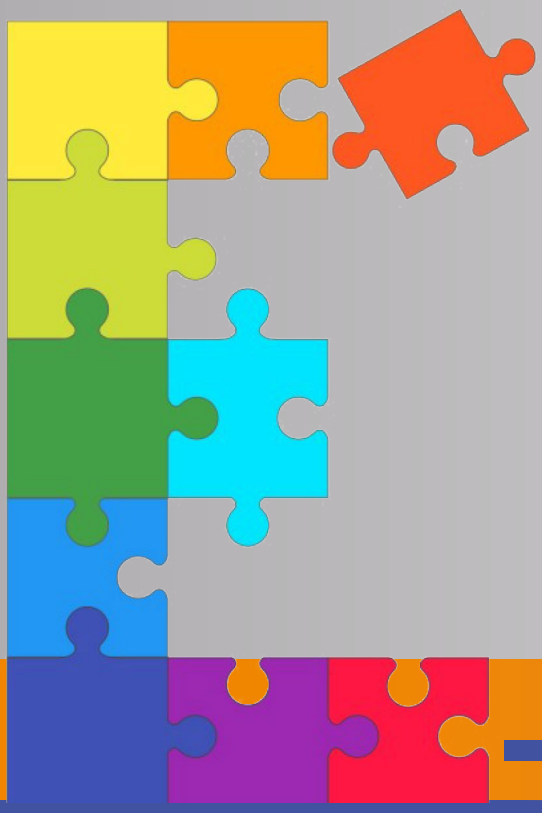


Vincent Heuveline
Philipp Kling
Florian Heuschkel
Sophie G. Habinger
Cora F. Krömer
(Hrsg.)

-Science-Tage 2025

Research Data Management: Challenges in a Changing World

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

E-Science-Tage 2025

E-Science-Tage 2025

Research Data Management: Challenges in a Changing World

Herausgegeben von

Vincent Heuveline, Philipp Kling, Florian Heuschkel,

Sophie G. Habinger, Cora F. Krömer

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 https://dnb.dnb.de abrufbar.

Dieses Werk ist unter der Creative-Commons-Lizenz CC BY-SA 4.0 veröffentlicht. Die Umschlaggestaltung unterliegt der Creative-Commons-Lizenz CC BY-ND 4.0.

Publiziert bei heiBOOKS, 2025

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

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-1652
doi: https://doi.org/10.11588/heibooks.1652

Text © 2025, das Copyright der Texte liegt bei den jeweiligen Verfasser:innen.

ISBN 978-3-911056-52-6 (Softcover)
ISBN 978-3-911056-51-9 (PDF)

Inhaltsverzeichnis

I Grußworte **11**

Grußwort der Ministerin für Wissenschaft, Forschung und Kunst Baden-Württemberg
Petra Olschowski 13

Grußwort der Rektorin der Universität Heidelberg
Frauke Melchior 15

Vorwort der Herausgeber*innen
Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, Cora F. Krömer 18

II Wissenschaftliche Beiträge **21**

Data Management in INF Projects of Collaborative Research Centres: Building Bridges Between Research, Infrastructure and Practice
Oliver Koepler, Iryna Mozgova, Florian Nürnberger, Christoph Steinbeck, Jürgen Pleiss 23

MaRDMO: Future Gateway to FAIR Mathematical Data
Marco Reidelbach 40

Data-Driven Community Standards for Interdisciplinary Heterogeneous Information Networks
Mattis thor Straten, Steffen Strohm, Florian Thiery, Matthias Renz 54

Systematic Assessment of the Prerequisites for Institutional RDM Services at a German University of Applied Sciences
Paulina Anna Dąbrowska, Esther Schneidenbach, Eske Heister, Lena Simon, Wojtek Dabrowski 77

Institutional Data Science Services at FDZ UB Mannheim: Enhancing Research Data Management
Renat Shigapov, Irene Schumm, Thomas Schmidt, Jan Kamlah, Larissa Will 94

Forschungsdatenmanagement in Nordrhein-Westfalen: Das Landeskonzept FDM und die Verstetigung der Landesinitiative fdm.nrw
Magdalene Cyra, Matthias Fingerhuth, Lioba Schreyer, Christina Schröder 102

