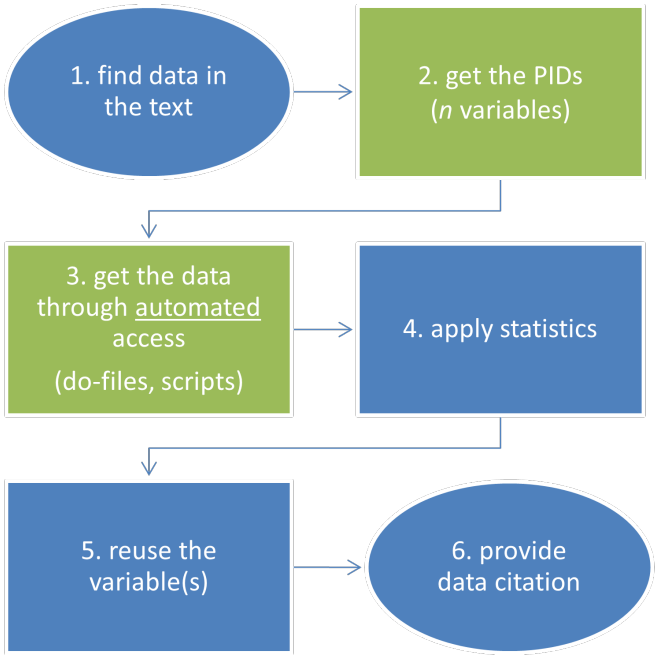


Figure 3: Accessing and reusing dataset elements with a PID.

3 PID registration

The PID service developed in the context of KonsortSWD is a technical solution aiming to make data findability and accessibility on the lower granularity level of studies, here survey variables in the Social Sciences, more efficient. To use the service, data holders (such as research data centers) must be registered in advance and authenticated within the PID registration service. Since a study may contain numerous dataset elements, an automated method for bulk PID registration is available. Using a script or integrated software within the documentation tool, all elements can be registered within the service. The request of PIDs is a task for data providers. In order to get as many PIDs as needed, the data provider must submit in the registration service a minimal set of metadata (Bach, Klas, and Mutschke 2023), including the suggested PID, landing page, original dataset PID

(commonly, a DOI), and other relevant metadata fields to identify each dataset element. The registration service then validates the metadata, confirms the registered study PID, and stores the metadata. Finally, the data provider includes the PID on each dataset element's landing page for citation purposes. As many variables exist within a study, an automated way to register PIDs as bulk is available. All variables can be registered through a script or integrated software in the infrastructure's documentation tool, which means the registration of many variables at once. To this end, any data provider can register an arbitrary number of variables (bulk registration) through one REST API endpoint using a REST client. This process automates the registration, avoiding much work for the data provider. See the first service report (Klas et al. 2022) for details.

The PID registration for lower-level elements assign Handle[3] PIDs, supported by the third-party Persistent Identifier Consortium for eResearch (ePIC)[4] API[5] registration service. The system is conceived to provide a general, maintainable, and scalable infrastructure that enables the registration of PIDs to the level of attributes.

4 Use cases in the Social Sciences

In the following we discuss use cases in the Social Sciences collected in the context of KonsortSWD, demonstrating the benefit of having PIDs on dataset element level. The PIDs registration service benefit research data centres (RDCs), fostering credibility results and ensuring the sustainable reusability of data. RDCs directly benefit from PIDs as they enable citation tracking and impact measurement, linking articles using the same dataset elements. It empowers the RDC's authority by demonstrating a commitment to best practices, enhancing its reputation in the research community by adopting recommendations to support PIDs at multiple granularity levels, such as the European Open Science Cloud (EOSC) PID policy (European Commission. Directorate General for Research and Innovation and Board. 2020). Furthermore, it promotes digital connections among researchers, organisations, and research outputs. Explicit relations between these elements are possible and favour the formation of a network into a knowledge graph representation. The agreed use cases are selected partners institutions participating in the KonsortSWD and play an essential role in shaping the service, testing the concept, and providing helpful feedback on the RDC daily activities associated with the Persistent Identifiers. The following use case descriptions are the Higher Education Analytical Data System (HEADS)[6] project from the German Center for Higher Education Research and

3 Handle System technology resolves PIDs such as Handles and DOIs. The Handle System was developed by Corporation for National Research Initiatives (CNRI) and is currently administered and maintained by the DONA Foundation. Handle.Net Registry (HNR) Homepage: `https://www.handle.net`.

4 ePIC is an international consortium provides a reliable Handle-based PID infrastructure for research data. ePIC has currently nine members and it is open for any center that stores scientific/research data. ePic Homepage: `http://www.pidconsortium.net`.

5 ePic documentation Homepage: `https://doc.pidconsortium.eu/docs`.

6 The German Center for Higher Education Research and Science Research (DZHW) Homepage: `https://www.dzhw.eu/gmbh/index_html`.

Science Studies (DZHW); the GESIS Search[7] from the Leibniz Institute for the Social Sciences (GESIS), including the GESIS harmonisation tool: QuestionLink[8], the German Socio-Economic Panel (SOEP-Core) (Liebig et al. 2022), a longitudinal study from the German Institute for Economic Research (DIW), and the Qualiservice[9] qualitative data collection from the University of Bremen.

4.1 HEADS project from the DZHW

For the Higher Education Analytical Data System (HEADS) project at the German Center for Higher Education Research and Science Studies (DZHW) A standard data citation system is essential to make HEADS results widely usable and citable. The PID (Persistent Identifier) system is particularly well-suited for this purpose, as it assigns PIDs to (1) individual variables and (2) comprehensive information packages, which include a central reporting variable ("indicator") and related multivariate analyses conducted in HEADS. Both professionals and the interested public will benefit from using and citing data with PIDs. The dependent variable comprises several items from a larger theoretical construct. Each variable classified as "indicator" (see Figure 4) named ziwahr01 to ziwahr5 gets a PID.

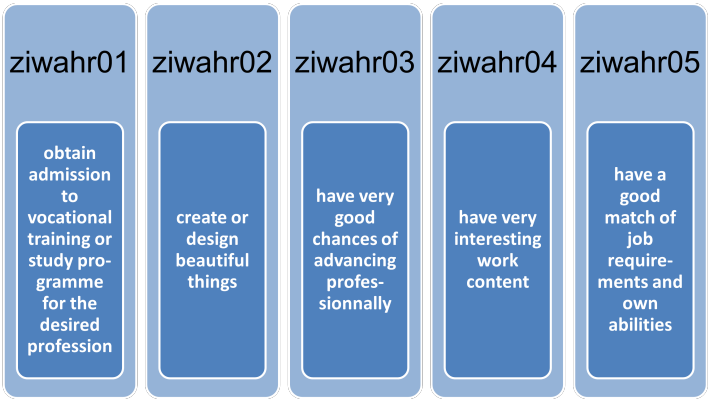


Figure 4: PIDs for each indicator variables.

In this use case, there are two ways to assign a PID: dependent variables can consist of multiple items, allowing for various intersections. Numerous independent variables enable differentiation of the dependent variable according to subgroups: gender, educational background, migration background, and school types. In this example, the dependent variable pertains to the target group's goals and achievements, operationalized as five "ziwahr" variables. These variables are analyzed concerning the independent variables of gender, educational background, migration background, and school types (see Figure 5).

7 GESIS – Leibniz Institute for the Social Sciences Homepage: `https://www.gesis.org/home`.

8 QuestionLink Homepage: `https://www.gesis.org/angebot/daten-aufbereiten-und-analysieren/question-link`.

9 Qualiservice - the data service center for qualitative social science research data is a data service center for archiving and sharing qualitative research data in the social sciences. Qualiservice Homepage `https://www.qualiservice.org`.

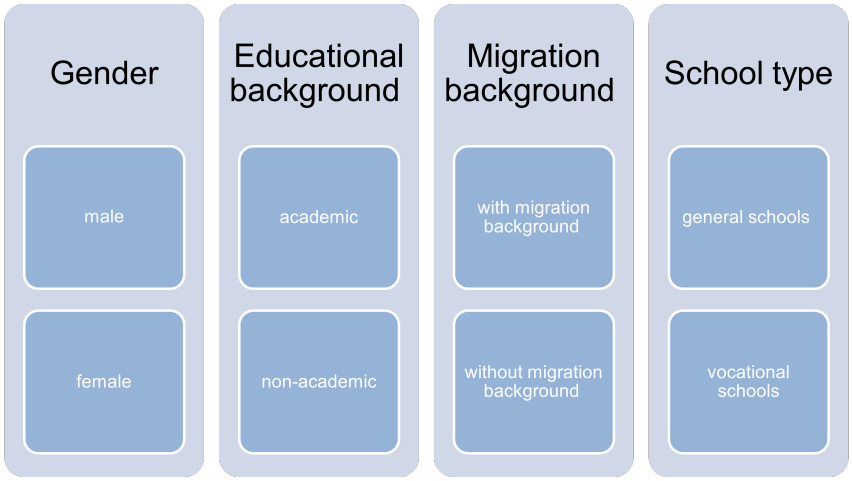


Figure 5: PIDs for each differentiation variables.

An information package is provided, displaying the values of the dependent variables (ziwahr01-05) according to the differentiation variables (gender, educational background, migration background, school types). For users, the critical aspect is the availability of a permanent location where the information, differentiated for subgroups or various differentiation variables, can be found and cited. It is immaterial whether users wish to use only the value of a subgroup, compare subgroups, or compare differences based on different differentiation variables. Consequently, a PID should be assigned to the information package to accommodate these various purposes (Figure 6).

As the PID is always connected to a "Landing Page", the data holder can define the level of data to be assigned a PID and describe it on the landing page.

4.2 GESIS Search

The GESIS Search[10] is a platform providing an integrated search across more than 6,500 national and international quantitative social science studies (mainly survey studies), more than 500,000 variables from those studies as well as instruments & tools and open access publications. The GESIS Search also provides links between diverse types of entities. However, PIDs are not assigned to variables so far. Applying PIDs to variables focuses on enhancing precise citation for secondary data analysis, as well as improving data discoverability and accessibility through automated access. With the accumulation of large datasets across studies and waves[11], dataset versions change over time. PIDs offer a more accurate way to distinguish repeated variables across the years, while ensuring direct access through automated features.

10 The GESIS Search is a Search Portal provided by GESIS to find information about social science research data and publications. GESIS Search Homepage: `https://search.gesis.org`.

11 Waves are different points in time when data is collected in a research study. Waves are typically associated with longitudinal studies, which involve the repeated observation of the same subjects over time.

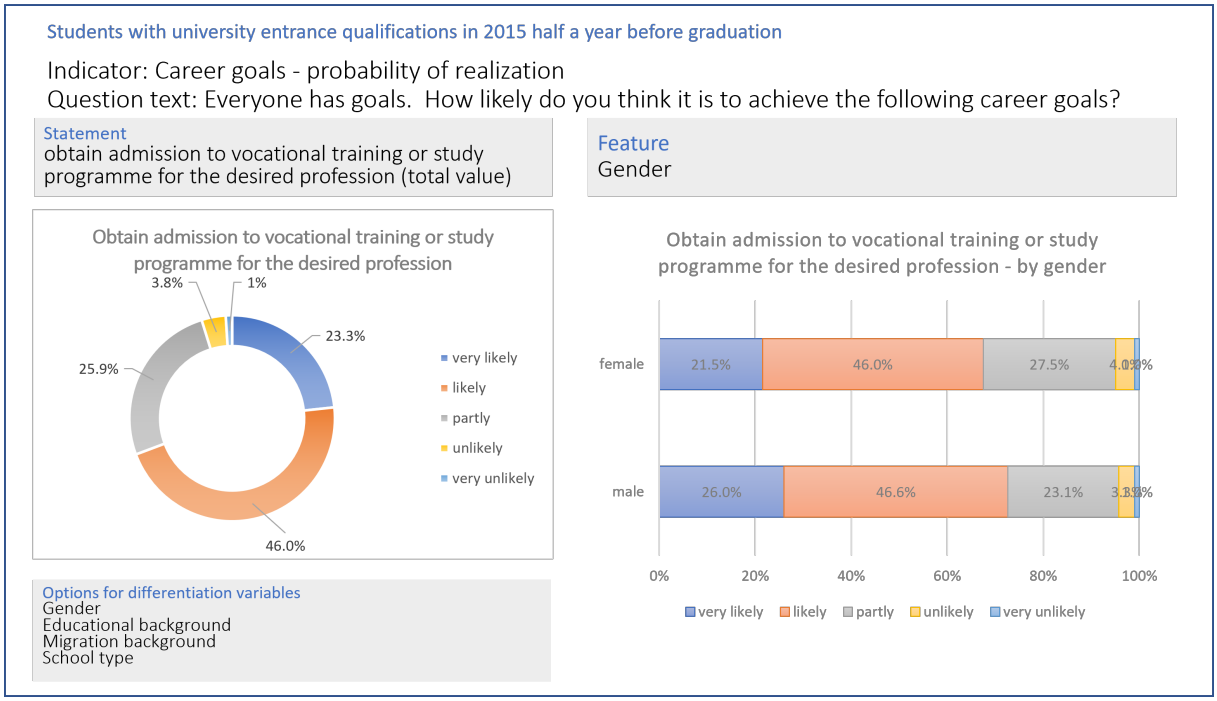


Figure 6: PIDs for information packages.

The direct access feature (automated access, i.e., by a computer program) is particularly relevant for GESIS for harmonisation tools within their code packages. These tools offer scripts and do-files that calculate response scales and provide harmonized measures and equivalent measures for similar variables across different studies, rather than providing data directly. Researchers are responsible for accessing datasets themselves from data providers. By assigning unique identifiers to each variable and embedding the variable's PIDs in these codes, it would simplify the use of numerous harmonized variables on the same topic from distinct sources. GESIS leads several projects that utilize such harmonisation tools.

4.2.1 GESIS harmonisation tool: QuestionLink:

QuestionLink is a tool that harmonizes sixty-eight political interest variables from seven measurement instruments: ALLBUScompact, GLES, GPANEL, ISSP (1990), ESS, NEPS, and SOEP. It helps researchers find and pool German data on political interest constructs over time and across large survey programs. However, accessing relevant data and applying the correct recording script require researchers to identify and retrieve the specific variables in source datasets. PIDs at the variable level would enhance QuestionLink and simplify its use.

For the selected instruments used sixty-eight political interest variables for harmonisation purposes, some surveys have the same variable name for multiple years, while others have different names per wave. Since question formulations and response labels may slightly differ across data collections while registering the same concepts, there is an elevated risk

Table 1: QuestionLink Example 1: Variable name similarity in the ALLBUS survey.

Survey	Instrument	Wave	Year range	Variable name	DOI
ALLBUS	ALLBUS B 10pt	cumulation	1982 – 1988	pa02	10.4232/1.13775
ALLBUS	ALLBUS A 5pt	cumulation	1980 – 2018	pa02a	10.4232/1.13775

of ambiguous citation when differentiating variables used in the same survey over the years.

Table 1 highlights the similarity in variable names within the ALLBUS instruments across different year ranges. The PID (DOI) identifies only the dataset, not the variable, and their names are similar.

Assigning PIDs for pre-harmonisation variables in the QuestionLink tool uniquely identifies individual variables per survey, wave, and year, preventing ambiguity due to repetitive variable names, which can be confusing and misleading.

4.3 SOEP-Core from DIW

The use case SOEP-Core[12] from the German Institute for Economic Research (DIW) encompasses various sub-samples and questionnaires related to households and individuals' members from Germans living in the former eastern and western German states, but also foreign citizens, and immigrants residing in Germany. Questions focus on finances, utilities, and general living conditions, while personal questionnaires explore work life, leisure activities, political interests, new family members, children's education, and youth-specific topics. The SOEP-Core consists of 101.574 variables, available from 560 data collections, distributed in 21.280 questions, and 309 instruments. The DIW office in Berlin documents the SOEP-Core complexity information extensively, covering topics, survey design, data editions, and distribution files. Panel data's interface[13] offers data and variable details, including variable landing pages. For example, the variable "*Interest in Politics*" (see Figure 7), with a landing page displaying variable values and timeline relations.

The SOEP-Core features a complex data structure with numerous datasets and variables across long-term investigations. Assigning a PID to identify these variables would lead the institute to utilize machine-actionable features to track and monitor the scientific output of specific variables. As the study covers various themes, PIDs enable tracking variable usage by subject and target types, such as household or individual-related information

12 German Socio-Economic Panel Study (SOEP-Core) Homepage: `https://paneldata.org/soep-core`.

13 Variable bip/bip_171: Interest in Politics from the Panel data: `https://paneldata.org/soep-core/datasets/bip/bip_171`.

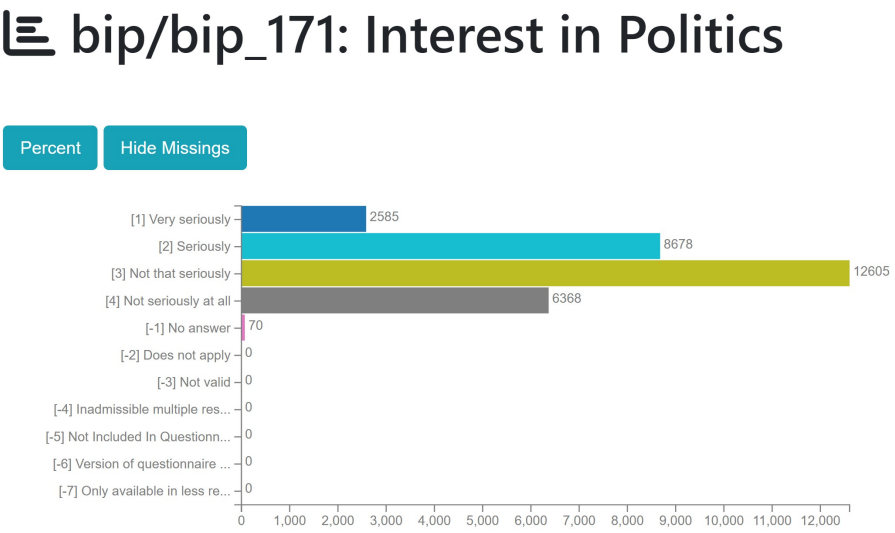


Figure 7: Variable graph: bip/bip_171: *Interest in Politics*.

in academic publications. This detailed tracking supports better evaluation of dataset usage, leading to improved decision-making regarding data services. Also, search and connect information from other datasets using the same variable under different labels. This approach is especially relevant for variables used in the harmonisation process within the same study or across different studies, such as the *Political Interest* variables used in the QuestionLink tool. Many variables are documented already and have a unique landing page. Registering this page as the variable's PID landing page would expedite the PID registration process, automatically linking to related datasets and enhancing findability.

4.4 Qualiservice

The use case Qualiservice is a data service centre from the University of Bremen. Its research data primarily includes qualitative interview transcripts and contextual data, which document the primary data collected during the research process. Qualiservice structures and standardizes data through metadata, registering elements according to the DDI 3.2 standard. PIDs (DOIs) are now assigned to identify dataset elements rather than study levels. A data collection, or dataset, can encompass various data types (interviews, observations), file versions (video, text, audio), or be organized by specific survey methods. PIDs for elements are assigned at the file level to distinguish between similar data types and file names, offering a direct method for citing, identifying, and accessing the target dataset element.

Considering the complex data structure encompassing a wide range of data files and formats, assigning a PID to each dataset element will simplify FAIR management of Qualiservice data. Relevant use cases include that Qualiservice's data structure varies from tabular data due to attributes like text, videos, and descriptive data. It demonstrates the PID service flexibility, including also qualitative data. PIDs can be assigned to identify

the dataset element that is considered most relevant to the data provider. Data files grouped in sets, such as datasets or collections, can maintain PIDs to identify related data from a specific study. Also, assigning a PID at the file level can be beneficial for disambiguating similar data files, given their data types and file naming similarities. This unique identifier gives end-users a straightforward method for citing, identifying, and accessing the target file. Since PIDs are machine-actionable elements, Qualiservice can leverage automatic features when assigning PIDs at the file level, such as linking related files within the same study or from different studies and directly accessing them.

Each use case has its unique aspects, as illustrated in previous examples. However, common advantages and benefits for RDCs and their users are consolidated in the conclusion section.

5 Conclusions

PIDs for lower granularity levels enhance FAIR data management, enabling Research Data Centers (RDCs) to benefit from the machine-actionable features of PIDs. By efficiently promoting data findability and accessibility at lower granularity levels, RDCs can make more informed decisions regarding services based on data utility, streamline data governance activities, and potentially reveal relationships between dataset elements across studies and datasets. This information lays the groundwork for knowledge graph visualization and fosters digital connections among researchers, organizations, and research outputs. Additionally, PIDs simplify harmonisation processes, which are often costly and time-consuming.

Regarding FAIRness, PIDs for lower granularity levels offer numerous benefits for researchers and data providers. They simplify FAIR data usage by providing unique identifiers for data elements below the study level, such as survey variables. PIDs enable referencing and retrieval of individual elements and metadata retrieval for data elements below the study level. They also help disambiguate data citations, promote safe and accurate data citations, and enhance recognition of produced data. Furthermore, PIDs foster credibility in research results and ensure the sustainable reusability of data while reducing documentation complexity. Additionally, PIDs offer feasible identification for various data types, including non-rectangular data attributes such as text, videos, and descriptive data.

Adopting PIDs to reference research data and their associated entities promotes FAIR data usage, as it significantly improves data findability, allows for more straightforward and, under certain conditions, automated access to data. Moreover, it enhances interoperability on a large scale by connecting dataset elements and other individual components, encourages data reuse, and simplifies the reproducibility of research. These benefits contribute to a more effective and efficient research ecosystem that fosters collaboration and knowledge sharing.

Future work. Dataset elements are interdependent and connected across studies and research outputs. In our approach, relationships between elements such as variables and

other attributes can be established, which we intend to incorporate in a knowledge graph representation of Social Science survey studies. PIDs, being machine-actionable, serve as the technical bridges that adhere to the FAIR principles, thus enhance the traceability of research results. By creating these connections and fostering a more comprehensive network, we can effectively improve the organization, accessibility, and overall understanding of research outcomes in these disciplines.

Acknowledgements

KonsortSWD is funded by the German Research Foundation (DFG) within the framework of the NFDI – project number: 442494171.

References

Bach, Janete Saldanha, Claus-Peter Klas, and Peter Mutschke. 2023. *KonsortSWD Measure 5.1: metadata schema extended report.* DOI: https://doi.org/10.5281/ZENODO.7588902.

Bensmann, Felix, Andrea Papenmeier, Dagmar Kern, Benjamin Zapilko, and Stefan Dietze. 2020. "Semantic Annotation, Representation and Linking of Survey Data". In *Semantic Systems. In the Era of Knowledge Graphs,* 53–69. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-59833-4_4.

Data Citation Synthesis Group. 2014. "Joint Declaration of Data Citation Principles". DOI: https://doi.org/10.25490/A97F-EGYK.

European Commission. Directorate General for Research and Innovation and EOSC Executive Board. 2020. *A Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC).* Publications Office. DOI: https://doi.org/10.2777/926037.

Gregory, Kathleen, Anton Boudreau Ninkov, Chantal Ripp, Emma Roblin, Isabella Peters, and Stefanie Haustein. 2023. *Tracing data: A survey investigating disciplinary differences in data citation.* DOI: https://doi.org/10.5281/ZENODO.7555266.

Klas, Claus-Peter, and Oliver Hopt. 2022. "DDI Variable Documentation and data access using R". DOI: https://doi.org/10.5281/zenodo.7408629.

Klas, Claus-Peter, Janete Saldanha Bach, and Peter Mutschke. 2023. "GESIS Use case Variable publication and citation & Fine granular access to research data". DOI: https://doi.org/10.5281/ZENODO.7750031.

Klas, Claus-Peter, Matthäus Zloch, Janete Saldanha Bach, Erdal Baran, and Peter Mutschke. 2022. *KonsortSWD Measure 5.1: PID Service for variables report.* DOI: https://doi.org/10.5281/ZENODO.6397367.

Liebig, Stefan, Jan Goebel, Markus Grabka, Carsten Schröder, Sabine Zinn, Charlotte Bartels, Andreas Franken, et al. 2022. *Sozio-oekonomisches Panel, Daten der Jahre 1984-2020 (SOEP-Core, v37, Onsite Edition).* DOI: https://doi.org/10.5684/soep.core.v37o.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Datenmanagementplan und Publikation von Forschungsdaten im Projekt „Emissionsminderung Nutztierhaltung“ EmiMin: Planung und Realität – Umsetzbarkeit von Forschungsdatenmanagement

Ewald Grimm[1], Birte Lindstädt[2], Katrin Wagner[1], Roman Riedel[2]

[1]Kuratorium für Technik und Bauwesen in der Landwirtschaft e.V.;
[2]ZB MED – Informationszentrum Lebenswissenschaften

Im Rahmen der E-Science-Tage 2019 wurde das Projekt „Emissionsminderung Nutztierhaltung – Einzelmaßnahmen“ (EmiMin[1]) als ein Anwendungsbeispiel für das Forschungsdatenmanagement vorgestellt. Ziel des Projekts ist die messtechnische Untersuchung der Wirksamkeit neuer baulich-technischer Maßnahmen zur Emissionsminderung in der Nutztierhaltung. Darüber hinaus sollten die Forschenden während des Projekts in Fragen des Forschungsdatenmanagements begleitet sowie Empfehlungen und Tools für den Umgang mit Forschungsdaten erarbeitet werden. Neben fünf Instituten aus dem Forschungsfeld der landwirtschaftlichen Verfahrenstechnik beteiligt sich das ZB MED – Informationszentrum Lebenswissenschaften als „Forschungsdatenmanager“ an der Durchführung des Projekts.

1 Umsetzung

Die Forschungsdaten werden von Projektbeginn an in einen Managementprozess eingebunden. Dieser umfasst die Standardisierung der Messungen über ein vorgegebenes Protokoll sowie die Dokumentation der Mess- und Begleitdaten in einer Datenbank des Kuratoriums für Technik und Bauwesen in der Landwirtschaft e.V. (KTBL). Mit Blick auf die gute wissenschaftliche Praxis werden ausgewählte Forschungsdaten gemäß der FAIR-Prinzipien (Wilkinson u. a. 2016) im Fachrepositorium Lebenswissenschaften (FRL) publiziert und somit nachnutzbar gemacht. Um das Forschungsdatenmanagement von Projektbeginn an umfassend zu begleiten und zu unterstützen, wurde mithilfe des DFG-geförderten Tools

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.c18064

1 https://www.ktbl.de/themen/emimin

Research Data Management Organizer (RDMO[2]) kooperativ ein Datenmanagementplan erstellt.

Hierfür erlaubt RDMO4Life[3] als Instanz von ZB MED das Erfassen aller relevanten Planungsinformationen in Datenmanagementplänen sowie die Verwaltung aller Datenmanagementaufgaben über den gesamten Datenlebenszyklus hinweg.

An die Erstellung eines projektbegleitenden Datenmanagementplans und die Publikation der Mess- und Begleitdatendaten wurde eine Reihe von Zielen und Erwartungen geknüpft. Beispielsweise war die Weiterentwicklung von Tools für das Forschungsdatenmanagement in engem Austausch mit Forschenden aufseiten der Infrastruktur wichtig. Die Forschenden erwarteten u.a. Mehrwerte im Hinblick auf Reputationsbildung durch Datenpublikationen sowie den Erhalt nachnutzbarer, fachspezifisch angepasster Tools und Vorlagen für weitere Drittmittelanträge, um Forderungen von Förderern bezüglich des Datenmanagements entsprechen zu können.

Der Beitrag stellt in einer Art Soll-Ist-Vergleich die geplanten Ziele und Erwartungen der tatsächlichen Umsetzung gegenüber. Dabei werden Hürden aufgezeigt und Empfehlungen abgeleitet. Unterschieden wird zwischen der Arbeit an den eigentlichen Informationsinfrastrukturen, dem Fragenkatalog in RDMO4Life bzw. dem fachspezifischen Metadatenschema im Fachrepositorium Lebenswissenschaft, sowie dem Prozess der Zusammenarbeit von Infrastruktur und Forschenden.

Im Hinblick auf die Informationsinfrastrukturen wurde der generische RDMO-Fragenkatalog durch fachspezifische Anpassungen im Rahmen von RDMO4life weiterentwickelt und dadurch ein Datenmanagementplan für aktives Forschungsdatenmanagement in der Agrartechnik bereitgestellt. So wurden beispielsweise Fragen zum Testdesign, zu Art und Anzahl der in den Betrieben gehaltenen Tierarten, zu in den Messverfahren gemessenen Emissionsvariablen oder zur Methode der Ermittlung des Luftvolumenstroms hinzugefügt. Ebenso wurden die Bedingungen am Messort, die Einfluss auf das Ergebnis haben können – beispielsweise die Fütterung, die Hauptwindrichtung oder die Belüftung – thematisiert. Durch die Weiterentwicklung des Datenmanagementplans wurde das Bewusstsein der Forschenden für die Themen Forschungsdatenmanagement und Datenmanagementpläne geschärft und die unterschiedlichen Blickwinkel der Forschenden sowie der Forschungsdatenmanager deutlich. Auch zeigte sich, dass die umfassende Darstellung eines Projektes in einem Datenmanagementplan eine hohe Detailtreue und Ausdifferenzierung bedarf. Darüber hinaus sind für die Entwicklung eines Template für die ganze Agrarwissenschaft die Berücksichtigung von Besonderheiten, Anliegen und Richtlinien anderer Teildisziplinen notwendig.

Auch das Metadatenschema im Repositorium wurde in Abstimmung mit den Forschenden fachspezifisch angepasst. In gemeinsamen Treffen mit allen Projektbeteiligten wurde das Metadatenschema in diesem Zuge um sogenannte projektspezifische Metadaten erweitert. So wurde das Schema um Angaben u.a. zur Produktionsrichtung, zum Haltungsverfahren,

2 `https://rdmorganiser.github.io`
3 `https://rdmo.publisso.de`

möglichen Minderungsmaßnahmen oder zum Emissionsmessverfahren angereichert. Diese Angaben dienen dazu, den Datensatz verständlicher sowie kognitiv zugänglicher und damit im Sinne der FAIR-Prinzipien leichter nachnutzbar zu machen.

Aber auch das bereits bestehende Metadatenschema für Forschungsdaten – im internen Sprachgebrauch in Abgrenzung zu den neu konzipierten projektspezifischen Metadaten als generische Metadaten bezeichnet – wurde erweitert. Fortan ist es z.B. möglich, den Autoren und Mitwirkenden eines Datensatzes Rollen zuzuweisen, um die jeweilige Funktion innerhalb des Projekts bzw. bei der Erstellung des Datensatzes noch treffender benennen zu können. Darüber hinaus sind durch die Erweiterung zusätzliche Erhebungsmethoden auswählbar, um nicht zuletzt agrarwissenschaftliche Forschung entsprechend in den Metadaten abbilden zu können.

Es zeigte sich, dass für eine angemessene Beschreibung eines Datensatzes viele neue Metadatenfelder notwendig sind. Entsprechend müssen hinsichtlich thematisch anders gelagerter agrarwissenschaftlicher Forschung zusätzliche Metadatenfelder entwickelt werden. Die Diskussionen um die Erweiterung des Metadatenschemas sorgten dafür, dass Forschende ein erstes Gespür für eine Datenpublikation bekommen und ZB MED einen Eindruck davon, was Forschenden mit Blick auf eine Publikation ihrer Daten wichtig ist.

2 Erkenntnisse

In Bezug auf den Austauschprozess konnten die Forschenden durch die Arbeit am Forschungsdatenmanagement über die gesamte Projektlaufzeit hinweg, das z.B. in Workshops gemeinsam realisiert wurde, praxisnah an das Thema Forschungsdatenmanagement herangeführt und mit Tools und Herangehensweisen für kommende Projekte ausgestattet werden. ZB MED wiederum konnte Erkenntnisse zum Bedarf von Forschenden gewinnen und Beratungs- und Servicedienstleistungen zur Verfügung stellen, diese evaluieren und gemäß dem Bedarf anpassen. Als eine wesentliche Hürde in diesem Prozess entpuppte sich beispielsweise das Verständnis für eine Datenpublikation im Unterschied zu einer Textpublikation. Um dem zu begegnen und zur Abstimmung des Workflows bei der Publikation der Forschungsdaten im FRL wurden Daten eines Teilprojekts testweise in einer Pilotpublikation veröffentlicht.

Auch der geplante automatisierte Transfer aller Publikationsmetadaten aus der KTBL-Datenbank erwies sich sowohl inhaltlich als auch organisatorisch als aufwendiger Prozess, in dem zunächst das Verständnis auf Seiten der Forschenden und ITler für die Erfordernisse einer Datenpublikation und die zugehörigen Metadaten geschaffen werden musste.

Danksagung

Die Förderung des Projekts erfolgte aus Mitteln des Zweckvermögens des Bundes bei der Landwirtschaftlichen Rentenbank sowie aus eigenen finanziellen Anteilen der beteiligten

Organisationen vor Ort. Zusätzlich erhielt das Projekt eine Förderung des Bundesministeriums für Ernährung und Landwirtschaft.

Literaturverzeichnis

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Automating DOI Registration with DataCite API

Giuditta Parolini, Falko Glöckler

Research Data Management Services, Museum für Naturkunde Berlin – Leibniz Institute for Evolution and Biodiversity Science

Automation is a main trend in all contexts in which data are a valuable asset. Sustained growth in the amount of available data, greater interest in extracting knowledge from data, and the significant increase in data regulations and policies are constantly shifting the balance from human-based to automated data management solutions. Research data management is also experiencing this automation trend. The tasks of research data management teams have been constantly expanding in recent years. More demanding funder policies regarding research data and growing needs on the side of researchers, who are dealing with scientific data in quantities never experienced before, are constantly increasing the workload of research data management teams. Automation becomes, therefore, a necessary choice to avoid limiting or reducing the services provided to users.

The paper will discuss an example of automation implemented to meet the significant increase in DOI (Digital Object Identifier) registration requests for data publications received at the Museum für Naturkunde Berlin (MfN). It will present the web application that the MfN research data management team has been building to automate the DOI registration process with the provider DataCite. Running in a Docker container, a Django web application calls the DataCite API and interoperates with MfN institutional databases to prompt the DOI creation and to collect the metadata that are required for DOI registration. The paper will reflect on the code development process and the design of the user interface. It will also examine the opportunities that the web application offers for the work of the MfN research data management team, such as increased processing speed for DOI requests and improved quality control in the data publication process.

1 Introduction

Automation is a main trend in all contexts in which data are a valuable asset (Hobart 2020). Sustained growth in the amount of available data, greater interest in extracting knowledge from data, and significant increase in data regulations and policies are constantly shifting the balance from human-based to automated data management solutions.

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.c18065

Research data management is also experiencing the automation trend. Automated solutions for the creation and constant update of data management plans and rule-based management of research data pipelines are just two examples of this (Chard et al. 2017; Miksa, Oblasser, and Rauber 2021). The potential for automation is not limited to researchers' daily work. At the data publication stage, there is still room for increasing automation.

Persistent identifiers (PIDs) in the form of DOIs (Digital Object Identifiers) are now well-established tools to promote and enhance citation, sharing, and reuse of research results (texts, datasets, media, etc.; Liu 2021). Once assigned a DOI, a research output has a permanent handle and the DOI will always resolve properly. In addition, descriptive metadata are associated with the DOI and provide relevant information about the object to which the DOI has been assigned. Registration agencies, such as Crossref and DataCite, provide the technical infrastructure for DOI services and all data publishers need to collaborate with one of these agencies to successfully register DOIs for their publications.

As DOIs are both machine and human-readable, they are a relevant element to make research data FAIR (Findable, Accessible, Interoperable, Reusable). Therefore, more and more often libraries and data management services of universities and research institutions register DOIs for the organisation research outputs. In many cases the process is still carried out manually using the web interfaces made available by DOI registration agencies, but there is the necessity to rethink the balance between people, processes, and technologies to cope with increasing amounts of DOIs requests.

The paper will discuss the web application that the research data management team at the Museum für Naturkunde Berlin (MfN) has been building to automate the DOI registration process with the provider DataCite. Running in a Docker container, a Django web application calls the DataCite API and interoperates with MfN institutional databases to prompt the DOI creation and to collect the metadata that are required for DOI registration. The paper will begin with a brief description of the data publication work at MfN and the current DOI workflow, continue with a description of the software architecture and user experience of the web application developed for automating the DOI registration process, and conclude with a reflection on the role of automation in current research data management.

2 MfN data publications

At MfN the research data management team is responsible for the pipelines of the organisation's data publications. The data publication process established by the institution follows the FAIR principles (Wilkinson et al. 2016). DOIs are registered to enhance data findability, accessibility, and reuse. MfN staff who request a DOI, for instance, can make available a dataset alongside a published paper without relying on paid services offered by academic publishers. In addition, researchers do not need to concern themselves with data availability in the institutional repository, as this is managed on their behalf by the research data management team. Once registered, the DOI will permanently identify

the dataset that can be referenced with a stable URL, and cited and shared with other researchers.

MfN registers DOIs with the global non-profit organisation DataCite, one of the main registration agencies for research data.[1] MfN benefits from DataCite's services via a cooperation agreement with the public organisation Deutsche Zentralbibliothek für Medizin (ZB MED). ZB MED is an official DataCite member and represents a consortium of German institutions engaged in life sciences research. ZB MED assigns prefixes to its consortium members and MfN has been assigned the prefix 10.7479. The MfN DOI `https://doi.org/10.7479/2j13-v254` identifies, for instance, a set of audio recordings of birds singing behaviour. The dataset is part of MfN Animal Sound Archive (Tierstimmenarchiv), which is one of the oldest and most extensive collections of animal sound recordings existing worldwide. It was started at MfN in the 1950s and now includes about one hundred and twenty thousand records.[2]

Using DataCite, DOIs can be created in three possible states: draft, registered, findable. DOIs in draft state are not yet publicly available and can still be deleted even though the DOI is reserved, DOI in registered state cannot be deleted anymore, but their metadata are only available to subscribers of DataCite, findable DOIs are officially registered and their metadata are publicly available. As evident in Figure 1, the number of DOIs registered by MfN has considerably increased over the years and it has more than doubled in 2022 compared to the total registrations in the previous year. This is not really surprising given the overall trend. If DataCite assigned just over one hundred and fifty thousand DOIs in 2011, this number has grown to over four millions in 2020 and the growth trend continues uninterrupted as the most recent statistics show.[3]

DOIs popularity is definitely good news for a FAIR management of research data, but poses practical challenges to the institutions that internally manage their data publications and register DOIs for them. The manual process of registering a DOI is time-consuming. Data managers need to collect the metadata associated to the dataset from internal systems and then transfer these metadata to the external service provider to register the DOI. In addition, creating a landing page to which the DOI redirects imposes further steps. The landing page needs to display properly all the information related to the dataset and stored in the descriptive metadata associated with the DOI. At the same time, the landing page must give access to all the files (e.g., csv files, txt files, media files, etc.) associated to the dataset. This requires to connect all the information related to the DOI also within institutional databases. Therefore, the DOI workflow typically involves multiple IT systems, such as institutional databases for data and metadata and the user interface offered by the DOI provider. Data managers need to be familiar with all these IT systems and switch from one to the other to keep the metadata updated and to complete the registration process.

1 `https://datacite.org`

2 `https://www.tierstimmenarchiv.de/webinterface`

3 See DataCite 2020 Annual Report (`https://datacite.org/wp-content/uploads/2023/06/DataCite-2020-Annual-Report.pdf`) and DataCite current statistics (`https://stats.datacite.org`).

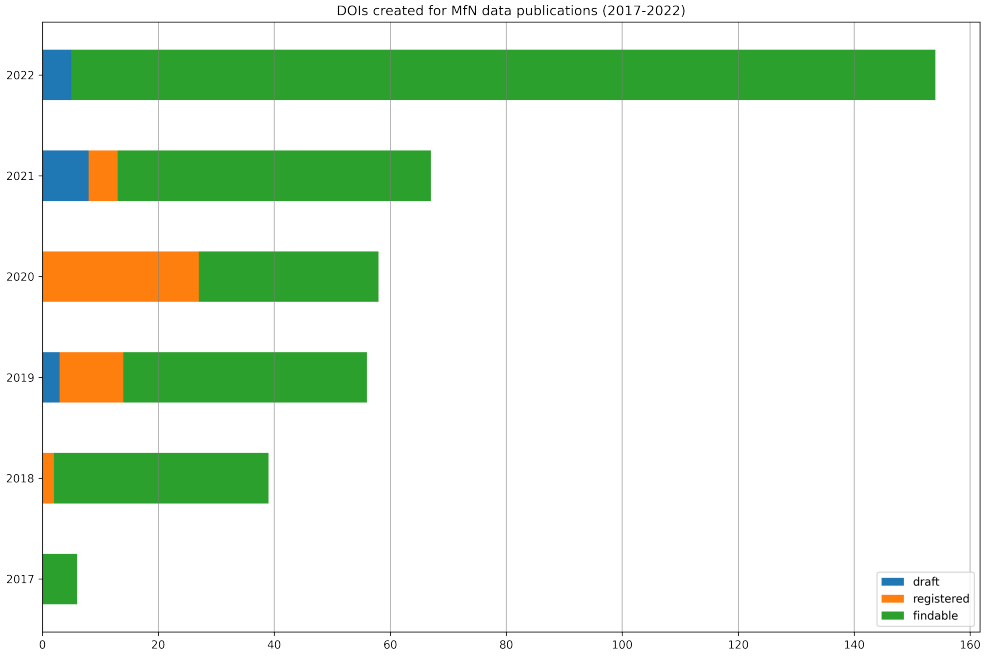


Figure 1: DOIs created for MfN data publications (2017-2022).

At MfN the DOI registration process begins with the creation of a DOI request by a member of the institution (Figure 2). The request is made by filling in a dedicated webform in the internal survey tool. The information collected includes metadata such as dataset author and abstract, license, necessity to set an embargo on the data, etc. The form is received by the data management team that checks manually the request and accepts it as it is or requires further information about the dataset for which the DOI is needed. If the request is accepted, the data manager creates a draft DOI using the web interface DataCite Fabrica.[4] The draft DOI is then added to the survey form originally submitted and the form data are transferred to MfN metadata storage, which is a MySQL database. The data transfer is also necessary to generate a landing page in the institutional repository where the dataset and metadata will be publicly available. To make the dataset available, the data files (e.g., csv files, media files, etc.) need to be uploaded to MfN digital asset management system, and each asset must be linked to the respective record of the MfN metadata database. Once this step is completed, the dataset will be properly displayed and available on the landing page created for the DOI. At this point, it will be possible to export the metadata in XML format from the landing page. These metadata will be used to add the descriptive metadata to the draft DOI record using DataCite Fabrica user interface. Once the metadata have been added, the status of the DOI can be changed to registered or findable by the data manager.

4 `https://doi.datacite.org`

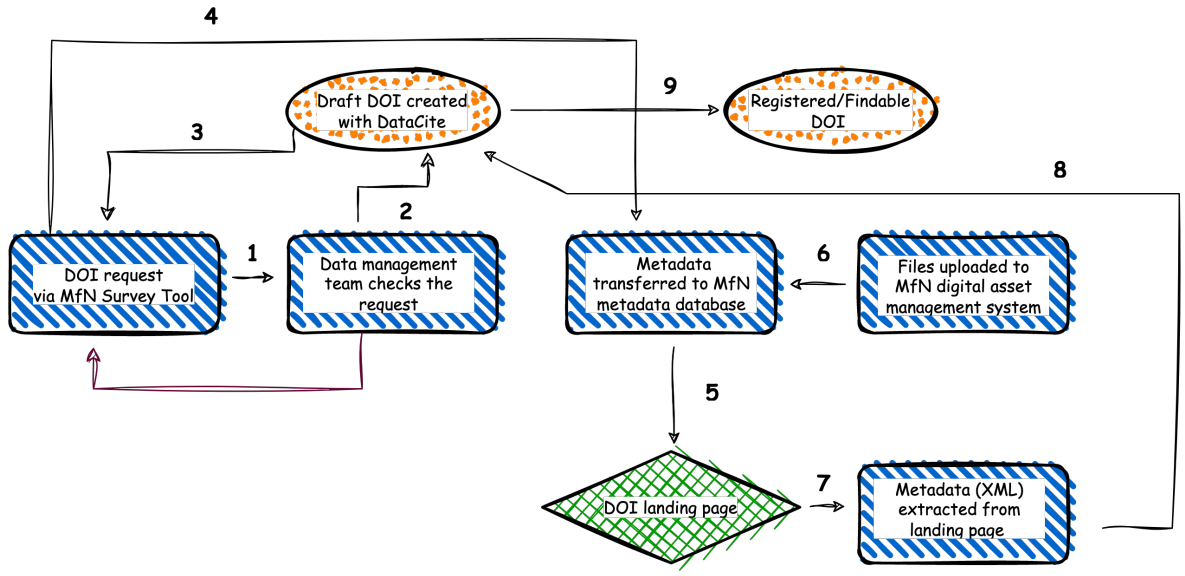


Figure 2: DOI workflow at MfN.

As evident from the description, the process of minting and registering a DOI with DataCite for MfN data publications has multiple steps and for the large part it is a human-centred process.[5] The significant increase in DOI requests registered in the past few years and the interest in eliminating possible human errors during DOI creation and metadata transfer impose to rethink the current procedure and encourage solutions that automate at least a few of the steps in the DOI creation process. The availability of a well-documented and reliable API for DataCite Fabrica suggested to start the automation process of the MfN DOI workflow from the interaction with the DataCite service. DOI creation and metadata registration and update are all tasks that can be easily carried out via the DataCite API rather than the DataCite user interface, as explained below in detail.

3 Automation opportunities offered by APIs

APIs (API is a shorthand for Application Programming Interface) are computer routines that allow communication between software components via requests and responses. Most APIs have methods that allow to perform basic operations on data. Standard operations are creating data (POST method), updating data (PUT method), retrieving data (GET method), and destroying data (DELETE method). Data are usually transferred to/from the API in JSON or XML format. APIs have a long history in computer science, but their popularity has greatly increased with the development of the World Wide Web – nowadays, most APIs are web APIs – because they allow easy interaction with remote resources using standardised protocols. Building and deploying APIs is now a key compo-

5 In the workflow described, only the metadata transfer to the metadata database (Step 4) and the creation of the landing page (Step 5) are triggered by human intervention, but otherwise automated.

nent of web development and a solution preferred by many organisations for transferring and receiving data.

DataCite API is a REST (REpresentational State Transfer) API (Masse 2011). Currently, there are two versions of the DataCite API, a Public API for non-authenticated users and a Member API for authenticated users.[6] Non-authenticated users can browse, query, and retrieve all findable DOIs and related metadata using the Public API. Authenticated users can also register DOIs, add and update DOI metadata, manage prefixes, etc. with the Member API. In short, DataCite Member API allows to automate all operations required to mint and register a DOI and to add and update the related metadata.[7] To facilitate members interested in working with the API rather than with the user interface, DataCite offers rich documentation and the opportunity to carry out development work with a test API. In the test API, all methods of the production API are available and users can experiment with all possible DOI statuses (draft, registered, findable) without interfering with the work of the production API and without cluttering the DOI register of the institution with dummy records. Both a test API and a test web interface are made available by DataCite for development purposes.[8] In addition, working with the DataCite API is facilitated by the availability of REST clients that provide all methods (e.g., POST, PUT, UPDATE, DELETE) to interact with the API. In the development work carried out at MfN, the Python API client wrapper for the DataCite API was used, as Python is the language of choice for many of the software applications currently developed by the institution.[9]

Relying on the DataCite API, at MfN it has been possible to automate all the interaction steps with the DOI registration service, starting with the creation of the DOI in draft state and proceeding with the transmission of the DOI metadata. The data manager can now request a draft DOI from within the institution's internal services and add the metadata available in the DOI request form to the draft DOI without needing to extract from the landing page the XML version of the metadata. As the data manager wanted to be able to check everything before making a DOI findable, all the DOIs minted with DataCite API are in draft status. Only when the data manager has completed the entire process of metadata transfer and file upload within MfN internal databases, and all the elements on the landing page are correctly displayed, the DOI status is manually modified in registered or findable, depending on whether there is an embargo on the dataset or not.

4 Software architecture and user experience

Two main criteria have been considered in designing the code for calling the DataCite API to mint and register DOIs automatically. The first one was the easy integration with the existing IT infrastructure at MfN and the microservice architecture that has guided its development (Bucchiarone et al. 2020). The second criterion was the necessity

6 The API url does not change. It is always `https://api.datacite.org`.
7 `https://support.datacite.org/docs/api`
8 `https://api.test.datacite.org` and `https://doi.test.datacite.org`
9 `https://datacite.readthedocs.io/en/latest/_modules/datacite/rest_client.html`

to create a user interface, as the data manager is not expected to have programming skills. Due to these requirements, it was decided to develop a Django web application running in a Docker container. The web framework Django was selected because it is Python-based, offers fast and secure development due to its several pre-coded features, and fulfils all the front-end (logging, messaging, etc.) and back-end (database connection, availability of web forms, etc.) requirements for the web application to interface with the DataCite API (Mele and Belderbos 2022).[10] In addition, it was easy to integrate with Docker, the technology of choice for running applications isolated in virtual software containers at MfN (Mouat 2015). Figure 3 is a schematic illustration of how the Django application works. The user interface allows log-in and log-out operations, access to the Django admin interface, and provides data managers with a webform to fill in to trigger the DOI registration process with DataCite. Via the front end, the user is notified of errors, warnings, and successfully completed DOI registrations. The back end of the web application manages all the interaction processes with DataCite and with the MfN databases essential for the DOI request and the functioning of the Django application.

Before triggering the DOI registration process, the back end queries the MfN Survey Database and checks that the submission number (an internal identifier for the DOI request form) is correct. If the submission number is not correct, the user is notified via the front-end interface. When a valid user request is received, it is the back end that collects the metadata for the DOI registration from the survey form with the given submission number, registers the required data (submission number, username, date and time of request creation) in the database associated to the Django admin interface, and calls the API to request the draft DOI and to put the metadata payload in JSON format. All the calls to the DataCite API are done using the available Python client mentioned above. The client greatly facilitates working with the API because it has all the required methods to transfer data to and from the DataCite service. In addition, the client enables to switch easily from the DataCite development API to the DataCite production API. This contributes to make the code ready for deployment quickly.

In the development of the Django application, a priority was to offer an effective and intuitive user interface to the MfN data manager. The application front end allows user authentication and, for the data management team, also enables access to the Django admin interface, where the submission numbers already processed can be checked. As the web application is designed only to register DOIs with DataCite, the web form that triggers the DOI registration process is immediately displayed to authenticated users. In the form, the data manager needs to provide only two pieces of information: their username and the submission number of the DOI request form in the MfN survey database (Figure 4). The submission number is the piece of information essential for the entire process. By using this number, it is possible to collect all the relevant metadata from the request form and start the process for minting the DOI with DataCite. As the submission number is the linking piece of information in the process, its existence in the MfN Survey Database is immediately checked, when the form is submitted. If the submission number

10 `https://www.djangoproject.com`

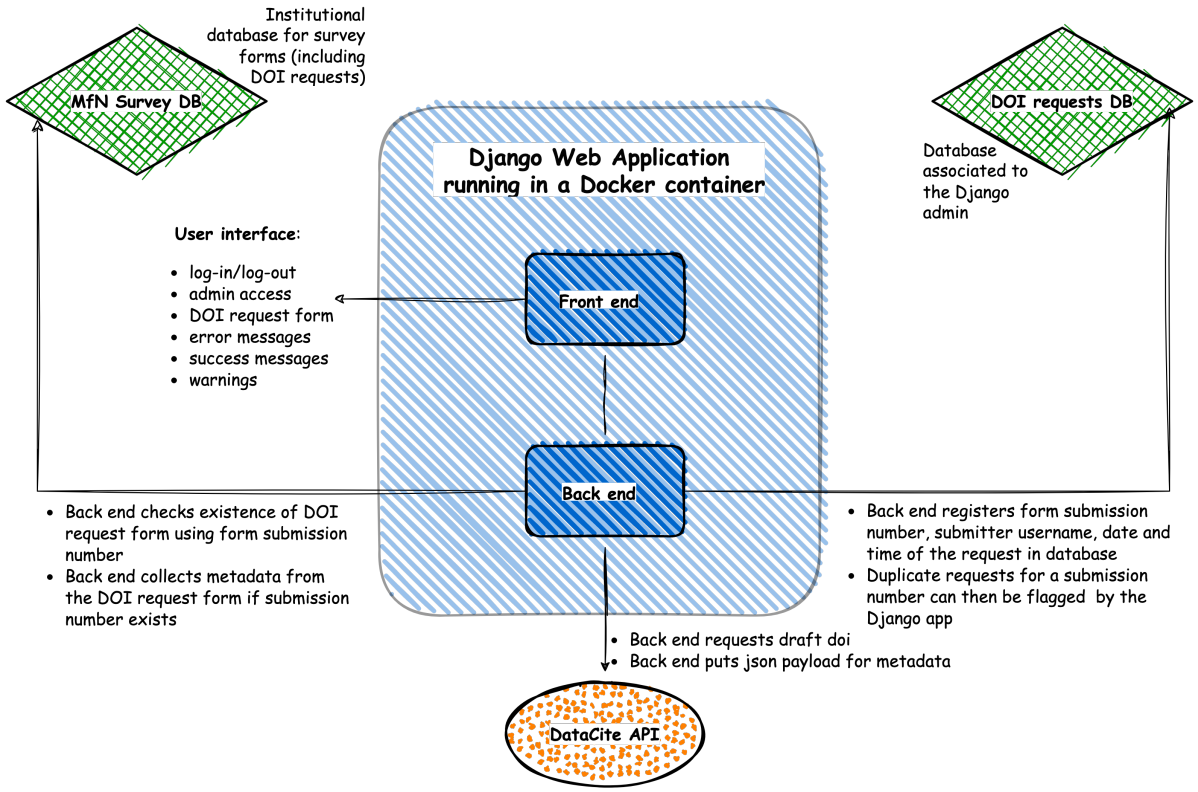


Figure 3: Code structure.

is not found among the DOI request forms in the database, an error message is generated and displayed to the user.

Once the user has typed a valid submission number, the application calls the DataCite API to generate a DOI and assign to the generated DOI the metadata extracted from the DOI request form available in the MfN Survey Database. Once the DOI has been successfully minted, a success message is displayed to the user (Figure 5 (left)). The success message contains information on the DOI created and the related submission number for which the DOI has been created. This allows the data manager full control of the process. In addition, using a button provided in the user interface, the data manager can easily copy the DOI with a click and paste it in the original survey form, the last passage before the form data can be transferred to MfN metadata database (see Figure 2). If the DOI has been generated for a submission number already present in the database associated to the Django admin interface, the user receives a warning message and can check for possible duplicates (Figure 5 (right)). On the success page, a link is available to return to the submission form, so that the data manager can restart the process immediately, if more DOIs need to be minted at the same time. By using the Django application developed, the DOI process can be carried out by the MfN data manager without using DataCite Fabrica user interface. This reduces at least a few manual steps in the DOI workflow described in Figure 2. As a result, there is not only some time-saving, but also better quality control

because the metadata are now associated to the DOI at the time of creation eliminating any risk for mismatching of metadata and DOI.

All the technologies used in the development of the web application (Docker, Django, etc.) are free to use for educational and non-commercial organisations. The DOI request process implemented by the application uses the available DataCite python client and it is portable and reusable by any organisation that needs to register/update a DOI via DataCite API. However, the way metadata are read from the survey database is specific to MfN and cannot be generalised to other institutions. Nonetheless, the tools used to read the metadata (sqlalchemy, mySQL connectors, etc.) are widely popular to integrate database queries in python scripts and can be recommended as a handy solution to work with DataCite client, regardless of the database type and configuration. The documentation for the sqlalchemy library (`https://www.sqlalchemy.org`) provides all the required information for connection to popular databases, even databases not explicitly considered here, such as PostgreSQL.

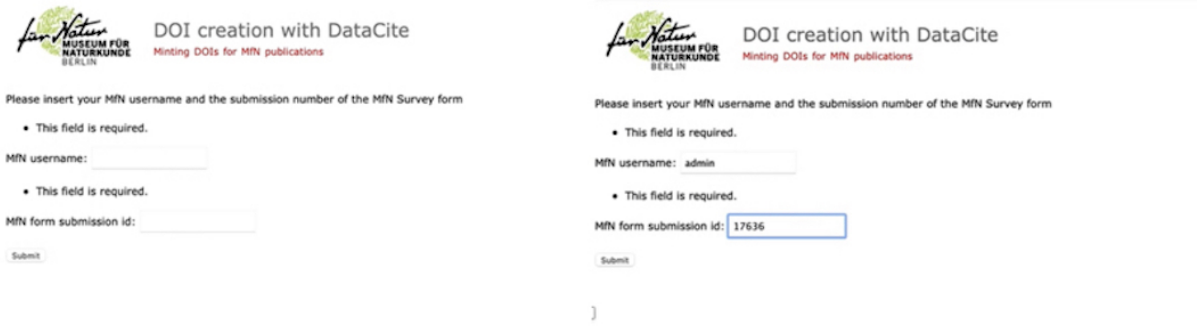


Figure 4: User interface. From left to right: 1) Django form for requesting a DOI; 2) Django form filled in. The images were taken during code development.

5 Research data management and automation

More demanding funder policies regarding research data and growing needs on the side of researchers, who are dealing with scientific data in quantities never experienced before, are constantly increasing the workload of research data management teams. Surveys conducted among library staff in several countries suggest that, alongside traditional advisory roles, such as preparation of data management plans or consultations on copyright and intellectual property, research data management teams are devoting more and more of their time to run data repositories, compile data catalogues, create metadata, rescue data, invest in long-term data preservation and data quality checks (Cox et al. 2019).

Usually, the workload increase has not been matched by a corresponding increase in staff and financial resources for research data management teams. Automation becomes, therefore, a necessary choice to avoid limiting or reducing the services that research data management teams provide to users. For services that are constantly requested, as it

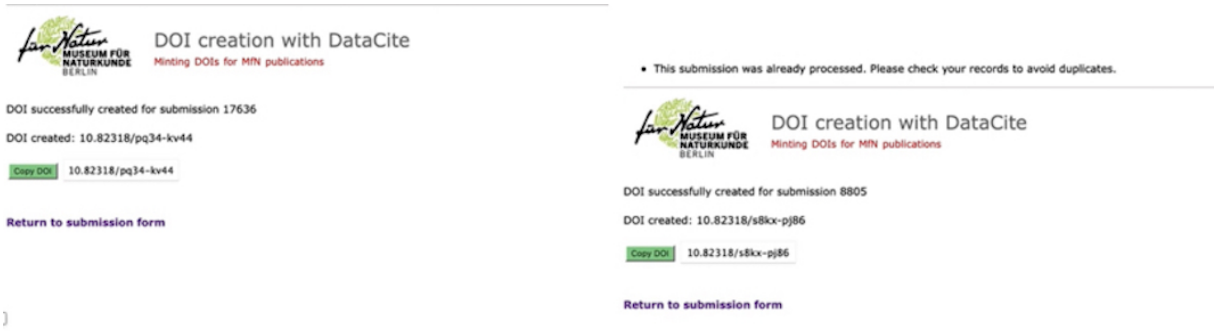


Figure 5: User interface. From left to right: 1) Success message for DOI creation related to the submission in Figure 4 (right); 2) Success message for DOI creation related to a submission number already present in the Django admin database. The user is warned about a possible duplicate. The images were taken during code development. The DOI prefix refers to the test repository made available to MfN by DataCite.

is the case of DOI registration at MfN, the initial time invested in developing code for automation is easily paid back by the time saved during daily operations. In addition, code development can be greatly facilitated by taking advantage of software already available and by following development strategies that are easy to implement and highly portable. This has been the philosophy behind the code development for automating the DOI process at MfN. Calling the DataCite API has been straightforward thanks to the already available API Python client. Furthermore, the microservices approach followed by the institution has made easier the integration of the new web application in the already existing IT ecosystem at MfN. The use of the web framework Django for creating the application has further accelerated the coding work. Each Django project already contains a template for the back end and the front end and this template can be customised and reused without the necessity to write all the code from scratch.

The arguments in favour of automation, however, are not confined to the limited resources of research data management teams and their increased workload. The concept of automation is interlinked with the concept of quality and this not just in manufacturing, which is the industrial sector in which modern automation procedures have their roots (Nof 2009). Key features of quality control, such as efficiency, productivity, and reliability, are all at stake in automation and this holds true also for automation in research data management. By shifting tasks from people to technologies in DOI registration we have set in motion a process that, on the one hand, allows data managers to be more efficient and productive in carrying out this task because they are able to achieve the goal with fewer steps, and, on the other hand, we have removed potential sources of mistakes that exist in manual operations, such as associating the wrong metadata to a DOI. The time that the data manager saves because there is no need to log in and out of multiple IT systems and to perform manual operations (like copying and pasting data from one system to the other) can be devoted to a much better use. At MfN the initial request of the user is the key source of metadata information for the DOI. To ensure the highest

quality in a FAIR publication process, it is, therefore, crucial that the data manager has sufficient time to review the request form in detail and, if necessary, interact with the user pointing out mistakes or inconsistencies in the metadata provided.

6 Conclusions

We have described an example of process automation in research data management implemented to meet the significant increase in DOI registration requests received at MfN. This is just an example of the benefits that task automation allows in research data management. Over time, we expect that there will be more interventions of this kind for shifting recurrent tasks, which do not require human control, from people to technologies.

We conceive this automation process as going through incremental steps, as suggested by an Agile approach to software development (Shore and Warden 2021). The web application described here is not a final product, but the first step of a more ambitious plan. Once the current version has satisfactorily performed in production, we plan to automate further. Currently, the data manager needs to manually copy the DOI minted into the request form in the MfN Survey Database (Figure 2. Step 3) and trigger the process of metadata transfer to the MfN metadata storage by filling in a web form with the submission number (Figure 2. Step 4). These two steps can be further automated by giving writing rights to the databases to the Django application that carries out the DOI registration process. In a further refinement, we also envision to automatically update the metadata registered in DataCite if there are changes in the MfN metadata database for the information related to a DOI. Further developments on the table also include the possibility to mint a DOI with a specific suffix and not just to mint a random DOI. This is a convenient feature when the data manager needs to create DOIs for a collection of related items or when a DOI must be assigned to versions of the same item that only differ in language.

Overall, we feel that there is a compelling argument in terms of time economy and improved quality control for automating tasks in research data management. At all stages of the data management process, there is both an increase in the quantity of research data available and in the tasks that need to be performed on these data to comply with funders' regulation, legal requirements, best practices to ensure FAIRness, etc. It is hardly realistic to think that a manual approach can be a solution to this transformation, given also that this trend is not going to stop, but rather it will continue with possibly an even more sustained growth in coming years. Only technologies can make a difference and free up valuable time for carrying out operations that really require human intervention in research data management, such as advisory tasks.

Acknowledgements

The authors thank the participants in the E-Science-Tage conference for the comments and suggestions received in Heidelberg. They are also grateful to the MfN colleagues, Mareike

Petersen and Caitlin Thorn, for the feedback received while developing the application user interface and user experience. The code development and the participation of one of the authors in the E-Science-Tage conference were kindly sponsored by MfN with internal funding.

References

Bucchiarone, Antonio, Nicola Dragoni, Schahram Dustdar, Patricia Lago, Manuel Mazzara, Victor Rivera, and Andrey Sadovykh, editors. 2020. *Microservices.* Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-31646-4.

Chard, Ryan, Kyle Chard, Jason Alt, Dilworth Y. Parkinson, Steve Tuecke, and Ian Foster. 2017. "Ripple: Home Automation for Research Data Management". In *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW),* 389–394. IEEE. DOI: https://doi.org/10.1109/ICDCSW.2017.30.

Cox, Andrew M., Mary Anne Kennan, Liz Lyon, Stephen Pinfield, and Laura Sbaffi. 2019. "Maturing research data services and the transformation of academic libraries". *Journal of Documentation* 75 (6): 1432–1462. DOI: https://doi.org/10.1108/JD-12-2018-0211.

Hobart, Mark. 2020. "The Business Case for Automating Data Management". *Journal of ICT Standardization* 8 (3): 199–216. DOI: https://doi.org/10.13052/jicts2245-800X.832.

Liu, Jia. 2021. "Digital Object Identifier (DOI) and DOI Services: An Overview". *Libri* 71 (4): 349–360. DOI: https://doi.org/10.1515/libri-2020-0018.

Masse, Mark. 2011. *REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces.* O'Reilly Media, Incorporated. ISBN: 9781449319908.

Mele, Antonio, and Bob Belderbos. 2022. *Django 4 by Example: Build Powerful and Reliable Python Web Applications from Scratch: Build Powerful and Reliable Python Web Applications from Scratch.* 4th edition. Packt Publishing, Limited. ISBN: 9781801813051.

Miksa, Tomasz, Simon Oblasser, and Andreas Rauber. 2021. "Automating Research Data Management Using Machine-Actionable Data Management Plans". *ACM Transactions on Management Information Systems* 13 (2): 1–22. DOI: https://doi.org/10.1145/3490396.

Mouat, Adrian. 2015. *Using Docker Developing and Deploying Software with Containers: Developing and Deploying Software with Containers.* O'Reilly Media, Incorporated. ISBN: 9781491915929.

Nof, Shimon Y., editor. 2009. *Springer Handbook of Automation.* Springer Handbooks. Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-540-78831-7.

Shore, James, and Shane Warden. 2021. *Art of Agile Development.* 2nd edition. 530. O'Reilly Media, Incorporated. ISBN: 9781492080695.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Research Data Policies in Scientific Journals – a Case Study

Gertraud Novotny, Thomas Seyffertitz

University Library, Vienna University of Economics and Business

In recent years, scientific journal publishers as well as scholarly associations - have begun to develop framework guidelines for handling the research data underlying a publication. Such data guidelines and recommendations have become increasingly important for researchers, as they have to be taken into account as part of the research or publication process. At the same time, publishing in renowned journals is an essential part of an academic career. For this reason, the Vienna University of Economics and Business (WU) has been awarding prizes[1]. for publications by WU researchers in one of the journals listed in the so-called "WU Star Journal List" for some time. The selection of the listed journals is based on international rankings of top scientific journal in different and covers research fields that are of high importance to the WU. In the following we provide a qualitative analysis of data policies along some a priori defined categories.

1 Introduction

The objective of our work is a formal and content-related analysis of the data policies selected from the "WU Star Journal List"[2]. We distinguish between two levels, that of the publishing company or research organization and that of the specific journal. The policies are to be analyzed according to various criteria and categorized in terms of their strength or assertiveness (i.e., to what extent do the data guidelines have recommendatory or mandatory elements). Formal criteria include attributes related to presentation, editing, timing and clarity. The content attributes deal with scope, minimum requirements, comprehensibility, and strength of the policies. This will include identifying whether there are similarities in data policies for different journals of the same disciplines. Subsequently, the task is to prepare the insights gained from the analysis for consulting services on the topic of data policies and thus to support researchers.

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.c18066

1 For details see `https://www.wu.ac.at/en/research/research-units-at-wu/internal-awards/`; WU's performance bonuses for academic staff are regulated by an Operational Agreement on Performance Bonuses and Examination Rates for Academic Staff.

2 `https://www.wu.ac.at/fileadmin/wu/h/research/wu_starjournalliste.pdf`

Table 1: Data policy framework spectrum of four large journal publishers (Seyffertitz 2023). Note: DAS = Data Availability Statement/Data Access Statement; PID = Permanent Identifier (e.g., DOI).

Publisher	**Minimum of information required on data sets as stated by publisher/journal** (all features **encouraged**)		(all features **required**)
John Wiley & Sons Inc. (2022)	- DAS	⟷	- DAS - Peer review of data
Taylor & Francis (2018)	- DAS	⟷	- DAS - PID for data - Data citation
Elsevier (2022)	- Data deposit in a relevant data repository - Citing this dataset in the article	⟷	- Data deposit - Data citation and linking (or a DAS) - Peer review of data prior to publication
Springer Nature (2023)	- Data sharing - Data citation	⟷	- Data sharing - Evidence of data sharing - Peer review of data

Not far too long-ago scholarly journal publishers have started developing (data) policies around the sharing or publication of research data underlying the manuscripts they are publishing. Such guidelines sometimes are termed "data policy" or "research data policy"[3] or similar. Some publishers or scholarly associations refer authors to different data repositories in their policies or guidelines or recommend searching a directory of data repositories to find a suitable data archive for the relevant research data[4]. As this may be important to the authors submitting their publication, it is worth looking at some journal data policies here. For example, Springer Nature developed a framework for the research data policies of all its journals (Hrynaszkiewicz et al. 2017). The Data policy standardisation and implementation Interest Group (IG) of the Research Data Alliance further developed this framework around existing scholarly publishers' research data policies of Springer Nature, Elsevier, Wiley, and PLOS (Hrynaszkiewicz et al. 2020). Table. 1 provides a summary of research data guidelines (or policies) of some scholarly publishing companies in a condensed manner.

There are differences in the extent to which the policies cover aspects of data management in their frameworks. In addition, terminology and wording vary from one to the other publisher in some way, but overall, the framework guidelines share some common core features. For an overview on general data policy features see Hrynaszkiewicz et al. (2020). The design and content may vary depending on the journal's research topics covered and

3 We will use these terms synonymously in the text.

4 See for example, `https://authorservices.taylorandfrancis.com/data-sharing/share-your-data/repositories`

Table 2: Overview of selected studies of journal data policies.

Study authors	**Research fields covered by the journals investigated (sample size)**	**Main objective/interest**
Andreoli-Versbach and Mueller-Langer (2014)	economics (147)	data sharing policies
Vlaeminck and Herrmann (2015)	economics (346)	characteristics of journals with data policies
O'Reilly and Herndon (2015)	social-sciences (100)	data sharing policies comparison (2003/2015)
Crosas et al. (2018)	social-sciences (291)	data policies and instructions
Rousi and Laakso (2020)	operation research (40)	data sharing policies (2019)
Vlaeminck (2021)	economics (353)	disclosure of research data, reproducibility

research tradition in the scientific community concerned. In developing the strength of a data policy, the characteristics can be phrased as (i) "it is recommended" or (ii) "it is expected" to (iii) "is required" or "must". It may also depend on the journal editorial board what type of policy is implemented at the journal level. For example, the "Data Availability Statement" (DAS) is one of the main features often provided as minimum mandatory instrument. It is a statement about where and how the underlying data in a published article can be found.

In contrast to other works, we do not review a complete set of journals of a certain discipline or subdiscipline, rather we have a given list of journals from different research fields represented at WU, except for law[5], which is not included due to its different publication tradition. Table. 2 summarizes some important studies of journal data policies in the social sciences, especially in economics and cover usually top-ranked journals in the field (according to predefined indicators, as for example citation metrics).

2 Research design

Methodologically, we refer to the deductive-qualitative content analysis (see Schneijderberg et al. 2022). Hence, we use a priori defined thematic categories (see Table. 3). In the next step the content of the discovered data guidelines is analyzed along these categories. In a search, we first tried to discover the guidelines or policies of the journals in our sample. For this search process we first looked at the website of the journal's publisher. As some journals may have more than one website (i.e., example one website is operated by the journal owner like for example, a scholarly society or a university, and the other

5 This covers two research units at WU: the Department of Private Law and the Department of Public Law and Tax Law.

Table 3: Four thematic data policy elements (categories) and what the usually cover.

Thematic categories	**Short description of the category content**
Data citation	What does the policy say about the provision of the citation of underlying data in the submitted paper (i.e., is a DOI required referring to the data)?
Reproducibility	Statement about reproducibility, its requirements, recommendations, and how-to descriptions
Data sharing	Statements about data sharing in general, mandatory or only recommending
Data availability statement	Is a data availability statement (DAS) required or not and what information must/should be provided in the DAS?

one is maintained by the publisher)[6], both web pages have been included in the analysis (if applicable).

2.1 Defining the thematic categories of analysis

For the present analysis, we are interested in whether publishers or editors make statements about these categories as part of their policies (i.e., constitute an element of the data policy), and if so, how can the statements be characterized or interpreted: Are they recommendatory or mandatory in nature. This will allow us to classify a journal's data policy as rather weak or strong.

We classify data policies or data guidelines weak if (i) statements for the a priori defined thematic categories are either missing or (ii) if these statements are mostly of a recommending character while not going into much detail (e.g., providing information on repository selection, or practical information on what ingredients a data availability statement should consist of). We regard policies as strong if for these categories, statements are provided and the majority of the categories provide mandatory statements.

2.2 Defining formal categories of analysis

While the thematic categories are important for characterizing the content of the data policies, the formal categories are important to characterize the ease-of-access of the data policy. Therefore, we defined some formal categories. These categories comprise format,

6 See for example the journal International Economic Review, which is owned by the Economics Department of the University of Pennsylvania and the Osaka University Institute of Social and Economic Research Association (`https://economics.sas.upenn.edu/ier`) whilst being operated and published by John Wiley & Sons Inc. (`https://onlinelibrary.wiley.com/journal/14682354`). Sometimes the reader may find some statement like "Published on behalf of the University of …" or similar.

Table 4: Formal characteristics of data policies.

Formal categories	**Description of the category**
Location and clarity	Where are the guidelines/policies located: • website of the publisher • website of the professional association/learned society or university. • are the data related guidelines easily findable?
Naming Representation	How are the data policies termed; are there standards? • explicit: separate data policy document • included: in author- or submission guidelines or other policy documents (e.g., ethical guidelines)
Format and preparation	In which format are the guidelines presented: • plain HTML, • pdf, Word etc. • embedded media files (e.g., video)
Timing	When do data guidelines become effective for the author – at the time of: • submission • acceptance (conditional) • publication

presentation, location, naming, timing of guidelines. In Table 4 we describe them in more detail.

2.3 Discovering the data policies

Data policies or guidelines sometimes are part of the guide for authors or submission guidelines. Some journals also have a specific section on their website in which they provide requirements regarding reproducibility and data disclosure. In some cases, these instructions were part of publishing ethics guidelines of the journal website (publishing company) or could be found at the website of the scholarly association.

3 Analysis and results

At the time of analysis[7], we identified three (two of the same association) out of the 34 journals for which we could not identify some kind of journal data policy or information on research data management. The rest covers a spectrum from rather very weak data policies to strong data policies. Interestingly, from eight journals (almost 25% of the investigated sample) that have a strong data policy according to our above definition, seven of them refer to economics, and one to the sciences. Concerning the formal characteristics of the data guidelines we found 18 journals providing their data guidelines in a separate document (i.e., being a single html document or a pdf-file) apart from the traditional submission or author guidelines. The naming of these explicit data guidelines is rather heterogeneous throughout the different journals. We could identify different terms appearing in the document title, as for example, "*Data and Code Sharing Policy*", "*Reporting standards*" or "*Data and Code Availability Policy*" just to mention a few. The remaining 13 journals included their data guidelines in the submission or author guidelines where data guidelines were highlighted using chapter headings.

Overall, the majority of journals analyzed set some minimum requirements in their policies. The thematic categories show a predominantly recommendatory character. e.g., stating that "*authors should provide data upon request*". A mandatory element that we found in some journals was the requirement for authors to provide the underlying data and code to the editor for reproducibility purposes. This does not mean that the data must be made openly accessible. Concerning comprehensibility of data guidelines, some (usually the stronger ones) are concise in their statements and provide clear information what is expected or *what* must be provided (e.g., data, DAS, DOI or other PID). Further, more developed and strict policies provide information under what *conditions* (type of paper: simulation, experiment, etc.) as well as at which *point in time* (submission, acceptance, publication) the policy or one of its elements becomes effective.

For some journals, we found rather new features, which will be explicitly shown in the following. First, three of the 34 journals allow authors for the optional provision of so-called pseudo-data. Especially in finance, the data used often are licensed from data providers like Datastream[8] or Bloomberg[9], proprietary in nature, and hence cannot be published or shared. Other journals do have a so-called data editor that we will explain briefly.

3.1 Special feature: use of pseudo data sets

Three journals – all related to the subject of finance according to the SSCI-index provide the possibility of using so called pseudo-data sets:

7 The content of the data policies/author guidelines etc. analyzed in this study, has been accessed and collected from the journals' websites in the period of February 9 to February 20, 2023.

8 Owned by REFINITIV (`https://www.refinitiv.com/en`; part of the London Stock Exchange Group Business).

9 `https://www.bloomberg.com/professional/product/data`

- **Journal of Finance**[10]: "… *Authors are also encouraged to include the data along with the source code if public posting of the data does not violate copyright or confidentiality agreements. If the authors choose not to provide the data,* ***they must include a pseudo-data set*** *that* ***illustrates the format of the files read by the code so that users can understand and check the functionality of the code****.* …" In their related FAQs[11] they provide more information on this: "… *not required that the code produces meaningful results based on the pseudo-data set. But pseudo-data should illustrate the format of the data that are read by the programs (e.g., dimension of the data, numbers or strings, etc.) to help a user understand the code. The* ***pseudo-data should not include any data that is protected by copyright or confidentiality agreements.*** *…*".

- **Journal of Financial Economics**[12]: "… *For those cases in which the* ***data cannot be disclosed****, the authors* ***must supply pseudodataset(s)*** *to demonstrate that the code runs. It is the authors' responsibility to ensure that the code works on the pseudo-dataset or the actual dataset if the data can be disclosed. …*".

- **Review of Financial Studies**[13]: "… *We encourage authors to include the data along with the source code. However, if the authors choose not to provide the data, or if they are restricted from doing so because of copyright or confidentiality agreements,* ***they are required to include a pseudo-data set*** *to illustrate the format of the files read by the code so that users can better understand the code. …*".

3.2 Special feature: the role of so-called data editors

Five journals analyzed provide a so-called data editor within their editorial board. Their roles comprise inter alia securing the data policy compliance, advising authors in several data related aspects and sometimes may verify for reproducibility of provided data. Below we provide some of their duties and responsibilities, extracted from the policy guidelines. For details we provide the links to the relevant documents at the journal websites (see footnotes):

- **Management Science**[14]: "… *The task of the Management Science Data Editor is to help authors to get their published research compliant with the Management Science Data Policy. … advises authors of published papers on the data, materials, and information to be provided to allow other researchers to replicate the original results. …*"

10 See `https://afajof.org/wp-content/uploads/files/policies-and-guidelines/CodePolicy.pdf`.
11 See `https://afajof.org/wp-content/uploads/files/policies-and-guidelines/CodePolicy_FAQ.pdf`.
12 See `https://www.jfinec.com/data-and-code-sharing-policy` and `https://www.jfinec.com/data-and-code`.
13 See `http://rfssfs.org/code-sharing-policy`.
14 `https://pubsonline.informs.org/page/mnsc/editorial-statement`

- **MIS Quarterly**[15]: "… *The new role of Transparency Editors will be similar in spirit to those at other journals, such as the American Economic Review's "Data Editor," the Journal of Personality and Social Psychology's "Methods and Statistics Associate Editors," and the INFORMS Journal of Data Science's "Reproducibility Editor. … Data Editor and her/his team will also occasionally verify that all results reported in an accepted paper can indeed be reproduced using the provided data, materials, and information. …*"
- **American Economic Review and Journal of Economic Literature**[16]: "… *AEA data editor will assess compliance with this policy, and will verify the accuracy of the information prior to acceptance by the Editor. … private (not to be published) version of the data should be provided to the AEA Data Editor and/or a designated third-party replicator who can provide a third-party reproducibility report. … After the data and code deposit is accepted by the AEA Data Editor, it will become the version of record associated with the paper. … Data Editor will assess suitability of any such repositories and archives. …*"
- **Review of Economic Studies**[17]: "… *The Review of Economic Studies endorses DCAS, the Data and Code Availability Standard [v1.0], and its data and code availability policy is compatible with DCAS. … Requests for exemptions should be clearly stated when the article is first submitted. The article will then be reviewed at the discretion of the Managing Editors and the Data Editor. Exceptions will not be considered later …*."

4 Conclusions and outlook

The aim of this case study was to analyze journal data policies of 34 scholarly journals from the fields of economics, business, management, finance, and science. For this reason, formal and thematic categories were defined along which the data guidelines were characterized. Results presented showed that almost all journals provided some form of data guidelines. Concerning the thematic categories about three quarters of the journals exhibited a more recommending character in their data policy. Three journals did not have any kind of data guidelines at the time of analysis.

A further important outcome is that data guidelines may become effective at different times: some journals require authors to deposit their underlying research data at the time of submission, other at the time of conditional acceptance of the submitted paper. This has an important implication for authors: they should become familiar with the data guidelines of the journal in which they intend to publish at an early stage.

Journals investigated in this study increasingly require authors to provide data and code in an appropriate form for reproducibility purposes. This implies that the authors are

15 `https://misq.umn.edu/research-transparency` → Following the research transparency approach as explained in the editorial by Burton-Jones et al. (2021).
16 `https://www.aeaweb.org/journals/data/data-code-policy`
17 `https://restud.dataeditor.group/before`

allowed to provide the data to the editors of the journal or the reviewers. This might become problematic when using licensed (and copyright protected) data for the research[18]. Hence a few journals have provided the option of using so called pseudo-data sets. Others have appointed a so-called data editor in their editorial board to monitor compliance with the relevant data guidelines.

So, authors are well-advised to look for (explicit) data policy documents or related guidelines, like for example publication ethics documents. However, sometimes such policy documents cannot be found easily at the publishers' website. Then author guidelines or submission guidelines may be a first starting point. Another option would be to contact a journals' editorial office. As indicated by our results, it is not uncommon to find elements of a "research data policy" somewhere in the submission or author guidelines, without any referral from outside the document. Overall, our results indicate that publishing data driven research in top scholarly journals increasingly requires authors to address data and code as early as the planning stage of a publication.

To raise researchers' awareness on the relevance of journal data policies for their publishing process, we will continue to extend our service portfolio in two ways: first, we will continue to collect and analyze the data policies of the journals in the WU-Star-Journal list. This list will be revised and extended in 2023 and will cover more than 70 international top journals covering the WU's research topics. Second, it is planned to design a workshop introducing important issues of data guidelines in scholarly journals, providing practical examples, tips and potential pitfalls. One of the main lessons learned within the work presented – together with the experiences from our previous consultations – is the necessity of already looking at the guidelines of the respective journal before submission. Preferably authors should take care of this already when writing the article – and if necessary, contacting a data editor (when available) already before the planned submission. This might be the case, for example, if the data are licensed and publication requires the provision of the data or in the case of sensitive data etc. Based on our findings and previous studies, we would therefore expect journal data policies becoming stricter, probably more detailed, and maybe more complex in future as the trend towards more open science seems to continue and the amount of data still increasing tremendously. Furthermore, methods in data analysis are improving, and development in the field of artificial intelligence will have a major impact on the research and publication process.

ORCID:

- Gertraud Novotny https://orcid.org/0000-0002-8816-4936
- Thomas Seyffertitz https://orcid.org/0000-0002-7444-6780

18 The problem may also arise if data are confidential (e.g., non-disclosure agreement has been signed).

References

Andreoli-Versbach, Patrick, and Frank Mueller-Langer. 2014. "Open access to data: An ideal professed but not practised". *Research Policy* 43 (9): 1621–1633. DOI: https://doi.org/10.1016/j.respol.2014.04.008.

Burton-Jones, Andrew, Wai Fong Boh, Eivor Oborn, and Balaji Padmanabhan. 2021. "Editor's Comments: Advancing Research Transparency at MIS Quarterly: A Pluralistic Approach". *MIS Quarterly* 45 (2): iii–xviii. ISSN: 0276-7783, visited on August 17, 2023. https://misq.umn.edu/misq/downloads/download/editorial/746/.

Crosas, Mercè, Julian Gautier, Sebastian Karcher, Dessi Kirilova, Gerard Otalora, and Abigail Schwartz. 2018. "Data policies of highly-ranked social science journals". DOI: https://doi.org/10.31235/osf.io/9h7ay.

Elsevier. 2022. "Research Data Guidelines". Visited on September 1, 2023. https://www.elsevier.com/authors/tools-and-resources/research-data/data-guidelines.

Hrynaszkiewicz, Iain, Aliaksandr Birukou, Mathias Astell, Sowmya Swaminathan, Amye Kenall, and Varsha Khodiyar. 2017. "Standardising and Harmonising Research Data Policy in Scholary Publishing". *International Journal of Digital Curation* 12 (1): 65–71. DOI: https://doi.org/10.2218/ijdc.v12i1.531.

Hrynaszkiewicz, Iain, Natasha Simons, Azhar Hussain, Rebecca Grant, and Simon Goudie. 2020. "Developing a Research Data Policy Framework for All Journals and Publishers". *Data Science Journal* 19 (1): 5. DOI: https://doi.org/10.5334/dsj-2020-005.

John Wiley & Sons Inc. 2022. "Wiley's Data Sharing Policies". Visited on September 26, 2022. https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html.