The Triathlon Relay of Storage Space Provisioning
Ilona Lang, Marcel Nellesen, Marius Politze 112

Staying Flexible and Transparent in a Changing World: Agile RDM Service Development
Petar Hristov, Katja Jansen, Ilona Lang 123

Ein standardisierter Datenmanagementplan als Hilfsmittel für niederschwelliges Forschungsdatenmanagement. Die Lösung für alle Herausforderungen?
Harald Kaluza 134

How Are Research Data Referenced? The Use Case of the Research Data Repository RADAR
Dorothea Strecker, Kerstin Soltau, Felix Bach 146

Auslotung des Potenzials von Fachreferent:innen als FDM-affine Unterstützungsgruppe in einer forschungsintensiven Hochschullandschaft
Roberto Cozatl 160

ROADS to Reusable Datasets: Potenziale und Herausforderungen bei der Datennachnutzung in der Hochschullehre
Anna Daudrich 171

Challenge Accepted? Möglichkeiten und Grenzen eines NFDI-Serviceverzeichnisses am Beispiel des NFDI-N(HAW)igators
Eva Eilert, Martin Reiter 184

Pitfalls in the Field – Rechtliche Herausforderungen beim Einsatz von ML in der Agrosystemforschung
Constantin Breß, Lea Sophie Singson 197

Resilient Hosting of Research Data Management Services
Halima Saker, Simon Pirkl, Jens Krüger, Holger Gauza, Suvasini Thangaraj, Ursula Eberhardt, Alexander Kirbis, Lucas Beuter . . . 216

Efficient Data Curation by Automating Workflows
Florian Fritze, Dorothea Iglezakis, Sarbani Roy, Björn Selent, Karoline Weinspach 235

Bremen's Interdisciplinary Data Competence Center DataNord: An Integrative Approach for Data Literacy in Research
Lena Steinmann, Tanja Hörner, Uta Bohnebeck, Rolf Drechsler, Frank Oliver Glöckner, Iris Pigeot 249

Terminology Services in the DACH Region Landscape – What Are the Essential Requirements?
Roman Baum, Syphax Bouazouni, Jan Fillies, Naouel Karam, Oliver Koepler, Pooya Oladazimi, Julia Sasse 266

Forschungsethik & Datenschutz – unterstützende Geschwister im Forschungsdatenmanagement. Eine Einordnung aus Perspektive der Pflege- und Sozialwissenschaften
Andrea Kuhn 284

FDM Meets HAW: Lessons Learned bei der Begleitung von Projekten der angewandten Forschung
Sarah Boelter, Romy Meyer, Wesley Preßler, Lucie Schmidt, Amélie Oberdorfer, Christoph Schmidt 299

Secure Infrastructures and Protection of Personal Data: Approaches for the Development of the NFDIxCS Research Data Management Container
Patrick Brunner, Stefan Bavendiek 314

Kooperation zwischen Universität und Informationsinfrastruktur – das Forschungsdatenrepositorium RADAR für Brandenburg
Maria Büttner, Boris Jacob, Daniela Mertzen, Christian Riedel, Kerstin Soltau 332

Beitrag zur Leistungsmessung bei der Evaluation einer mit RISE-DE an einer Hochschule für angewandte Wissenschaften erstellten Forschungsdatenstrategie
Ian Wolff, Claus Spiecker, Daniela Mertzen 347

The Impact of Digitalisation on Science: Workflows, Outputs, and Trust
Felix Bach, Linda Nierling, Christian Bonatto Minella, Angelina Sophie Dähms 365

Nachnutzbar, aber nicht reproduzierbar. Abgeleitete Textformate für die Arbeit mit urheberrechtsbewehrten Texten
Philippe Genêt, Lukas Weimer 386

Shaping the Data Steward Community in Germany
Jens Dierkes, Daniela Hausen, Ute Trautwein-Bruns 394

Perplexity-Inspired Metasearch-Based Alternatives to FAIR GPT: Open-source AI Consultants for Research Data Management
Thomas Schmidt, Renat Shigapov, Jan Kamlah, Irene Schumm 402

Licensing Research Software Made Easy: Introducing the License Checker Tool
Dorothea Iglezakis, Iqra Imran 409

Die GesundFDM Spring School: Praxisnaher Erwerb von Datenkompetenzen für Studierende
Katharina Koch, Stephanie Werner, Robert Werth, Nadine Probol, Eike Quilling 416

Modulare standortübergreifende Schulungsangebote: Das FDM-Curriculum der Universitätsallianz Ruhr
Julia Stapels, Tuba Güden-Silber, Jessica Stegemann 424

Open Science in den Humanities: Mehr Spielräume und Nachvollziehbarkeit durch Galaxy
Daniela Schneider, Jan Leendertse 432

Bridging the Troubled Waters of Peer Review with ing.grid: Towards Openness, Fairness and F.A.I.R.-ness
Agnes Kleinhans, Michaela Leštáková, Kevin Logan, Peter F. Pelz 439

The CREATIVE Project – Customizing a Generic Repository for Domain Scientists and Creating a Data Steward Network
Sibylle K. Hassler, Peter Braesicke, Mirko Mälicke, Jörg Meyer, Carlos Zuleta Salmon 447

FIZ-OAI: Eine skalierbare Open-Source-Implementierung zur Aggregation und Dissemination von Metadaten mit OAI-PMH
Stefan Hofmann, Kerstin Soltau, Michael Hoppe, Felix Bach 454

Uniting the DIFferences With the DALIA Interchange Format (DIF) – a Minimal Metadata Specification for the DALIA OER Platform
Jonathan D. Geiger, Petra C. Steiner, Abdelmoneim Amer Desouki, Henrika M. Hüppe, Jan-Michael Haugwitz 460

FORTH-BW: Erarbeitung und Bereitstellung eines bedarfsorientierten FDM-Workflows für HAWs
Karoline Herrmann, Ina Elisabeth Klingele, Lioba Petter, Carsten Herbes 468

LLM4DDC: Adopting Large Language Models for Research Data Classification Using Dewey Decimal Classification
Gautam Kishore Shahi, Renat Shigapov, Oliver Hummel 476

New Developments in RADAR: Safeguarding and Publishing Research Data for Long-term Usability
Kerstin Soltau, Stefan Hofmann, Sandra Göller, Felix Bach 485

Was machen eigentlich Data Champions?
Anne Gärtner, Kristina Grüttemeier, Markus Rump, Johanna Vompras . . 491

Automated Metadata Extraction Compliant with Machine-actionable Software Management Plans
Dhwani Solanki, Suhasini Venkatesh, Dietrich Rebholz-Schuhmann, Leyla Jael Castro 499

elAPI: A Powerful, Extensible API Client for eLabFTW
Mahadi Xion, Alexander Haller, Philipp Kling, Martin Baumann, Nick Kepper 505

DataPLANT Services Design – Considerations Towards a Common NFDI Landscape
Dirk von Suchodoletz, Marcel Tschöpe, Jonathan Bauer, Kevin Schneider, Timo Mühlhaus 515

Veranstalter
Landesinitiative für Forschungsdatenmanagement in Baden-Württemberg . . 523

Teil I

Grußworte

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

Petra Olschowski

Das Thema der diesjährigen E-Science-Tage lautet „Research Data Management: Challenges in a Changing World“. Diese Veränderungen betreffen alle Ebenen der Wissenschaft – das spüren wir jeden Tag. Das Tempo der aktuellen Entwicklung bedeutet enorme Herausforderungen. Dass damit aber genauso Chancen verbunden sind, wird auch und vor allem im Wissenschaftsbetrieb immer wieder deutlich. Gerade Wissenschaftlerinnen und Wissenschaftler sind es, die mit viel Neugier und Mut lösungsorientiert neue Wege suchen und andere Perspektiven eröffnen – und sie tun dies mit viel Erfolg hier in Baden-Württemberg.

Die Aufgabe der Wissenschaftspolitik ist es, für diese Prozesse die richtigen Rahmenbedingungen zu schaffen und sie laufend den neuen Gegebenheiten anzupassen. Das gilt ganz besonders für die Bereiche Künstliche Intelligenz, Höchstleistungsrechnen und Forschungsdaten.