O'Reilly, Robert, and Joel Herndon. 2015. *Data Sharing Policies in Social Sciences Academic Journals: Evolving Expectations of Data Sharing as a Form of Scholarly Communication.* Version V2. DOI: https://doi.org/10.15139/S3/12157.

Rousi, Antti M., and Mikael Laakso. 2020. "Journal research data sharing policies: a study of highly-cited journals in neuroscience, physics, and operations research". *Scientometrics* 124 (1): 131–152. DOI: https://doi.org/10.1007/s11192-020-03467-9.

Seyffertitz, Thomas. 2023. "Research data repositories and what to consider when choosing one for deposit". DOI: https://doi.org/10.5281/zenodo.7716474.

Springer Nature. 2023. "Research data policies". Visited on September 6, 2023. https://www.springernature.com/gp/authors/research-data-policy/research-data-policy-types.

Taylor & Francis. 2018. "Data sharing policies". Visited on September 1, 2023. https://authorservices.taylorandfrancis.com/wp-content/uploads/2019/04/Author-Services-Data-sharing-policies.pdf.

Vlaeminck, Sven. 2021. "Dawning of a new age? Economics journals' data policies on the test bench". *LIBER Quarterly: The Journal of the Association of European Research Libraries* 31 (1): 1–29. DOI: https://doi.org/10.53377/lq.10940.

Vlaeminck, Sven, and Lisa-Kristin Herrmann. 2015. "Data Policies and Data Archives: A New Paradigm for Academic Publishing in Economic Sciences?" In *Proceedings of the 19th International Conference on Electronic Publishing,* edited by Birgit Schmidt and Milena Dobreva, 145–155. Amsterdam: IOC Press. DOI: https://doi.org/10.3233/978-1-61499-562-3-145.

Appendix: List of 34 scholarly journal titles

Date of analysis: data policies/author guidelines etc. have been accessed and collected from the journals' websites from **February 9 to February 20, 2023**. The table contains the 34 journal titles from the WU-Star-Journal-List (2016)[19]. The table columns are as follows:

- Journal title: the full title of the journal
- Homepage Journal: Link to the website of the journal (usually at the hosting publisher or scholarly society)
- SSCI[20]: Classifier from the Web of Science Master Journal List (https://mjl.clarivate.com)
- Link to the Journal data policy: either at the publisher's website or at the website of the respective editing scholarly association or university.
- Top factor[21]: only overall score if available; not all journals are covered in the TOP Factor (https://topfactor.org)

19 https://www.wu.ac.at/fileadmin/wu/h/research/wu_starjournalliste.pdf; *Last accessed on May 10th, 2023.*

20 Social Science Citation Index.

21 Transparency and Openness Promotion: The TOP Guidelines were created by journals, funders, and societies to align scientific ideals with practices and are provided by the Center of Open Science (https://www.cos.io/initiatives/top-guidelines).

Table 5: List of 34 scholarly journal titles.

Journal Title	**Homepage Journal**	**SSCI (WoS)**	**Publisher/ Association**	**Link to Journal Data Policy**	**Top Factor**
Academy of Management Journal	https://aom.org/research/journals/journal	Business | Management	Academy of Management	-	0
Academy of Management Review	https://journals.aom.org/journal/amr	Management | Business	Academy of Management	-	n/a
Accounting Review	https://aaahq.org/Research/Journals/The-Accounting-Review	Business, Finance	American Accounting Association	http://aaahq.org/portals/0/documents/about/policies&proceduresmanual/aaa%20publications%20ethics%20policy%20-%20data%20integrity.pdf	n/a
Accounting, Organizations and Society	https://www.sciencedirect.com/journal/accounting-organizations-and-society	Business, Finance	Elsevier	https://www.elsevier.com/journals/accounting-organizations-and-society/0361-3682/guide-for-authors	1
Administrative Science Quarterly	https://journals.sagepub.com/home/asq	Management | Business	SAGE	https://journals.sagepub.com/pb-assets/cmscontent/ASQ/2.%20Data%20and%20Methods%20Transparency-1674844749.pdf	0
American Economic Review	https://www.aeaweb.org/journals/aer	Economics	American Economic Association	https://www.aeaweb.org/journals/aer/about-aer/editorial-policy und ab Juli 2023: https://www.aeaweb.org/journals/data/data-legality-policy	9
Econometrica	https://onlinelibrary.wiley.com/journal/14680262	Economics | Social Sciences, Mathematical Methods	Econometric Society	https://www.econometricsociety.org/publications/econometrica/information-authors/editorial-procedures-and-policies	6

Economic Journal	https://academic.oup.com/ej?login=false	Economics	Oxford University Press	https://academic.oup.com/ej/pages/General_Instructions#data	7
Information Systems Research	https://pubsonline.informs.org/journal/isre	Management | Information Science & Library Science	INFORMS	https://pubsonline.informs.org/page/isre/guidelines-for-ethical-behavior-in-publishing	n/a
International Economic Review	https://onlinelibrary.wiley.com/journal/14682354	Economics	Wiley	https://economics.sas.upenn.edu/ier/submissions/data-availability-policy	0
International Journal of Research in Marketing	https://www.sciencedirect.com/journal/international-journal-of-research-in-marketing	Business	Elsevier	https://www.elsevier.com/journals/international-journal-of-research-in-marketing/0167-8116/guide-for-authors https://www.elsevier.com/authors/tools-and-resources/research-data/data-base-linking	n/a
Journal of Accounting and Economics	https://www.sciencedirect.com/journal/journal-of-accounting-and-economics	Business, Finance | Economics	Elsevier	https://www.elsevier.com/journals/journal-of-accounting-and-economics/0165-4101/guide-for-authors	1
Journal of Accounting Research	https://onlinelibrary.wiley.com/journal/1475679X	Business, Finance	Wiley	https://www.chicagobooth.edu/-/media/research/arc/docs/journal/jardatapolicyasof112022.pdf	n/a
Journal of Consumer Research	https://consumerresearcher.com/about	Business	Oxford University Press	https://consumerresearcher.com/research-ethics	5
Journal of Econometrics	https://www.journals.elsevier.com/journal-of-econometrics	Social Sciences, Mathematical Methods | Economics	Elsevier	https://www.elsevier.com/journals/journal-of-econometrics/0304-4076/guide-for-authors	1
Journal of Economic Literature	https://www.aeaweb.org/journals/jel	Economics	American Economic Association	https://www.aeaweb.org/journals/data/data-code-policy	7

Journal of Economic Theory	https://www.journals.elsevier.com/journal-of-economic-theory	Economics	Elsevier	https://www.elsevier.com/journals/journal-of-economic-theory/0022-0531/guide-for-authors	1
Journal of Finance	https://onlinelibrary.wiley.com/journal/15406261	Economics | Business, Finance	Wiley	https://afajof.org/wp-content/uploads/files/policies-and-guidelines/CodePolicy.pdf	0
Journal of Financial Economics	https://www.sciencedirect.com/journal/journal-of-financial-economics	Business, Finance | Economics	Elsevier	https://www.elsevier.com/journals/journal-of-financial-economics/0304-405X/guide-for-authors	1
Journal of Marketing	https://journals.sagepub.com/home/jmx	Business	SAGE	https://journals.sagepub.com/author-instructions/JMX#2.6 https://www.ama.org/ama-journals-editorial-policies-procedures/	0
Journal of Marketing Research	https://journals.sagepub.com/home/mrj	Business	SAGE	https://www.ama.org/ama-journals-editorial-policies-procedures/ https://journals.sagepub.com/author-instructions/JMX#2.6	0
Journal of Political Economy	https://www.journals.uchicago.edu/toc/jpe/current	Economics	University of Chicago Press	https://www.journals.uchicago.edu/journals/jpe/datapolicy	4
Management Science	https://pubsonline.informs.org/journal/mnsc	Management	INFORMS	https://pubsonline.informs.org/page/mnsc/datapolicy	4
Marketing Science (MS)	https://pubsonline.informs.org/journal/mksc	Business	INFORMS	https://pubsonline.informs.org/page/mksc/replicationpolicy	n/a

MIS Quarterly	https://misq.umn.edu/	Management | Information, Science & Library Science	"Carlson School of Management - University of Minnesota"	https://www.misq.org/code-of-conduct https://misq.umn.edu/research-transparency https://www.misq.org/skin/frontend/default/misq/pdf/MSGuidelines/AIS_Code_of_Research_Conduct.pdf	4
Nature	https://www.nature.com/	Science	Springer	https://www.nature.com/nature/editorial-policies/reporting-standards#availability-of-data%20%20Reporting%20standards%20and%20availability%20of%20data,%20materials,%20code%20and%20protocols	9
Operations-Research	https://pubsonline.informs.org/journal/opre	Management	INFORMS	https://pubsonline.informs.org/page/opre/guidelines-for-ethical-behavior-in-publishing	n/a
Organization Science	https://pubsonline.informs.org/journal/orsc	Management	INFORMS	https://pubsonline.informs.org/page/orsc/guidelines-for-ethical-behavior-in-publishing	0
Quarterly Journal of Economics	https://academic.oup.com/qje	Economics	Oxford University Press	https://academic.oup.com/qje/pages/Data_Policy	4
RAND Journal of Economics	https://www.rje.org/	Economics	Wiley	https://authorservices.wiley.com/ethics-guidelines/index.html#11	0
Review of Economic Studies	https://www.restud.com/	Economics Econometrics	London School of Economics and Political Science; Oxford University Press	https://restud.github.io/data-editor/before/#data-availability-policy	6

Review of Financial Studies	https://academic.oup.com/rfs	Business, Finance | Economics	Oxford University Press	http://rfssfs.org/code-sharing-policy/	2
Science	https://www.science.org/	Science	American Association for the Advancement of Science	https://www.science.org/content/page/science-journals-editorial-policies#publication-policies	11
Strategic Management Journal	https://www.strategicmanagement.net/smj/overview/overview	Management | Business	Wiley	https://onlinelibrary.wiley.com/pb-assets/assets/10970266/SMJ_Author_Instructions_January_2022-1641850654740.pdf	3

DataPLANT – Harnessing the Power of Ontologies for FAIR Research Data Management

Kathryn Dumschott[1], Hannah Dörpholz[1], Kevin Frey[2], Marcel Tschöpe[3], Heinrich Lukas Weil[2], Timo Mühlhaus[2], Dirk von Suchodoletz[3], Björn Usadel[14], Angela Kranz[1]

[1]IBG-4 Bioinformatics, BioSC, Forschungszentrum Jülich;
[2]Computational Systems Biology, RPTU University of Kaiserslautern;
[3]Computer Center, University of Freiburg, Freiburg im Breisgau;
[4]Institute for Biological Data Science, CEPLAS, Heinrich Heine University, Düsseldorf

The NFDI funded DataPLANT consortium aims to provide a sustainable and user-friendly data management platform for the fundamental plant research community. DataPLANT has developed tools and services that encourage open and collaborative research, facilitate the annotation of metadata, and unify the use of ontologies. DataPLANT aims to establish a foundation that enables scientists to effortlessly use and access specific ontologies as well as to expand ontologies with missing terms in order to increase the FAIRness of their research data.

The center of DataPLANT´s developments is the Annotated Research Context (ARC), a data-centric approach to capturing and structuring the entire research cycle. As a structural ontology, the ARC container ontology is designed to help researchers contextualize their data within the ARC and easily compare it to other public ARCs, facilitating the linking of already acquired information to gain knowledge and answer new research questions. Metadata annotation within ARCs is supported by the Swate tool. Swate is linked to the Swate database (SwateDB), which stores a collection of ontologies to facilitate standardized metadata annotation. This collection includes a selection of established ontologies as well as the DataPLANT biology ontology (DPBO), a "broker ontology', which contains missing terms that do not yet appear in known, established ontologies.

1 Introduction

In recent decades, proper research data management (RDM) has become increasingly important with the advent of high throughput methods such as omics (transcriptomics, proteomics, etc.) and imaging. While these methods are incredibly useful for how much information they can provide a researcher, the sheer volume of data produced

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.c18067

requires additional support and infrastructure to facilitate the correct collection, processing and storage of the data. Additionally, data must be properly integrated before any meaningful interpretation can take place. For this reason, the DataPLANT DataHUB (Bauer et al. 2023), an RDM platform that enables the proper storage and annotation of data, increasing its reusability, is important to the fundamental plant science community. As part of the National Research Data Initiative (NFDI; Hartl, Wössner, and Sure-Vetter 2021; Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V. 2023), the DataPLANT (Martins Rodrigues et al. 2021; DataPLANT Consortium 2023) consortium aims to support plant scientists in managing their research data, including data organization, storage and metadata annotation, while adhering to the FAIR principles (Wilkinson et al. 2016). To do so, DataPLANT has developed tools and services that work together seamlessly to store data and annotate metadata quickly and efficiently. Git based versioning (Git community 2023) is employed within the DataPLANT DataHUB to provide complete transparency and version control of all tools, thereby encouraging community contribution and engagement. To encourage the reusability and interoperability of data, ontologies are incorporated into the platform in two ways: as a structural ontology and as an ontology service (Figure 1). By harnessing the potential of ontologies, DataPLANT is able to improve data standardization and metadata annotation, all the while facilitating the linking of previously acquired data for novel discoveries.

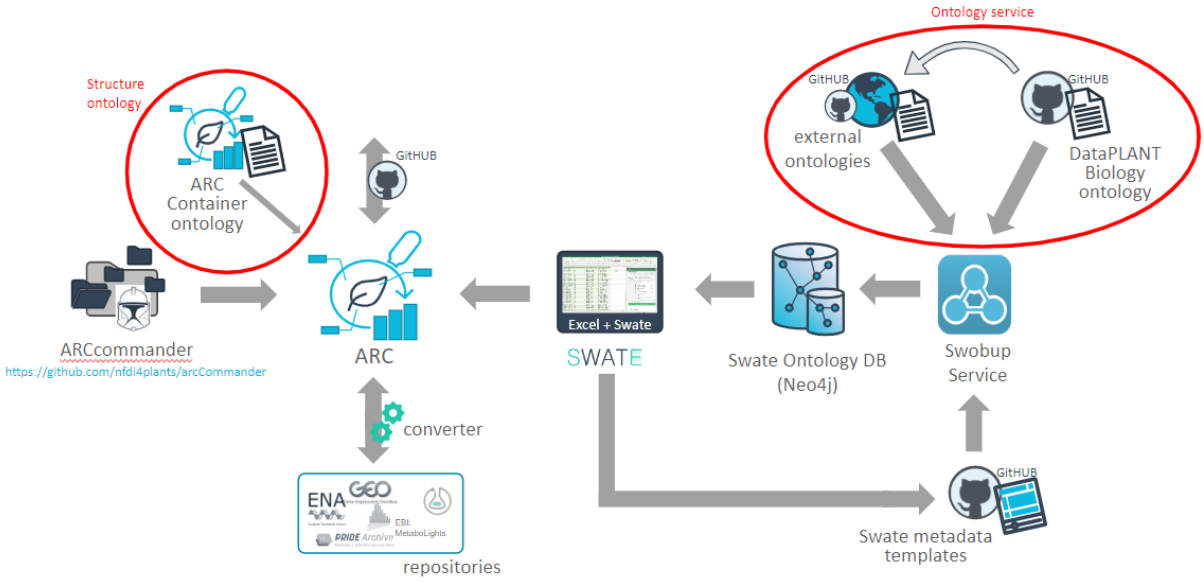


Figure 1: Outline of the DataPLANT DataHUB tools and services. Ontologies are incorporated in two main locations to facilitate FAIR plant research data.

In recent years, ontologies have become important tools for the standardization of data annotation and data reusability in the plant sciences, promoting FAIR principles. By providing unique identifiers for concepts within a domain and describing relationships between them, ontologies ensure that data is structured and can be easily understood by both humans and machines, allowing for machine-based reasoning (Walls et al. 2012). The recorded relationships between terms allows data to not only be machine-readable, but

also facilitates data to be processed in a biologically relevant way, which in turn enables the integration of different data sets.

The DataPLANT consortium harnesses this unique power of ontologies to assist plant scientists in managing their research data in a sustainable and FAIR way. While the platform described here is currently focused on plant research data, the concept can be easily adjusted to meet the requirements and needs of other scientific disciplines. DataPLANT encourages FAIR RDM for all and is therefore actively promoting the presented ontology concept to other scientific communities.

2 The ARC container ontology

FAIR digital objects (FDO) are at the core of all considerations and developments in DataPLANT. To implement a strict data centric approach for RDM, the Annotated Research Context (AC; see Garth et al. 2022; NFDI4Plants 2023a) was designed to capture and structure the complete research cycle to meet the FAIR requirements with low friction for the individual researcher in plant biology. ARCs are self-contained structures that include all biological, measurement, and computational data, as well as relevant metadata, produced during a scientific investigation. Components of ISA (investigation, study and assay; ISA Tools 2023), as well as the CWL (common workflow language; CWL Project 2023) runs and workflows are incorporated to increase the shareability of data. Currently under development, the ARC container ontology (NFDI4Plants 2023b) represents the metadata structure of an ARC (Figure 2). Within the ARC container ontology, the three sub-categories "investigation", "study" and "assay" are represented as a hierarchical structure, with each branch incorporating the required metadata defined by the ISA model. These include "ontology sources' under investigation, "study protocols and materials" under study and "assay technologies and data files" under assay. Major ISA concepts such as "person", "publication" and "ontology source" are represented as classes, while attributes relating to each class are represented as data properties. The individual classes are connected through object properties, creating semantic context and giving the ontology its structure.

The ARC container ontology is important for information inference, or the ability to integrate data from multiple sources. This facilitates the identification of connections and patterns that might not be apparent when investigating individual datasets. This, in turn, enables the linking of already acquired information to gain novel insights and answer new research questions. For instance, if two experiments use varying methods to measure the same set of variables, the ARC container ontology can help to align the data and make meaningful comparisons between them. Another advantage of using the ARC container ontology for information inference is the ability to apply reasoning and inference algorithms to the data. This approach can reveal implicit relationships and dependencies that are not explicitly stated in the metadata, leading to a deeper understanding of the data.

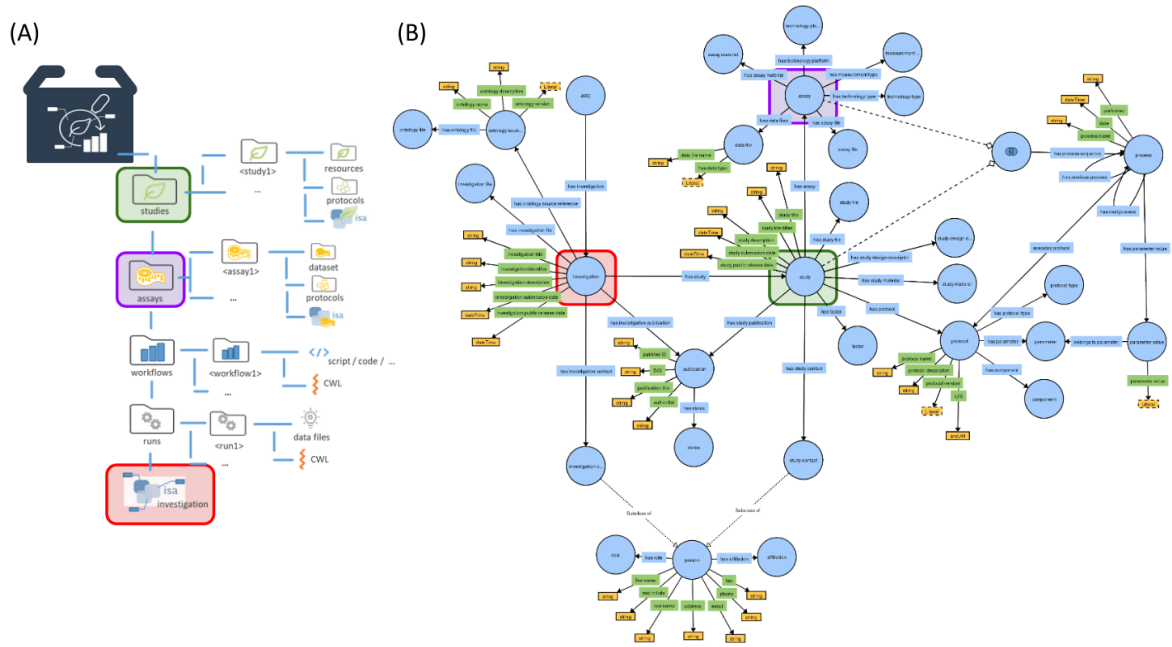


Figure 2: (A) The Annotated Research Context (ARC) is the central component of the DataPLANT DataHUB, where research data is structured and stored. (B) Excerpt of the ARC container ontology, which represents the metadata structure of the ARC. Classes are shown in blue circles, object properties are shown in blue rectangles, datatype properties and their datatype ranges are shown in green and yellow rectangles, respectively. The ontology was visualized using WebVOWL (`https://service.tib.eu/webvowl`).

3 The DataPLANT ontology service

A crucial component of RDM is the complete and proper annotation of metadata. Metadata, or the data about data, gives the context and background of studies and assays performed within an investigation. Raw data is essentially meaningless without accompanying metadata that provides information about the sample acquisition process and machine parameters used for measurement. For this reason, DataPLANT has developed a suite of seamlessly integrated tools for the annotation of metadata that work together to encourage plant researchers to annotate their data as early as possible in the experimental timeline (Mühlhaus et al. 2022).

Swate (NFDI4Plants 2023d), DataPLANT's own ontology-driven metadata annotation tool, simplifies the annotation of assay and study metadata for plant researchers. Swate is available both online and as a plugin for Microsoft Excel, enabling users to incorporate standard spreadsheet features into their metadata annotation protocol, making it convenient and user-friendly. Users can easily modify their ISA-compliant sheets by adding or removing building blocks in order to describe the necessary metadata in a clear and concise way. Building blocks can cover relevant characteristics, parameters, factors, com-

ponents and protocols. When a building block is added to the spreadsheet as a header, it is tagged with the appropriate ontology term, referred to as the "parent term". To facilitate data input, users can then utilize the "Ontology term search" tab to fill in the values corresponding to the building block heading. These values are also tagged with the appropriate ontology term, known as "child term" (Figure 3). This functionality allows Swate to not only tag the headings of metadata sheets with ontology terms, but to also tag the respective column values, ensuring comprehensive and structured annotation. While Swate provides pre-suggested terms for selection, users have the flexibility to add their own terms not linked to a current ontology term. This flexibility accommodates diverse research needs while still promoting the essential task of metadata annotation. To assist scientists in initiating the metadata annotation process, Swate also provides a variety of different templates which already include building blocks based on metadata required by repositories or minimum information standards (see section 4).

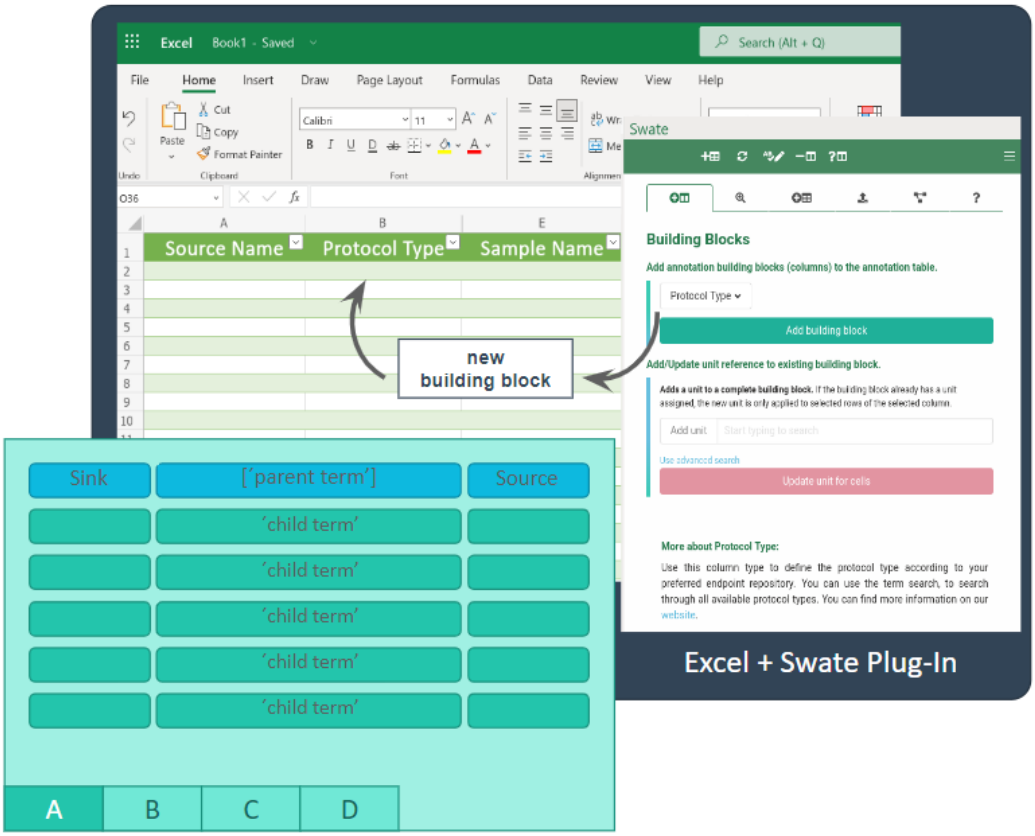


Figure 3: Swate incorporates ontologies within the SwateDB to facilitate the annotation of study and assay metadata.

To adequately describe the metadata essential for plant science experiments and investigations, a diverse range of ontologies spanning various domains is required. The DataPLANT ontology team collaborates closely with the plant science community to identify ontologies that encompass a substantial number of terms required for metadata annotation of plant-related experiments. Selected ontologies are imported into a database

called SwateDB. Currently, included ontologies cover domains related to the plant sciences, such as the Plant Ontology (PO; see Walls et al. 2012), more general biology and chemistry domains, such as the Chemical Entities of Biological Interest (CheBI; see Degtyarenko et al. 2007), as well as technical terminologies required for the acquisition and analysis of biological data. The full list of external ontologies currently included in SwateDB can be found in the nfdi4plants_ontology Github repository (`https://github.com/nfdi4plants/nfdi4plants_ontology/blob/main/ext_ontologies.include`; see NFDI4Plants 2023c). While the included external ontologies are mostly applicable for fundamental plant research, further ontologies relevant for other scientific disciplines (e.g. microbiology) can easily be added to SwateDB and be used for metadata annotation. In addition to established ontologies, DataPLANT has developed its own DataPLANT Biology Ontology (DPBO), which is also included in SwateDB. The DPBO serves two primary goals: addressing the ontology gap by collecting missing vocabulary required to annotate metadata and acting as a middleman between the researcher and the main ontology provider. Community contributions are encouraged and suggestions for new terms or improvements to already existing terms can be easily submitted via the DataPLANT helpdesk[1] or the Issues tab located at the nfdi4plants_ontology repository[2]. Once a term suggestion or improvement has been submitted, the ontology team incorporates the information into the OBO file (Figure 4). This process involves assigning a unique ID to the term, adding information such as the term definition or relevant synonyms, and incorporating the term into the ontology structure (for example, via is_a to delineate parent-child relations). As terms are suggested by the fundamental plant research community, the DataPLANT ontology team actively liaises with the main ontology providers, serving as a broker between the researchers and the ontology providers.

Once the OBO file has been saved, the term undergoes an automatic process facilitated by the Swate OBO Updater (Swobup; see NFDI4Plants 2023f), enabling its integration into SwateDB and subsequent availability within Swate (Figure 5). During this automatic step, Swobup reads OWL-compatible files retrieved in an update in a Git repository (Git community 2023), and applies the changes to the graph database (Neo4j[3]). With Git, the ontology source files are tracked, ensuring the integrity and versioning of the files. This approach enables community engagement in ontology development through standard mechanisms successfully employed in open source software development. Consequently, ontology changes and updates can be swiftly incorporated into SwateDB, making them readily available in Swate. This includes changes to the DPBO as well as the collection of external ontologies, although, in contrast to the external ontologies, terms deleted from the DPBO will also be removed from SwateDB. The file versioning feature further facilitates easy reversal to previous file versions, which can seamlessly be incorporated into Swate via Swobup.

1 `https://helpdesk.nfdi4plants.org`
2 `https://github.com/nfdi4plants/nfdi4plants_ontology`
3 `https://neo4j.com`

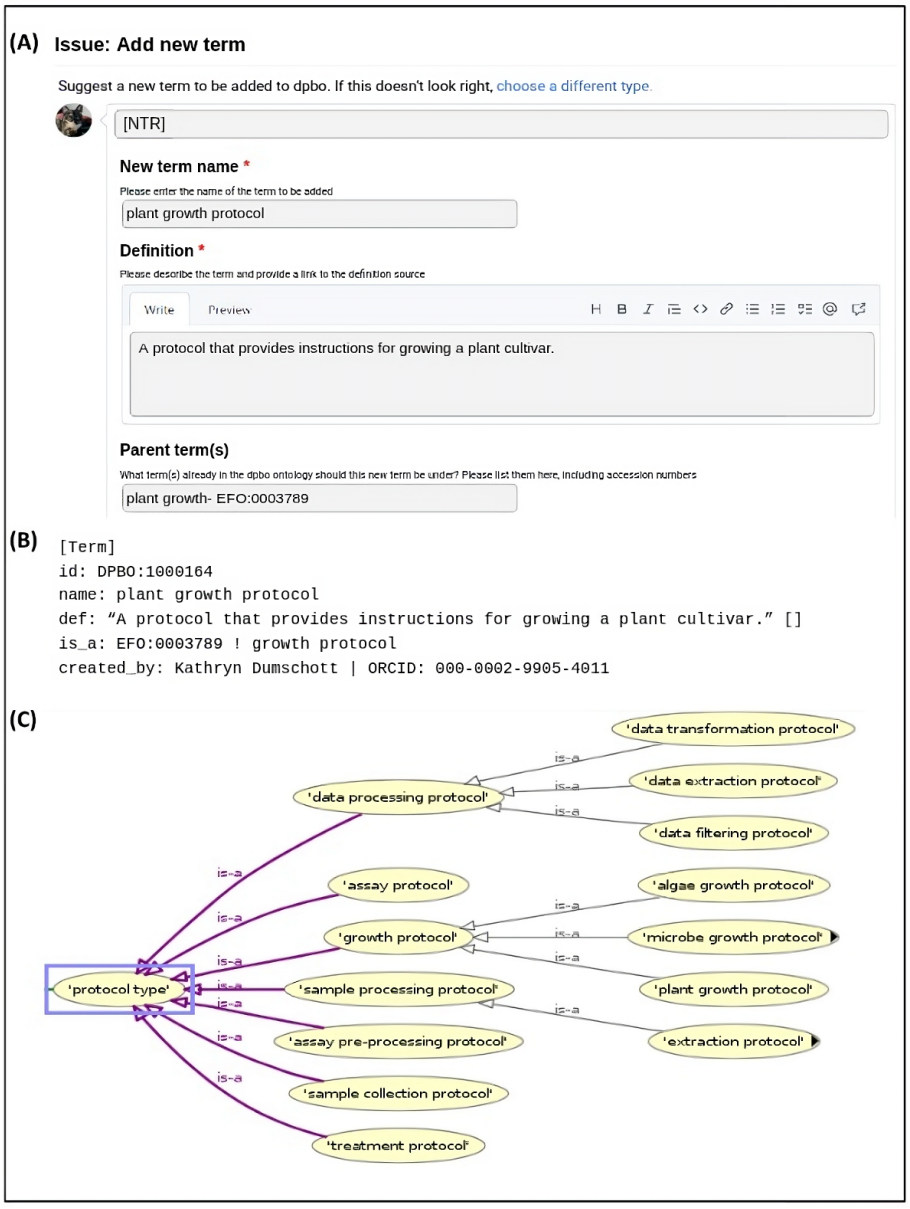


Figure 4: The DPBO is curated with the help of the scientific community (A) users can submit suggestions for new ontology terms or additions to already existing terms via the Issues tab at `https://github.com/nfdi4plants/nfdi4plants_ontology` or the DataPLANT Helpdesk (`https://helpdesk.nfdi4plants.org`) (B) the DataPLANT ontology team incorporates the term into DPBO (C) once the DPBO file has been updated with the new term, it is written to SwateDB by Swobup, making the term readily available within Swate.

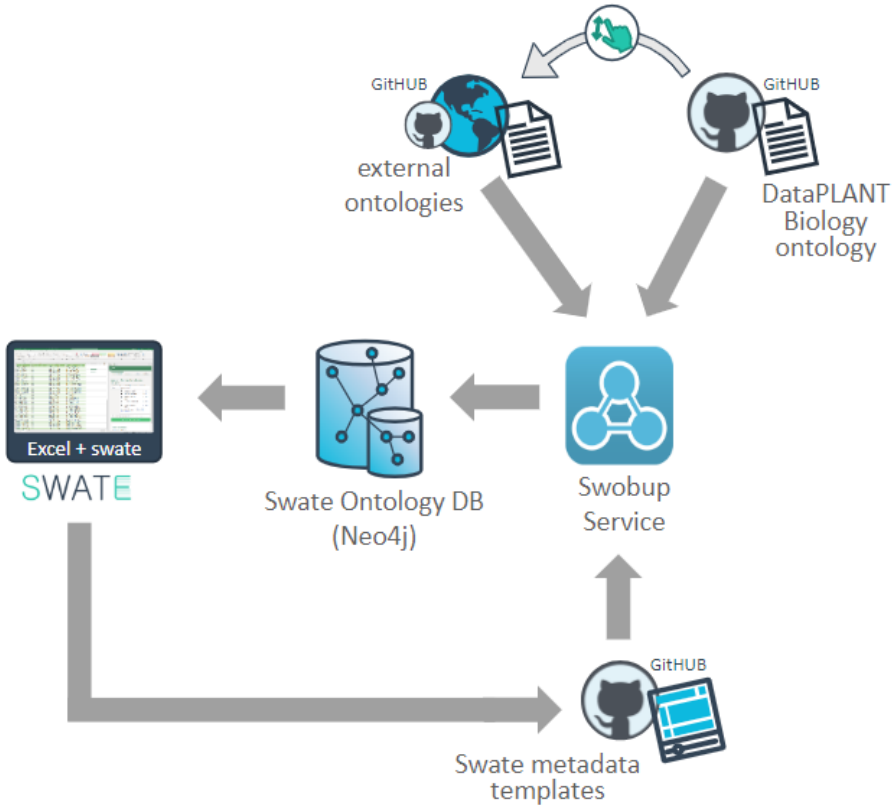


Figure 5: The Swobup service and SwateDB work together to make the annotation of metadata sheets within Swate possible.

4 Templates

A crucial first step in FAIR RDM is knowing what metadata should be included when annotating a study or assay, before even knowing what ontologies are relevant. Necessary metadata can be found in checklists required by repositories when uploading data, or else based on minimum information standards such as MIAPPE (Papoutsoglou et al. 2020), MIAME (Brazma et al. 2001), or MIMARKS (Yilmaz et al. 2011). To assist researchers in beginning the process of metadata annotation, Swate provides a collection of templates, which outline the metadata required by the repository or minimum information standard, located under the "Templates tab" in Swate. Researchers can search for relevant templates or filter templates by tags. When a template is selected, the researcher is returned to the main page of Swate, where all the building blocks contained in the template are displayed. The researcher can then click "Add template" to add the template building block to the Excel sheet. Selected templates can be used as-is or else amended to fit the exact need of the researcher, so only building blocks not yet included in the table will be added, and any unnecessary building blocks can be easily removed. To ensure total control and flexibility over the metadata annotation process, users also have the option of creating and saving their own metadata templates. These can reflect the specific experiments or

protocol performed during an investigation and can be shared with other members in the group to help standardize metadata annotation within institutes and to encourage new members to do so as early in the investigation timeline as possible. A template is created in the same way as Swate metadata sheets (described above), but with an additional SwateMetadataSheet, which must be filled out to create a function template. Once the template has been added to the repository, it is incorporated into SwateDB via the same Swobup process described above. A detailed description of how to create customized templates can be found in the Swate template Github repository (NFDI4Plants 2023e). As with all of DataPLANT's tools and standards, the use of Git when creating and uploading templates ensures data integrity and versioning control.

5 Conclusions

With the DataPLANT DataHUB, the DataPLANT consortium aims to provide a sustainable and well-annotated data management platform that encourages open and collaborative research and facilitates annotation of metadata. The consortium has developed flexible, adaptable tools and services that incorporate ontologies in two ways to support researchers in increasing the FAIRness of their plant-related research data.

Firstly, the ARC Container ontology is an organizational representation of the data-centric Annotated Research Context (ARC). While still under development, the ARC ontology already incorporates components of ISA (investigation, study and assay) and is being actively expanded to conclude CWL runs and workflows. The ontology plays a crucial role for information inference and enables the linking of acquired data to gain novel insights that may not be apparent when investigating individual datasets.

Furthermore, ontologies are incorporated within the ontology service, which supports the annotation of metadata sheets via the Swate and Swobup tools. The DBPO acts as a broker ontology, giving researchers the opportunity to efficiently add new terms that may be absent from established ontologies. The DataPLANT ontology team then acts as the middle man, curating the terms and suggesting them back to the relevant established ontologies. In addition to ontologies, a series of templates is provided to aid researchers in beginning their journey into metadata annotation. These are based on checklists included in repositories and minimum information standards. Researchers can select available templates, subsections of available templates, or else design their own based on experiments or protocols commonly performed during their research.

Most importantly, the tools and services within the platform are data centric and serve to improve the user experience while implementing current state of the art practices and versioning control. The flexibility of the DataPLANT concept means it can be easily adapted to fit the needs of researchers in other scientific domains. In the spirit of open source, DataPLANT encourages users to openly contribute and collaborate, thereby continuously improving the RDM landscape for all.

Acknowledgements

We acknowledge support for DataPLANT 442077441 through the German National Research Data Initiative (NFDI 7/1) and CEPLAS is supported by Deutsche Forschungsgemeinschaft within the Excellence Initiative (EXC 1028) and under Germany's Excellence Strategy – EXC 2048/1 – project 390686111.

References

Bauer, Jonathan, Marcel Tschöpe, Julian Weidhase, Timo Mühlhaus, Christoph Garth, Gajendra Doniparthi, Holger Gauza, Louisa Perelo, Cristina Martins Rodrigues, and Dirk von Suchodoletz. 2023. "From DataPLANT's DataHUB to DataPUB(lication)". In *International Workshop on Science Gateways.* Accepted for publication.

Brazma, Alvis, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, et al. 2001. "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data". *Nature Genetics* 29 (4): 365–371. DOI: https://doi.org/10.1038/ng1201-365.

CWL Project. 2023. "Common Workflow Language". Visited on May 30, 2023. https://www.commonwl.org/.

DataPLANT Consortium. 2023. "DataPLANT". Visited on May 30, 2023. https://www.nfdi4plants.de/.

Degtyarenko, Kirill, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcantara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. "ChEBI: a database and ontology for chemical entities of biological interest". *Nucleic Acids Research* 36 (Database): D344–D350. DOI: https://doi.org/10.1093/nar/gkm791.

Garth, Christoph, Jonas Lukasczyk, Timo Mühlhaus, Benedikt Venn, Jens Krüger, Kolja Glogowski, Cristina Martins Rodrigues, and Dirk Von Suchodoletz. 2022. "Immutable yet evolving: ARCs for permanent sharing in the research data-time continuum". In *E-Science-Tage 2021: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 366–373. heiBOOKS. DOI: https://doi.org/10.11588/heibooks.979.c13751.

Git community. 2023. "Git". Visited on May 30, 2023. https://git-scm.com/.

Hartl, Nathalie, Elena Wössner, and York Sure-Vetter. 2021. "Nationale Forschungsdateninfrastruktur (NFDI)". *Informatik Spektrum* 44 (5): 370–373. DOI: https://doi.org/10.1007/s00287-021-01392-6.

ISA Tools. 2023. "ISA Model and Serialization Specifications". Visited on May 30, 2023. https://isa-specs.readthedocs.io/en/latest/isamodel.html.

Martins Rodrigues, Cristina, Dirk von Suchodoletz, Timo Mühlhaus, Jens Krüger, and Björn Usadel. 2021. "DataPLANT – Ein NFDI-Konsortium der Pflanzen-Grundlagenforschung". *Bausteine Forschungsdatenmanagement,* number 2 (2): 46–56. DOI: https://doi.org/10.17192/bfdm.2021.2.8335. https://bausteine-fdm.de/article/view/8335.

Mühlhaus, Timo, Dominik Brillhaus, Marcel Tschöpe, Oliver Maus, Björn Grüning, Christoph Garth, Cristina Martins Rodrigues, and Dirk Von Suchodoletz. 2022. "DataPLANT – Tools and Services to structure the Data Jungle for fundamental plant researchers". In *E-Science-Tage 2021: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 132–145. heiBOOKS. DOI: https://doi.org/10.11588/heibooks.979.c13724.

NFDI4Plants. 2023a. "Annotated Research Contexts specification". Visited on May 30, 2023. https://github.com/nfdi4plants/ARC-specification.

———. 2023b. "ARC Ontology". Visited on May 30, 2023. https://github.com/nfdi4plants/ARC_ontology.

———. 2023c. "NFDI4Plants Ontology - An intermediate ontology for plants used by DataPLANT to fill the ontology gap". Visited on May 30, 2023. https://github.com/nfdi4plants/nfdi4plants_ontology.

———. 2023d. "Swate (Excel Add-In for annotation of experimental data and computational workflows". Visited on May 30, 2023. https://github.com/nfdi4plants/Swate.

———. 2023e. "Swate Templates". Visited on May 30, 2023. https://github.com/nfdi4plants/Swate-templates.

———. 2023f. "Swobup". Visited on May 30, 2023. https://github.com/nfdi4plants/Swobup.

Papoutsoglou, Evangelia A., Daniel Faria, Daniel Arend, Elizabeth Arnaud, Ioannis N. Athanasiadis, Inês Chaves, Frederik Coppens, et al. 2020. "Enabling reusability of plant phenomic datasets with MIAPPE 1.1". *New Phytologist* 227 (1): 260–273. DOI: https://doi.org/10.1111/nph.16544.

Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V. 2023. "Nationale Forschungsdaten Infrastruktur (NFDI)". Visited on September 4, 2023. https://www.nfdi.de/.

Walls, Ramona L., Balaji Athreya, Laurel Cooper, Justin Elser, Maria A. Gandolfo, Pankaj Jaiswal, Christopher J. Mungall, et al. 2012. "Ontologies as integrative tools for plant science". *American Journal of Botany* 99 (8): 1263–1275. DOI: https://doi.org/10.3732%2Fajb.1200222.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Yilmaz, Pelin, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, et al. 2011. "Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications". *Nature Biotechnology* 29 (5): 415–420. DOI: https://doi.org/10.1038/nbt.1823.

Data Repositories 4Culture – Bedarfsorientierte Forschungsdatenrepositorien für den Kulturbereich

Alexandra Büttner [1], Sandra Göller [2], Peggy Große [2], Kerstin Soltau [2]

[1]Universitätsbibliothek, Universität Heidelberg;
[2]FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur

Forschende der kulturwissenschaftlichen Disziplinen an Universitäten und anderen Kultureinrichtungen stehen zunehmend vor der Aufgabe, ihre digitalen Forschungsdaten nachhaltig sichern zu müssen. Der Beitrag stellt zwei Repositorien für Forschungsdaten aus dem Bereich des kulturellen Erbes vor: arthistoricum.net@heiDATA und RADAR4Culture. Beide Angebote ermöglichen Wissenschaftler:innen ihre Daten nach den FAIR-Prinzipien zu publizieren und langfristig verfügbar zu machen. Durch den Einsatz von DOIs (Digital Object Identifier) sind die Daten dauerhaft zitierfähig, gezielt verlinkbar und als eigenständige wissenschaftliche Leistungen sichtbar. Die abgelegten Forschungsdaten werden durch Metadaten, z. T. unter Verwendung von Normdaten, beschrieben, deren Indexierung eine maximale Dissemination der Daten möglich macht. Sowohl arthistoricum.net@heiDATA als auch RADAR4Culture gehören zum Portfolio an empfohlenen Diensten des DFG-geförderten Konsortiums NFDI4Culture.

Das Konsortium konzentriert sich vor allem auf Forschungsdaten aus der Architektur, den Kunst-, Musik-, Theater-, Tanz-, Film- und Medienwissenschaften und bietet Forschenden und Infrastrukturbetreibenden eine bedarfsorientierte Unterstützung und Beratung zu allen Phasen des Datenlebenszyklus an.

1 Einleitung

Das Konsortium für Forschungsdaten materieller und immaterieller Kulturgüter – NFDI4Culture[1] widmet sich dem Aufbau einer bedarfsgerechten Infrastruktur für Forschungsdaten im Kulturbereich. NFDI4Culture wird durch die Deutsche Forschungsgemeinschaft (DFG) gefördert und ist Teil der Nationalen Forschungsdateninfrastruktur e.V. (NFDI)[2].

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.c18068

1 https://nfdi4culture.de; *Zuletzt aufgerufen am 17. August 2023.*
2 https://www.nfdi.de; *Zuletzt aufgerufen am 17. August 2023.*

Im Fokus stehen bei NFDI4Culture die Bedürfnisse von Wissenschaftler:innen der Fachbereiche Kunstgeschichte und Architektur, Musikwissenschaft, Darstellende Kunst sowie der Film- und Medienwissenschaften. Darüber hinaus steht NFDI4Culture im engen Austausch mit weiteren Konsortien[3], insbesondere der Geistes- und Sozialwissenschaften, wie z.B. BERD@NFDI[4], KonsortSWD[5], NFDI4Memory[6], NFDI4Objects[7] und Text+[8].

NFDI4Culture hat seine Aufgaben entlang des Forschungsdatenlebenszyklus unterteilt und insgesamt sieben operative Bereiche geschaffen (Abbildung 1). Die ersten sechs Arbeitsbereiche, sog. Task Areas (TA), widmen sich ausschließlich den Prozessen und Phasen innerhalb des Forschungsdatenlebenszyklus: (TA1) Digitalisierung und Anreicherung digitaler Kulturgüter; (TA2) Standards, Datenqualität und Kuratierung; (TA3) Forschungswerkzeuge und Datendienste; (TA4) Datenpublikation und Langzeitarchivierung; (TA5) Übergreifende technische, ethische und rechtliche Aktivitäten; (TA6) Cultural Research Data Academy. Der siebte Arbeitsbereich (TA7) Governance und Administration umfasst den gesamten Zyklus und übernimmt die administrativen Aufgaben sowie die Kommunikations- und Entscheidungsprozesse innerhalb des Konsortiums (Altenhöner u. a. 2020).

Im Bereich des kulturellen Erbes nimmt die Komplexität der Forschungsdaten sowie der einzelnen Schritte im Zuge des Datenlebenszyklus stetig zu und stellt sowohl Repositorienbetreibende als auch Forschende vor wachsende Herausforderungen. Durch die Optimierung von Diensten in den 4Culture-Fachbereichen reagiert der Aufgabenbereich (TA4) Datenpublikation und Langzeitarchivierung[9] darauf und unterstützt Forschende aller 4Culture-Disziplinen. In enger Zusammenarbeit mit den Fachcommunities sollen über die Projektlaufzeit –vorläufig auf fünf Jahre festgelegt – hinweg Lücken in der bestehenden Forschungsdateninfrastruktur geschlossen sowie nachhaltige und zuverlässige Dienste etabliert werden. In einem ersten Schritt wurde der Community eine Kuratierte Repositorienliste[10] mit einem Überblick über bereits bestehende, fachspezifische und generische Publikations- bzw. Archivierungsangebote bereitgestellt. Vorausgegangen war eine detaillierte Kartierung aktuell vorhandener Infrastrukturen – unabhängig davon, ob diese von einer der für Culture verantwortlichen Projektpartner selbst oder von einer anderen Einrichtung im 4Culture-Umfeld betrieben wird. Durch die Zusammenarbeit mit dem Fachin-

3 `https://www.nfdi.de/konsortien`; *Zuletzt aufgerufen am 17. August 2023.*
4 BERD@NFDI: NFDI für Betriebswirtschaftslehre, Volkswirtschaftslehre und verwandte Daten. Siehe `https://www.berd-nfdi.de`; *Zuletzt aufgerufen am 17. August 2023.*
5 KonsortSWD: Konsortium für die Sozial-, Bildungs-, Verhaltens- und Wirtschaftswissenschaften. Siehe `https://www.konsortswd.de`; *Zuletzt aufgerufen am 17. August 2023.*
6 NFDI4Memory: Konsortium für historisch arbeitende Geisteswissenschaften. Siehe `https://4memory.de`; *Zuletzt aufgerufen am 17. August 2023.*
7 NFDI4Objects: Forschungsdateninfrastruktur für die materiellen Hinterlassenschaften der Menschheitsgeschichte. Siehe `https://www.nfdi4objects.net`; *Zuletzt aufgerufen am 17. August 2023.*
8 Text+: Sprach- und textbasierte Forschungsdateninfrastruktur. Siehe `https://www.text-plus.org`; *Zuletzt aufgerufen am 17. August 2023.*
9 `https://nfdi4culture.de/de/ueber-uns/aufgabenbereiche/aufgabenbereich-4.html`; *Zuletzt aufgerufen am 17. August 2023.*
10 `https://nfdi4culture.de/de/ressourcen/repositorien.html`; *Zuletzt aufgerufen am 17. August 2023.*

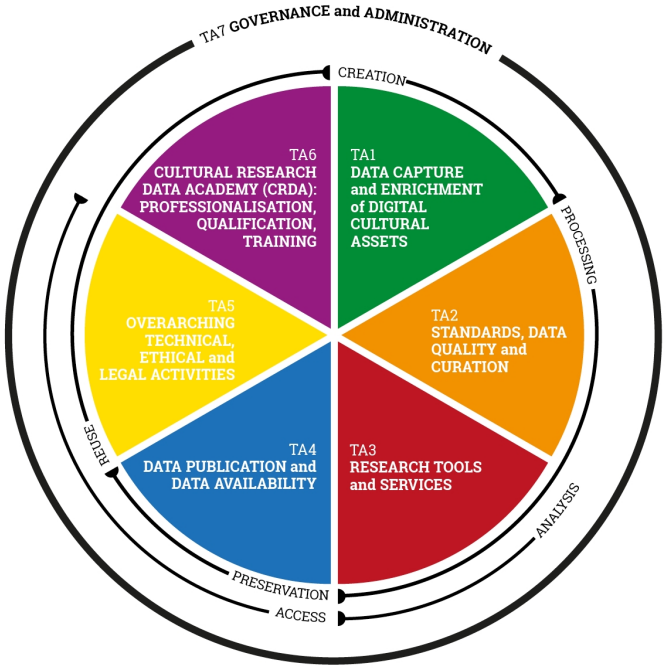


Abbildung 1: NFDI4Culture – Forschungsdatenlebenszyklus und Aufgabenbereiche/Task Areas (NFDI4Culture 2023).

formationsdienst (FID) für Kunst, Design und Fotografie – arthistoricum.net[11], dem Fachinformationsdienst Musikwissenschaft – musiconn – Für vernetzte Musikwissenschaft[12], dem Fachinformationsdienst BAUdigital[13] sowie dem Fachinformationsdienst für Medien-, Kommunikations- und Filmwissenschaft – adlr.link[14] existieren bereits etablierte Angebote u.a. für Textpublikationen und Bilddateien. Der FID Kunst – arthistoricum.net bietet beispielsweise mit ART-Dok[15] ein Open-Access-Repositorium für Aufsätze, Monographien und Rezensionen an. E-Books können auf arthistoricum.net–ART-Books (inkl. Print-on-

11 `https://www.arthistoricum.net`; *Zuletzt aufgerufen am 17. August 2023.*
12 `https://musik.fid-lizenzen.de`; *Zuletzt aufgerufen am 17. August 2023.*
13 `https://www.fid-bau.de`; *Zuletzt aufgerufen am 17. August 2023.*
14 `https://www.ub.uni-leipzig.de/forschungsbibliothek/projekte/projekte-chronologisch-alle/fachinformationsdienst-fuer-medien-und-kommunikationswissenschaft`; *Zuletzt aufgerufen am 17. August 2023.*
15 `https://www.arthistoricum.net/publizieren/art-dok`; *Zuletzt aufgerufen am 17. August 2023.*

Demand-Angebot) online gestellt und im Multimediarepositorium heidICON u.a. zugehörige Bild-, AV- und 3D-Dateien veröffentlicht werden. Auch die anderen FIDs bieten vergleichbare Angebote an, wie z.B. media/rep[16], das Open-Access-Repositorium für medienwissenschaftliche Publikationen, oder musiconn.publish[17] für die Veröffentlichung und langfristige Archivierung musikwissenschaftlicher Fachliteratur.

Bei der Analyse der bestehenden Publikations- und Archivierungslandschaft in den 4Culture-Disziplinen wurden dennoch nicht nur für einzelne Fachbereiche, sondern auch bezüglich bestimmter Dateiformate Lücken im Dienstangebot festgestellt. Neben Text- und Bilddaten besteht in der 4Culture-Community zunehmend der Bedarf, AV-Daten, 3D-Daten sowie strukturierte Daten (z. B. Tabellenstrukturen) und Metadaten (Markup-Languages) zu veröffentlichen bzw. langfristig verfügbar zu halten. Im Folgenden werden zwei Repositorien aus dem Bereich des kulturellen Erbes vorgestellt, die diese Bedarfe aus der Community aufgreifen und eine Publikations- und Archivierungsmöglichkeit für komplexe Forschungsdaten nach den FAIR-Prinzipien (Wilkinson u. a. 2016) anbieten: arthistoricum.net@heiDATA und RADAR4Culture.

2 arthistoricum.net@heiDATA

Im Bereich Kunstgeschichte besteht neben den oben erwähnten Diensten ein Open-Access-Publikations- und Archivierungsangebot für genau diese komplexeren Forschungsdaten. arthistoricum.net@heiDATA ist das Forschungsdatenrepositorium für kunstwissenschaftliche Daten und richtet sich an Kunsthistoriker:innen weltweit (Abbildung 2). Das Repositorium ist ein Angebot der Universitätsbibliothek Heidelberg im Rahmen von arthistoricum.net, gefördert durch die Deutsche Forschungsgemeinschaft (DFG). arthistoricum.net@heiDATA ist Teil des universitären Forschungsdatenrepositoriums heiDATA[18], das vom Kompetenzzentrum Forschungsdaten (KFD), einer gemeinsamen Serviceeinrichtung des Universitätsrechenzentrums (URZ) und der Universitätsbibliothek Heidelberg, betrieben wird (Apel u. a. 2018). Die institutionelle Veröffentlichungsplattform für Open-Research-Data basiert auf der an der Harvard University entwickelten Open-Source-Software Dataverse. Bis heute gibt es rund 100 Dataverse-Installationen weltweit, die von den konstanten Software-Weiterentwicklungen innerhalb der Dataverse Project-Community profitieren.[19]

Grundvoraussetzung für erfolgreiche Forschung ist ein effizientes und nachhaltiges Datenmanagement. Im Kontext von arthistoricum.net werden Forschende zu allen Aspekten des Forschungsdatenlebenszyklus und insbesondere zum Management ihrer kunstwissenschaftlichen Daten bis hin zur Publikation und Archivierung beraten. Alle auf arthistoricum.net@heiDATA veröffentlichten Forschungsdaten werden im universitären Langzeitarchivsystem heiARCHIVE[20] nachhaltig archiviert. Der Publikations- und Langzeitarchi-

16 `https://mediarep.org`; *Zuletzt aufgerufen am 17. August 2023.*
17 `https://musiconn.qucosa.de`; *Zuletzt aufgerufen am 17. August 2023.*
18 `https://heidata.uni-heidelberg.de`; *Zuletzt aufgerufen am 17. August 2023.*
19 `https://dataverse.org`; *Zuletzt aufgerufen am 15. Mai 2023.*
20 `https://heiarchive.uni-heidelberg.de`; *Zuletzt aufgerufen am 17. August 2023.*

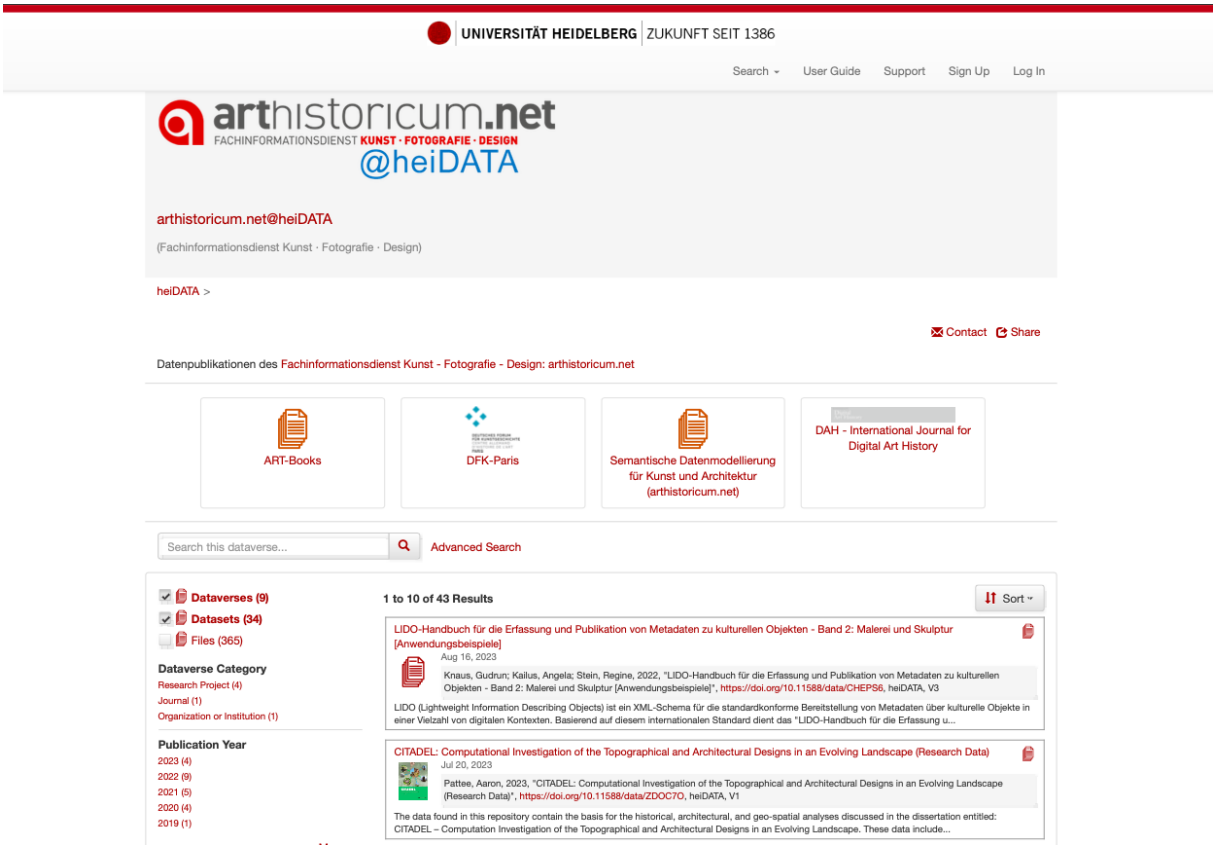


Abbildung 2: arthistoricum.net@heiDATA – Forschungsdatenrepositorium für kunstwissenschaftliche Daten / Task Areas (NFDI4Culture 2023).

vierungsservice ist für Kunstwissenschaftler:innen im Rahmen von arthistoricum.net kostenfrei. Bedingung für die Veröffentlichung auf arthistoricum.net@heiDATA ist aktuell die Verknüpfung der Forschungsdaten mit einer wissenschaftlichen Textveröffentlichung auf einer der Publikationsplattformen von arthistoricum.net. Alle Forschungsdaten sind im Open Access zugänglich und gemäß der FAIR-Prinzipien publiziert um die Sichtbarkeit, Transparenz und Reproduzierbarkeit der veröffentlichten Daten zu gewährleisten. Das Forschungsdatenrepositorium arthistoricum.net@heiDATA ist in re3data, der Registry of Research Data Repositories, die einen Überblick über die internationale Landschaft von Datenrepositorien bietet, als fachspezifisches Repositorium gelistet (Re3data.Org 2021).

arthistoricum.net@heiDATA und die Struktur eines Dataverse-Repositoriums

Der Aufbau eines Dataverse-Repositoriums bietet die Möglichkeit, Forschungsdaten sowohl auf Sammlungs-, als auch auf Dateiebene gezielt mit relevanten Projekten und Textpublikationen zu verknüpfen. Ein Dataverse ist in einzelne Sub-Dataverses unterteilt, sog. Sammlungen von Datasets. Das Layout eines Dataverse kann leicht angepasst und mit einem Logo oder einem kurzen Teaser-Text versehen werden. Dies ermöglicht z. B. Forschungsgruppen bei größeren Datenmengen ein eigenes Branding ihrer Publikation. Datasets wiederum sind Sammlungen von einzelnen Data-Files. Sowohl Datasets als

auch einzelne Files erhalten mit dem DOI (Digital Object Identifier) persistente Identifier und sind somit nach der Veröffentlichung direkt referenzierbar und zitierfähig. Auf der Ebene der Datasets wird die Vergabe standardisierter und fachspezifischer Metadaten ermöglicht. So können die Inhalte der Datasets mit folgenden, auch disziplinspezifischen Standards beschrieben werden, wie z. B. GND[21], Wikidata[22], AAT[23] oder zur fachlichen Einordnung mit LCSH[24]. Für Personen wird neben VIAF[25], ISNI[26] und GND auch die ORCID[27] als Identifier unterstützt.

Auf arthistoricum.net@heiDATA werden alle Daten stets in einen Forschungskontext eingebettet und mit einer kunstwissenschaftlichen Textveröffentlichung auf arthistoricum.net verknüpft. In dem Metadatenfeld „Related Publication" wird der Zitierhinweis mit persistentem DOI auf die relevante Publikation eingetragen, um so zu garantieren, dass die multimedialen und multilokalen Publikationen aufeinander verweisen. Auch von der Textpublikation wird auf die zugrunde liegenden Forschungsdaten per DOI zurückverwiesen. Hier können zum Beispiel Aufsätze auf ART-Dok, Monographien auf ART-Books, oder auch Artikel aus E-Journals auf arthistoricum.net verlinkt werden.[28] Die Beschreibung des Aufbaus eines Dataverse-Repositoriums soll zeigen, wie die Forschungsdaten strukturiert publiziert und auf verschiedenen Ebenen gemäß der FAIR-Prinzipien direkt adressiert und zitiert werden können.

arthistoricum.net@heiDATA und die FAIR-Prinzipien

Ziel des Forschungsdatenmanagements allgemein ist die Aufbereitung der Daten gemäß der FAIR-Data-Prinzipien (Findable, Accessible, Interoperable und Reusable).[29] Wie die Publikationen von Forschungsdaten auf arthistoricum.net@heiDATA den FAIR-Prinzipien entsprechen, soll im Folgenden noch näher erläutert werden. Die Zusammenschau der Umsetzung der FAIR-Prinzipien bei arthistoricum.net@heiDATA in Abbildung 3 kann im Prinzip auf alle Dataverse Repositorien übertragen werden (Crosas 2019). Die Auffindbarkeit – Findability – wird, wie bereits erwähnt, zum einen durch die Dokumentation und Beschreibung der Forschungsdaten anhand standardisierter Metadaten und Normda-

21 GND (Gemeinsame Normdatei): `https://gnd.network`; *Zuletzt aufgerufen am 17. August 2023.*

22 `https://www.wikidata.org`; *Zuletzt aufgerufen am 17. August 2023.*

23 AAT (Art & Architecture Thesaurus – Getty Research Institute): `https://www.getty.edu/research/tools/vocabularies/aat`; *Zuletzt aufgerufen am 17. August 2023.*

24 LCSH (Library of Congress Subject Headings): `https://id.loc.gov/authorities/subjects.html`; *Zuletzt aufgerufen am 17. August 2023.*

25 VIAF (Virtual International Authority File): `https://viaf.org`; *Zuletzt aufgerufen am 17. August 2023.*

26 ISNI (International Standard Name Identifier): `https://isni.org`; *Zuletzt aufgerufen am 17. August 2023.*

27 ORCID (Open Researcher and Contributor ID): `https://orcid.org`; *Zuletzt aufgerufen am 17. August 2023.*

28 Die Publikationsplattform ART-Books basiert auf Open Monograph Press (OMP) (`https://pkp.sfu.ca/software/omp/`; *Zuletzt aufgerufen am 17. August 2023.*) und zur Publikation der E-Journals wird als technische Plattform Open Journal Systems (OJS; `https://pkp.sfu.ca/software/ojs/`; *Zuletzt aufgerufen am 17. August 2023.*) eingesetzt.

29 Go FAIR – FAIR Principles: `https://www.go-fair.org/fair-principles`; *Zuletzt aufgerufen am 17. August 2023.*

ten unterstützt. Zum anderen garantieren die DOIs ein präzises Referenzieren der Daten – auf mehreren Ebenen – sowie einen persistenten Zugriff. Auch Forschungsgruppen oder Projekte können z. B. auf ihren Websites oder Berichten direkt auf ihr Dataverse, also die Sammlungen der Datasets, oder via DOI auf die Datasets verweisen.

Durch die Katalogisierung der Forschungsdaten im Bibliotheksverbund K10Plus[30] wird sichergestellt, dass die Daten nicht nur national, sondern auch international z. B. im WorldCat[31]indexiert werden. Für die internationale kunsthistorische Forschung werden die Daten wiederum im Art Discovery Group Catalogue[32], basierend auf dem WorldCat, nachgewiesen. Darüber hinaus werden die Daten auch in weiteren Katalogen und Datenbanken indexiert, wie z. B. KVK[33], BASE[34] oder B2Find EUDAT[35]. Die Indexierung der Forschungsdaten in weiteren Datenbanken ist zudem über ein API-Harvest möglich.

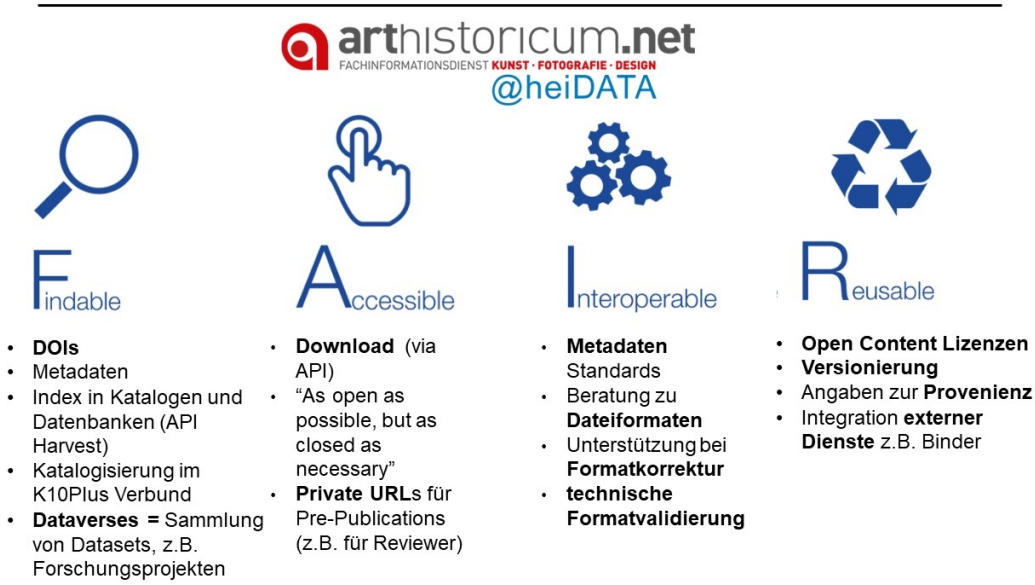


Abbildung 3: arthistoricum.net@heiDATA und die FAIR-Prinzipien.

Forschungsdaten auf arthistoricum.net@heiDATA sind frei zugänglich – Accessible – und können nicht nur über den Browser, sondern auch via API heruntergeladen werden. Für Fördermittelanträge haben Forschende die Möglichkeit, vor der Datenpublikation private URLs für Datasets zu generieren, um beispielsweise Gutachter:innen einen Einblick in das Vorhaben oder die noch unveröffentlichten Files zu geben.

30 K10Plus: `https://www.bszgbv.de/services/k10plus`; *Zuletzt aufgerufen am 17. August 2023.*
31 WorldCat: `https://www.worldcat.org`; *Zuletzt aufgerufen am 17. August 2023.*
32 Art Discovery Group Catalogue: `https://artlibraries.on.worldcat.org/discovery`; *Zuletzt aufgerufen am 17. August 2023.*
33 KVK (Karlsruher Virtueller Katalog): `https://kvk.bibliothek.kit.edu`; *Zuletzt aufgerufen am 17. August 2023.*
34 BASE (Bielefeld Academic Search Engine): `https://www.base-search.net`; *Zuletzt aufgerufen am 17. August 2023.*
35 B2Find EUDAT: `https://b2find.eudat.eu`; *Zuletzt aufgerufen am 17. August 2023.*

Die Interoperabilität der Daten wird auf Metadaten-Ebene durch die Unterstützung einschlägiger Metadatenstandards, wie z. B. DataCite[36], Dublin Core[37], DDI[38] und weitere, unterstützt. Im Rahmen von arthistoricum.net erhalten Datengebende Beratung zu passenden Dateiformaten sowie Unterstützung bei Dateiformatkorrektur oder der technischen Formatvalidierung.

Die Publikation im Forschungsdatenrepositorium erfolgt unter geeigneten Open-Content-Lizenzen. Für Daten werden die international anerkannten Creative-Commons-Lizenzen[39] empfohlen, insbesondere CC0 und CC-BY. Für Software gibt es eigene Softwarelizenzen, wie z.B. die GNU General Public License oder die GNU Lesser General Public License[40]. arthistoricum.net@heiDATA unterstützt ebenfalls mit weiteren Features die Reusability der Daten.

Die Versionierung ermöglicht einen transparenten Einblick in den aktuellsten Stand der Forschungsdaten sowie die Historie des Files innerhalb eines Datasets. So können neueste Forschungsergebnisse unkompliziert publiziert werden und der aktuellste Forschungsstand auch eindeutig referenziert werden. Auf File-Ebene können neben einem Abstract und Metadaten auch Angaben zur Provenienz der Files gemacht werden, um die Herkunft der Daten eindeutig zu dokumentieren. Mit ReadMe-Dateien können die Metadaten um weitere relevante Dokumentation ergänzt und die Datenstruktur einzelner Files veranschaulicht werden. Zudem können beispielsweise publizierte Jupyter-Notebooks[41] über den DOI direkt in Binder[42] ausgeführt werden. Einige dieser Funktionen und Möglichkeiten, welche publizierte Daten wiederverwendbar – Reusable – halten, stehen in der Kunstgeschichte noch ganz am Anfang und werden nur vereinzelt in der Community umgesetzt. Doch die Möglichkeiten sind gegeben und bieten Potenzial. Auch das Bewusstsein, kunstwissenschaftliche Daten gemäß der FAIR-Prinzipien zu publizieren, rückt zunehmend in den Vordergrund. arthistoricum.net@heiDATA wird als disziplinspezifisches Repositorium für Open Data auf FAIRsharing gelistet (FAIRsharing Team 2018).

Im Unterschied zu generischen Forschungsdatenrepositorien liegt bei dem fachspezifischen Repositorium arthistoricum.net@heiDATA ein Fokus darauf, die publizierten Daten in der entsprechenden Forschungsumgebung sichtbar und auffindbar zu machen. Hierfür ist auch die enge Verknüpfung mit den wissenschaftlichen Textbeiträgen von Bedeutung. Da der Fokus von arthistoricum.net@heiDATA naturgemäß auf kunstwissenschaftlichen Daten liegt und NFDI4Culture auf den gesamten Kulturbereich ausgerichtet ist, bietet in Ergänzung RADAR4Culture eine geeignete Publikationsplattform für Forschungsdaten aus der gesamten 4Culture-Community.

36 DataCite: `https://datacite.org`; *Zuletzt aufgerufen am 17. August 2023.*
37 Dublin Core: `https://www.dublincore.org`; *Zuletzt aufgerufen am 17. August 2023.*
38 DDI (Data Documentation Initiative): `https://ddialliance.org`; *Zuletzt aufgerufen am 17. August 2023.*
39 `https://creativecommons.org`; *Zuletzt aufgerufen am 17. August 2023.*
40 `https://www.gnu.org/licenses/licenses.html`; *Zuletzt aufgerufen am 17. August 2023.*
41 `https://jupyter.org`; *Zuletzt aufgerufen am 17. August 2023.*
42 `https://mybinder.org`; *Zuletzt aufgerufen am 17. August 2023.*

3 RADAR4Culture

Mit RADAR4Culture[43] bietet das FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur seit Frühjahr 2022 einen niedrigschwelligen, einfach zu nutzenden und kostenlosen Dienst zur nachhaltigen Veröffentlichung und Sicherung kulturwissenschaftlicher Forschungsdaten.

Das Angebot ist im Rahmen der Beteiligung von FIZ Karlsruhe als Mitantragsteller im Fachkonsortium NFDI4Culture entstanden. Es basiert auf dem bewährten Forschungsdatenrepositorium RADAR Cloud, ein auf die Bedürfnisse von Forschenden ausgelegter Dienst, der aus einem DFG-Projekt (2013–2016) hervorging (Brophy und Razum 2017). Die Repositorien-Software von RADAR wird von FIZ Karlsruhe entwickelt und auf einer sicheren Cloud-Plattform DSGVO-konform betrieben. Der Dienst wird vorrangig von akademischen Einrichtungen für das institutionelle Forschungsdatenmanagement eingesetzt. Seinem Anspruch an Nachhaltigkeit entsprechend, adressiert das Leibniz-Institut mit RADAR4Culture nun direkt die Forschenden in den Kulturwissenschaften.

Dem Aufbau des Dienstes gingen Analysen innerhalb des NFDI4Culture-Aufgabenbereichs Datenpublikation und Langzeitarchivierung (TA4) voraus. Diese stellten die besondere Bedeutung disziplinspezifischer Fachrepositorien zur Publikation und Langzeitarchivierung digitaler Forschungsdaten für die 4Culture-Community heraus. Sie zeigten jedoch gleichzeitig, dass noch nicht alle relevanten Anwendungsfälle im Bereich der Kulturwissenschaften abgedeckt werden und dass diesem zusätzlichen Bedarf durch ein generisch ausgerichtetes Datenrepositorium wie RADAR entsprochen werden kann. Dieser Ansatz steht im Einklang mit dem Bestreben der NFDI, möglichst vorhandene Infrastrukturen und Dienste zu nutzen, weiterzuentwickeln und diese interoperabel zu gestalten (Soltau und Razum 2021). In diesem Sinne zielt auch FIZ Karlsruhe darauf ab, RADAR4Culture zukünftig sukzessive an die spezifischen Bedarfe der Disziplinen des materiellen und immateriellen Kulturerbes anzupassen. RADAR4Culture ist ebenso wie RADAR Cloud in re3data gelistet (Re3data.Org 2022, 2017).

RADAR4Culture – Anwendungsbeispiele und Fakten
RADAR4Culture ist ein generisches und formatagnostisches Datenrepositorium, über das Forschungsdaten aus allen kulturwissenschaftlichen Disziplinen sicher und zuverlässig gemäß den FAIR-Prinzipien publiziert werden können (Abbildung 4). Das Angebot ergänzt das Spektrum bereits existierender Fachrepositorien im 4Culture-Bereich und eignet sich insbesondere für diejenigen Daten, die nicht den Spezifikationen anderer Repositorien entsprechen, z. B. Datensätze,

- für die noch kein passgenaues, fachspezifisches Repositorium verfügbar ist, oder
- die aufgrund ihrer Vielfalt an Datentypen in kein einzelnes etabliertes Repositorium passen, oder
- die interdisziplinärer Art sind.

43 `https://radar4culture.radar-service.eu`; *Zuletzt aufgerufen am 17. August 2023.*

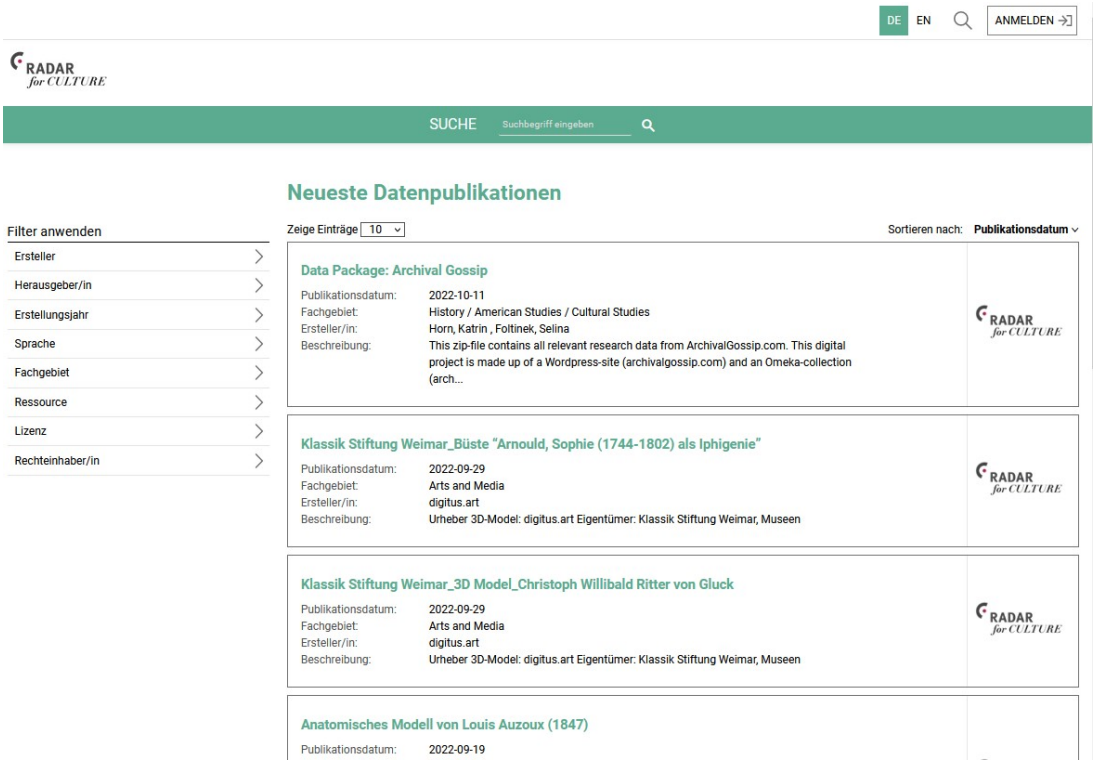


Abbildung 4: RADAR4Culture Portal.

Da kein Vertrag mit FIZ Karlsruhe geschlossen werden muss, ist RADAR4Culture äußerst niedrigschwellig zu nutzen und kann von Forschenden unabhängig von der institutionellen Zugehörigkeit in Anspruch genommen werden. Aktuell steht das Angebot allerdings ausschließlich Nutzenden in Deutschland zur Verfügung.

Für Forschende fallen in RADAR4Culture zudem weder Nutzungs- noch Publikationsgebühren an. Da FIZ Karlsruhe die Kosten für das Speicherkontingent und die Gebühren für die DOI-Registrierung der Forschungsdatensätze aus seinem NFDI4Culture-Förderbudget übernimmt, ist die Nutzung bis zu einem Speichervolumen von derzeit max.10 GB kostenfrei.

RADAR4Culture ist dabei als Self-Service ausgelegt: Datenupload, Metadaten-Annotation, Kuratierung, Publikation und die Verantwortung für die Preservation Policy übernehmen jeweils die Wissenschaftler:innen selbst.

RADAR4Culture – Kernfunktionen

Mit RADAR4Culture können Forschende die Daten ihrer Studien und Projekte über ein Webportal hochladen, zu Datenpaketen zusammenstellen, mit Metadaten beschreiben, begutachten lassen und dauerhaft öffentlich zugänglich machen.

RADAR4Culture erlaubt die Publikation aller Datentypen und -formate. Alle Forschungsdatensätze werden für mindestens 25 Jahre öffentlich verfügbar vorgehalten und über diesen Zeitraum hinweg georedundant in drei Kopien physikalisch erhalten. Jeder publizierte

Datensatz erhält einen DataCite-DOI und ist damit dauerhaft identifizierbar, referenzierbar und zitierbar. Falls notwendig, beispielsweise um Auflagen von Fachzeitschriften zu erfüllen, kann der DOI bereits vor der Publikation reserviert oder eine Datenpublikation optional mit einer Embargofrist von ein bis 12 Monaten verzögert werden. Im letzteren Fall wird das Datenpaket erst mit Ablauf der Sperrfrist zugänglich gemacht, während beschreibende Metadaten bereits unmittelbar nach der Publikation öffentlich sichtbar sind. Falls ein zeitlich begrenztes Embargo im Einzelfall zu kurz greift, kann die Publikationsoption mit „unbegrenztem Embargo“ gewählt werden. Hier verbleiben die eigentlichen Forschungsdaten dauerhaft für die Öffentlichkeit unzugänglich, können jedoch über eine Anfrage- bzw. Freigabeoption individuell mit anderen RADAR-Nutzer:innen geteilt werden. RADAR4Culture unterstützt zudem einen Review-Prozess vor der Datenpublikation. Sobald das Datenpaket in den Status „in Begutachtung“ gesetzt wird, ist es nicht weiter bearbeitbar und kann über einen gleichzeitig erzeugten, sicheren Link für externe Gutachter:innen freigegeben werden.

Die Metadaten-Annotation kann bequem über einen formularbasierten Editor oder per Upload einer XML-Datei durchgeführt werden. Zusätzlich steht eine REST-basierte RADAR-API zur Verfügung, die den vollständigen Funktionsumfang des RADAR4Culture-Fronten-ds abbildet und so zum Beispiel für die automatisierte Übergabe von Metadaten verwendet werden kann. Die Beschreibung mit Metadaten kann nicht nur für das RADAR-Datenpaket (hier: verpflichtend) erfolgen, sondern optional auch für darin enthaltene Einzeldateien und Verzeichnisse.

Das RADAR-Metadatenschema[44] basiert auf dem DataCite-Metadatenschema[45], einem weitverbreiteten und disziplinunabhängigen Standard zur Beschreibung von Datensätzen und ist kompatibel mit Dublin Core. Es ist disziplinagnostisch und als Standard auf FAIRsharing gelistet.(FAIRsharing Team 2023) Das Schema enthält zehn Pflichtfelder, unter anderem die Grundanforderungen für die DOI-Registrierung in Übereinstimmung mit dem DataCite-Schema. Daneben stehen zusätzlich 13 optionale Parameter zur Verfügung. Das RADAR-Metadatenschema bietet eine Kombination von Freitextfeldern, kontrollierten Listen und Auswahloptionen für standardisierte bzw. normierte Einträge. Letztere sind über Schnittstellen als Vorschlags- bzw. Auswahllisten in die Benutzeroberfläche integriert: ORCID für Personenangaben, ROR[46] für Institutionsangaben, Crossref Funder Registry[47] für Angaben zur Förderorganisation und – seit kurzem und vorangetrieben insbesondere aufgrund des Bedarfs der 4Culture-Community – die GND im Schlagwort-Feld. Vor jeder Publikation ist verpflichtend eine Lizenz zu wählen, welche die Nachnutzungsrechte am Datensatz definiert. Hierfür stehen nicht nur gängige Lizenztypen für Forschungsdaten, wie z. B. Creative Commons[48] Lizenzen, sondern auch Lizenzen für Forschungssoftware zur Verfügung. RADAR4Culture unterstützt außerdem die Relatierung von Forschungsdaten mit verwandten digitalen Ressourcen, beispielsweise Zeitschriftenartikeln oder Buch-

44 `https://radar.products.fiz-karlsruhe.de/de/radarfeatures/radar-metadatenschema`; *Zuletzt aufgerufen am 17. August 2023.*
45 `https://schema.datacite.org`; *Zuletzt aufgerufen am 17. August 2023.*
46 ROR (Research Organization Registry): `https://ror.org`; *Zuletzt aufgerufen am 17. August 2023.*
47 `https://www.crossref.org/services/funder-registry`; *Zuletzt aufgerufen am 17. August 2023.*
48 `https://creativecommons.org`; *Zuletzt aufgerufen am 17. August 2023.*

publikationen, über das Metadatenfeld „Verwandter Identifikator“. Diese Verknüpfungen zu verwandten Ressourcen sind ebenso wie normierte und standardisierte Daten auf der Landingpage eines Datensatzes als Links integriert und zusätzlich in den Metadaten als persistente Identifikatoren nachhaltig und maschinenlesbar gespeichert.

Speziell in einem fachspezifischen Kontext wie dem des kulturellen Erbes läuft der etablierte disziplinagnostische Annotationsansatz jedoch Gefahr, zu kurz zu greifen. RADAR4Culture wird dem Bedarf nach Flexibilisierung des generischen RADAR-Metadatenschemas bereits gerecht und unterstützt das Hinterlegen von fachspezifischen Schemata und den Upload disziplinspezifischer Metadaten als XML-Dateien. Aktuell erarbeitet FIZ Karlsruhe eine Möglichkeit, Nutzer:innen fachspezifische und individuelle Metadatenschemata einfach und bequem erstellen und ihre Forschungsdaten über eine nutzerfreundliche Eingabemaske annotieren zu lassen.

Nach jeder Publikation werden die deskriptiven Metadaten in verschiedenen Formaten indexiert und öffentlich über DataCite sowie zusätzlich über den OAI-Provider von FIZ Karlsruhe[49] zum Harvesting angeboten, so dass Dissemination und Auffindbarkeit eines jeden RADAR4Culture-Datensatzes sichergestellt werden. Alle RADAR4Culture-Datensätze können darüber hinaus von Dritten als komplettes Set geharvestet[50] und so beispielweise auf fachspezifischen Portalen integriert werden.

RADAR4Culture und die FAIR-Prinzipien

In den FAIR-Prinzipien werden Kriterien definiert, um Forschungsdaten auffindbar, zugänglich, interoperabel und nachnutzbar zu machen. Eine wachsende Anzahl an wissenschaftspolitischen Akteuren, z.B. im Bereich der Forschungsförderung, unterstützen die Forderung nach FAIR Data. Ziel ist es, Forschungsdaten so aufzubereiten und zugänglich zu machen, dass sie für Menschen und Maschinen optimal nutzbar sind und existierende Datenbestände – sofern technische und rechtliche Rahmenbedingungen es zulassen – für neue Forschungsfragen wiederverwendet werden können. RADAR4Culture unterstützt, wie das zugrundeliegende Repositorium RADAR, die FAIR-Prinzipien mit verschiedenen Maßnahmen und Dienstmerkmalen. FIZ Karlsruhe arbeitet kontinuierlich an der Optimierung der FAIRness seiner Dienstangebote. Abbildung 5 zeigt eine Zusammenschau der Umsetzung der FAIR-Prinzipien bei RADAR.

RADAR4Culture Workflow

Forschende aus dem Bereich der Kulturwissenschaften, z. B. an öffentlich geförderten Forschungseinrichtungen, (Kunst-)Hochschulen, nicht-kommerziellen Akademien, Galerien, Bibliotheken, Archiven und Museen in Deutschland können sich bei Interesse am RADAR4Culture-Publikationsangebot an FIZ Karlsruhe wenden. FIZ Karlsruhe berät Forschende individuell und richtet bei Passung des Datensatzes für den Anwendungsbe-

49 `https://radar.products.fiz-karlsruhe.de/de/radarfeatures/radar-oai-provider`; *Zuletzt aufgerufen am 17. August 2023.*

50 Harvesting von RADAR4Culture Metadaten als Set in den drei Formaten (RADAR / DataCite / Dublin Core): `https://radar.products.fiz-karlsruhe.de/de/nachricht/radar4chem-und-radar4culture-direkt-im-zugriff`; *Zuletzt aufgerufen am 17. August 2023.*

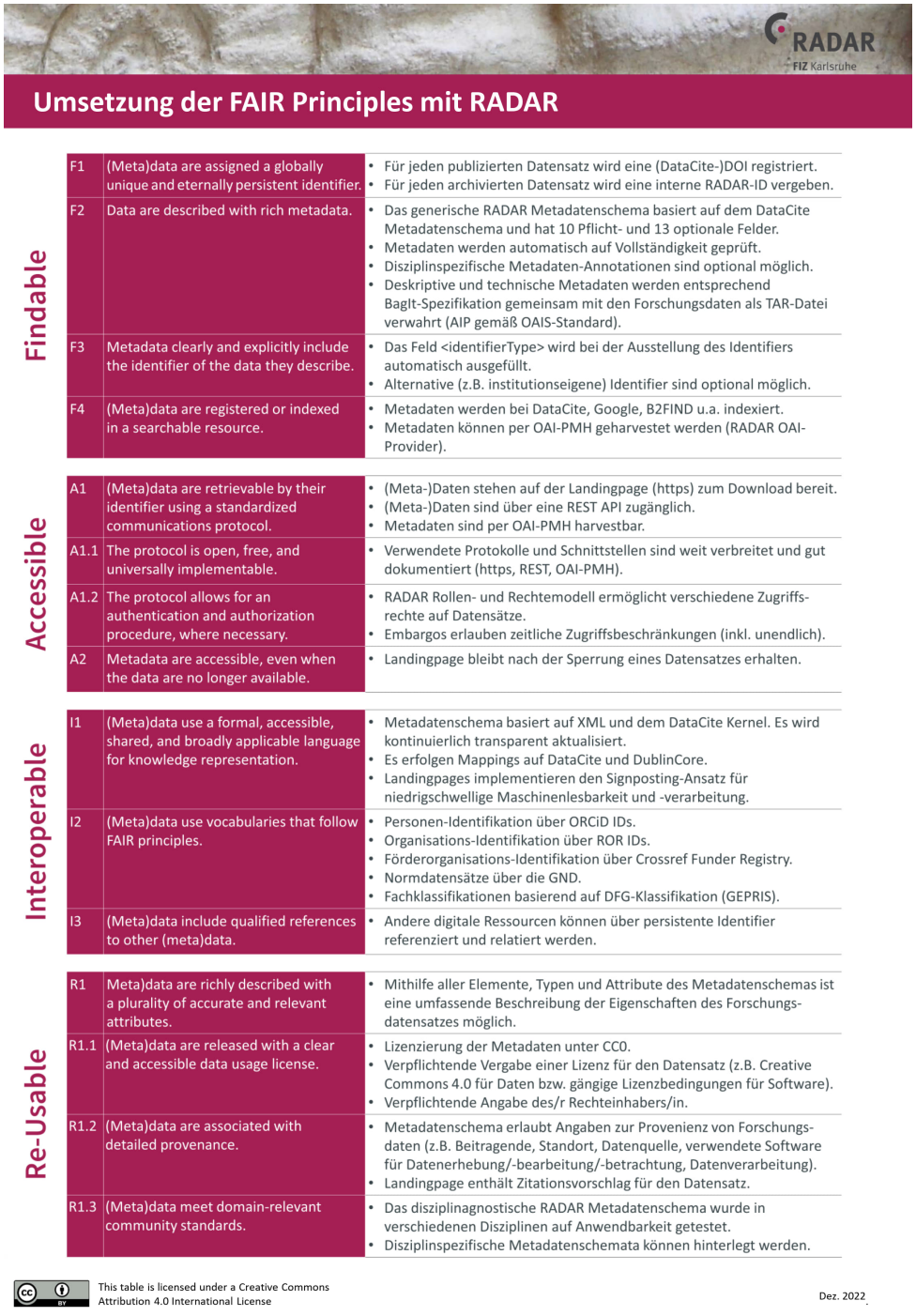


Abbildung 5: Umsetzung der FAIR-Prinzipien mit RADAR.

reich von RADAR4Culture einen dezidierten Arbeitsbereich ein, in dem die Daten für die Publikation aufbereitet werden können.

Ein Quickstart-Guide für Datengeberinnen und Datengeber (Soltau und Goeller 2023) sowie eine Handreichung zu personenbezogenen Daten (Soltau 2023) bieten den Forschen-

den einen schnellen Einblick in den Workflow von RADAR4Culture. Sobald ein Datensatz publikationsbereit ist, kann er durch die Forschenden selbst publiziert werden. Derzeit ist das maximale Speichervolumen auf 10 GB pro Forschungsprojekt begrenzt. Vor der Datenpublikation müssen Datengebende online den Lizenz- und Nutzungshinweisen von RADAR4Culture[51] zustimmen. Mit der Publikation erhält jeder Datensatz einen DOI und ist somit auch gleich zitierbar.

4 Zusammenfassung

Die vorgestellten Repositorien sind Teil des Angebotes von NFDI4Culture. Darüber hinaus werden Wissenschaftler:innen durch das 4Culture-Helpdesk[52] zu allen Fragen rund um das Thema Forschungsdaten beraten. Ergänzend dazu wurde eine FAIR-Clearing-Stelle[53] eingerichtet, die Forschende vor allem in der Planungsphase eines Projektes, bei der Auswahl und Umsetzung entsprechender Standards und Strategien sowie bei der Entwicklung von Datenmanagementplänen unterstützt. Weitere Dienste und Services werden künftig in der Registry für Forschungswerkzeuge und Datendienste[54] verzeichnet sein. Die Registry ist ein über das NFDI4Culture-Portal zugänglicher Service, in dem Metadaten zu bestehenden Forschungswerkzeugen und Datendiensten, die speziell für die Kulturwissenschaften geeignet sind, zu finden sind. Die vorgestellten Repositorien und die Angebote zur Beratung sollen die NFDI4Culture-Community zukünftig befähigen, die für die Forschung so wichtigen Datenressourcen gemäß der FAIR-Prinzipien langfristig zu sichern.

Danksagung

Mit besonderem Dank für die Unterstützung durch Maria Effinger, Jochen Apel und Sabrina Herzog. NFDI4Culture wird durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Projekt-ID 441958017 gefördert.

ORCID:

- Alexandra Büttner https://orcid.org/0000-0002-4950-0941
- Sandra Göller https://orcid.org/0000-0003-4553-3671
- Peggy Große https://orcid.org/0000-0003-1181-6219
- Kerstin Soltau https://orcid.org/0000-0002-6368-1929

51 `https://radar.products.fiz-karlsruhe.de/sites/default/files/radar/docs/terms/Lizenz_und_Nutzungshinweise_fuer_Datengeber_RADAR4Culture.pdf`; *Zuletzt aufgerufen am 17. August 2023.*

52 `https://nfdi4culture.de/services/details/culture-helpdesk.html`; *Zuletzt aufgerufen am 17. August 2023.*

53 `https://nfdi4culture.de/de/ueber-uns/aufgabenbereiche/aufgabenbereich-2.html`; *Zuletzt aufgerufen am 17. August 2023.*

54 `https://nfdi4culture.de/de/ressourcen/registry.html`; *Zuletzt aufgerufen am 17. August 2023.*

Literaturverzeichnis

Altenhöner, Reinhard, Ina Blümel, Franziska Boehm, Jens Bove, Katrin Bicher, Christian Bracht, Ortrun Brand u. a. 2020. „NFDI4Culture - Consortium for research data on material and immaterial cultural heritage“. *Research Ideas and Outcomes* 6. DOI: https://doi.org/10.3897/rio.6.e57036.

Apel, Jochen, Fabian Gebhart, Leonhard Maylein und Martin Wlotzka. 2018. „Offene Forschungsdaten an der Universität Heidelberg: von generischen institutionellen Repositorien zu fach- und projektspezifischen Diensten“. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB.* DOI: https://doi.org/10.5282/O-BIB/2018H2S61-71.

Brophy, Ena, und Matthias Razum. 2017. „RADAR: A Research Data Management Repository for Long Tail Data“. In *E-Science-Tage 2017: Forschungsdaten managen.* heiBOOKS. DOI: https://doi.org/10.11588/HEIBOOKS.285.C3874.

Crosas, Mercè. 2019. „The FAIR Guiding Principles: Implementation in Dataverse“. Besucht am 15. Mai 2023. https://scholar.harvard.edu/sites/scholar.harvard.edu/files/mercecrosas/files/fairdata-dataverse-mercecrosas.pdf.

FAIRsharing Team. 2018. *FAIRsharing record for: arthistoricum.net@heiDATA.* DOI: https://doi.org/10.25504/FAIRSHARING.VBXAEP.

———. 2023. *FAIRsharing record for: RADAR Metadata Schema.* DOI: https://doi.org/10.25504/FAIRsharing.e26f92.

NFDI4Culture. 2023. „NFDI4Culture – Data Lifecylce und Task Areas“. Besucht am 15. Mai 2023. https://nfdi4culture.de/about-us/task-areas.html.

Re3data.Org. 2017. *RADAR.* DOI: https://doi.org/10.17616/R3ZX96.

———. 2021. *arthistoricum.net@heiDATA.* DOI: https://doi.org/10.17616/R31NJMWV.

———. 2022. *RADAR4Culture.* DOI: https://doi.org/10.17616/R31NJNAZ.

Soltau, Kerstin. 2023. *RADAR4Culture: Handreichung zu personenbezogenen Daten.* Technischer Bericht. NFDI4Culture. DOI: https://doi.org/10.5281/zenodo.8221495.

Soltau, Kerstin, und Sandra Goeller. 2023. *RADAR4Culture: Quickstart-Guide für Datengeberinnen und Datengeber [deutsch].* Zenodo. DOI: https://doi.org/10.5281/zenodo.8221340.

Soltau, Kerstin, und Matthias Razum. 2021. „Veränderung als Konstante: RADAR etabliert sich als flexibler Baustein im Forschungsdatenmanagement“. *b.i.t. online Heft* 24 (2): 152–162.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Herausforderungen beim Aufbau eines föderierten Datenrepositoriums auf Basis von InvenioRDM

Dirk von Suchodoletz [1], Jonathan Bauer [1], Marcel Tschöpe[1], Holger Gauza [2], Michael Derntl[3], Steve Kaminski[3]

[1]Universität Freiburg, Rechenzentrum;
[2]Universität Tübingen, Zentrum für Datenverarbeitung;
[3]Universität Tübingen, Digital Humanities Center

Forschungsdaten sind Produkte und Rohstoffe von und für Forschung gleichermaßen. Deren Veröffentlichung zeugt nicht nur von Forschungsleistung, sondern auch von guter wissenschaftlicher Praxis und ist im Sinne von Open Data und Open Science. Umso wichtiger sind der Aufbau und die dauerhafte Einrichtung von Datenpublikationsrepositorien, die Forschende bei der Veröffentlichung niederschwellig unterstützen, Workflows zur Qualitätssicherung beinhalten und die Auffind- und Zitierbarkeit von Forschungsdaten realisieren. Dies bedingt eine enge Integration in die jeweiligen Prozesse der Fach-Communities, um Doppelarbeiten und -eingaben seitens der Forschenden zu vermeiden. Forschungsleistungen werden langfristig identifizierbar und transparent mit dem Scholarly Record der einzelnen Beteiligten verknüpft. Zu diesen Zwecken wird an den Universitäten Tübingen und Freiburg die Plattform InvenioRDM eingesetzt. Um Daten dauerhaft und georedundant gesichert vorzuhalten, wird auf die für wissenschaftliche Daten ausgelegten, föderierten Speichersysteme von bwSFS aufgebaut. Organisatorisch vernetzt wird das Datenpublikationsrepositorium mit den Aktivitäten im Rahmen der Nationalen Forschungsdateninfrastruktur DataPLANT und des Science Data Centers BioDATEN des Landes Baden-Württemberg.

1 Einleitung

An Universitäten bildet Forschung eine zentrale Säule des institutionellen Selbstverständnisses und des gesetzlichen Auftrags. Dabei ist die Digitalisierung der Forschungs- und Arbeitsprozesse ist in allen Wissenschaftsdisziplinen allgegenwärtig. Digitale Werkzeuge und Arbeitsabläufe gehören für die meisten Forschenden mittlerweile zum Standard ihrer Forschung. Sie benötigen hierfür Forschungsinfrastrukturen, die zunehmend auf IT setzen und dabei Anforderungen von Forschungsförderern berücksichtigen. Forschungsdaten

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.c18069

sind Produkte und Rohstoffe von und für Forschung gleichermaßen und deren Veröffentlichung zeugt nicht nur von Forschungsleistung, sondern auch vom verantwortungsvollen Umgang mit Daten (Deutsche Forschungsgemeinschaft e.V. 2013; Suchodoletz u. a. 2021). Forschungsdaten bilden nicht nur „Beiwerk“ einer Publikation, sondern erhöhen die wissenschaftlichen Anerkennung und Reputation der Forschenden. Eine Publikation von Forschungsdaten unterstreicht das Bekenntnis zur Selbstverpflichtung der Wissenschaft zu Open Data und Open Science. Das Forschungsdatenmanagement (FDM), die nachhaltige und zukunftsorientierte Organisation von Forschungsdaten, ihre Bereitstellung und optimalerweiser Publikation, rückt damit in den Fokus zeitgemäßer Forschungsunterstützung.

Umso wichtiger ist der Aufbau und der dauerhafte Betrieb von Repositorien, welche die Forschenden bei der Veröffentlichung niedrigschwellig unterstützen, Workflows zur Qualitätssicherung beinhalten und die Auffind- und Zitierbarkeit von Forschungsdaten realisieren. Damit wird die allgemeine Bereitstellung von Forschungsdaten zur Nachnutzung wesentlicher Bestandteil des Lebenszyklus von Forschungsdaten und überhaupt erst ermöglicht. Eine zitierbare Veröffentlichung der Daten und die Verknüpfung dieser mit den involvierten Forschenden belegt transparentes Handeln und ordnet die Forschungsleistung vergleichbar zu traditionellen Publikationen den involvierten Personen und Einrichtungen zu. Forschung wird reproduzier- und überprüfbar. Die Entwicklung wird durch die Erwartungen einer steigenden Anzahl von Forschungsförderern beschleunigt, die bereits in der Beantragung die Dokumentation und Planung eines umsichtigen und nachhaltigen Umgangs mit Forschungsdaten wünschen oder voraussetzen (Leendertse, Mocken und Suchodoletz 2019). Diese Erwartungen und Vorgaben finden sich unter anderem in den Open-Access-Policies von Hochschulen und Forschungsinstituten sowie in den Leitlinien zur Sicherung guter wissenschaftlicher Praxis der DFG (Deutsche Forschungsgemeinschaft e.V. 2013), den FAIR-Prinzipien (Wilkinson u. a. 2016) als auch in vielen Data Policies von wissenschaftlichen Zeitschriften und Verlagen.

Die Herausforderungen für die Bereitstellung eines Datenrepositoriums gehen über die bloße Auswahl einer technischen Basis unter Berücksichtigung des jeweils lokalen Systemkontextes hinaus. Vielmehr müssen organisatorische und rechtliche Fragen geklärt und die notwendigen Voraussetzungen geschaffen werden. Gleichzeitig sind zur Erfüllung der Aufbewahrungsfristen der Fördergeber tragfähige und nachhaltige Betriebskonzepte zu erarbeiten. Primär befördert wurden die Auswahl und Bereitstellung eines Datenrepositoriums durch die Fach-Community-Projekte BioDATEN, MoMaF, BERD@BW und SDC4Lit im Rahmen der Science Data Center (SDC) Initiative des Landes Baden-Würtemberg, ausgeführt in (Axtmann u. a. 2021), und später DataPLANT, welches eines der geförderten Konsortien in der Nationalen Forschungsdateninfrastruktur ist.

Die nutzbringende Veröffentlichung von Forschungsdaten erfordert die Annotation mit einschlägigen wissenschaftlichen Metadaten, die über die Anforderungen des DataCite-Schemas zur DOI-Registrierung hinausgehen. Solche Schemata werden in Kooperation mit den Forschenden erarbeitet und sollten mit den Daten leicht abrufbar bereit liegen. Der Einsatz von InvenioRDM bietet erhebliches Potential als Ergänzung von etablierten Publikationssystemen. Die weitergehende Integration in Daten-Workflows – auch in über-

greifenden Kooperationen und bei international agierenden Forschungs-Communities – ist kein Selbstläufer, und gerade organisatorische Aspekte sind nicht zu vernachlässigen.

In diesem Beitrag werden aus standortübergreifender Betreiberperspektive zentrale Herausforderungen und Lösungsansätze des Einsatzes und der Integration von InvenioRDM dargelegt. Das beinhaltet die Erarbeitung eines Anforderungskatalogs für Repositorien sowie auf dessen Basis die Auswahl einer technischen Plattform zur Umsetzung (Abschnitt 3). Hierfür erfolgte eine enge Koordination und Kooperation aller beteiligter Akteure auf den verschiedenen Ebenen. Das betrifft Kontakte zur Entwickler-Community ebenso wie die Abstimmung mit den Infrastrukturbetreibern für das bwSFS (Storage-for-Science; Suchodoletz, Hahn, Bauer u. a. 2022) und den Einrichtungen, die das organisatorische Gerüst für die Nutzerauthentifizierung und die DOI-Schnittstelle bereitstellen (Abschnitt 2). Die Verpflichtung zu einer langfristigen Verfügbarkeit von Veröffentlichungen wirkt sich ebenfalls auf den technischen Aufbau der Repositorien aus und benötigt ein zukunftssicheres Betriebskonzept (Abschnitte 4 und 5). Erste wichtige Schritte auf diesem Weg sind im SDC BioDATEN in enger Abstimmung mit den Beteiligten aus Bibliotheken, Rechenzentren und Anwendenden erfolgt.

2 Organisatorische Grundlagen

Forschungsdatenmanagement ist keine rein technische Aufgabe, sondern erfordert den Einbezug und die Abstimmung mit den wesentlichen Stakeholdern als Akteure an der Universität. Dazu zählen die Forschenden als Datenproduzierende, Rechenzentren und Bibliotheken als Datenmanager und die Universitätsleitung als oberstes Organ und Dienstherr. In diesem Rahmen werden Anforderungen an FDM ausgehandelt und umgesetzt. Hierbei sind insbesondere Aspekte wie Nachhaltigkeit und Recht sowie sich ergebende Anforderungen und eventuelle neue Aufgaben zu thematisieren. Zentral ist ebenfalls die Aushandlung zwischen der Forderung nach Open Data und dem Schutz sensibler Daten.

Nachhaltigkeit Die Frage nach einem nachhaltigen Umgang mit Forschungsdaten aus Betreibersicht hat mindestens zwei Dimensionen: Verfügbarkeit und Kosten. Die Zusage und Sicherstellung einer langfristigen Verfügbarkeit von Forschungsdaten resultiert zwangsläufig in Aufwendungen, die gegebenenfalls über die Projektlaufzeit hinaus gehen und trotzdem geplant werden müssen (Leendertse und Suchodoletz 2020). Relevante Kostenfaktoren sind dabei das Mengengerüst, der jeweils notwendige Aufwand und die Betreuung einer Datenpublikationsplattform. Da viele Forschungs-Communities diese Fragen im Rahmen der NFDI angehen, sind hier weitere Erkenntnisse zu erwarten. Gleichzeitig muss die Verfügbarkeit von Forschungsdaten bei technischem oder wirtschaftlichem Ausfall des Repositoriums mitgedacht werden. InvenioRDM unterstützt Betreiber hinsichtlich eines möglichst wirtschaftlichen Umgangs mit Ressourcen durch zwei zentrale Aspekte: Die Anbindung an das Speichersystem bwSFS und somit eine (geo-)redundate Speicherung von Daten mittels S3 sowie den Aufbau von Communities[1] innerhalb von InvenioRDM,

1 Communities fungieren als Mandanten, die eigene Workflows und Policies definieren können. Gleichzeitig können Communities die Sichtbarkeit von Datensätzen beschränken.

um den Betreuungsaufwand zu zentralisieren und Doppelarbeiten zu vermeiden. Der Anspruch an eine dauerhafte Verfügbarkeit erfordert zwingend ein überprüftes Konzept für ein *Disaster Recovery* und nach Möglichkeit alternative Betriebszenarien bei dauerhafter Nichtverfügbarkeit der Datenpublikationsplattform. Für diesen Fall kann eine leichtgewichtige Alternative eingerichtet werden, die ausgehend von einer für jede Publikation generierten statischen Landingpage die S3-Datenobjekte direkt referenziert und ohne InvenioRDM bereitstellt.

Forschungsdatenpolicies, Datenüberlassungsverträge und Lizenzen Die Nachnutzbarkeit von Forschungsdaten ist nur mit Veröffentlichung unter einer möglichst offenen Lizenz gegeben. Entsprechend empfiehlt der Arbeitskreis Forschungsdatenmanagement in Baden-Württemberg (AK FDM) die Bereitstellung von Daten unter CC-BY und die Bereitstellung von Metadaten unter CC0 (Brettschneider u. a. 2021). Um diese Empfehlung zu befördern, wurden diese Lizenzen als Standardwert im Publikationsprozess hinterlegt. Dieser Prozess muss außerdem einen Datenüberlassungsvertrag beinhalten, der sowohl den Betreibern als auch den Forschenden praktikable Rechte einräumt. Diese Verträge leiten sich am besten von vorgelagerten Forschungsdatenpolicies der jeweiligen Institutionen ab. Die Universität Freiburg hat hierzu ihre Forschungsdatenpolicy überarbeitet und die Verantwortlichkeiten der Forschenden sowie der Universität definiert. An dieser Stelle wurde gleichzeitig die Verwendung der ORCID für Personen und DOIs für Daten festgehalten (Albert-Ludwigs-Universität Freiburg, Rektorat 2022). Die nachhaltige Nutzung von Forschungsdaten beruht nicht nur auf technischen Betrieb einer Plattform, sondern erfordert Einsatz und Abstimmung mit den beteiligten Einrichtungen einschließlich der höchsten Leitungsebene und der akademischen Gremien der Universitäten.

ORCID Die Nutzung der persönlichen ORCID-iD erlaubt die Identifizierung von Forschenden und die automatische Anreicherung der jeweiligen Scholarly Records um die veröffentlichten (Daten-)Publikationen. Forschende haben weitgehende Rechte bei der Freigabe ihrer Daten. Bisher wird die Authentizität der ORCID-Halter nicht verifiziert. Hier wäre es in Zukunft an den lokalen Identity-Providern der Einrichtungen, dafür zu sorgen, dass nur geprüfte ORCID-iDs beim Login übergeben werden oder eine Liste mit überprüften ORCID-iDs erstellt wird. Auf diese Weise lässt sich das Potential eines breiten Einsatzes von ORCID erschließen und gleichzeitig Missbrauch vermeiden.

Schutz sensibler Daten Insbesondere öffentliche Forschungsförderer setzen aus Gründen der Nachnutzung und Forschungstransparenz zunehmend auf offene Wissenschaft (Open Science und Open Scholarship).[2] Sie sind bestrebt, dieses Vorgehen als Standard zu etablieren. Hierzu zählt eine weitreichende Verfügbarkeit der entstehenden Forschungsdaten. Die Forschenden sollten im Sinne von Open Access verpflichtet werden, ihre Daten nach einer gewissen Zeit und in Abhängigkeit bestimmter Parameter (welche vom Pro-

2 Die Universitäten reagieren auf diese Anforderungen, indem sie entsprechende Unterstützungsangebote entwickeln und ihre eigenen Policies anpassen (Albert-Ludwigs-Universität Freiburg, Rektorat 2022). Siehe zudem `https://www.ub.uni-freiburg.de/unterstuetzung/elektronisch-publizieren` und die Open Access Resolution der Universität Freiburg `https://www.ub.uni-freiburg.de/unterstuetzung/elektronisch-publizieren/open-access/open-access-resolution-der-universitaet`.

jekt, den Fördergebern, und der Policy der Community abhängen können) bereitzustellen. Dieses Ziel wird klar von der Einsicht bestimmt, dass sich offene Wissenschaft nicht pauschal erzwingen lässt. Insbesondere können Gründe vorliegen, die in bestimmten Fällen eine Einschränkung der Zugänglichkeit nahelegen. So ist eine gewisse Zurückhaltung zu akzeptieren, wenn in bestimmten Forschungsfeldern eine breite Zugänglichkeit der Daten die Gefährdung des Forschungsgegenstands bedeutet. Der Einsatz beispielsweise von Sperrfristen (*Embargo*) oder weiteren Einschränkungen muss mit den betroffenen Forschenden ausgehandelt werden.[3]

3 Technische Grundlagen

Anforderungen an ein Repositorium Mit Blick auf den Lebenszyklus von Forschungsdaten kommen Repositorien an dessen Ende zum Einsatz, um die Grundlage für eine Nachnutzung und Referenzierbarkeit der Daten im Sinne der FAIR-Prinzipien (Wilkinson u. a. 2016) zu schaffen und die Anforderungen von Fördergebern zu erfüllen. Gerade in den Lebenswissenschaften und mit Blick auf die heterogene Community des SDC BioDATEN wurden folgende Kriterien erarbeitet:

- Umgang mit großen Datenpaketen, die nicht über klassische Repositorien für textuelle Ressourcen abgedeckt werden
- Anbindung an einen Registrar für persistente Identifier wie DataCite für eine referenzierbare Datenpublikation sowie eine niederschwellige Unterstützung der Forschenden und Beitrag zur Datenqualität
- Nachhaltige Perspektive im Sinne der Weiterentwicklung und technologische Anschlussfähigkeit sowohl zu Speichertechnologien als auch zur bestehenden Systemlandschaft
- Flexibilität hinsichtlich annotierbarer Metadaten und Community-Anforderungen für eine verbesserte Auffindbarkeit von Forschungsdaten und Arbeitsunterstützung

Diese und weitere Kriterien wurden gemeinsam mit den anderen SDCs BERD@BW, SDC4Lit und MoMaF in einen Anforderungskatalog zusammengeführt (Axtmann u. a. 2021).

Entscheidung für InvenioRDM Nach Prüfung mehrerer Optionen haben sich BioDATEN und DataPLANT für den Einsatz von InvenioRDM entschieden, wobei mehrere Aspekte ausschlaggebend waren. Die konsequente Open-Source-Entwicklung von InvenioRDM wird von einer aktiven großen internationalen Community aus Universitäten unter der Leitung des CERNs getragen. Deshalb wird eine gute Perspektive für eine langfristige (Weiter-)Entwicklung erwartet. Die Universitäten Freiburg und Tübingen beteiligen sich aktiv an der Entwicklung, sind offizielle Entwicklungspartner und entsprechend Teil der Entwickler-Community. Das Invenio-Framework hat seine Leistungsfähigkeit jenseits

3 Während einzelne Personen oder Gruppen durch medizinische, sozialwissenschaftliche oder psychologische Studien offenbart werden könnten, sind auf anderen Gebieten beispielsweise seltene Spezies oder wertvolle Höhlenmalereien durch Geolokalisierung gefährdet.

eines Proof-of-Concepts bereits mit Zenodo unter Beweis gestellt.[4] Technische Kernaspekte liegen in der Benutzerfreundlichkeit der Weboberfläche, Erweiterbarkeit hinsichtlich der Integration von Metadatenschemata und Vokabularen zur Annotation der Forschungsdaten sowie der Flexibilität in Bezug auf verwendbare Speichertechnologien. Für letzteres ist die Unterstützung von Object Storage via S3 zukunftsfähig und skalierbar. S3 wird föderiert durch den von der DFG und vom Land Baden-Württemberg finanzierten Dienst bwSFS (Suchodoletz u. a. 2019; Suchodoletz, Hahn, Bauer u. a. 2022) bereitgestellt.

Die Anbindung mehrerer *Authentication and Authorization Infrastructures* (AAI) reduziert auf Seite der Betreiber den notwendigen Aufwand für die Pflege einer eigenen Benutzerverwaltung und erlaubt es den Nutzenden, bereits vorhandene Zugangsdaten, beispielsweise ihrer Heimateinrichtung, oder ORCID zu verwenden. Der Einsatz von Schnittstellen zu DataCite und ORCID erlaubt die Vergabe von DOIs und mittels *Auto-Profile Update* von DataCite kann die Übertragung in das ORCID-Profil automatisiert werden. Auf diese Weise werden die Empfehlungen der ORCID-Integration umgesetzt (Suchodoletz u. a. 2020). Die Mandantenfähigkeit von InvenioRDM ermöglicht den Aufbau von Communities samt Integration von Workflows zur Qualitätssicherung im Peer-Review-Verfahren.

Forschung findet weltweit vernetzt in unterschiedlichsten Kooperations- und Interaktionsbeziehungen statt. Ergebnisse, die an anderen Einrichtungen produziert werden, bilden die Fragestellung für die eigenen Forschenden und umgekehrt. Diesem verteilten Charakter muss ein Datenpublikationssystem Rechnung tragen. Die eScience-Strategie des Landes Baden-Württemberg fordert deshalb unter anderem die Such- und Auffindbarkeit von Forschungsdaten über mehrere Repositorien hinweg, welche die Grenzen der einzelnen Einrichtung und Fachdisziplinen überwindet[5]. Für solche übergreifenden Initiativen zu Suchportalen existieren bereits verschiedene Ansätze, wie beispielsweise re3data.org[6]. Hierfür hat sich der gemeinsame Standard OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) mit einem technisch sehr einfachen Verfahren zwischen einem oder mehreren Data- und Service-Providern etabliert. Über die OAI-PMH- und REST-Schnittstellen von InvenioRDM können andere Forschungsinformationssysteme die publizierten Datensätze systematisch und automatisch sammeln und referenzieren. Durch Abfragen (*harvesting*) werden die Daten zusammengetragen und zu einem konsolidierten Suchindex aggregiert. Die Metadaten sind in ihrer Struktur nicht durch OAI-PMH spezifiziert, so dass beispielsweise verschiedene disziplinspezifische Datenformate angeboten werden können. Für ein Mindestmaß an Interoperabilität sollte jeder Daten-Provider sinnvollerweise Publikationsmetadaten z.B. nach Dublin Core (DC) unterstützen. InvenioRDM implementiert daher die Bereitstellung solcher Metadaten über OAI-PMH.

4 Zenodo (`https://zenodo.org`) ist eine etablierte Datenpublikationsplattform des CERN.

5 Siehe hierzu `https://mwk.baden-wuerttemberg.de/de/forschung/forschungslandschaft/e-science`.

6 Zentraler Repository-Aggregator: `https://www.re3data.org`.

4 Aufbau und Betrieb institutioneller Repositorien

Wissenschafts-Communities betreiben FDM jeweils angepasst an die Erfordernisse ihrer Forschungsdisziplin auf unterschiedliche Weise. Viele Forschungsdaten laufen noch nicht in nationalen oder internationalen Datenzentren der jeweiligen Communities zusammen, sondern verteilen sich auf Repositorien von Journals, Fachdatenbanken und generische Publikationsplattformen wie Zenodo. Diese heterogene Landschaft fragmentiert sich weiter durch community-spezifische Verfahren und Standards. Bereits bestehende Strukturen sollen an den Standorten Tübingen und Freiburg nicht dupliziert werden. Das Ziel besteht in der Bereitstellung von Infrastruktur vor Ort unter Einbindung in übergeordnete Kontexte. Hierbei geht es um die fachspezifische Betreuung mehrerer Communities in *einer* zentral betreuten Instanz, wie es unter anderem für BioDATEN und DataPLANT umgesetzt wird (Martins Rodrigues u. a. 2021). Die Forschungsdaten bilden einen Nachweis der Forschungsaktivität einer Universität und ihrer zugehörigen Forschenden. Der Nachweis und die Recherche in Forschungsergebnissen sollte an einer Stelle erfolgen, unabhängig vom Speicherort der eigentlichen Daten. Die Datensätze sollten hierfür mittels persistenter Identifikatoren bzw. Handles (z.B. DOI, URN) referenzierbar und darüber erreichbar sein. Daraus ergeben sich auf lokaler Ebene zunächst verschiedene Szenarien für die Universität: Die Forschungsergebnisse (Dokumente und/oder Forschungsdaten) sind im institutionellen Repositorium abgelegt und werden dort mit Metadaten beschrieben. Der Nachweis wird an einer zentraler Stelle der Universität geführt, beispielsweise im Forschungsinformationssystem bzw. der Universitätsbibliografie der jeweiligen Bibliothek. Wenn entsprechende fachspezifische Systeme bereits existieren, könnten diese wegen des spezifischen Harvestings von Forschungsdaten eine sinnvolle und bessere Alternative sein. Wünschenswert ist darüber hinaus die automatische Anreicherung des mit der ORCID-iD verknüpften Scholarly Records um die DOIs der Datenpublikationen.

Ein Beispiel für die Abstimmung und Anbindung an institutionelle Partner ist der notwendige und kostenpflichtige Bezug von DOIs durch eine direkte oder indirekte Anbindung an DataCite. Die DOIs für die BioDATEN-Community werden beispielsweise über die Bibliothek der Universität Tübingen bezogen, was eine Klärung des Kostengerüsts notwendig macht. Gleichzeitig sollte eine Qualitätskontrolle der eingereichten Daten erfolgen, welche optimalerweise durch Vertreter aus der wissenschaftlichen Community in der Rolle von Data Stewards erledigt wird (Suchodoletz, Mühlhaus u. a. 2022).

Einsatz an der Universität Tübingen Das Digital Humanities Center der Universität Tübingen[7] betreibt das generische institutionelle Forschungsdatenrepositorium FDAT,[8] das den Anforderungen von Drittmittelgebern entspricht. Der Anspruch liegt darin, als institutionelles Repositorium ein disziplinübergreifendes Angebot für Forschungsdatenmanagement samt Beratungsschwerpunkt auf den Geistes- und Sozialwissenschaften zu schaffen. Das Ziel ist es, einen nachhaltigen Umgang mit Forschungsdaten zu fördern (Abbildung 1 und 2). Eine Säule der Nachhaltigkeit liegt in der Anbindung an eine namhafte Institution und der Einsatz von bwSFS als technische Speicherschicht. Der nachhaltige und

7 `https://dh-center.uni-tuebingen.de`
8 `https://fdat.uni-tuebingen.de`

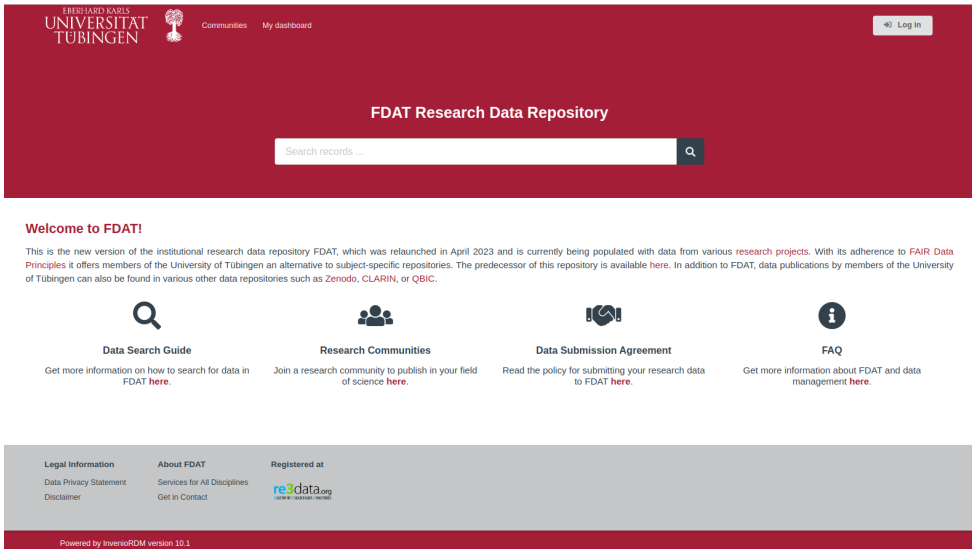


Abbildung 1: Startseite des Forschungsdatenrepositoriums FDAT.

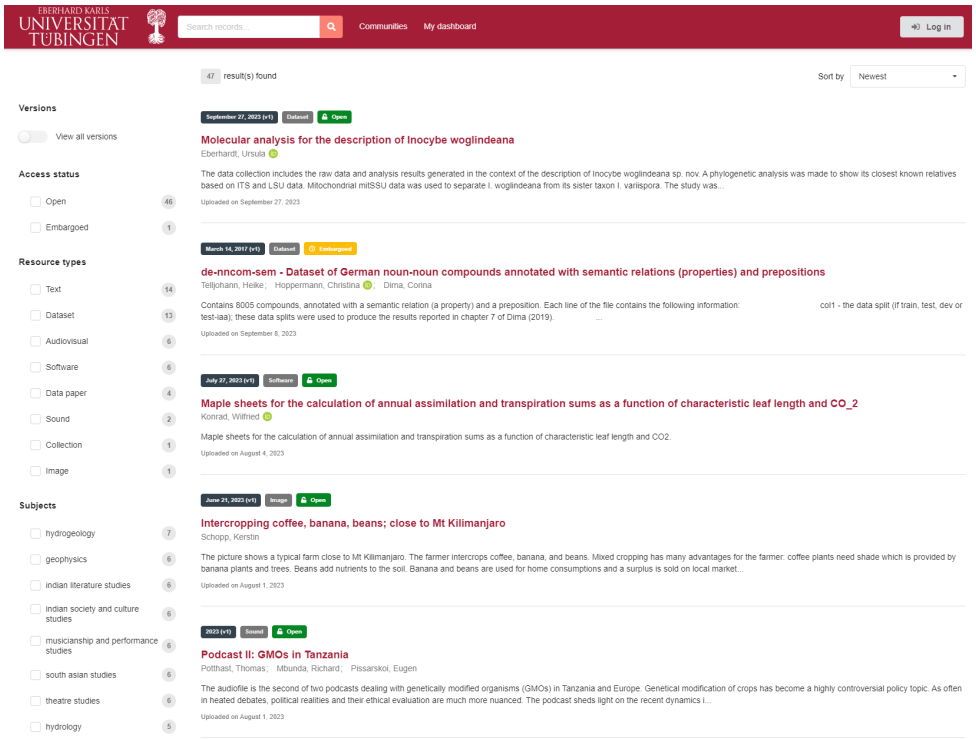


Abbildung 2: Listenansicht publizierter Forschungsdaten.

langfristige Umgang mit Forschungsdaten ist eine organisatorische Herausforderung, wird jedoch in Zertifizierungsprozessen wie jenem von CoreTrustSeal gefordert.[9] Der Einsatz von InvenioRDM baut auf den bereits vorliegenden Erfahrungen und den organisatorischen Vorarbeiten auf und führt diese konsequent weiter.[10] Das technische Grundgerüst InvenioRDM bildet die generische Plattform und eine Ausdifferenzierung erfolgt durch den Aufbau von Communities mit eigenständigen Kurationsworkflow und Qualitätskriterien. Eine solche Community wurde auch von BioDATEN etabliert und dient zur Datenpublikation nach entsprechender Kuration (Abbildung 5 und 6).

Einsatz an der Universität Freiburg Die Universität Freiburg strebt an, im Sinne von Open Data und Open Science eine einfache Publikation von Forschungsdaten zu befördern. Hierzu bildet InvenioRDM die Grundlage für einen neu eingerichteten Publikationsservice „FreiData“ der Research Data Management Group (RDMG),[11] welcher allgemeine, disziplinübergreifende Bedarfe von Forschenden verschiedener Exzellenz-Cluster und Projekte ohne wohletablierte Community-Repositorien bedient (Abbildung 3). Es ergänzt an dieser Stelle das seit längerem etablierte FreiDok+ um die Ablagemöglichkeit insbesondere größerer Forschungsdaten. Komplementär zu bestehenden Repositorien der einzelnen Fach-Communities soll damit eine Lücke im bisherigen Angebot geschlossen werden. Hierzu kooperieren das Rechenzentrum und die Universitätsbibliothek, um eine dauerhafte Bereitstellung dieses Dienstes zu erlauben und dabei Doppelarbeiten und -eingaben seitens der Forschenden zu vermeiden, sowie deren Forschungsleistungen langfristig identifizierbar zu machen und automatisch mit ihrem Scholarly Record zu verknüpfen. Diese Informationen sollen zudem in weiteren Schritten in das neu entstehende Forschungsinformationssystem der Universität einfließen.

Für die langfristige Zuordnung und das Provenance Tracking von Daten folgt die Universität Freiburg der Empfehlung des AK FDM (Suchodoletz u. a. 2020) mit der inzwischen verpflichtenden Verwendung von ORCID-iDs durch das wissenschaftliche Personal.

4.1 Workflow-Integration

Eine Stärke von InvenioRDM liegt in der Bereitstellung von Schnittstellen zur Integration in die bestehende Systemlandschaft und zur Anbindung von Workflows (Abbildung 6). Das Framework bietet eine umfangreiche REST-API, die jegliche Funktion der Plattform umfasst. Dadurch können eingebaute Features wie der Community-basierte Kurations-Workflow in Drittsystemen verwendet werden, ohne diese dort neu zu implementieren. Ein solcher Workflow besteht beispielsweise in der Übernahme von Datenpaketen zur Publikation aus einer Versionierungsplattform im Rahmen von DataPLANT. Hier werden Datenpakete als Annotated Research Contexts (ARC)[12] in GitLab gehalten und bei einer Veröffentlichung per API an InvenioRDM übergeben (Bauer u. a. 2023). Unter Ver-

9 Für weitere Informationen zum Zertifizierungsprozess siehe `https://www.coretrustseal.org`.
10 `https://dh-center.uni-tuebingen.de/fdat-policy/agreement.html`
11 `https://rdmg.uni-freiburg.de`
12 Für darüberhinausgehende weitere Informationen vergleiche `https://www.nfdi4plants.de/content/learn-more/annotated-research-context.html` bzw. Suchodoletz u. a. (2020).

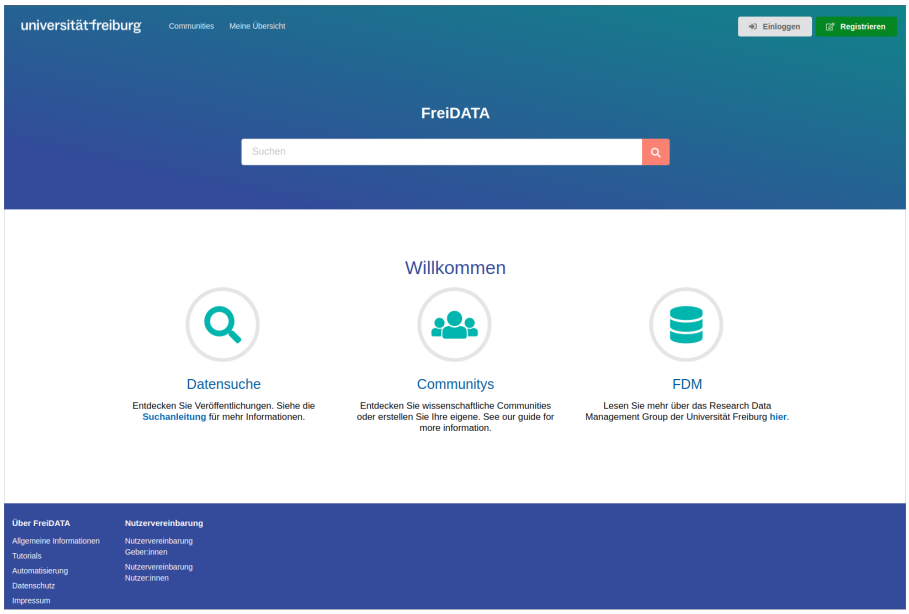


Abbildung 3: FreiData als institutionelles Repositorium der Universität.

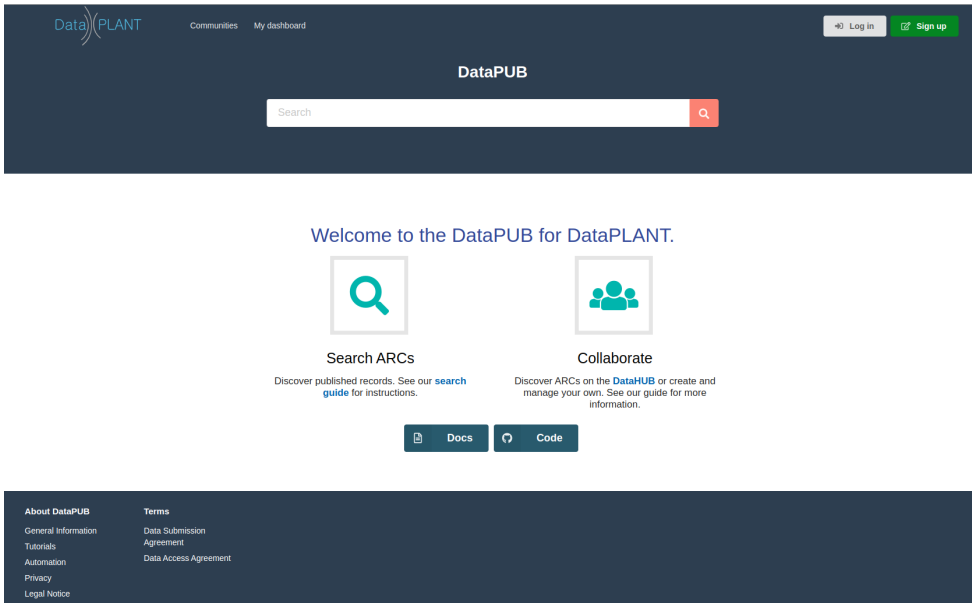


Abbildung 4: DataPUB als Community-Repositorium für DataPLANT.

wendung der GitLab-CI werden ARCs automatisch auf die Vollständigkeit der zu den Daten zugehörigen Metadaten geprüft und den Nutzenden wird ein Feedback in Form eines Badges auf der Webseite angezeigt. Die Nutzenden können validierte ARCs für die Veröffentlichung vorbereiten, was zur Erstellung eines *Drafts* in InvenioRDM führt. Die Drafts werden automatisch durch ein sogenanntes *Submission Request* einer Community zugeordnet und müssen von einem zuständigen Data Steward in der Rolle *Curator* akzeptiert werden. ARCs werden auf diese Weise nochmal einer manuellen Qualitätskontrolle unterzogen. Sobald die Daten für die Publikation freigegeben wurden, werden diese veröffentlicht und ein DOI registriert. Hierdurch wird der Stand der Daten fixiert und referenzierbar. Außerdem können über die OAI-PMH Schnittstelle die Metadaten der veröffentlichen Datensätze in Forschungsinformationssysteme übertragen werden. Der Einsatz der REST-API ermöglicht darüber hinaus die Anbindung an weitere Plattformen im Kontext von bwHPC und Galaxy[13]. Hierdurch werden Plattformen für die Datenproduktion in Form von wissenschaftlichen Workflows und Versionierungsplattformen mit einer Plattform für die Datenpublikation und langfristige Datenhaltung verknüpft.

5 Bisherige Erfahrungen und Ausblick

Der Einsatz von InvenioRDM bietet großes Potential als Ergänzung von etablierten Publikationssystemen der Bibliotheken, ist aber bei weitem kein Selbstläufer. Gerade organisatorische Aspekte sind nicht zu vernachlässigen. Die Verpflichtung zu einer langfristigen Verfügbarkeit von Forschungsdatenpublikationen erfordern ein zukunftssicheres organisatorisches Betriebskonzept und eine zukunftsfähige technologische Speicherschicht. In Baden-Württemberg wird mit Stand Mitte 2023 überlegt, ob ein erweiterter Object-Storage-Verbund im Rahmen von bwSFS eine sehr zuverlässige und effiziente Speicherschicht verteilt über vier Universitätsstandorte in Tübingen, Freiburg, Stuttgart und Hohenheim für diese Zwecke geschaffen werden kann.

Während die Verknüpfung der persistenten Identifier DOI und ORCID verhältnismäßig einfach und mit überschaubarem Koordinationsaufwand gelingen kann, ist die Integration in bereits bestehende Systemlandschaften eine größere Aufgabe, die in der jeweiligen Einrichtung geleistet werden muss. Der Einsatz von InvenioRDM erfordert daher eine enge Koordination und Kooperation der beteiligten Akteure auf allen Ebenen. Das betrifft den Kontakt zur Entwickler-Community ebenso wie Abstimmung mit den Infrastrukturbetreibern, hier bwSFS, und den Einrichtungen, die das organisatorische Gerüst für die Nutzerauthentifizierung und die DOI-Schnittstelle bereitstellen. Gleichzeitig müssen die Forschenden einbezogen und ihre Bedarfe berücksichtigt werden.

Die Umsetzung des Forschungsdatenmanagements und speziell des Datenpublikations-Workflows benötigt eine nachhaltige Finanzierung und Ausstattung nicht nur wegen der erwartbaren erheblichen Zunahme der Datenmengen. Ein Publikationssystem wird ebenso wie andere Unterstützungssysteme für die Wissenschaft zu einer zentralen Infrastruktur mit entsprechendem Finanzierungsbedarf und langfristiger Verpflichtung über den eige-

13 `https://galaxyproject.org`

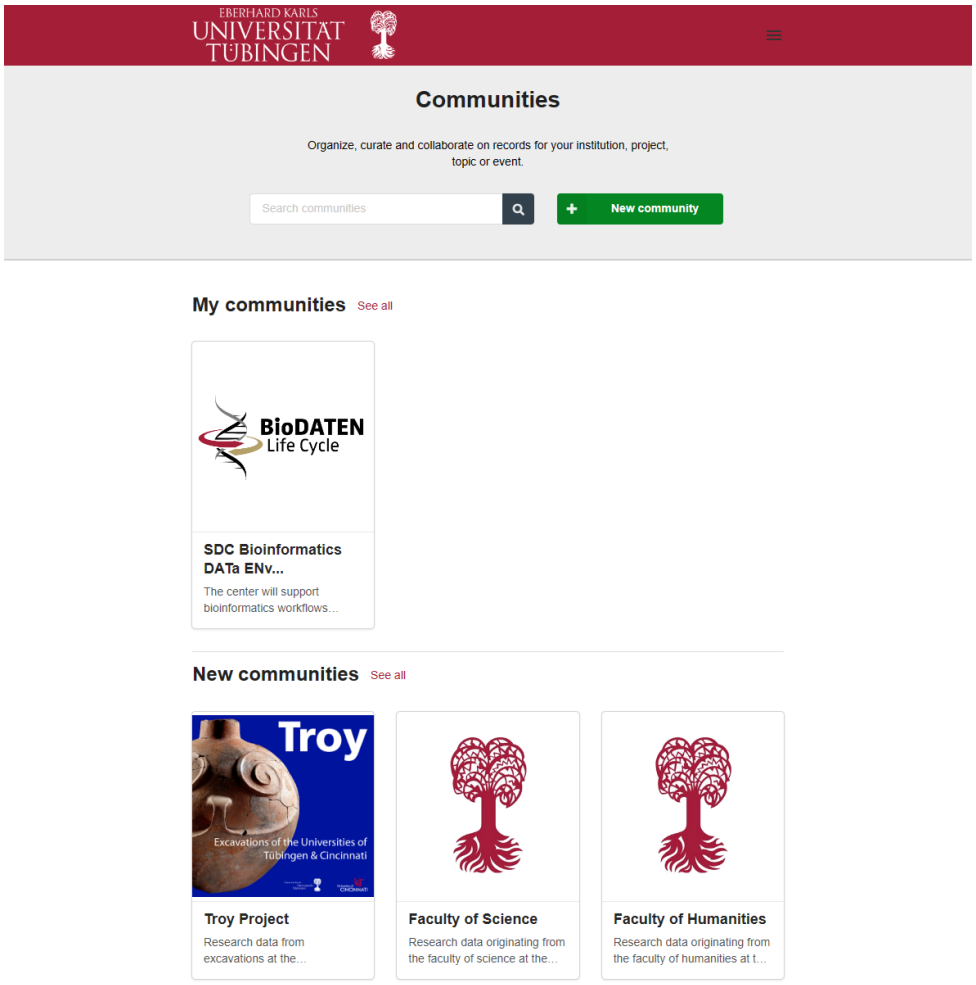


Abbildung 5: Darstellung und Auswahl von Communities in FDAT.

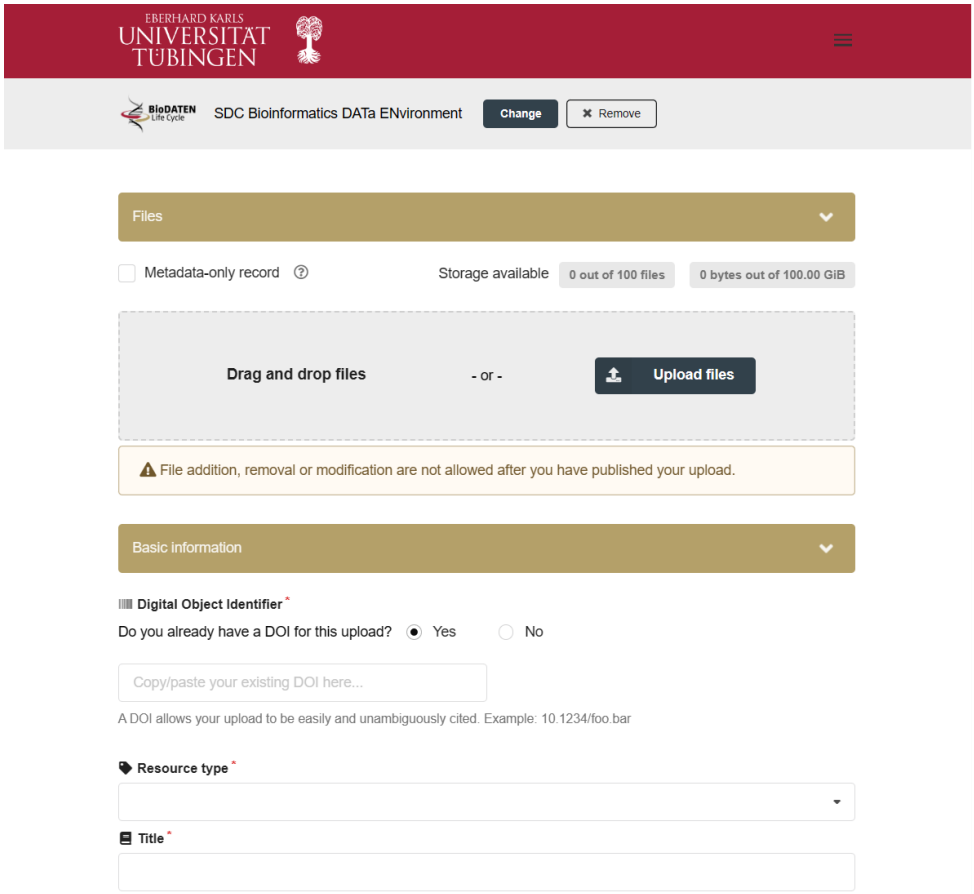


Abbildung 6: Umsetzung eines Publikationsworkflows der BioDATEN-Community.

nen Standort hinaus. Ein Datenrepositorium wie InvenioRDM benötigt zukunftssichere Produktentwicklung, die zumindest in der derzeitigen Konstellation aus Open Source und starkem Akteur über die nächsten Jahre sichergestellt sein sollte. Hier kann es sinnvoll sein, gemeinsame Foren von anwendenden Einrichtungen zu schaffen, die sich regelmäßig beispielsweise zu Möglichkeiten und Beispielen der API-Programmierung austauschen.

Da ein nicht unerheblicher Teil zukünftiger Kosten von der Datenmenge abhängt (Leendertse und Suchodoletz 2020), sind Optionen zur Beteiligung der Nutzenden insbesondere bei erheblichen Speicherbedarfen vorzusehen. Prinzipbedingt beinhaltet FDM ein nachhaltiges Engagement der beteiligten Parteien, mindestens jedoch der beauftragten Institutionen für die langfristige Speicherung der Daten. Zwischen den verschiedenen Wissenschafts-Communities, den Betreibern wie Rechenzentrum und Universitätsbibliothek sowie den Mittelgebern wie Universität oder Forschungsförderer muss daher ein sinnvoller Ausgleich der Interessen und Kosten organisiert werden.

Danksagung

Wir danken dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg für die Unterstützung des Science Data Centers BioDATEN im Rahmen der Digitalisierungsstrategie digital@bw und die Co-Finanzierung der bwSFS-Speicherinfrastruktur. bwSFS wird ebenfalls durch die Deutsche Forschungsgemeinschaft DFG gefördert: GZ: INST 37/1046-1 FUGG, GZ: INST 37/1047-1 LAGG, GZ: INST 39/1099-1 FUGG, GZ: INST 39/1098-1 LAGG. Das Konsortium DataPLANT wird durch die Deutsche Forschungsgemeinschaft DFG gefördert: NFDI 7/1 – 442077441 auf Basis der Bund-Länder-Vereinbarung zum Aufbau einer nationalen Forschungsdateninfrastruktur vom 26. November 2018 finanziert.

ORCID:

- Dirk von Suchodoletz https://orcid.org/0000-0002-4382-5104
- Jonathan Bauer https://orcid.org/0000-0002-5624-2055
- Holger Gauza https://orcid.org/0000-0003-0191-3680

Literaturverzeichnis

Albert-Ludwigs-Universität Freiburg, Rektorat. 2022. *Policy zum Umgang mit Forschungsdaten an der Universität Freiburg.* DOI: `https://doi.org/10.6094/UNIFR/231612`. `https://freidok.uni-freiburg.de/data/231612`.

Axtmann, Alexandra, Felix Bach, Jonathan Bauer, André Blessing, Thomas Bönisch, Nina Buck, Holger Gauza u. a. 2021. „Kriterien für die Auswahl einer Softwarelösung für den Betrieb eines Repositoriums für Forschungsdaten“. *Bausteine Forschungsdatenmanagement,* Nr. 3: 14–26. DOI: https://doi.org/10.17192/bfdm.2021.3.8348. https://bausteine-fdm.de/article/view/8348.

Bauer, Jonathan, Marcel Tschöpe, Julian Weidhase, Timo Mühlhaus, Christoph Garth, Gajendra Doniparthi, Holger Gauza, Louisa Perelo, Cristina Martins Rodrigues und Dirk von Suchodoletz. 2023. „From DataPLANT's DataHUB to DataPUB(lication)“. In *International Workshop on Science Gateways.* Accepted for publication.

Brettschneider, Peter, Alexandra Axtmann, Elisabeth Böker und Dirk von Suchodoletz. 2021. „Offene Lizenzen für Forschungsdaten: Rechtliche Bewertung und Praxistauglichkeit verbreiteter Lizenzmodelle“. *O-Bib. Das Offene Bibliotheksjournal* 8 (3): 1–22. DOI: https://doi.org/10.5282/o-bib/5749. https://www.o-bib.de/bib/article/view/5749.

Deutsche Forschungsgemeinschaft e.V. 2013. *Sicherung guter wissenschaftlicher Praxis.* Wiley Online Library. ISBN: 978-3-527-33703-3. DOI: https://doi.org/10.1002/9783527679188. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527679188.

Leendertse, Jan, Susanne Mocken und Dirk von Suchodoletz. 2019. „Datenmanagementpläne zur Strukturierung von Forschungsvorhaben“. *Bausteine Forschungsdatenmanagement,* Nr. 2: 4–9. DOI: https://doi.org/10.17192/bfdm.2019.2.8003.

Leendertse, Jan, und Dirk von Suchodoletz. 2020. „Kosten und Aufwände von Forschungsdatenmanagement“. *Bausteine Forschungsdatenmanagement,* Nr. 1 (1): 1–7. DOI: https://doi.org/10.17192/bfdm.2020.1.8246. https://bausteine-fdm.de/article/view/8246.

Martins Rodrigues, Cristina, Dirk von Suchodoletz, Timo Mühlhaus, Jens Krüger und Björn Usadel. 2021. „DataPLANT – Ein NFDI-Konsortium der Pflanzen-Grundlagenforschung“. *Bausteine Forschungsdatenmanagement,* Nr. 2 (2): 46–56. DOI: https://doi.org/10.17192/bfdm.2021.2.8335. https://bausteine-fdm.de/article/view/8335.

Suchodoletz, Dirk von, Elisabeth Böker, Peter Brettschneider und Franziska Rapp. 2020. „Entwicklung in Baden-Württemberg: ORCID und ROR IDs als Standard für langfristige Personen-und Institutionen-Identifier“. *Bausteine Forschungsdatenmanagement,* Nr. 2: 80–88. DOI: https://doi.org/10.17192/bfdm.2020.2.8272. https://doi.org/10.17192/bfdm.2020.2.8272.

Suchodoletz, Dirk von, Peter Brettschneider, Elisabeth Böker, Jochen Apel, Dorothea Iglezakis, Karsten Schmidt und Gabriel Schneider. 2021. *Leitfaden – Verantwortungsvoller Umgang mit Forschungsdaten.* DOI: https://doi.org/10.5281/zenodo.4907422. https://doi.org/10.5281/zenodo.4907422.

Suchodoletz, Dirk von, Ulrich Hahn, Jonathan Bauer, Kolja Glogowski und Mark Seifert. 2022. „Storage for Science – Aktueller Stand und anstehende Entwicklungen eines verteilten FDM-Systems“. In *E-Science-Tage 2021: Share Your Research Data,* herausgegeben von Vincent Heuveline und Nina Bisheh, 298–305. Heidelberg: heiBOOKS. DOI: https://doi.org/10.11588/heibooks.979.c13741.

Suchodoletz, Dirk von, Ulrich Hahn, Bernd Wiebelt, Kolja Glogowski und Mark Seifert. 2019. „Storage infrastructures to support advanced scientific workflows: Towards research data management aware storage infrastructures“. In *Proceedings of the 5th bwHPC Symposium: HPC Activities in Baden-Württemberg,* Freiburg, September 2018, herausgegeben von Michael Janczyk, Dirk von Suchodoletz und Bernd Wiebelt, 263–279. TLP, Tübingen. DOI: https://doi.org/10.15496/publikation-29058. http://hdl.handle.net/10900/87672.

Suchodoletz, Dirk von, Timo Mühlhaus, Dominik Brilhaus, Hajira Jabeen, Björn Usadel, Jens Krüger, Holger Gauza und Cristina Martins Rodrigues. 2022. „Data Stewards as ambassadors between the NFDI and the community“. In *E-Science-Tage 2021: Share Your Research Data,* herausgegeben von Vincent Heuveline und Nina Bisheh, 358–365. Heidelberg: heiBOOKS. DOI: https://doi.org/10.11588/heibooks.979.c13750.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Datensammlung in der Romanistik – Eine Analyse von Normierung und Standardisierung in E-Mails

Laura Bothe, Sybille Große

Romanisches Seminar, Universität Heidelberg

Den E-Mailaustausch gibt es seit nunmehr 50 Jahren (Delfa 2021). Er ist heute ein weltweit anerkanntes Kommunikationsmedium, sowohl in formellen als auch informellen Kommunikationskontexten. Gerade deshalb stellt sich aus linguistischer Sicht die Frage nach der Standardisierung und Normierung von E-Mails (Große 2012). Nach einer Studie von *Statista* wurden im Jahr 2021 ca. 319 Milliarden E-Mails am Tag verschickt (Statista 2023). Tendenz steigend. Trotzdem ist das Interesse der Linguisten an E-Mails in den letzten Jahren zurückgegangen. Das Aufkommen der sozialen Medien (Rentel und Schröder 2018) und der Wechsel vom informellen auf einen formelleren Gebrauch der E-Mail (Souchier u. a. 2019), der sich auf die Kommunikationssituationen und verwandten Versprachlichungsstrategien auswirkt, können hierfür als Erklärung herangezogen werden. Zudem haben sich die Fragen der Datenerhebung und -verarbeitung im Rahmen der Analyse von internetbasierter Kommunikation (IBK) auf der Grundlage von sich in den *Digital Humanities* etablierenden Standards in den letzten Jahren zu einer immer größeren Herausforderung entwickelt (Beißwenger 2017). Gleichzeitig wird die Datenerhebung in sozialen Medien wie *Twitter* oder *Telegram* von den Unternehmen durch die vorgegebenen Privatsphäre-Regelungen und entsprechenden Download-Tools vereinfacht. Ein umfangreiches, zeitlich relevantes, zahlreiche Schreiber:innen umfassendes und aus formellen und informellen E-Mails bestehendes Korpus anzulegen, ist bereits durch den restriktiveren Zugang zu den Daten mit erheblichem Mehraufwand verbunden. Da es keine zentrale Sammelstelle gibt, an der die E-Mails abrufbar sind, ist die Forschung in diesem Bereich im Wesentlichen auf Daten-Spenden angewiesen. Dabei stoßen wir auch im akademischen Kontext auf großes Misstrauen potenzieller Spender:innen im Bereich der Datensicherheit. Tatsächlich beinhalten E-Mails persönliche und personenbezogene, also sogenannte „sensible Daten“, deren Löschung und Anonymisierung gerade bei multilingualen Daten, die in Emails keinesfalls ausgeschlossen werden können, eine Hürde darstellen kann.

Zwar besteht durch die Digitalisierung gleichfalls in der linguistischen Forschung die Möglichkeit immer komplexere Daten in Korpora zusammenzufassen, dieses Unterfangen setzt jedoch eine Infrastruktur und Datenmodelle voraus, die dieses unterstützen. Am Beispiel

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.c18070 (Freier Zugang – alle Rechte vorbehalten)

des Projekts *Zwischen Briefen und E-Mails: Dynamiken der Normierung und Standardisierung* werden die Herausforderungen thematisiert, denen wir begegnen, wenn es um die Akquise, Verarbeitung und Publikation von mehrsprachigen IBK-Daten geht.

1 Einleitung

50 Jahre E-Mail-Verkehr

Die Kommunikation per E-Mail findet nunmehr seit 50 Jahren statt. Diese computerbasierte, zuerst für professionelle Kontexte entwickelte Methode der Nachrichtenübermittlung hat sich zwischen dem Ende der 1970er und den 1990er Jahren zu einer eher informellen bzw. persönlichen Kommunikationsform entwickelt (Rentel und Schröder 2018). Die in den letzten Jahrzehnten sichtbar gewordene materielle und soziale Metamorphose, die der Brief in Richtung der E-Mail durchlief (ebd.), führte zu einer starken Heterogenität der einzelnen Mails z.B. in ihren Formalitätsgraden und den Kommunikationssituationen. In einer Prognose von 2022 geht Statista (2023) davon aus, dass im Jahr 2023 347 Milliarden E-Mails geschrieben werden – Tendenz steigend. Dieser steigenden Kommunikation zum Trotz ist die linguistische Auseinandersetzung bzw. das Interesse an der linguistisch-kommunikativen Erforschung dieser Form der Kommunikation rückläufig. Nach einer relativ großen Zahl von europäischen Arbeiten um die Jahrtausend-Wende (Baron 1998; López Alonso und Seré 2003; Ziegler und Dürscheid 2007; Anis 1999) ist das Forschungsinteresse an der E-Mail-Kommunikation seither zurückgegangen. Gerade die aktuelle Heterogenität und die Entwicklung der E-Mails hin zu einer oft in distanzierteren, in professionellen Kontexten genutzten Kommunikationsform wirft Fragen rund um die Standardisierung von E-Mails auf: In welchen Schreibsituationen wird eine Standardisierung am deutlichsten manifest? Wie zeigt sie sich im sprachlichen Gebrauch in den einzelnen romanischen Sprachen und hier vor allem im Spanischen, Portugiesischen und Französischen? Gibt es neben den impliziten Normen auch explizite Festschreibungen einer empfohlenen Verwendung? Was wissen wir über die Anwendung spezifischer einzelsprachlicher oder auch diskursiver Empfehlungen?

Derartige Fragen lassen sich nur auf Grundlage einer exhaustiven Menge an Forschungsdaten analysieren. Hier liegt ein Grund, warum die E-Mail-Kommunikation nach einem ersten lebhaften Forschungsinteresse aus dem linguistischen Forschungsfokus rückte. Der aufkommende prominentere Datenschutz und die damit verbundenen Einschränkungen in der Datenakquise erschweren die Erstellung umfassender E-Mail-Korpora. Die Hürden und Herausforderungen auf dem Weg zu einem mehrsprachigen Korpus aus spanischen, französischen und portugiesischen E-Mails, das zur Beantwortung der genannten Forschungsfragen dienen könnte, sind Gegenstand der vorliegenden Abhandlung.

2 Erstellung von Korpora der E-Mail-Kommunikation

2.1 Herausforderungen

Um Normierungs- bzw. Standardisierungsprozesse in E-Mails quantitativ zu erforschen, sollte das zu erstellende Korpus möglichst syn-, aber auch diachrone Daten enthalten, um auch einzelne Perioden der Standardisierung nachzeichnen zu können. Eine hohe Anzahl von Schreiber:innen aus den drei Sprachräumen ist die Voraussetzung, um wechselnde Schreibsituationen sowie die eingangs erwähnte Heterogenität abbilden zu können. Unsere Daten müssen Metadaten wie Sozioprofession, Generationenzugehörigkeit, Formalität des Schreibkontextes etc., beinhalten, um Variation, Innovation, Wandel und Standardisierung an verschiedenen Schreibsituationen nachzuvollziehen und benennen zu können.

Bei der Datenerhebung stellen sich demzufolge bereits zwei Herausforderungen. Eine erste ist die rechtliche Herausforderung des Speicherns und Verarbeitens jener Daten, die durch den Gesetzgeber geschützt sind und personenbeziehbare Informationen beinhalten (siehe §3 BDSG bzw. auf Europäischer Ebene Artikel 5 der Europäischen Datenschutz-Grundverordnung). Die zweite Herausforderung ist ethischer Natur. In E-Mails kommen nicht nur personenbeziehbare, sondern zugleich auch persönliche Informationen, wie Haltungen und Meinungen zum Ausdruck. Dies führt gerade bei der Datenakquise zu Zurückhaltung in der Zustimmung möglicher Spender:innen zur Nutzung der Daten als linguistische Forschungsgrundlage. Die Datensammlung und -verarbeitung ist nur im Einklang mit einer Datenanonymisierung möglich, um einerseits im rechtlichen Rahmen zu forschen und andererseits, den Spender:innen das nötige Vertrauen in unsere Forschung und den Schutz ihrer Daten zu bieten.

2.2 Die Sammlung erster Test-Daten

In einem Pilotprojekt haben wir 2021 zunächst zur Spende von E-Mails bei französischen und spanischen Stiftungen und Vereinen, die im sprachlichen Bereich agieren, aufgerufen. Da ein E-Mailverlauf häufig aus mehr als nur einer Nachricht mit einem Absender und einem Empfänger besteht, stellte sich das Einholen der sogenannten informierten Einwilligung als schwierig heraus. Im Regelfall sind in einer E-Mailkonversation mehrere Personen direkt oder indirekt involviert, deren Einverständnis nur schwer zu erhalten ist.

Trotz dieser Hürde konnten wir ein Sample von ca. 1000 französischen und spanischen E-Mails nutzen, um einen ersten Schritt in die Richtung einer automatisierten Datenverarbeitung zu gehen. Eine solche bietet uns die Möglichkeit, das Risiko für uns, aber zugleich für die Spender:innen. zu minimieren und Daten gesetzeskonform und ethisch sammeln sowie verarbeiten zu können.

3 Ein Weg in Richtung austomatisierten Datenverarbeitung

3.1 Der anonymizer zur Anonymisierung mehrsprachiger Daten

Zur Datenverarbeitung haben wir die universitäre Infrastruktur der Universität Heidelberg bemüht. Hier hat uns das *Scientific Software Center* (SSC) bei der Erstellung eines Algorithmus zur Anonymisierung unterstützt. Das SSC ist dem *Interdisziplinären Zentrum für Wissenschaftliches Rechnen* (IWR) der Universität Heidelberg angegliedert.

Der entwickelte Algorithmus ist als beta version auf github zu finden (Git Hub Repository: `https://github.com/ssciwr/anonymize`). Es handelt sich um einen Prototypen in das .eml Dateien eingelesen und.txt Dateien ausgeworfen werden. Ein erstes Modul säubert den Umgebungstext (An, Von, Datum, Betreff etc.) und extrahiert Sätze mit Satzerkennungstool *SpaCy* (Montani u. a. 2023), auf Grundlage der NLP-Modelle „fr_core_news_sm“ „es_core_news_sm“.

Ein zweites Modul wendet das Named-Entity-Recognition (NER)-Tools *Stanza* auf die Sätze an. Pro Satz werden also mit *Stanza* (Qi u. a. 2020) persönliche Daten (*Named Entities*) extrahiert und durch die jeweiligen Entitäten-Namen (Person, Organisation, Orte) ersetzt. Da *Stanza* auf einsprachigen Modellen basiert, bedarf der Algorithmus demzufolge einer Voreinstellung für die gewünschte Sprache. Da die NER für die einzelnen Sprachen auf verschiedenartig trainierten Modellen basiert, sind die Anonymisierungs-Ergebnisse für unsere französischen und spanischen Test-E-Mails unterschiedlich ausgefallen.

3.2 Technische Hürden der Anonymisierung

Einige Hürden stellten sich uns bei der Anonymisierung mit den tools *SpaCy* und *Stanza* in den Weg. Da die Satztrennung erheblich zu einer korrekten Erkennung von NER beiträgt, beginnt die Schwierigkeit bei der korrekten Erfassung eines Satzes in den E-Mails. Wie bereits eingangs ausgeführt, sind E-Mails allerdings eine heterogene Kommunikationsform und vereinen Merkmale aus verschiedenen Textsorten und Diskurstraditionen, wie Brief, Textnachricht oder administratives Schreiben. Sie können sowohl Spuren von distanzkommunikativen Schreibens als auch Nähe-Markierungen (Koch und Oesterreicher 1985) aufweisen. Beide tools, *SpaCy* als auch *Stanza*, sind auf distanzsprachlichen Modellen trainiert, was zu zahlreichen Fehlern insbesondere in den französischsprachigen E-Mails führte.

Ein weiteres Hindernis ist der Umgebungstext, der in unseren Test-E-Mails uneinheitlich und deshalb schwierig zu säubern war. Da dieser stark von den sprachlichen Voreinstellungen des jeweiligen, die E-Mails generierenden PC's abhängt, müssen an diesem Punkt mehrsprachliche Umgebungstexte noch stärker berücksichtigt werden.

Bei der Benutzung von *Stanza* wird darüber hinaus ein generelles Problem der Behandlung von Zahlen unterschiedlicher Formate deutlich. Sowohl Postleitzahlen als auch Telefonnummern werden in der beta-Version nicht anonymisiert.

Auch Signaturen sind sehr vielgestaltig aufgebaut und können so unterschiedlich gut extrahiert werden. Eine voreingestellte Signatur kann Informationen, z.B. Grußformeln, enthalten, die für die Analyse wichtig sind. Sie kann aber auch gesponsorte Werbungen aufweisen, die statistische Zugriffe verfälschen dürften, da der Wordcount von ihnen betroffen ist. Da die Erkennung personenbezogener Daten in *Stanza* auf einem Datenmodell aus Wikipe-dia-Artikeln (*wikiner*) basiert, kommt es außerdem vor, dass die NER zu sensibel ist. Zu viele Entitäten oder auch zu wenige werden als Personen, Organisationen oder Orte ausgewiesen. Im Gegensatz zu dem französischen Wikipedia-Modell, basiert das spanische NER-Modell in *Stanza* auf Medientexten. Es konnte festgestellt werden, dass die Anonymisierung unserer spanischen Testdaten weniger Fehler in Bezug auf die Sensibilität enthielt.

3.3 Die transformers Datenbank als ein Lösungsansatz

Ein erster Lösungsansatz, den wir gemeinsam mit dem SSC gefunden haben, ist die Benutzung der *transformers* Datenbank. Anders als *Stanza* ist die *transformers*-Datenbank mehrsprachig, weshalb die Sprache nicht mehr ausgewählt werden muss; das Tool basiert auf einem generalisierten multilingualen Korpus und ist so auf verschiedensprachige Datensets anwendbar. Die ersten Tests mit der NER aus der Datenbank zeigten sowohl in den französischen, als auch in den spanischen E-Mails verbesserte Ergebnisse in der *Named Entity Recognition*. Gemeinsam mit einem stärkeren Fokus auf die Säuberung der Rohdaten, erscheint der Wechsel des NER-Tools als für das zukünftige Projekt vielversprechend. Eine zweite Möglichkeit wäre ein eigenes *One shot training*. Hier würde ein von den Ingenieur:innen des SSC entwickeltes Modell basierend auf von uns annotierten Daten trainiert. Diese Lösung bedeutet allerdings einen erhöhten Aufwand im Projekt, da eine umfangreichere Menge Daten manuell annotiert werden müsste.

4 Zur Datenveröffentlichung

Ähnlich wie in Beißwenger u. a. (2017) beschrieben, stellt sich auch für unser Projekt im Anschluss an die Verarbeitung der Daten die ethische und rechtliche Frage nach der Veröffentlichung sprachgebrauchsbezogenen Daten in der linguistischen Forschung. Diese stellt sich im Übrigen zugleich bei Projekten, die mit *Twitter* oder *Telegram*-Nachrichten als Datengrundlage arbeiten. Die Nutzer stimmen bereits bei der Anmeldung auf den einschlägigen Plattformen der Verarbeitung ihrer Daten zu – häufig ohne sich dessen im Detail bewusst zu sein. Es bleiben persönliche und teils auch personenbeziehbare Daten, die wir für die Forschung benutzen und somit auch in Ausschnitten veröffentlichen möchten. Aus dem von Beisswenger et al. vorgestellten Rechtsgutachten zum Dortmunder Chatkorpus geht hervor, dass Daten aus der internet-basierten Kommunikation durch die sehr unterschiedlichen enthaltenen personenbezogenen Informationen rechtlich anders behandelt werden sollten. In diesem Beispiel, das im Bereich des Datenschutzes auch auf unser Projekt übertragbar ist, wurden so Subkorpora zu verschiedenen Kommunikati-

onssituationen gebildet, um die Daten kontextbezogen auf ihre Personenbeziehbarkeit zu prüfen.

Zur rechtlichen Komponente kommt noch die Methodische hinzu: verschiedene Kontexte müssen zwingend klassifiziert werden und würden im Anschluss datenschutzrechtlich nuanciert behandelt werden. Im Chatkorpus von Beisswenger et al. wurden so Daten mit besonders hoher Sensitivität aus z.B. psychosozialen Beratungen vollständig aus dem Korpus gelöscht. Bereits die Erfassung dieser Daten gilt laut des Gutachtens als unzulässig, wenn sie, unter anderem, zum Zwecke der Veröffentlichung dient. Das für das Chatkorpus in Auftrag gegebene Rechtsgutachten weist zudem darauf hin, dass bei der Erhebung bereits ein konkreter Zweck für die Datensammlung angegeben werden müsse (Montani u. a. 2023), hier wird die Forschung bislang nicht unter § 28 Abs. 3 Nr. 4 BDSG (Archivzwecken im öffentlichen Interesse) geführt.

5 Die zukünftige Fusion der Datensammlung und Datenverarbeitung in einer Spendenwebseite

Zukünftig soll aus dem Projekt eine Spendenwebseite im Netz werden, bei der E-Mail-Spender:innen E-Mails direkt hochladen und so keine Interaktion von den Spender:innen mit den Forschenden mehr nötig ist. Während andere E-Mail-Korpora auf Datensets aus online Archiven basieren, wie das in der Germanistik ansässige Projekt *CodE Alltag* (Krieg-Holz u. a. 2016), sind Spendenwebseiten gerade bei internetbasierter Kommunikation bereits an anderer Stelle genutzt worden. So konnte das Projekt *What's up*, (Ueberwasser und Stark 2017) 617 Chatverläufe sammeln. Derzeit läuft ein weiteres Projekt der Universitäten Lothringen, Lüttich und Strasbourg bei dem Audionachrichten gesammelt werden (Glikman und Fauth 2022).

Die von uns angedachte Webseite soll auch die umgehende Anonymisierung der Daten ermöglichen. Dazu planen wir, auf dem Pilotprojekt aufzubauen und den in Zusammenarbeit mit dem SSC entwickelten Algorithmus in die Webseite zu integrieren. Des Weiteren sollen die Nutzer:innen der Webseite während des *uploads* Angaben zu ihrer Person und der Schreibsituation machen können. Die Spender:innen sollen unmittelbar die Möglichkeit erhalten, die Daten zur Verarbeitung an die Universität freizugeben und damit zugleich eine sogenannte *Informierte Einwilligung* für den Verlauf erteilen. Wir stellen im Gegenzug Transparenz in Bezug auf die Datennutzung her, indem wir unser Projekt vorstellen und die Spender:innen über die Art der Weiterverarbeitung informieren. So können wir Metadaten sammeln, die wir zu unserer Auswertung verwenden, aber gleichzeitig technisch absichern, dass es sich um nutzbare Daten handelt. In einem letzten Schritt kann mit Hilfe der Metadaten und Einwilligungen die Möglichkeit einer Veröffentlichung unseres Korpus rechtlich geprüft werden.

Danksagung

Besonderer Dank gilt dem Scientific Software Center der Universität Heidelberg und speziell Frau Inga Ulusoy für die Entwicklung des *anonymizer* Algorithmus.

Literaturverzeichnis

Anis, Jacques. 1999. *Internet communication et langue française.* 191. Paris: Hermes Sciences Publications. ISBN: 2-7462-0063-5.

Baron, Naomi S. 1998. „Letters by phone or speech by other means: the linguistics of email". *Language and Communication* 18 (2): 133–170. DOI: https://doi.org/10.1016/S0271-5309(98)00005-6.

Beißwenger, Michael, Hrsg. 2017. *Empirische Erforschung internetbasierter Kommunikation.* Berlin, Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110567786.

Beißwenger, Michael, Harald Lüngen, Jan Schallaböck, John H. Weitzmann, Axel Herold, Pawel Kamocki, Angelika Storrer und Julia Wildgans. 2017. „Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens". In *Empirische Erforschung internetbasierter Kommunikation,* 7–46. Berlin, Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110567786-002.