Seit 2008 haben wir in Baden-Württemberg eine Landesstrategie zum Höchstleistungsrechnen. Im Jahr 2016 ist das CyberValley für KI ins Leben gerufen worden. An der Landesstrategie Forschungsdaten wird seit dem Jahr 2023 erfolgreich gearbeitet. Wir sehen, wie diese Bereiche eng verwoben sind: Forschungsdaten sind die Grundlage für innovative Forschung, unerlässlich für eine verlässliche Qualitätsprüfung der Ergebnisse und die Basis für eine vertrauenswürdige KI.

Unsere Hochschulen und Einrichtungen in Baden-Württemberg spielen eine wichtige Rolle in der Nationalen Forschungsdateninfrastruktur (NFDI). Die Evaluation der NFDI durch den Wissenschaftsrat erbrachte ein insgesamt positives Ergebnis und wir arbeiten an der produktiven Umsetzung für die Fortführung nach 2028.

Als Land Baden-Württemberg werden wir in diesem Jahr den letzten Baustein unserer Landestrategie Forschungsdaten umsetzen. Mit diesem Baustein wird der Aufbau von Governance-Strukturen, technischer Infrastruktur und die Einstellung und Qualifizierung von Personal ermöglicht.

Sowohl das Land als auch der Bund setzen auf Open Data, übergreifendes Datenmanagement, breite Datenverfügbarkeit und die Vernetzung der Forschungseinrichtungen. Wir

begrüßen die Fortschritte bei der European Open Science Cloud und wir kooperieren international.

Wir führen verstärkt Diskussionen zur Forschungssicherheit. Denn Daten sind extrem anfällig für Missbrauch und Manipulation. Das Ziel muss es sein, Forschungssicherheit zu gewährleisten, ohne jedoch die Freiheit der Wissenschaft, Kooperationen und den offenen Austausch zu gefährden. Denn ohne diese ist Wissenschaft nicht denkbar.

Auch und gerade angesichts der aktuellen außenpolitischen Bedrohungen müssen wir diese Freiheit sichern und in Balance mit den nötigen Sicherheitsmaßnahmen bringen. Sonst kann Wissenschaft ihre Rolle zur Lösung der drängenden Probleme nicht erfüllen. Das Thema der Tagung haben Sie aus meiner Sicht genau richtig gewählt.

Das internationale Panorama mag teils übermächtig wirken, die E-Science-Tage werden Ihnen dennoch wieder einmal den Rahmen für viele anregende Diskussionen und einen inspirierenden Austausch bieten.

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

Grußwort der Rektorin der Universität Heidelberg

Frauke Melchior

Sehr geehrte Damen und Herren, liebe Gäste von nah und fern,

es ist mir eine große Freude, Sie heute an der Universität Heidelberg zu den E-Science-Tagen 2025 willkommen zu heißen. Uns erwartet ein inspirierender Austausch über ein Thema, das für die Zukunft der Wissenschaft von zentraler Bedeutung ist: Forschungsdatenmanagement. Ein Thema, das bei rapide wachsenden Datenmengen in einer sich dramatisch wandelnden Welt sehr viel komplexer ist, als es vielleicht noch vor wenigen Jahren schien.

Das Konferenzmotto „Research Data Management: Challenges in a Changing World“ könnte daher kaum treffender gewählt sein. In einer Welt, in der wissenschaftliche Erkenntnisse zunehmend datengetrieben sind, steht das Forschungsdatenmanagement vor spannenden Herausforderungen. Die Anforderungen an Speicherung, Sicherheit, Nachnutzbarkeit und Verfügbarkeit von Forschungsdaten entwickeln sich stetig weiter – und mit ihnen auch die Verantwortung, diese Daten effizient zu verwalten.

Forschung ist heute internationaler, interdisziplinärer und digitaler denn je, und hierfür gilt es, Lösungen zu finden, die dieser neuen Realität gerecht werden. Gleichzeitig bringt der technologische Wandel – von Cloud-Technologien über künstliche Intelligenz bis hin zu Open-Science-Initiativen – neue Möglichkeiten mit sich, die wir nutzen können, um Forschung transparenter, reproduzierbarer und nachhaltiger zu gestalten.