Delfa, Christina Vela. 2021. *La comunicación por correo electrónico: análisis discursivo de la correspondencia digital.* Madrid,Frankfurt: Iberoamericana;Vervuert.

Glikman, Julie, und Camille Fauth. 2022. „Un nouvel accès à la parole spontanée : les vocaux". In *XXXIVe Journées d'Études sur la Parole – JEP 2022.* ISCA. DOI: https://doi.org/10.21437/JEP.2022-17.

Große, Sybille. 2012. „Sprache und Öffentlichkeit in realen und virtuellen Räumen". Kap. Französische E-Mails: Briefmodelle im Abschwung, herausgegeben von Annette Gerstenberg, Claudia Polzin-Haumann und Dietmar Osthus, 126–139. Romanistischer Verlag. ISBN: 978-3-86143-202-9.

Koch, Peter, und Wulf Oesterreicher. 1985. „Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte". *Romanistisches Jahrbuch* 36 (1): 15–43. DOI: https://doi.org/10.1515/9783110244922.15.

Krieg-Holz, Ulrike, Christian Schuschnig, Franz Matthies, Benjamin Redling und Udo Hahn. 2016. „CodE Alltag: A German-Language E-Mail Corpus". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 2543–2550. Portorož, Slovenia: European Language Resources Association (ELRA).

López Alonso, Covadonga, und Arlette Seré, Hrsg. 2003. *Nuevos Generos Discursivos: Los Textos Electronicos.* 219. Madrid: Biblioteca nueva. ISBN: 978-8497422017.

Montani, Ines, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann u. a. 2023. *explosion/spaCy: v3.5.2: Pretraining improvements, bug fixes for spans and spancat and more.* DOI: https://doi.org/10.5281/zenodo.7820813.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton und Christopher D. Manning. 2020. „Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.acl-demos.14.

Rentel, Nadine, und Tilman Schröder. 2018. *Sprache und digitale Medien.* Berlin: Peter Lang Verlag. DOI: https://doi.org/10.3726/b12951.

Souchier, Emmanuel, Étienne Candel, Gustavo Gomez-Mejia und Valérie Jeanne-Perrier. 2019. *Le numérique comme écriture. Théories et méthodes d'analyse.* Paris: Armand Collin. ISBN: 978-2-200-61858-2.

Statista. 2023. „Number of sent and received e-mails per day worldwide from 2017 to 2026". Besucht am 15. Mai 2023. https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/.

Ueberwasser, Simone, und Elisabeth Stark. 2017. „What's up, Switzerland? A corpus-based research project in a multilingual country". *Linguistik Online* 84 (5). DOI: https://doi.org/10.13092/lo.84.3849.

Ziegler, Arne, und Christa Dürscheid, Hrsg. 2007. *Kommunikationsform E-Mail.* Tübingen: Stauffenburg Verlag Brigitte Narr GmbH. ISBN: 978-3-86057-686-1.

Carrots and Sticks: Motivating with Storage for Good RDM – Science Led Allocation of Research Data Storage Resources within an Integrated RDM System

Ilona Lang, Marcel Nellesen, Lukas C. Bossert, Marius Politze

IT Center, RWTH Aachen University

Storage space is valuable and there are many researchers who need to store their research data (also demanded by the Good Scientific Practice (GSP); Deutsche Forschungsgemeinschaft e.V. 2019). Most existing storage distribution systems are ad-hoc, require (internal) transfer of funds, or do not scale on institutional or even national level. Most importantly, the value for the scientific community often remains unaddressed. Starting with our existing data management platform Coscine we adapted the Joint Application Review and Dispatch Service (JARDS; Janetzko 2019), a tool already utilized within many computing centers within Germany to handle applications for computing time. Hence, our aim is to unify applications for scientific IT resources and lighten the process of formalities management.

1 Introduction

The structured organization of research data is eminent to research projects. And since metadata are more and more required to fulfill the requirements of e.g., FAIR principles (Wilkinson et al. 2016) and/or GSP, researchers are confronted not only with the task to find a suitable storage system for their data along with the metadata, but also they need to find a system with enough storage capacity for ongoing and finalized projects. At the RWTH Aachen University, we support researchers with the research data management platform Coscine (Politze et al. 2020). Coscine enables researchers to store their data along with all needed and demanded customized metadata. It further provides sufficient storage capacity in a secured and through Coscine easily accessible and manageable way. To achieve this, Coscine combines (decentral) data storage systems with a metadata management (Schmitz and Politze 2018; Politze et al. 2020). Technically it leverages persistent identifier (PID; Kálmán, Kurzawe, and Schwardmann 2012; Krämer, Politze, and Schmitz 2016) and linked data technologies on multiple levels: projects, storage

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.c18071

resources and individual files and applies the FAIR Digital Object (FDO; Smedt, Koureas, and Wittenburg 2020) concept and Data Catalog Vocabulary (DCAT; Maali and Erickson 2014, cf. Figure 1). One of the core storage systems behind Coscine is Research Data Storage (RDS; Eifert, Claus, and Lopez 2018), a geo-redundant object storage system that is provided by a consortium of universities for all researchers within the federal state of North Rhine-Westphalia and their collaboration partners within the National Research Data Infrastructure (NFDI).

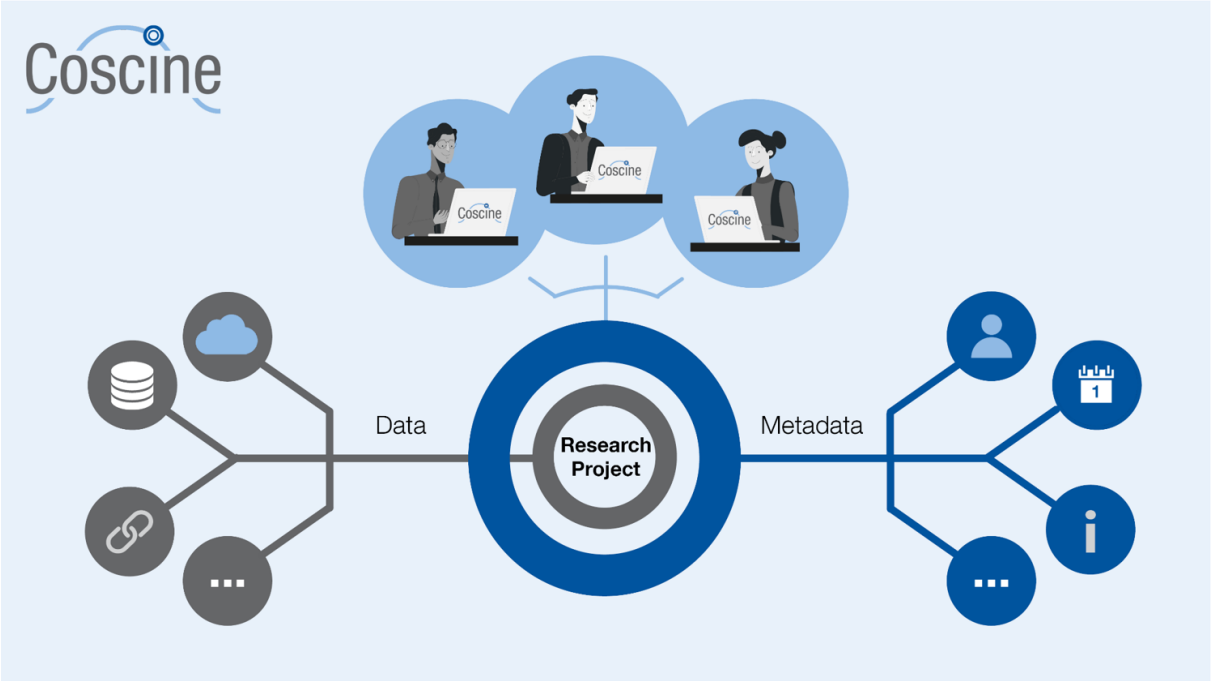


Figure 1: Data and metadata in Coscine.

Depending on the project and the data, the required amount of storage capacity varies strongly. Furthermore, the knowledge on research data management is also individual, and therefore research data is handled, organized, and annotated differently. To support researchers in their needs and at the same time ensure that the data and the corresponding metadata is managed correctly, the research data management (RDM) team not only consults and offers workshops, and when it comes to the request for storage capacity we created a digital process that guides researchers through the steps of describing their research project and how they intend to manage the corresponding research data. Hence, the carrots and sticks metaphor: rewarding good RDM practices with access to data storage systems. The required application process is conducted by the use of JARDS. JARDS is already used in the context of applying for computer time at various high performance computing (HPC) systems in Germany.

2 Workflow

As a prerequisite, researchers need to create a project in Coscine. At this point they already have to provide certain meta information about the project (description, time frame, collaborating people and institutions etc.). Having collected the meta information for a research project grants a limited default storage quota that can be used directly. When it comes to extend this default storage capacity for a project, researchers can follow a science led application process close to the peer review of a research contribution. Researchers will find a documentation on the pages of Coscine how to proceed[1].

In the following sections, we will describe the workflow of the required steps for the application and review process. After describing the preparation and submission steps, we will explain the formal evaluation and as well technical as scientific review of the application. Further on, we talk about the resource allocation and monitoring steps and how the storage capacity is included in the reporting.

2.1 Project preparation and proposal submission

In Coscine researchers can create various forms of resources in which the data is stored. The resource types not only differentiate in how data is mainly uploaded and annotated by metadata (RDS-Web: via a web interface or a custom Application Programming Interface (API) that enforces metadata quality, RDS-S3: via the widely used S3 protocol) but also on the persistent integrity of once uploaded data (RDS-WORM: write-once-read-many-storage that does not allow changes once a file is stored).

In JARDS the different resource types are represented since they require different information from the researcher. For resource types that ensure correct handling of metadata and other good RDM practices the form is simpler (especially in the case of RDS-Web), the more specific the requirements of the researchers are the more information they must provide. The most information currently is required for the resource type RDS-WORM, since incorrect use of the resource will block valuable storage space for 10 or more years.

After researchers have identified a resource type that matches their requirements based on the flow diagram that is shown in Figure 2, the application process is initiated. Through the system, researchers can file an arbitrary amount of storage applications for one or multiple projects within Coscine, as they assume being appropriate for their scientific workflow. JARDS offers an overview of the current status of these applications (cf. Figure 3). For getting some context about the project and contact information, a very first question demands the project title, description and PI or PC (cf. Figure 4).

The application workflow ensures that on the one hand the application provides scientific value and on the other hand that there is at least a basic data management plan (DMP). As such, important questions are what kind of data is created or processed within this project. Are there any special requirements for the data that need to be considered, e.g.,

1 `https://docs.coscine.de/en/projects/storage/`; *Last accessed on May 15th, 2023.*

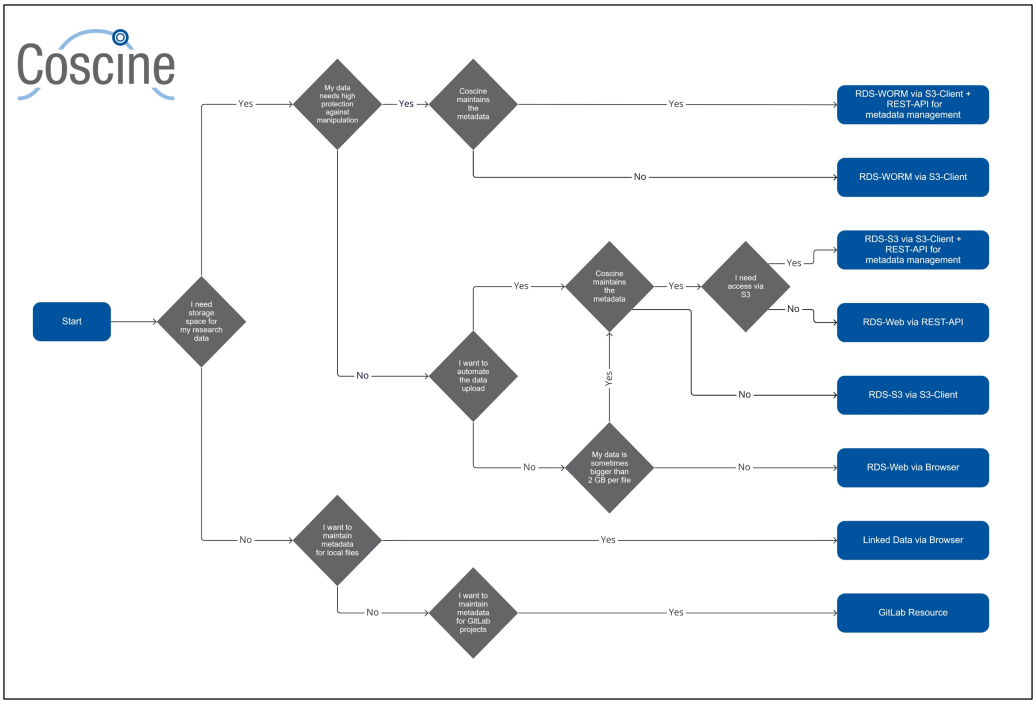


Figure 2: Application selection.

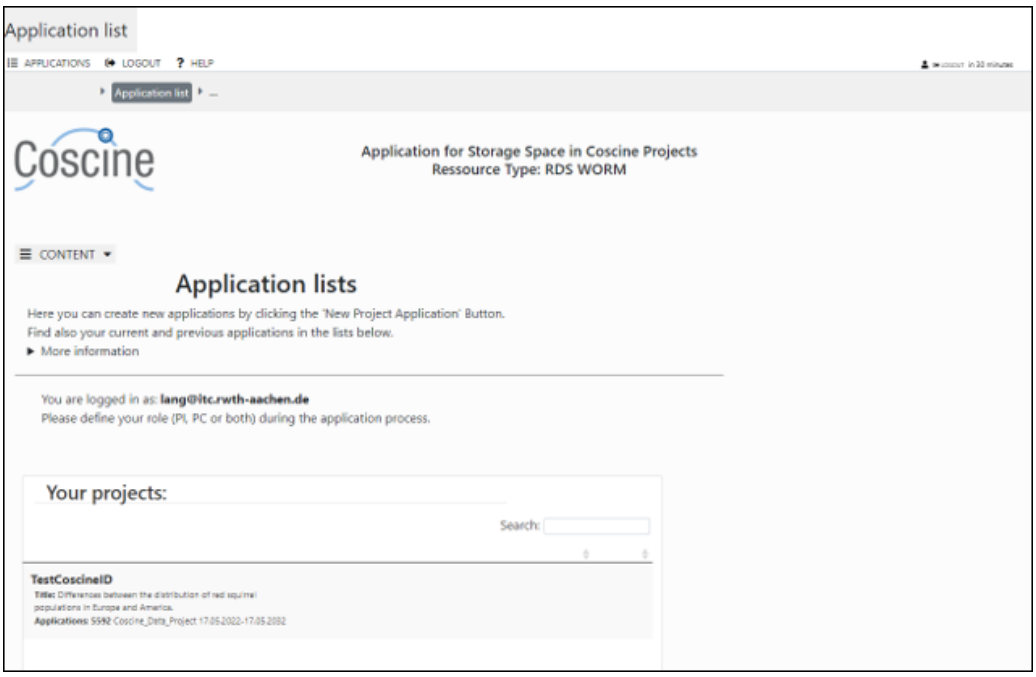


Figure 3: Applying for storage space in JARDS: Application Creation.

data privacy concerns, ensuring the raw data cannot be altered or the usage of distributed data sources. A core part is about the internal structure of the research data and how researchers intend to handle or organize it.

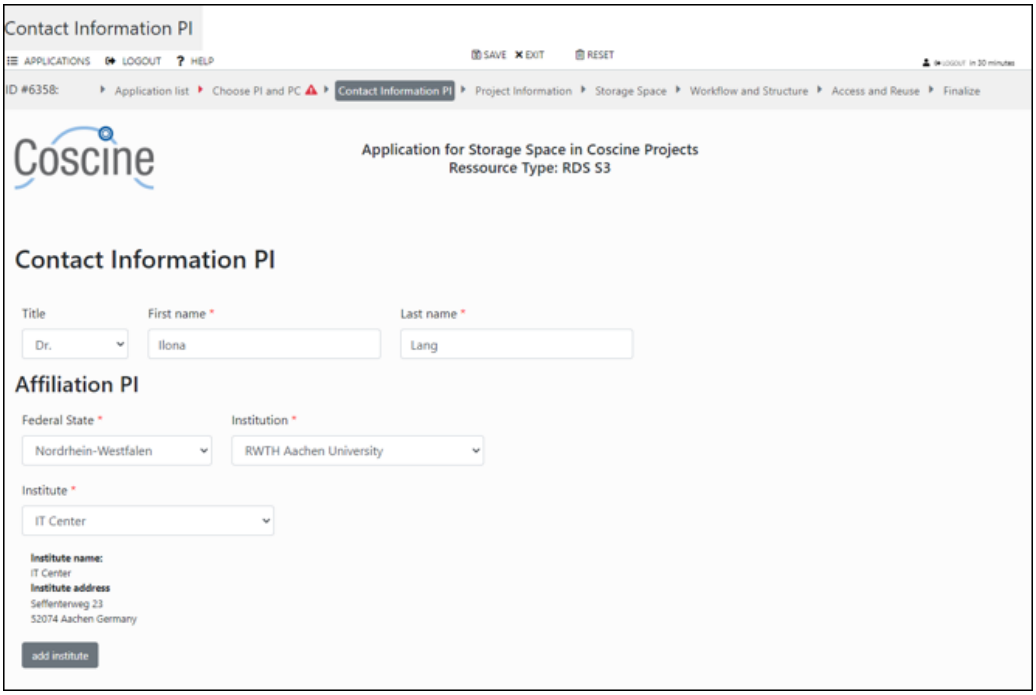


Figure 4: Applying for storage space in JARDS: PI Information.

The crucial question is about the amount of storage capacity. The default quota for an RDS-Web resource is 100 GB. For the resource type RDS-S3 and RDS-WORM, there is no default quota. Researchers can name any figure, but it needs to be plausible in respect of the described project and handling of data.

All the given information are part of the technical, and in case of the request of more than 125 TB storage capacity, also scientific review.

2.2 Formal evaluation, technical and scientific review

After the researcher has submitted the formal request for storage capacity by filling out all required fields regarding the project and data handling, the review process is initiated. As a first step, the proposal is formally evaluated: This means that it is checked first, if the applicant is eligible to request a storage capacity and second, whether the answers are complete and contain all needed information. This step is conducted by members of the universities' RDM team, and the formal evaluation typically takes between one or two days. Once the evaluation is done, in the next stage the technical and scientific review is performed (cf. Figure 5). Within the technical review, staff of the local RDM team will review the application for technical feasibility with special focus on the proposed data and metadata management. In case of problems or questions, the principal investigator (PI) and/or the person of contact (PC) of the project are contacted to provide the missing information or to adjust the plan to ensure good research data management practices. This is roughly equivalent to a data management plan review. Usually, this step takes about one week.

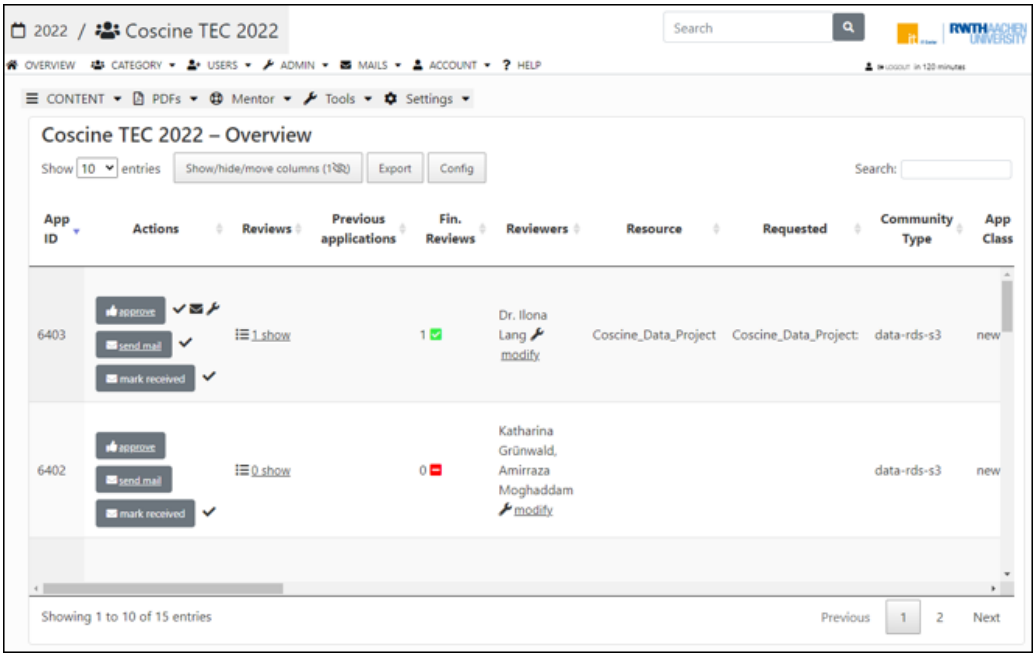


Figure 5: Review component.

Based on the amount of storage space that is requested, the review process can be extended with a third step: the scientific review. When researchers require more than 125 TB, a single-blind review of the project application by up to three independent domain scientists from German universities or other research facilities is performed. These domain scientists can suggest adjustments to both the envisioned process and the requested storage space. Because of these external dependencies, this process takes between four and six weeks for applications for RDS-Web and up to three months for applications for RDS-S3 and RDS-WORM.

2.3 Resource allocation and monitoring

After the review process is completed, the requests will be either rejected or approved. In the latter case, quota will be granted. In case special configurations were requested, a training or counselling is offered to the applicant to ensure correct usage of the system. This approach offers a unique possibility for the universities' RDM team to get into contact with the heavy users of data storage infrastructures and to increase digital literacy and competences in a targeted manner.

When the review process is finalized, the application is approved, the requested resources are assigned within Coscine to the project of the applicant. Since the review process can take longer for larger applications, a preliminary initial quota can be provided for certain categories. This enables the researches to set up their workflows with the storage systems while waiting for the final review of their application. After the application was approved, the quota will be extended and the size of the initially created resources can be easily adjusted within Coscine (cf. Figure 6). The PI/PC can add further users to their projects

within Coscine at any time and can also monitor their available and utilized quota at any time.

In case the originally requested resources are not sufficient, an extension can be requested. The original application can be used as a base for the application for an extension, which will then be reviewed as described above. JARDS also provides an additional option: if small amounts of additional quota are required, a project can be extended once to grant an additional 25 % of the original quota. This small extension does not require a complete review process.

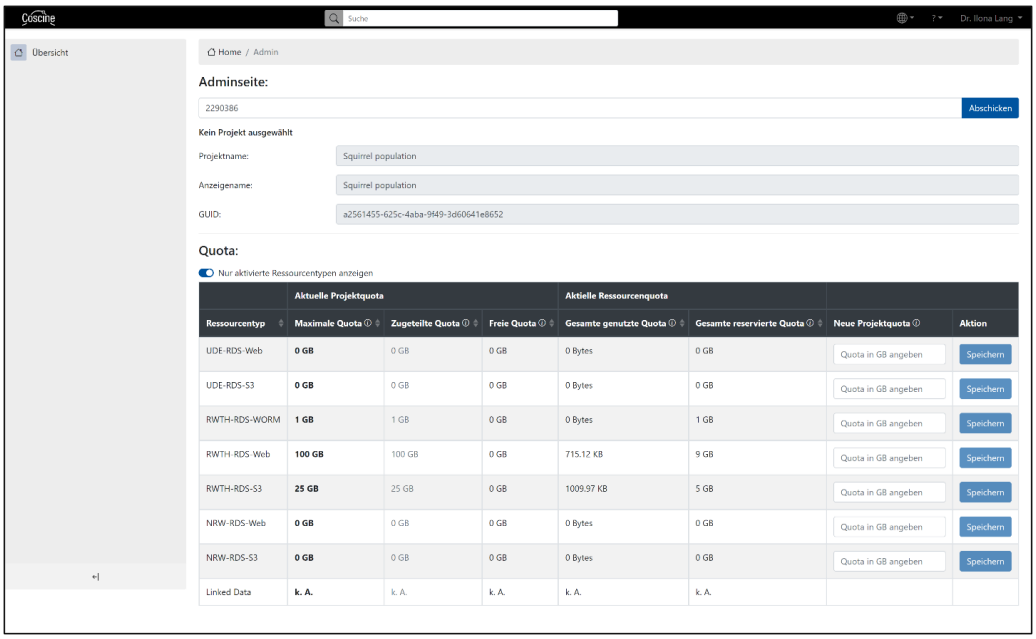


Figure 6: Quota management.

2.4 Reporting

JARDS also provides the option for the RDM staff to manage existing projects within the project component. Within this component the users can see all their approved projects, and the granted resources. There is an option for system operators to automatically report the amount of utilized resources, so the PI can monitor the still available resources and request more storage space if required. The component also offers different options for operators and managers of the storage system, e.g. there are regular status reports and a final report can be requested from the researchers. The researchers are contacted through mail and can upload these reports within JARDS. In addition, publications that were created as parts of the research project can be entered within the component as well.

3 Conclusion and outlook

The workflow presented in the previous section can easily be extended to include different resource types. This can be other storage systems, computation time on a high-performance computing system, or any other IT resource. Through the science led review process, all applicants are treated equally throughout the entire process. This not only ensures a quality standard but could also enable comparability between different applications, in case a strongly limited resource is managed with the system. The system is scalable and the number of operators and reviewers for each resource can be adjusted according to the requests. Additionally, this allows the allocation and provision of statewide available storage resources, such as RDS, according to uniform criteria by the science led management concept within national service offerings like Coscine.nrw[2]. In addition to management, this supports the storage of research data according to the FAIR principles. This improves participation opportunities of smaller universities in these scientific (storage) infrastructures and thus increases the economic efficiency of the invested resources in the long term.

The presented approach forces researchers to think about their data and the corresponding metadata from the start of the project. It also provides a unique opportunity for the universities' RDM team to reach out to heavy data users and supply them with targeted information about the systems used, or to build tailored offers to enhance digital literacy. The process has several similarities to the submission and review of scientific papers, and therefore is familiar to the researchers. Another advantage is that many researchers are already familiar with the utilized software and its functions, since they use the same software to apply for computing time projects on many HPC clusters in Germany. This allows an easier adaptation of the software for the researchers and can give HPC centers the possibility to combine applications for computing time projects and data projects.

Acknowledgements

The work was partially supported with resources granted by NFDI4Ing, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 442146713, NFDI-MatWerk, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 460247524, and FAIR Data Spaces, funded by the German Federal Ministry of Education and Research (BMBF) under funding reference FAIRDS11.

References

Deutsche Forschungsgemeinschaft e.V. 2019. *Guidelines for Safeguarding Good Research Practice: Code of Conduct.* Bonn, Germany. https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp_en.pdf.

2 https://www.dh.nrw/kooperationen/Coscine.nrw-100; *Last accessed on May 15th, 2023.*

Eifert, Thomas, Florian Claus, and Ania Lopez. 2018. *Research Data Storage (RDS): Verteilte Speicherinfrastruktur für Forschungsdatenmanagement: Gemeinsamer Antrag (öffentliche Fassung) im DFG-Programm "Großgeräte der Länder": RWTH Aachen University (Konsortialführer), Fachhochschule Aachen, Ruhr-Universität Bochum, Technische Universität Dortmund, Universität Duisburg-Essen, Universität zu Köln.* Technical report. DOI: https://doi.org/10.18154/RWTH-2021-04541.

Janetzko, Florian. 2019. "JARDS Ein Softwarewerkzeug zur Handhabung von Ressourcenvergabeprozessen". In *ZKI-AK Supercomputing Herbsttagung.* https://juser.fz-juelich.de/record/868324.

Kálmán, Tibor, Daniel Kurzawe, and Ulrich Schwardmann. 2012. "European Persistent Identifier Consortium - PIDs für die Wissenschaft". In *Langzeitarchivierung von Forschungsdaten: Standards und disziplinspezifische Lösungen,* edited by Reinhard Altenhöner and Claudia Oellers, pages 151–164. Berlin, Germany: Scivero Verl. ISBN: 978-3-944417-00-4.

Krämer, Florian, Marius Politze, and Dominik Schmitz. 2016. *Empowering the Usage of Persistent Identifiers (PID) in Local Research Processes by Providing a Service and Integration Infrastructure.* In collaboration with RD Alliance. Garching, Germany.

Maali, Fadi, and John Erickson, editors. 2014. *Data Catalog Vocabulary (DCAT).* W3C. Visited on June 10, 2018. http://www.w3.org/TR/vocab-dcat/.

Politze, Marius, Florian Claus, Bela Darius Brenger, Mohammad Amin Yazdi, Benedikt Paul Anton Heinrichs, and Annett Schwarz. 2020. "How to Manage IT Resources in Research Projects? Towards a Collaborative Scientific Integration Environment". *European journal of higher education IT* 1 (2020/1): 5. ISSN: 2519-1764. DOI: https://doi.org/10.18154/RWTH-2020-11948. https://publications.rwth-aachen.de/record/808269.

Schmitz, Dominik, and Marius Politze. 2018. "Forschungsdaten managen – Bausteine für eine dezentrale, forschungsnahe Unterstützung". *o-bib. Das offene Bibliotheksjournal* 5 (3): 76–91. DOI: https://doi.org/10.5282/o-bib/2018H3S76-91.

Smedt, Koenraad de, Dimitris Koureas, and Peter Wittenburg. 2020. "FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units". PII: publications8020021, *Publications* 8 (2): 21. DOI: https://doi.org/10.3390/publications8020021.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Cat4KIT: A Cross-institutional Data Catalog Framework for the FAIRification of Environmental Research Data

Mostafa Hadizadeh[1], Christof Lorenz[1], Sabine Barthlott[1], Romy Fösig[1], Uğur Çayoğlu[2], Robert Ulrich[3], Felix Bach[4]

[1]Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology;
[2]Steinbuch Centre for Computing, Karlsruhe Institute of Technology;
[3]KIT Library, Karlsruhe Institute of Technology;
[4]Leibniz Institute for Information Infrastructure, FIZ Karlsruhe

A contemporary and flexible Research Data Management (RDM) framework is required to make environmental research data Findable, Accessible, Interoperable, and Reusable (FAIR) and, thus, provide the foundation for open and reproducible earth system sciences. While datasets accompanying scientific articles are typically published through large data repositories such as Pangaea, Zenodo, or RADAR4KIT, intermediate, day-to-day, or actively used data is often still exchanged through simple cloud storage services and email. However, despite the FAIR principles emphasizing the need for openly findable and accessible data, it is often confined to closed and restricted infrastructures and local file systems.

Therefore, our research project, Cat4KIT, aims to develop a cross-institutional catalog and RDM framework to FAIRify such day-to-day research data. The framework consists of four modules with different tasks: (1) providing access to data on storage systems through well-defined and standardized interfaces, (2) harvesting and transforming (meta)data into consistent and standardized formats, (3) making (meta)data publicly accessible using well-defined and standardized catalog services and interfaces, and (4) enabling users to search, filter, and explore data from decentralized research data infrastructures. Each module is developed and implemented within an inter-institutional consortium comprising scientists, software developers, and potential end-users. This approach ensures that our framework is applicable to a wide range of research data, from multi-dimensional climate model outputs to high-frequency in-situ measurements.

We place emphasis on the application of existing open-source solutions and community standards for data interfaces, (meta)data schemes, and catalog services such as the Spatio-Temporal Assets Catalog (STAC). This approach ensures easy integration of research data

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.c18072

into the Cat4KIT framework and facilitates straightforward extension to other research data infrastructures.

1 Introduction

Nowadays, numerous real-world applications generate voluminous quantities of precise and ambiguous data from a broad variety of rich data sources at a rapid rate, and utilizing big data heralds the beginning of a new era of rising production (Hampton et al. 2013). This holds particularly true for the environmental sciences, where advances in modeling, *in situ* observation, and remote sensing systems, as well as the rapid growth of applications in the field of citizen science, have led to a massive increase in the number and volume of environmental data (e.g., Buytaert et al. 2012). And collaborative projects with partners spread across the world, as well as the increasing attention to such information from non-scientific communities, require this data to be remotely available and accessible via standardized and well-communicated interfaces. All this is aggravated by the release of the FAIR-principles (Wilkinson et al. 2016), which are becoming more and more mandatory in research projects or publications. In particular, the enrichment of data with consistent and well-defined metadata and unique indexes, as well as the public provision of this information via standardized interfaces, adds another level of complexity to modern research data management (RDM).

While there are initiatives that aim at the *FAIRification* of research data (e.g., Jacobsen et al. 2020; Kersloot et al. 2022), the usual way is still to publish data via dedicated data repositories. However, despite the need for such open and freely accessible data as well as the recognition that we urgently need to develop concepts for rewarding the publication of data (e.g., Pierce et al. 2019), this task is still assumed to be an additional (and often cumbersome) step at the end of a study or research project. And even if data is made publicly available, its exploitation is often rather limited: it is described and presented with properties that are relevant for domain experts (data producers) but that are not properly understood and reusable by other scientific communities (Annane et al. 2022) or the metadata is limited to a generic minimum description without crucial information and guidelines for proper usage of the underlying data. Furthermore, most modern domain-specific repositories like, e.g., PANGAEA[1] (Diepenbroek et al. 2002) or the World Data Center for Climate[2] (WDCC) at the German Climate Computing Center (DKRZ) or their generic counterparts like, e.g., ZENODO (European Organization For Nuclear Research and OpenAIRE 2013) or RADAR4KIT[3], only allow for the download of full datasets instead of allowing users to interact with the data, e.g., for subsetting or visualization. On the contrary, currently, it is common practice to transfer intermediate, day-to-day, or regularly accessed data through readily available cloud storage services and email, rather than utilizing a dedicated data portal or metadata service for the purpose

1 `http://www.pangaea.de`
2 `https://www.wdc-climate.de`
3 `https://radar.kit.edu`

of sharing, filtering, and exploring data from the Institute of Meteorology and Climate Research (IMK) at the Karlsruhe Institute of Technology (KIT).

Environmental scientists and data producers are hence facing a severe challenge: while the need for FAIR and collaborative research data is increasing, there are only limited frameworks and tools that help to make particularly intermediate and day-to-day data openly available and accessible according to the FAIR principles.

Within the research project Cat4KIT, we hence want to develop a catalog framework, that allows for a simple and straightforward provision of environmental research data. This is achieved by the development and implementation of four independent but interlinked modules:

- Dedicated data services take research data from existing storage systems at the KIT and make this data remotely accessible via standard interfaces
- A (meta)data harvester that involves systematically scanning data services and collecting metadata attributes in a standardized manner
- A dedicated catalog service allows for the interaction (search, filter, and modify) of the collected metadata
- A portal service aims at the user-friendly presentation of the collected metadata

An overview of these different modules and how they are interlinked is presented in Figure 1.

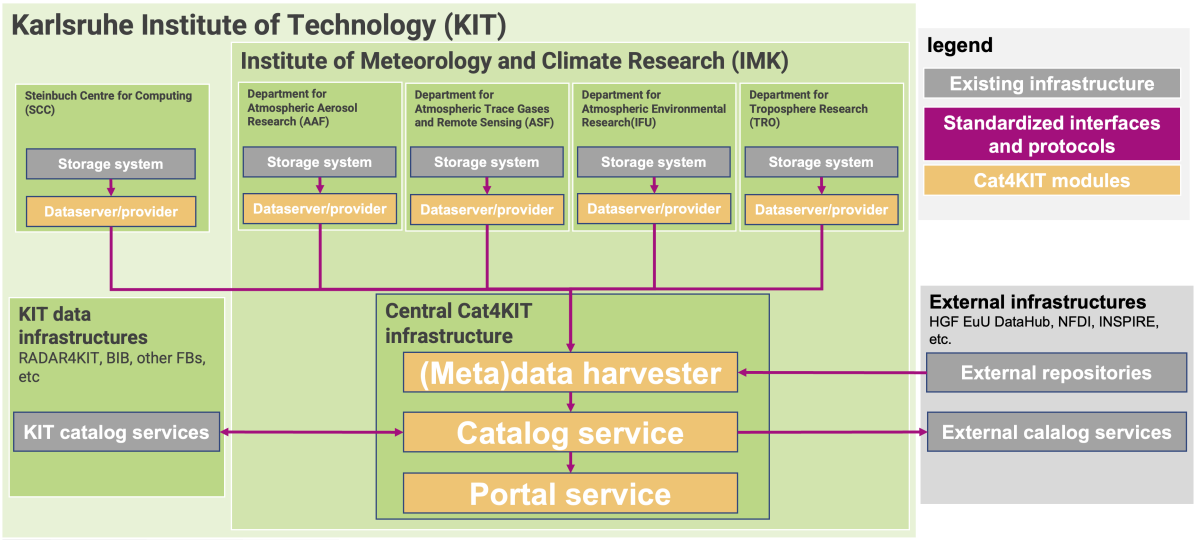


Figure 1: Schematic representation of the components in Cat4KIT project.

Funded by the *Exzellenzuniversitätsvorhaben Forschungsdatenmanagement* of the Karlsruhe Institute of Technology (KIT), The collaborative endeavor encompasses the cooperation of four KITs departments within the Institute of Meteorology and Climate Research, the Steinbuch Centre for Computing, and the KIT Library. In order to keep the entry

barrier as low as possible, Cat4KIT is hence build upon existing storage systems and infrastructure components at the participating institutes so that data procurers can use our framework without changing their established workflows. We further focus on widely used interfaces and community standards to ensure a simple and straightforward linkage to other catalog infrastructures, both within the KIT and with external repositories and catalog services (see Figure 1).

2 Components of the Cat4KIT framework

Our Cat4KIT-framework is based on a software stack that consists of existing open-source tools as well as complementing in-house developments, particularly for harvesting and ingesting the metadata from the different data sources. Each of these modules will be based on one (or multiple) interlinked Docker-containers which simplify the implementation of (sub)modules of Cat4KIT in other infrastructures. In the following, we will discuss each of the modules and tools in more detail.

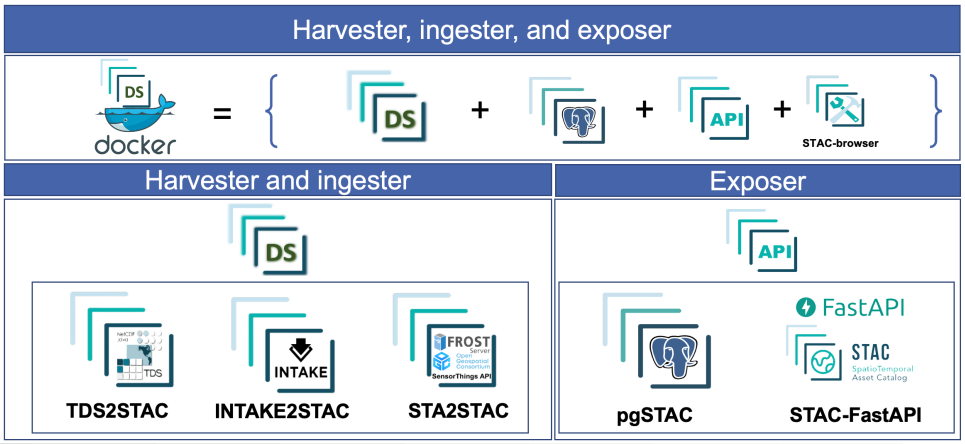


Figure 2: Components of the harvester, ingester, and exposer module.

2.1 Data service/provider

The first task of Cat4KIT is to make data on common and existing storage systems externally available. We currently focus on storage systems that are heavily used by the participating institutes and that is particularly the Large Scale Data Facility (LSDF) at KITs Steinbuch Centre for Computing as well as the BeeGFS (ThinkParQ GmbH 2023) and S3-compliant object storage implementations at Department for Atmospheric Environmental Research at Institute of Meteorology and Climate Research (IMK-IFU). In doing so, we want to ensure that a) we include a wide range of existing data into our Cat4KIT framework and b) users can easily integrate their data without changing their existing workflows.

However, as data from environmental sciences is usually highly heterogeneous, we need to apply dedicated services for different data types. In Cat4KIT, we hence distinguish

between multi-dimensional (e.g., from remote sensing or modeling systems) and one-dimensional data (e.g., from environmental sensor systems). For most of our multi-dimensional data, we apply a THREDDS-Data-Server (TDS; Domenico et al. 2002) while selected (high-volume) datasets (e.g., from climate models or high-resolution remote sensing systems) are also made available via so-called Intake-Catalogues[4]. Access to and interaction with one-dimensional data is realized via the Open Geospatial Consortium (OGC) SensorThings API (Liang, Huang, and Khalafbeigi 2016), which is provided by the so-called FROST Dataserver[5].

THREDDS Developed by the unidata community[6], THREDDS is a tailor-made data server for publishing (Network Common Data Format (NetCDF) data via various interfaces. As NetCDF is the quasi-standard many domains of environmental sciences, THREDDS-Server is widely used in the community for providing access to research data. Right now, we include data and catalogs from two TDS-instances at IMK-IFU[7] and Department for Atmospheric Trace Gases and Remote Sensing at Institute of Meteorology and Climate Research(IMK-ASF)[8]. But it is planned to implement further THREDDS servers both within but also outside of KIT.

Within TDS, data is organized in *catalogues*. These catalogs support both the static linking of datasets and their dynamic creation via so-called *DatasetScans*. TDS also features a wide range of interfaces with which users can interact with the data. Here, we only present some of the interfaces that are relevant within the Cat4KIT-framework:

- The Open-source Project for a Network Data Access Protocol (OPeNDAP)[9] allows for simple remote access to the data in the NetCDF-files. There are various libraries for most programming languages and environments that support data access via OpenDAP. Hence, this interface is used more and more to realize, e.g., workflows with decentralized data storage.

- The Web Mapping Service[10] (WMS) from the OGC allows for the server-side visualization of data. Due to its long history and ease of use, it is the quasi-standard for the visualization of geospatial raster data.

- ncISO [11] allows for the construction of ISO 19115[12] conformal metadata from NetCDF-files. Being the metadata standard for various communities, ISO 19115 features a wide range of attributes and information that (usually) has to be collected manually. Via ncISO, however, this information can be extracted automat-

4 https://github.com/intake/intake
5 https://github.com/FraunhoferIOSB/FROST-Server
6 https://www.unidata.ucar.edu
7 https://thredds.imk-ifu.kit.edu/thredds/catalog.html
8 https://thredds.imk-asf.kit.edu
9 https://www.opendap.org
10 https://www.ogc.org/standard/wms
11 https://artifacts.unidata.ucar.edu/service/rest/repository/browse/unidata-all/EDS/nciso
12 https://www.iso.org/standard/53798.html

ically from the NetCDFs which greatly simplifies the construction of standardized metadata.

Despite all these features, TDS currently does not include catalog services like, e.g., OGCs Catalog Service for the Web (CSW) or OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). Hence, if we want to implement searching or filtering capabilities, we need to harvest the (meta)data from our THREDDS-servers into dedicated catalog frameworks.

Intake Intake is a lightweight Python package that allows to load data from a variety of formats and sources into well-known containers such as Pandas dataframes (The pandas development team 2020) or Xarray DataSets (Hoyer and Hamman 2017), etc. Data is usually collected in catalogs and each item within such a catalog contains all information that is required to interact with the data. It hence removes the need for a potential end-user to know about the exact storage location/system or data format. Moreover, Intake allows for the generation of simple catalogs with data of different formats and across different institutions. It hence can be a crucial element in the development of reproducible workflows with data on decentralized storage systems. In Cat4KIT, we apply such Intake catalogs manly for high-volume model and remote sensing data that typically lies on cloud-optimised storage systems like, e.g., S3-conformal object storage.

FROST The Fraunhofer Open Source SensorThings API Server (FROST) is the reference implementation of OGCs SensorThings API (STA). With its roots in the Internet of Things(IoT), STA is tailor-made for one-dimensional data from typical sensor systems. At its core, the STA consists of an interface (a set of commands to interact with the (meta)data) and a dedicated datamodel. This datamodel is constructed around *Things*, which could be, according to STA, "anything in the physical or information world that can be uniquely identified and integrated into communication networks". It further provides entities for *Location* (directly attached to a *Thing*), *Sensor*, and *ObservedProperty*. When combining a *Thing* with a *Sensor* and an *Observed Property*, we obtain a so-called *DataStream*, which defines the container into which data is written. In Cat4KIT, we apply FROST-Servers mainly for providing access to one-dimenisonal data from various observation and monitoring systems that are operated at the IMKs. But as the implementation of FROST-Servers and the SensorThings API has just started, we only include data from the instance at IMK-IFU[13], that, so far, only holds some preliminary data.

2.2 Metadata catalog framework

As our catalog framework, we apply the SpatioTemporal Assets Catalog (STAC)[14], that has its origin in the collaboration of satellite imagery providers. In the last years, the user base and community around STAC has increased substantially and it is now applied by a wide range of data providers as their main catalog framework. There has also grown a

13 `https://sensorthings.imk-ifu.kit.edu`
14 `https://stacspec.org`

large ecosystem[15] around STAC with extensions and plugins for all kinds of applications ranging from visualization solutions for geospatial data over dedicated databases for STAC to client libraries for major programming languages. Today, STAC describes itself as "common language to describe geospatial information, so it can more easily be worked with, indexed, and discovered".[16]. In that sense, it is a tailormade framework for realizing a catalog and data exploration infrastructure with a minimal set of mandatory metadata. Because at its core, STAC is build around Items that are simple GeoJSONs with all necessary information about a dataset. In general, a STAC Item only requires a title, some information about the bounding box as well as some temporal information. And this information can easily be retrieved from all data servers within Cat4KIT (see section 2.1) so that we can easily generate a consistent set of STAC Items. Multiple STAC Items, that share properties and metadata, can be combined in STAC Collections. And finally, a STAC Catalog is constructed of one or multiple Collections and (Sub)Catalogs.

As many of the datasets included in Cat4KIT are typical raster datasets with multiple variables, we further make use of the DataCube[17] extension. This extension allows to automatically retrieve and add information about variables and dimensions within a dataset. This, again, reduces the need for manual metadata curation of the final STAC Items.

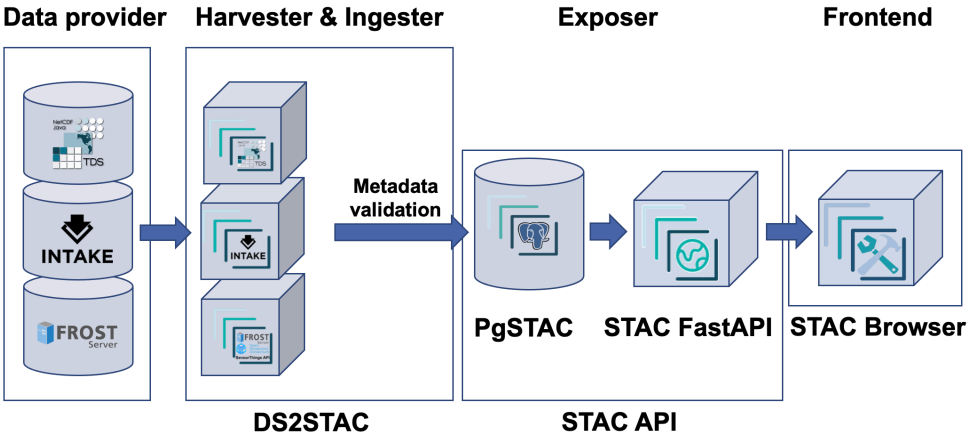


Figure 3: Workflow of the Cat4KIT system.

2.3 Harvester, ingester and exposer

Once the data is available via different data server (see section 2.1, the next step is to retrieve or *harvest* the respective metadata and ingest it into a consistent STAC database. This step is done by the *harvester/ingester* while the *exposer* is responsible for making the catalog data accessible from the outside (see Figure 2).

15 `https://stacindex.org/ecosystem`
16 `https://stacspec.org/en`
17 `https://github.com/stac-extensions/datacube`

For the harvester and ingester, we have developed the so-called *DS2STAC*[18] (Data Servers/-Services to STAC metadata catalog) package. DS2STAC includes three tailored subpackages for scanning and harvesting datasets. Each of these packages retrieves the geospatial and temporal information that is necessary for creating STAC items:

- TDS2STAC[19] (harvester for THREDDS Server): STAC Items via THREDDS datasets, coordinates, variables and temporal information via ncISO and WMS (depending on availability)
- INTAKE2STAC[20] (harvesrer for Intake catalogs): STAC Items via Intake catalog entries, coordinates, variables and temporal information via STAC DataCube extension and Xarray
- STA2STAC[21] (harvester for OGC SensorThings API): STAC Items via *Things*, coordinates via *Location*-entities, temporal information via *pheonomenonTime* in the *Datastream*-entities

In its current implementation, DS2STAC is run manually but it is planned to add a scheduler so that we can harvest from each data source, e.g., at pre-defined times or when new data has been added or modified. Once new metadata is retrieved, it is first passed through a *Validator*[22] that checks the data for the availability of all mandatory elements (temporal and spatial coordinates, general description, description of available data services). If this validation is passed, the new entry is moved forward to the *Exposer*-module (see below). If not, it is envisaged that the data provider receives a notification with details about missing or erroneous elements.

For TDS2STAC, the generation of STAC Collections and Catalogs (see section 2.2) is triggered automatically and resembles the structure of the underlying THREDDS catalogues. For STA2STAC and INTAKE2STAC, this is still a manual process that needs some further refinements and hierarchies directly at the data source.

In the next step, we are going to develop tailored harvesters for other data sources and repositories. Furthermore, the harvester for the SensorThings API is still work in progress, as we are currently developing specific STA metadata profiles and data models for environmental sciences. As these profiles will also feature new entities particularly for a more detailed grouping of STA-*Things* (e.g., via projects or networks), they will allow for, e.g, the direct harvesting and mapping into STAC Collections and Catalogs.

The *exposer* is based on the STAC API[23], which is implemented via the STAC FastAPI[24] with the *pgSTAC*[25] storage backend. We further plan to implement data within our

18 https://codebase.helmholtz.cloud/cat4kit/ds2stac
19 https://tds2stac.readthedocs.io
20 https://intake2stac.readthedocs.io
21 https://sta2stac.readthedocs.io
22 https://stac-validator.readthedocs.io
23 https://github.com/radiantearth/stac-api-spec
24 https://github.com/stac-utils/stac-fastapi
25 https://github.com/stac-utils/pgstac

Cat4KIT infrastructure into other higher-level repositories and portals. Hence, as a demonstrator, we currently develop a dedicated STAC exposer for the so-called *Earth Data Portal*[26] of the Helmholtz Research Field Earth and Environment[27]. The STAC API furhter allows, due to the growing ecosystem and user basis of STAC, to seamlessly integrate our Cat4KIT catalogues into other third-party software like QGIS[28].

2.4 Data portal and frontend

Finally, once metadata is available via the *exposer*, it is presented in a data portal and graphical user interface that based on the STAC-Browser[29]. Here, a common site for a STAC Item allows to present information about the geographic location on a map, the temporal span as well as further available metadata of the underlying dataset. As the latest version of the STAC-Browser is fully consistent with the STAC API, it further allows for a direct searching and filtering of the catalogs.

3 Working with Cat4KIT - a first concept

One of the key objectives of our Cat4KIT framework is to make the integration of data as simple as possible. Thus, data producers should be able to stick to their established workflows while taking advantage from the external accessibility and interactivity. A typical algorithm using Cat4KIT, hence, should look like

1. User/data provider stores data on an integrated storage system (raster data as CF-conformal NetCDF, one-dimensional data in database with STA interface)
2. Only for THREDDS and Intake: Cat4KIT-admin adds dataset to THREDDS- or Intake-catalogues
3. Automatic harvesting of metadata (esp. spatial and temporal coordinates as well as data services) from newly created entries
4. Automatic generation of STAC items and integration into STAC database
5. Automatic exposing of newly generated STAC items via STAC API
6. Presentation of harvested information, dimensions and variables as well as available data services in STAC Browser

Users can now find/search/explore the newly integrated data in the STAC Browser. The respective description also includes links to available data services like WMS, OpeNDAP, or STA and hence allow for direct and tailored access to the data, e.g., for visualization, extraction or analysis.

26 `https://earth-data.de`
27 `https://www.helmholtz.de/en/research/research-fields/earth-and-environment`
28 `https://stac-utils.github.io/qgis-stac-plugin`
29 `https://github.com/radiantearth/stac-browser`

4 Conclusions and Outlook

In this paper, we present our current concept for a catalog framework, that should help researchers and data providers to make their data FAIR with a focus on findability and accessibility. Once fully operational, it will allow for an easy integration of (meta)data by automatic harvesting from several data sources. By that, there is no need to change existing and established workflows; instead, data providers will benefit from the added accessibility and interactivity via standard interfaces and data services as well as the straightforward integration into higher-level data and catalog infrastructures via the STAC API. This is further supported by the heavy usage of open-source tools and interfaces that are widely applied in the scientific community. In particular, if data is available via one of the (currently) three supported data sources (THREDDS, Intake, SensorThings API), an integration is straightforward and does not need any further manual configuration.

One key aspect of Cat4KIT is the integration of (meta)data from highly decentralized storage and infrastructure systems within the four departments of the Institute of Meteorology and Climate Research at the Karlsruhe Institute of Technology, as well as the Steinbuch Centre for Computing and the KIT Library. This should also allow for a simple and straightforward integration of other internal and external repositories and data sources. In order to timely consider this integration, we are already in contact with interested parties and potential users from external institutions.

Currently, we are in the testing-phase of the DS2STAC-module, which contains tailored harvesters for each of the three data sources. While the harvesting, in general, is working as supposed, particularly the division into STAC Catalogs and Collections still requires some further refinements and conventions (like, e.g., the concrete meaning of Collections and Catalogs for different use cases). It is important to acknowledge that following the launch of the initial version, the users' perspectives on infrastructure improvements will be gathered through the feedback system in Cat4KIT infrastructure. Subsequently, these perspectives will be reviewed and incorporated into future developments.

Overall, it is planned that a first running version of the Cat4KIT-framework is available from Q4 2023. During the coming months, we will hence focus on the linkage of the different modules, the implementation of further data sources as well as continuously enhancing the number of integrated datasets.

After its launch, Cat4KIT will be the central entry point for searching, filtering and exploring data from the Institute of Meteorology and Climate Research at the KIT. It will hence provide a substantial contribution towards the FAIRification of research data and further foster an open research from environmental sciences.

Acknowledgements

The Cat4KIT project has been funded by the *Exzellenzuniversitäts-Vorhaben Research Data Management* of the Karlsruhe Institute of Technology, Germany. We further thank

our colleagues Dr. Philipp Sebastian Sommer and Linda Baldewein from the Helmholtz-Zentrum Hereon for many inspiring and fruitful discussions about our framework. Additionally, we would like to express our appreciation to the anonymous reviewers for their constructive feedback, which has greatly improved the quality of this paper.

References

Annane, Amina, Mouna Kamel, Cassia Trojahn, Nathalie Aussenac-Gilles, Catherine Comparot, and Christophe Baehr. 2022. "Towards the FAIRification of Meteorological Data: A Meteorological Semantic Model", 81–93. DOI: https://doi.org/10.1007/978-3-030-98876-0_7.

Buytaert, Wouter, Selene Baez, Macarena Bustamante, and Art Dewulf. 2012. "Web-Based Environmental Simulation: Bridging the Gap between Scientific Modeling and Decision-Making". *Environmental Science & Technology* 46 (4): 1971–1976. ISSN: 0013-936X. DOI: https://doi.org/10.1021/es2031278.

Diepenbroek, Michael, Hannes Grobe, Manfred Reinke, Uwe Schindler, Reiner Schlitzer, Rainer Sieger, and Gerold Wefer. 2002. "PANGAEA—an information system for environmental sciences". *Computers & Geosciences* 28 (10): 1201–1210. DOI: https://doi.org/10.1016/S0098-3004(02)00039-0.

Domenico, Ben, John Caron, Ethan Davis, Robb Kambic, and Stefano Nativi. 2002. "Thematic real-time environmental distributed data services (THREDDS): Incorporating interactive analysis tools into NSDL". Accessed: May 19, 2023. https://docs.unidata.ucar.edu/tds/5.4/userguide/index.html.

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

Hampton, Stephanie E., Carly A. Strasser, Joshua J. Tewksbury, Wendy K. Gram, Amber E. Budden, Archer L. Batcheller, Clifford S. Duke, and John H. Porter. 2013. "Big data and the future of ecology". *Frontiers in Ecology and the Environment* 11 (3): 156–162. ISSN: 1540-9295. DOI: https://doi.org/10.1890/120103.

Hoyer, Stephan, and Joe Hamman. 2017. "xarray: N-D labeled arrays and datasets in Python". *Journal of Open Research Software* 5 (1). DOI: https://doi.org/10.5334/jors.148.

Jacobsen, Annika, Rajaram Kaliyaperumal, Luiz Olavo Bonino da Silva Santos, Barend Mons, Erik Schultes, Marco Roos, and Mark Thompson. 2020. "A Generic Workflow for the Data FAIRification Process". *Data Intelligence* 2 (1-2): 56–65. ISSN: 2641-435X. DOI: https://doi.org/10.1162/dint_a_00028.

Kersloot, Martijn G., Ameen Abu-Hanna, Ronald Cornet, and Derk L. Arts. 2022. "Perceptions and behavior of clinical researchers and research support staff regarding data FAIRification". *Scientific Data* 9 (1): 241. ISSN: 2052-4463. DOI: https://doi.org/10.1038/s41597-022-01325-2.

Liang, Steve, Chih-Yuan Huang, and Tania Khalafbeigi. 2016. "OGC SensorThings API Part 1: Sensing, Version 1.0". Accessed: May 19, 2023. https://docs.ogc.org/is/18-088/18-088.html.

Pierce, Heather H., Anurupa Dev, Emily Statham, and Barbara E. Bierer. 2019. "Credit data generators for data reuse". *Nature* 570 (7759): 30–32. ISSN: 0028-0836. DOI: https://doi.org/10.1038/d41586-019-01715-4.

The pandas development team. 2020. *pandas-dev/pandas: Pandas.* Version 2.1.0. DOI: https://doi.org/10.5281/zenodo.3509134.

ThinkParQ GmbH. 2023. "BeeGFS". Visited on August 23, 2023. https://www.beegfs.io/c/.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

bwVisu: A Scalable Remote Service for Interactive Data Processing and Training for Scientists

Erik Schnetter[1], Carlo Antonio Beretta[2], Martin Baumann[1], Sabine Richling[1], Florian Heuschkel[1], Thomas Kuner[2]

[1]University Computing Centre, Heidelberg University;
[2]Department for Anatomy and Cell Biology, Heidelberg University

bwVisu is an easy-to-use, web-based platform for interactive, remote, GUI-based access to scientific work environments that require large compute and storage resources. We explain the technical architecture of bwVisu, the spectrum of available applications, and the typical way of working with the system. With use cases from biological/medical research and teaching we illustrate the advantages and user-friendliness of bwVisu. Finally, we give an outlook on further usage scenarios and possible future development of bwVisu.

1 Introduction

In the era of data sciences, researchers face various challenges that demand large storage capacities, powerful computing hardware, and interactive data exploration capabilities. However, traditional approaches often struggle to meet all these requirements, leading to issues such as lengthy data transfer times, inadequate local hardware capabilities, and limited interactivity in supercomputing environments, which typically rely on command-line skills.

To tackle these challenges, the bwVisu platform (Schridde, Baumann, and Heuveline 2017; Alpay et al. 2020) was developed as a comprehensive solution and gradually improved during the last years. bwVisu is a user-friendly, web-based platform that enables remote access to scientific work environments that require access to large compute und storage resources. It offers a collection of pre-built scientific applications designed to run on powerful hardware such as high-performance computing (HPC) systems which are typically connected to large storage systems. The bwVisu platform has a user interface that allows the selection of applications, the request of compute resources, and the direct web-based execution of graphical interactive applications. Users can interact with the deployed applications in real time through their web browser without the need for special software or hardware.

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.c18073

The bwVisu platform hosted as a service at the University Computing Centre of Heidelberg University benefits from access to bwForCluster Helix[1] and Scientific Data Storage SDS@hd[2]. bwForCluster Helix offers high-performance computing resources and SDS@hd provides a robust storage infrastructure for handling large scientific datasets.

These services are available as "Landesdienste" to researchers at universities from Baden-Württemberg and are part of Baden-Württemberg's concept for High Performance Computing, Data Intensive Computing and Large Scale Scientific Data Management, see Schneider et al. (2019). bwVisu cannot be used as persistent storage services. Instead, bwVisu conceptionally integrates storage services like SDS@hd for this purspose. SDS@hd is one of Baden-Württembergs research storage service with connections and processes for research data management, e.g. transfer data to publication platforms or archives and to foster collaboration among researchers, see Richling et al. (2022). Thus, bwVisu enhances FAIR-based[3] principles in the open sciences by further removing barriers to research data that are due to required specialized IT-skill sets and thus promoting existing platforms for open data exchange.

This paper provides a detailed description of the bwVisu platform including its features and technical architecture (Section 2) as well as the specific implementation of bwVisu as a service at Heidelberg University, the bwVisu workflow, and the availiable applications (Section 3).

In this work we focus on two primary use cases: scientific work and teaching (Section 4). In the scientific work scenario, we highlight how bwVisu addresses challenges related to image analysis, modeling, and simulation. In this case, the bwVisu platform facilitates interactive exploration and efficient utilization of computing resources. In the teaching context, we demonstrate how bwVisu enhances educational environments by providing students with access to powerful scientific tools and fostering collaborative learning experiences.

Furthermore, we engage in a discussion of the presented features, evaluating their benefits and limitations (Section 5). We also outline future developments and planned enhancements, illustrating our ongoing efforts to improve the platform and to meet evolving user requirements (Section 6).

2 bwVisu Architecture

This section explains the technical architecture of bwVisu. It provides an overview of the general structure and describes the individual components in detail. The focus is on the current implementation of bwVisu and the considerations for specific design decisions.

1 https://www.urz.uni-heidelberg.de/en/service-catalogue/high-performance-computing/bwforcluster-helix

2 https://www.urz.uni-heidelberg.de/en/service-catalogue/storage/sdshd-scientific-data-storage

3 https://www.go-fair.org/fair-principles

The technical architecture of bwVisu consists of three layers: frontend, middleware, and backend. The layers communicate via different Application Programming Interfaces (APIs) (Fielding 2000a), namely via RESTfull (Fielding 2000b) HTTP calls or backend specific API calls, see Figure 1. Each layer is discussed in more detail below.

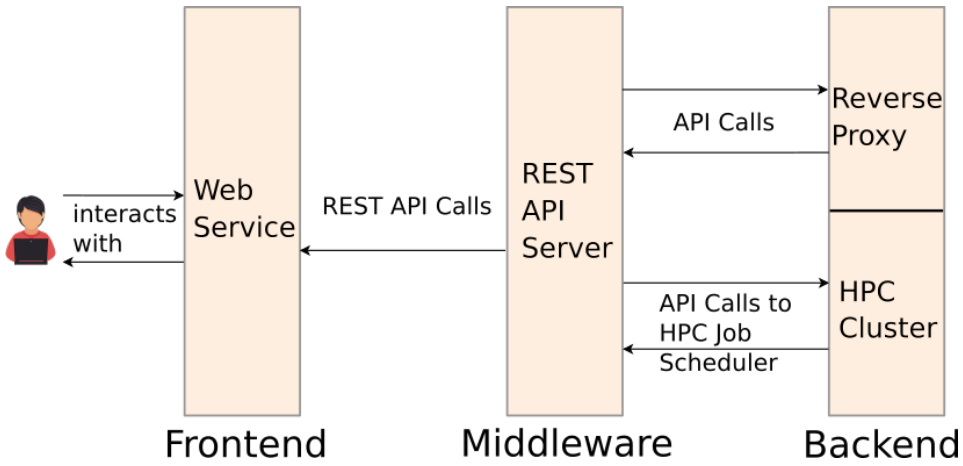


Figure 1: Technical architecture of bwVisu.

2.1 Frontend

The frontend of bwVisu is designed as a classic web application and implemented using the Python web framework Django[4]. It provides users with a web interface, including a login area. Users can start and stop applications offered by bwVisu through this interface. The frontend allows users to request dedicated backend resources for a selected application, such as the number of CPU cores and GPUs. The resource requests are communicated by the middleware to the backend. The frontend also displays links to the streaming endpoints provided by the middleware that allows users to interact with the graphical user interface of the started application in real time through their browser.

Django was chosen because of its suitability for implementing a classic web application in a small group of developers. Django has many built-in features such as user authentication or database management systems that make it possible to significantly reduce development time. Currently, authentication via the LDAP protocol is supported only, but due to the flexible authentication backend[5] of Django, different authentication methods can be easily implemented.

2.2 Middleware

The middleware is implemented as a REST API server using the Python web framework Flask[6]. Flask was chosen for its lightweight nature, which makes it an ideal choice for API

4 `https://www.djangoproject.com`
5 `https://docs.djangoproject.com/en/4.2/topics/auth/customizing`
6 `https://flask.palletsprojects.com/en/2.3.x`

servers that typically require only a few features of a full web development framework. The middleware manages the status of launched applications and provides access to them via JSON-based REST endpoints. Furthermore, it interacts with backend systems via their API interfaces to implement various features of bwVisu. The tasks of the middleware are as follows:

Launching an application: When an application is launched via the associated middleware API endpoint, a generic submit script is populated with application-specific data and submitted as a compute job to the scheduler of the backend HPC system. The submit script loads necessary software modules, starts the remote visualization tool Xpra[7], if the application cannot render its graphical output to the browser itself, and initiates the application within an isolated Singularity[8]/Apptainer[9] container which also provides access to storage systems and other resources needed by the application in said container. After that, the middleware assigns free ports from a pool of reserved ports to bwVisu and creates dynamic port forwarding rules on the reverse web proxy in the backend for the streaming endpoints opened by the application on the HPC system.

Aborting an application: When an application is aborted via the associated middleware API endpoint, the corresponding job on the HPC system is aborted, and the dynamically generated port forwarding rules on the reverse web proxy are removed.

Monitoring running jobs: The middleware regularly retrieves information about running bwVisu jobs on the backend HPC system and stores it in an internal SQLite database.

2.3 Backend

bwVisu uses as backend an HPC system that provides the powerful hardware necessary to run the scientific applications provided by bwVisu as well as a reverse web proxy to manage the port forwarding rules for the streaming endpoints. In this context, a reverse web proxy is a web server that retrieves resources for web clients from other servers on an internal network of the HPC system in order to keep the interface of the internal network to the internet as small as possible and thus protect internal resources from uncontrolled or unauthorized access from the internet.

An HPC system typically provides access to compute and storage resources via a job scheduler or workload manager (e.g. Slurm[10]). The middleware interacts with the job scheduler either by an exposed API or – as a fallback – by a generic shell script which invokes the required commands provided by the job scheduler.

The HPC system must also host the Singularity containers of the bwVisu applications and provide the Singularity and Xpra software.

7 `https://www.xpra.org`
8 `https://sylabs.io/singularity`
9 `https://apptainer.org`
10 `https://slurm.schedmd.com`

To enable dynamic generation of port forwarding rules, bwVisu needs a reverse web proxy that is capable of adapting its configuration on the fly, which is essential for the dynamic nature of the port forwarding rules generated by the middleware. Currently, only Traefik[11] is supported as reverse web proxy, however, support for other proxies can be implemented as well if required. Traefik was selected because its ability to dynamically change the firewall configuration is its key feature, which is critical for managing dynamic port forwarding rules.

3 bwVisu at Heidelberg University

bwVisu is successfully deployed as a service at Heidelberg University with bwForCluster Helix as backend HPC system. bwForCluster Helix is operated by the University Computing Centre as a service for Baden-Württemberg Universities mainly for research in the fields of Structural and Systems Biology, Medical Science, Soft Matter, Computational Humanities as well as method development in scientific computing. The cluster is equipped with about 20,000 AMD EPYC Milan CPU cores, around 200 NVIDIA Ampere Tensor Core GPUs (A100 and A40), and a high-performance parallel file system with a flash tier for high-speed data access.

As a special feature, bwForCluster Helix provides native access to SDS@hd, a storage service for large research data. SDS@hd is directly accessible from a mount point on bwForCluster Helix and is also available for use in bwVisu applications. Thus, the users of bwVisu can directly benefit from massive storage capacities of SDS@hd in the order 20 PB.

3.1 Workflow

To get a better idea of the features of bwVisu, we will now describe the typical workflow in bwVisu from a user's perspective (Figure 2). See also Beretta (2023a) and Beretta (2023b) for recorded user sessions.

Choosing applications and cluster resources: Users choose an application from a list of available options and specify the resources required to run it on the backend HPC system, such as time duration and the number of GPUs. The middleware receives the name of the application along with the resource request via a REST API call.

Starting applications and streaming graphical output: The chosen application is launched on the backend HPC system by the middleware, which launches a submit script corresponding to the chosen application and submits it to the job scheduler. Due to a resource reservation for bwVisu on bwForCluster Helix, a bwVisu job will start immediately, as long as the resources are available. After launching the application on the cluster, the middleware opens the assigned streaming ports via the reverse web proxy. Information regarding the job such as the public web address of the streaming endpoints

11 `https://traefik.io/traefik`

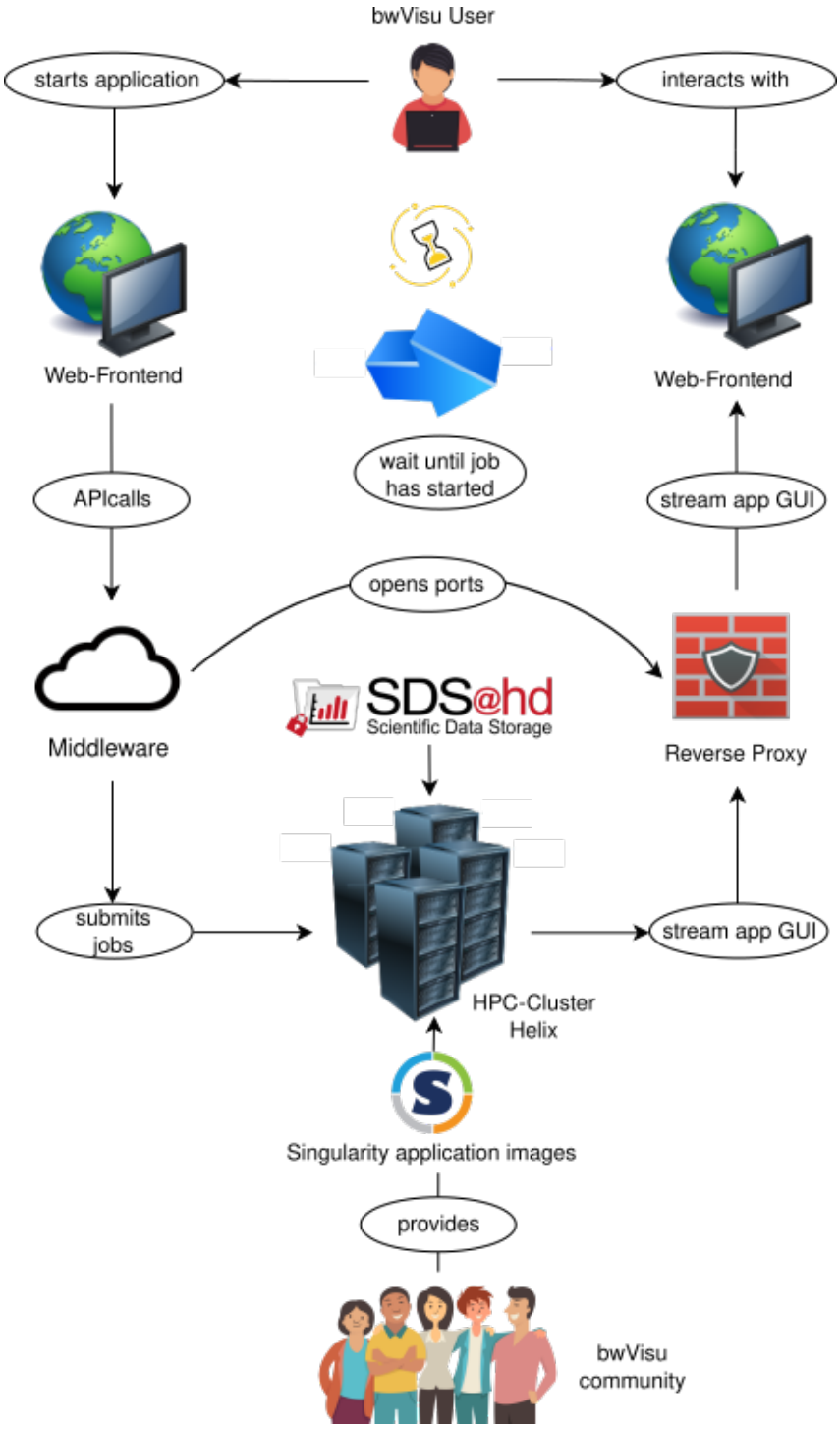


Figure 2: bwVisu workflow at Heidelberg University.

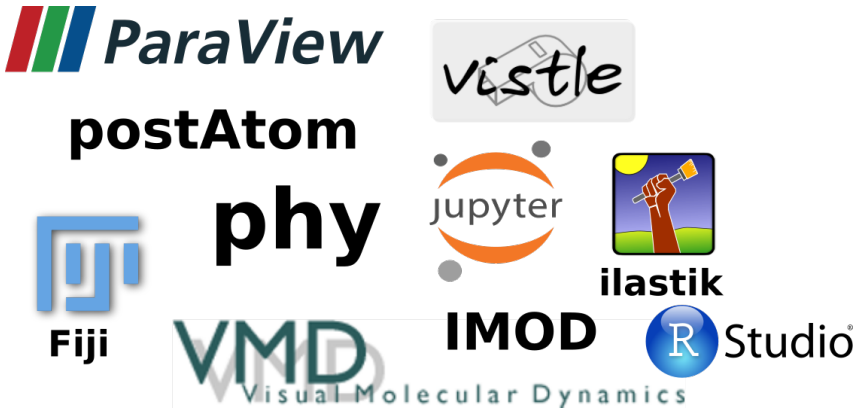


Figure 3: Available bwVisu applications at Heidelberg University.

is made available to the frontend through REST API endpoints. The streaming endpoints for accessing the application's graphical user interface is provided in the web interface as an URL. When this URL is opened by the user, the application's interface is streamed in real-time to the user's web browser, allowing direct interaction with the remote application.

Stopping applications: The application can be stopped either manually by the user or automatically when the specified time limit will be reached. If the user manually cancels the application, the frontend sends an API call to the middleware, which then cancels the associated job on the backend HPC system. In this case, or if the job scheduler reports a timeout for the associated job, the middleware closes the relevant ports via the reverse proxy.

3.2 Applications

bwVisu aims to facilitate the utilization of interactive software in the field of Data Science for scientists. To achieve this, it offers a diverse range of scientific applications commonly used across various disciplines. Figure 3 shows the currently available applications. Expansion of the application repertoire is planned. Deploying additional applications on bwVisu is straightforward due to the minimal requirements imposed by the platform. Essentially, any Linux container-compatible application can be deployed on bwVisu. However, it should be noted that applications designed for multi-user systems may require reconfiguration, as bwVisu exclusively initiates applications for individual users. Based on our experience, such reconfiguration is typically feasible and is usually a matter of just disabling the login mask of the application. Nonetheless, it is possible that certain programs may possess operating requirements that are incompatible with bwVisu.

It is worth reiterating the significant advantage provided by bwVisu in terms of instant availability of scientific applications for scientists who may not possess extensive IT expertise. Additionally, the utilization of powerful hardware without the need for installation, deployment, or complex command-line operations further enhances user productivity.

4 Use Cases

4.1 bwVisu for Science

Microscopy is one of the most common methods to investigate biological processes in life sciences and delivers images that can be challenging to quantify. Yet, image quantification is essential to reveal biological mechanisms, in particular when multimodal imaging approaches are used. In the last decade many open source tools have been developed to facilitate image analysis using conventional approaches (e.g. Imagej/Fiji; see Schindelin et al. 2012) and more recently, machine learning (e.g. ilastik pixel classification; see Berg et al. 2019) or deep learning methods (e.g. noise2void, CARE, and stardit; see Krull, Buchholz, and Jug 2019; Weigert et al. 2018; Schmidt et al. 2018, respectively). Image analysis requires computational resources and, when data analysis demands state-of-the-art tools, programming skills. The rapid and continuous development of artificial intelligence-based deep learning tools for image analysis leads to new powerful ways to precisely quantify images and to enable analyses of complex imaging data that was previously not accessible. In this study, we extensively tested bwVisu functionality using:

- StarDist Jupyter Notebook (Beretta 2023b)
- Ilastik Pixel Classification Workflow (Beretta 2023a)

We used on bwVisu a Jupyter Notebook developed to segment cell nuclei in three dimensional image stacks by training the StarDist deep learning model on GPU resources. bwVisu allows to run the training and the inference directly on cluster resources accessing large dataset of images from SDS@hd storage. The user can install the necessary dependencies to run any Jupyter Notebooks in a Conda environment. bwVisu is a user interactive tool that can be used to train a machine learning classifier on CPU resources with high memory consumption. In our study, we train the ilastik pixel classification workflow (Berg et al. 2019). bwVisu users can run the ilastik workflow on cluster resources and add labels on images in an interactive process. When the training is finished the user can directly batch process large data located on the SDS@hd storage. Both examples show how bwVisu can be used to run applications that require demanding CPU and GPU resources in a user-friendly environment.

4.2 bwVisu for Teaching

Teaching image analysis can be a challenge in particular when GPU resources are needed. bwVisu can be used as centralized infrastructure for running workshops and courses to train scientists on traditional computer vision applications and AI tools overcoming local hardware limitations, for example. There is a large number of methods, tools and applications that could benefit from bwVisu as a teaching platform, since the necessary hardware resources can be reserved and made available on the powerful HPC system for the duration of a course. Additionally, there are processes for the creation of a working area for a course or series of events as well as for the user and group management includ-

ing a right-/role-concept. Course materials can be easily accessible directly from SDS@hd storage creating a common ground to share data during courses and workshops.

In the near future, bwVisu will be used to organize several workshops to train Heidelberg University scientists on state-of-the-art image analysis tools and it will also become a centralized infrastructure to share image analysis tools and knowledge. Furthermore, a course series is in preparation for students using Jupyter on GPUs for different applications via bwVisu. There is a strong interest in using bwVisu for further teaching events at Heidelberg University.

5 Discussion on other technologies

While bwVisu offers a valuable and comprehensive solution for teaching and computational based research, it is important to be aware that alternative platforms exist that serve similar purposes within the scientific community. These alternatives can provide additional features and functionalities that may complement or enhance the teaching experience in consideration of one or another feature.

Subsequently, we will give an overview of some alternative platforms and a comparison of features which might be interesting for operating sites that want to take an informed decision about which platform might meet their specific needs the best.

KASM[12] is a web-based streaming platform focused on providing containerized desktop applications. Notably, KASM introduces a custom open-source VNC server[13] to enhance streaming performance over conventional VNC-based alternatives[14]. Additionally, KASM offers built-in support for educational scenarios, featuring integrated chat functionality and session sharing capabilities tailored for educational contexts. However, it's important to note that Kasm Technologies Inc, the entity behind KASM, hasn't open-sourced all components, leading to licensing costs for adopting KASM. Moreover, KASM lacks integration with existing software in the HPC landscape. By design, there is no integration into an HPC infrastructure for KASM.

Apache Guacamole[15] is an open-source remote desktop gateway aiming to provide a unified access point for diverse server-side streaming technologies through an HTML client and API. A key feature of Apache Guacamole is its browser-based HTML client, enabling user access to stream backend server applications exclusively via web browsers. Additionally, it abstracts various streaming technologies, such as VNC and RDP[16], offering operating sites the flexibility to select their preferred streaming technology. Similar to KASM, Apache Guacamole lacks inherent integration into the HPC landscape.

12 `https://kasmweb.com`
13 `https://kasmweb.com/kasmvnc`
14 `https://en.wikipedia.org/wiki/Virtual_Network_Computing`
15 `https://guacamole.apache.org`
16 `https://en.wikipedia.org/wiki/Remote_Desktop_Protocol`

Open OnDemand[17] is an open-source web application designed to offer user-friendly access to HPC resources for scientific end-users. It provides a web-based interface to access resources and stream applications on HPC clusters, with its standout feature being seamless integration within the HPC landscape. Notably, there exists a wide community of HPC sites that operate Open OnDemand, allowing new adopters to tap into existing knowledge within this community. Furthermore, Open OnDemand's open-source code base has undergone multiple security audits, reducing potential security vulnerabilities.

As previously mentioned, bwVisu aims to offer user-friendly web-based access to HPC resources, allowing for the mapping of interactive scenarios with significant resource demands. When considering the aforementioned platforms in this context, the authors present the following evaluation results from their perspective:

The performance advantages of KASM's open-source VPC server over alternative streaming solutions, such as Xpra used by bwVisu, are not currently clear. Further investigation would be necessary if there is a demand for improvement in bwVisu. KASM's chat and session sharing features might enrich teaching scenarios. However, KASM's licensing model introduces cost considerations and might lead to potential limitations within its field of application.

Apache Guacamole's streaming abstraction results in a high flexibility, which might be useful for some operators. In bwvisu, streaming performance has not been rated as needing improvement thus far. Despite the strengths of KASM and Apache Guacamole, their lack of HPC integration necessitates extra effort and tailored solutions for that use-case.

This is where Open OnDemand emerges. Both bwVisu and Open OnDemand offer distinct advantages for integration with, or utilization within, HPC environments. While we think bwVisu's simple architecture may offer benefits in security and long-term maintainability, Open OnDemand's mature code base, larger feature set, and community support are noteworthy. This platform represents the most promising technology for the authors, which will be further evaluated in the future (see next section).

Overall, determining the "best platform" is likely something that can only be evaluated on a case-by-case basis, using specific and precise use cases.

6 Summary and outlook

The utilization of bwVisu for teaching classes in the field related to data science can be justified based on several key reasons. Firstly, it allows instructors and students to easily bring their own custom applications, enabling tailored and specific educational experiences. Additionally, the availability of ready-made applications eliminates the need for complex installations, reducing technical barriers and enabling efficient use of class time. Moreover, bwVisu leverages powerful hardware resources, such as bwForCluster Helix, which eliminates the burden of managing computing environments and provides students with access to high-performance computing capabilities. Furthermore, the platform's

17 `https://openondemand.org`

integration with large cloud or network storage, such as SDS@hd, facilitates seamless collaboration and data sharing, promoting a collaborative and interactive learning environment. By offering these features, bwVisu offers a compelling rationale for its adoption in data science education, enhancing teaching effectiveness and fostering a conducive learning experience.

While the core features of bwVisu are already implemented, there are specific areas that could benefit from improvements, the introduction of new features, and the exploration of potential alternative approaches. To enhance debugging capabilities and to support the typical workflow of scientists, there are plans to make the outputs of bwVisu job executed on the backend system available to users in the frontend. This functionality would not only aid the operators of the bwVisu service in implementing new applications but also facilitate better debugging capabilities. Additionally, in order to foster the growth of the bwVisu user community and the development of bwVisu applications, users will have the opportunity to locally test and customize new applications for bwVisu using newly provided developer tools. These applications can then be directly uploaded in the frontend for review by the bwVisu operators. Furthermore, access control will be implemented, granting users the opportunity to assume responsibility by becoming maintainers for specific bwVisu applications. In addition to the introduction of new features, it is essential to conduct further benchmarking, tests, and the collection of metrics for typical use-cases. This evaluation aims to assess the robustness of the current implementation when faced with high network latency or a high volume of concurrent user requests and to provide further steps for improvement.

As motivated in the previous section, it is crucial to explore alternative platforms and technologies for remote visualization and remote desktop software in the scientific community. These may offer unique features and capabilities that can complement or enhance the functionality of bwVisu. By examining and testing the integration of these alternative approaches with bwVisu, or exploring how bwVisu can contribute to existing projects, a climate of mutual collaboration can be fostered. When undertaking this process, it is important to consult use cases from both research and teaching domains. This ensures that any subsequent developments are appropriate and contribute to the enhancement of bwVisu's value for research and teaching.

Acknowledgements

The authors acknowledge support by the state of Baden-Württemberg through bwVisu and bwHPC as well as the bwForCluster Helix and the storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grants and INST 35/1503-1 FUGG and INST 35/1597-1 FUGG. Furthermore, the authors gratefully acknowledge the Federal Ministry of Education and Research (BMBF) and the MWK within the framework of the Excellence Strategy of the Federal and State Governments of Germany and the DFG in form of a CRC1158 that granted bwVisu hardware.

References

Alpay, Aksel, Karsten Hanser, Egzon Miftari, Dennis Schridde, Sabine Richling, Martin Baumann, Filip Sadlo, and Vincent Heuveline. 2020. "bwVisu: A Scalable Remote Visualization Service and its Application to Flow Visualization". In *E-Science-Tage 2019: Data to Knowledge,* edited by Vincent Heuveline, Fabian Gerhart, and Nina Bisheh, 173–184. heiBOOKS. DOI: https://doi.org/10.11588/heibooks.598.c8426.

Beretta, Carlo Antonio. 2023a. "ilastik Pixel Classification Workflow". Visited on September 25, 2023. DOI: https://doi.org/10.11588/heidicon/1747437. https://heidicon.ub.uni-heidelberg.de/#/detail/1747436.

———. 2023b. "StarDist Jupyter Notebook". Visited on September 25, 2023. DOI: https://doi.org/10.11588/heidicon/1747437. https://heidicon.ub.uni-heidelberg.de/#/detail/1747438.

Berg, Stuart, Dominik Kutra, Thorben Kroeger, Christoph N. Straehle, Bernhard X. Kausler, Carsten Haubold, Martin Schiegg, et al. 2019. "ilastik: interactive machine learning for (bio)image analysis". *Nature Methods.* ISSN: 1548-7105. DOI: https://doi.org/10.1038/s41592-019-0582-9.

Fielding, Roy Thomas. 2000a. *Architectural styles and the design of network-based software architectures.* Chapter 6.5.1: Advantages of a Network-based API. University of California, Irvine.

———. 2000b. *Architectural styles and the design of network-based software architectures.* Chapter 5: Representational State Transfer (REST). University of California, Irvine.

Krull, Alexander, Tim-Oliver Buchholz, and Florian Jug. 2019. "Noise2void-learning denoising from single noisy images". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2129–2137. DOI: https://doi.org/10.48550/arXiv.1811.10980.

Richling, Sabine, Sven Siebler, Alexander Balz, and Martin Kühl Robert und Baumann. 2022. "Managing large research data with SDS@hd". In *E-Science-Tage 2021: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 421–427. heiBOOKS. DOI: https://doi.org/10.11588/heibooks.979.c13759.

Schindelin, Johannes, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al. 2012. "Fiji: an open-source platform for biological-image analysis". *Nature methods* 9 (7): 676–682. DOI: https://doi.org/10.1038/nmeth.2019.

Schmidt, Uwe, Martin Weigert, Coleman Broaddus, and Gene Myers. 2018. "Cell Detection with Star-Convex Polygons". In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II,* 265–273. DOI: https://doi.org/10.1007/978-3-030-00934-2_30.

Schneider, Gerhard, Vincent Heuveline, Karl-Wilhelm Horstmann, Bernhard Neumair, Petra Hätscher, Josef Kolbitsch, Simone Rehm, Michael Resch, Thomas Walter, Stefan Wesner, et al. 2019. "Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das High Performance Computing (HPC), Data Intensive Computing (DIC) und Large Scale Scientific Data Management (LS^2 DM)". In *Proceedings of the 5th bwHPC Symposium.* Universität Tübingen. DOI: https://doi.org/10.15496/publikation-29040.

Schridde, Dennis, Martin Baumann, and Vincent Heuveline. 2017. "Skalierbare und flexible Arbeitsumgebungen für Data-Driven Sciences". In *E-Science-Tage 2017: Forschungsdaten managen,* edited by Jonas Kratzke and Vincent Heuveline, 153–166. heiBOOKS. DOI: https://doi.org/10.11588/heibooks.285.c3887.

Weigert, Martin, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Jain Akanksha, Benjamin Wilhelm, et al. 2018. "Content-aware image restoration: pushing the limits of fluorescence microscopy". *Nature Methods:* 1090–1097. DOI: https://doi.org/10.1038/s41592-018-0216-7.

Wege aus der Verantwortungsdiffusion – Vermittelnde Angebote des Forschungsdatenmanagements zwischen Top-Down und Bottom-Up

Jan Leendertse [*1], Dirk von Suchodoletz [1], Saher Semaan [2]

[1]Universität Freiburg, Rechenzentrum;
[2]Universitätsbibliothek Freiburg ;
[*]Korrespondierender Autor: jan.leendertse@rz.uni-freiburg.de

Forschungsdatenmanagement erhält aktuell viel Aufmerksamkeit, auch in Form von Förderungen. Es hat sich eine Szene etabliert, in der Wissenschaftler:innen den Umgang mit Forschungsdaten reflektieren, Dienste erdacht, konzipiert und erprobt werden, Kooperationen eingefädelt und Projekte ersonnen werden. Von prominenten Ausnahmen abgesehen ist ein selbstverständlicher Umgang mit Forschungsdaten im wissenschaftlichen Tagesgeschäft nicht angekommen. Eine stabile Planung von unterstützenden Dienste auf nachhaltiger Infrastruktur ist für Hochschulen, Fachgesellschaften oder andere potenzielle Betreiber kaum leistbar. Der übliche Weg einer Verlagerung in Projekte bearbeitet Aspekte, die durch die Antragslogik in den Vordergrund rücken, grundsätzlichen Lösungen aber nicht näher kommen. Dieses Paper stellt kurz die Beschreibung von „wicked problem“ vor, das in anderen Politikfeldern als Erkläransatz für das Scheitern in komplexen Problemlagen verwendet wird. Es erläutert, wie eine Anwendung dieses Ansatzes auf das Forschungsdatenmanagement einige der typischerweise festgestellten Hürden in ein neues Licht rückt. Aus der Praxis der Autoren werden konkrete Beispiele vorgestellt, die als Maßnahme in einem „wicked problem“ in einen neuen, fruchtbaren Begründungszusammenhang gestellt werden. Die Kommunikation mit Stakeholdern des Forschungsdatenmanagement, insbesondere die Einbeziehung von Forschenden bei der Entwicklung und Etablierung von Diensten, erweist sich in diesem Ansatz als Schlüsselfaktor.

1 Einleitung

Forschungsdatenmanagement operiert im Spannungsfeld personell und thematisch hochdynamischer Einrichtungen und einer auf Langfristigkeit angelegte Nachnutz- und Nachvollziehbarkeit von wissenschaftlichen Ergebnissen (Brenger u. a. 2020; Meyer-Doerpinghaus

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.c18074

und Tröger 2015). Hochschulen zeichnen sich sowohl durch Forschungsprozesse, die aus sich heraus auf Innovation und damit Dynamik angelegt sind als auch durch hohe Fluktuation wissenschaftlichen Personals aus. Forschungsdatenmanagement (FDM) hingegen muss langfristig planen und reicht deshalb teilweise weit über den Abschluss einzelner Forschungsvorhaben hinaus. Die Umsetzung von Nachhaltigkeit erfordert somit eine Ausdehnung des Kontexts über organisatorische Grenzen und Projektlaufzeiten hinaus. Mit Projekten enden typischerweise Verträge und damit personelle Zuordnungen von Verantwortlichkeiten. Nachhaltigkeit muss sicherstellen, dass Forschungsdatenmanagement von den Beteiligten in einem größeren Kontext als dem eigenen Projekt betrieben wird.

Nachhaltigkeit bedeutet für das Forschungsdatenmanagement Verbindlichkeit der entsprechenden Dienste, beispielsweise für die langfristige Speicherung und Bereitstellung von (publizierten) Daten und die damit verbundenen personellen und technischen Aufwendungen (Leendertse und Suchodoletz 2020). Ein zentraler Erfolgsfaktor ist hier die effiziente Verwendung von Ressourcen, die von den Beteiligten eingesetzt werden. Das ist ein auszugestaltender Aushandlungsprozess zwischen der jeweiligen Einrichtung und ihren forschenden Entitäten (Leendertse und Suchodoletz 2021).

Forschungsdatenmanagement operiert prinzipbedingt „zwischen allen Stühlen“, was wegen der oft unübersichtlichen Gemengelage und vielen beteiligten Akteuren zu einer Diffusion von Verantwortung führt. Dieses wird in einigen Bereichen durch die zunehmende Tendenz vielfacher Rückversicherung und nachlassender Verantwortungsübernahme der Entscheidungsebene verstärkt. Zu berücksichtigende gesetzliche Anforderungen wie Datenschutz oder die Beachtung der verschiedenen Policies der Universität zu Open Science und Open Data, die Erwartungen seitens der Guten Wissenschaftlichen Praxis (Deutsche Forschungsgemeinschaft e.V. 2013) und der Umgang mit Partnern aus Dritteinrichtungen lässt Forschende die Selbstwirksamkeit und damit den Antrieb verlieren, FDM in ihren Arbeitsabläufen zu integrieren.

Das Papier versucht Ansätze zu finden, typische sich regelmäßig stellende Herausforderungen im FDM zu analysieren und diese zu adressieren (Brenger u. a. 2020). Alle diese Herausforderungen lassen sich durch ihren Multi-Stakeholder-Charakter und Beteiligungen über die eigene Einrichtung hinaus in einem noch ungewohnten Aufgabenfeld beschreiben. Neue Herangehensweisen sind zu erproben und erfolgreiche Lösungen in der Breite zu etablieren (Suchodoletz u. a. 2021). Nach einem Erklärungsansatz zum „wicked problem“ wird dieses anhand von Beispielen aus dem Umfeld der „Research Data Management Group“ (RDMG) an der Universität Freiburg ausgeführt. Diese umfassen Umgang mit sensiblen Daten, Finanzierung von FDM-Personal oder der Aufbau von Datenrepositorien (Suchodoletz u. a. 2021; Bauer u. a. 2023).

2 Erkläransatz „wicked problem“

Seit den 1960er Jahren fand der Begriff „wicked problem“ Eingang bei der Betrachtung von Problemfeldern, die sich in ihrer Komplexität einer lösungsorientierten Erfassung entzogen. In den Jahrzehnten seit 2000 wurde dieser Begriff etwas schärfer gefasst und

als ein Ansatz gesehen, wie komplexe und wenig greifbare Politikfelder bearbeitet werden können. Neben allgemeinen Attributen interdependenter Probleme mit einer Vielzahl von Protagonisten werden in einem der ersten Beträge zu „wicked problem“ in Rittel (1972) folgende Attribute zugeschrieben:

- Es gibt keine Definition des Problems, die von allen Beteiligten geteilt wird. Der Versuch von Lösungen legt eher weitere Aspekte des Grundproblems offen.
- Es gibt keine klare Beschreibung dessen, was als Erfolg anzusehen ist.
- Ein Zustand im Sinne einer finalen Lösung („stopping rule“) ist nicht erreichbar.
- Aus Sicht einzelner Beteiligter führt ein iteratives Vorgehen nicht zu einem Erfolg im Sinne von „richtig“ versus „falsch“.
- Für die Beschreibung des Problemfeldes gibt es mehrere Erklärungsansätze, die nicht als „richtig“ oder „falsch“ gegeneinander abgewogen werden können. Lediglich ein „besser“ oder „schlechter“ ist denkbar.
- Ein „wicked problem“ kann Ausdruck eines weiteren, tieferliegendes Problem sein, das nicht diskutiert wird oder oder gar zur Seite geschoben.
- Lösungen können nicht getestet werden. Das Erarbeiten einer Lösung verändert bereits das Problemfeld in einer Weise, die ein Zurückgehen auf einen Ausgangszustand nicht erlaubt.
- Für eine Lösung bieten sich Maßnahmen an, die als „one-shot-solution“ bezeichnet werden. Sie müssen ohne Testphase auf der Basis einer systemischen Entscheidung und mit bewusst eingegangenen Risiken durch Verantwortliche umgesetzt werden.

Im Anschluss an die letzte Eigenschaft wird sogar postuliert, das Arbeiten an Lösungen sei Teil der Problemformulierung. Mit der Umsetzung von Maßnahmen treten häufig weitere Aspekte des Problemfeldes zutage.

Die kursorisch genannten Kennzeichen für ein „wicked problem“ finden sich beim Forschungsdatenmanagement wieder. Es entzieht sich einer allseits anerkannten Definition schon allein aus der Uneinigkeit, was im Forschungsprozess als „Daten“ anzusehen ist. Kriterien für den Erfolg eines Forschungsdatenmanagements lassen sich nur bedingt ausmachen. Für die Anforderung, Daten über Mindestlaufzeiten oder länger aufzubewahren für unbekannte Nachnutzungen, ist keine stichhaltige Deklaration als Erfolg möglich, eine finale Lösung ist wegen der unklaren Nachnutzung oder Replikationsanforderung nicht exakt fassbar. Die Replikationskrise (Frias-Navarro u. a. 2020) als eine Begründung für die Notwendigkeit eines Forschungsdatenmanagements deutet weitere Problemschichten an. Die Vorgaben aus den Grundsätzen zur guten wissenschaftlichen Praxis führt je nach Fachgepflogenheiten bei Wissenschaftler:innen zu unterschiedlichen Umsetzungen, falls es als Aufgabe nicht gar verdrängt wird. Hochschulen, Forschungseinrichtungen etc. als ermöglichende Institutionen, die Investitionen mit langfristiger Bindung tätigen, berücksichtigen in ihren Entscheidungen Aspekte, die von Wissenschaftler:innen nicht von Belang sind.

Das gegenseitige Verständnis für die anderen Sphären steht durch Sachzwänge in Gefahr auseinanderzulaufen.

Die Leitungen von Universitäten und Fakultäten sind daran interessiert, die verfügbaren personellen und infrastrukturellen Ressourcen effizient einzusetzen. Sie ist auf die Professuren und Forschungsverbünde angewiesen, um Ziele wie sichtbare Forschungsstärke und Exzellenzwürdigkeit erreichen und erhalten zu können. Dazu benötigt sie einen Rahmen, mit dem sie deren Aktivitäten so lenkt, dass daraus ein zielgerichtetes Handeln für die Gesamtuniversität ergibt. Die Herausforderung auf Leitungsebene besteht darin, mit begrenzten Ressourcen unterschiedliche Bedarfe verschiedener Disziplinen abdecken zu können und für zusätzliche Mittel geeignete Andockstellen zu schaffen. Hierfür ist die Erarbeitung von Direktiven für die Etablierung von entsprechenden Diensten notwendig.

3 Fallbeispiele aus dem Forschungsdatenmanagement

Aus der Beschreibung von Problemfeldern als „wicked“ zieht Rittel (1972) Schlüsse zur Verfolgung gewohnter Lösungsansätze. Die übliche Sequenzierung von zunächst Faktensammlung unter Zuhilfenahme wissenschaftlicher Erhebungsmethoden zur Absteckung eines Problems mit dann folgender Erarbeitung eines Lösungsverfahrens ist nicht anwendbar. Allein das Herstellen eines Konsenses bei der Problemformulierung erweist sich als praktisch unmöglich. Das Umsetzen von Lösungen legt neue Constraints offen und beruht auf impliziten Entscheidungen, die das Feld verändern. Die Unmöglichkeit, sich einer finalen Lösung testweise oder iterativ zu nähern, führt zu Maßnahmen, bezeichnet als „one-shot-solution“, für die die Ausführenden bewusst persönliche Risiken eingehen müssen. Sie sind nicht vollständig berechenbar und erfordern eine systemische Sicht über eine Einzelperspektive hinaus. Lösungen sind Teil der Argumentation in der Realität, erfordern gleichzeitig eine Begründung über Motivation und Rational. Dieses Vorgehen steht im Kontrast zur Begründungslogik, mit der in Projekten Förderwürdigkeit belegt werden muss. Dort wird die Formulierung einer Problemlage erwartet, die mit einer Sequenz von Maßnahmen nachweisbar bearbeitet werden soll.

Im Folgenden sollen daher Situationen vorgestellt werden, in denen sich Merkmale von Verfahrenheit finden. Zu ihnen werden Lösungen präsentiert, die aus einer perspektivischen Verengung herauszufinden versuchen. Die Beispiele sind als typische Anwendungsfälle ausgewählt, um exemplarisch zu verdeutlichen, wie über eine sequenzielle Bearbeitung von Problemstellungen hinausgedacht werden muss.

3.1 Entwicklung von institutionellen Policies

Das Abfassen eines grundlegenden Dokuments gilt in Planungshilfen für ein institutionelles Forschungsdatenmanagement als essenziell für einen verbindlichen Einstieg. Solche zentralen Dokumente werden gerne als Policy oder Leitlinie aufgesetzt und dienen dem Setzen eines Rahmens für die eigene Institution. Die Notwendigkeit der Existenz einer Policy für das Forschungsdatenmanagement wird von der DFG (Deutsche Forschungsge-

meinschaft e.V. 2019, S. 9) im Kodex zur guten wissenschaftlichen Praxis (GWP) postuliert (Hiemenz und Kuberek 2019). Die DFG sieht den Prozess, der zur Verabschiedung führt, am Beginn zu dessen Institutionalisierung. Der Kodex zur GWP richtet sich gleichermaßen an Forschende und Einrichtungen, weil beider Beitrag notwendig ist, um das Ziel, Forschungsdaten für Replikation oder Nachnutzung bereitzustellen, zu erreichen. Für das Abfassen einer Policy bildet der Aufriss des Problemfelds die Grundlage für das Skizzieren der Organisation und des grundsätzlichen Vorgehens. Sobald in der Entstehung Stakeholder mit ihren Perspektiven und Interessen beteiligt sind, werden Unterschiede sichtbar. Selbst nach einer formal erfolgten Annahme ist eine Akzeptanz in der Gemeinschaft der Forschenden nicht sicher, zumal Policies ein schwaches Binnenrecht sind, das kaum durchsetzbar ist. Dies ist ein Indiz für die Annahme, selbst nach einem Aushandlungsprozess noch nicht zu einer allgemein getragenen Problembeschreibung gefunden zu haben.

3.2 Koordinierte Nutzung externer Dienste

In einem weiteren Beispiel benötigt ein Projekt Zugriff auf Daten einer europäischen Behörde. Diese hat dafür einen definierten Prozess, dessen Ziel explizit in der Unterstützung von Forschung besteht. In diesem Prozess muss das beantragende Forschungsvorhaben in einer ersten Stufe das Endorsement der Hochschule oder Forschungseinrichtung beibringen, bevor das eigentliche Vorhaben inhaltlich geprüft und eine Freigabe auf die gewünschten Daten erteilt wird. Bei der Erstanfrage hat ein Projekt der Universität diesen Prozess begonnen, um festzustellen, dass eine andere Professur der gleichen Einrichtung diesen Weg bereits bis zum Ende beschritten hat. Im ersten Schritt wurde jedoch nicht das Endorsement der Universität organisiert, das wiederum eine Prüfschleife durch Rechtsabteilung und Genehmigungsschleife durch zeichnungsberechtigte Gremien, sprich Rektorat, erfordert hätte. Diesen Aufwand nicht in Betracht ziehend hat die Professur mit der früheren Anfrage bei der Behörde sich selbst als die Organisation vorgestellt, die in der ersten Stufe durch die europäische Behörde geprüft wird. Damit ist der Weg für das als zweites anfragende Forschungsprojekt, das nicht auf das bereits stattgegebene Endorsement zurückgreifen kann, versperrt.

Es zeigt sich hier ein isolierendes Bearbeiten von Problemstellungen ohne koordinierende Berücksichtigung potenziell weiterer Interessen innerhalb der Universität. Ein Ausweg lässt sich hier nur finden, indem eine Basis der Koordination gefunden wird, die nicht auf die Abarbeitung vorhandener Fälle reduziert ist und prospektiv Maßnahmen ergreift zum Nutzen eines potenziell größeren Kreises.

3.3 Bearbeitung sensibler Daten

In vielen Forschungsgebieten fallen unterschiedliche Daten sensiblen Charakters an, die nicht ohne weiteres Dritten zur Verfügung gestellt werden können und vor dem unbefugten Zugriff zu schützen sind. Dezentrale Ansätze, die von einzelnen Akteuren im Feld unternommen werden, führen hier nicht weit. In sehr vielen Fällen sollen Daten kollaborativ

erhoben und verarbeitet werden und benötigen dazu oft Speichersysteme und Rechenkapazitäten, die über die lokal vorhandene Ausstattung hinausgehen. Ebenso verfügen die wenigsten Einrichtungen über starke Absicherung ihrer lokalen Arbeitsumgebungen, die zudem in der Menge in einer größeren Forschungseinrichtung durch den Datenschutzbeauftragten sinnvoll zu bewerten und zuzulassen sind. Hier bietet der stark formalisierte Ansatz einer Zertifizierung der betroffenen IT-Infrastrukturen nach beispielsweise ISO bzw. BSI eine Chance, die bestehenden Blockaden zu beheben und eine weiterreichende Lösung zu implementieren. Es schafft den notwendigen Rahmen einer innerbetrieblichen Organisation und Ausrichtung des Service-Providers. Dieses ist zwar anfänglich mit Mehraufwänden verbunden, sorgt aber dann dafür, dass nicht jede neue Anfrage zur Verarbeitung sensibler Daten erneute Aktivitäten und einzelfallbezogene Vorkehrungen anstößt. Für Datenschutzbeauftragte, Förder- und Auftraggeber von Forschung wird ein formal nachvollziehbarer Rahmen geschaffen, der eine zügige Klärung und Zustimmung erlaubt. Das vereinfacht und beschleunigt wiederum den Antragsprozess für nachfolgende Forschende und Projekte.

In einem längeren Prozess, angestoßen durch die Mitarbeit im de.NBI-Cloud-Verbund, konnte der Prozess einer ersten Zertifizierung nach ISO/IEC 27001:2017 in 2021 am Rechenzentrum der Universität Freiburg abgeschlossen werden. Das Rechenzentrum belegt mit dieser Zertifizierung das professionelle Management eines der zentralen Maschinensäle, die auch für das Hosting von Maschinen Dritter genutzt werden. Dort werden neben den de.NBI Cloud-Diensten weitere Systeme wie Speicher und HPC für die Forschung und Kooperationsprojekte untergebracht. Der Maschinensaal gehört zu den Basisinfrastrukturen, um strategische Ziele der Universität im Bereich der Forschungsunterstützung zu realisieren.

Auf einer skalierenden und professionell gemanagten HPC- und Cloudinfrastruktur lassen sich Forschungsdaten in allen Phasen ihres Zyklus bearbeiten und archivieren bzw. publizieren (Suchodoletz u. a. 2022; Bauer u. a. 2023; Janczyk, Suchodoletz und Wiebelt 2019). Besonders Daten mit Personenbezug – sonst in Life-Sciences, Digital-Humanities oder Verhaltenswissenschaften eine Herausforderung – finden im Maschinensaal eine IT mit belastbaren technisch-organisatorischen Maßnahmen vor, wie sie von datenschutzrechtlichen Vorgaben gefordert werden. Forschende können in Drittmittelanträgen auf eine funktionierende Architektur verweisen, die als Unterbau für fachbezogene IT-Anwendungen universitätsintern gebucht werden kann. Die Universitätsleitung weiß mit der Zertifizierung um die Solidität zentraler Dienste, die für die Zukunftssicherung des hohen Drittmittelanteils am Haushalt der Universität notwendig sind.

Am Freiburger Rechenzentrum wurde die Asymmetrie der anfallenden Aufwände für die Zertifizierung dadurch gelöst, dass die Kosten für die Durchführung der konkreten Zertifizierung und späteren Rezertifizierung und dafür notwendigen Personalaufwände durch das deutschlandweite Verbundprojekt de.NBI-Cloud getragen wurden. Die Praxis der Zertifizierung wird inzwischen von einer zunehmenden Anzahl von Rechenzentren an Forschungseinrichtungen übernommen. Gleichzeitig liefert das Verfahren Impulse für Best-Practices des IT-Betriebs des gesamten Campusses.

3.4 Verarbeitung externer Forschungsdaten mit formaler Risikoübernahme

Die Verfahrenheit zeigt sich häufig in Anfragen externer Forschender, die ihre Daten auf einem System außerhalb ihrer Heimateinrichtung verarbeiten wollen. Mit bwHPC gibt es ein Verbund von Tier3-Clustern in Baden-Württemberg, die fokussiert auf bestimmte Disziplinen optimierte Hardware- und Softwareumgebungen anbieten. In diesem Clusterverbund fragen Forschende anderer Landesuniversitäten regelmäßig an, ob Berechnungen in Freiburg, also an einem anderen Standort, ausgeführt werden können. Dazu ist die Übertragung von Forschungsdaten notwendig. Sollen in diesem Zusammenhang sensible Daten verarbeitet werden, wird es für die Forschenden unübersichtlich. Sie benötigen eine Freigabe von ihrem lokalen Datenschutzbeauftragten, der für seine Entscheidung eine Übersicht getroffener Schutzmaßnahmen am anderen Standort anfordern wird und einen Vertrag zur Auftragsdatenverarbeitung anfragt oder selbst vorschlägt. Ein solcher Vertrag wird ebenso von der Rechtsabteilung oder dem Datenschutzbeauftragten in Freiburg zu prüfen sein, die selbst keine Kenntnis von den Schutzmaßnahmen vor Ort haben und das Forschungsvorhaben mit den von den externen Forschenden zur Verarbeitung beabsichtigten Daten nicht kennen. Für die Kaskade von Abstimmungen gibt es formal keine koordinierende Stelle, die die externen Anfragen aufnimmt, den Bedarf analysiert, mit den Schutzmaßnahmen vor Ort abgleicht, einen Vertrag zur Auftragsdatenverarbeitung so aufsetzt, dass er für alle Seiten akzeptierbar ist, und für eine Zeichnungsberechtigung sorgt, mit der ein gültiger Vertrag geschlossen werden kann. Das Ergebnis sind erhebliche Verzögerungen und Aufwände auf allen Seiten. Die Unübersichtlichkeit führt zu Ausweichbewegungen, dem Abschieben der Verantwortung und bei vielen der Beteiligten zur Resignation. Im Ergebnis sind diverse Forschungsvorhaben nicht durchgeführt worden, oder die Forschenden sind erhebliche persönliche Risiken eingegangen.

Für eine Lösung sind unterschiedliche Expertisen verlangt, die nicht sequenziell abgearbeitet werden können. Weil die Folgen – in diesem Fall die Nichtbeachtung gesetzlicher Vorschriften im Datenschutz mit wenig Ausurteilungen – für Forschende und Cloudbetreiber nicht einschätzbar sind, kommt es zur beschriebenen Blockade.

Um aus dieser Lage herauszufinden, wurde auf der Managementebene des Cloudstandorts in Freiburg ein prototypischer Vertrag für eine Auftragsdatenverarbeitung entwickelt, mit der lokal zuständigen Rechtsabteilung abgestimmt, eine pauschale Zeichnungsberechtigung eingeschränkt auf die Abzeichnung solcher Verträge ist im Rektorat beantragt. Um solcherart vorbereitet externe Forschungdaten zu verarbeiten, hat sich die Leitung des Rechenzentrums in einem bewussten und dokumentierten Akt mit dem Risiko auseinandergesetzt und eine Übernahme formal beschlossen. Die im vorigen Abschnitt erläuterte Zertifizierung setzt den Rahmen für eine solche formale Risikoübernahme.

3.5 Personelle Reorganisation der neuen Tätigkeitsfelder im Forschungsdatenmanagement

Für die Ausdifferenzierung eines institutionellen FDM wird der vermehrte Einsatz von Personal als entscheidend gesehen, dessen Tätigkeit mit „Data-Stewardship“ umschrieben wird. In Stellenausschreibungen finden sich variierende Profile je nach ausschreibender Einrichtung oder nach Fachkontext. Für die Quantifizierung des tatsächlich benötigten Bedarfs gibt es kaum Erfahrungswissen. In Forschungsprojekten oder kleineren Professuren ist dennoch absehbar, dass Data-Stewards nicht durchgehend oder nur in Bruchteilen eines Vollzeitäquivalents benötigt werden (Förstner u. a. 2023). Aus diesem Faktum ergeben sich Folgeeffekte, die zu einer nicht adäquaten Umsetzung der Aufgaben des „Data-Stewardship“ führen. Fachlich gebotene Stellen werden erst gar nicht ausgeschrieben, sie werden mit anderen Stellenprofilen vermischt oder nicht adäquat qualifiziertes Personal versucht sich an der Aufgabe. Was anderswo als Human-Resource-Management eine signifikante Entwicklung genommen hat, wird in dezentralen wissenschaftlichen Einrichtungen nicht überall den erforderlichen Stand erreichen. Die Verschiebungen im Personalmanagement durch den höheren IT-Anteil, ohne aber als reines IT-Personal angesehen zu werden, legt Rückstände in der Ausbildung von Personal zur Wissenschaftsunterstützung offen und zeigt die Defizite einer zersplitterten Rekrutierung mit Einschränkung auf den eigenen Bedarf. Diese Mängel erkennend, wird an der Universität Freiburg ein Pool von Teilstellen konzipiert. Das Personal für diesen Pool soll dediziert nur für „Data-Stewardship“ angestellt werden. Eine Differenzierung, die fachliche und methodische Diversität abdeckt, muss über eine kritische Masse von Stellenhülsen möglich gemacht werden. Mit einer größeren Zahl von Data-Stewards in einem Pool sollen mittlere und kleine Bedarfe bedient werden, die bei einer eigenen Rekrutierung zu qualifikationsmindernden Effekten führen.

Für die Etablierung eines solchen Pools ist die Einlösung des Nachhaltigkeitsversprechens essentiell. Projekte mit zu kleiner Ressourcenausstattung und zeitlicher Befristung werden Mühe haben, ein solches Nachhaltigkeitsversprechen einzulösen.

4 Zusammendenken der verschiedenen Ebenen

Die verschiedenen und für die einzelnen Wissenschaftler:innen oft nicht komplett zu erfassenden strukturellen Ebenen im FDM verleiten zur Verantwortungsdiffusion. An dieser Stelle können geeignet geschaffene und miteinander eng sich abstimmende organisatorische Einheiten helfen, die Erwartungen der Gesamtinstitution an ein nachhaltiges Forschungsdatenmanagement (Top-Down) mit den Erwartungen der Forschenden nach umfassender und zielgerichteter Unterstützung (Bottom-Up) zusammenzubringen. Hierbei kann eine klare Aufgabenverteilung helfen, die an die Bedürfnisse der jeweiligen Fach-Communities angepasst ist (Suchodoletz u. a. 2021).

Die Kunst einer Einrichtung muss darin bestehen, dort die Verantwortung zu übernehmen, wo in den Fachdisziplinen noch keine ausreichenden Strukturen etabliert sind oder eine klare lokale Verankerung erforderlich ist. So sollten Aufgaben, wie eine einführende, gene-

relle FDM-Unterstützung – organisatorische Aspekte, Antragsunterstützung, Umsetzung von Policies, Abstimmung von übergreifender Kooperation – vor Ort geleistet werden. Hierzu zählen ebenfalls die Betreuung der universitären Infrastrukturen für das FDM mit Schulungen und lokalem Helpdesk.

Fachspezifische Belange, wie der Einsatz moderner (Software-)Werkzeuge, Standardisierung von Metadaten, Unterstützung komplexer Workflows sowie etliche rechtliche Fragen sind am besten auf der Ebene der jeweiligen NFDI-Fachkonsortien (Kraft u. a. 2021; Martins Rodrigues u. a. 2021) oder internationaler Fachgemeinschaften aufgehoben. In der derzeitigen Einführungs- und Entwicklungsphase lassen sich im Moment noch die bereits dargestellten Elemente von „wicked problems“ beobachten, da breit aufgestellte Forschungseinrichtungen wie Universitäten in ganz verschiedenen Rollen involviert sind:

- Direkte Akteure als (Co-)Applicant Institutions, um die Entwicklungen des FDM in bestimmten Fachgebieten führend voranzutreiben
- Vielfach Participants in diversen, aber bei weitem nicht allen Fachdisziplinen, um zumeist in bestimmten Aspekten beizutragen und verschiedene fachliche Sichten einzubringen
- Anbieter von Diensten und Infrastruktur für NFDI-Konsortien
- Mitglied als Organisation im NFDI e.V.
- Nutznießende in verschiedenen Fachdisziplinen von aktuellen und zukünftigen Entwicklungen in der NFDI

Das resultiert bisher in einer sehr unterschiedlichen Wahrnehmung der Rolle und Funktion der NFDI und führt vielfach zu eher verhaltener Mitarbeit. Die Durchdringung und damit Einbindung dieser Ebene ist wegen der wahrgenommenen Komplexität und Unübersichtlichkeit oft gering und entfaltet daher nur bedingt Wirkung. Eine Forschungseinrichtung muss dazu ihre eigene Rolle definieren und in ihren Strategieprozessen abstimmen, was sie selbst anbieten und vorantreiben wollen und welche Dienstleistungen sie von dritter Stelle beziehen. Gerade für rechtliche Fragestellungen offeriert die NFDI eine Chance der drastischen Entlastung der mit der Vielfalt und Tiefe überforderten eigenen Justiziariate. Durch die Auffächerung in fachliche Konsortien lassen sich ähnliche Problemstellungen viel leichter bündeln und so schneller, standardisierter und qualifizierter beantworten.

An der Universität Freiburg finden sich die angesprochenen Elemente wieder wie unterschiedliche Geschwindigkeiten in Fachbereichen, unübersichtliche Servicelandschaft, komplexe Anbindung an die NFDI über unabhängig voneinander operierende Beteiligte und überlastete Rechtsabteilung. Die „Research Data Management Group“ operiert in diesem unscharfen Bereich, stellt gleichzeitig die Verbindung zu Landesinitiativen wie den Science Data Center und dem Arbeitskreis Forschungsdatenmanagement her sowie verfolgt die Entwicklungen in der NFDI. Neben der Abstimmung der Ebenen untereinander benötigt es eine zielführende Verteilung der Aufgaben (Suchodoletz u. a. 2021). So sollte die lokale Ebene immer erster Anlaufpunkt sein, bei fachspezifischen Belange jedoch

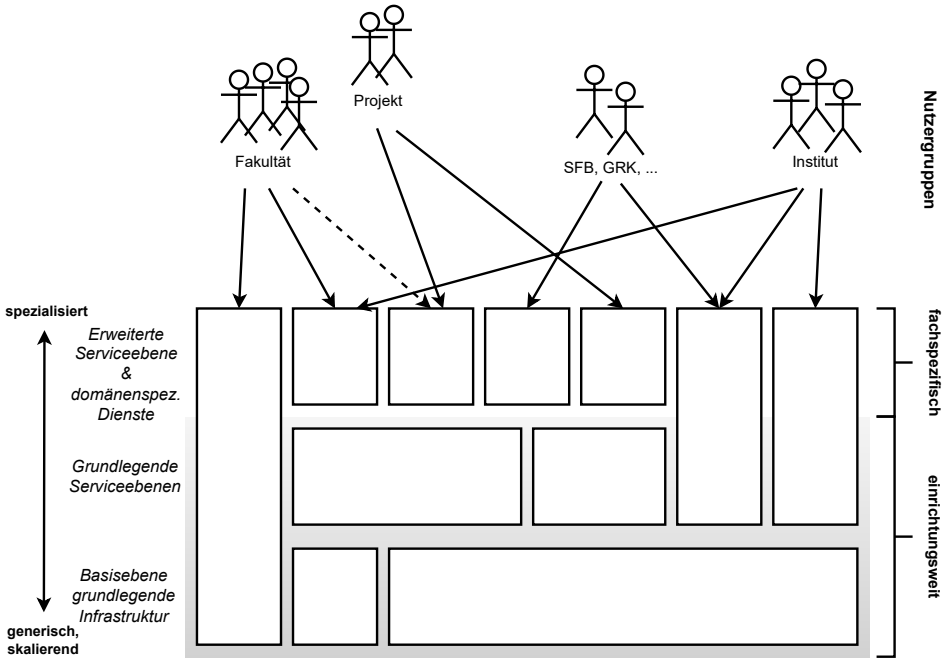


Abbildung 1: Denkbare Versorgungsebenen in der Unterstützung eines institutionellen Forschungsdatenmanagements unter Einbeziehung lokaler und community-spezifischer Angebote.

schnell und nachvollziehbar auf besser aufgestellt übergeordneten Ebenen verwiesenwerden (vgl. Abbildung 1).

Die auf den verschiedenen universitären Ebenen sowie der NFDI begonnene übergreifende Kooperation realisiert ein Lernen von anderen, erlaubt die gemeinsame Entwicklung von Ideen und die Vermeidung von Doppelarbeit. Dieses reicht von der Übernahme „erfolgreicher“ Texte bis hin zur Vermeidung der Reproduktion von Infrastrukturen, die an anderer Stelle schon nachhaltig etabliert wurden.

5 Fazit

Aus der Beobachtung, bei der Institutionalisierung des Forschungsdatenmanagements wieder und wieder an Grenzen zu stoßen, ohne eine eingrenzbare, bearbeitbare Ursache als Ausgangspunkt nehmen zu können, wurde das Modell des „wicked problem“ herangezogen. Dieses bietet sich an, um eine allgemeinere Erklärung für dieses Phänomen zu bieten.

Die Beteiligten aus Universitätsbibliothek, Rechenzentrum und Forschungsunterstützung an der Universität Freiburg haben sich dem komplexen Thema FDM über dieses Erklärmodell genähert, um zu einem neuen Verständnis zu finden. Es soll helfen, die festgestellten Grenzen mit nachhaltigen Maßnahmen zu überwinden. Allen Beteiligten war der Wille gemeinsam, das Thema an der Einrichtung zu verankern und damit dem Anspruch einer Forschungsuniversität gerecht zu werden. Die Suche richtet sich auf Handlungsmuster,

für die eine höhere Wahrscheinlichkeit angenommen wird, zu nachhaltigen Lösungen zu finden. Die vorgestellten Fallbeispiele aus der eigenen Praxis wurden analysiert, ob sie die Merkmale aufweisen, ihnen eine höhere Erfolgswahrscheinlichkeit zuzusprechen. Insbesondere wurde Hinweisen auf potenzielle tieferliegende Problemschichten nachgegangen, die das Forschungsdatenmanagement beeinflussen, auf unscharfe Problemformulierungen und auf die Notwendigkeit, risikobehaftete Maßnahmen anzugehen, die sich einer iterativen Korrektur entziehen. Dieses Vorgehen zeigt die Grenzen des Ansatzes einer projektbasierten Erschließung von neuen Arbeitsfeldern. Solche weisen eine Abfolge von eingegrenzten Problemformulierungen respektive Statusbeschreibungen mit folgenden Arbeitspaketen auf, die in einen definierten und scheinbar bekannten Erfolg münden. Die Fallbeispiele zeigen exemplarisch, wie die Annahme, von einer präzise fassbaren Ausgangslage in ein berechenbares und abschließbares Handeln zu finden, die Schwierigkeit negiert, bei allen Beteiligten ein gemeinsames Verständnis herzustellen. Es hilft nichts: Weil Forschungsdatenmanagement verhältnismäßig neu ist, liegt das Argumentieren im Handeln. Ein Handeln nach Argumentieren oder Auslagern in ignorierbare Projekte wird weniger erfolgreich sein.

FDM muss durch gezielte Unterstützung und Steuerung aus ihrer Wahrnehmung als zusätzliche bürokratische Belastung im Forschungsalltag herausgeholt werden. Neben klarer institutioneller Verankerung (Albert-Ludwigs-Universität Freiburg, Rektorat 2022) kann dieses auch durch beherzte Verantwortungsübernahme und das Ausprobieren neuer Wege und Herangehensweisen geschehen.

Hochgradige Kommunikation, die Einbindung der verschiedenen Stakeholder und insbesondere auch das Nachverfolgen schwieriger Problemstellungen außerhalb des primären Zuständigkeitsbereichs ist dabei wesentlich für eine fortlaufende Verständigung mit dem Ziel, zu eine allgemein anerkannteren Problemformulierung zu finden. Das hilft den Kulturwandel voranzubringen und die Motivation für Forschende zu erhöhen. Für viele ist der Vorteil von offen verfügbaren Forschungsdaten als „First-Mover“ schwer greifbar, so dass die Motivation hierzu entsprechend verhalten ausfällt.

Danksagung

Wir danken dem Land Baden-Württemberg für die Unterstützung des Science Data Centers BioDATEN und bwSFS-Infrastruktur. bwSFS wird durch die Deutsche Forschungsgemeinschaft DFG gefördert: GZ: INST 37/1046-1 FUGG, GZ: INST 37/1047-1 LAGG, GZ: INST 39/1099-1 FUGG, GZ: INST 39/1098-1 LAGG. Zusätzlich danken die Autoren dem Ministerium für Bildung und Forschung für die Co-Finanzierung der ISO-Zertifizierung (031 A538A de.NBI-RBC). Die NFDI wird durch die Deutsche Forschungsgemeinschaft DFG auf Basis der Bund-Länder-Vereinbarung zum Aufbau einer nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018 finanziert.

ORCID:

- Jan Leendertse https://orcid.org/0000-0001-5676-493X
- Dirk von Suchodoletz https://orcid.org/0000-0002-4382-5104
- Saher Semaan https://orcid.org/0000-0001-7487-7348

Literaturverzeichnis

Albert-Ludwigs-Universität Freiburg, Rektorat. 2022. *Policy zum Umgang mit Forschungsdaten an der Universität Freiburg.* DOI: https://doi.org/10.6094/UNIFR/231612. https://freidok.uni-freiburg.de/data/231612.

Bauer, Jonathan, Michael Derntl, Holger Gauza, Steve Kaminski und Dirk von Suchodoletz. 2023. *Herausforderungen beim Aufbau eines föderierten Datenrepositoriums auf Basis von InvenioRDM.* Vortrag und eingereichter Aufsatz für die E-Science-Tage 2023 in Heidelberg.

Brenger, Bela, Marina Lemaire, Jens Ludwig, Janna Neumann, Stephanie Rehwald und Jessica Stegemann. 2020. „FDM am Standort: von der initialen Idee zum dauerhaften Service: Rückblicke auf die DINI/nestor-Workshopreihe". *Bausteine Forschungsdatenmanagement,* Nr. 1: 53–68. DOI: https://doi.org/10.17192/bfdm.2020.1.8168. https://bausteine-fdm.de/article/view/8168.

Deutsche Forschungsgemeinschaft e.V. 2013. *Sicherung guter wissenschaftlicher Praxis.* Wiley Online Library. ISBN: 978-3-527-33703-3. DOI: https://doi.org/10.1002/9783527679188. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527679188.

———. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis.* Zenodo. DOI: https://doi.org/10.5281/zenodo.6472827.

Förstner, Konrad Ulrich, Eva Seidlmayer, Jens Dierkes, Ralf Depping, Fabian Hoffmann und Birte Lindstädt. 2023. *Ergebnisse des Projektes DataStewForschung unterstützen - Empfehlungen für Data Stewardship an akademischen Forschungsinstitutionen.* PUBLISSO. DOI: https://doi.org/10.4126/FRL01-006441397. https://repository.publisso.de/resource/frl:6441397.

Frias-Navarro, Dolores, Juan Pascual-Llobell, Marcos Pascual-Soler, Jose Perezgonzalez und Jose Berrios-Riquelme. 2020. „Replication crisis or an opportunity to improve scientific production?" *European Journal of Education* 55 (4): 618–631. DOI: https://doi.org/10.1111/ejed.12417. https://onlinelibrary.wiley.com/doi/abs/10.1111/ejed.12417.

Hiemenz, Bea, und Monika Kuberek. 2019. *Strategischer Leitfaden zur Etablierung einer institutionellen Forschungsdaten-Policy.* Herausgegeben von Technische Universität Berlin und Technische Universität Berlin. Technische Universität Berlin. DOI: https://doi.org/10.14279/DEPOSITONCE-8412. https://depositonce.tu-berlin.de/handle/11303/9354.

Janczyk, Michael, Dirk von Suchodoletz und Bernd Wiebelt, Hrsg. 2019. *Proceedings of the 5th bwHPC Symposium: HPC Activities in Baden-Württemberg.* TLP, Tübingen. ISBN: 978-3-946552-26-0. DOI: https://doi.org/10.15496/publikation-29062. http://hdl.handle.net/10900/87676.

Kraft, Sophie, Angela Schmalen, Hendrik Seitz-Moskaliuk, York Sure-Vetter, Jennifer Knebes, Eva Lübke und Elena Wössner. 2021. „Nationale Forschungsdateninfrastruktur (NFDI) e. V.: Aufbau und Ziele". *Bausteine Forschungsdatenmanagement,* Nr. 2: 1–9. DOI: https://doi.org/10.17192/bfdm.2021.2.8332. https://bausteine-fdm.de/article/view/8332.

Leendertse, Jan, und Dirk von Suchodoletz. 2020. „Kosten und Aufwände von Forschungsdatenmanagement". *Bausteine Forschungsdatenmanagement,* Nr. 1 (1): 1–7. DOI: https://doi.org/10.17192/bfdm.2020.1.8246. https://bausteine-fdm.de/article/view/8246.

———. 2021. „Überraschung: Ohne Forschende geht es nicht. Digitalisierung und Forschungsdatenmanagement". *Wissenschaftsmanagement* 2021:48–55. https://shop.lemmens.de/produkt/ueberraschung-ohne-forschende-geht-es-nicht/.

Martins Rodrigues, Cristina, Dirk von Suchodoletz, Timo Mühlhaus, Jens Krüger und Björn Usadel. 2021. „DataPLANT – Ein NFDI-Konsortium der Pflanzen-Grundlagenforschung". *Bausteine Forschungsdatenmanagement,* Nr. 2 (2): 46–56. DOI: https://doi.org/10.17192/bfdm.2021.2.8335. https://bausteine-fdm.de/article/view/8335.

Meyer-Doerpinghaus, Ulrich, und Beate Tröger. 2015. „Forschungsdatenmanagement als Herausforderung für Hochschulen und Hochschulbibliotheken". *o-bib. Das offene Bibliotheksjournal/Herausgeber VDB* 2 (4): 65–72. DOI: https://doi.org/10.5282/o-bib/2015H4S65-72.

Rittel, Horst. 1972. „On the Planning Crisis: Systems Analysis of the 'First and Second Generations'". In *Human and Energy Factors in Urban Planning: A Systems Approach,* 389–396. Dordrecht: Springer Netherlands.

Suchodoletz, Dirk von, Ulrich Hahn, Jonathan Bauer, Kolja Glogowski und Mark Seifert. 2022. „Storage for Science – Aktueller Stand und anstehende Entwicklungen eines verteilten FDM-Systems". In *E-Science-Tage 2021: Share Your Research Data,* herausgegeben von Vincent Heuveline und Nina Bisheh, 298–305. Heidelberg: heiBOOKS. DOI: https://doi.org/10.11588/heibooks.979.c13741.

Suchodoletz, Dirk von, Jan Leendertse, Klaus Rechert, Rafael Gieschke, Saher Semaan, Björn Goldammer und Dimitri Tolkatsch. 2021. „Developing a Holistic Research Data Management Strategy for a University. Making Preservation Planning and Long Term Access First Grade Citizens“. In *iPRES 2021. 17th International Conference on Digital Preservation. Proceedings.* Beijing. DOI: https://doi.org/10.17605/OSF.IO/ACNSR. https://osf.io/ACNSR.

How to Choose a Research Data Repository Software? Experience Report

Nina Buck[1],Volodymyr Kushnarenko[1], Björn Schembera[1], Mona Ulrich[2], Heinz Werner Kramski[2], Andreas Ganzenmüller[1], Jan Hess[2], Alexander Holz[2], André Blessing[4], Pascal Hein[3], Kerstin Jung[4], Nicolas Schenk[2], Claus-Michael Schlesinger[3], Thomas Bönisch[1], Roland S. Kamzelak[2], Jonas Kuhn[4], Gabriel Viehhauser[3]

[1]High-Performance Computing Center Stuttgart (HLRS), University of Stuttgart;
[21]German Literature Archive Marbach (DLA);
[3]Institute for Literary Studies / Department of Digital Humanities (ILW), University of Stuttgart;
[4]Institute for Natural Language Processing (IMS), University of Stuttgart

In the age of digital transformation, scientific and social interest for data and data products is constantly on the rise. The volume as well as the variety of digital research data is increasing significantly. This raises the question about the management of this data. For example, storing data so that it is presented transparently, freely accessible and subsequently available for re-use to comply with good scientific practice. Research data repositories provide solutions to these issues and foster compliance with the FAIR principles.

Considering the variety of available software products, it is sometimes difficult to identify a fitting solution for a specific use case. This paper shares our experiences during the process of assessing, choosing and implementing a research data management repository. We provide a brief reflection about standard repository software in contrast to in-house development, describe several software solutions and their features, show software testing results and provide recommendations for assessing and choosing a suitable solution. This paper is aimed in particular for researchers, projects and institutions searching for a suitable software solution to set up and run a repository.

1 Introduction

Within the SDC4Lit project (Science Data Center for Literature)[1] a sustainable repository for net literature and born digitals aims to be built. For this purpose, a suitable open-source research data repository software is needed. Considering the variety of repository

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.c18075

1 https://www.sdc4lit.de

software nowadays, it is sometimes difficult to choose the right one. In this paper we want to share our experience, how to assess a repository to your needs. The paper starts with a brief reflection about a standard research data repository software in contrast to in-house development (section 2), explains the role of requirements for choosing or developing repositories (section 3), introduces several software solutions and their features (section 4), shows some of our software testing results (section 4) and provides recommendations on how to choose a suitable repository software pertaining to a use case based on our experience (section 5). In addition to this paper, we provide a catalogue with general requirements regarding repository software and specific requirements pertaining to the use case in SDC4Lit. This paper and the catalogue can be used as a starting point for other groups and projects who plan to set up and run a research data repository.

2 Standard software or individual solution

Setting up a repository raises the question whether already existing software solutions (standard software) should be used or if it is necessary to develop a new and individual solution from scratch.

The use of a standard software has the advantage that further development and support are more extensively ensured than in the case of individual development, especially through the larger user community of the standard software (Winkler 2008). A ready-to-use solution is usually offered as open source and free of charge and has many already integrated standard functions. A usage of such a software is usually not complicated and quick start is also possible, even if some local adjustments might be necessary. To have it even easier, some proprietary, tailor-made solutions available on the market can be considered as well. They are very convenient, flexible and easy to use, but are usually tied to a pricing model, which can sometimes be unacceptable.

Development of own repository applications is recommended only in cases when none of the offered software solutions can meet individual requirements (ibid.). It is a customised software that is ideally adapted to the local requirements. However, the development of such software is very time-consuming and cost-inefficient, so using and customizing standard software whenever possible is a clear advantage.

3 Requirements

The technical implementation and operation of a repository necessitates selecting suitable software packages according to a number of requirements. These requirements can be defined based on the use cases and workflows of individual projects or institutions. Besides common requirements like the assignment of persistent identifiers (PID), the implementation of a specific metadata model or the availability of application programming interfaces (API), which are available in almost all repository software, sometimes there are also very special requirements such as rendering of directory structures or the capability to operate extremely large files, which not every software is able to fulfill.

There is no common checklist for the selection of a software solution. It always depends on the individual requirements of the institution running the repository. However, there is already a published article that describes possible criteria and their relevance for the selection of repository software (Axtmann et al. 2021a) as well as a catalogue of general requirements for the repository software (Axtmann et al. 2021b) which both can be used as a guidance. In addition to this we provide a catalogue of special requirements for the repository for net literature and born digitals as a supplement to this paper (Buck et al. 2023), which also includes a mechanism to create a ranking of repository software based on fulfillment of the requirements.

4 Standard Software overview

There are a number of standard repository software. Some are widely spread, making their names already known, others are less common or used in only one or several countries. An overview of standard currently used repository software can be found here:

- Registry of Research Data Repository[2]
- Directory of Open Access Repositories (Open DOAR)[3]

Some popular free of charge repository solutions are DSpace[4], Dataverse[5], Fedora[6] and its Framework Islandora[7]. Two others – MyCoRe[8] and Invenio[9] – have been established in German-speaking countries. The list of available and also free of charge repository software can be also extended with Samvera[10], OPUS[11], EPrints[12], Software Heritage[13], CKAN[14], AtoM[15], LibreCat[16], etc. Below follows a brief overview of the software and their features, which were tested as part of the project SDC4Lit.

The overview and results described in this paper are based on the research about repository software made in 2020. Therefore the current version of the software may offer a different feature set as described below. However, the described procedure for selecting a repository software based on individual requirements is valid on general principle.

2 `https://www.re3data.org/metrics/software`; *Last accessed on May 23rd, 2022.*
3 `http://v2.sherpa.ac.uk/view/repository_visualisations/1.html`; *Last accessed on April 25th, 2022.*
4 `https://duraspace.org/dspace`; *Last accessed on April 25th, 2022.*
5 `https://dataverse.org`; *Last accessed on April 25th, 2022.*
6 `https://duraspace.org/fedora`; *Last accessed on April 25th, 2022.*
7 `https://www.islandora.ca`; *Last accessed on April 25th, 2022.*
8 `https://www.mycore.de`; *Last accessed on April 25th, 2022.*
9 `https://invenio-software.org`; *Last accessed on April 25th, 2022.*
10 `https://samvera.github.io/introduction.html`; *Last accessed on May 27th, 2022.*
11 `http://www.opus-repository.org`; *Last accessed on May 27th, 2022.*
12 `https://wiki.eprints.org/w/Introduction`; *Last accessed on May 27th, 2022.*
13 `https://archive.softwareheritage.org`; *Last accessed on May 27th, 2022.*
14 `https://ckan.org`; *Last accessed on May 27th, 2022.*
15 `https://www.accesstomemory.org/de`; *Last accessed on May 27th, 2022.*
16 `https://github.com/LibreCat/LibreCat`; *Last accessed on May 27th, 2022.*

4.1 DSpace

DSpace was originally developed for document management in 2002 at the Massachusetts Institute of Technology (MIT) and the research department of Hewlett-Packard (HP) and is currently managed by the non-profit organisation DuraSpace. Various service providers[17] offer commercial development of the code. DSpace is used by more than 1000 organisations, making it the most widely used standard software for repositories, which speaks for a large and active community.

DSpace is free, easy to install and fully customisable. DSpace supports the OAI Protocol for Metadata Harvesting OAI-PMH, comes with a Solr-based search mechanism and internal checksum verification. Also Shibboleth integration is possible via a plug-in.

DSpace 6.0 was investigated. Late summer 2021 a new version DSpace 7.0 was released. It has a completely new interface and some new features.

How DSpace works can be tried at a demo version[18]. Several login credentials are provided, so it is not necessary to create own account.

4.2 Fedora

Fedora was developed at the University of Virginia and Cornell University. The project is led by the Fedora Leadership Group and overseen by the non-profit organisation DuraSpace.

Fedora is a robust, modular and freely available repository software for managing digital content. It is primarily used in libraries, universities and other research and academic institutions as a data repository for document servers. Fedora enables access to very large and complex digital collections. It features very high flexibility and supports all metadata schemas. It has a RESTful API, checksum verification and enables Shibboleth integration. However, Fedora is only a minimal environment that needs to be significantly extended for productive use (see section on Islandora).

Fedora 5.1.0 was investigated. Meanwhile Fedora 6 was released. Unfortunately, there was no demo instance offered.

4.3 Islandora

Islandora was originally developed at the University of Prince Edward Island (UPEI) in 2009 and is now used by over 300 institutions.

Islandora is an open source software framework based on Fedora and content management system Drupal. It can be extended with several modules developed by the Drupal community. The software comes with a default configuration, which provides basic functionality.

17 `https://duraspace.org/dspace/resources/service-providers`; *Last accessed on April 25th, 2022.*

18 `https://demo7.dspace.org`; *Last accessed on May 23rd, 2022.*

Islandora is widely used and has a strong community. However, the software product is very complex and consists of many microservices. The installation went not smoothly and a high maintenance and operation effort was estimated.

There is no native possibility to transfer a directory tree on a hard disk into a repository. The directory structures must be simulated via "member of" relationships.

Islandora 7 and 8 were tested. Nowadays version 9 is available. A demo version[19] with provided login credentials could also be tried out.

4.4 Invenio

Invenio was originally developed at CERN in 2002 as a document server. It has been further developed and nowadays consists of three products[20]:

- InvenioRDM – a repository/document management platform,
- InvenioILS – an integrated library system,
- Invenio Framework – a code library to build large-scale information systems such as InvenioRDM and InvenioILS.

Invenio is an open access software and is designed to work with huge amount of data as well as large datasets. Metadata from various sources can be integrated. Invenio provides a PID-store and resolver that allows you to use a preferred PID scheme to identify records. Invenio uses Elasticsearch as a search engine.

Two Invenio instances were installed and tested: Invenio Framework and InvenioRDM. Despite numerous tutorials, the installation went not without problems. Configuration and extension of the repository was time-consuming and required additional programming.

At the time of testing, InvenioRDM was not a finished product and had only a few functionalities. The operation of this software was done mostly via command line. A Web-Interface was not fully available during the testing.

Invenio Framework is very modular and allows a wide range of applications to be served. The operation of this software via web interface was very limited. File upload was done via the command line. Mapping of the data hierarchy was also possible, as paths, but there was no direct implementation on the web interface.

Only for InvenioRDM is a demo version[21] available.

19 `https://sandbox.islandora.ca`; *Last accessed on May 23rd, 2022.*
20 Invenio Homepage, `https://invenio-software.org`; *Last accessed on April 25th, 2022.*
21 `https://inveniordm.web.cern.ch`; *Last accessed on May 23rd, 2022.*

4.5 MyCoRe

MyCoRe was developed at the University of Duisburg-Essen. The office is located at the Regional Computer Centre of the University of Hamburg. MyCoRe is mainly distributed within Germany. The community is not very large, but is always ready to help.

MyCoRe framework provides all basic functions of document and publication servers. Own web applications can be developed through adaptations. An integrated image viewer is provided. Metadata models are customisable and extensible. Persistent identifiers ensure permanent access to the data.

MIR (MODS Industrial Repository) is an application that can be installed out-of-the-box. It provides all typical repository functions and can be used productively immediately. Adaptations are only foreseen for the layout and web content.

MyCoRe LTS 2019.06.04 was tested. The installation is slightly complicated as there are no step-by-step instructions and many components such as Solr, Apache Tomcat, etc. have to be configured by oneself. The web interface can be fully customised. During the upload of data it is possible to drag and drop entire directories and also lists of directories and files into the corresponding place and keep the original structure of the data.

Every year a new Long Term Support (LTS) version is released. A demo version[22] with several login credentials is also available.

4.6 Dataverse

Dataverse originates from the Institute for Quantitative Social Science (IQSS) in a collaboration with the Harvard University Library and Harvard University Information Technology. It is an open source software, has a supportive developer community and is distributed worldwide.

Dataverse repository software has two types of "data containers": dataverses and datasets. Dataverses (logical dataverses within a Dataverse installation) are the organisational structure of the repository. Datasets are the organisational structure below dataverses. They contain files and associated descriptive metadata. Datasets are nestable and can imitate directory structures. All files inside of datasets can be represented as a directory structure via the activated Tree-view.

Dataverse supports Search, Data Access, Metrics and Native APIs, as well as SWORD, OAI-PMH and Solr. It provides versioning and data citation. Authentication is possible via local accounts as well as via Shibboleth, ORCID, Google or GitHub. Roles and rights can be defined for each dataverse or dataset. The File Previewer and some other tools can be additionally integrated, which provides an opportunity to extend the functionality of the repository.

22 `https://www.mycore.de/mir/content/index.xml`; *Last accessed on May 23rd, 2022.*

Software versions since 5.3 were tested. Following the installation guide[23], it is quick and easy to install Dataverse. Upgrading to the next versions is not difficult either.

A demo instance of Dataverse[24] is also available.

4.7 Testing and Results

At first glance, there are no significant differences between repository software solutions. Therefore, the search for suitable software should be more precise and targeted explicitly at the area of application. But how should it be done exactly?

For a more detailed investigation, it is recommended to install an eligible software and evaluate it according to the collected requirements. On the one hand, you can see how the installation proceeds and how the further operation of the repository could look like. On the other hand, the specified requirements can be precisely tested to see whether they are fully, partially or not fulfilled.

5 Own experience

SDC4Lit was created with the aim to reflect on the requirements that digital literature places on its archiving, research, and mediation, and to implement appropriate solutions for a sustainable data life-cycle for literary research and mediation in the long term. In the course of this a long-term repository for digital literature is to be established. It should serve as a central repository for genuinely digital literary materials, literature on the net and parts of the author's inheritances, born-digitals. Literature on the net collection includes archived websites, literary blogs or online magazines related to the modern German literature. This collection consists of about 500 sources and results in a data volume of about 9TB with annual growth of 1TB. Born-digitals collection consists of about 75 inventors and represents a total data volume of about 2.8TB, stored in different formats and distributed over about 2000 data carriers.

In order to find the most suitable software solution for SDC4Lit, several potential use cases and resulting requirements were collected. These requirements (Buck et al. 2023) are the part of the catalogues of requirements (Axtmann et al. 2021b) mentioned in section 3. Some of the most essential requirements in SDC4Lit were the rendering of directory structures and possibility to allocate a persistent identifier for each file in a dataset, which are not common for research data repositories.

Based on requirements, it is now clear what should be included in a software package. At first, documentations of software products were examined to see if one or another requirement might be fulfilled. Since not all repositories were up to date, we limited our search to DSpace, Dataverse, Fedora, Islandora, MyCoRe and Invenio. All these

23 `https://guides.dataverse.org/en/latest/installation/index.html`; *Last accessed on June 1st, 2022.*

24 `https://demo.dataverse.org`; *Last accessed on May 23rd, 2022.*

repository solutions are widely spread and have a big community. They are continuously developed and improved. But not all of them could fulfill our requirements. Thus the choice had to be reduced again.

To evaluate the installation process and all requirements, Dataverse, MyCoRe, Islandora and Invenio were installed. Afterwards each software was weight up by assigning points for fulfillment of requirements. In the end, Dataverse met our essential requirements and suited slightly better than other candidates, therefore it was chosen as a repository software for the SDC4Lit data.

6 Summary

The variety of open source repository software is large. Each product has many features, most are the same or similar, but some are only available in one or another software. If you want to build your own repository, you need to consider many aspects, because developing a repository from scratch turned out to be difficult, and good repository products are already available. During our project, we tested several repository software and created a list of requirements that is aligned to our data, that helped us to select the most suitable software for our data. For this reason, collected references and catalogues of requirements presented in this paper can serve as a starting point and decision-making guide for choosing a proper repository software for project-specific data, regardless of the discipline from which the data originate.

Acknowledgements

This research was done in scope of the SDC4Lit project, funded by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK).

References

Axtmann, Alexandra, Felix Bach, Jonathan Bauer, André Blessing, Thomas Bönisch, Nina Buck, Holger Gauza, et al. 2021a. "Kriterien für die Auswahl einer Softwarelösung für den Betrieb eines Repositoriums für Forschungsdaten". *Bausteine Forschungsdatenmanagement,* number 3: 14–26. DOI: https://doi.org/10.17192/bfdm.2021.3.8348. https://bausteine-fdm.de/article/view/8348.

———. 2021b. *Kriterien für die Auswahl einer Softwarelösung für den Betrieb eines Repositoriums für Forschungsdaten.* DOI: https://doi.org/10.5281/zenodo.5562885.

Buck, Nina, Volodymyr Kushnarenko, Björn Schembera, Mona Ulrich, Heinz Werner Kramski, Andreas Ganzenmüller, Jan Hess, et al. 2023. *How to choose a research data repository software? Experience report. Table of requirements.* DOI: https://doi.org/10.5281/zenodo.7656573.

Winkler, Marco. 2008. “Langzeitarchivierung von Online-Publikationen digitaler Repositorien”. Diplomarbeit, Fachhochschule Potsdam.

Quo venis? Metadata for Common Scientific ASCII Files

Muhammed Bayram, Frank Tristram

Karlsruhe Institute of Technology

Currently most scientific information is stored in files and some in databases or repositories. The diversity of file formats is a challenge for all software systems and automated workflows that try to process the information therein. While the final analysis level of scientific content gets a lot of attention in context of machine learning and data analysis, the other information layers are less in focus, although they are the basis for the other progress. There is no efficient way we can link and compare scientific information over scales, disciplines or sometimes even different labs in the same discipline, if we cannot merge information automatically. Here we show a perspective to structure and enrich unstructured data with metadata as a prerequisite step to combine information from heterogeneous sources for the example of ASCII files. Currently it is trivial to bring thousands of file formats into one storage device, but pretty hard to provide the content of a file on a higher level than a stream of characters. For each specific format a special converter is built and for unspecific formats like .dat or .txt there is even no solution to clarify what converter is needed. The first steps from an unknown file format to "information" need to cut the file into meaningful sub-objects, interpreting a file more as a container. This is very analogue to a "table of content" for a book. We argue that a meaningful file segmentation, that allows to provide object names and object types (like table or key-value pairs), headers and other relations as defined and interlinked data objects and not as plain string is a major step. On such objects, an information crawler can pull information out of dark data archives.

1 Introduction

Clusters of excellence and other large research collaboration must solve the problem of creating an added value from bringing together diverse data, expertise, infrastructures, and motivations. This is mainly relevant for linking cross-disciplinary research to a core topic, which is more difficult, if data needs to be connected across different institutes and reused interoperably. In experimental clusters, hundreds of measurement devices, software solutions and objectives are scientifically connected with each other, which makes an overarching interpretation of all data difficult. Interfaces and standardizations are often

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.c18076

missing as well as overlapping knowledge and experiences (Stocker et al. 2020). A common starting point is data that is stored on central storage systems but is not FAIR. In order to make every piece of information available to every cluster member, the representation of this data needs to be standardized. The HDF5 format offers a flexible container for structuring all conceivable data in a standardized way and making it available. However, it is not straightforward to convert an arbitrary ASCII file into HDF5, such that there is a major improvement in the amount of structuredness.

2 Reading of ASCII Files

While many specific file formats like .tiff or .pdf at least offer metadata about the creation of the file, plain ASCII files offer no information. Even worse, there is no common interpreter that can, for example, distinguish between different .dat or .txt structures to extract specific information. We are developing a method that one could call file segmentation. As in the image segmentation, this process does not deliver the final file analysis to answer a specific question (e.g. to identify a person or animal), but an intermediate result consisting of distinct, condensed components. Our most important insight was, that an information extraction can never be complete, but needs to be "fit for purpose" like the current image analysis (see Figure 1). It does not deliver all information about an image, but enough to answer questions like "What is in there?" and might add value for one or several specific information dimensions (from which there are practically infinite).

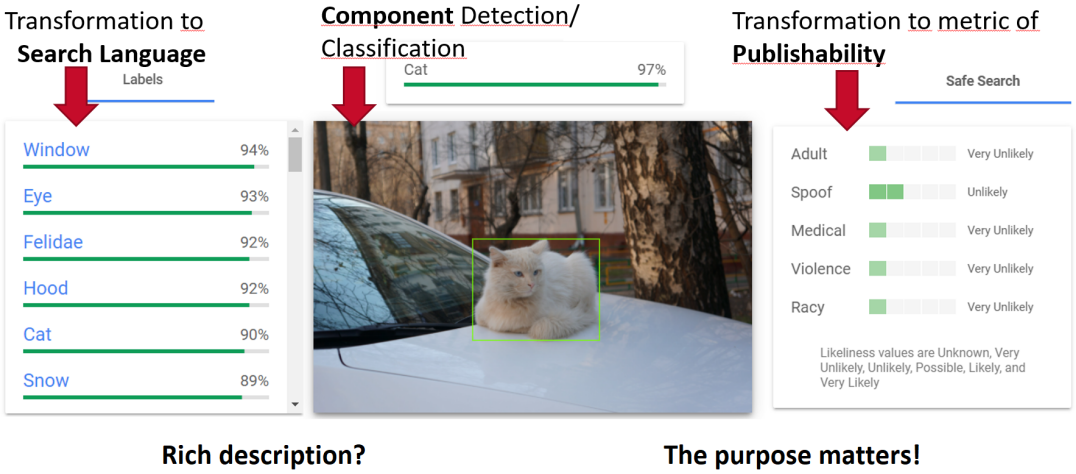


Figure 1: Google Information extraction from images: It still is pretty incomplete, but serves some intended purposes.

Of general importance is a component detection. The components in an ASCII file can be a table, a set of information (like an address), a collection of key-value pairs, some longer text block etc. We identify describing metadata for these components. We then not only know that there is a distinct object, but also object properties, especially focusing on the underlying data structure. Imagine someone would store a 2D-Array of integers into an ASCII file. We try to guess this initial data structure and put this into an HDF5. This intermediate result within the HDF5-file opens up several interesting perspectives:

- For each file format "subclass" (e.g. a special .dat file coming out of a device) a unique fingerprint of the structural subcomponents can be built.
- Values inside very different files can be visualized, compared and analyzed by existing HDF5 tools without the need to write wrappers for each file format.
- It is possible to create a unique interface with persisting functionality, almost independent from the underlying formats (or versions).

We are motivated by these points, but saw these can be of interest for other communities, too. In any use of this, one needs to check the output structure manually to control if it is like one would expect. However, this check is still much less effort, then creating standardized structures "by hand" and can be applied automatically on future use cases. In Figure 2, you can see, that already the most simple example creates an ambigious structural result between a table and key-value pairs.

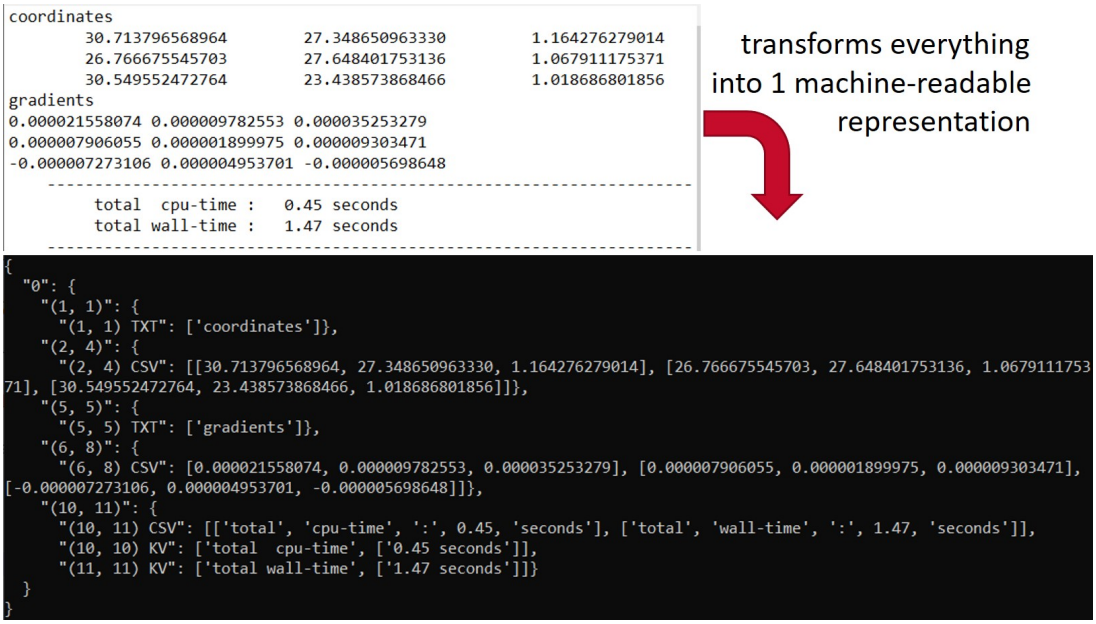


Figure 2: Most basic example what is done by converting a human readable ASCII file into a nested machine-readable form. Here, the machine detects two options to interpret the CPU time content: It could be a table with 5 columns or two key-value pairs. Final decisions must always be done by the user.

3 Harvesting Zenodo

We have downloaded 10.000 scientific data files from material science from Zenodo, to run some real-word analysis with our new information extractor. We only downloaded .json, .txt, .dat and .csv files. We included "prestructured" files: .csv(n=3126), .json(n=259) as well as "unstructured" files: .dat (n=1693) and .txt(n=4935) However, we could only identify 9860 encodings (98.6%) automatically. Manuel inspection revealed, that most remaining files, were likely to be misnamed. So e.g. a readme.txt was in fact a readme.doc

file that could be opened with any common text processing program, but counts as "unreadable encoding" in this statistic. We found that the unstructured formats had very different origins and compositions. For them we identified over 100 structurally different fingerprints. Clustering along these fingerprints created large subclusters of e.g. "readme-similar" files, bullet-lists, data tables and combinations of such elements. We also found a long tail of structures that seem to exist just in one data source.

4 Conclusion

The diversity of subformats raises the question, how to standardize a description of such files. Therefore we attend the E-Science-Tage to discuss about real applications in this area. The library as as well as the computing center community showed interest in what we are doing and more future work, because currently, there is simply nothing that could harvest information from unstructured files and perform some structuring. From these discussions and our own experience we suggest here, to develop a metadata schema that is close to what we know as a "table of content" for books. A table of content is in principle a nested short description of meaningful blocks. Although such blocks are somewhat arbitrary, they are a step forward for structuring and findability. Probably a web portal that provides something like this table of content for any uploaded ASCII (and some more) files, could force a standardization in this untouched area by practical usage. Another idea for a use case that came up, was to apply this on REST-APIs for repositories to identify what you are really looking for. Each API is similar but a bit different and therefore cumbersome to connect to in a way that returns the object of interest automatically - and persistent, even if something within the repository changes. These are just very recent ideas how to move on from where we are and it is not unlikely that a solution in some years will exactly work like this, whether it is developed by us or by others.

Acknowledgements

The work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – 2082/1 – 390761711.

References

Stocker, Markus, Louise Darroch, Rolf Krahl, Ted Habermann, Anusuriya Devaraju, Ulrich Schwardmann, Claudio D'Onofrio, and Ingemar Häggström. 2020. "Persistent Identification of Instruments". *Data Science Journal* 19 (1): 18. DOI: https://doi.org/10.5334/dsj-2020-018.

A Machine-actionable Workflow for the Publication of Climate Impact Research Data from the ISIMIP Project

Jochen Klar, Matthias Mengel

Potsdam-Institut für Klimafolgenforschung (PIK)

The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) aims to provide a quantitative and cross-sectoral synthesis of the different impacts of climate change. To this end, a simulation protocol defines a set of common experiments that are valid across sectors. Modelling groups around the world run their simulations following this protocol using climate and socio-economic forcing data provided by ISIMIP. The impact model output data is collected by the ISIMIP team at PIK and made publicly available via the ISIMIP repository. In total, more than 100 TB of up-to-date climate impact simulations are freely available. The data is broadly used by the international scientific community, but also by economic and civil society actors.

The workflow of data submission and publication was considerably improved by the the introduction of a machine-actionable protocol and the development of several interlinked software tools for data maintenance, quality control and data publication. While the specific implementation is tailored to ISIMIP, the general ideas should be transferable to other projects.

1 Introduction

The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP; Frieler et al. 2017) is a community-driven climate impact modeling initiative that aims to contribute to a quantitative and cross-sectoral synthesis of the various impacts of climate change, including associated uncertainties. It is designed as a continuous model intercomparison and improvement process for climate impact models and is supported by the international climate impact research community. ISIMIP is organized into simulation rounds, for which a simulation protocol specifies a set of common experiments. The protocol further describes a set of climate and direct human forcing data to be used as input data for all ISIMIP simulations. Based on this information, modelling groups from different sectors (e.g. agriculture, biomes, water) perform simulations using various climate impact mod-

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.c18077

els. After the simulations are performed, the data is collected by the ISIMIP data team, quality controlled and eventually published on the ISIMIP Repository. From there, it can be freely accessed for further research and analyses. The data is widely used within academia, but also by companies and civil society. ISIMIP was initiated by the Potsdam Institute for Climate Impact Research (PIK) and the International Institute for Applied Systems Analysis (IIASA).

In this paper we describe the data publication workflow from the modelling groups to the end user in detail (see Figure 1 for an overview). This workflow ensures a high data quality and follows the FAIR data principles (Wilkinson et al. 2016).

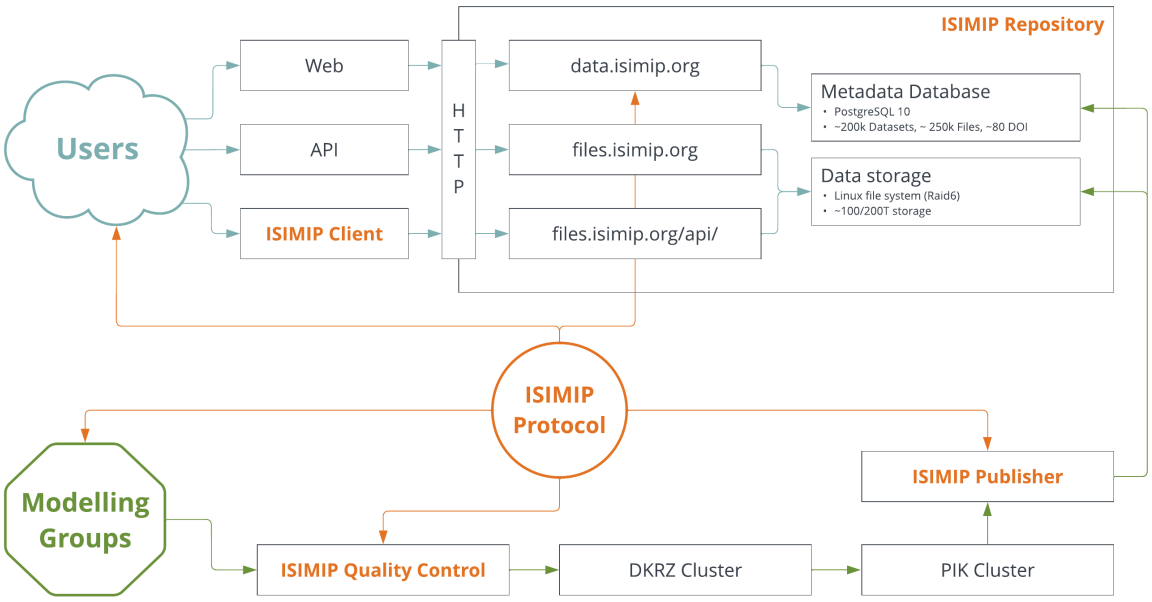


Figure 1: Overview of the ISIMIP publication workflow: After the simulations are performed by the modelling groups, the data is checked using the ISIMIP Quality Control tool. Only if the data passes all checks, it is made internally available to the modellers community at the ISIMIP space at the German Climate Computing Centre (DKRZ) and the PIK cluster. Finally, it is published on the ISIMIP Repository, using the ISIMIP Publisher. From there, the data can be retrieved by users all over the world, via the web page of the Repository, using it's API, or with the ISIMIP Client library. All components of the workflow are using information from the machine-actionable ISIMIP Protocol.

2 The machine actionable Protocol

The simulation protocol is used by the modeling groups contributing to ISIMIP to set up their simulation runs. In the past, the simulation protocol was a text document of ca. 100 pages, which was collaboratively edited using Microsoft Word. The protocol was published as a PDF document and subsequently updated every few months. It took considerable

effort to manage the different changes in the protocol and it was not possible to update the protocol in an appropriate frequency to follow the development of the models in the different sectors. Since the chapters for the different sectors where edited independently, inconsistencies in the output variables occurred frequently.

To address these issues, the protocol for the current simulation rounds was implemented using a novel approach. Instead of one large document, the information is now stored in a set of smaller documents. The text part of the protocol, which contains mainly the introduction and notes on specific sectors, is written in markdown. Structured information is stored in JSON files. Both file formats are plain ASCII and are therefore perfectly suited to be version-controlled using Git. A set of Python scripts was developed to create a human readable, interactive web page[1] and a set of static sector specific, machine readable JSON files[2]. Using GitHub Actions we are able to run the build process every time a new commit is pushed to the GitHub repository. In addition to the JSON definitions, the protocol contains file name patterns for each sector. These patterns are regular expressions which convert the file names of the input and output files to a dictionary of specifiers (cp. Figure 2).

Figure 2: Metadata extraction using regular expressions: The file name consist of a set of specifiers which are defined in the ISIMIP Protocol. Using regular expressions we validate the file names and extract the specifiers into a dictionary.

1 ISIMIP protocol: `https://protocol.isimip.org`.
2 e.g. `https://protocol.isimip.org/definitions/ISIMIP3a/OutputData/water_global.json`.

3 Quality control

After the model groups finished their simulations, the data, which is stored in the NetCDF[3] format, is transferred to the ISIMIP data space at the German Climate Computing Centre (DKRZ), where the ISIMIP team performs different quality control checks on the data to ensure that the data products are in agreement with the specifications of the protocol. In the past, this was done using a number of different scripts which were hard to maintain. In order to connect the quality control with the new, actionable protocol, we developed a dedicated command-line tool written in the Python programming language. Its source code is available on GitHub[4] and it is published on the Python Packaging Index[5]. The tool can be used independently by the modelling groups to find and fix most errors *before* uploading terrabytes of data.

With only the *protocol path*, which specifies simulation round, data product, and sector[6] as argument, the tool is able to fetch the *current* sector specific machine readable protocol via the internet. The tool then scans the current directory and its subdirectories recursively. For each file individually, it checks its filename against the file pattern from the protocol. Only if a filename matches, the file is opened and a series of checks (NetCDF data model, compression, dimensions, grid, global attributes, variables, units) is performed. The tool is also able to perform limited checks on the actual data stored in the files. As part of the protocol, each variable has a minimum and maximum value assigned to it. Examples include negative water consumption or unnaturally high tree heights.

If one of the steps fails or any other information needs to be communicated to the user, the program gives a meaningful response, so the users can alter the files accordingly. Some of the less severe errors can be fixed by the tool itself (if the user so chooses).

4 Data publication

After the data is checked and an optional embargo period has passed, the data is published on the ISIMIP Repository. From a technical point of view, the repository contains three main components:

1. A file server, which can be accessed via the internet using a simple web server[7]. As of 2023, the server has a storage capacity of about 200 TB.

2. A relational database, which contains metadata entries for all files on the file server, as well as entries for *datasets*, in which files with the same variable and from the same experiment, but for different decades are combined.

3 Network Common Data Form (NetCDF):`https://www.unidata.ucar.edu/software/netcdf`.
4 ISIMIP Quality Control on GitHub: `https://github.com/ISI-MIP/isimip-qc`.
5 ISIMIP Quality Control on PyPI: `https://pypi.org/project/isimip-qc`.
6 e.g. `ISIMIP3a/OutputData/agriculture` for the agriculture sector in the ISIMIP3a simulation round.
7 ISIMIP files server: `https://files.isimip.org`.

3. The main repository web page[8], which allows for a convenient search of the metadata database and guides the user to the corresponding downloads from the file server.

The publication process contains a number of processing steps, which are performed using the specifically developed command-line-tool ISIMIP Publisher[9]. First, the files to be published are copied to a working directory on the file server. There, the file name patterns (cp. Section 2) are used to extract the metadata from the file names. Files are then logically combined to datasets, so that all files for the same experiment and variable, but different years are contained in one dataset. Each dataset and file is assigned a UUID as unique persistent internal identifier (called `isimip_id`) and a checksum is computed for each file. For each dataset and file, an entry is created in the database, containing the described metadata, a version string based on the date (e.g. `20230101`), and the licence under which the dataset is published. In the final step of the publication process, the files are moved to the public directory on the server, and the entries in the database are included in searches on the repository website.

The Repository assigns Digital Object Identifiers (DOI) to all datasets corresponding to a specific sector and simulation round (e.g., Marcé et al. 2022). The DOI are registered with DataCite (PIK is a member of the German DataCite consortium). The metadata according to the DataCite Metadata Schema 4.4 is manually prepared and stored in the database with a link to all datasets that are referenced by the DOI. This rigid link between the DOI and the specific version of the data ensures traceability and reproducibility.

When datasets need to be replaced because problems were discovered or improved data is available, the old datasets are archived. The metadata for archived datasets remains in the database, a corresponding page remains online, and the datasets are available on request. When the new dataset is published, the repository links the new and the old dataset together. When files are replaced, a new DOI is created for the sector. The old DOI page stays online and a reference to the new DOI is added. A caveat system[10] provides the users with information about changes to the data.

To further improve data access, a *Configure Download* option can be used to perform operations on the server, e.g. a cut-out of a specific country or region or the extraction of a time series for a point as CSV. This functionality is provided by a dedicated web service[11]. The Repository can also be accessed by it's API and a dedicated Python client library[12]. This programmatic access can be used in scripts and Jupyter notebooks to search and download large sets of data directly.

8 ISIMIP Repository: `https://data.isimip.org`.
9 ISIMIP Publisher on GitHub: `https://github.com/ISI-MIP/isimip-publisher`.
10 ISIMIP Caveats and Updates: `https://data.isimip.org/caveats`.
11 ISIMIP Files API on GitHub: `https://github.com/ISI-MIP/isimip-files-api`.
12 ISIMIP Client on GitHub: `https://github.com/ISI-MIP/isimip-client`.

5 Discussion and Outlook

As the world's largest data archive of model-based climate impact data, ISIMIP output data is used by a diverse audience inside and outside of academia for research and analyses. It is therefore crucial to ensure a high-quality of the published data, both formally and content-wise. Using the machine-readable protocol as *single source of truth* for quality control, metadata extraction and data publication has greatly improved our workflow. Using a git-based workflow allows us to adjust and track the changes suggested by the impact modeling sectors. Changes only need to be made in one place, and the history of the protocol is available in a transparent and open way.