Hier an der Universität Heidelberg haben wir diese Themen früh erkannt und gezielt Strukturen geschaffen, um Forschende bestmöglich zu unterstützen. Die Research Data Unit (RDU) ist ein herausragendes Beispiel dafür: Sie bündelt die Expertise aus der Universitätsbibliothek und dem Universitätsrechenzentrum, um Forschende in allen Phasen des Forschungsdatenmanagements umfassend zu beraten und zu unterstützen. Neben der strategischen Beratung stellt die RDU auch Werkzeuge und Schulungen bereit, um die Herausforderungen der Datennutzung und -archivierung zugunsten der Nachnutzbarkeit wertvoller Forschungsdaten zu bewältigen.

Ebenso möchte ich den Dienst „Scientific Data Storage“ hervorheben, kurz SDS@hd, der als Landesdienst am Universitätsrechenzentrum angesiedelt ist und seit vielen Jahren eine

wichtige Rolle für die Forschenden in Baden-Württemberg einnimmt. SDS@hd stellt eine Basis-Technologie dar, um Forschungsdaten zu speichern, zu verarbeiten und zu teilen. Ein weiteres bedeutendes Werkzeug in unserer Infrastruktur ist der Dienst eLabFTW, der für die Digitalisierung im Labor – genauer gesagt für die Nutzung elektronischer Laborbücher – steht. eLabFTW steht als universitärer Dienst seit einem Jahr produktiv zur Verfügung und wird bereits von rund 1.000 Nutzer:innen geschätzt. Mit diesen Angeboten wollen wir als Universität Verantwortung übernehmen, um nachhaltige und praktikable Lösungen für den wissenschaftlichen Alltag bereitzustellen.

Doch Forschungsdatenmanagement ist keine isolierte Aufgabe einzelner Institutionen. Vielmehr ist es eine gesamtgesellschaftliche Herausforderung, die koordinierte Antworten auf Landes-, Bundes- und Europaebene erfordert. Wir engagieren wir uns daher aktiv in der Landesinitiative Forschungsdatenmanagement in Baden-Württemberg, oder kurz bwFDM, die eine zentrale Rolle bei der Entwicklung und Implementierung nachhaltiger und disziplinspezifischer Strategien für das Forschungsdatenmanagement spielt. Das Projekt bwFDM bringt zahlreiche Universitäten und Forschungseinrichtungen des Landes zusammen mit dem Ziel, Best Practices zu etablieren und Forschenden einen verlässlichen Rahmen für den Umgang mit Forschungsdaten zu bieten. Auch die Vernetzung mit der Nationalen Forschungsdateninfrastruktur (NFDI) gehört selbstverständlich dazu.

Natürlich gibt es auf dem Weg zu flächendeckendem und transparentem Zugang zu kurierten Forschungsdaten noch viele Herausforderungen. Eine davon betrifft das Stichwort Sensitive Daten und bezieht sich auf die Balance von Open Science und Datenschutz. Gerade in Bereichen mit hochsensiblen Daten wie in der Medizin müssen wir sicherstellen, dass Daten zwar FAIR – also auffindbar, zugänglich, interoperabel und wiederverwendbar[1] – sind, dabei aber die Privatsphäre von Individuen gewahrt bleibt. Dies erfordert technologische Lösungen, aber auch ethische und rechtliche Abwägungen, die über die Grenzen einer einzelnen Universität, einer Forschungscommunity und auch über nationale Grenzen hinausgehen.

Ein weiteres zentrales Thema ist der Einfluss künstlicher Intelligenz auf das Forschungsdatenmanagement. KI-gestützte Methoden können helfen, große Datenmengen zu analysieren und Muster zu erkennen, die für den wissenschaftlichen Fortschritt entscheidend sind. Gleichzeitig stehen wir vor der Frage, wie wir den Ursprung und die Qualität der verwendeten Daten sicherstellen und wie wir Transparenz und Reproduzierbarkeit in KI-gestützten Forschungsprozessen gewährleisten können. Die rasanten Fortschritte in der KI-Technologie eröffnen einerseits neue Möglichkeiten zur Optimierung von Datenanalysen, verlangen aber andererseits auch nach einem kritischen Diskurs über ethische, rechtliche und methodische Standards sowie Fragen der Datensouveränität.