The presented workflow has proven useful to ensure the conformance of the provided data with the simulation protocol throughout the whole data publication chain. In particular the possibility for modellers to check their output files prior to the upload to the ISIMIP data space, using the same tools as in the final quality control step, has been productively adopted by several modelling groups.

Although the work presented is tailored to the ISIMIP project, we believe it can be seen as a best practice example for similar collaborations that have a common set of simulations experiments and a data curation process. The combination of a machine-readable protocol, automatic deployment on the Internet using Continuous Integration (CI), and subsequent use by tools and services can be useful for many data intensive research areas. All the components presented are available as open source software and can be adopted for other contexts.

In the future, we will extend the quality control process, which currently only checks for formal deviations from the ISIMIP protocol and some minimum or maximum violations, to include a quality assessment of the data content itself. We are working on a new tool[13], which is able to automatically check model outputs against an ensemble of already published models. This will allow us to identify and correct not only formal errors to formats and naming conventions, but errors in the data itself.

Acknowledgements

This research has received funding from the German Federal Ministry Ministry of Education and Research (BMBF) under the research project ISIAccess (16QK05) and from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821010.

References

Frieler, Katja, Stefan Lange, Franziska Piontek, Christopher P. O. Reyer, Jacob Schewe, Lila Warszawski, Fang Zhao, et al. 2017. "Assessing the impacts of 1.5°C global

13 ISIMIP Quality Assessment on GitHub: `https://github.com/ISI-MIP/isimip-qa`.

warming – simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b)". *Geoscientific Model Development* 10 (12): 4321–4345. DOI: https://doi.org/10.5194/gmd-10-4321-2017.

Marcé, Rafael, Donald Pierson, Daniel Mercado-Bettin, Wim Thiery, Sebastiano Piccolroaz, Bronwyn Woodward, Richard Iestyn Woolway, et al. 2022. *ISIMIP2b Simulation Data from the Local Lakes Sector.* Version 1.0. DOI: https://doi.org/10.48364/ISIMIP.563533.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Empowering Data at Leeds Beckett University: Understanding Institutional Needs and Applying Best Practice

Amy Campbell

Leeds Beckett University

1 Introduction

Whilst historically a teaching-focused institution, research output is growing exponentially at Leeds Beckett University (LBU). Therefore, LBU's Library and Student Services (LSS) must be ambitious in developing the breadth and quality of its support to researchers, whilst pragmatically recognising curtailing factors such as budgets and staff capacity. Many university library services recognise the challenge of meeting the changing needs of their institution and will develop strategies to adapt.

Open Data is the latest frontier of Open Science but detailed understanding of existing LBU researcher knowledge on research data has been absent. Meanwhile, research-intensive institutions have greater experience of delivering Research Data Management (RDM) support compared to LBU, so gaining an understanding of best practice from them would be beneficial for shaping future services.

As part of an Arts and Humanities Research Council and Research Libraries UK (AHRC-RLUK) Professional Practice Fellowship, this study aimed to conduct a literature review on RDM support, research LBU researcher knowledge and support needs regarding Open Data, gain best practice from other UK universities on supporting RDM, and then make recommendations on university RDM service development.

2 Literature Review

Research Data Management (RDM) describes activities which manage research data through the lifecycle of a project and in the last decade it has become a strategic priority for universities (Cox and Pinfield 2013; Oo et al. 2021; Andrikopoulou, Rowley, and Walton 2021). Open Data is data made freely available for anyone to use under licence and is based on the principal that publicly funded research should be made publicly available.

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.c18078

Open Data has practical, methodological, and potentially ethical issues, requiring detailed consideration by researchers even prior to starting a research project. Therefore, good RDM training is essential for "maximising the potential of Open Data" (Childs et al. 2014, p.154).

Parsons et al. (2011) suggest socio-cultural factors pose a greater barrier to Open Data than technical issues, and Chen et al. (2018, p.113) echo the need for a cultural shift towards the "...pursuit of reusability..." in researcher practice. Although funder mandates have galvanised academic libraries to promote and support Open Data, incentivising researchers to share their data requires both cultural change at institutional level and discipline-specific support (Woods and Pinfield 2022). This suggests publishers, funders and research institutions should seek to normalise data sharing behaviour and support the development of communities of practice where researchers share skills, mentor each other, and collaborate with Open Data (see Levine et al. 2020).

Librarians require a proactive approach to Continued Professional Development and managerial support to access relevant training to be both competent and confident when delivering RDM training and influencing researcher behaviour (Rachlin 2022). Best practice in RDM training has been identified (see Oo et al. 2021) and implementing it has the potential to not only increase Open Science behaviours, but also build a new identity for library services (Andrikopoulou, Rowley, and Walton 2021; Childs et al. 2014).

3 Methods

This mixed-methods study had two elements:

- **Open Data Questionnaire**: To understand Open Data knowledge and identify training and support needs, an internal, online questionnaire was promoted between Aug-Oct 2022 to all LBU researchers. 11 quantitative questions were asked, plus opportunity for participants to add qualitative detail to responses.
- **Best Practice in RDM Support Interviews**: To gain an understanding of best practice in supporting research data across the sector, online and in-person meetings were held between Oct-Nov 2022 with library teams supporting Research Data Management at four UK Higher Education institutions: University of Leeds, University of Sheffield, De Montfort University and Edge Hill University.

4 Results

4.1 Open Data Questionnaire

51 responses were received from all but one of Leeds Beckett's nine academic Schools. Whilst most researchers knew what Open Data was (67%), the majority had not used it for their own research (57%). Most researchers said they required further guidance or support, with the most popular types being "Practical guidance on how to make data

open", "A named person/service to go to for support" and "Practice guidance on the potential risks to making data open".

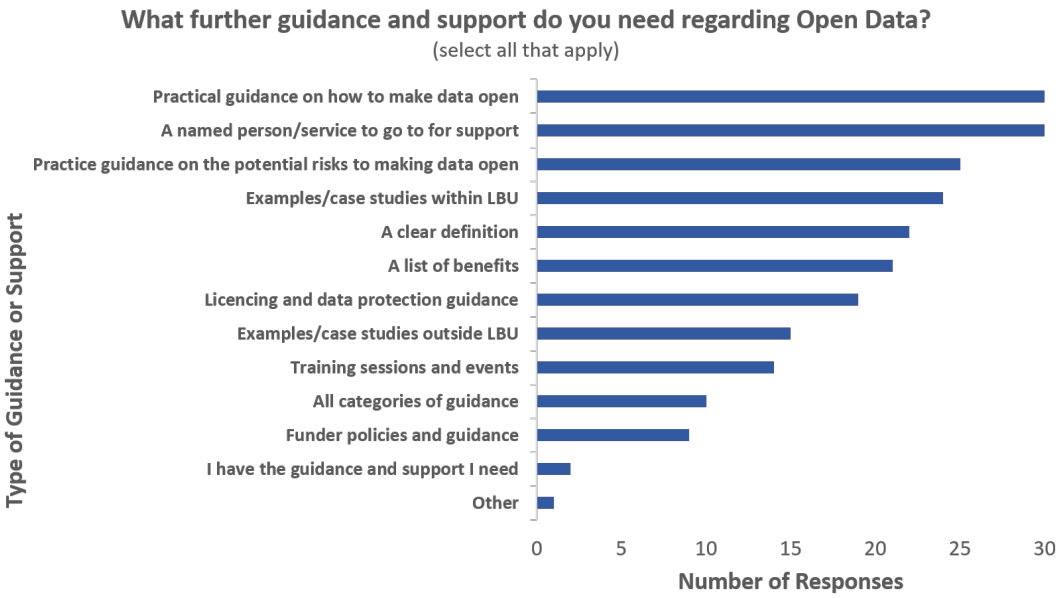


Figure 1: Questionnaire results for question "What further guidance do you need regarding Open Data?".

The key barriers cited to making data open were "concerns about how my data will be used by others" and "lack of support and guidance". Other barriers included not knowing who to go to for support and "lack of infrastructure". Eight respondents were concerned about the "robustness" of their data. Several qualitative responses to this question were received, revealing ethical concerns about making participant details open and beliefs about open data being irrelevant to the Humanities.

Respondents were asked "Do you think open data will increase collaborative research opportunities for you in the next 3 years?" and the majority said they were "Unsure" (49%) although 31% responded "Yes".

4.2 Best Practice in Research Data Management (RDM) Support Interviews

Between the four UK universities interviewed, there were significant differences in RDM team sizes, demonstrating scalability for service provision, dependant on budgets and institutional needs. Removing jargon and "thinking like a researcher" were key recommendations from De Montfort University. Supporting researchers through discipline-specific examples and demonstrating positive outcomes gained buy-in and improved engagement. Online support (e.g. webpages, short videos, tutorials) met most researcher support needs, enabling the single staff member at De Montfort University to focus on complex queries.

The University of Sheffield (UoS) and University of Leeds (UoL), as research-intensive universities, had significantly larger teams to support RDM, but still needed strategies to manage demand. The UoS provided comprehensive RDM training and website guidance. Teaching sessions on Data Management Plans, for example, were considered valuable to reduce future RDM issues, especially for postgraduate and early career researchers. UoS recommend their researchers deposit their data in a discipline-specific repository where possible but use their institutional data repository if necessary. This in beneficial in reducing staff time in processing datasets internally, but means the institution has less knowledge and oversight of data being produced by their researchers.

At UoL, a large and dedicated team support RDM, providing online guidance, training and 1-2-1 sessions. They have multiple repositories for data and other outputs, creating complexity in workflows, processes, and the need for technical skills. Meeting individual researchers to understand their support needs was valuable but time intensive. The institution supports researchers from a wide range of disciplines and a key challenge was ensuring the service was equally supportive of all. Overall, best practice was revealed to include detailed, multifaceted online guidance, training sessions pitched for different knowledge levels, and specialist RDM staff to handle individual researcher enquiries from multiple disciplines.

5 Discussion and Recommendations

The Open Data questionnaire findings provide the first ever insight into LBU researchers' knowledge and needs. They suggest that whilst there was a good, general awareness of Open Data amongst participants, Library and Student Services (LSS) should provide practical guidance on making data open, as recommended in the systematic review by Oo et al. (2021).

Interestingly, respondents from all academic Schools gave a variety of answers, suggesting knowledge and interest was uneven regardless of discipline. Providing tailored sessions for different levels of understanding (e.g. beginners/intermediate/advanced) would be beneficial to meet institutional needs and as suggested by Oo et al. (2021), likely successful at increasing knowledge. The best practice interviews also showed this to be beneficial for RDM support. The challenge for library staff in any institution, and as discussed in Rachlin (2022), is to provide this breadth of training within staff resources and ensuring those staff feel confident and competent to provide training.

LBU implemented an instance of the Figshare repository platform in Autumn 2022, providing institutional infrastructure for research data, but the Open Data questionnaire results revealed participants were often unaware of it. Meanwhile, best practice interviews suggested discipline-specific data repositories were often preferable to institutional repositories. Therefore, a challenge for library services is to develop the knowledge and skills to support researchers to deposit data in the most suitable place for their specific needs. Furthermore, findings of this study suggest that practical guidance on identifying repositories and depositing data is essential. This study highlights the need for increased

Table 1: Recommendations for Research Data Management (RDM) service development at Leeds Beckett University.

Recommendation Theme	**Recommendation Description**
1. Training and Support	1.1. Create practical RDM guidance that covers the whole research lifecycle for the Library service's webpages e.g. • FAIR principles • Writing a DMP • Benefits of Open Data • How to identify and use discipline-specific data repositories 1.2. Run training sessions pitched at different levels of knowledge for staff and Postgraduate Research students 1.3. Create short videos of the above practical guidance 1.4. Offer Schools tailored RDM training sessions
2. Institutional Relationships	2.1. Partner with Schools to: • Identify and recruit data 'champions' • Understand discipline-specific needs • Identify potential case studies of good practice • Recruit researchers to review Library support and guidance • Develop links with administrative staff to aid communication and knowledge-sharing 2.2. Engage with the senior University staff on Open Science research culture
3. Communication	3.1. Position the Library as the key source of research data support using the improved guidance and training: • Utilise internal newsletter to communicate support • Write up case studies and promote

emphasis on ethical issues in research data, including guidance and training on how to consider Open Data prior to gaining ethical approval and when writing DMPs. This needs to be relevant and tailored to all disciplines where feasible.

Surprisingly, findings demonstrate there are potential benefits to being a smaller, less research-intensive institution. Compared with some universities interviewed, LBU has fewer systems to manage and a smaller number of disciplines to support. Therefore, for LBU and similar universities, there may be opportunity, for example, to identify key researchers in different Schools to have in-depth conversations to understand their challenges and concerns around RDM. This would develop a library service's existing role of providing holistic support to the research community. For LBU, this deeper knowledge would build on the results of the Open Data questionnaire and further inform the support provided.

In conclusion, the literature review, Open Data questionnaire results and best practice interviews enable a clear set of recommendations on RDM service development to be made. These are grouped under themes of Training and Support, Institutional Relationships, and Communication. Key recommendations are provided in Table 1.

References

Andrikopoulou, Angeliki, Jennifer Rowley, and Geoff Walton. 2021. "Research Data Management (RDM) and the Evolving Identity of Academic Libraries and Librarians: A Literature Review". *New Review of Academic Librarianship* 28 (4): 349–365. DOI: https://doi.org/10.1080/13614533.2021.1964549.

Chen, Xiaoli, Sünje Dallmeier-Tiessen, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, et al. 2018. "Open is not enough". *Nature Physics* 15 (2): 113–119. DOI: https://doi.org/10.1038/s41567-018-0342-2.

Childs, Sue, Julie McLeod, Elizabeth Lomas, and Glenda Cook. 2014. "Opening research data: issues and opportunities". Edited by Dr Anne Thurston. *Records Management Journal* 24 (2): 142–162. DOI: https://doi.org/10.1108/RMJ-01-2014-0005.

Cox, Andrew M., and Stephen Pinfield. 2013. "Research data management and libraries: Current activities and future priorities". *Journal of Librarianship and Information Science* 46 (4): 299–316. DOI: https://doi.org/10.1177/0961000613492542.

Levine, Robert M., Kristen E. Fogaren, Johna E. Rudzin, Christopher J. Russoniello, Dax C. Soule, and Justine M. Whitaker. 2020. "Open Data, Collaborative Working Platforms, and Interdisciplinary Collaboration: Building an Early Career Scientist Community of Practice to Leverage Ocean Observatories Initiative Data to Address Critical Questions in Marine Science". *Frontiers in Marine Science* 7. DOI: https://doi.org/10.3389/fmars.2020.593512.

Oo, Cherry Zin, Adrian W. Chew, Adeline L. H. Wong, Joanne Gladding, and Cecilia Stenstrom. 2021. "Delineating the successful features of research data management training: a systematic review". *International Journal for Academic Development* 27 (3): 249–264. DOI: https://doi.org/10.1080/1360144x.2021.1898399.

Parsons, Mark A., Øystein Godøy, Ellsworth LeDrew, Taco F. de Bruin, Bruno Danis, Scott Tomlinson, and David Carlson. 2011. "A conceptual framework for managing very diverse data for complex, interdisciplinary science". *Journal of Information Science* 37 (6): 555–569. DOI: https://doi.org/10.1177/0165551511412705.

Rachlin, David J. 2022. "Academic Librarians and Research Data Services: Preparation and Attitudes Revisited". *Internet Reference Services Quarterly* 26 (4): 199–211. DOI: https://doi.org/10.1080/10875301.2022.2072042.

Woods, Helen Buckley, and Stephen Pinfield. 2022. "Incentivising research data sharing: a scoping review". *Wellcome Open Research* 6:355. DOI: https://doi.org/10.12688/wellcomeopenres.17286.2.

NFDI4DS – NFDI for Data Science and Artificial Intelligence

Sonja Schimmler

Fraunhofer FOKUS

NFDI4DataScience (NFDI4DS) supports researchers along all stages of the research data lifecycle to conduct their research in line with the FAIR principles. An infrastructure is developed targeting researchers from a wide range of disciplines working in data science and artificial intelligence.

By regularly conducting interviews and surveys, NFDI4DS identifies the needs and challenges of researchers from various disciplines regarding data science and artificial intelligence, keeping ethical, legal, and social aspects in mind. Those identified needs and challenges are continuously addressed by picking up existing services, developing new ones and integrating them into the NFDI4DS infrastructure. By systematically adding digital objects (articles, data, models, workflows, scripts/code, etc.) to the NFDI4DS research knowledge graph within the infrastructure, transparency, reproducibility, and fairness are steadily improved. Support structures, including interactive learning materials and community events, accompany the whole process.

This short paper gives an overview of NFDI4DS and its work programme. It provides details about its approach to address the current challenges. It also gives an overview of the services planned, and how they are meant to interact.

1 Introduction

The past years have seen a paradigm shift, with computational methods increasingly relying on data-driven and often deep learning-based approaches, leading to the establishment of data science as a discipline driven by advances in the field of computer science. Transparency, reproducibility and fairness have become crucial challenges for data science (DS) and artificial intelligence (AI) due to the complexity of contemporary DS methods, often relying on a combination of scripts/code, workflows, models, and data.

The *vision* of NFDI4DS is to support all steps of the complex and interdisciplinary research data lifecycle, including collecting/creating, processing, analyzing, preserving, accessing, and reusing resources in DS and AI. The *overarching objective* of NFDI4DS is the

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.c18079

development, establishment, and sustainment of a national research data infrastructure (NFDI) for the DS and AI community in Germany. This will also deliver benefits for a wider community requiring data analytics solutions, within the NFDI and beyond. The *key idea* is to work towards increasing the transparency, reproducibility and fairness of DS and AI projects, by making all digital objects available, interlinking them, and offering innovative tools and services.

2 Challenges and Approach

Sharing scientific knowledge is not just about publishing articles. Instead, it involves documenting the entire research data lifecycle and providing a multitude of digital objects in compliance with the FAIR principles by making them findable, accessible, interoperable and reusable. As DS and AI are continuously evolving, the methods used become more complex, and it is difficult to maintain transparency, reproducibility, and fairness in research. Challenges related to ethical, legal, or social aspects further limit the willingness and/or ability of researchers to conduct, archive, or publish their research in line with the FAIR principles.

NFDI4DS[1] is part of the NFDI initiative to build a German National Research Data Infrastructure. It supports all stages of the complex and interdisciplinary research data lifecycle to enable the efficient and effective reuse of research data and other digital objects. Additionally, the consortium steadily contributes to establishing best practices in research, fostering open science to enable researchers to make full use of valuable resources.

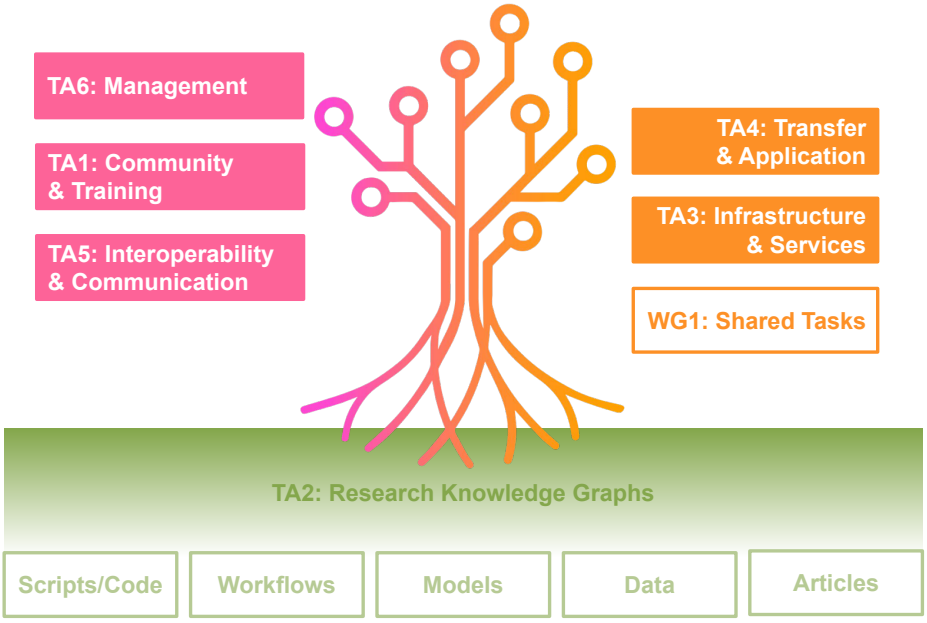


Figure 1: NFDI4DS Task Areas.

1 https://www.nfdi4datascience.de

NFDI4DS is organized around six task areas (see Figure 1): (1) Community and Training, (2) Research Knowledge Graphs, (3) Infrastructure and Services, (4) Transfer and Application, (5) Interoperability and Cooperation, and (6) Management. In addition, working groups are temporarily set up on further important topics such as (1) Shared Tasks.

NFDI4DS intends to represent the DS and AI community in academia, which is an interdisciplinary field rooted in computer science. The consortium currently focuses on four DS intense application areas: (1) language technology and natural language processing, (2) biomedical research and clinical decision-making, (3) information sciences and (4) social sciences. Further application areas are involved via speedboat projects later on.

By regularly conducting interviews, insights are gathered on the needs and challenges of researchers, especially about ethical, legal, and social aspects. Systematically conducted surveys identify gaps for new implementations, as well as tools and services that already exist and are useful for the NFDI4DS infrastructure. As DS and AI are important in many disciplines, with often contradicting requirements, building an infrastructure is a collaborative effort involving the scientific communities.

Support structures are accompanying the whole process, addressing the identified needs and challenges. Interactive training materials such as educational videos are provided, and community events such as challenges are organized.

By regularly providing benchmark datasets and fostering joint work on shared tasks interdisciplinary as well as domain-specific services and solutions are achieved. Each shared task focuses on a specific aspect of the research data lifecycle and has the goal to initiate a concrete service that is being integrated into the NFDI4DS infrastructure later on.

3 Core Services

The core services focus on six main areas around digital objects: collecting/creating, processing, analysing, preserving, accessing, and reusing (see Figure 2). Digital objects include artefacts beyond articles, such as data, models, workflows, and scripts/code.

The NFDI4DS infrastructure is based on a number of already existing software components and already well-established tools and services, which target different phases of the research data lifecycle. There are also new technologies, tools, and services uncovered regularly which will continuously be integrated in the NFDI4DS infrastructure. Our service integration strategy is inspired by the EOSC Interoperability Framework, which is based on: (1) persistent identification using PIDs, such as DOI, ORCID or authoritative URIs, (2) authentication and authorisation (AAI) adhering to common standards, (3) semantic interoperability using RDF, vocabularies and ontologies, and (4) API integration based on REST principles.

The NFDI4DS research knowledge graph forms the basis of the infrastructure, providing details about digital objects and their interrelation. Key elements of the infrastructure are the NFDI4DS gateway and portal as well as the NFDI4DS registries and repositories.

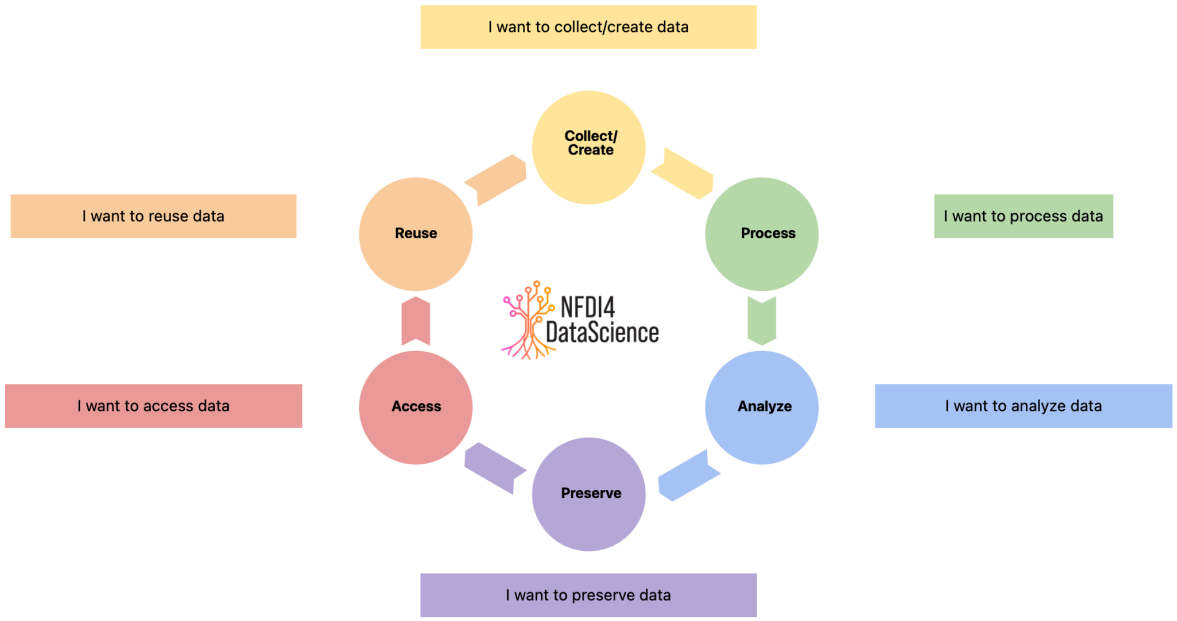


Figure 2: Core Services Dashboard.

Digital objects are harmonized, aggregated, and preserved via the repositories and exposed via the registries and the gateway and portal.

NFDI4DS Research Knowledge Graph. The research knowledge graph will entail automatically extracted metadata about resource relations. The component makes use of *the Open Research Knowledge Graph (ORKG)*, a service for semantically describing research contributions in a knowledge graph. The semantic descriptions of articles are crowd-sourced from authors and researchers leveraging NLP of articles. The component also utilizes *the GESIS Knowledge Graph Infrastructure*, which consists of tools and pipelines for constructing actual research knowledge graphs of research information, metadata and primary research data.

NFDI4DS Registries and Repositories. The consortium aims at providing registries and repositories for different digital objects. One registry being integrated is *the DBLP Computer Science Bibliography*, an open bibliographic data base, search engine, and knowledge graph on computer science publications.

NFDI4DS Gateway and Portal. Through a unified and intuitive search interface, users are enabled to query a wide range of scientific databases such as DBLP, Zenodo, and OpenAlex. While the gateway queries APIs in an ad-hoc fashion, the portal provides a harvesting-based service. The component makes use of *the Data Management Platform Piveau*, which provides services and pipelines for harvesting data and metadata from various sources saving it into a knowledge graph utilizing semantic linking.

NFDI4DS Tools and Services. The consortium aims at integrating different tools and services, including a JupyterHub instance. To facilitate the further analysis and visual-

isation of digital objects contributed by the participating services, they can be directly loaded into a JupyterHub instance for further programmatic processing. The component utilizes *GESIS Notebooks*, an online reproducibility service for FAIR digital objects. Its main components are a BinderHub, a JupyterHub, and a place to publish, explore, try out, and learn about DS and AI methods.

NFDI4DS Compute Infrastructure. The consortium aims at providing a compute infrastructure. While most of the NFDI4DS tools and services will be hosted in a cloud infrastructure located at various sites of the partners, some DS and AI tasks require access to specialized high-performance computing (HPC) resources such as GPU accelerators and large memory capacities.

4 Conclusions

This short paper gives an overview of NFDI4DS and its work programme. It provides details about its approach to address the current challenges. It also gives an overview of the services planned, and how they are meant to interact.

Acknowledgements

This joint project received funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number: NFDI4DataScience (460234259). We thank all project partners for their contributions.

Mit maßgeschneiderten Metadatenprofilen zu validierten und nachhaltigen Forschungsdaten

Matthias Grönewald [1], Nils Preuß [2]

[1]Universitäts- und Landesbibliothek, Technische Universität Darmstadt;
[2]Institut für Fluidsystemtechnik, Technische Universität Darmstadt

Zeitgemäßes Forschungsdatenmanagement (FDM) beinhaltet zunehmend auch die Integration reichhaltiger, maschinennutzbarer Metadaten, allgemein zur Sicherstellung wissenschaftlicher Qualität insbesondere aber im Kontext von Reproduzierbarkeit und Nachnutzung. Bestehende Standards umfassen meist nur deskriptive Metadaten und die in umfassenderen Metadatenschemata enthaltenen Informationen sind in der Regel weder standardisiert noch maschinennutzbar. Fachspezifische Metadaten sind jedoch notwendig, um Forschung präzise und reichhaltig zu dokumentieren. Die Abläufe und Werkzeuge dafür sind jedoch nicht umfassend verfügbar. Ein vielversprechender Ansatz ist die Anwendung von Metadatenprofilen, die es ermöglichen hochspezifische Terminologien, aufbauend auf bestehenden Community-Standards, in flexible und interoperable Metadatenbeschreibungen zu überführen. Basierend auf etablierten Technologien ermöglichen Metadatenprofile eine Lösung zum Gestalten und Verarbeiten von komplexen, maschinennutzbaren und letztlich FAIRen Metadaten.

Anhand eines Beispiels aus den Ingenieurswissenschaften, wird die Datenvalidierung mittels Metadatenprofilen basierend auf kontrolliertem Vokabular gezeigt. Dieser Prozess kann dann zu fast jedem Zeitpunkt im Lebenszyklus von Forschungsdaten genutzt werden. Ein Anwendungsbeispiel demonstriert außerdem die sich daraus ergebenden Möglichkeiten im Bereich der Datenanalyse bzw. der Archivierung.

Erst die Kombination aus praktischer Integration in die Forschungslandschaft in Verbindung mit der Umsetzung in verschiedenen FDM-Projekten, Initiativen und Werkzeugen ermöglicht die für eine Standardisierung notwendigen Synergien. Damit wird die Forschung durch höhere Datenqualität gefördert, sowie für die nachhaltige Bewahrung von Forschungsinhalten durch spezifischere Dokumentation ein Mehrwert gebildet.

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.c18080

1 Einleitung, Ziele und Bedarf

Metadaten, allgemein oft als „Daten über Daten“ oder „Daten die Daten beschreiben“ (Furner 2019) charakterisiert, sind in der Forschung von entscheidender Bedeutung. Sie beschreiben Zusammenhänge, wie zeitliche Bezüge, administrative Zuordnung zu Projekten oder auch Urheberschaft. Neben solchen weit verbreiteten deskriptiven, administrativen oder bibliographischen Metadaten, die sich in der Forschungslandschaft weitgehend standardisiert etabliert haben, sind unterschiedliche Informationen über eingesetzte Methoden, Software, Materialien und Abläufe (Abbildung 1) notwendig um Forschung nachvollziehbar und reproduzierbar zu machen (Deutsche Forschungsgemeinschaft e.V. 2019). Damit sind sie zwar von mindestens genauso entscheidender Bedeutung, werden jedoch in der Praxis allerdings weit weniger wahrgenommen.

Um erstgenannte Metadaten zu erfassen und beschreiben liegen standardisierte Terminologien und Metadatenschemata vor[1], auch im Bezug auf Forschungsdaten (Grönewald u. a. 2023a; Albertoni u. a. 2023; DataCite Metadata Working Group 2021), für nachfolgend genannte fachspezifische Metadaten jedoch nicht. Auch ist nicht klar ob die moderne stark heterogene Forschung, selbst innerhalb einer Disziplin eine solche Standardisierung zulässt. Gerade das Forschungsdatenmanagement von großen Datenmengen stellt allerdings an Dokumentation und Datenbeschreibung umfassende Herausforderungen. Forschende, die damit konfrontiert sind, gestalten dann eigene Lösungen, teilweise mit hoch individuellen Metadatenschemata, die in Ausgestaltung und Betrieb aufwendig und in der Regel nicht auf andere Forschende übertragbar sind. Letzteres ist besonders mit Blick auf Nachnutzung, aber auch Archivierung, von Forschungsdaten nachteilig. Im Rahmen des DFG-Projekts AIMS (Grönewald u. a. 2023b) wurde eine Softwareplattform entwickelt, die es erlaubt stattdessen interoperable und trotzdem hochspezifische Metadatenprofile zu gestalten und zu teilen.

Metadatenprofile, technisch als Applikationsprofile umgesetzt, erlauben es durch die Beschreibung von Anforderungen Eigenschaften von Metadatensätzen festzulegen bzw. gemäß diesen zu validieren (Coyle 2017). Neben dem Bedarf bei der Handhabung, Analyse und Dokumentation von Daten in der aktiven Forschung, sind fachspezifischen Metadaten insbesondere auch notwendig zur Erstellung und Veröffentlichung von Datensätzen gemäß den FAIR Prinzipien (Wilkinson u. a. 2016). Am deutlichsten wird das an der Nachnutzbarkeit von Forschungsdaten, die durch detaillierte Metainformationen insbesondere verbessert wird, aber durch die Bereitstellung von Metadatenprofilen gemäß etablierter Standards (RDF, SHACL, etc.) werden auch Auffindbarkeit, Zugänglichkeit und Interoperabilität gestärkt. Letztlich umfasst der Anwendungsbereich damit den gesamten Forschungsdatenlebenszyklus. Anhand von Beispielen aus den Ingenieurwissenschaften wird gezeigt, in welcher Weise maßgeschneiderte Metadatenprofile es erlauben Forschungsmetadaten gemäß den FAIR-Prinzipen zu beschreiben und diese in im aktiven Forschungsdatenmanagement zur Datenvalidierung zu nutzen.

1 Vgl. (a) `http://dublincore.org/documents/dcmi-terms/`; (b) `http://www.rdaregistry.info/Elements/u/#`, (c) `https://github.com/tibonto/DFG-Fachsystematik-Ontology`, uvm., *zuletzt aufgerufen am 12. Mai 2023.*

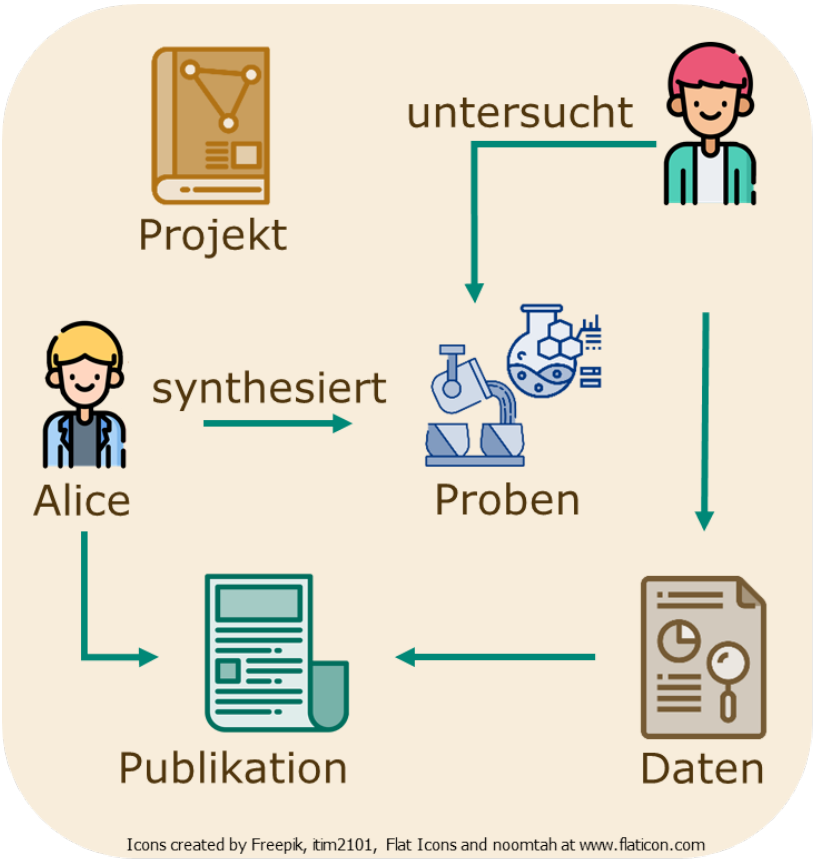


Abbildung 1: Forschungsmetadaten - Neben deskriptiven und administrativen Angaben zu Projektbezug, Forschenden oder Forschungsdisziplin, stellen die Bezügen zwischen einzelnen Forschungsdaten, sowie Metadaten über angewandte Parameter, eingesetzte Geräte oder Analyseverfahren, wichtige Informationen zum Verständnis der eigentlichen Inhalte dar. In aller Regel unterliegen sie hochspezifischen disziplinabhängigen Bedingungen. Es handelt sich um fachspezifische Forschungsmetadaten.

2 Reichaltige fachspezifische Metadaten erlauben FAIRe Inhalte

Für die (Nach-) Nutzung von Forschungsdaten, das Vorhandensein von beschreibenden Forschungsmetadaten zentral. Sie werden benötigt um eine Bewertung der Beschaffenheit der Daten und ihrer Eignung für das jeweilige Nutzungsszenario zu erlauben, und nehmen daher auch im Kontext der FAIR-Prinzipien eine besondere Stellung ein.

Es ist also zweckdienlich, zunächst die Erfüllung formaler Kriterien sicherzustellen, d.h. zu validieren. Dies wird bereits am einfachen Beispiel eines Messwertes ersichtlich: Unabhängig vom Nutzungsszenario ist es für ingenieurwissenschaftliche Versuche unerlässlich, dass für jede Messgröße die physikalische Dimension (Länge, Masse, Zeit, etc.), und für jeden Messwert die physikalische Einheit (Meter, Kilogramm, Sekunde, etc.) unmissverständlich angegeben ist. Gleiches gilt etwa für den Zeitpunkt und die genaue Messstelle, außerdem die Herkunft des Messwertes, etwa Messinstrument und Messverfahren, bzw. Messunsicherheit. Ist die Erfüllung solcher Kriterien nicht sichergestellt, ist eine Beurteilung der Eignung der Daten für ein bestimmtes Nutzungsszenario nicht möglich oder zumindest deutlich erschwert.

```
soil:MeasurementShape
    a sh:NodeShape ;
    sh:name "measurement"@en ;
    sh:property [
       sh:path qudt:hasQuantityKind ;
       sh:name "quantity"@en ;
       sh:minCount 1 ;
       sh:maxCount 1 ;
       sh:class qudt:QuantityKind ;
    ] ;
    sh:property [
       sh:path qudt:applicableUnit ;
       sh:name "unit"@en ;
       sh:minCount 1 ;
       sh:maxCount 1 ;
       sh:class qudt:Unit ;
    ] .
```

```
ex:TemperatureReading
    qudt:hasQuantityKind quantitykind:Temperature ;
    qudt:applicableUnit unit:DEG_C ;
    schema:value "20.54513"^^xsd:float .
```

Abbildung 2: Links das Metadatenprofil einer Messung, rechts ein valider Messwert eines Temperatursensors. Das Profil bedingt dabei die zwei Eigenschaften, das Vorhandensein einer Größe samt Einheit.

Wenn gleich der manuelle Aufwand und die benötigte fachliche Expertise zur letztlichen Bewertung von Daten nur schwer reduzierbar ist, kann die Überprüfung von formalen Kriterien, beispielsweise das bloße Vorhandensein bestimmter bewertungsrelevanter Informationen, aber auch anderer Aspekte der FAIR-Prinzipien maschinell stattfinden. Für eine solche automatisierte Validierung werden formalisierte Vorgaben benötigt, z.B. in Form von maschinennutzbaren Metadatenprofilen. Im Rahmen des vorgestellten Projekts werden RDF (Schreiber und Raimond 2023) und SHACL (Knublauch und Kontokostas 2023) hierfür als Basis-Technologien eingesetzt. Hierdurch werden bereits einzelne FAIR Prinzipien (Wilkinson u. a. 2016) adressiert (im folgenden referenziert durch ihre Nummerierung gemäß der Veröffentlichung, z.B. F1 für das erste FAIR Prinzip).

Der Einsatz einer formalen, zugänglichen, gemeinsamen und breit anwendbaren Sprache zur Wissensdarstellung ist sichergestellt (I1). Zugleich sind die Beschränkung auf Vokabu-

lare und Terminologien, die den FAIR-Prinzipien genügen (I2), und Verweise auf andere (Meta-) Daten (I3) einfach realisierbar.

Darüber hinaus kann analog zum oben erläuterten Beispiel die Angabe von Metadaten überhaupt (F2), bis hin zu bestimmten relevanten Attributen (R1), inkl. Nutzungslizenz (R1.2) und Herkunft (R1.2) sichergestellt werden. Dies ist individuell auf Konventionen und Standards der jeweiligen Fachrichtung anpassbar (R1.3), hier der Ingenieurwissenschaften. Zusätzlich sind weiter Spezialisierungen, beispielsweise für den Werkzeugmaschinenbaus sind umsetzbar und bleiben interoperabel.

Die Angabe von Identifiern für Datensätze (F4), sowie speziell die Nutzung von global einzigartigen und persistenten Identifiern wie DOI, Handle, w3id, etc. (F1) lassen sich durch geeignete Metadatenprofile überprüfen. In Verbindung mit geeigneten Repositorien können schließlich die übrigen Aspekte der FAIR-Prinzipien abgedeckt werden: Indexierung und Suche (F3), Abruf via Identifier (A1) und standardisierte, offene Protokolle (A1.1), inkl. Authentifizierung und Autorisierung (A1.2), Zugänglichkeit von Metadaten unabhängig von den Daten selbst (A2).

Natürlich ist dies nicht nur auf Daten und ihre Metadaten anwendbar, sondern gilt in gleichem Maße für die Metadatenprofile selbst: als Ressourcen verstanden, sollen sie den FAIR-Prinzipien ebenso genügen.

3 Datenvalidierung mittels formalisierter semantischer Metadaten

Eine direkte Anwendung findet sich nun z.B. bei der Validierung von Datensätzen. Hier Sensordaten eines Temperatursensors, bestehend aus einer Größe Temperature in der Einheit Celsius, gezeigt auf der rechten Seite der Abbildung 2. Ein zugehöriges Profil, dass allgemein einen Messwert beschreibt ist links dargestellt, es verlangt genau eine Größe mit genau einer Einheit. Nicht gezeigt ist hier die Verknüpfung mit allgemeineren Profilen, die wie im obigen Abschnitt beschrieben allgemeinere Aspekte der FAIR-Prinzipien validieren.

Das in Abbildung 2 links gezeigte Profil validiert auch Daten die von einem Lasertracker generiert werden, hier für die Größe Länge und die Einheit Meter, siehe Abbildung 3.

```
ex:TargetPosition
    qudt:hasQuantityKind quantitykind:Length ;
    qudt:applicableUnit unit:M ;
    schema:value (
        "-0.226959617"^^xsd:float
        "-0.3047850568"^^xsd:float
        "15.10344412"^^xsd:float
    ) .
```

Abbildung 3: Das in Abbildung 2 gezeigte Profil einer Messung validiert auch den hier gezeigten Messwert eines Lasertrackers, da die Eigenschaften (Vorhandensein von Größe und Einheit) ebenso erfüllt sind.

Durch eine hierarchische und modulare Modellierung der Metadatenprofile wird eine Individualisierung von Metadatenprofilen ermöglicht. Abbildung 4 zeigt das Profil eines Temperaturmesswertes, basierend auf dem allgemeineren Messwertprofil von zuvor.

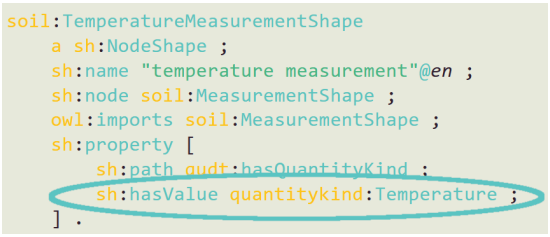


Abbildung 4: Wird zusätzlich die Art der Größe des in Abbildung 1 gezeigten Profils eingschränkt, entsteht ein interoperables aber spezifischeres Profil, das nur noch den in Abbildung 1 gezeigten Messwert, nicht jedoch den des Lasertrackers aus Abbildung 2 validiert.

Hier ist nun zusätzlich vorgeschrieben, dass die physikalische Größe nicht nur angegeben sein muss, sondern außerdem der Art Temperatur sein muss, wodurch das Profil die Temperaturdaten validiert, aber nicht die des Lasertrackers. Es lassen sich also hochspezifische Beschreibungen erstellen, die trotzdem interoperabel sind, da sie gemeinsame Terme und Profile verwenden. Darüber hinaus sind sie maschinennutzbar und so in eine automatisierte Datenvalidierung integrierbar.

Mögliche Ziele kann dabei die Prüfung der Datenqualität anhand von formalen Kriterien bei Datenübergabe innerhalb von Projektstrukturen, vor Erfassung in Repositorien oder Austausch zwischen Forschungsinstitutionen sein, also sowohl noch im Bereich der Forschung, als auch später bei Fragen der Verfügbarmachung und Archivierung.

4 Zusammenfassung & Möglichkeiten in der Zukunft

Metadatenprofile sind auf Basis von RDF und SHACL technologisch umsetzbar und individualisierbar. Sie ermöglichen die automatisierte Überprüfung formaler Kriterien, und bilden damit einen wichtigen Baustein zur breiten Umsetzung der FAIR-Prinzipien und fachspezifischer Standards darüber hinaus. Speziell am Beispiel der Datenvalidierung von Sensorwerten kann gezeigt werden, dass damit hochspezifische Metadaten, wie die Beschreibung physikalischer Größen validierbar erfasst werden können und dabei sowohl untereinander interoperabel, gleichzeitig konform zu bestehenden Normen gehalten werden können. Damit leisten Metadatenprofile auf Basis gemeinsamer kontrollierter Terminologien einen wesentliche Beitrag für die Implementierung der FAIR Prinzipien, insbesondere der Gewährleistung von Interoperabilität.

Das benötigte Hintergrundwissen zur Modellierung, sowie erforderliche Kenntnisse über geeignete Vokabulare, Terminologien und existierende Profile bzw. Informationsmodelle bilden dennoch eine signifikante Einstiegshürde für Forschende die es zu überwinden gilt. Dieser Herausforderung widmen sich aktuell verschiedene Initiativen. Im Projekt AIMS website ist eines der Ziele die Entwicklung einer Plattform zum Gestalten und Teilen von

Metadatenprofilen zur Anwendung in den Ingenieurwissenschaften mit dem expliziten Fokus diese Hürde zu senken.

Eine entsprechende Verbreitung verbunden mit Bemühungen aus den Fachcommunities trägt so zu einer Standardisierung und Harmonisierung bei. Als Beispiel dafür wird die entwickelte Software im Kontext der Metadatendienste des Konsortiums NFDI4Ing (Schwarz und Anthofer 2023) zukünftig als communityweite Dienstleistung angeboten. Bei ausreichender Integration erlauben die Analyse von angewandten Terminologien, Modellierungskonzepten und Akzeptanzverteilung in Zukunft damit noch bessere Anpassungen der Dienste an die Communities und Aussagen über die entscheidenden Kristallisationspunkte einer Standardisierung, die es zu fördern gilt. Die Datenvalidierung ist dabei nur eine mögliche Anwendung, die jedoch großes Potential zu Integration in verschiedene Strukturen im Rahmen der Datenqualitätsanalyse und institutionellen Speicherung bietet. Der Mehrwert reichhaltiger und maschinennutzbarer Metadaten liegt dabei nicht nur bei Einrichtungen wie den Bibliotheken, sondern auch den Forschenden in der Praxis selbst. Niederschwellig gestaltbare Metadatenapplikationsprofile und die zugehörige Infrastrukturlandschaft bilden dabei einen wichtiges Hilfsmittel.

Danksagung

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) -Projektnummer 432233186. Die Universitäts- und Landesbibliothek Darmstad (ULB), das IT Center der RWTH Aachen University (ITC), der Lehrstuhl für Fluidsystemtechnik (FST), sowie das Werkzeugmaschinenlabor der RWTH Aachen University (WZL) sind Partner in der Umsetzung dieses Projekts.

ORCID:

- Matthias Grönewald https://orcid.org/0000-0002-3480-9102
- Nils Preuß https://orcid.org/0000-0002-6793-8533

Literaturverzeichnis

Albertoni, Riccardo, David Browning, Simon Cox, Alejandra Gonzalez Beltran, Andrea Perego und Peter Winstanley. 2023. „Data Catalog Vocabulary (DCAT) – Version 3“. Besucht am 12. Mai. https://www.w3.org/TR/vocab-dcat-3/.

Coyle, Karen. 2017. „Application Profiles“. In *Advances in Web Technologies and Engineering,* 1–15. IGI Global. DOI: https://doi.org/10.4018/978-1-5225-2221-8.ch001. https://doi.org/10.4018/978-1-5225-2221-8.ch001.

DataCite Metadata Working Group. 2021. „DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4“. DOI: https://doi.org/10.14454/3W3Z-SA82. https://schema.datacite.org/meta/kernel-4.4/.

Deutsche Forschungsgemeinschaft e.V. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis.* Zenodo. DOI: https://doi.org/10.5281/zenodo.6472827.

Furner, Jonathan. 2019. „Definitions of 'Metadata': A Brief Survey of International Standards“. *Journal of the Association for Information Science and Technology* 71 (6). DOI: https://doi.org/10.1002/asi.24295. https://doi.org/10.1002/asi.24295.

Grönewald, Matthias, Marc Fuhrmans, Nils Preuss, Benedikt Heinrichs, Sousan Homaeipour und Matthias Bodenbenner. 2023a. „Dataset Structured Data | Google Search Central | Documentation | Google Developers“. Besucht am 12. Mai. https://developers.google.com/search/docs/appearance/structured-data/dataset.

———. 2023b. „DFG-Projekt AIMS - Website“. Besucht am 12. Mai. https://www.aims-projekt.de/.

Knublauch, Holger, und Dimitris Kontokostas. 2023. „Shapes Constraint Language (SHACL)“. Besucht am 12. Mai. https://www.w3.org/TR/shacl/.

Schreiber, Guus, und Yves Raimond. 2023. „RDF 1.1 Primer“. Besucht am 12. Mai. https://www.w3.org/TR/rdf11-primer/.

Schwarz, Annett, und Verena Anthofer. 2023. „NFDI4Ing - Website“. Besucht am 12. Mai. https://nfdi4ing.de/.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Automated Software Metadata Conversion and Publication Based on CodeMeta

Marie Houillon[1], Jochen Klar[2], Tomas Stary[1], Axel Loewe[1]

[1]Karlsruhe Institute of Technology (KIT);
[2]Independent Software Developer

Reproducible research requires publication of software together with appropriate metadata. Different metadata standards exist for different steps in the research software publication process: the Citation File Format (CFF) became very popular to provide information on how users are supposed to cite the software, DataCite is one of the established standards for research data archiving and CodeMeta is an extension of schema.org specifically tailored to research software. If research software developers must maintain a whole set of metadata files in different formats with largely overlapping content, it poses a risk both to data consistency and to adoption of good software publication practices. Therefore, we developed pipelines that put developers in a position to only maintain a CodeMeta file, from which CFF and DataCite files are automatically generated. These pipelines can easily be integrated in continuous integration and deployment environments. They also provide tools for software publication via tagged releases, creation of BagIt and BagPack files and publication on the research data repository RADAR.

1 Introduction

Research software development is a fundamental aspect in research (Anzt et al. 2021), and it is now acknowledged that the FAIR principles (Findable, Accessible, Interoperable, Reproducible; Wilkinson et al. 2016), historically established for research data, should also be applied to research software (Chue Hong et al. 2021). In particular, reproducible research requires that software and its associated metadata can be found easily by both machines and humans, and that they are retrievable via standardised protocols. In this context, several metadata standards are widely used across the scientific community:

- the Citation File Format (CFF; Druskat et al. 2021)[1] aims to indicate to users how to cite a software package

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.c18081

1 CFF: https://citation-file-format.github.io.

- DataCite[2] (DataCite Metadata Working Group 2021) is a standard Metadata schema for archiving digital assets

- CodeMeta[3] (Jones et al. 2017) is an extension of `schema.org` created to standardize the exchange of software metadata across repositories and organizations

All of these standards serve specific purposes and several of them are required to cover the whole software lifecycle. However, research software developers should ideally not be burdened with maintaining multiple metadata files in different formats and largely overlapping content. This poses a risk to data consistency and to adoption of good software publication practices.

Therefore, we have developed a framework, named *openCARP-CI*, which allows developers to easily create and maintain the metadata associated to research software, by only maintaining a CodeMeta file from which CFF and DataCite files are automatically generated. The framework also allows publishing software according to the FAIR principles: releases with persistent identifiers can be created, archived and published on the open research data repository RADAR.

2 Description of Components

2.1 The openCARP-CI environment

The openCARP-CI package (Houillon et al. 2023) is part of the openCARP Collaborative Development Environment (Bach et al. 2022), an advanced technical infrastructure for collaborative research software projects based on GitLab[4]. It is composed of a set of Python scripts around the publication and long-term preservation of software repositories (see Figure 1). These tasks can be performed automatically when being integrated in GitLab Continuous Integration and Deployment (CI/CD) pipelines.

The openCARP-CI was created for the openCARP simulation software (Plank et al. 2021) but has its own separated repository and can be adopted for any project including research software hosted on GitLab. It complements efforts by other teams (such as the HERMES Project[5]) that aim to simplify publication workflow of research software together with rich metadata.

In the next section, we describe the different pipelines related to metadata management and software publication available in openCARP-CI.

2 DataCite: `https://schema.datacite.org`.
3 CodeMeta: `https://codemeta.github.io`.
4 GitLab: `https://about.gitlab.com`.
5 HERMES: `https://project.software-metadata.pub`.

Table 1: Components of openCARP-CI.

Script	Functionality
`create_cff`	generates Citation File Format (CFF) metadata file
`prepare_release`	updates *version* and *dateModified* in metadata
`create_release`	creates release in GitLab
`create_datacite`	generates DataCite metadata file
`create_bag`	creates BagIt package
`create_bagpack`	adds DataCite XML to BagIt
`prepare_radar`	reserves DOI on RADAR
`create_radar`	creates archive and uploads it to RADAR
`run_markdown_pipeline`	updates Grav CMS website
`run_bibtex_pipeline`	treats BibTex file for publications on website
`run_docstring_pipeline`	extracts docstrings from Python scripts

2.2 Automated metadata conversion

In order to ensure the coherence of metadata across different metadata file formats and to remove the burden of copying and maintaining redundant metadata information in several files, openCARP-CI offers scripts that convert metadata expressed in the CodeMeta standard to other metadata formats. As a consequence, developers only need to maintain `codemeta.json` as the unique metadata file for their software.

To generate the initial codemeta.json file, the CodeMeta Generator[6] can be used. Then, the script `create_cff` generates a Citation File Format (CFF) metadata file from the CodeMeta file (Druskat et al. 2021). The script `create_datacite` generates a DataCite XML file from the CodeMeta file.

```
build-datacite:
  stage: build
  image: python:3.9
  before_script:
  - pip install git+https://git.opencarp.org/openCARP/openCARP-CI.git
  script:
  - create_datacite
  artifacts:
    paths:
    - $DATACITE_PATH
    expire_in: 2 hrs
```

Figure 1: Example of a Gitlab CI job for automated creation of the DataCite metadata file.

6 CodeMeta-Generator: `https://codemeta.github.io/codemeta-generator`.

2.3 Creation of releases

A software release associated with a version number can be created on GitLab using the scripts `prepare_release` and `create_release`. The script `prepare_release` updates the CodeMeta file with a given version number and date. When using the script as part of a CI pipeline, this information is taken from the *tag* of the release and the current date. The script `create_release` actually creates the software release on GitLab using its API.

2.4 Creation of archives

openCARP-CI allows creating software packages destined to persistent long-term storage in research data repositories. These archives are created using the BagIt File Packaging Format[7], which is designed for reliable storage and transfer of arbitrary digital content.

The script `create_bag` creates a BagIt package containing the given assets, using the Python package `bagit-python`[8]. The script `create_bagpack` adds a DataCite XML file to the BagIt package as recommended by the RDA Research Data Repository Interoperability WG (RDA Research Data Repository Interoperability WG 2018).

2.5 Software publication

`prepare_radar` and `create_radar`, can be used to publish the software in the research data repository service RADAR[9]. In the RADAR repositories, datasets are assigned a persistent DOI (Digital Object Identifier) and published in accordance with the FAIR principles.

The script `prepare_radar` assigns a DOI and a RADAR ID to the dataset and adds them to its metadata (`codemeta.json`). The script `create_radar` creates the release in the RADAR service. This is done in a two step process, where first a *dataset* is created in RADAR, which contains the metadata. Then, in a second step, the different assets of the release (e.g. the source code and different compiled binaries) are uploaded.

2.6 Integration with the project website

An additional feature of openCARP-CI is publication of relevant information on a web page managed with the Grav content management system (CMS)[10].
The scripts `run_markdown_pipeline`, `run_bibtex_pipeline` and `run_docstring_pipeline` can be used for this purpose if desired.

7 BagIt description: `https://www.rfc-editor.org/rfc/rfc8493`.
8 bagit-python repository: `https://github.com/LibraryOfCongress/bagit-python`.
9 RADAR: `https://radar.products.fiz-karlsruhe.de/en`.
10 Grav CMS: `https://getgrav.org`.

3 Pipeline setup in a software repository

3.1 Prerequisites

The pipelines provided in openCARP-CI can be set up directly in any software project which fulfills the following conditions:

- The project's repository is under version control using Git and hosted in a GitLab instance
- A Docker runner is configured for the project's GitLab CI pipelines
- For optional publiction on RADAR, credentials have to be provided

3.2 Integration in GitLab CI pipelines

The CI scripts can be included in any GitLab project using the following process.
For projects hosted on GitHub, adaptations are required[11].

- In the project repository, go to *Settings* → *Access Tokens*, and create a token with the role *Maintainer* and scopes *api* and *write_repository*. Copy the token value.
- Go to *Settings* → *CI/CD* → *Variables* and choose *Add Variable*. Create a masked variable named `PUSH_TOKEN` and as a value, paste the copied token.
- Create a variable with key `PRIVATE_TOKEN` and as a value enter `$PUSH_TOKEN`.
- Copy the GitLab CI configuration files (`.gitlab-ci.yml` and `.gitlab/`) from the openCARP-CI repository to your software repository and adapt them to your needs. You can deactivate the release on RADAR by setting `ENABLE_RADAR` to `false` in `.gitlab-ci.yml`.
- Create a commit with the tag `pre-vX.Y`. The CI jobs will update metadata and create a release commit with the tag `vX.Y`.

4 Conclusions

The package openCARP-CI provides tools for automatic metadata conversion and software publication according to the FAIR principles, which can be automated in CI/CD pipelines on the GitLab development platform. After the initial setup, the user maintains a single metadata file in CodeMeta format. Other metadata formats are automatically generated from this file. The releases and supporting files are archived automatically for every new version of the software.

11 GitLab CI/CD to GitHub Actions: `https://docs.github.com/en/actions/migrating-to-github-actions/manual-migrations/migrating-from-gitlab-cicd-to-github-actions`.

We believe that the automated metadata conversion based on CodeMeta can be a useful tool for many research software developers and can facilitate the adoption of good software publication practices by reducing the effort for developers.

Acknowledgements

We gratefully acknowledge support by Deutsche Forschungsgemeinschaft (DFG, projects LO2093/1-1 and LO2093/9-1) and Karlsruhe Institute of Technology (KIT). This project has received funding from the European High-Performance Computing Joint Undertaking EuroHPC (JU) under grant agreement No 955495. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Italy, Germany, Austria, Norway, Switzerland.

References

Anzt, H, F Bach, S Druskat, F Löffler, A Loewe, BY Renard, G Seemann, et al. 2021. "An environment for sustainable research software in Germany and beyond: current state, open challenges, and call for action". *F1000Research* 9 (295). DOI: https://doi.org/10.12688/f1000research.23224.2.

Bach, Felix, Jochen Klar, Axel Loewe, Jorge Sánchez, Gunnar Seemann, Yung-Lin Huang, and Robert Ulrich. 2022. "The openCARP CDE: Concept for and implementation of a sustainable collaborative development environment for research software". *Bausteine Forschungsdatenmanagement,* number 1: 64–84. DOI: https://doi.org/10.17192/bfdm.2022.1.8368.

Chue Hong, Neil P., Daniel S. Katz, Michelle Barker, Anna-Lena Lamprecht, Carlos Martinez, Fotis E. Psomopoulos, Jen Harrow, et al. 2021. "FAIR Principles for Research Software (FAIR4RS Principles)". DOI: https://doi.org/10.15497/RDA00068.

DataCite Metadata Working Group. 2021. "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4". DOI: https://doi.org/10.14454/3W3Z-SA82. https://schema.datacite.org/meta/kernel-4.4/.

Druskat, Stephan, Jurriaan H. Spaaks, Neil Chue Hong, Robert Haines, James Baker, Spencer Bliven, Egon Willighagen, David Pérez-Suárez, and Olexandr Konovalov. 2021. *Citation File Format.* Version 1.2.0. DOI: https://doi.org/10.5281/zenodo.5171937.

Houillon, Marie, Jochen Klar, Axel Loewe, Tomas Stary, and openCARP consortium. 2023. *openCARP-CI.* DOI: https://doi.org/10.35097/974.

Jones, Matthew B., Carl Boettjiger, Abby Cabunoc Mayes, Arfon Smith, Peter Slaughter, Kyle Niemeyer, Yolanda Gil Gil, et al. 2017. "CodeMeta: an exchange schema for software metadata. Version 2.0." Edited by KNB Data Repository. DOI: https://doi.org/10.5063/schema/codemeta-2.0.

Plank, Gernot, Axel Loewe, Aurel Neic, Christoph Augustin, Yung-Lin Huang, Matthias A.F. Gsell, Elias Karabelas, et al. 2021. "The openCARP simulation environment for cardiac electrophysiology". *Computer Methods and Programs in Biomedicine* 208:106223. DOI: https://doi.org/10.1016/j.cmpb.2021.106223.

RDA Research Data Repository Interoperability WG. 2018. *Research Data Repository Interoperability WG Final Recommendations.* DOI: https://doi.org/10.15497/RDA00025.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Reifegradmodell für die Verwaltung des Datenzugriffs

Max Leo Wawer, Roland Lachmayer

Institut für Produktentwicklung und Gerätebau, Leibniz Universität Hannover

Durch die Anforderung Forschungsergebnisse langfristig verfügbar und nachnutzbar zu machen, werden Forschende in der datenintensiver werdenden ingenieurwissenschaftlichen Forschung mit einer Vielzahl an Leitlinien und Richtlinien für den Umgang mit Forschungsdaten konfrontiert, welche es seitens der Forschenden umzusetzen gilt. Der Umgang mit forschungsbezogenen Daten erstreckt sich entlang des Datenlebenszyklus, von der Datenmanagementplanung bis zur Datennachnutzung. Als Forschungsdatenmanagement (FDM) werden dabei alle Maßnahmen entlang des Datenlebenszyklus verstanden, um Daten nachnutzbar, nachvollziehbar und nachprüfbar zu machen. Die zusätzlichen Anforderungen durch die Umsetzung des FDMs stellt die Forschenden vor die Herausforderung Kenntnisse in allen Bereichen des FDMs für eine adäquate Umsetzung zu besitzen. Dies betrifft viele neue Prozesse und Aktivitäten, um den Umgang mit Forschungsdaten effektiv zu gestalten. Ein wichtiger Aspekt ist es dabei Daten zugänglich zu machen. Zugänglich gemachte Daten haben den Vorteil die Effizienz der Forschung durch eine Nachnutzbarkeit zu erhöhen und Forschungsergebnisse nachvollziehbarer zu gestalten. Hier gilt es in den Forschungsprojekten und auf Organisationsebene Prozesse zu definieren, die den Anforderungen der ingenieurwissenschaftlichen Forschung entsprechen.

Eine Betrachtung definierter Ziele und Praktiken für die Verwaltung des Datenzugriffs kann die Prozessgestaltung auf Projekt- oder Organisationsebene unterstützen und den genannten Mehrwert zugänglich gemachter Daten durch definierte Prozesse bestärken. Durch die Entwicklung und den Einsatz von Reifegradmodellen können bestehende Prozesse auf Projekt- oder Organisationsebene bewertet werden. Zudem lassen sich Ziele zur Verbesserung der Prozesse aufzeigen, mit denen sich neue Handlungsoptionen abhängig einer Reifestufe ableiten lassen.

Reifegradmodelle sind dabei ein für das Prozessmanagement und die Qualitätsverbesserung einsetzbares Werkzeug und dienen als Lösungsansatz für die Verbesserung und Umsetzung definierter Prozesse.

In diesem Beitrag wird ein entwickeltes Reifegradmodell für die Verwaltung des Datenzugriffs in Forschungsprojekten dargestellt. Jede Reifestufe enthält eine Anzahl definierter

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.c18082

Ziele und Praktiken entsprechend einer gegebenen Reifegradcharakteristik. Dieses Modell dient der Bewertung und der Verbesserung von Prozessen, um Daten zugänglich zu machen.

1 Einleitung

Mit steigenden Anforderungen an die Wissenschaft im Umgang mit Forschungsdaten werden Forschende mit einer Vielzahl an neuen Aufgaben und Pflichten konfrontiert. Hinzu kommt, dass auch Fördergeber Ansprüche in Hinblick auf zugänglich gemachte Forschungsdaten stellen, um Forschungsergebnisse nachvollziehbar und reproduzierbar durch zugänglich gemachte Forschungsdaten zu gestalten (Deutsche Forschungsgemeinschaft e.V. 2019). Die Existenz von Initiativen wie der deutschen „Nationalen Forschungsdateninfrastruktur“ (NFDI) zeigt die Bedeutung von Relevanz und Steuerung des Datenmanagements in Forschungsprojekten (Hartl, Wössner und Sure-Vetter 2021).

Forschungsdaten haben dabei eine hohe Relevanz im wissenschaftlichen Erkenntnisprozess, da in diesem mittels Datenerzeugung und anschließender Datenanalyse neue Erkenntnisse identifiziert werden. Dabei bilden Daten die Grundlage für neues Wissen und können disziplinübergreifend nachgenutzt werden, wenn diese zugänglich gemacht werden. Hierbei gilt es projektspezifische Richtlinien, welche Einfluss auf den Umgang mit Forschungsdaten und das Zugänglichmachen haben zu berücksichtigen. In den Ingenieurwissenschaften stehen Forschende diesbezüglich vor neuen Herausforderungen, die das Verhalten zur Datenpublikation und Wiederverwendung externer Daten beeinflussen (Joo und Kim 2017; Dierend u. a. 2023). Im Sinne der guten wissenschaftlichen Praxis (Deutsche Forschungsgemeinschaft e.V. 2019) ist hier ein Trend zu der Veröffentlichung von Forschungsdaten zu erwarten.

Um Forschende bei der Umsetzung des Datenzugriffs zu unterstützen und den aktuellen Stand in Forschungsprojekten zu bewerten, gibt es eine Vielzahl an möglichen Bewertungsmethoden. So kann die Qualität des Forschungsdatenmanagements (FDM) in Forschungsprojekten mittels Reifegradmodellen bewertet werden. Diese ermöglichen durch definierte Ziele auf einzelnen Reifestufen eine differenzierte Bewertung der Prozesse und haben den Vorteil, dass sie Verbesserungsmöglichkeiten durch nachgelagerte Reifestufen entlang eines Evolutionspfades bieten können. Die beiden vielmals zitierten Modelle, welche die Grundlage für viele weitere entwickelte Reifegradmodelle darstellen, sind das „Capability Maturity Model Integration“ (CMMI), ehemals CMM (CMMI Product Team 2010) und die ISO 33001 (ISO/IEC 33001 2015), ehemals ISO 15504 (SPICE). So haben sich Reifegradmodelle in der qualitativen Bewertung von Prozessen etabliert und finden auch immer mehr Eingang in den Bereich des FDMs (Oppenländer u. a. 2017).

Das in dieser Arbeit entwickelte Modell soll Forschenden im Forschungsprozess eine Bewertung über den aktuellen Stand des FDMs in Hinblick auf das Zugänglichmachen von Daten ermöglichen und durch nachgelagerte Reifestufen Verbesserungspotentiale aufzeigen. Ziel ist es, die Forschungsdaten nachnutzbar und nachvollziehbar zugänglich zu ma-

chen und eine Verbesserung der Umsetzung des Datenzugriffs in Forschungsprojekten zu ermöglichen.

2 Bestehende Arbeiten

Reifegradmodelle stellen eine qualitative Bewertungsmethode von Objekten dar, zumeist mit dem Fokus auf Prozesse. Sie setzen sich aus einer Folge diskreter Reifestufen zusammen, welche einen gewünschten Entwicklungspfad von einem Anfangsstadium bis hin zu einer vollumfänglichen Reife darstellen. Die Modelle können dabei als Maßstab für die Bewertung entsprechend eines definierten Entwicklungspfades dienen. Je Reifestufe werden Merkmale und Kriterien definiert, welche zur Erreichung einer Reifestufe erfüllt sein müssen. Nachgelagerte Reifestufen mit beschriebenen Merkmalen und Kriterien stellen ausgehend des Ist-Zustands ein kontinuierliches Verbesserungspotential dar (Paulk u. a. 1993). Die Modelle lassen sich für eigene Anwendungsfälle definieren und bieten daher einen Fokus für interne Strategien. Dabei können gesamte Organisationen oder einzelne definierte Prozessbereiche bewertet werden (Becker, Knackstedt und Pöppelbuß 2009). Auch das FDM kann dabei als Prozess- und Objektbereich im Forschungsprozess gesehen werden und dahingehend auf die qualitative Umsetzung bewertet werden. Für den Bereich des FDMs gibt es eine Vielzahl entwickelter Reifegradmodelle, mit welchen das FDM hinsichtlich verschiedener Anhaltspunkte und Dimensionen bewertet und verbessert werden kann (Oppenländer u. a. 2017; Lehmann und Odebrecht 2023; Proença und Borbinha 2018).

Ein viel zitiertes Beispiel aus diesem Bereich ist das „Capability Maturity Model for Research Data Management“ (CMM-RDM) von QIN (Qin, Crowston und Kirkland 2014), in welchem die Reifegradstruktur und Anzahl der Reifestufen des CMMI adaptiert wird. In diesem entwickelten Reifegradmodell werden eine Vielzahl von Querschnittsthemen wie beispielsweise *Training, Metadaten, Policies*, in fünf definierten Schlüsselprozessbereichen adressiert, welche wiederum in vier einzelne Teilbereiche aufgeteilt werden. Die Reifestufen werden dabei für jeden identifizierten Teilbereich definiert und kurz mit einem Satz beschrieben. Auch werden Materialien für die einzelnen Teilbereiche bereitgestellt. Das Modell sieht so die Bewertung des ganzheitlichen FDMs in Forschungsorganisationen vor, ohne aber eine direkte Bewertungsmethode und Anwendung des Modells festzulegen.

Aus dem Bereich des FDMs wurde von der Research Data Alliance (RDA) ein Reifegradmodell der FAIR-Prinzipien entwickelt. Mit dem Modell kann die Einhaltung der FAIR-Prinzipien, nach Wilkenson et al.

(Wilkinson u. a. 2016), bewertet werden, indem diese nach ihrer Priorität (*Essential, Important, Useful*) kategorisiert wurden. Das Modell ist normativ gestaltet und es werden die Prinzipien, Prioritäten und Bewertungsmethoden von Daten definiert. Die Kategorisierung der Prinzipien in dem Modell zeigt, dass die Bereiche *Findable* und *Accessible* von grundlegenderer Bedeutung für das Erreichen von „FAIRness“ sind als die Bereiche *Interoperable* und *Reusable* (Research Data Alliance FAIR Data Maturity Model Working Group 2020).

Das im PODMAN Projekt entwickelte Referenzmodell „DIAMANT“ (Designing an Information Architecture for Data Management Technologies; Gerhards u. a. 2020), sieht eine anwendungsbezogene Bewertung einzelner Dienste und Services im Bereich des FDMs vor. Bei dem DIAMANT-Modell steht dabei die operative Sicht und technische Infrastruktur im FDM bezogen auf gesamte Forschungseinrichtungen im Vordergrund. Ein definierter FDM-Referenzprozess stellt dabei die Grundlage für die Integration des FDMs dar. Die Bewertung des FDM-Portfolios findet dabei durch einen IST- /SOLL-Abgleich definierter Kompetenzen für definierte FDM-Funktionen statt.

Mit dem in dieser Arbeit erstellten Reifegradmodell will man das FDM als parallellaufendes Prozesssystem im Forschungsprojekt integrieren und die Umsetzung des Datenzugriffs direkt im Forschungsprozess adressieren. So wird ein Ansatz von einem forschungsorientierten Reifegradmodell forciert, welches direkt von Forschenden angewendet werden kann. In dieser Arbeit wird dabei ein Reifegradmodell für die Verwaltung des Datenzugriffs in Forschungsprojekten vorgestellt.

3 Der Datenzugriff im Forschungsprozess

Bei der Betrachtung des Forschungsprozesses erfolgt das Publizieren von Forschungsergebnissen nach der Analyse hinsichtlich der Beantwortung von wissenschaftlichen Fragestellungen (Kowalczyk 2018; Patel 2011; Maxwell 2015). Eingebettet in ein Forschungsprojekt, läuft der beschriebene Forschungsprozess innerhalb eines Projektes mehrmals ab. So werden in einem Forschungsprojekt mehrere Forschungsergebnisse auf Grundlage erhobener und analysierter Daten publiziert. Bei der Betrachtung des FDMs durch das Heranziehen verschiedener Datenlebenszyklusmodelle erfolgt der Datenzugriff nach der Analyse von erhobenen Daten, gefolgt von einer Datennachnutzung verfügbar gemachter Daten (Wissik und Ďurčo 2015; Wolf und Leppla 2020). Dabei lässt sich das FDM (orientiert am Datenlebenszyklus) in den Forschungsprozess integrieren (RfII - Rat für Informationsinfrastrukturen 2019; Minn und Lemaire 2017), wobei die Nachnutzung während des Forschungsprozesses in der Phase der Datenerhebung im Sinne einer Sammlung bereits erhobener Daten und deren Nachnutzung stattfindet. Auch lassen sich sowohl Rohdaten (nach der Datenerhebung) und analysierte Daten jeweils beide archivieren und zugänglich machen. So lässt sich das FDM in den Forschungsprozess integrieren und erweitert diesen um datenmanagementbezogene Aktivitäten und Inhalte. Die Phase der Publikation im Sinne einer Textpublikation von Forschungsergebnissen wird erweitert durch die zusammenhängende Verfügbarmachung zugrundeliegender Daten der publizierten Forschungsergebnisse. Dies stärkt die Nachvollziehbarkeit und Nachprüfbarkeit von publizierten Forschungsergebnissen. Auch können im Sinne der Nachnutzbarkeit Daten zugänglich gemacht werden, welche nicht direkt in einem Zusammenhang mit Textpublikationen stehen, aber für die Forschungscommunity einen nachnutzbaren Wert haben. Im Rahmen des Datenzugriffs müssen Forschende bestimmen, welche Daten zugänglich gemacht werden sollen. Hierbei geht es um die Auswahl der Forschungsdaten, um diese nachnutzen zu können. Es spielen Kriterien, wie die Verifizierung von Forschungsergebnissen, das Potential von nicht wiederholbaren Studien oder gegebene Nachnutzungsszenarien eine Rolle bei der Daten-

auswahl von zugänglich zumachenden Daten. Auch muss definiert werden, wer auf die Daten zugreifen darf und wie die Daten nachgenutzt werden dürfen. Dies hat meist einen projektpartnerspezifischen Hintergrund. Auch muss das technische System, die Plattform worüber die Daten zugänglich gemacht werden sollen (bspw. ein Datenrepositorium), bestimmt werden (Ludwig und Enke 2013). Für einen definierten Zugang der Daten können offene oder geschlossene Zugangsplattformen ausgewählt werden, mit welchen der Rahmen der Zugänglichkeit bestimmt werden kann. Auch kann die Nachnutzbarkeit der Daten durch die Angabe von Lizenzen eingegrenzt werden. Um Daten nachnutzbar zu gestalten, ist die Kontextualisierung und Datenprovenienz von bedeutender Rolle. Die Daten werden aufbereitet und durch weitere (Meta-)daten und entsprechende Dateiformate nachnutzbar und für weitere Forschungszwecke einsetzbar gestaltet. Hierbei ist es relevant, durch weitere Datendokumentationen die Daten mit zugehörigen, der Forschung entsprechenden Materialien (bspw. Rahmenbedingungen der Datenerhebung, Analyseverfahren) in einen Kontext zu setzen (Eynden u. a. 2011).

4 Entwickeltes Reifegradmodell

Die in den vorherigen Abschnitten beschriebenen Inhalte zeigen den Anwendungsbereich des entwickelten Reifegradmodells auf. So sollen die Reifegradmodelle von Forschenden im Forschungsprojekt während ihrer Forschung Anwendung finden. Dabei wird eine forschungsprozessorientierte Sichtweise auf die Inhalte und Anwendung des Reifegradmodells gelegt. Die entwickelte Reifegradcharakteristik folgt dabei den Zielen des FDM. In Abbildung 1 ist die definierte Reifegradcharakteristik der fünf Reifestufen mit zugehörigen Merkmalen der jeweiligen Reifestufe dargestellt. Die Reifestufen orientieren sich dabei an den Aufbau des CMMI (vgl. CMMI Product Team 2010) und sind speziell für die Ziele des Datenzugriffs definiert (Deutsche Forschungsgemeinschaft e.V. 2019; RfII - Rat für Informationsinfrastrukturen 2019; Iglezakis und Schembera 2018).

Reifestufe 1 „Einstieg“ beschreibt Projekte, in denen noch keine Inhalte und Prozesse für die Verwaltung des Datenzugriffs auf Projektebene definiert wurden. Hier erfolgt ein Datenzugriff reaktiv und intuitiv, beispielsweise auf Nachfrage anderer Forschenden, und hängt somit vom Engagement des Einzelnen ab.

Reifestufe 2 „Geführt“ beschreibt Projekte, in denen auf Projektebene die Verwaltung des Datenzugriffs definiert und geregelt ist. Es wird bestimmt welche Daten wie zugänglich gemacht werden und wie diese nachgenutzt werden dürfen. Das technische System wird bestimmt, mit welchem die Daten zugänglich gemacht werden. Die Daten werden zudem mit einem Unique Identifier versehen, sodass sie sich auffinden lassen und für ein definiertes Nutzerfeld zugänglich sind.

Reifestufe 3 „Definiert“ baut auf die Inhalte der vorherigen Reifestufe auf, nun kommt aber hinzu, dass das Zugänglichmachen von Daten an die zugehörige Fachcommunity ausgerichtet wird und eine inhaltliche Nachnutzbarkeit und Weiterverarbeitung der Daten gewährleistet wird. Die Daten werden mit fachspezifischen Metadaten beschrieben und in entsprechende Dateiformate überführt, sodass die Daten integrierbar sind und die Inter-

Abbildung 1: Reifegradcharakteristik und zugehörige Merkmale.

operabilität und Nachnutzbarkeit der Daten gesichert wird. Auf dieser Reifestufe kommt die ausreichende Kontextualisierung der Forschungsdaten durch zugehörige Materialien und entsprechende Metadatenstandards hinzu.

Reifestufe 4 „Quantitativ Geführt“ beschreibt die Umsetzung des Datenzugriffs mit qualitativen und quantitativen Sicherungsmaßnahmen. So werden Daten vor dem Zugänglichmachen auf Ihre Vollständigkeit und Korrektheit hin geprüft. Erst nach positiver Prüfung erfolgt die Verfügbarmachung.

Reifestufe 5 „Optimierend“ beschreibt Projekte, in denen die Aktivitäten und Inhalte für das Zugänglichmachen von Daten proaktiv und fortlaufend verbessert werden. Die Projekte entwickeln in diesem Zusammenhang Best Practices und weitere Inhalte, welche mit der Fachcommunity geteilt werden.

Die einzelnen Reifestufen werden zusätzlich mit definierten Zielen und zugehörigen Praktiken näher beschrieben, dabei werden die in Kapitel 3 beschriebenen Umfänge und Aktivitäten für das Zugänglichmachen von Forschungsdaten wieder aufgegriffen und den Reifestufen entsprechend der entwickelten Reifegradcharakteristik zugeordnet. Auf diese Weise ergibt sich eine Checkliste mit definierten Zielen je Reifestufe. Diese Checklisten können von Forschenden selbst angewendet werden und attestieren eine Reifestufe bei Zielerfüllung aller Ziele einer jeweiligen Reifestufe. Durch nachgelagerte Reifestufen wird zudem ein Ausblick auf Verbesserungen in Hinblick auf das Zugänglichmachen von Daten gegeben.

5 Anwendung des Reifegradmodells in Forschungsprojekten

Das entwickelte Reifegradmodell wurde zur Reifegradbestimmung in Form einer Checkliste (Wawer 2023) in drei verschiedenen Forschungsprojekten angewendet. Dabei konnten

die Forschenden durch die Checkliste eigenständig die Erfüllung der Ziele je Reifestufe bestätigen. Eine Reifestufe wird erreicht, wenn alle auf der Reifestufe definierten Ziele im Rahmen des Datenzugriffs erfüllt werden. Die Ergebnisse der Reifegradbestimmung werden im Folgenden näher dargestellt.

In einem ersten Anwendungsbeispiel wurde das Reifegradmodell in einem Forschungsprojekt angewendet, welches in einem institutsübergreifenden Sonderforschungsbereich aus dem Bereich Maschinenbau eingebettet ist. Der Sonderforschungsbereich setzt sich dabei aus verschiedenen Forschungsfeldern zusammen, in denen eine Vielzahl heterogener Daten erhoben und hinsichtlich gemeinsamer Forschungsziele analysiert werden (Mozgova u. a. 2020). Innerhalb des Sonderforschungsbereichs bestehen Richtlinien für den Umgang mit Forschungsdaten und Inhalte zum FDM werden entwickelt. Bei der Reifegradbestimmung kam dahingehend eine Reifegradbestimmung der Reifestufe 2 „Geführt“ als Ergebnis heraus. Die Forschungsdaten, welche im Zusammenhang mit publizierten Forschungsergebnissen stehen, werden entsprechend projektinterner Richtlinien zugänglich gemacht. Es ist definiert, welche Daten wie zugänglich gemacht werden sollen. Die Daten sind in einem Datenrepositorium für andere Forschende auffindbar und zugänglich und werden mit allgemeinen Metadaten beschrieben. Jedoch sind noch keine forschungsspezifischen Metadaten für die entwickelten Softwaremodelle zur weiteren Beschreibung bestimmt worden, mit welchen die Forschungsdaten dieses Forschungsfeldes spezifisch beschrieben werden, so dass eine inhaltliche Nachnutzung und Weiterverarbeitung gewährleistet werden kann. Dahingehend wird die Identifizierung und Entwicklung forschungsspezifischer Metadaten angestrebt, um die Nachnutzbarkeit der Daten zu erhöhen.

In einem weiteren Anwendungsfall kam das Reifegradmodell in einem landesgeförderten Forschungsprojekt zum Einsatz. Das Ausfüllen der Checkliste ergab dabei eine Reifegradbestimmung der Reifestufe 2 „Geführt“. In dem Forschungsprojekt werden institutsinterne Richtlinien zum Umgang mit Forschungsdaten für das Zugänglichmachen von Daten angewendet, indem die Daten intern zugänglich gemacht werden. Bis zum Projektende wird angestrebt, die entwickelte Software der Forschungscommunity frei zugänglich zu machen. Dahingehend werden Metadaten und zugehörige Daten zur Nachnutzbarkeit der entwickelten Software identifiziert und verknüpft. Auch sollen die entwickelten Softwaremodelle interoperabel gestaltet werden, sodass sie für weitere Forschungszwecke nachgenutzt und weiterentwickelt werden können. Hierfür sollen entsprechende Dateiformate berücksichtigt werden.

Außerdem wurde das Reifegradmodell in einem Forschungsprojekt angewendet, welches in enger Zusammenarbeit mit Industriepartnern steht. Im Rahmen dieses Forschungsprojektes werden durch Feldexperimente eine Vielzahl an Daten erhoben und hinsichtlich definierter wissenschaftlicher Fragestellungen analysiert und Lösungen entwickelt. Die Anwendung des Reifegradmodells ergab eine Reifegradbestimmung der Reifestufe 1 „Einstieg“. Forschungsergebnisse werden publiziert, jedoch werden zugrundeliegende Forschungsdaten nicht proaktiv zugänglich gemacht. Auf Nachfrage werden Daten für die Nachnutzung an andere Forschende unter Berücksichtigung vorliegender Richtlinien weitergeleitet. Jedoch wurde in diesem Forschungsprojekt noch kein weiterer Umgang für den Datenzugriff von Daten definiert. Es wird dahingehend angestrebt, die Ziele der Reifestufe 2 zu erfüllen,

indem bestimmt wird, welche Forschungsdaten in dem Forschungsprojekt im Folgenden zugänglich gemacht werden und welche Zugangsplattformen dafür verwendet werden sollen.