Nicht zuletzt bleibt die langfristige Finanzierung dieser Infrastruktur eine Herausforderung. Forschungsdatenmanagement ist eine Grundvoraussetzung für exzellente Wissenschaft, findet bisweilen aber nicht genug Beachtung. Es liegt an uns allen, in Wissen-

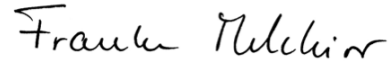
1 FAIR: Findable, Accessible, Interoperable, Reusable

schaft, Politik und Gesellschaft, die Notwendigkeit nachhaltiger Investitionen in diese Infrastruktur zu betonen und entsprechende Lösungen zu entwickeln. Die enge Zusammenarbeit zwischen Forschungseinrichtungen, Zuwendungsgeber:innen und politischen Entscheidungsträger:innen ist hier unerlässlich. Nur wenn es uns gelingt, langfristige, stabile Finanzierungsmodelle zu etablieren, können wir gewährleisten, dass wertvolle Daten nicht verloren gehen und wissenschaftliche Innovationen weiterhin auf einer soliden Grundlage aufbauen.

Zum Abschluss möchte ich unserem Förderer, dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg, und natürlich ganz besonders dem Organisationsteam der E-Science-Tage herzlich danken. Die Konferenz wird vom Projekt bwFDM veranstaltet, und von Mitarbeitenden der Partnereinrichtungen Karlsruher Institut für Technologie, der Universität Konstanz und der Universität Heidelberg organisiert. Ihr Engagement und Ihre Expertise tragen maßgeblich dazu bei, dass wir uns hier gemeinsam über die Zukunft des Forschungsdatenmanagements austauschen können. Ohne Ihren unermüdlichen Einsatz wäre eine Veranstaltung dieser Qualität und dieses Umfangs nicht möglich.

Die E-Science-Tage bieten Gelegenheit, all diese Themen intensiv zu diskutieren, voneinander zu lernen und gemeinsam an zukunftsfähigen Lösungen zu arbeiten. Ich freue mich auf die kommenden Vorträge und Diskussionen und bin überzeugt, dass diese Konferenz erneut wertvolle Impulse für die Weiterentwicklung des Forschungsdatenmanagements setzen wird.

Mit diesen Gedanken wünsche ich Ihnen allen eine erfolgreiche und erkenntnisreiche Tagung, neue Impulse für Ihre eigene Forschung und einen produktiven Austausch! Vielen Dank!



Prof. Dr. Frauke Melchior
Rektorin der Universität Heidelberg

Vorwort der Herausgeber*innen

Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, Cora F. Krömer

Liebe Leserinnen und Leser des Tagungsbandes der E-Science-Tage 2025,

die Universität Heidelberg durfte erneut, nun das fünfte Mal in Folge, Gastgeberin der E-Science-Tage sein. Diese werden von der Landesinitiative für Forschungsdatenmanagement in Baden-Württemberg – bwFDM – unter Beteiligung des Karlsruher Instituts für Technologie und der Universität Konstanz veranstaltet. In diesem Jahr konnten wir knapp 400 Teilnehmerinnen und Teilnehmer willkommen heißen, worüber wir uns besonders gefreut haben. Zugleich werten wir das auch als Zeichen, dass die Aktualität und das Interesse an den Fragestellungen rund um das Forschungsdatenmanagement immer weiter steigen. Ein besonderes Augenmerk war und ist den Herausgeberinnen und Herausgebern daher auch, dem Diskurs hierzu und gegebenenfalls auch zu kontroversen Thesen einen Raum zu geben.