Die dargestellten Anwendungsbeispiele zeigen, dass das Zugänglichmachen von Daten in den Forschungsprojekten nach aktuellem Stand noch auf einer niedrigen Reifestufe („Einstieg“ oder „Geführt“) stattfindet. Dies liegt daran, dass in den Forschungsprojekten noch keine forschungsspezifischen Metadaten entsprechend der Forschungsfelder definiert wurden oder sich Standards in den Forschungsbereichen etabliert haben, welche angewendet werden können. Zum aktuellen Zeitpunkt erfolgt das Zugänglichmachen von Daten noch zumeist intuitiv oder es werden Forschungsdaten mit minimalen Aufwänden auffindbar und zugänglich gestaltet, ohne eine fachweite Nachnutzbarkeit durch eine Kontextualisierung der Daten zu berücksichtigen und zu ermöglichen. Durch gegebene und etablierte Standards in den Forschungsbereichen lassen sich aber dahingehend die Ziele höherer Reifestufen erreichen.

6 Zusammenfassung

Das vorgestellte Reifegradmodell ist eine erste Version des Gesamtmodells zur Bewertung des FDMs in Forschungsprojekten. So werden in Zukunft noch weitere Reifegradmodelle für die Phasen des Forschungsprozesses im Bereich des FDMs entwickelt. Das dargestellte Reifegradmodell zielt auf die Verfügbarmachung von Forschungsdaten ab und ermöglicht die Bewertung der Verwaltung des Datenzugriffs in Forschungsprojekten. Es werden definierte Ziele und Praktiken auf einzelnen Reifestufen aufgezeigt, sodass von Forschenden eine eigenständige Bewertung durch Checklisten möglich ist.

Zudem werden Verbesserungsmöglichkeiten durch Ziele nachgelagerte Reifestufen aufgezeigt, wodurch die Umsetzung verbessert werden kann. Dabei zeigt die Reifestufencharakteristik einen Evolutionspfad zur schrittweisen Verbesserung auf. So wird ein intuitiver Datenzugriff in einer nächsten Reifestufe projektintern definiert, sodass die Daten grundlegend auffindbar und zugänglich sind. Auf einer nächsten Reifestufe wird die Interoperabilität und inhaltliche Nachnutzbarkeit der Daten in der Forschungscommunity adressiert. Orientiert an dem CMMI finden in der nächsten Reifestufe Qualitätssicherungsmaßnahmen zur Verwaltung des Datenzugriffs Anwendung, sodass ein Datenzugriff beispielsweise erst bei erfolgreicher Prüfung erfolgt. Mit Reifestufe 5 wird zuletzt die Weiterentwicklung von FDM-Lösungen und Best Practices im Bereich des Datenzugriffs adressiert.

Die Fallbeispiele aus drei verschiedenen Forschungsprojekten haben die Einsatzfähigkeit der Reifegradmodelle unter Beweis gestellt. Die Anwendung des Modells hat dabei gezeigt, dass sich die resultierenden Reifegradbewertungen vornehmlich noch auf den Reifestufen 1 und 2 befinden. In den Forschungsprojekten gibt es die Bestrebungen, sich im weiteren Projektverlauf bis zum Projektende verbessern zu wollen und die Ziele der nachgelagerten Reifestufen zu erfüllen. Hierbei spielen auch der Projektumfang und die Forschungsmethodik eine wichtige Rolle, die Einfluss darauf haben, wie Forschungsdaten während der Projektlaufzeit zugänglich gemacht werden. Darüber hinaus sollten auch Fallbeispiele hö-

heren Reifegrades identifiziert werden, um aus den Aktivitäten und Umsetzungen dieser Projekte relevante Inhalte für die Reifegradmodelle identifizieren zu können. So kann auch die Reifegradbewertung der Stufe 3 und höher in Forschungsprojekten aufgezeigt werden. Trotz des normativen Charakters des Reifegradmodells sollen neben den bereits beschriebenen Zielen und Praktiken noch zusätzliche Hilfestellungen und zugehörige Materialien mit beigefügt werden, um Verbesserungen entsprechend der Ziele nachgelagerter Reifestufen vereinfacht vornehmen zu können. Alle Informationen über die Weiterentwicklung des Reifegradmodells und der Betrachtung weiterer relevanter Inhalte im Bezug zum FDM werden über Gitlab verfügbar gemacht.[1]

Danksagung

Max Leo Wawer und Roland Lachmayer möchten sich bei Bund, Ländern und bei der Gemeinsamen Wissenschaftskonferenz (GWK) für die Förderung und Unterstützung im Rahmen des Konsortiums NFDI4Ing bedanken. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 442146713.

Literaturverzeichnis

Becker, Jörg, Ralf Knackstedt und Jens Pöppelbuß. 2009. „Developing Maturity Models for IT Management“. *Business & Information Systems Engineering* 1:213–222. DOI: https://doi.org/10.1007/s12599-009-0044-5.

CMMI Product Team. 2010. *CMMI for Development, Version 1.3.* Technischer Bericht. Software Engineering Institute, Carnegie Mellon University. DOI: https://doi.org/10.1184/R1/6572342.v1. http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=9661.

Deutsche Forschungsgemeinschaft e.V. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis.* Zenodo. DOI: https://doi.org/10.5281/zenodo.6472827.

Dierend, Hauke, Osman Altun, Iryna Mozgova und Roland Lachmayer. 2023. „Management of Research Field Data Within the Concept of Digital Twin“. In *Advances in System-Integrated Intelligence,* herausgegeben von Maurizio Valle, Dirk Lehmhus, Christian Gianoglio, Edoardo Ragusa, Lucia Seminara, Stefan Bosse, Ali Ibrahim und Klaus-Dieter Thoben, 205–214. Cham: Springer International Publishing. ISBN: 978-3-031-16281-7.

Eynden, Veerle Van den, Louise Cortia nd MatthewWoollard, Libby Bishop und Laurence Horton. 2011. *Managing and sharing data: best practice for researchers.* UK Data Archive. ISBN: 1-904059-78-3.

Gerhards, Lea, Marina Lemaire, Stefan Kellendonk und André Förster. 2020. „Das DIAMANT-Modell 2.0“. DOI: https://doi.org/10.25353/UBTR-XXXX-F5D2-FFFB.

1 https://git.rwth-aachen.de/nfdi4ing/s-1/fdm-reifegradmodelle

Hartl, Nathalie, Elena Wössner und York Sure-Vetter. 2021. „Nationale Forschungsdateninfrastruktur (NFDI)“. *Informatik Spektrum* 44 (5): 370–373. DOI: https://doi.org/10.1007/s00287-021-01392-6.

Iglezakis, Dorothea, und Björn Schembera. 2018. „Anforderungen der Ingenieurwissenschaften an das Forschungsdatenmanagement der Universität Stuttgart - Ergebnisse der Bedarfsanalyse des Projektes DIPL-ING“. *o-bib. Das offene Bibliotheksjournal* 5 (3): 45–60. DOI: https://doi.org/10.5282/o-bib/2018H3S46-60.

ISO/IEC 33001. 2015. *Information technology — Process assessment — Concepts and terminology.* Technischer Bericht. ISO/IEC 33001:2015.

Joo, Yeon Kyoung, und Youngseek Kim. 2017. „Engineering researchers' data reuse behaviours: a structural equation modelling approach“. *The Electronic Library* 35 (6): 1141–1161. DOI: https://doi.org/10.1108/EL-08-2016-0163.

Kowalczyk, Stacy T. 2018. „Modelling the Research Data Lifecycle“. *International Journal of Digital Curation* 12 (2): 331–361. DOI: https://doi.org/10.2218/ijdc.v12i2.429.

Lehmann, Anna, und Carolin Odebrecht. 2023. „Reifegradmodelle im Forschungsdatenmanagement – IT-Prozessoptimierung im Wissenschaftsbetrieb“. *Information – Wissenschaft & Praxis* 74 (1): 9–21. DOI: https://doi.org/10.1515/iwp-2022-2249.

Ludwig, Jens, und Harry Enke, Hrsg. 2013. *Leitfaden zum Forschungsdaten-Management: Handreichungen aus dem WissGrid-Projekt.* Glückstadt: Werner Hülsbusch. ISBN: 978-3-86488-032-2.

Maxwell, Dan. 2015. „The Research Lifecycle as a Strategic Roadmap“. *Journal of Library Administration* 56 (2): 111–123. DOI: https://doi.org/10.1080/01930826.2015.1105041.

Minn, Gisela, und Marina Lemaire. 2017. „Forschungsdatenmanagement in den Geisteswissenschaften“. In *Universität Trier eSciences Working Papers,* 32. 03. Universität Trier Servicezentrum eSciences.

Mozgova, Iryna, Oliver Koepler, Angelina Kraft, Roland Lachmayer und Sören Auer. 2020. „Research Data Management System for a large Collaborative Project“. In *Balancing Innovation and operation.* The Design Society. DOI: https://doi.org/10.35199/NORDDESIGN2020.48.

Oppenländer, Jonas, Falko Glöckler, Jana Hoffmann und Claudia Müller-Birn. 2017. „Reifegradmodelle für ein integriertes Forschungsdatenmanagement in multidisziplinären Forschungsorganisationen“. In *E-Science-Tage 2017: Forschungsdaten managen,* herausgegeben von Jonas Kratzke und Vincent Heuveline, 53–64. heiBOOKS. DOI: https://doi.org/10.11588/heibooks.285.377.

Patel, Manjula. 2011. „An Idealised Scientific Research Activity Lifecycle Model“. University of Bath.

Paulk, Mark C., Bill Curtis, Mary Beth Chrissis und Charles V. Weber. 1993. „Capability Maturity Model, Version 1.1“. *IEEE Software* 10 (4): 18–27. ISSN: 0740-7459. DOI: https://doi.org/10.1109/52.219617.

Proença, Diogo, und José Borbinha. 2018. „Maturity Models for Data and Information Management“. In *Digital Libraries for Open Knowledge,* herausgegeben von Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David und João Correia Lopes, 81–93. Cham: Springer International Publishing. ISBN: 978-3-030-00066-0.

Qin, Jian, Kevin Crowston und Arden Kirkland. 2014. *A Capability Maturity Model for Research Data Management.* Technischer Bericht. NY: School of Information Studies, Syracuse University.

Research Data Alliance FAIR Data Maturity Model Working Group. 2020. „FAIR Data Maturity Model: specification and guidelines“. DOI: https://doi.org/10.15497/RDA00050.

RfII - Rat für Informationsinfrastrukturen. 2019. *RfII-Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel – November 2019.* https://rfii.de/?p=4043.

Wawer, Max Leo. 2023. *Reifegradmodell für das Management des Datenzugriffs.* Presentation. DOI: https://doi.org/10.5281/zenodo.7730556.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Wissik, Tanja, und Matej Ďurčo. 2015. „Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions“. In *Linköping Electronic Conference Proceedings.* 123.

Wolf, Armin Harry, und Cindy Leppla. 2020. „Harmonisierung von Datenlebenszyklus-Modellen: Nutzung von Synergien für optimierte Anwendungen im FDM“. *Bausteine Forschungsdatenmanagement,* Nr. 2: 1–19. DOI: https://doi.org/10.17192/BFDM.2020.2.8281.

Das Data Science Center an der Universität Bremen: Interdisziplinärer Knotenpunkt und Service-Infrastruktur für die datenintensive Forschung

Lena Steinmann*, Heike Thöricht, Sandra Zänkert, Rolf Drechsler

Data Science Center, Universität Bremen;
*Korrespondierende Autorin: lena.steinmann@uni-bremen.de

Das Data Science Center (DSC) der Universität Bremen ist ein interdisziplinäres Institut und dient als Knotenpunkt für die datenintensive Forschung. Mit seinen Aktivitäten bricht das DSC disziplinäre Silos auf und stärkt die kooperative Forschung, steigert die Datenkompetenzen von Forschenden und ermöglicht eine wertschöpfende Nutzung von Daten in allen Forschungsbereichen. Es bietet umfassende Angebote und Services für datenintensiv Forschende wie Trainings, Beratungsangebote, eine IT-Infrastruktur und finanzielle Unterstützung. Das DSC kann als institutionelles Best-Practice für andere Standorte dienen, um eine strukturelle Zusammenführung und Zentralisierung von Forschungsdatenmanagement (FDM) und Data Science umzusetzen und so die Effizienz im Forschungsprozess zu steigern.

1 Die Bedeutung eines Data Science Centers

Der Wissenschaftsrat hat bereits in einem Positionspapier im Jahr 2020 die Errichtung von Data Science Centern als interdisziplinäre Forschungsstrukturen, die auch Möglichkeiten zur Weiterbildung und Kompetenzentwicklung bieten, empfohlen (Wissenschaftsrat 2020). Der Hintergrund dieser Empfehlung ist, dass die datenintensive Forschung mittlerweile Einzug in alle Disziplinen gehalten hat und das Wissenschaftssystem auf die einhergehenden Herausforderungen wie der Schnelllebigkeit von Datenangeboten, Hard- und Software-Entwicklungen sowie dem immer schnelleren Aufkommen neuer Methoden stellen muss. Um auf diese Entwicklungen angemessen zu reagieren, sind neue Strukturen sowie Beratungs- und Lernangebote erforderlich. Data Science Center können dabei eine wichtige Rolle einnehmen, indem sie geeignete Rahmenbedingungen für datenintensiv Forschende schaffen und zur strategischen Positionierung der Hochschulen beitragen.

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.c18083

Die Universität hat bereits 2019 mit Unterstützung des Landes Bremen das Data Science Center (DSC@UB; im Folgenden durch DSC abgekürzt) als interdisziplinäres Institut und Knotenpunkt für die datenintensive Forschung gegründet. Die übergeordneten Ziele des DSC sind die wertschöpfende Nutzung von Daten in allen Forschungsbereichen nachhaltig zu ermöglichen und den kulturellen Wandel im Sinne einer „FAIR Data-Kultur“[1] voranzutreiben. Dazu bündelt das DSC die Kompetenzen von Forschenden unterschiedlicher Wissenschaftsbereiche in seinem Mitgliedernetzwerk. So werden die Vernetzung und der Transfer von Datenkompetenzen über fachliche Grenzen hinweg gefördert, disziplinäre Silos aufgebrochen und neue Wege für die kooperative Forschung geschaffen. Darüber hinaus stärkt das DSC durch gezielte Trainings die Datenkompetenzen von Forschenden und bietet umfassende Forschungsdatenmanagement (FDM)- und Data-Science-Services für datenintensiv Forschende, die in Abschnitt 2 näher erläutert werden. Durch die enge Verzahnung von FDM- und Data-Science-Kompetenzen ist es möglich, Forschende während des gesamten Datenlebenszyklus zu unterstützen und so eine bestmögliche Wertschöpfung aus Daten zu erzielen (siehe auch Steinmann und Drechsler 2021, für weitere Informationen). Das DSC repräsentiert somit einen wichtigen strukturellen Pfeiler für die datenintensive Forschung an der Universität Bremen und fungiert auch als Bindeglied zwischen Forschenden und zentralen Infrastruktureinrichtungen (wie Bibliotheken oder Datenzentren). Auf regionaler Ebene bildet es gemeinsam mit dem Promovierenden-Trainingsprogramm „Data Train – Training in Research Data Management and Data Science“ (Hörner u. a. 2021) und im Zusammenspiel mit den Bremer Konsortien der Nationalen Forschungsdateninfrastruktur (NFDI) einen entscheidenden Baustein im Leitprojekt „Forschungsdatenmanagement und Data Science“ der U Bremen Research Alliance – ein Kooperationsnetzwerk der Universität und zwölf außeruniversitärer Institute.

2 Angebote und Services

Das DSC bietet vielfältige Angebote und Services, die angepasst an die Bedarfe der Forschenden kontinuierlich weiterentwickelt werden. Im Folgenden stellen wir diese eingeordnet in die drei Bereiche Datenmanagement, Data Science und Kulturwandel vor (für eine Zusammenfassung siehe Abbildung 1).

2.1 Angebote und Services im Bereich Datenmanagement

FDM-Beratung: Data Stewards unterstützen Forschende der Universität Bremen sowie der außeruniversitären Institute der U Bremen Research Alliance beim effektiven und nachhaltigen Management ihrer Forschungsdaten. Dabei arbeiten sie eng mit dem FDM-Beratungspersonal der Staats- und Universitätsbibliothek (SuUB) und des Referats für Forschung und wissenschaftlicher Nachwuchs der Universität Bremen sowie mit den in Bremen ansässigen NFDI-Konsortien und den Datenzentren PANGAEA und Qualiservice zusammen. Die FDM-Beratung erfolgt sowohl in der Antragsphase als auch forschungsbe-

1 Gemäß der FAIR-Prinzipien sollen Forschungsdaten auffindbar (findable), zugänglich (accessible), interoperabel (interoperable) und nachnutzbar (reusable) sein (Wilkinson u. a. 2016).

gleitend. Typische Themen sind dabei Datenmanagementpläne, Datenorganisation, Datenspeicherung, FAIR-Prinzipien und die Nachnutzung von Forschungsdaten.

FDM-Trainings: Neben den Beratungstätigkeiten entwickeln die Data Stewards auch gezielte FDM-Trainings. Hierbei liegt der Fokus auf "hands-onFormaten, die sich insbesondere an Promovierende und Postdocs richten. Bisher wurden Themen wie die Erstellung von Datenmanagementplänen (DMP) für Anträge, die Nutzung von Git und die Reproduzierbarkeit von Daten in Form eines „Reproducibility Hackathon" adressiert. Die Trainings werden regelmäßig angeboten und angepasst an die Bedarfe um weitere Formate ergänzt.

Brücken bauen: Das DSC nimmt eine wichtige Rolle bei der Verknüpfung von Forschenden mit den zahlreichen FDM-Strukturen in der Bremer Forschungslandschaft ein. Ziel ist es, die bestehenden Angebote der unterschiedlichen Einrichtungen für die Zielgruppe sichtbar zu machen und Hürden bei der Nutzung zu verringern. Die Data Stewards agieren dabei im „Front Office" als direkte Ansprechpartner:innen für Forschende, während die NFDI-Infrastruktur und Datenzentren im „Back Office" die Daten entgegennehmen und zugänglich machen.

Peer2Peer Austausch: Der Austausch innerhalb der FDM-Community wird auf mehreren Ebenen gefördert. Zum einen werden lokale Community-Events in ungezwungener Atmosphäre für den kollegialen Austausch organisiert, wie z.B. În Love With Data - A Coffee and Cake Get-Together of the Data-Steward-Network Bremen"während der Love Data Week 2022. Zum anderen arbeitet das DSC aktiv in verschiedenen FDM-Netzwerken mit, wie dem lokalen Data-Steward-Netzwerk der U Bremen Research Alliance oder der DINI/nestor UAG-Schulungen/Fortbildung auf nationalem Level. Darüber hinaus beteiligt es sich am fachlichen Austausch auf den Workshops der NFDI oder FDM-Landesinitiativen und bei FDM-Konferenzen und -Vernetzungsevents wie dem Open Science Festival oder Data Stewardship goes Germany.

2.2 Angebote und Services im Bereich Data Science

Data-Science-Beratung: Das DSC unterstützt anwendungsorientierte Forschende bei der Einbindung von Data-Science-Komponenten in Anträgen und deren Umsetzung im Forschungsprozess. Dabei bietet es Beratung zu der Wahl der richtigen Methoden und Hardware, dem Aufbau einer passenden System-/IT-Architektur, der (Weiter-)Entwicklung von Algorithmen und dem Training neuronaler Netze. Als Partner für Anträge bietet es außerdem die Möglichkeit zur zentralen Ansiedlung von Data Scientists und Data Stewards zur bestmöglichen Bündelung und Nutzung von Kompetenzen.

Data-Science-Trainings: Data-Science-Trainings werden mit Fokus auf die am Standort Bremen häufig verwendeten Programmiersprachen Python und R angeboten. Thematisch werden Einführungskurse aber auch fortgeschrittene Workshops zu Maschinellem Lernen, Datenvisualisierung oder Zeitreihenanalyse angeboten. Ergänzend zur bereitgestellten IT-Infrastruktur gibt es regelmäßige Einführungskurse für die Nutzer:innen.

Abbildung 1: Zusammenfassende Darstellung der Angebote und Services des DSC bezogen auf die drei Bereiche Datenmanagement, Data Science und Förderung des Kulturwandel.

DSC Seed Grant: Über den DSC Seed Grant können Forschende finanzielle Unterstützung zur Umsetzung ihrer Data-Science-Vorhaben beantragen. Förderfähig sind beispielsweise datenwissenschaftliche Pilotstudien, die Durchführung von Konferenzen und Workshops in Bremen sowie Forschungsaufenthalte zur Vernetzung und Weiterentwicklung von Forschungsvorhaben. Die geförderten Vorhaben decken eine breite Spanne an datenwissenschaftlichen Themen ab mit Beteiligung von Forschenden aus unterschiedlichsten Disziplinen, wie Wirtschafts-, Sprach-, Geo-, Sozial-, und Kulturwissenschaften sowie der Informatik und Mathematik. Der DSC Seed Grants stärkt damit die datenwissenschaftliche, kooperative Forschung.

IT-Infrastruktur: Die spezialisierte IT-Infrastruktur ist auf die Durchführung von Data-Science-Anwendungen wie maschinelles Lernen ausgelegt. Sie bietet Kapazitäten für Virtualisierung, Speicherung und High-Performance-Computing, um das Arbeiten mit aufwendigen Algorithmen zu ermöglichen. Für Anwender:innen dient sie dabei als leicht zugängliche Struktur mit deutlich höherer Leistung als ein herkömmlicher PC. Somit schließt

sie die Lücke zwischen Rechenressourcen auf Arbeitsgruppenlevel und nationalen Supercomputern wie das System des Norddeutschen Verbunds für Hoch- und Höchstleistungsrechnen (HLRN). Im Sinne der Ressourcenschonung und -bündelung, ist die Architektur auf Skalierbarkeit ausgelegt und kann flexibel erweitert werden. Dies bietet die Möglichkeit flexibel auf die Bedürfnisse und Anforderungen unserer Nutzer:innen zu reagieren.

2.3 Angebote und Services zur Förderung des Kulturwandels

Strategieentwicklung: Das DSC unterstützt die Universität Bremen bei der Entwicklung von Strategien in Bezug auf Datenkompetenzen, Beratungsstrukturen, FDM-Leitlinien und technische Infrastruktur. Dabei wurde aus der Zusammenarbeit in der U Bremen Research Alliance im Jahr 2021 ein Whitepaper erstellt, das den aktuellen Stand sowie zukünftige Handlungsfelder für die disziplinübergreifende Etablierung eines kooperativen FDM skizziert (Pigeot u. a. 2021).

Wissenstransfer: Um einen wechselseitigen Transfer von Wissen zwischen Wissenschaft, Gesellschaft, Wirtschaft und Politik zu ermöglichen, engagiert sich das DSC bei der Umsetzung unterschiedlicher Wissenschaftskommunikationsformate. Hierzu gehören informationsvermittelnde Formate wie öffentlichen Vorlesungen oder digitale Beiträge (z.B. auf Social-Media-Kanälen oder Blogs) sowie dialogorientierte Formate wie Podiumsdiskussionen, Roundtables oder World Cafés.

Vernetzung: Die wissenschaftliche Vernetzung und der datenwissenschaftliche Austausch werden auf verschiedenen Ebenen gefördert. Lokal haben Forschende die Möglichkeit, Teil des Mitgliedernetzwerks zu werden und so ihre datenwissenschaftlichen Kompetenzen sichtbar zu machen. Seminarreihen wie das „Data Science Forum“ bieten eine Plattform für den wissenschaftlichen Dialog. Auf nationaler Ebene engagiert sich das DSC durch die Durchführung von oder den Beitrag zu Community-Events und Workshops. Im internationalen Kontext wird die Vernetzung von Forschenden und der Aufbau neuer Forschungskooperationen durch länderübergreifende Austauschformate wie dem „Bremen-Cardiff Data Science Workshop“ gefördert. Ziel dieser Aktivitäten ist es, die interinstitutionelle, disziplinübergreifende Zusammenarbeit zu fördern und den wechselseitigen Kompetenztransfer zu ermöglichen.

Bewusstsein schaffen: Um Forschende für einen FAIRen Umgang mit Daten zu sensibilisieren und ihnen die damit verbundenen Vorteile für sich und ihre Forschung aufzuzeigen, bietet das DSC unterschiedliche niederschwellige Informationsangebote. Dazu gehört beispielweise das hybride Coffee-Lecture-Format „Data Snacks“, das in kurzweiligen Sessions zu aktuellen Themen, wichtigen datenwissenschaftlichen Aspekten und hilfreichen Diensten informiert. Darüber hinaus steht das DSC für Expert:innen-Inputs bei Early Career Events, an Instituten oder Fachbereichen sowie für Lehrveranstaltungen zur Verfügung. Durch die aktive Ansprache von Forschenden wird eine individuelle Sensibilisierung für einen verantwortungsvollen Umgang mit Daten erreicht.

3 Ausblick

Das DSC hat sich bereits jetzt als bedeutende Institution an der Universität Bremen und in der Bremer Forschungslandschaft positioniert und kann für andere Regionen als Modell für die institutionelle Zusammenführung von FDM- und Data-Science-Kompetenzen dienen. Im nächsten Schritt soll die die Zusammenarbeit und der Kompetenzaustausch mit anderen „Data Science Centern“ und Data-Science-Initiativen vertieft werden. Dafür organisiert das DSC im Rahmen der INFORMATIK 2023 einen Community-Workshop zum Thema „Aktuelle Entwicklungen und Perspektiven (an Hochschulen) im Bereich Data Science“.[2]. Die Angebote und Aktivitäten des DSC werden kontinuierlich weiterentwickelt, basierend auf den Bedürfnissen der Nutzenden. Die Bedarfserfassung erfolgt beispielsweise durch quantitative Umfragen und im Dialog mit Forschenden z.B. in Beratungen. Die Wirksamkeit der Maßnahmen wird über verschiedene Kennzahlen bewertet, darunter die Anzahl und fachliche Diversität der Nutzer:innen in Bezug auf Beratungen, Trainings, Seed Grants, IT-Infrastruktur, Netzwerkevents und Wissenstransferformate. Auch die Anzahl der unterstützten Drittmittelanträge und Forschungsprojekte sowie die eingeworbenen Drittmittel werden berücksichtigt. Die Qualität der Trainings wird außerdem über Evaluationen sichergestellt. Zusätzlich betrachtet das DSC die Teilnahme an (inter)nationalen Community-Events sowie die Veröffentlichung von Materialien und Konzepten (wie Trainingsinhalte und Coffee Lectures) zur Nachnutzung als essenziell, um einen standortübergreifenden Kompetenztransfer zu ermöglichen und den kulturellen Wandel voranzutreiben.

Um der steigenden Nachfrage nach Unterstützungsangeboten gerecht zu werden, ist perspektivisch eine Erweiterung der personellen Unterstützungsstruktur geplant. Hierbei sollen zusätzliche zentral koordinierte Data Scientists und Data Stewards für unterschiedliche Disziplinen eingebunden werden. Eine Finanzierung kann beispielsweise durch Verbundprojekte wie Exzellenzcluster, Sonderforschungsbereiche oder Graduiertenkollegs sichergestellt werden.

Danksagung

Die Autor:innen danken dem Land Bremen für die Förderung des Data Science Centers.

Literaturverzeichnis

Hörner, Tanja, Iris Pigeot, Frank Oliver Glöckner und Rolf Drechsler. 2021. „Disziplinübergreifendes Modell zur Ausbildung von Forschungsdatenmanagement und Data Science Kompetenzen: 'Data Train – Training in Research Data Management and Data Science'“. *Bausteine Forschungsdatenmanagement,* Nr. 3: 56–69. DOI: `https://doi.org/10.17192/bfdm.2021.3.8343`.

2 `https://informatik2023.gi.de` und `https://www.dsc-ub.de/GI-data-science-workshop.php`; *Zuletzt aufgerufen am 11.Mai 2023.*

Pigeot, Iris, Frank Oliver Glöckner, Rolf Drechsler, Tanja Hörner, Derk Hergen Schönfeld, Lena Steinmann und Björn Oliver Schmidt. 2021. *Etablierung eines kooperativen Forschungsdatenmanagements in der U Bremen Research Alliance.* Technischer Bericht. U Bremen Research Alliance. DOI: https://doi.org/10.5281/zenodo.4775371.

Steinmann, Lena, und Rolf Drechsler. 2021. „Verzahnung von Data Stewardship und Data Science – Wege und Perspektiven". *Bausteine Forschungsdatenmanagement,* Nr. 3: 82–91. DOI: https://doi.org/10.17192/bfdm.2021.3.8342.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Wissenschaftsrat. 2020. „Zum Wandel in den Wissenschaften durch datenintensive Forschung". Besucht am 11. Mai 2023. https://www.wissenschaftsrat.de/download/2020/8667-20.html.

Leibniz Data Manager – Data Management Across Various Research Data Repositories

Angelina Kraft [1], Anna Beer [2], Mauricio Brunet [1], Ahmad Sakor [1], Maria-Esther Vidal [1,3]

[1]TIB – Leibniz Information Centre for Science and Technology;
[2]University of Hildesheim, University Library;
[3]Leibniz University Hannover & L3S Research Center

The Leibniz Data Manager (LDM) is a research data management system that resorts to Semantic Web technologies to empower FAIR principles. LDM supports searching and exploring research data across various digital repositories by providing a (meta-)data management layer for digital sources based on the web-based data catalog software CKAN (Comprehensive Knowledge Archive Network). The LDM allows users to preview research data, e.g., tables, audio-visual material like AutoCAD files or 2D and 3D data, or live programming code via Jupyter Notebook(s) so that their potential for re-use can be easily evaluated. LDM is available as a Docker container, enabling the installation of local LDM distributions to assist research data management in different phases of the research data life cycle. LDM is publicly accessible at `https://service.tib.eu/ldmservice`.

1 Introduction

The FAIR principles (Findable, Accessible, Interoperable, Reusable; Wilkinson et al. 2016; Hodson et al. 2018) and reproducible guidelines aim at advising the publication of scientific digital objects. Additionally, research data repositories and code repositories provide the basis for supporting researchers during publication and validation processes. In this process, research data repositories (i.e., listed via re3data.org) offer the possibility to explore published digital objects.

However, their scope does not allow for holistic management of scientific objects and research data so that computational transparency is endured over time. The data ecosystem of research data repositories consists of various available categories and types: Discipline-specific repositories, interdisciplinary repositories, institutional repositories, and mixtures thereof. With this heterogeneity comes large variations in data and metadata standards, APIs, file formats, license information, archival and publication guidelines, terms of re-use, and others. This is also the reason why a search across multiple repositories is considered

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.c18084`

Table 1: Key LDM features in the Research Data Life Cycle.

Steps in the Data Life Cycle	LDM features
Data collection	Data selection from repositories, metadata is mapped
Data curation	Definition of various access privileges
Data analysis	E.g. Jupyter Notebook(s) enable the execution of live code
Data re-use	Data visualizations enable preview without download

a time-consuming task to be carried out by researchers who want to re-use data but are still determining where to look for it. Geisler et al. (2021) discuss a list of requirements to be met in a data ecosystem. Knowledge-driven data ecosystems are positioned as frameworks for enhancing transparency in data exchange.

The Leibniz Data Manager (LDM) provides a tool that can aid in the transition from a publication- or article-based to an information-based (linked-data) research workflow (Table 1). This happens by further developing a CKAN-based software distribution that allows indexing metadata and data across digital repositories, using the existing vocabularies like the Data Catalog Vocabulary (DCAT) to map metadata standards. Currently, LDM connects three pilot repositories as a proof-of-concept. With this technology in place, an intuitive user interface allows performing a data search for relevant and related data sets across the connected repositories, screening for relevant data and ultimately taking another step towards the reproducibility of science.

2 Architecture

The Leibniz Data Manager (Figure 1) solves interoperability across repositories and integrates data sets published in other repositories. The LDM is available as an open source software distribution since April 2021.[1] An up-to-date user documentation and maintenance documentation is available and linked at the homepage, and requests for functionality updates of the LDM are managed via GitHub.

LDM offers a simple, small-scale, and open software distribution which can connect digital repositories in a way such that data sets and other digital scientific objects will stay in their respective repositories, while LDM provides an integrated view of the metadata. Figure 1 depicts the main components for research data management and analysis (modified after Beer et al. 2022b). Data collections are derived from data sets in heterogeneous formats; also, data catalogs can be integrated from existing repositories (e.g., the research data repository of the Leibniz University Hanover). Metadata describing the data set is collected from the data provider; and existing vocabularies, e.g., DCAT (Albertoni et al. 2023) and DataCite (DataCite Metadata Working Group 2021) are utilized to describe the metadata. A newly created data collection is uniquely and persistently

1 `https://service.tib.eu/ldmservice`

identified by generating a Digital Object Identifier (DOI). The user can define a scheduler for synchronizing the data collection with the other data set providers (Chamanara et al. 2019). Lastly, a user can describe the access regulations. Once a data set is part of the LDM catalog, data and metadata are created and synchronized according to the schedule defined during the data creation step.

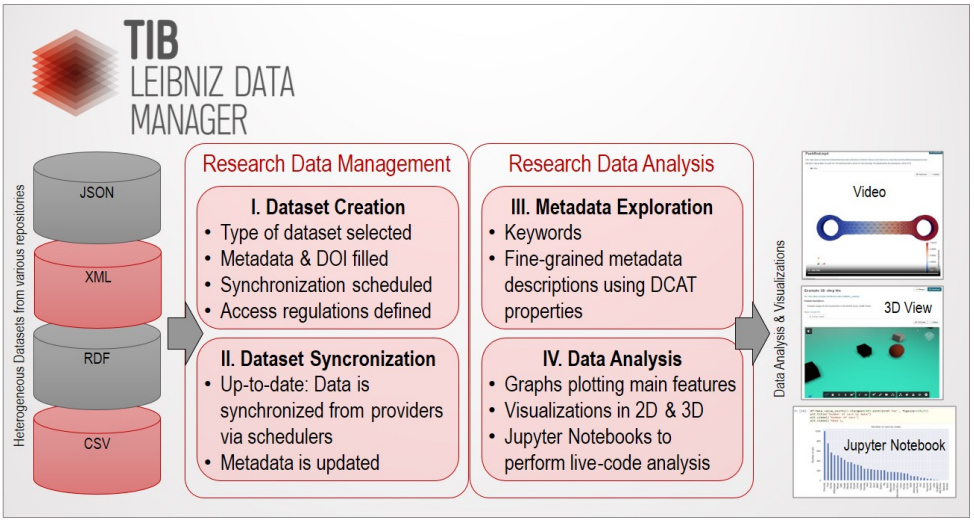


Figure 1: The Leibniz Data Manager: An adaptive RDM system which supports data search, exploration, analysis and visualization across various research data repositories.

With the LDM other information infrastructure providers may offer data "showcases" by semantically connecting existing data catalogs and repositories (Beer et al. 2022a). This enables researchers to find and analyze published digital research objects that may be of relevance for their own research. LDM offers Web APIs for traversing the LDM catalog and uploading new digital objects. Moreover, it enables the publication of data services and live code over LDM data sets. Following Linked Data principles, the LDM catalog is modeled as an RDF knowledge graph. A SPARQL endpoint allows for querying the RDF factual statements of all LDM digital objects.

Another main use of LDM is its application as a data management training tool for researchers. As a small-scale software distribution, LDM is easily installed at a local computer for training purposes. Demonstrating data sets are also available for this purpose. In 2023, LDM-training sessions were created and implemented within Master and PhD courses provided by the Leibniz University Hanover. An example are the courses "Responsible Research Data Management" in the doctoral program BIOMEDAS (BIOMEDAS Management 2023). BIOMEDAS was developed as a cross-university PhD program within the Academy of the Translational Alliance of Lower Saxony (TRAIN).

3 Support

TIB operates the hosting and long-term availability of LDM, which includes the data collection and analysis service and the availability of code and documentation on GitHub.

This also includes server infrastructure, storage, and personnel for server and application administration.

4 Conclusion and Outlook

The Leibniz Data Manager is a research data management system that supports searching and exploring research data across various repositories. LDM follows the FAIR data principles and resorts to standard vocabularies to represent metadata about research digital objects. Furthermore, LDM features a (meta-)data management layer for digital objects, which enables researchers to preview research data, e.g., tables, audio-visual material like AutoCAD files or 2D and 3D data, or live programming code via Jupyter Notebook(s) so that their potential for reuse can be easily evaluated. In a next step, an LDM extension with the most promising FAIR developments and recommendations – e.g., indexing of FAIR Digital Objects (Smedt, Koureas, and Wittenburg 2020) and FAIR Signposting (FAIRsharing Team 2023) – is planned. This will contribute to a better machine-actionability and semantic interoperability of the LDM from a data provider and data consumer point of view towards a knowledge-driven ecosystem of research data.

Acknowledgements

The project is funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the LIS Funding Programme *e-Research Technologies* (grant no. 438302-423). This article is based on the accepted poster for the E-Science-Tage 2023 (DOI: `https://doi.org/10.11588/heidok.00033144`).

ORCID:

- Angelina Kraft https://orcid.org/0000-0002-6454-335X
- Anna Beer https://orcid.org/0000-0002-3447-0575
- Mauricio Brunet https://orcid.org/0000-0001-9576-8845
- Ahmad Sakor https://orcid.org/0000-0001-8007-7021
- Maria-Esther Vidal https://orcid.org/0000-0003-1160-8727

References

Albertoni, Riccardo, David Browning, Simon Cox, Alejandra Gonzalez Beltran, Andrea Perego, and Peter Winstanley. 2023. "Data Catalog Vocabulary (DCAT) – Version 2". `https://www.w3.org/TR/vocab-dcat-2/`.

Beer, Anna, Mauricio Brunet, Vibhav Srivastava, and Maria-Esther Vidal. 2022a. *Dataset: LDM Demo.* Technical report. Leibniz Information Centre For Science and Technology University Library (TIB). DOI: https://doi.org/10.57702/wb5bok01.

———. 2022b. "Leibniz Data Manager – A Research Data Management System". In *The Semantic Web: ESWC 2022 Satellite Events,* 73–77. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-031-11609-4_14.

BIOMEDAS Management. 2023. "Translation Alliance Lower Saxony (TRAIN)". Visited on April 27, 2022. https://translationsallianz.de/train-academy/biomedas.

Chamanara, Javad, Angelina Kraft, Sören Auer, and Oliver Koepler. 2019. "Towards Semantic Integration of Federated Research Data". *Datenbank-Spektrum* 19 (2): 87–94. DOI: https://doi.org/10.1007/s13222-019-00315-w.

DataCite Metadata Working Group. 2021. "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4". DOI: https://doi.org/10.14454/3W3Z-SA82. https://schema.datacite.org/meta/kernel-4.4/.

FAIRsharing Team. 2023. *FAIR Signposting.* DOI: https://doi.org/10.25504/FAIRsharing.d7622d.

Geisler, Sandra, Maria-Esther Vidal, Cinzia Cappiello, Bernadette Farias Lóscio, Avigdor Gal, Matthias Jarke, Maurizio Lenzerini, et al. 2021. "Knowledge-Driven Data Ecosystems Toward Data Transparency". *Journal of Data and Information Quality* 14 (1): 1–12. DOI: https://doi.org/10.1145/3467022.

Hodson, Simon, Sarah Jones, Sandra Collins, Françoise Genova, Natalie Harrower, Leif Laaksonen, Daniel Mietchen, Rūta Petrauskaité, and Peter Wittenburg. 2018. *Turning FAIR data into reality: interim report from the European Commission Expert Group on FAIR data.* Technical report. European Commission Expert Group on FAIR data. DOI: https://doi.org/10.5281/zenodo.1285272.

Smedt, Koenraad de, Dimitris Koureas, and Peter Wittenburg. 2020. "FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units". PII: publications8020021, *Publications* 8 (2): 21. DOI: https://doi.org/10.3390/publications8020021.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Schöne neue Laborwelt – Elektronische Laborbücher digitalisieren die Labordokumentation

Bert Zulauf, Nina Knipprath

Zentrum für Informations- und Medientechnologie (ZIM), Heinrich-Heine-Universität Düsseldorf

An deutschen Universitäten und Fachhochschulen erschließt sich aktuell im Bereich der Struktur und des Aufbaus von Laborarbeit eine neue Arbeitsweise: Dokumentation mittels elektronischer Laborbücher (ELB). Forschende produzieren und nutzen eine bunte Mischung an Daten wie Grafikdateien, Formeln, Tabellen oder Mikroskopie-Dateien. An der Heinrich-Heine-Universität Düsseldorf (HHU) findet regelmäßig Beratung und Unterstützung von Forschenden im Bereich des Forschungsdatenmanagements (FDM) statt. Neben Hilfestellungen bei der Antragsstellung bieten wir auch eine Unterstützung mittels diverser FDM-Tools an, unter anderem auch mit elektronischen Laborbüchern. Warum es sinnvoll ist, elektronische Laborbücher einzusetzen, wie unser Weg damit bisher verlaufen ist und was aktuell zu überwindende Hürden sind, stellen wir in diesem Bericht dar.

1 Einleitung

Das Zentrum für Informations- und Medientechnologie (ZIM) ist Teil des Forschungsdatenkompetenzzentrums (FDMK) der HHU und stellte Forschenden bereits seit einigen Jahren das lizenzbasierte ELB „Labfolder“ in einer eigenen Hosting Variante zur Verfügung. Im Zuge des vom deutschen Bundesministerium für Bildung und Forschung (BMBF) geförderten Kooperationsprojekts „FoDaKo“ (Knipprath u. a. 2020), einer Zusammenarbeit von den drei nordrhein-westfälischen (NRW) Universitäten Düsseldorf, Wuppertal und Siegen, bauen wir nach Evaluation von verschiedenen Softwarelösungen seit 2018 am ZIM mit „eLabFTW“ eine Open Source Lösung auf.

Aus infrastruktureller Sicht haben wir seitdem maßgeblich eLabFTW gefördert, da es besser zu den Anforderungen der Forschenden passt. Denn Labfolder geht mit Lizenzkosten einher und ist damit nur schwer für kurzfristige Forschungsprojekte mit Gastwissenschaftlern sinnvoll einsetzbar, ganz zu schweigen von einem kostengünstigen Einsatz in der Lehre, den wir aber ebenso vorantreiben wollten.

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.c18085

Auf dem Campus der Heinrich-Heine-Universität wurde ein Neubau mit 70 % Laborflächen für die Biowissenschaften errichtet, der mittlerweile bezogen ist. Ein guter Zeitpunkt, die Digitalisierung der Labore mithilfe elektronischer Laborbücher weiter voranzutreiben und in die Breite zu tragen.

2 Erste Anforderungen an ein ELB

Beide Entscheidungen für die Softwarelösungen fielen aufgrund von Wünschen der Forschenden speziell nach diesen ELB. Ein wichtiges Kriterium aus Infrastruktursicht war es, vorerst zu einer generischen Lösung zu kommen, mit der eine möglichst große Zielgruppe Forschender gut arbeiten kann. Ein weiterer wichtiger Punkt war, dass es keiner zusätzlichen Software bedarf, um ein ELB zu benutzen, sondern der Service komplett webbasiert angeboten werden kann. Nach einer gleichzeitigen Pilotphase beider Laborbücher entschied man sich dazu, den Fokus auf die Open Source Lösung zu setzen, und nur noch eLabFTW zu betreiben, da diese mehr Anklang unter den Forschenden fand. Deswegen stellte die HHU die Hosting Variante von Labfolder Anfang 2022 ein. Der „ELN-Wegweiser“ der Deutschen Zentralbibliothek für Medizin (ZB-MED; Adam u. a. 2023), sowie der neu erschienene „ELN-Finder“[1], sind sehr gute Leitlinien, um sich für eine Softwarelösung zu entscheiden.

Die Laborarbeit sollte gut strukturiert dokumentiert werden können. Gleichzeitig wollen Forschende auch, vor allem aber in der aktiven Praxisphase, schnell Informationen in einem unstrukturierten Bereich ablegen können, um sie in der Nacharbeit sauber zu dokumentieren. So stellt also die Nutzerseite verschiedene Arten von Anforderungen an ein ELB: Neben den technischen Anforderungen wie Benutzerfreundlichkeit, der Verwendung bekannter Dateiformate und der ständigen digitalen Verfügbarkeit der Daten existiert auch der Wunsch nach Möglichkeit seine Daten komplett wieder aus der ELB Software zu entnehmen und mit anderen Tools weiterzuarbeiten. Hinzu kommen spezifische fachliche Anforderungen, welche sich unter Umständen stark unterscheiden können. Während die Biologen und die Physiker mit dem generischen Tool eLabFTW zufrieden sind, haben die Chemiker speziellere Anforderungen an ein ELB. Sie benötigen beispielsweise einen spezifischen chemischen Formeleditor, damit die Molmasse automatisch übernommen werden kann und damit eine Struktursuche möglich ist. Außerdem können die Messdaten zwar hoch- und heruntergeladen werden, jedoch gibt es kein direktes Öffnen in einer geeigneten Software. Eine der regulatorischen Anforderungen, gerade bei Drittmittelprojekten, ist die Einhaltung der guten wissenschaftlichen Praxis und guten Laborpraxis. Dazu gehört ebenso die sichere Aufbewahrung der Forschungsdaten, wie auch das Sicherstellen der Verfügbarkeit und Auffindbarkeit dieser.

1 `https://eln-finder.ulb.tu-darmstadt.de/home`; *Zuletzt aufgerufen am 20. April 2023.*

2.1 Einbindung ins Curriculum

Trotz der stetig wachsenden Benutzung besteht häufig eine interne Schwelle das elektronische Laborbuch auch einzusetzen. Die Motivation der Mitarbeitenden ist hier der wesentliche Punkt, an dem angesetzt werden muss. So sind wir vom ZIM bereits mit Forschenden im engen Austausch, das Thema schon früh Studierenden nahezubringen und attraktiv als Teil des Curriculums darzustellen, um damit die positive Haltung von Studierenden gegenüber Digitalisierungsprojekten auch bei der Digitalisierung von Laborarbeiten zu nutzen.

Mit dem Ziel einen Kulturwechsel hin zu elektronischen Laborbüchern zu gestalten, führte man, als Pilotprojekt in der Neurobiologie, die Master-Studierenden des Kurses „Cellular and molecular analysis of brain function“ mit Hilfe von Videotutorials in die Laborbuchsoftware ein, um unter Laborbedingungen erste Schritte in fluorometrischem Imaging zu erlernen, d.h. wie Daten generiert und ausgewertet werden sollen. Studierende können leichter in die Forschungsarbeit eingebunden werden, da über das ELB Einträge leicht freigegeben werden können. In enger Zusammenarbeit mit den Lehrenden kann so das Laborbuch für die Lehrzwecke optimiert werden. So lernen junge Wissenschaftler früh den Umgang mit elektronischen Laborbüchern und werden in der Forschung an diese neuen Standards anknüpfen.

Um die verschiedenen Anforderungen von Forschung und Lehre bedienen zu können, haben wir eine eigene Instanz von eLabFTW speziell für die Lehre bereitgestellt. Hier ist es den Lehrenden wichtig, dass Testdaten auch wieder gelöscht werden können, und dass Studierende während des Ausprobierens keine wichtigen Daten der Forschungsdokumentation verändern. Trotzdem können Experimente aus dem System der Lehre über die Export & Import Funktionalität in die Hauptinstanz übernommen werden.

3 Vor-Ort-Unterstützung und gemeinsame regionale Zusammenarbeit

Um grundsätzliche Bedarfe von Forschenden zu identifizieren hat uns die Gründung einer NRW-weiten Arbeitsgruppe zum Thema ELB geholfen.[2] Die Kollaboration der Infrastruktureinrichtungen hat uns voneinander lernen lassen: So erhielten wir Hilfe bei der Implementierung von einer zentralen Authentifizierungsmethode, und helfen anderen beim Einsatz in der Lehre oder wenn es um konzeptionelle Fragestellungen geht. Unter anderem wird in dieser Arbeitsgruppe über die technische Einführung, sowie IT-Sicherheitskonzepte und Schnittstellen zu anderer Forschungsdateninfrastruktur diskutiert.

Die Vernetzung der verschiedenen NRW-Einrichtungen bietet auch einen guten Überblick, denn neben eLabFTW verwenden diese Einrichtungen u.a. auch Labfolder, eLabJournal

2 `https://wiki.hhu.de/x/vAMdB`; *Zuletzt aufgerufen am 20. April 2023.*

oder RSpace als Laborbuchlösungen[3]. Forschende aus ganz Deutschland wurden auf unsere Pionierarbeit aufmerksam und traten in Kontakt mit uns oder der Arbeitsgruppe. So wuchs die Gemeinschaft auf mittlerweile 100 Mitglieder an. Zusammen haben wir einen Leitfaden entwickelt, der Infrastruktureinrichtungen bei der Einführung von elektronischen Laborbüchern helfen soll (Adam u. a. 2023).

Gleichzeitig wurde im Kooperationsprojekt OER.DigiChem.nrw das Thema der elektronischen Laborbücher in der Chemie speziell mit Fokus auf Studierende erweitert. Im Rahmen dieser Kooperation erstellen die HHU Düsseldorf, die Bergische Universität Wuppertal und die Technische Hochschule Köln zusammen freie Bildungsmaterialien (OER) für Studierende und entwickeln Materialien mit denen intentionales Lernen in der Studieneingangsphase besser berücksichtigt werden kann (Mertineit u. a. 2021). Die Tutorialvideos zur Nutzung von eLabFTW können über das Portal ORCA.nrw abgerufen werden, und sind sehr gut geeignet zur Gestaltung oder Ergänzung von fachspezifischen Lernräumen zum Thema ELB innerhalb von Learning Management Systemen.

3.1 Rahmenbedingungen an der HHU

Wir erhielten einen besseren Überblick der Forschungsszenarien am eigenen Standort, indem wir verschiedene Forschende aus unterschiedlichen Fachbereichen, wie beispielsweise aus der Physik, der Biologie und der Medizin, eingeladen haben, um über ihre Anforderungen zu sprechen, und um Use cases zu entwickeln (Zulauf und Knipprath 2019). Hier konnten wir gemeinsame Rahmenbedingungen schaffen, was den Service eLabFTW betrifft. Aktuell treffen wir uns monatlich mit Power-Usern, um über Probleme und Möglichkeiten in Bezug auf die Nutzung von elektronischen Laborbüchern an der HHU zu sprechen.

Zur Etablierung des Services wurden Gespräche mit dem Personalrat geführt, erste Versionen von Verfahrensverzeichnissen erstellt und über Datenschutzbetrachtungen nachgedacht. Man muss den Spagat zwischen den Datenschutzrechten der Einzelnen und der guten wissenschaftlichen Praxis meistern, also die Löschmöglichkeit der personenbezogenen Daten gegen Aufbewahrungs- und Nachweispflicht abwägen. Gelöst haben wir das über eine Zustimmung der Nutzenden beim Login über Shibboleth, sowie über den Punkt der Labordokumentation und Aufbewahrungsfrist in den Nutzungsbedingungen. Es muss demnach eine regelmäßige Sensibilisierung der Mitarbeitenden durchgeführt werden, damit die korrekte Dokumentation der im Laborbuch zugelassenen Forschungsdaten gewährleistet ist. Außerdem liegt es in der Verantwortung der Team-Admins, regelmäßig zu prüfen, ob die vorgegebene Aufbewahrungsfrist der gespeicherten Daten erreicht wurde, und diese somit gelöscht werden sollten.

Unveränderlichkeit wird innerhalb eLabFTW durch eine Zeitstempel-Möglichkeit gewährleistet. Hier können Forschende ihre Arbeit zu diesem Augenblick beweissicher festhalten. Eine digitale Möglichkeit Aufzeichnungen über gentechnische Arbeiten nach Gentechnik

3 `https://www.forschungsdaten.org/index.php/Elektronische_Laborbücher`; *Zuletzt aufgerufen am 20. April 2023.*

Aufzeichnungsverordnung (das sogenannte „Formblatt-Z“) ebenfalls elektronisch zu speichern, wird momentan an unserer Universität diskutiert und entwickelt.

Aktuell erheben wir auch Kennzahlen im Kontext des FDMK, um unsere Forschenden gezielter unterstützen zu können, und entwickelten in einem Projekt unserer Auszubildenden Fachinformatiker ein Tool, welches unter anderem die Benutzer von eLabFTW in Fakultäten und Rollen einteilt. Hauptnutzer ist die Mathematisch-Naturwissenschaftliche Fakultät mit 400 Mitarbeitenden und 150 Studierenden, gefolgt von der Medizinischen Fakultät mit 230 Mitarbeitenden und 50 Studierenden, welche sich insgesamt in 88 Teams befinden und 25.000 Experimente angelegt haben.

3.2 Anwenderbeispiele

Das FDMK begleitet den SFB1208 sehr intensiv in Bezug auf das Forschungsdatenmanagement. Die Gruppe hat die Nutzung von eLabFTW als verpflichtend eingeführt und entwickelt aktuell in engem Austausch mit uns eine Möglichkeit um Daten aus dem elektronischen Laborbuch zu extrahieren und in einem Archiv zu speichern. Für Forschungsdaten bieten wir bereits ein institutionelles Repositorium auf Basis der Open Source Software DSpace an der HHU an. Jetzt haben Forschende des SFB1208 das Tool LISTER - „Life Science Experiment Metadata Parser“ (Musyaffa, Rapp und Gohlke 2023) programmiert, welches über die eLabFTW API einen Export mit zusätzlichen Metadaten qualifiziert und als JSON Datei an eine Schnittstelle unseres separaten, auf DSpace basierenden Archives, weitergibt. Dort werden die Metadaten sowohl in die DSpace Metadaten übernommen wie auch als Bestandteil der zu archivierenden Daten im S3 Objektstore bzw. Langzeitspeicher abgelegt. Durch die Vorgehensweise sind die Daten nachhaltig und robust auch ohne DSpace als Repository-Anwendung verwendbar. Ein Ziel ist es auch im Laborbuch-Archiv sehr spezifisch, ähnlich wie in der Laborbuchplattform, nach Experimenten suchen zu können. Daneben wird die Plattform entlastet und das Vorhaben unterstützt die gute wissenschaftliche Praxis.

Ein weiteres Anwendungsbeispiel findet sich im Bereich der vorklinischen Studien der Medizin. Dort wird eLabFTW als zentrale Schnittstelle für die Komponenten des Qualitätsmanagementsystems und die digitale Speicherung von Laborjournalen verwendet. Es wurden Standardarbeitsanweisungen (Standard Operating Procedure - SOP) sowie eine verwaltete Biobank für die sichere Langzeitlagerung von Bioproben eingeführt (Hewera u. a. 2020).

Bereits erzielter Nutzen kann am Beispiel der Neurobiologie sehr gut dargestellt werden: Hier wird die Chemikaliendatenbank innerhalb von eLabFTW nach einer festgelegten Vorgabe von allen gepflegt. Jede der 400 Chemikalien hat einen eigenen Eintrag in eLabFTW welcher eine Tabelle beeinhaltet, in der auch eingetragen werden kann, falls diese Chemikalie nachbestellt werden muss. Die technischen Assistenten haben so über ein Lesezeichen direkt eine Einkaufsliste zur Hand. Zudem können die Forschenden die Chemikalien direkt mit ihren Experimenten verknüpften und haben die H- und P-Sätze (Gefahren und Sicherheitshinweise) schnell auf einen Blick. Durch die Einführung von elektronischen

Laborbüchern konnte so auch die viel komfortablere Arbeitsweise mit einer Chemikaliendatenbank etabliert werden. Neben den Chemikalien verwaltet die Arbeitsgruppe auch ihre optischen Elemente, Protokolle und Listen sowie Präparationslösungen über eLabFTW. Durch die Nutzung von eLabFTW ergab sich eine große Zeitersparnis, weil keine bzw. weniger Wiederholungen gemacht werden müssen, Experimente und weitere Daten schneller wiederauffindbar sind, und man leichter an alte Projekte anschließen kann.

Das Feedback der sehr unterschiedlichen Forschergruppen ist zum einen Inspiration für andere Forschende und bietet für uns einen wertvollen Input für die Weiterentwicklung der Plattform. Die Forschenden stehen nicht nur mit den Infrastrukturmitarbeitenden und untereinander in ständigem Austausch, sondern wir beteiligen uns auch aktiv an der eLabFTW Community und halten regelmäßig Rücksprache mit dem Hauptentwickler und seiner Firma, um die Open-Source Software kontinuierlich weiterzuentwickeln und an neue Gegebenheiten anzupassen.

Literaturverzeichnis

Adam, Beatrix, Lukas C. Bossert, Magdalene Cyra, Matthias Grönewald, Stephan Janosch, Nina Knipprath, Birte Lindstädt u. a. 2023. „Raus aus dem Kladdenchaos: Elektronische Laborbücher als zentrale Dienstleistung – Erfahrungen und Empfehlungen“. DOI: https://doi.org/10.5281/zenodo.7529588.

Hewera, Michael, Ann-Christin Nickel, Nina Knipprath, Sajjad Muhammad, Xiaolong Fan, Hans-Jakob Steiger, Daniel Hänggi und Ulf Dietrich Kahlert. 2020. „An inexpensive and easy-to-implement approach to a Quality Management System for an academic research lab“. Version 2; peer review: 2 approved, 1 approved with reservations, *F1000Research* 9:660. DOI: https://doi.org/10.12688/f1000research.24494.2.

Knipprath, Nina, Torsten Rathmann, Jessica Stegemann, Bert Zulauf, Thomas von Rekowski und Maurice Schleußinger. 2020. „Schlussbericht FoDaKo - Forschungsdatenmanagement in Kooperation“. DOI: https://doi.org/10.25819/ubsi/728.

Mertineit, Ann-Kathrin, Klaus Schaper, Claudia Bohrmann-Linde, Dirk Burdinski, Bert Zulauf, Nico Meuter, Hans-Niklas Hackrath, Richard Kremer und Nina Knipprath. 2021. „Collaborative development of open educational resources for building competencies in the use of digital tools in chemistry“. In *ICERI2021 Proceedings.* IATED. DOI: https://doi.org/10.21125/iceri.2021.0325.

Musyaffa, Fathoni A., Kirsten Rapp und Holger Gohlke. 2023. „LISTER: Semi-automatic metadata extraction from annotated experiment documentation in eLabFTW“. DOI: https://doi.org/10.1101/2023.02.20.529231.

Zulauf, Bert, und Nina Knipprath. 2019. „Electronic Lab Notebooks - early research practice in teaching“. DOI: https://doi.org/10.5281/zenodo.4787258.

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.

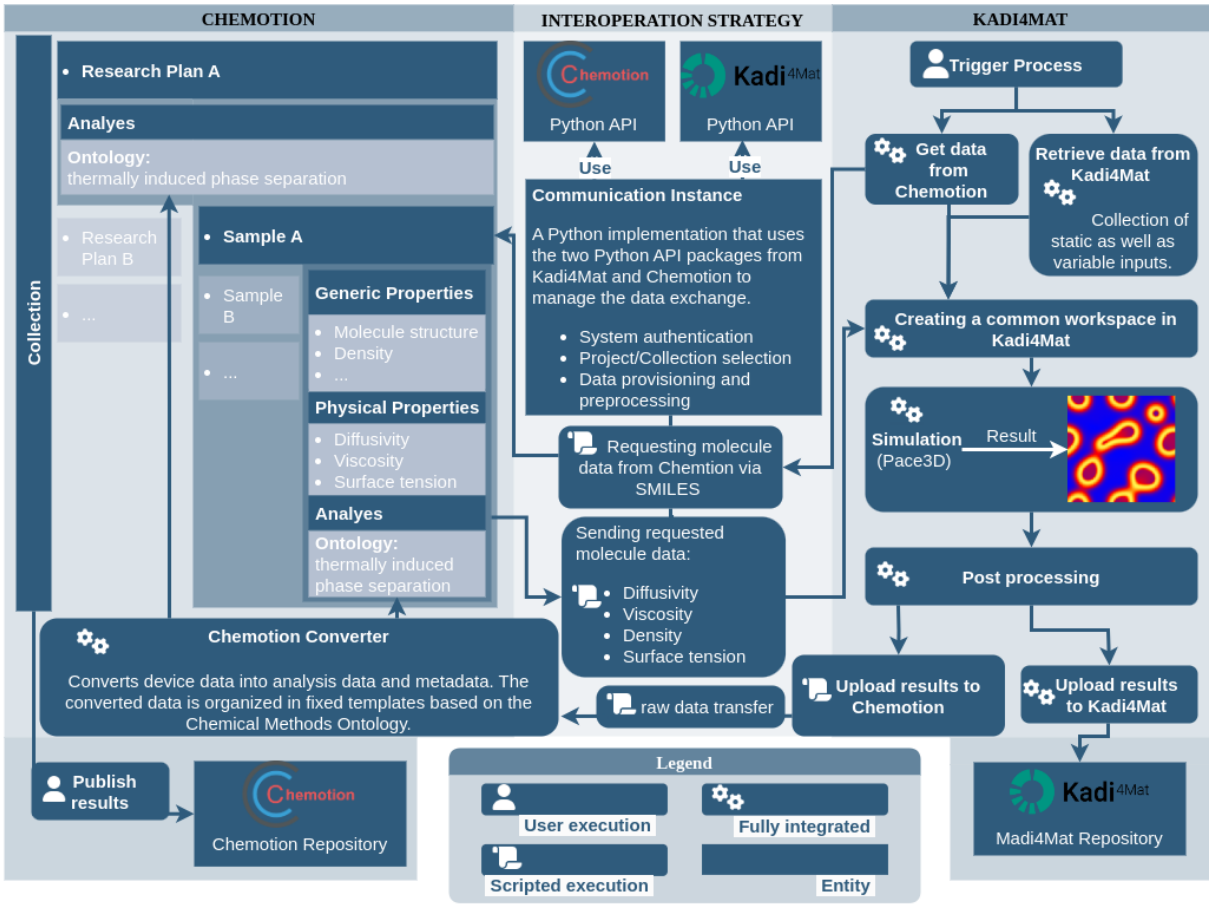


Figure 1: Interaction strategy between the Kadi4Mat and Chemotion RDM infrastructures. The legend explains the different levels of integration for each process, from user execution, via scripted execution, to full integration.

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.

Stärkung von FDM-Services im Verbund – Ergebnisse einer Bedarfserhebung

Angela Ariza de Schellenberger[1], Evgeny Bobrov[1], Kerstin Helbig[2], Denise Jäckel[2], Monika Kuberek[3], Lea-Sophie Orozco Prado[4], Elisabeth Maria Schlagberger[3], Sibylle Söring[4], Britta Steinke[4]

[1]QUEST Center for Responsible Research, Berlin Institute of Health, Charité – Universitätsmedizin Berlin;
[2]Computer- und Medienservice, Humboldt-Universität zu Berlin;
[3]Universitätsbibliothek, Technische Universität Berlin;
[4]Universitätsbibliothek, Freie Universität Berlin

Um Forschende bestmöglich im Forschungsdatenmanagement (FDM) zu unterstützen, sind zentrale Dienstleistungen und Werkzeuge unverzichtbar. Neben den bestehenden institutionellen FDM-Services gilt es dabei auch, Potentiale für gemeinsame Dienste im regionalen Verbund zu eruieren, um ein nachhaltiges FDM umzusetzen. Ein entscheidender Bestandteil bei der Entwicklung entsprechender Konzepte ist die systematische Einbeziehung der Forschenden, um die Services nah an ihren Bedürfnissen auszurichten, eine größtmögliche Identifikation mit zu entwickelnden Maßnahmen zu gewährleisten und spezifische Anforderungen unterschiedlicher Fachdisziplinien berücksichtigen zu können. Zentral für den anforderungsorientierten Auf- und Ausbau von FDM-Services ist darüber hinaus die kontinuierliche Abstimmung im regionalen Verbund, z.B. durch eine systematische Bedarfsermittlung.

In diesem Kontext haben die vier in der Berlin University Alliance (BUA) zusammengeschlossenen Institutionen Freie Universität Berlin, Humboldt-Universität zu Berlin, Technische Universität Berlin und Charité – Universitätsmedizin Berlin erstmalig gemeinsam für das Bundesland Berlin eine Online-Befragung zu FDM-Desideraten unter 975 Forschenden mit einem einrichtungsübergreifenden Fragenkatalog durchgeführt. Ziel war es, Feedback zu vorhandenen Angeboten und gewünschten Services und Werkzeugen zu erhalten sowie Potentiale für gemeinsame Dienste im Rahmen des Verbundes zu ermitteln. Inhaltlich adressierte die Umfrage verschiedene Aspekte der Forschungs(daten)praxis, aber auch Anreizmechanismen und Unterstützungsformate. Insgesamt bestätigt die verbundweite Bedarfserhebung zum FDM den bereits vielfach formulierten Bedarf an Personalressourcen sowie an nachhaltiger IT-Infrastruktur. Im Hinblick auf technische Werkzeuge werden vor allem Tools für die Erstellung von Datenmanagementplänen, den Datenaustausch zwischen verschiedenen Institutionen und Systemen sowie die Datenorganisation,

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.c18087

-anonymisierung und -bereinigung benötigt. Neben technischer Infrastruktur wird vor allem Best-Practice-Wissen als hilfreich für die Umsetzung von FDM im Forschungsalltag angesehen. Darüber hinaus wird die Notwendigkeit einer erhöhten Sichtbarkeit bereits bestehender FDM-Services an den Einrichtungen deutlich. Ein von den Einrichtungen gemeinsam auszuarbeitendes Konzeptpapier wird auf Basis der Erhebungsergebnisse Empfehlungen für Kern-Services der BUA für das FDM formulieren.

1 Projektkontext

Das Projekt „Concept Development for Collaborative Research Data Management Services" (kurz: BUA-FDM[1]) wird im Rahmen der Exzellenzstrategie für Bund und Länder im Kontext der Berlin University Alliance (BUA[2]), einem Verbund der drei Berliner Universitäten (Freie Universität Berlin, Humboldt-Universität zu Berlin, Technische Universität Berlin) sowie der Charité - Universitätsmedizin Berlin, gefördert. Die BUA verfolgt fünf zentrale Ziele (Objectives), die unter anderem die gemeinsame Forschung zu gesellschaftlichen Herausforderungen und die Bündelung der Berliner Expertise zur Bewertung und Entwicklung allgemeiner Standards für Qualität und Bewertung von Forschung adressieren. Das Projekt BUA-FDM ist Teil des Objective 5 „Sharing Resources"[3], das sich den Aufbau eines BUA-weiten Netzwerks für Forschungsdienstleistungen und -infrastrukturen zum Ziel gesetzt hat. Insgesamt sollen die Stärken der vier Partnerinnen in der BUA gebündelt, der Wissenschaftsstandort Berlin gemeinsam weiterentwickelt sowie die Offenheit, Transparenz und Reproduzierbarkeit der Forschung im Sinne der FAIR-Prinzipien gefördert werden.

2 Das Projekt

Ziel von BUA-FDM ist die Stärkung von Services und Informationsstrukturen des Forschungsdatenmanagements (FDM) innerhalb der BUA (siehe für die Schwerpunkte Abbildung 1). Dabei steht die Konzeptentwicklung zum nachhaltigen Aufbau von Kompetenz, Expertise und Ressourcen zum Thema FDM für Forschende und Multiplikator:innen im Zentrum, um bereits bestehende Ressourcen bedarfsgetrieben bestmöglich einzusetzen und Synergieeffekte optimal zu nutzen. Das Projekt soll die verschiedenen FDM-Strategieprozesse an den einzelnen Einrichtungen begleiten und unterstützen sowie Empfehlungen zur FDM-Strategieentwicklung sowie zum Aufbau gemeinsamer FDM-Services innerhalb der BUA geben.

Bis zum Ende der 34-monatigen Projektlaufzeit im Dezember 2023 erarbeitet das Projekt Konzepte zur BUA-weiten Förderung, nachhaltigen Etablierung und Unterstützung von Services im Handlungsfeld FDM. Zu Projektbeginn wurde eine Selbstevaluation mit

1 `https://www.berlin-university-alliance.de/commitments/sharing-resources/fdm/index.html`; *Zuletzt aufgerufen am 14. April, 2023.*

2 `https://www.berlin-university-alliance.de/`; *Zuletzt aufgerufen am 14. April, 2023.*

3 `https://www.berlin-university-alliance.de/commitments/sharing-resources/index.html`; *Zuletzt aufgerufen am 14. April, 2023.*

Abbildung 1: Übersicht Projektschwerpunkte.

RISE-DE (Hartmann, Jacob und Weiß 2019; Ariza de Schellenberger u. a. 2023), einem Referenzmodell für Strategieprozesse im institutionellen FDM, durchgeführt, um den Ist- und den Soll-Zustand im FDM an den jeweiligen Einrichtungen zu evaluieren. Daran anschließend wurde eine standortspezifische Bestands- und Bedarfserhebung konzipiert, um die Anforderungen von Forschenden, Exzellenzclustern, Sonderforschungsbereichen und anderen Verbundvorhaben zu ermitteln (Jäckel, Helbig und Odebrecht 2022; Kuberek, Schlagberger und Steinke 2022; Taubitz, Bobrov und De Schellenberger 2022). Zusätzlich erfolgten im ersten Halbjahr 2022 verschiedene Expert:innenrunden mit Einrichtungen und Verbünden im bundesdeutschen und europäischen Raum, die aufgrund ihrer organisatorisch-strukturellen Aufstellung oder ihrer fachspezifischen Ausrichtung als Good- bzw. Best-Practice-Modelle für kollaborative Servicestrukturen dienen können. Weitere Schwerpunkte des Projekts sind die Identifikation FDM-bezogener Communities, z.B. durch themen- und fachspezifische Workshops, die Konzeptionalisierung eines als Koordinations- und Vernetzungsstelle im Handlungsfeld FDM der BUA dienendes FDM-Büros, und die Entwicklung von Empfehlungen für nachhaltige, kollaborativ nutzbare FDM-Services, -Tools und Infrastrukturen.

2.1 Bestands- und Bedarfserhebung zum FDM

Um die Desiderate im Bereich FDM der Forschenden zu sammeln und einen Überblick über mögliche Lücken im Service-Portfolio der Einrichtungen zu erhalten, wurde im Winter 2021/22 eine Bestands- und Bedarfserhebung zum FDM parallel an den vier BUA-Einrichtungen durchgeführt. Die Ergebnisse bilden dabei die zentrale Grundlage für die Entwicklung bedarfsorientierter standortspezifischer und standortübergreifender Beratungs-, Schulungs-, Kommunikations- und technischer Serviceleistungen. Ein entscheidender Bestandteil bei der Entwicklung entsprechender Konzepte ist die Einbeziehung der Forschenden sowie eine koordinierte Abstimmung im regionalen Verbund. Ziel der Befragung war es, zu ermitteln, welche Services

1. aktuell an den jeweiligen Standorten bekannt sind und genutzt werden

2. an allen Standorten gewünscht, aber noch nicht angeboten werden

3. im Verbund als fruchtbar erachtet werden, um qualitätsvolle Forschung auch institutionsübergreifend zu ermöglichen.

Dazu wurde um Feedback zu vorhandenen Angeboten oder gewünschten Services und Werkzeugen der Einrichtungen gebeten, um Bedarfe für die Umsetzung eines nachhaltigen FDMs und das Potential gemeinsamer Dienste im Rahmen der BUA zu eruieren.

Die parallel durchgeführten jeweiligen Erhebungen in den vier Einrichtungen enthielten sowohl ein gemeinsam entwickeltes einrichtungsübergreifendes generisches als auch ein institutionsspezifisches Fragenset. Zielgruppe der Online-Befragung waren alle Forschenden der Einrichtungen, insbesondere auch Beschäftigte in Verbundvorhaben (z.B. Sonderforschungsbereiche oder Exzellenzcluster) sowie forschungsunterstützendes Personal. Die Bewerbung der Umfrage erfolgte über institutionelle Mailinglisten und Twitterkanäle sowie bei der Charité über personalisierte Umfragelinks. Die institutionsspezifischen E-Mails richteten sich an verschiedene Statusgruppen, an Fakultäten, Fachbereiche, Dekanate und Forschungsdekane, relevante Zentraleinrichtungen wie Rechenzentren und Forschungsabteilungen sowie Sprecher:innen bzw. Koordinator:innen der Sonderforschungsbereiche und Exzellenzcluster. Zusätzlich wurde in den BUA-News, auf der BUA-FDM-Projekt-Website und durch Erinnerungs-E-Mails auf die Umfrage aufmerksam gemacht. An der Befragung nahmen insgesamt 975 Personen teil (davon 135 Teilnehmende an der Freien Universität Berlin, 162 Teilnehmende an der Humboldt-Universität zu Berlin, 207 Teilnehmende an der Technischen Universität Berlin, 471 Teilnehmende an der Charité - Universitätsmedizin Berlin).

Aufgrund der insgesamt geringen Rückläufe sind die Ergebnisse als nicht repräsentativ einzustufen. Dennoch können sie auf Trends hinweisen und Tendenzen abgeleitet werden.

2.2 Ergebnisse der Umfrage

Die Ergebnisse der Befragung lassen sich in den folgenden Punkten kurz zusammenfassen:

1. Die Mehrheit der Teilnehmenden ist im Rahmen der eigenen Forschungspraxis mit dem Thema FDM in Berührung gekommen, wodurch die hohe Relevanz des FDM im eigenen Forschungsumfeld deutlich wird.

2. Es besteht die Notwendigkeit zielgruppenspezifischer Angebote, um alle Statusgruppen differenziert ansprechen zu können. item Wichtige Inzentivierungsmaßnahmen im Handlungsfeld FDM sind die Schaffung von Personalressourcen für FDM-Aufgaben und die Erhöhung der Sichtbarkeit von Datenpublikationen.

3. Wissen über Best Practices, der Ausbau von IT-Infrastrukturen sowie Personalressourcen sehen die Teilnehmenden als besonders hilfreich für die praktische Umsetzung von FDM an.

4. Die FDM-Richtlinien des unmittelbaren Forschungsumfelds sind am bekanntesten und auch am häufigsten umgesetzt, im Gegensatz zu den institutionellen Forschungsdaten-Policies.

5. Die in den vier Einrichtungen derzeit angebotenen technischen Werkzeuge decken die Bedarfe in weiten Teilen des Forschungsdatenlebenszyklus- wie Projektmanage-

ment, Datenaustausch, -organisation, -dokumentation, -anonymisierung, -analyse und -bereinigung sowie Versionierung, Visualisierung und Publikation nicht ausreichend ab.

6. Es besteht ein hoher Bedarf an institutionsübergreifenden Services (z.B. Schulungen) und Infrastrukturen (z.B. Dienste für kooperatives Arbeiten und weitere FDM-Software).

Insgesamt hat die Umfrage einige Lücken im FDM-Service-Portfolio der BUA-Einrichtungen bestätigt, insbesondere hinsichtlich technischer Lösungen zur Planung des FDMs sowie zum kollaborativen aktiven Datenmanagement. Zudem bedarf es einer verbesserten Kommunikation und Bewerbung vorhandener FDM-Services, geltender Standards und Best Practices. Empfehlungen, um die Lücken zu schließen und Verbesserungen durchzuführen, werden im Laufe des Projektes folgen. Ein ausführlicher Bericht zur Erhebung und zur Auswertung der Ergebnisse ist auf Zenodo publiziert (Ariza de Schellenberger u. a. 2022).

3 Ausblick

In der verbleibenden Projektlaufzeit liegt der Fokus darauf, Konzepte zu entwickeln, um die ermittelten Desiderate für standortspezifische sowie standortübergreifende Beratungs-, Schulungs-, Kommunikations- und technische Serviceleistungen zu adressieren. Für die gemeinsame Weiterentwicklung qualitätsvoller Forschung im Verbund werden dafür ein übergreifendes Konzeptpapierpapier und Empfehlungen für Kernservices der BUA zur Verbesserung des FDM in Abstimmung mit allen vier Einrichtungen erarbeitet. Dies soll Kompetenzen, Expertisen und Ressourcen langfristig auf- sowie ausbauen und die Präsidien, wo möglich, in der Erarbeitung und Implementierung von FDM-Strategien unterstützen. Perspektivisch ist eine zentrale BUA-FDM-Koordinationsstelle geplant, die als zentrale Kontaktstelle dient, die Vernetzung FDM-relevanter Vorhaben, Initiativen und Akteur:innen am Wissenschaftsstandort Berlin stärkt und eine kontinuierliche Evaluation der Bedarfe, der FDM-Angebote und der entwickelten Strategien begleitet.

Danksagung

Die Berlin University Alliance wird gefördert vom Bundesministerium für Bildung und Forschung (BMBF) und dem Land Berlin im Rahmen der Exzellenzstrategie von Bund und Ländern. Das Förderkennzeichen des Projekts ist 501_CRDMS.

Literaturverzeichnis

Ariza de Schellenberger, Angela, Evgeny Bobrov, Kerstin Helbig, Denise Jäckel, Monika Kuberek, Lea-Sophie Orozco Prado, Elisabeth Maria Schlagberger, Sibylle Söring und Britta Steinke. 2022. *Bestands- und Bedarfserhebung zum Forschungsdatenma-*

nagement an den BUA-Einrichtungen. DOI: https://doi.org/10.5281/zenodo.7060446.

———. 2023. „Anwendung des FDM-Referenzmodells RISE-DE im Verbund“. *Bausteine Forschungsdatenmanagement,* Nr. 2. DOI: https://doi.org/10.17192/bfdm.2023.1.8551.

Hartmann, Niklas K., Boris Jacob und Nadin Weiß. 2019. *RISE-DE – Referenzmodell für Strategieprozesse im institutionellen Forschungsdatenmanagement.* Technischer Bericht. DOI: https://doi.org/10.5281/zenodo.3585556.

Jäckel, Denise, Kerstin Helbig und Carolin Odebrecht. 2022. „Desiderate zum Forschungsdatenmanagement 2013 und 2022“. *Information – Wissenschaft & Praxis* 73 (5-6): 265–276. DOI: https://doi.org/10.1515/iwp-2022-2239.

Kuberek, Monika, Elisabeth Maria Schlagberger und Britta Steinke. 2022. *Bestands- und Bedarfserhebung zum Forschungsdatenmanagement an der Technischen Universität Berlin: Ergebnisse der Online-Befragung 2021/22.* Technischer Bericht. Technische Universität Berlin. DOI: https://doi.org/10.14279/depositonce-16664.2.

Taubitz, Jan, Evgeny Bobrov und Angela Ariza De Schellenberger. 2022. *Auswertung der Bedarfserhebung zum Forschungsdatenmanagement an der Charité 2021/22 (Analysis of the survey on research data management at Charité 2021/22).* DOI: https://doi.org/10.5281/zenodo.7385548.

Ein Werkzeug zur XSD-basierten Metadatenannotation

Olaf Brandt[1], Holger Gauza[1], Jan Kaltenbach[1], Maximilian E. Müller[2], Gabriel Schneider[2], Claus Zinn[1]

[1]Universität Tübingen;
[2]Universität Konstanz

Der Umgang mit Forschungsdaten entlang des Forschungsdatenlebenszyklus erlangt immer mehr Relevanz und erfordert nicht nur eine gewissenhafte Planung und eine sichergestellte Speicherung der Forschungsdaten, sondern auch die Auszeichnung der Forschungsdaten mit geeigneten Metadaten, um deren Auffindbarkeit und Nachnutzbarkeit im Sinne der FAIR-Prinzipien umzusetzen (Wilkinson u. a. 2016). Aus diesem Grund stehen viele Dienste im Bereich des Forschungsdatenmanagements vor der Herausforderung, ihren Nutzer:innen ein Werkzeug anzubieten, mit dem sie ihre Daten mit passenden Metadatenstandards beschreiben können. Die Metadatenstandards werden dabei aber nur in den wenigsten Fällen durch die Dienstanbieter:innen selbst entwickelt, sondern durch verschiedene Konsortien, wie z.B. in den nationalen Forschungsdateninfrastrukturen (NFDI)[1], entwickelt und gepflegt. Die Dienstanbieter:innen müssen entsprechend auf Änderungen im Upstream reagieren und ihre Dienste und Werkzeuge anpassen. Gleichzeitig muss die Verwendung eines Werkzeugs zur Annotation von Metadaten für die Nutzer:innen unter Aspekten der Usability dahingehend ausgelegt sein, dass diese sich auf die einschlägigen Metadaten konzentrieren können und z.B. durch Dropdownmenüs und Hilfetexte unterstützt werden. Zusätzlich zur Usability tragen Dropdownmenüs mit hinterlegten fachspezifischen Vokabularen und Ontologien zur Qualität der Metadaten bei, da nicht nur Freitext vermieden, sondern auch semantische Verbindungen durch Übernahme von Uniform Resource Identifiern (URI) etc. ermöglicht werden. Unter Aspekten der Nachhaltigkeit ist es außerdem wünschenswert, dass ein Werkzeug nicht spezifisch für nur eine wissenschaftliche Community maßgeschneidert wird, sondern dass es potentiell durch Austausch von Metadatenschemata zur Erfassung relevanter Inforamtionen auf andere wissenschaftliche Disziplinen übertragen und angewendet und somit auch von anderen Communities genutzt werden kann. Es ergeben sich vier grundlegende Anforderungen an ein Werkzeug zur Metadatenannotation:

1. Reduzierter Betreuungsaufwand für Dienstanbieter:innen durch die einfache Integration bzw. den Austausch von Schemata. Integration von Ontologien und Voka-

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.c18088

1 Für weitere Informationen siehe https://www.nfdi.de.

bularen aus externen Quellen und eine einfache Integration in bestehende Systeme wie universitäre Idendity Provider (IDP) oder Publikationsplattformen.

2. Erhöhte Usability für die Nutzer:innen durch die Integration von Dropdownmenüs mit integrierter Autocomplete-Funktion und dynamische Reduktion der Schemakomplexität auf einschlägige Felder. Berücksichtigung von Anforderungen im Sinne der Barrierefreiheit.

3. Generischer Anspruch und Einsatzmöglichkeit durch den einfachen Austausch von Schemata zur Bereitstellung für weitere Communities. Entsprechend ein Verzicht auf Hardcoding und Nutzung von Standards.

4. Mehrwert für die Community durch die Generierung von maschienenlesbarem Output bei Erhalt semantischer Relationen sowie automatisierte Qualitätskontrolle durch automatische Validierung der Eingaben sowie Vorbereitung für eine Publikation der Forschungsdaten.

Im Folgenden wird ein Werkzeug zur XSD-basierten[2] Metadatenannotation vorgestellt, das einen generischen Ansatz bei der Bereitstellung von Formularen zur Metadatenbeschreibung verfolgt.[3] Anstatt ein spezifisches Metadatenschema zu hinterlegen, können jegliche XSD-basierten Schemata verwendet werden. Aus dieser formalen Beschreibung werden anschließend durch einen XSLT-Processor HTML-Formulare generiert, die Nutzer:innen webbasiert in einem Browser ausfüllen können. Hierdurch wird eine Trennung von Inhalt (XSD) und Form (HTML) erzeugt, die eine erhöhte Flexibilität durch den Austausch von XSD-Schemata bedeutet. Gleichzeitig lassen sich beispielsweise mehrere Schemata integrieren und nebeneinander befüllen, um so eine Metadatenbeschreibung zu erreichen, die exakt auf die Bedarfe der Nutzer:innen der Dienstanbieter:innen zugeschnitten sind. Eine automatisierte Validierung der Ergebnisse ist auf diese Weise ebenfalls möglich, wodurch Fehler direkt bei der Eingabe erkannt und korrigiert werden können. Der Aufwand bei der Implementierung von Aktualisierungen wird dadurch reduziert, dass diese sich direkt in die Formulare integrieren lassen, solange sie als XSD vorliegen. Eine Administrationsoberfläche ermöglicht Anpassungen an der Darstellung der Formulare hinsichtlich Reihenfolge und Sichtbarkeit von Elementen und den Eingabefeldern, um eine Fokussierung auf die einschlägigen Metadaten umzusetzen. Auf diese Weise werden Anpassungen in der XSD-Datei umgangen und gleichzeitig Flexibilität erzeugt. Das Annotationstool ist kompatibel mit allen IDPs, die OIDC[4] anbieten.

1 Reduzierter Betreuungsaufwand

Die Auszeichnung von Forschungsdaten mit Metadaten ist kein Selbstzweck, sondern zielt darauf ab, die Forschungsdaten im Sinne der FAIR-Prinzipien zu beschreiben, um so unter anderem Angaben für die Auffindbarkeit, die wissenschaftliche Nachnutzbarkeit und

2 XML Schema Definition, siehe `https://www.w3.org/TR/xmlschema11-1`.

3 Der verwendete XSLT-Processor basiert auf dem ursprünglichen Projekt XSD2HTML2XML von Meulendijk (2019).

4 Für weitere Informationen siehe `https://openid.net/connect`.

dauerhafte Referenzierbarkeit zu erfassen. Für die dauerhafte Referenzierbarkeit hat sich weitgehend die Vergabe von DOIs[5] durchgesetzt, woraus sich die Notwendigkeit der Erfassung von Metadaten nach dem DataCite-Schema ergibt.[6] Neben diesen deskriptiven Metadaten werden wissenschaftliche Metadaten benötigt, um Such- und Filterfunktion zu ermöglichen und Wissenschaftler:innen einen schnellen Überblick über die Forschungsdaten zu ermöglichen. Hierfür notwendige Schemata sind Anpassungen unterworfen, die entsprechend in einem Annotationswerkzeug nachgezogen werden müssen. Die Integration neuer Versionen oder gänzlich neuer Schemata erfordert bei dem hier vorgestelltem Werkzeug lediglich den Austausch oder die Bereitstellung einer neuen XSD, wodurch der Betreuungsaufwand stark reduziert wird. Im Rahmen des SDC BioDATEN wurde das DataCite-Schema 4.4 integriert und um wissenschaftliche Metadaten aus dem BioDATEN-Minimalschema ergänzt.[7] Diese Kombination erfüllt die projektinternen Anforderung hinsichtlich Publikation und Suchbarkeit von Forschungsdaten bei gleichzeitiger Anwendbarkeit auf mehrere Omics-Disziplinen. Um die Nutzer:innen beim Umgang mit dem Werkzeug zu unterstützen und die Qualität ihrer Eingaben zu erhöhen, wurde die Einbindung von Ontologien und Vokabularen über Dropdownmenüs mit Autocomplete-Funktion in das Werkzeug integriert. Hierfür wird die API von Bioportal[8] eingesetzt, um die Anforderungen an eine eigene Datenaufbereitung der hinterlegten Ontologien zu minimieren.[9] Über eine Adminstrationsoberfläche lässt sich konfigurieren, welche Metadatenfelder aus welchen Vokabularen befüllt werden sollen, siehe Abbildung 1. Hierdurch und durch die Integration der Bioportal API wird der Betreuungsaufwand minimiert und gleichzeitig die Datenqualität erhöht.

2 Mehrwert

Die Annotation von Metadaten ist ein wichtiger Baustein des Forschungsdatenlebenszyklus und bildet die Grundlage für die Auffindbarkeit und Nachnutzbarkeit der Forschungsdaten.[10] Entsprechend wurde bei der Entwicklung des Annotationswerkzeugs darauf Wert gelegt, dass die erfassten Metadaten eine Grundlage für anschließende Prozesse sind. Dies geschieht im Kontext von BioDATEN durch die Anbindung an eine Publikationsplattform auf Basis von InvenioRDM[11]. Die erfassten deskriptiven Metadaten erfüllen die Anforderungen des DataCite-Schemas 4.4[12] und bilden die Grundlage für die DOI-Registrierung. Für die Nutzer:innen bedeutet dies einen Verzicht auf die nochmalige Eingabe ihrer Daten. Während des Annotationsprozesses werden die Angaben validiert und wo immer möglich durch ein Vokabular bzw. eine Ontologie supplementiert. Somit kann durch den Verzicht

5 Digital Object Identifier, siehe `https://www.doi.org`.
6 Für die aktuelle Version des Schemas siehe `https://schema.datacite.org`.
7 Für weitere Informationen zum BioDATEN Minimalschema siehe `https://github.com/ubtue/BioDATEN-Minimalschema`.
8 Für weitere Informationen siehe `https://bioportal.bioontology.org`.
9 Einen alternativen Ansatz bietet der Dienst Semlookup der ZB MED, siehe hierzu Madan u. a. (2018).
10 Für weitere Informationen siehe `https://forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus`.
11 `https://inveniosoftware.org/products/rdm`
12 `https://datacite.org`

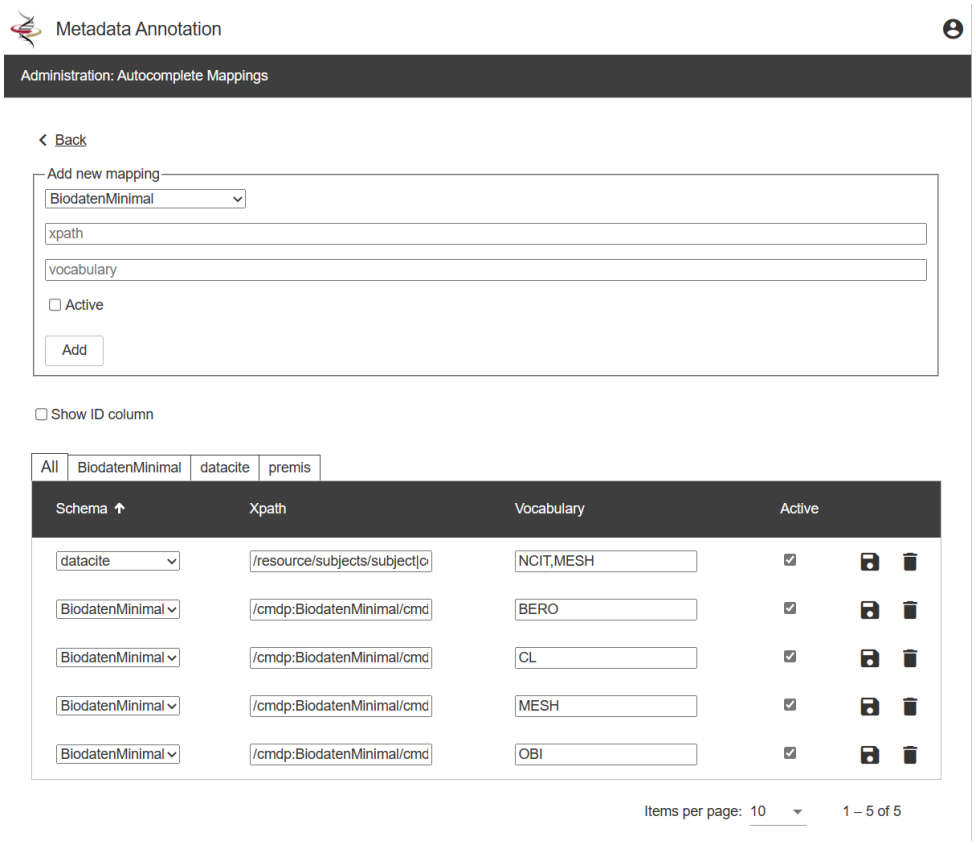


Abbildung 1: Screenshot der Administrationsoberfläche. Auswahl der Metadatenfelder mit zugeordnetem Vokabular.

auf Freitext die Qualität der Metadaten erhöht und die Erfassung von semantischen Relationen generiert werden.

3 Erhöhte Usability

Die Auszeichnung von Forschungsdaten bringt einen gewissen Mehraufwand für die Forscher:innen mit. Um diesen zu reduzieren, wurden Dropdownmenüs integriert und Vokabulare hinterlegt, siehe Abbildung 2. Eine weitere Maßnahme zur Verbesserung der Usability liegt in der fokussierten Integration einschlägiger Metadatenfelder unter Berücksichtigung des Anspruchs auf generische Einsetzbarkeit. Das Annotationswerkzeug unterstützt deshalb generell und speziell für das BioDATEN Minimalschema die Verwendung von konditional-obligatorischen Metadatenfeldern und entsprechende Abhängigkeiten, wie sie unter anderem durch das CMDI-Framework generiert und in XML abgebildet werden können (siehe Brandt u. a. 2021). Hierdurch werden Metadatenfelder nur dann angezeigt, sofern diese aufgrund definierter Abhängigkeiten von getätigten Eingaben relevant sind. Zusätzlich stand bei der Entwicklung des Annotationswerkzeugs die Berücksichtigung und Umsetzung digitaler Barrierefreiheit im Fokus.

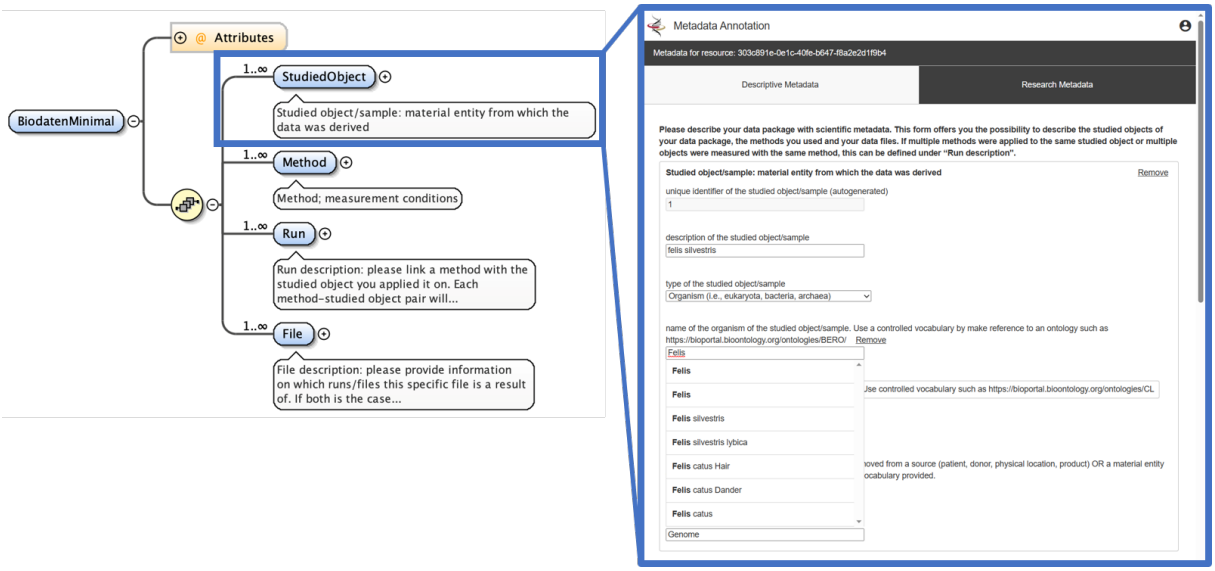


Abbildung 2: Screenshot der Annotationsoberfläche. Auswahl des untersuchten Organismus mit Dropdownmenü in der Beschreibung des „Studied Object“.

4 Generischer Anspruch

Das Annotationswerkzeug wurde im Rahmen des SDC BioDATEN mit Fokus auf Bioinformatik und mehrerer der sogenannten Omics-Disziplinen entwickelt.[13] Dennoch wurde bei der Entwicklung großer Wert darauf gelegt, das Annotationswerkzeug durch eine einfache Austauschbarkeit von Metadatenschemata breit und generisch anbieten zu können. Der Verzicht auf Hardcoding erlaubt in Kombination mit der Trennung von Funktion und Inhalt eine einfache Anpassung an andere Communities. Entsprechend kann das Annotationswerkzeug beispielsweise auch im Rahmen von bwHPC eingesetzt werden, wo ebenfalls der Bedarf nach einer generischen und anpassungsfähigen Lösung besteht. Die Grundlage für den Aufbau der webbasierten Annotationsoberfläche bilden vorhandene XSDs. XSDs haben den Vorteil, Metadaten-Standards abbilden und gleichzeitig die Validierung gegen diese Standards ermöglichen zu können.

5 Open Source Software

Das Tool wurde als Open-Source-Software entwickelt. Es steht unter der AGPL-3-Lizenz[14] zur Verfügung. Für das Frontend wurde das auf TypeScript basierte Webapplikationsframework Angular[15] verwendet. Der Backend-Service des Annotationstools wurde mit

13 Für weitere Informationen siehe `https://portal.biodaten.info`.
14 `https://www.gnu.org/licenses/agpl-3.0.de.html`
15 `https://angular.io`

dem Spring Boot Framework[16] erstellt. Die Einzelkomponenten liegen öffentlich zugänglich in GitHub-Repositorien: Backend[17], Frontend[18] und XSLT-Prozessor[19].

6 Fazit

Die Annotation von Forschungsdaten mit Metadaten erzeugt Aufwand bei den Forscher:innen und den Dienstanbieter:innen gleichermaßen. Erstere müssen Daten liefern und letztere müssen entsprechende Dienste bereitstellen. Das Ziel bei der Entwicklung des Tools liegt in der Entlastung beider Seiten gleichermaßen und der Schaffung von Mehrwert durch eine Annotation. Die Kombination aus XSD Input, menügeführter Administration, Integration von kontrollierten Vokabularen, Dropdownmenüs und der automatischen Validierung von Input resultiert in einem Werkzeug, das die eingangs genannten Anforderungen erfüllt. Der generische Ansatz bei der Erstellung des Werkzeugs und die konsequente Umsetzung von Open-Source sowie eine Nutzbarkeit ohne tiefes Expertenwissen verspricht generische Einsatzmöglichkeiten.

Literaturverzeichnis

Brandt, Olaf, Holger Gauza, Steve Kaminski, Mario Trojan, Thorsten Trippel und Johannes Werner. 2021. „Extending the CMDI Universe“. In *Linköping Electronic Conference Proceedings,* herausgegeben von Costanza Navarretta und Maria Eskevich, Bd. 180. DOI: https://doi.org/10.3384/ecp1806.

Madan, Sumit, Maksims Fiosins, Stefan Bonn und Juliane Fluck. 2018. „A Semantic Data Integration Methodology for Translational Neurodegenerative Disease Research“. https://doi.org/10.6084/m9.figshare.7339244.v1. See also https://semanticlookup.zbmed.de/, *Semantic Web Applications and Tools for Healthcare and Life Sciences.*

Meulendijk, Michiel. 2019. „XSD2HTML2XML“. Besucht am 6. September 2023. https://github.com/MichielCM/xsd2html2xml.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

16 https://spring.io
17 https://github.com/ubtue/BioDATEN-Metadata-Annotation-Backend
18 https://github.com/ubtue/BioDATEN-Metadata-Annotation-Tool
19 https://github.com/ubtue/BioDATEN-Metadata-XSLT-Processor

Standardized Metadata Collection to Reinforce Collaboration in Collaborative Research Centers

Manuel Watter, Laura Kahle, Birger Brunswiek, Urs A. Fichtner, Michelle Pfaffenlehner, Frank Werner, Denis Gebele, Harald Binder, Jochen Knaus

Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg

The availability of good metadata in referenced terminologies is a prerequisite for data interoperability and the associated reliable retrieval. This interoperability of data through their documentation is considered one of the more complex problems in the creation of FAIR datasets (Jacobsen et al. 2020; Guizzardi 2020).

Standardization of data collection depends not only on the field of research, but also on the object of research: while excellent standards such as SNOMED CT[1] have been established in clinical trials and medical routine care, this is usually not the case in basic biomedical science using cell cultures or animal models. This is also reflected in the organization of large-scale research projects such as Collaborative Research Centers: in addition to highly standardized data types, such as for genetic analyses, there is also long-tail data with sometimes individual signatures. In both cases, however, there is a need for a standardized description of the experimental set-up.

As any documentation of datasets is labor-intensive, it is often only of medium-term benefit to the researcher. Therefore, the additional workload is more likely to be accepted if there are clear guidelines, e.g., from data repositories. If data documentation is to be incentivized instead of forced, a reduction in the effort required to collect the data is certainly a prerequisite.

In our bottom-up approach, scientists are empowered to define minimal datasets that are iteratively aligned with existing terminologies and standards by RDM managers.

1 Data documentation

General description standards such as DataCite (DataCite Metadata Working Group 2021) help to document datasets at an administrative level. Due to the lack of structured

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.c18089

1 https://www.snomed.org

information from specific domains, this information is of limited use for further assessing the usability of a dataset.

We hypothesize that a "collage" of the useful parts of different standards and controlled vocabularies can keep the effort of collection low and thus increase adoption, without reducing the interoperability of the data sets for machine analysis too much.

2 Schematic integration of terminologies and ontologies

Transferring existing terminology can be difficult, even with a search function. Presenting exhaustive lists in a narrow use case can feel overwhelming and might waste a user's precious time. A complete hierarchy of possible tissue sources is not relevant to a cardiologist, for example.

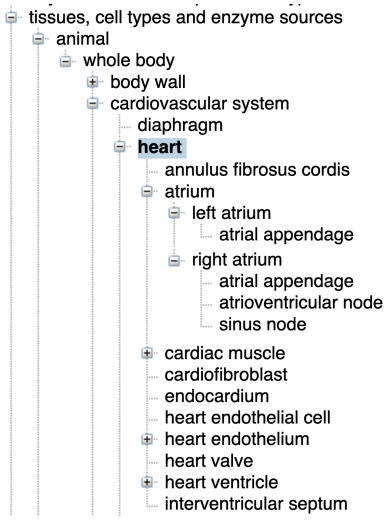


Figure 1: Example of a hierarchical vocabulary (Brenda Tissue Ontology BTO).

Accordingly, even when using terminologies, we propose a selection that is geared to the particular input case, covering it completely from a technical point of view, but reducing it to the minimally required areas (see Figure 1 for an example). Three strategies are potentially possible (Figure 2). Reducing the range of possible values is optimal for speeding up input without compromising the precision of the description and thus interoperability.

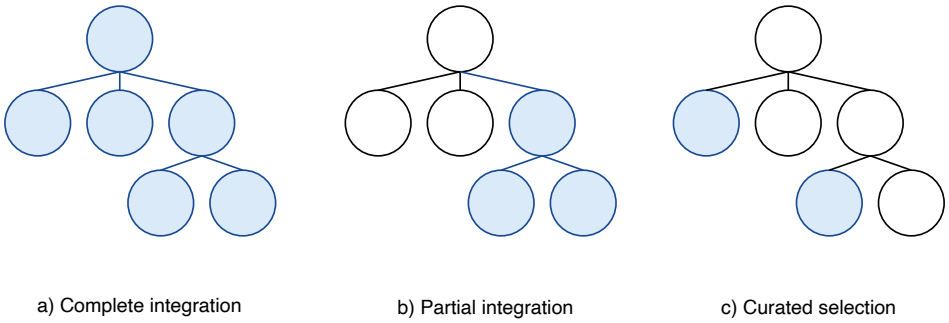


Figure 2: Options to integrate existing vocabularies such as taxonomies and ontologies.

While in Figure 2 a) in the entire terminology is browsable, b) only transfers a substructure, provided that suitable hierarchy levels or separation criteria are available. If only a few of the values ever occur in lab reality, a manually curated list (c)) is advantageous and can re-combine nodes scattered throughout the terminology.

3 Tools for scientists and data stewards

To enable scientists or data stewards to maintain data description structures themselves, a Microsoft Excel schema is provided (Figure 3). The familiarity with this tool lowers the barrier of entry. It is also particularly useful for locally specified lists that are typically already maintained in laboratories (e.g., antibodies, mouse lines), which can then be more easily compared to (potentially) existing standards during a subsequent revision step.

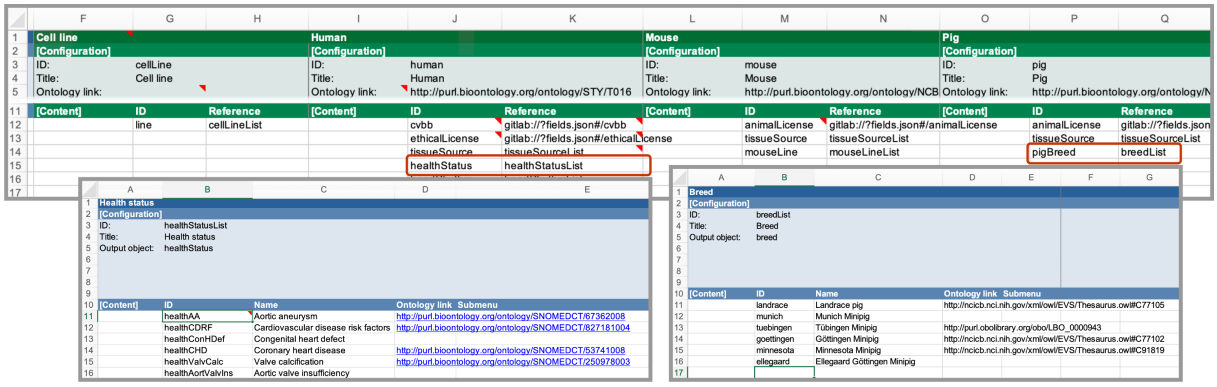


Figure 3: Example of defining structures and relationships of documentation entities using Excel.

4 Technical implementation

Our data documentation forms are embedded in our research data management system *fredato*, which is a thin wrapper over on-premises GitLab[2] and Nextcloud[3] instances. It stores the metadata directly in Git[4] repositories without a database, so they are always kept in sync with the research data and do not require explicit export processes, meaning no lock-in to our software and full user control.

The form definitions are also treated as data and exist as distributed JSON schema[5] definitions after being converted and merged from various sources (external vocabulary imports, local Excel lists, manual input) and displayed in the web frontend using the VJSF library[6] (see Figure 4). Once stored in their respective repositories, the metadata is automatically indexed in OpenSearch using GitLab Continuous Integration.

An example of processing a single aspect and the resulting internal representation is shown in Figure 5, an example of form logic defined in Excel is shown in Figure 6.

2 `https://about.gitlab.com`
3 `https://nextcloud.com`
4 `https://git-scm.com`
5 `https://json-schema.org`
6 `https://github.com/koumoul-dev/vuetify-jsonschema-form`; *Last accessed on May 5th, 2023.*

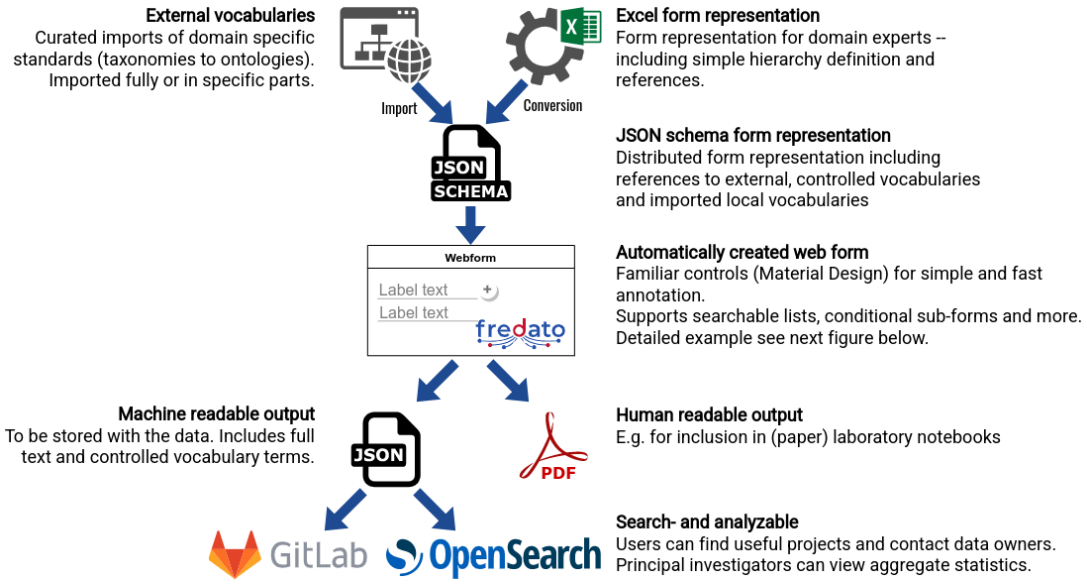


Figure 4: Workflow of form creation and processing using *fredato*.

5 Everyday use

Currently, for example, a template is available for the documentation of data sets in basic cardiological research, which was introduced as a recommendation in the Collaborative Research Centre 1425. In everyday life, there are two different procedures: Researchers use this to document the end of an experimental series and thus the generation of the raw data set in the laboratory. Alternatively, researchers, especially those who still work without an electronic lab book, use the metadata editor on a daily basis to document experimental progress. A copy function is available for this purpose, which only requires the new parts to be changed. A data set documentation is thus created from the compilation of the metadata of the individual laboratory days.

6 Discussion

When developing data documentation schemes in a bottom-up manner, it is advisable to include support from the research data management side in addition to the actual users, i.e. the subject experts. Both sides can benefit from each other, as knowledge of the need for reporting guidelines and data standards often needs to be built up by the subject experts. Ideally, candidates for local data stewards will emerge from this iterative process, greatly accelerating future collaborative efforts.

Our solution improves metadata interoperability, but does not produce fully machine-understandable grammars (Jacobsen et al. 2020). However, simply referencing published terminologies is usually not enough context for software agents to understand naming. The context can be re-created later in the export process by translating terminology look-ups into grammars.

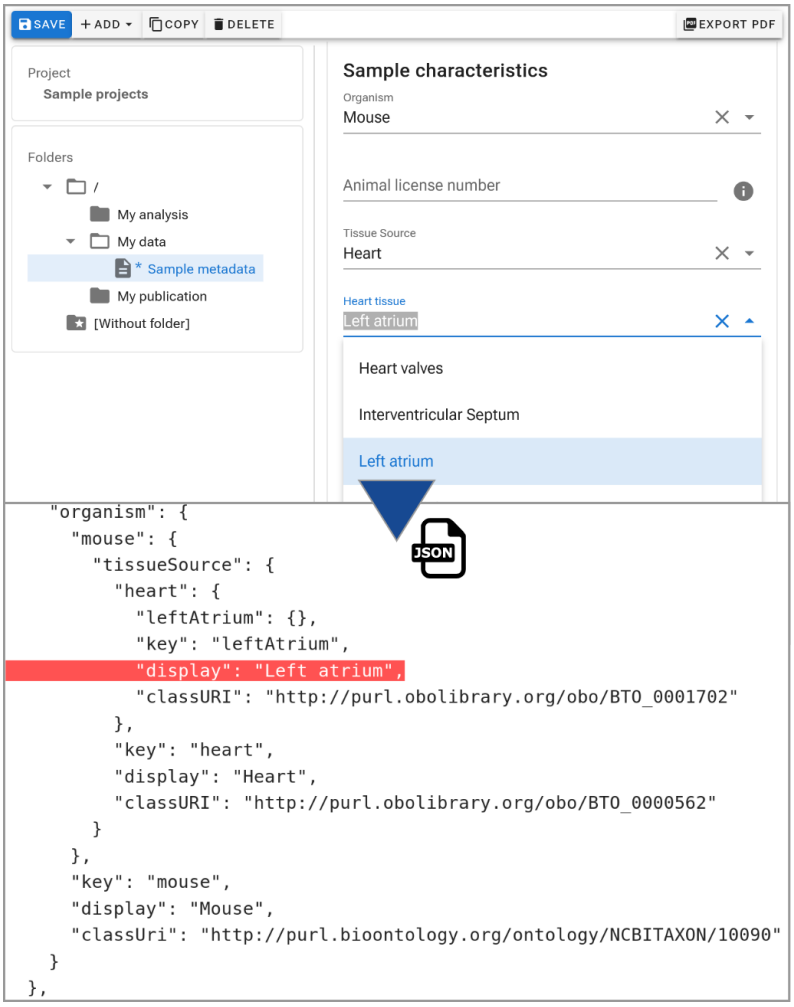


Figure 5: Web form and resulting JSON metadata file.

7 Conclusion

The burden of data documentation can be selectively reduced, without loss of technical interoperability, by presenting only the information from the terminologies that is necessary for a particular group based on standardized controlled vocabularies.

8 Author contributions

Manuel Watter developed the metadata editing and importing software and contributed to the original draft. Birger Brunswiek developed the metadata search software and added metadata indexing. Urs Fichtner and Michelle Pfaffenlehner contributed with writing - reviewing & editing. Denis Gebele, Laura Kahle and Frank Werner contributed to software testing. Harald Binder contributed with funding acquisition and monitoring. Jochen Knaus contributed to the conception, writing of the original draft and supervision.

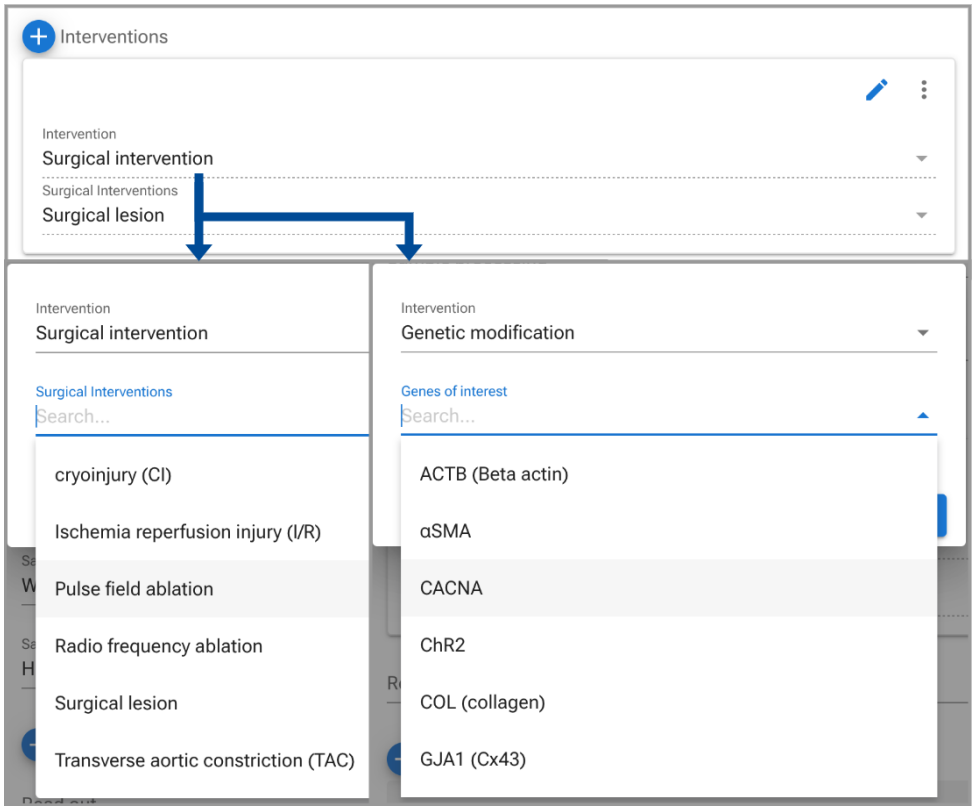


Figure 6: Example of a repeating field with conditional subfields.

Acknowledgements

This work is funded by the Collaborative Research Centers 1425 (project number #42268-1845), 1453 NephGen (#431984000), 1479 OncoEscape (#441891347) and TR-CRC 359 PILOT (#491676693), all funded by the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation).

References

DataCite Metadata Working Group. 2021. "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4". DOI: `https://doi.org/10.14454/3W3Z-SA82`. `https://schema.datacite.org/meta/kernel-4.4/`.

Guizzardi, Giancarlo. 2020. "Ontology, Ontologies and the 'I' of FAIR". *Data Intelligence* 2 (1-2): 181–191. DOI: `https://doi.org/10.1162/dint_a_00040`.

Jacobsen, Annika, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, et al. 2020. "FAIR Principles: Interpretations

and Implementation Considerations". *Data Intelligence* 2 (1-2): 10–29. DOI: https://doi.org/10.1162/dint_r_00024.

Bringing FAIR Bioimage Data Management into Practice: the Information Infrastructure for BioImage Data (I3D:bio) Project – bottom-up Community Support for Microscopy Data Sharing and Preservation.

Christian Schmidt[1], Michele Bortolomeazzi[1], Tom Boissonnet[2], Julia Dohle[3], Tobias Wernet[4], Janina Hanne[5], Roland Nitschke[4], Susanne Kunis[3], Karen Bernhardt[3], Stefanie Weidtkamp-Peters[2], Elisa Ferrando-May[1]

[1]Deutsches Krebsforschungszentrum, Heidelberg;
[2]Heinrich-Heine-Universität Düsseldorf;
[3]Universität Osnabrückk;
[4]Albert-Ludwigs-Universität Freiburg;
[5]German BioImaging – Gesellschaft für Mikroskopie und Bildanalyse e.V.

The practical adoption of the widely acknowledged FAIR (Findable, Accessible, Interoperable, Reusable) principles for research data and data stewardship requires researchers and infrastructure providers to work hand-in-hand. Microscopy-driven and image-analysis-driven research projects are particularly challenging concerning FAIR data handling, data sharing, and data preservation. The multitude of imaging modalities, the large number of vendor-specific proprietary file formats, the enormous sizes of high-dimensional bioimaging data files, and the vast array of software products for processing and analysis have led to the stigma of bioimaging data being cumbersome to handle. Community-driven solutions for image data management systems, file format translation libraries, and open-source software exist in the field of bioimaging. However, many solutions are only applicable in practice if researchers have access to institutional resources for hardware and technical support for implementation, maintenance, and training. To help overcome this hurdle to the practical implementation of FAIR data management in bioimaging, the Information Infrastructure for BioImage Data project (I3D:bio) started in 2022 as a DFG-funded, collaborative effort to provide bioimaging research data management (RDM) support at universities and research institutions. The work program is based on the exchange of experience among members of German BioImaging – Society for Microscopy and Image Analysis (GerBI-GMB), the Research Data Management for Microscopy (RDM4mic) group, and the NFDI4BIOIMAGE consortium participating in the National Research

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.c18090

Data Infrastructure (NFDI). The I3D:bio project focuses on data handling leveraging the open-source software OME Remote Objects (OMERO), offering direct and on-premise support for the implementation of bioimage data management, including storage concepts for long-term archiving, data sharing capabilities, and metadata enrichment. We are developing tailored training material and workshops. The I3D:bio project, moreover, aims to create a public resource for reusable metrology data in microscopy. Here we present the project progress, our support offers, and the lessons learned from use cases at the co-applicant sites and partner sites.

1 Introduction

Research across various disciplines relies on microscopy as a collection of techniques allowing insight into living or non-living matter with high spatial and temporal resolution. Data generated by modern microscopy is often characterized by large file sizes and considerable complexity regarding both the performed experiment and the file structure in which the acquired data is stored (Ouyang and Zimmer 2017). The lack of a standard file format in microscopy impedes reusability and accessibility for third-party users, while non-standardized or even lacking metadata about the technical setup and the experimental protocol are obstacles to the findability and interoperability of microscopy data. At present, the Open Microscopy Consortium's OME.TIFF format and the Bio-Formats translation library offer a partial solution in practice (Linkert et al. 2010). However, these approaches to data handling cannot scale with the increasing variability of imaging modalities and proprietary file formats by microscope vendors (OME 2019). Moreover, classical binary file formats are not well suited for modern microscopy and image analysis workflows that include remote storage, frequent access, and processing (Moore et al. 2021).

Large N-dimensional arrays within the image files require high network transfer rates and sufficiently large computer memory (RAM). The heterogeneity of microscopy modalities and file types is particularly challenging for individual researchers. While specialized staff at imaging core facilities train researchers for the proper use of microscope systems or even assist with the preparation and acquisition of bioimaging data, data management after the acquisition is often a sole user responsibility. Dedicated funding to support RDM for microscopy is often lacking in imaging core facilities and individual research laboratories. As a result, bioimaging data today is often hard to preserve and share in compliance with the FAIR principles (Wilkinson et al. 2016). To overcome these hurdles toward FAIRification of microscopy data, several members of the German bioimaging community, firmly grounded in the imaging core facility network German BioImaging – Society for Microscopy and Image Analysis (GerBI-GMB), initiated the "Information Infrastructure for BioImage Data" (I3D:bio), as a bottom-up project promoting the direct benefit of community-tested image data handling software as well as metadata annotation tools and guidelines.[1]

1 `https://www.i3dbio.de`

2 Resources

I3D:bio was proposed by imaging core facility managers as a collaboration of four German institutions. Regular exchange in the open community group "Research Data Management for Microscopy" (RDM4mic) identified common issues and pitfalls in microscopy data handling[2]. Several core facilities have implemented centralized instances of the image data management software OME Remote Objects (OMERO) between 2017 and 2022 (Burel et al. 2015; Zobel, Weischer, and Wendt 2022; Kunis, Bernhardt, and Hensel 2023).

The 2021 NFDI4BIOIMAGE community survey furthermore confirmed that OMERO was the so far best-known and most widely used image data management software in the (German) bioimaging community, evincing OMERO as common ground for starting a bioimage RDM initiative (Schmidt et al. 2022). While OMERO greatly facilitates structuring, sharing, and annotating imaging data as compared with classical file folder hierarchies, running OMERO as a facility requires investing in personnel capacity and IT resources that are not readily available for many core facilities in Germany. Therefore, a goal of I3D:bio is to capacitate core facilities for implementing and managing OMERO instances by offering support and guidance for this process. Additionally, core facility users must be trained to become proficient in the use of OMERO, which is based on object storage instead of file hierarchies. Questions of data ownership, storage security, or access to OMERO-hosted data for image processing and analysis were observed as impediments to the practical adoption by users. I3D:bio intends to provide best-practice, peer-reviewed training material as a resource for core-facility-centric user training in addition to the OMERO guides[3]. Additionally, annotation tools and guidance to comply with community-driven metadata enrichment standards are being tested in use cases and at the applicant sites. These efforts are orchestrated to align with the activities of the international bioimaging community. To reach these goals, the three work packages of I3D:bio focus, first, on the deployment of OMERO instances according to best practices as well as building a public OMERO-based database for microscopy metrology data, second, the identification of the suitable technical infrastructure requirements for image data storage in OMERO and metadata annotation software, and, third, the coordination with (international) partners, communication and training (Figure 1). The project duration is three years; a prolongation is possible

3 Support

I3D:bio lays groundwork and complements the goals of the collaborating NFDI4BIOIMAGE consortium[4], which is part of the German National Research Data Infrastructure (NFDI[5]). Several NFDI consortia from the life sciences support the I3D:bio project and have stated their intention to implement I3D:bio guidelines for bioimaging RDM work-

2 `https://german-bioimaging.github.io/RDM4mic.github.io`
3 `https://omero-guides.readthedocs.io/en/latest`
4 `https://nfdi4bioimage.de`
5 `https://nfdi.de`

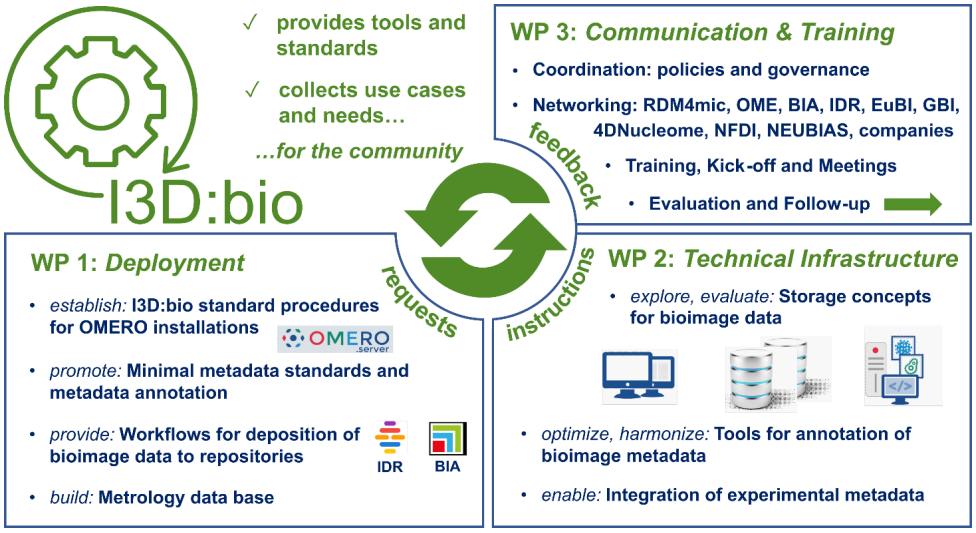


Figure 1: I3D:bio work packages. Taken from the I3D:bio project proposal.

flows in their field of research. To test and refine I3D:bio's approach to implementing OMERO at research institutions and universities, two naïve sites have been selected for the de novo installation of OMERO. The applicant institution German Cancer Research Center supports the project with in-kind contributions for the institution-wide adoption of OMERO in the field of cancer research. The Technical University of Dresden's Medical Faculty serves as an external naïve site for testing and refining I3D:bio-guided image data management workflows. Moreover, I3D:bio works in close collaboration with members of the RDM4mic group, who provide input for the project and constitute the connection to the broader target audience whose members may become integrated into the project as use cases. At the international level, I3D:bio collaborates with community partners like Euro-BioImaging[6], BioImaging North America[7], the Quality Assessment and Reproducibility of Instruments and Images in Light Microscopy (QUAREP-LiMi) group[8], and others. For example, two joint workshops for metadata annotation for bioimaging were offered together with BINA at the European Light Microscopy Initiative (ELMI) conference in 2023[9].

4 Conclusions

Within its first year, I3D:bio has collected community input via RDM4mic and several use case partners in and beyond Germany. The website[10] was established as a central knowledge resource on bioimaging RDM, including an overview of different topics and solutions, links to training resources, and an introduction to the complex topic of bioimaging

6 https://www.eurobioimaging.eu
7 https://www.bioimagingnorthamerica.org
8 https://quarep.org
9 https://elmi2023.eu
10 https://www.i3dbio.de

metadata. As a live assessment of FAIR bioimaging data collection and annotation, the I3D:bio team has organized the management of microscopy data produced during practical workshops at the Trends in Microscopy 2023 Conference and invited participants and workshop providers to annotate imaging data based on the Recommended Metadata for Biological Images (REMBI) guidelines (Sarkans et al. 2021) for OMERO-based bioimage data handling and metadata annotation.

The I3D:bio website offers a Help Desk, and community stakeholders are invited to contact I3D:bio via the website or via the community forum image.sc to collaborate as a partner or use case.

Acknowledgements

We thank the I3D:bio project partners, in particular, J. Moore and T. Zobel, and the RDM4mic group for input and collaboration. German BioImaging is acknowledged for its contribution of resources and community integration. This work is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation), project number 462231789.

References

Burel, Jean-Marie, Sébastien Besson, Colin Blackburn, Mark Carroll, Richard K. Ferguson, Helen Flynn, Kenneth Gillen, et al. 2015. "Publishing and sharing multi-dimensional image data with OMERO". *Mammalian Genome* 26 (9-10): 441–447. DOI: https://doi.org/10.1007/s00335-015-9587-6.

Kunis, Susanne, Karen Bernhardt, and Michael Hensel. 2023. "Setting up a data management infrastructure for bioimaging". *Biological Chemistry* 404 (5): 433–439. DOI: https://doi.org/10.1515/hsz-2022-0304.

Linkert, Melissa, Curtis T. Rueden, Chris Allan, Jean-Marie Burel, Will Moore, Andrew Patterson, Brian Loranger, et al. 2010. "Metadata matters: access to image data in the real world". *Journal of Cell Biology* 189 (5): 777–782. DOI: https://doi.org/10.1083/jcb.201004104.

Moore, Josh, Chris Allan, Sébastien Besson, Jean-Marie Burel, Erin Diel, David Gault, Kevin Kozlowski, et al. 2021. "OME-NGFF: a next-generation file format for expanding bioimaging data-access strategies". *Nature Methods* 18 (12): 1496–1498. DOI: https://doi.org/10.1038/s41592-021-01326-w.

OME. 2019. "OME's position regarding file formats (Blog post)". Visited on May 9, 2023. https://www.openmicroscopy.org/2019/06/25/formats.html.

Ouyang, Wei, and Christophe Zimmer. 2017. "The imaging tsunami: Computational opportunities and challenges". *Current Opinion in Systems Biology* 4:105–113. DOI: https://doi.org/10.1016/j.coisb.2017.07.011.

Sarkans, Ugis, Wah Chiu, Lucy Collinson, Michele C. Darrow, Jan Ellenberg, David Grunwald, Jean-Karim Hériché, et al. 2021. "REMBI: Recommended Metadata for Biological Images – enabling reuse of microscopy data in biology". *Nature Methods* 18 (12): 1418–1422. DOI: https://doi.org/10.1038/s41592-021-01166-8.

Schmidt, Christian, Janina Hanne, Josh Moore, Christian Meesters, Elisa Ferrando-May, and Stefanie Weidtkamp-Peters and. 2022. "Research data management for bioimaging: the 2021 NFDI4BIOIMAGE community survey". *F1000Research* 11:638. DOI: https://doi.org/10.12688/f1000research.121714.2.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Zobel, Thomas, Sarah Weischer, and Jens Wendt. 2022. *OMERO for microscopy research data management - A use case example from the Münster Imaging Network.* Technical report. Wiley Analytical Science. https://analyticalscience.wiley.com/do/10.1002/was.0004000267.

Implementation of an InfraStructure for dAta-BasEd Learning in environmental sciences (ISABEL)

Marcus Strobl[1], Elnaz Azmi[1], Balazs Bischof[2], Alexander Dolich[2], Sibylle K. Hassler[2,3], Mirko Mälicke[2], Ashish Manoj Jaseetha[2], Jörg Meyer[1], Achim Streit[1], Erwin Zehe[2]

[1]Steinbuch Centre for Computing, Karlsruhe Institute of Technology;
[2]Institute of Water and River Basin Management, Chair of Hydrology, Karlsruhe Institute of Technology;
[3]Institute of Meteorology and Climate Research, Atmospheric Trace Gases and Remote Sensing, Karlsruhe Institute of Technology

The quantity and diversity of digital environmental data are growing, but they are often inaccessible due to the lack of metadata, inconsistent formats, and local storage of data. ISABEL aims to solve this problem by advancing the V-FOR-WaTer virtual research environment (VRE), which provides a user-friendly web portal for scientists to access and share data from various sources. The portal includes tools for data processing, scaling, and complex analysis, with contributions from both developers and users. Shareable workflows ensure reproducible analysis, to advance research in hydrology and environmental sciences.

1 Introduction

Observational data serve as a fundamental building block for developing a deeper comprehension of ecological systems, either through data-driven approaches or by comparing the data with model predictions. Nevertheless, a significant portion of this data can be challenging to access and often lacks adequate metadata descriptions. Consequently, the data requires significant effort to be useful for science. Accessing, preparing and (pre)processing of this data can be incredibly time-consuming, particularly when attempting to combine datasets from various sources. In the end, the results are often not reproducible (Hutton et al. 2016; Stagge et al. 2019). The ISABEL project aims to improve the situation by providing findable, accessible, interoperable and re-usable (FAIR; Wilkinson et al. 2016) hydrological data and tools to its users through a single entry-point: the user-friendly V-FOR-WaTer web portal (Figure 1).

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.c18091

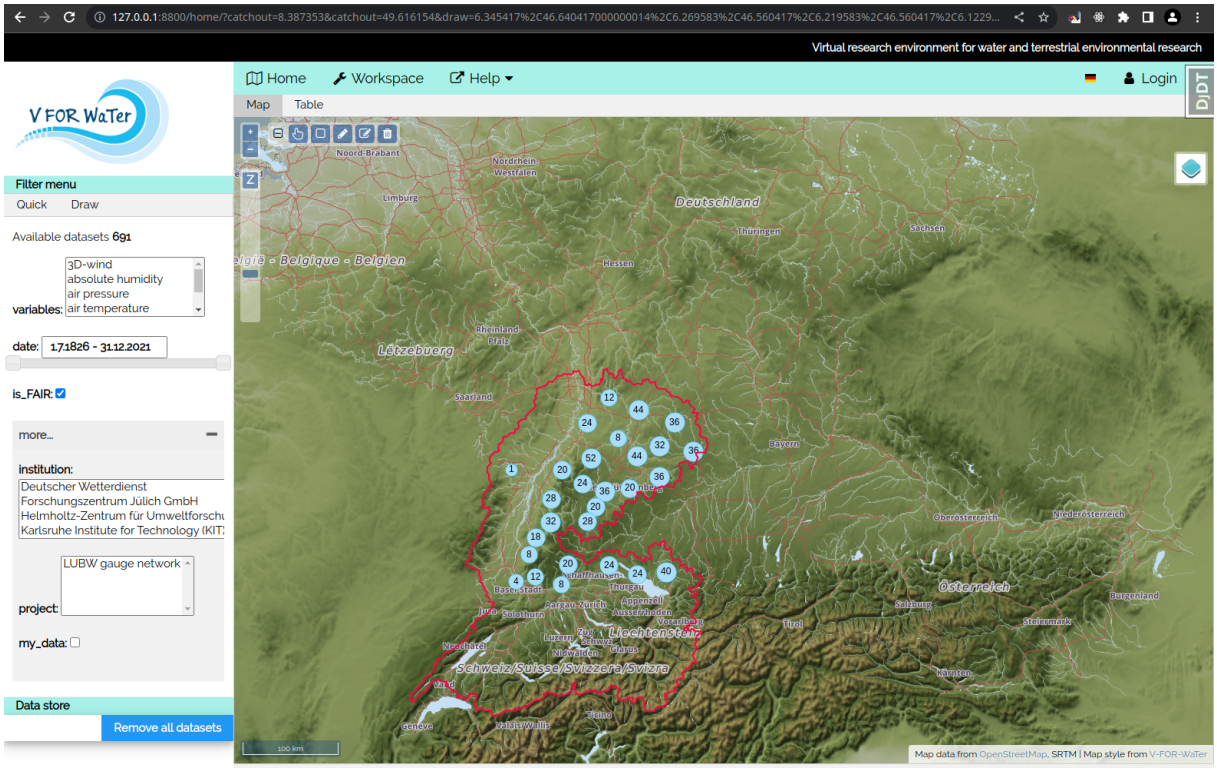


Figure 1: Screenshot of the V-FOR-WaTer web portal. Shown is the filter menu on the left, and data filtered within the upper Rhine catchment on a map.

2 The V-FOR-WaTer portal

Starting as part of the E-Science initiative of the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK, Ministry for Science, Research, and Arts Baden-Württemberg), V-FOR-WaTer has been developed to foster access and management of diverse hydrometeorological data and provide tools for preprocessing, standard hydrological procedures, and more sophisticated analyses (Strobl et al. 2022). Within the DFG-funded project ISABEL, we further develop the virtual research environment to (i) considerably expand its scientific scope, the toolbox and its user-friendliness, (ii) broaden the spectrum of hosted data to include data from state offices, complex data structures and important remote sensing products and (iii) provide access to data and tools in a modern, secure, and responsive web portal with GIS functions and a drag and drop functionality to connect tools and data for building workflows.

The portal already incorporates data from a variety of sources, including state offices and university projects. New data is gradually being added to the portal by the ISABEL team, and the data schema is continuously extended to accommodate new data types. Furthermore, an interface for open data repositories is being developed to make the most important datasets accessible in the Virtual Research Environment (VRE). This way, the web portal becomes a comprehensive resource for accessing hydrological data. The portal offers various features that facilitate the sharing of data in a metadata scheme. Users can

share their data in different common file formats, facilitating import and export of new data. The metadata can also be shared following ISO 19115 in a standardised way. In the current version, data is provided in CSV and XML formats. However, we are also preparing to support additional file formats for exportation, such as Shapefile, NetCDF, and JSON. Access management is implemented to protect critical data and maintain the ownership of unpublished data, ensuring that only authorized users can access it. In the final version, the portal will have interfaces with existing data repositories, allowing scientists to publish their data directly from the portal. To meet standards for data publication, we maintain a close collaboration with the GFZ Data Services repository to work on interfaces, both for accessing their published data and for enabling publication of data with a Digital Object Identifier (DOI) from V-FOR-WaTer in their repository. In the productive version, these features will facilitate the secure and convenient sharing of data with other members of the scientific community.

Data processing within the portal is facilitated through integrated tools for pre-processing and scaling of the data. Currently, the tools are implemented by the portal developers; in future versions, users will have the possibility to include their own tools as well, making the toolbox development a collaborative community effort. The toolbox already contains processes for geostatistics (Mälicke 2021), variogram analyses and kriging tools, hydrostatistics and visualization. More tools such as uncertainty package, GIS tools, data scaling, evaporation and Eddy covariance tools are being added within the scope of the ISABEL project. Workflows can be composed and customized easily via drag and drop. Users can also store their workflows, and the upcoming feature to share workflows will ensure reproducible data analysis. These features render the portal's data processing efficient and user-friendly.

In 2022 and 2023, we started two projects as case studies to actively use the portal to access data and contribute to the toolbox for testing. The aim is to ensure that V-FOR-WaTer covers a wide range of hydrological research and practical applications. These projects require integration of a variety of functions, data types and several user-developed packages. Their scientific focus is (i) hydrological model evaluation and associated uncertainty estimation (Manoj Jaseetha et al. 2023) and (ii) evaluating the potential of machine learning to support hydrological modelling. Given the challenging nature of these case studies, the processing and analysis workflows should be adaptable to a wide range of use cases.

3 Technical aspects

The design of the V-FOR-WaTer web portal follows well-known Geographic Information Systems (GIS) such as ArcGIS or QGIS, as the handling of such systems is intuitive among environmental scientists. The portal includes map-based operations, sophisticated data filters, workflows, and data visualization. The system provides an advanced metadata catalogue based on PostgreSQL (Mälicke and Dolich 2023), a fine-grained user and authentication management, workflows and tools for data visualization and data analysis. Under the hood, we put emphasis on a modular design through containerization (Figure 2), allowing for easy exchange of components and extensions of the portal. The

whole system is composed of open-source projects and is itself open source as well. The central building block is the secure and scalable Python web framework Django, which is well documented and actively supported by a large community. Interaction with the map is handled with the JavaScript library OpenLayers.

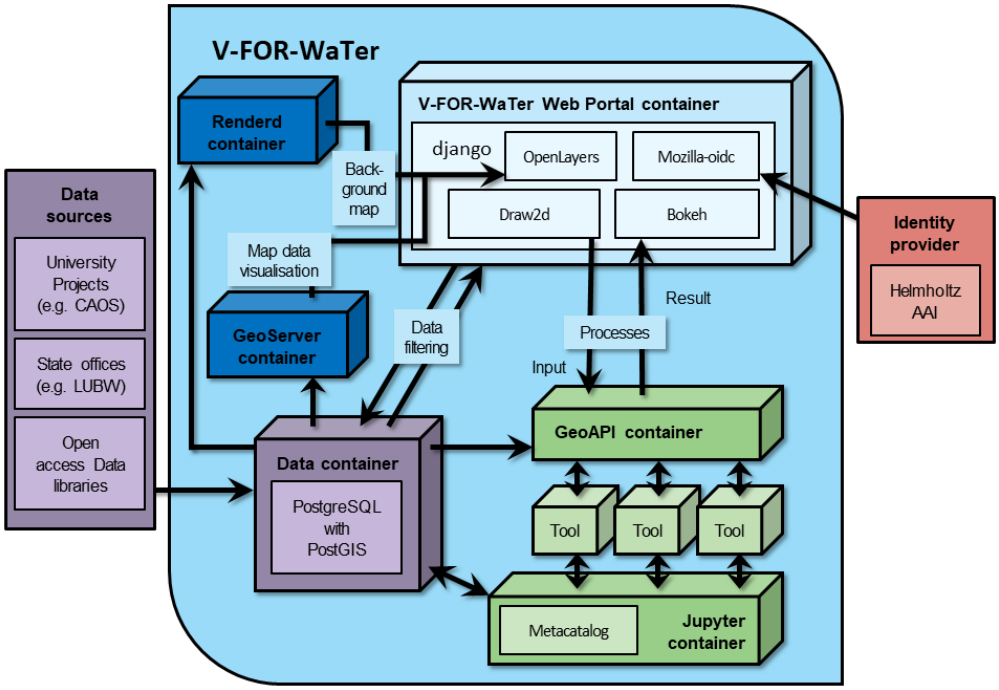


Figure 2: Current architecture of the V-FOR-WaTer portal. All components (3D boxes in the image) run in separate docker containers at Karlsruhe Institute of Technology (KIT).

While V-FOR-WaTer is intended to provide easy and open access to data and tools, restricted data access is necessary in some cases. This can be due to sensitive information contained in the data or a maximum 2-year embargo period imposed by the data owner for completing data analyses and publication. Consequently, no direct access to the data is provided. Instead, access requests are verified and redirected in Django. The authentication is facilitated by the federated Helmholtz Authentication and Authorization Infrastructure (AAI). Access to the metadata of all datasets is open for everyone and happens on the Web Portal through Django and GeoServer. The latter is used especially to visualize the position and extent of datasets on the map through a Web Feature Service (WFS).

The collection of tools provided in the V-FOR-WaTer web portal are accessed as API – Processes of the Open Geospatial Consortium (OGC), a common standard for web-based geo-applications. Using the OGC API – Processes standard ensures that the portal can be easily expanded with new tools and also enables direct access to the tools, e.g., from Jupyter Notebooks. In the Python backend, the V-FOR-WaTer toolbox already contains a set of example tools and packages, from simple hydrological signatures to comprehensive

variogram analyses. For the creation of shareable workflows we have developed a model builder, based on Draw2d.js, offering a drag-and-drop functionality to connect data and tools. A test instance for demonstrations is currently up and running at `https://portal.vforwater.de` (*Last accessed on May 12th, 2023*).

4 Conclusions

The V-FOR-WaTer web portal provides a centralized platform for scientists to access relevant data and tools, thereby greatly assisting them in searching, preparing, analysing, and publishing of data. By streamlining these processes, the portal facilitates the advancement of scientific knowledge and fosters reproducibility in research. In the future, the code of the web portal could be reused in other fields, where spatial information of their data is required and the visualization on a map is mandatory.

Acknowledgements

The ISABEL project is being funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – STR 1231/4-1, ZE 533/21-1.

References

Hutton, Christopher, Thorsten Wagener, Jim Freer, Dawei Han, Chris Duffy, and Berit Arheimer. 2016. "Most computational hydrology is not reproducible, so is it really science?" *Water Resources Research* 52 (10): 7548–7555. DOI: `https://doi.org/10.1002/2016WR019285`.

Mälicke, Mirko. 2021. *VForWaTer/hydrobox: Version 0.2.* Visited on September 6, 2023. DOI: `https://doi.org/10.5281/zenodo.4774860`.

Mälicke, Mirko, and Alexander Dolich. 2023. *VForWaTer/metacatalog: v0.8.0.* DOI: `https://doi.org/10.5281/zenodo.7643117`.

Manoj Jaseetha, Ashish, Franziska Villinger, Mirko Mälicke, Ralf Loritz, and Erwin Zehe. 2023. "Representative Hillslope Approach for Modeling Flash Flood Generation in Ungauged Catchments". DOI: `https://doi.org/10.5194/egusphere-egu23-6096`.

Stagge, James H., David E. Rosenberg, Adel M. Abdallah, Hadia Akbar, Nour A. Attallah, and Ryan James. 2019. "Assessing data availability and research reproducibility in hydrology and water resources". *Scientific Data* 6 (1). DOI: `https://doi.org/10.1038/sdata.2019.30`.

Strobl, Marcus, Elnaz Azmi, Sibylle K. Hassler, Mirko Mälicke, Jörg Meyer, Achim Streit, and Erwin Zehe. 2022. "V-FOR-WaTer – a virtual research environment for environmental research". In *E-Science-Tage 2021: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 394–398. heiBOOKS. DOI: https://doi.org/10.11588/heibooks.979.c13755.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Data Competence for Photonic Nanotechnologies

Jörg Meyer[1], Nigar Asadova[2], Dominik Beutel[3], Uğur Çayoğlu[1], Carsten Rockstuhl[2,3], Frank Tristram[4]

[1]Steinbuch Centre for Computing, Karlsruhe Institute of Technology;
[2]Institute of Nanotechnology, Karlsruhe Institute of Technology;
[3]Institute of Theoretical Solid State Physics, Karlsruhe Institute of Technology;
[4]3D Matter Made to Order, Karlsruhe Institute of Technology

In nano-optics, the optical properties of scattering objects can be expressed by so-called transition matrices (T-matrices). T-matrices are obtained by solving Maxwell's equations for given configurations under many different illumination conditions. The T-matrix can be extracted from the individual scattering response. Although this might be compute-intensive, T-matrices, as the outcome of these computations, are often not shared nor reused. This is a waste of resources and also does not permit addressing novel scientific questions. In this contribution, we present the strategies of the *data competence for photonic nanotechnologies* (DAPHONA) project to tackle this issue by suggesting a standardized file format for T-matrices and functionalities of a proper web service.

1 Introduction

In nano-optics we deal, among other things, with the optical properties of structures with a spatial extent comparable to or smaller than the optical wavelength. These scatterers have many applications, e.g., in imaging, sensors, or quantum technologies. All optical properties of these scatterers are captured by their transition matrix (T-matrix). This T-matrix describes how an illumination field is converted into a scattering field. T-matrices are the basis for describing complex nano-optical systems that are made out of many individual scatterers. These are systems that consist of coupled particles, a larger number of disordered particles, or structures made out of a periodic arrangement of identical objects. Properties of all these advanced photonic materials can be easily expressed once the T-matrix of the individual constituents is known. Currently, these T-matrices are repeatedly recalculated and not systematically reused. This wastes computational resources and also does not allow to address questions that could be answered on the basis of these data.

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.c18092

The presented project DAPHONA (Data competence for photonic nanotechnologies) aims to address this shortcoming. We will provide technologies to combine the geometric and material properties of an object and its optical properties, expressed by the T-matrix, into a unified data structure. This data will be used systematically to extract the T-matrix from the data set for a given object. Also, we want to identify objects that have predefined optical properties expressed by a T-matrix. Along these approaches, we will answer many novel questions that can be addressed by the data-driven approach formulated here.

In DAPHONA, we combine subject specific expertise with data expertise. We present our approach to the data model at an early stage of the project to allow coordination with the community and experts from research data management. Our approach is open, based on FAIR principles (Wilkinson et al. 2016), and will provide sustainable benefits to the entire community.

2 Transition Matrix Method

At the center of all considerations are the properties of the individual scattering objects. All, and really all, linear optical properties of such a scatterer can be represented in a highly aggregated form: the transition matrix (abbrev. T-matrix; Waterman 1965, 1971). The T-matrix generally describes how an illumination field with a fixed wavelength and expanded in a suitable basis system (represented with a vector, which contains the amplitudes of the basis functions) is converted into a scattered field, which is also expanded in an appropriate basis system (which, again results in a vector). The relation between illumination and scattered field is a vector-matrix multiplication, in which the illumination vector is merely multiplied by the T-matrix. Starting from the T-matrix as the central element, various ways have been developed in recent years to describe complex photonic systems. In particular, the optical properties of metasurfaces, metamaterials, or photonic crystals can be solved numerically or even analytically. The propagation properties of light in an infinite number of periodically arranged scatterers can be calculated on the basis of a band structure. The optical response of an array of thousands, sometimes even up to a million different objects can be calculated quickly if the T-matrix of each individual object is known.

3 Common File Format

We chose the Hierarchical Data Format (HDF5) as the storage format for T-matrices as it is commonly used by scientific communities. It is an open format, supported by open-source libraries in various programming languages on several platforms, and it is maintained by the HDF Group, a non-profit organization. The data model of HDF5 can represent heterogeneous data objects with various data types and in particular supports n-dimensional datasets. HDF5 is self-describing, i.e., allows to include metadata[1]. An HDF5 file describing a T-matrix needs to include the following minimal information:

1 `https://www.hdfgroup.org`; *Last accessed on September 9th, 2023.*

- T-matrix: i.e., a complex-valued matrix
- Modes: l, m, and the electric or magnetic polarization (l and m are indexes specifying the modes described by vector spherical waves)
- Name, description, and keywords
- ID as a unique identifier
- Frequency of electromagnetic waves
- Name and version of the software that created the HDF5 file

HDF5 does not natively support complex numbers. The real and imaginary parts are stored in separate arrays. The modes need to be associated with the rows and columns of the T-matrix. A full specification of the file format needs to include the applied definitions of the normalization of the modes allowing to define the used vector spherical waves unambiguously. Frequencies may be given as frequency, angular frequency, vacuum wavelength, vacuum wavenumber, or angular vacuum wavenumber that are related to each other. While this set of information is the minimum required more information should be specified in addition:

- Material: description of the isotropic, anisotropic, or bi-isotropic material
- Embedding material: allows to specify the embedding medium in case it differs from the vacuum
- Computation
- Geometry
- Mesh

An isotropic material can be specified by giving the relative permittivity and relative permeability. The metadata on the computation contains information on the used software, version, and parameters used to calculate the T-matrix. The description of the geometry of a scatterer is very important. In Table 1 example parameters are given that describe the geometry of simple 3D shapes. More complicated geometries and material distributions can be described by a mesh. The mesh itself should be specified in an existing standard file format like STL (StereoLithography file) (Roscoe 1988) or gsmh (file format of the tool Gsmh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities; Geuzaine and Remacle 2009).

4 Repository for T-Matrix Data

While the definition and establishment of a common file format for T-matrices is an important step for the reusability of T-matrix calculations, in addition, a web platform is needed with the functionality of a repository and data-specific search functionality. Repositories allow providing FAIR data. This includes publishing and sharing of data. New T-matrices may be uploaded and described by metadata. If needed, a persistent

Table 1: Example parameters to describe the geometry of simple three-dimensional shapes and their defining parameters.

Shape	Parameters	Comments
Sphere	radius	
Ellipsoid	radiusx, radiusy, radiusz	
Spheroid	radiusxy, radiusz	
Cylinder	radius, height	Rotational symmetry around z-axis
Cone	radius, height	Rotational symmetry around z-axis
Torus	major_radius, minor_radius	Rotational symmetry around z-axis
Cube	length	
Rectangular cuboid	lengthx, lengthy, lengthz	

identifier like a DOI could be assigned to a dataset. However, for scientists, it would be very useful to be able to search for domain-specific metadata. For example, a scientist might want to query all T-matrices for a given material and geometry. The search engine should not only return exact matches but should allow for fuzzy searches, i.e., also list results for similar query parameters. One objective of the DAPHONA project is to propose a suitable metric defining the similarity of T-matrices for this purpose.

5 Inverse Problem

For a given configuration (materials and geometry) the Maxwell's equations can be solved, and the T-matrix be derived. Inferring the materials and geometry for a given T-matrix is called the inverse problem. The solution of the inverse problem does not need to be unique and in general, is not simple to achieve. One approach to tackle this challenge is to apply machine learning. While supervised machine learning approaches proved to be very successful in many domains, they require large labeled training datasets. A benefit of the standardization and sharing efforts of the DAPHONA project is to compile such a training dataset for future research.

6 Conclusions

Nano optics has many practical applications and aims at the development of new materials with specific (optical) properties. In the project DAPHONA researchers with a background in nano physics and data science collaborate in order to develope a common file format to store and share T-matrices that contain all optical properties of new materials and structures. The calculation of these T-matrices is very computationally intensive. We presented ideas to define a common data format based on HDF5 allowing to share and reuse T-matrices to avoid the recalculation of the same or similar T-matrices. Besides a common file format, a repository with T-matrix-specific search functionalities is required

for the community. Once established a repository with T-matrices can be the basis for further research questions like trying to solve the inverse problem with supervised machine learning.
We present the ideas of DAPHONA at an early project phase in order to be able to coordinate efforts and integrate early feedback from the nano-optics and e-science communities.

Acknowledgements

We acknowledge support by the Federal Ministry of Education and Research (BMBF) within the project DAPHONA (16DKWN039).

References

Geuzaine, Christophe, and Jean-François Remacle. 2009. "Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities". *International Journal for Numerical Methods in Engineering* 79 (11): 1309–1331. DOI: https://doi.org/10.1002/nme.2579.

Roscoe, Larry. 1988. "Stereolithography interface specification". *America-3D Systems Inc* 27 (10).

Waterman, Peter C. 1965. "Matrix formulation of electromagnetic scattering". *Proceedings of the IEEE* 53 (8): 805–812. DOI: https://doi.org/10.1109/proc.1965.4058.

———. 1971. "Symmetry, Unitarity, and Geometry in Electromagnetic Scattering". *Physical Review D* 3 (4): 825–839. DOI: https://doi.org/10.1103/PhysRevD.3.825.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

Bayesian Optimization Framework for Data-driven Materials Design

Giovanna Tosato[1], Arnd Koeppe[1,2], Bai-Xiang Xu[4], Michael Selzer[1,2,3], Britta Nestler[1,2,3]

[1]Institute for Applied Materials – Microstructure Modelling and Simulation (IAM-MMS), Karlsruhe Institute of Technology (KIT);
[2]Institute of Nanotechnology – MicroStructure Simulation (INT-MSS), Karlsruhe Institute of Technology (KIT);
[3]Institute for Digital Materials Science (IDM), Karlsruhe University of Applied Sciences;
[4]Institute of Materials Science, Mechanics of Functional Materials, Technische Universität Darmstadt

Research data management systems enable efficient data-driven design and optimization across research disciplines. In materials science, advancing experimental design and optimizing materials with complex properties pushes the boundaries of research. For aqueous foams, the microstructure significantly influences the macroscopic properties, which can be characterized through experimental techniques or simulated using physics-based models. To address the complexity of tuning interrelated parameters in a vast design space, we propose an active learning-based data-driven framework.

This framework leverages Bayesian Optimization to guide the exploration of the search space, prioritizing informative experiments or simulations and minimizing the number of required evaluations. Implemented within the Kadi ecosystem, our workflow promotes data reuse and facilitates seamless collaboration.

1 Introduction

In materials design, optimizing material properties for specific applications is a challenging task, as these properties often exhibit complex behaviors. Effects at multiple scales influence macroscopic material behavior and function, e.g., microscopic pores affect stiffness and thermal insulation. The characterizing parameters of microstructures are often numerous and interconnected through relationships that are only partially understood. This high dimensionality and complex dependencies yield a vast design space, which poses the researchers the challenge of tuning several parameters simultaneously to gain the desired performance.

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.c18093

This multifaceted challenge is prominently evident in the study use case of liquid foam evolution, where additionally, the non-linear, time-series nature of the data has to be taken into account. To that end, machine learning can handle unknown, nonlinear, complex, and high-dimensional relationships, but often requires large amounts of data to recapture information.

The acquisition of this data frequently involves performing high numbers of resource-intensive experiments or simulations. Therefore, it is desirable to generate information-rich datasets while minimizing the number of required evaluations. Our strategy to generate small but comprehensive datasets is to exploit Bayesian Optimization and its memory of previously acquired data to guide the selection of the next most informative data point to be evaluated.

Advancing research within the context of e-science and open data, the need for research data management support becomes self-evident. Therefore, we operate within the Kadi ecosystem, our open-source research data infrastructure for materials science[1], which empowers our research following the FAIR principles. The Kadi4Mat repository (Brandt et al. 2021), the workflow engine KadiStudio (Griem et al. 2022), and the interface for AI, KadiAI (Koeppe and CIDS Team 2023) enable structured data storage and reproducible workflows for data analysis and visualization tasks, facilitating sharing and reproducibility.We enclose the Bayesian oracle into a Python framework and make it accessible from the Kadi ecosystem through a node-based graphical interface. This enables the development of "intelligent" study workflows that actively learn to investigate the parameter space.

2 Methods

In materials design, data-driven approaches often outperform traditional trial-and-error and even systematic a-priori planned studies (Himanen et al. 2019). However, the efficiency of data-driven approaches hinges on the ability to guide data collection effectively. Our solution to this problem consists of a design-of-experiments methodology leveraging Bayesian Optimization to efficiently explore and exploit the search space. We iteratively update our model and enrich our dataset by selecting the next most informative data point, guided by previous evaluations. This approach falls in the category of *active learning* (Settles 2009), a subfield of supervised machine learning where a component of the algorithm (an oracle) is responsible for choosing the training data. Motivated by the knowledge that not all samples are equally meaningful, we aim to obtain an expressive dataset that meets the requisites of quality, quantity, and variability while reducing resource usage.

In this work, we implement a query strategy based on a Bayesian oracle. Bayesian Optimization (BO) is known in statistics as a method for global optimization of black-box functions, usually employed in the case of expensive-to-evaluate functions.

1 `https://kadi.iam.kit.edu`

In resource-intensive materials simulations and experiments, microstructure-property relations represent an attractive use-case for BO. This aspect significantly contributes to the growing attention given to BO as an efficient tool for multidimensional optimization in materials design (Kotthoff, Wahab, and Johnson 2021; Kuhn et al. 2021; Zhao et al. 2023; Kulagin et al. 2023).

The main strength of the method lies in the ability to define, work with, and continuously update a cheap *surrogate function* which estimates the behavior of the real (expensive) objective function. The surrogate is a probabilistic model over functions mirroring our objective. A common choice is to use *Gaussian Processes*, which approximate the behavior of the datasets and provide uncertainty estimates. To cover the search space balancing exploration and exploitation, the most promising candidate is selected maximizing a customizable *acquisition function*, see Figure 1.

A customizable active learning framework is under development in the Computational Intelligence and Data Science framework CIDS (Koeppe et al. 2022) within the Kadi ecosystem.

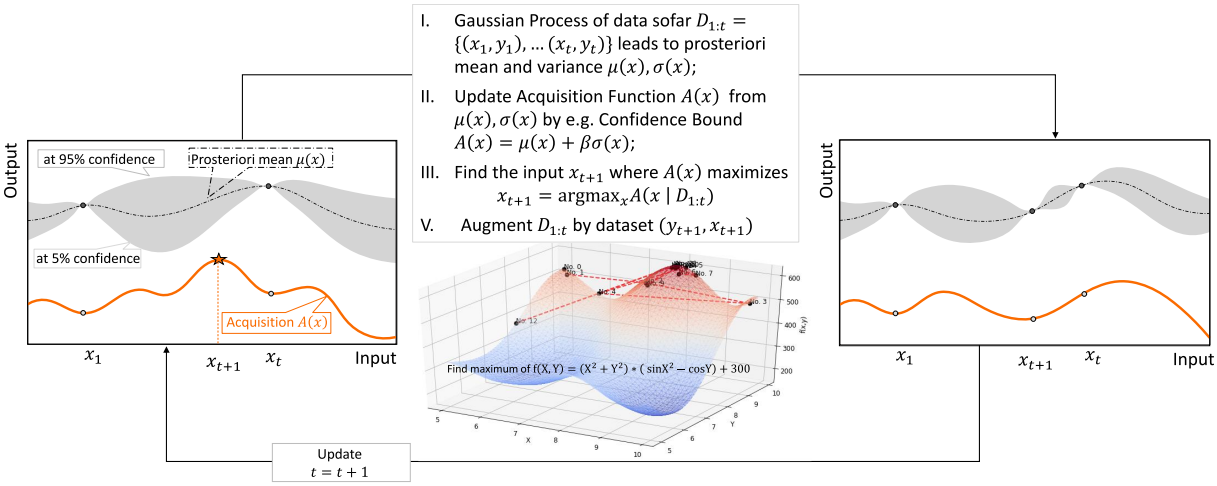


Figure 1: Example iteration of Bayesian Optimization.

3 Workflow

For the use-case of liquid foam evolution, Figure 2 shows how the whole data life cycle exploits and is embedded in Kadi4Mat, which manages every step of the research process, including generation, storage, usage, and sharing. For the generation of simulation data, a KadiStudio workflow automatizes the creation of initial parameters' settings and performs the designed simulation(s) employing the in-house framework Pace3D, Parallel Algorithms for Crystal Evolution in 3D (Hötzer et al. 2018).

A representative temporal evolution analysis of 2D liquid foams was performed on the BwUniCluster2.0 platform, utilizing 40 cores over a computational time of 35 hours. A single timestamp for such simulations occupies around 500MB of storage. To adequately

capture temporal evolution patterns, the study necessitates the examination of 50 to 200 frames for each configuration.

In contrast, the experimental data are generated with a Dynamic Foam Analyzer (DFA100, Krüss GmbH, Hamburg, Germany) from research partners (Jung et al. 2023). Significant observations can span a duration of up to 10 hours, while the equipment generates a complete status dump every 0.2 seconds, yielding a substantial volume of unprocessed data.

For storage, the results are uploaded to Kadi4Mat, imparting structure and enabling access for all the users working on the project.

Kadi allows heterogeneous data formats and corresponding metadata, to answer the needs of an interdisciplinary research field such as materials science. The Bayesian oracle framework has been developed with the same vision in mind and therefore adapts to different (meta)data types. To support fast metadata collection, Kadi offers the possibility to define templates, in combination with validation functionality and, for some type of metadata (e.g. creation time), automatic recording.

For data usage, KadiAI integrates and standardizes AI projects, work packages, and workflows (Koeppe et al. 2021) into the Kadi ecosystem, serving as an interface between RDM and machine learning applications. KadiAI is combined with the Python-based framework CIDS, designed to facilitate the development and implementation of learning algorithms in AI workflows. Thus, the Kadi4Mat ecosystem can leverage data-integrated AI to enable quantitative and qualitative data analysis.

For sharing, Kadi4Mat allows management of access and editing rights for each project, as well as for each specific record. Different roles can be established for working groups or specific registered users. This feature is designed to enhance collaborations among different groups and institutions, while upholding the imperative of controlled data access.

In the vision of a partially-automated laboratory, all data are to be processed in an automated and reproducible manner. Therefore, we develop workflows in KadiStudio that take care of the consequential execution of each research step within a graphical user interface (GUI), using an intuitive point-and-click mechanism. The Bayesian oracle has been implemented as (a set of) nodes, basic building blocks of our workflow system (Figure 3). To monitor the optimization progress, a visualization of the Bayesian oracle current and global state is available in Kadi4Mat.

4 Conclusions

Establishing a common workspace and the possibility of sharing research data on a routine basis has proven to be crucial, especially for collaboration across multiple research groups. Workflows have streamlined the research process, making it faster and providing a simple and intuitive way for researchers, including non-experts, to reproduce their work.

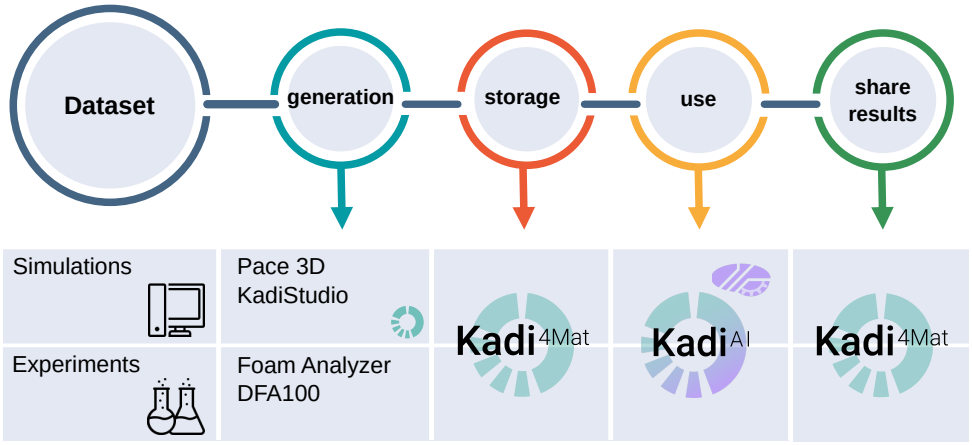


Figure 2: Kadi4Mat integration through the whole research process for the foam evolution use case.

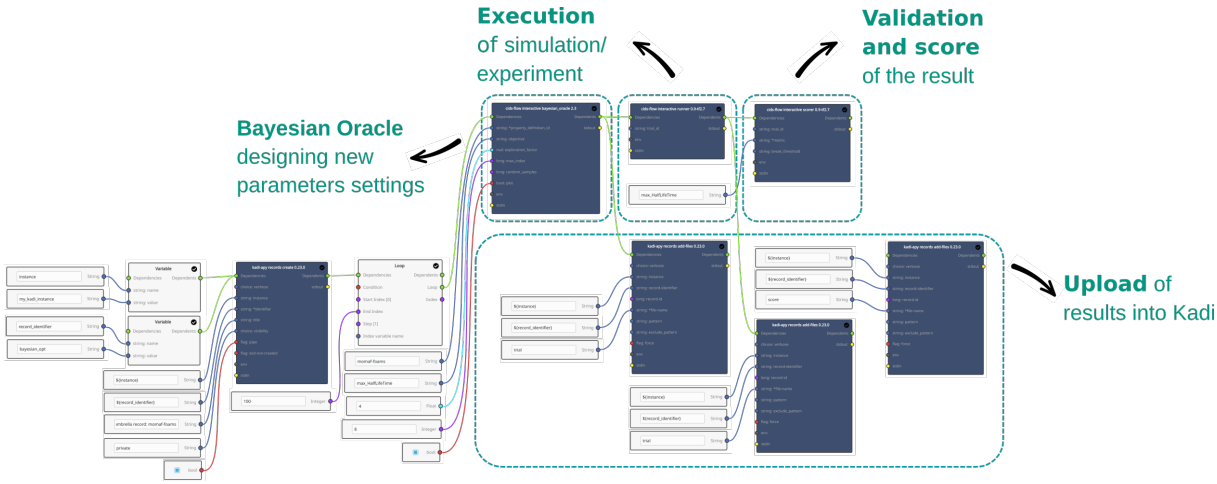


Figure 3: Example workflow implementation of the Bayesian oracle framework.

The possibility to store, share, and work with data in an intuitive manner plays a more and more important role in research. A structured working environment is provided for the day-to-day work of researchers from different groups, places, and disciplines.

Incorporating the Bayesian oracle in this workflow context, in the form of user-friendly nodes, provides a cost-effective solution accessible and utilizable for a broader range of users. Furthermore, this iterative approach presents a valuable opportunity to accelerate the process of materials' development.

Acknowledgements

This work is funded by the Ministry of Science, Research and Art Baden-Württemberg (MWK-BW) in the SDC MoMaF, with funds from the state digitization strategy digital@bw (project No. 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), 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.

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.

Himanen, Lauri, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. 2019. "Data-Driven Materials Science: Status, Challenges, and Perspectives". *Advanced Science* 6 (21). DOI: https://doi.org/10.1002/advs.201900808.

Hötzer, Johannes, Andreas Reiter, Henrik Hierl, Philipp Steinmetz, Michael Selzer, and Britta Nestler. 2018. "The parallel multi-physics phase-field framework Pace3D". *Journal of Computational Science* 26:1–12. DOI: https://doi.org/10.1016/j.jocs.2018.02.011.

Jung, Ole, Mike Barbeck, L U Fan, Fabian Korte, Cuifeng Zhao, Rumen Krastev, Sven Pantermehl, and Xin Xiong. 2023. "Republication: In Vitro and Ex Vivo Analysis of Collagen Foams for Soft and Hard Tissue Regeneration". *In Vivo* 37 (1): 320–328. DOI: https://doi.org/10.21873/invivo.13082.

Koeppe, Arnd, Franz Bamer, Michael Selzer, Britta Nestler, and Bernd Markert. 2021. "Workflow concepts to model nonlinear mechanics with computational intelligence". *PAMM* 21 (1). ISSN: 1617-7061. DOI: https://doi.org/10.1002/pamm.202100238.

———. 2022. "Explainable artificial intelligence for mechanics: physics-explaining neural networks for constitutive models". *Frontiers in Materials* 8:636. DOI: https://doi.org/10.48550/arXiv.2104.10683.

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

Kotthoff, Lars, Hud Wahab, and Patrick Johnson. 2021. *Bayesian Optimization in Materials Science: A Survey.* Technical report. Arxiv. DOI: https://doi.org/10.48550/arXiv.2108.00002.

Kuhn, Jannick, Jonathan Spitz, Petra Sonnweber-Ribic, Matti Schneider, and Thomas Böhlke. 2021. "Identifying material parameters in crystal plasticity by Bayesian optimization". *Optimization and Engineering* 23:1489–1523. ISSN: 15732924. DOI: https://doi.org/10.1007/s11081-021-09663-7.

Kulagin, Roman, Patrick Reiser, Kyryl Truskovskyi, Arnd Koeppe, Yan Beygelzimer, Yuri Estrin, Pascal Friederich, and Peter Gumbsch. 2023. "Lattice Metamaterials with Mesoscale Motifs: Exploration of Property Charts by Bayesian Optimization". *Advanced Engineering Materials* 25 (13). ISSN: 1438-1656. DOI: https://doi.org/10.1002/adem.202300048.

Settles, Burr. 2009. *Active Learning Literature Survey.* Technical report. https://burrsettles.com/pub/settles.activelearning.pdf.

Zhao, Yinghan, Patrick Altschuh, Jay Santoki, Lars Griem, Giovanna Tosato, Michael Selzer, Arnd Koeppe, and Britta Nestler. 2023. "Characterization of porous membranes using artificial neural networks". *Acta Materialia* 253:118922. ISSN: 1359-6454. DOI: https://doi.org/10.1016/j.actamat.2023.118922.

Veranstalter

Die **E-Science-Tage** werden vom Projekt bw2FDM unter Beteiligung des Karlsruher Instituts für Technologie, der Universität Konstanz und der Universität Heidelberg veranstaltet und vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg gefördert.

Baden-Württemberg

MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

Druck und Bindung
Books on Demand GmbH
In de Tarpen 42, 22848 Norderstedt

Die Forschung im digitalen Wandel benötigt ein koordiniertes Umfeld. Die E-Science-Tage 2023 boten Verantwortlichen und Interessierten die Möglichkeit, sich zu den aktuellen Entwicklungen im Bereich Forschungsdatenmanagement auszutauschen und zu vernetzen. Mit dem Thema „Empower Your Research – Preserve Your Data“ lag der Fokus auf der nachhaltigen Speicherung und der interdisziplinären sowie internationalen Verfügbarkeit von Forschungsdaten. In diesem Rahmen wurden die Bedeutung, die Risiken und der Nutzen einer transparenten Bereitstellung und Zugänglichkeit von Forschungsdaten thematisiert. Der vorliegende Tagungsband ist eine Sammlung von Vorträgen und Postern zu diesen Themen.

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

ISBN 978-3-948083-91-5