Nachdem die Themen der letzten Auflagen der E-Science-Tage am Lebenszyklus von Forschungsdaten orientiert waren – man denke unter anderem an die Stichworte Erzeugung, Analyse, Aufbereitung, und Archivierung –, betonte das Motto der diesjährigen Konferenz nun eine andere Perspektive. Das Motto „Research Data Management: Challenges in a Changing World“ trug den großen Herausforderungen der jüngeren Zeit Rechnung, vor denen auch Forschung und Wissenschaft stehen. Wie werden sich die Themen rund um KI weiterentwickeln, wie wird sich die (geo-)politische Landschaft entwickeln, wie wird die Forschungsdateninfrastruktur künftig finanziert werden? Genau zu diesen Fragestellungen liefert der vorliegende Tagungsband Anleitungen, Diskussionen, Hilfestellungen und Erfahrungsberichte. Dabei wird ein umfangreicher Überblick zum Forschungsdatenmanagement in Deutschland gegeben, indem neue Entwicklungen in den Kernthemen Nachnutzbarkeit, Auffindbarkeit und Langzeitarchivierung behandelt werden. Darüber hinaus wird jedoch auch verstärkt die Rolle von Data Stewards und die Perspektive der Hochschulen für angewandte Wissenschaften (HAWs) in den Fokus gerückt. Abgerundet werden diese Schwerpunkte durch eine große Diversität an Beiträgen zur NFDI, Datenmanagementplänen, sowie Ausarbeitungen zu Datenrecht und Datenethik (u. a.).

Als Herausgeberteam möchten wir gerne auch noch kurz auf die Bedeutung des Forschungsdatenmanagements eingehen, das für unsere Einrichtungen eine zentrale Rolle spielt. Dabei geht es nicht nur um Innovation, sondern um die Unterstützung der For-

schung durch ein adäquates Management von Forschungsdaten. Dieses ist heutzutage eine sine qua non Bedingung für erfolgreiche Forschungsaktivitäten.

Gerne möchten wir schließlich hervorheben, dass diese Konferenz nicht ohne die großzügige finanzielle Unterstützung des Ministeriums für Wissenschaft, Forschung und Kunst Baden-Württemberg möglich gewesen wäre. Dafür danken wir sehr herzlich. Auch all denjenigen, die mit ihren Beiträgen während der Konferenz oder nun mit ihren schriftlichen Beiträgen zum Gelingen dieses Tagungsbandes und des ganzen Formats beigetragen haben, sei herzlichst gedankt. Explizit möchten wir uns bei allen Gutachter*innen für ihr detailliertes Feedback zu allen Einreichungen bedanken – namentlich (aber nicht nur) bei:

- Alexandra Axtmann
- Felix Bach
- Sabine Barthlott
- Martin Baumann
- Isabella Bierenbaum
- Katarzyna Biernacka
- Valerie Boda
- Olaf Brandt
- Michael Derntl
- Matthias Fingerhuth
- Holger Gauza
- Yves Vincent Grossmann
- Alexander Haller
- Sibylle Haßler
- Canan Hastik
- Andreas Hübner
- Dorothea Iglezakis
- Arnd Koeppe
- Claudia Kramer
- Jens Krüger
- Konstantin Kuck
- Matthias Landwehr
- Anne Lauber-Rönsberg
- Jan Leendertse
- Marina Lemaire
- Birte Lindstädt
- Felix Magin
- Brigitte Mathiak
- Maximilian Müller
- Janna Neumann
- Jordy Orellana Figueroa
- Philipp Ost
- Corinna Rebmann
- Gabriel Schneider
- Irene Schumm
- Georg Schwesinger
- Björn Selent
- Michael Selzer
- Renat Shigapov
- Martin Spenger
- Arnost Stanzel
- Vasilka Stoilova
- Thorsten Trippel
- Robert Ulrich
- Dirk von Suchodoletz
- Kerstin Wedlich-Zachodin
- Cord Wiljes
- Larissa Will
- Wolf Zinke

Und nicht zuletzt steht hinter einer Veranstaltung wie den E-Science-Tagen auch ein ganzes Team, das organisiert, hinter den Kulissen gearbeitet und für einen reibungslosen Ablauf gesorgt hat – vielen Dank hierfür!

Wir hoffen, dass dieser Tagungsband den regen wissenschaftlichen Austausch der Konferenz weiterträgt, wünschen Ihnen, dass Sie viele neue Ideen und Anregungen mitnehmen können und freuen uns bereits jetzt auf die Neuauflage der E-Science-Tage im Jahr 2027!

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

Dr. Philipp Kling
Leitung des Organisationsteams der E-Science-Tage 2025

Florian Heuschkel
Mitglied des Organisationsteams der E-Science-Tage 2025

Dr. Sophie G. Habinger
Mitglied des Organisationsteams der E-Science-Tage 2025

Dr. Cora F. Krömer
Mitglied des Organisationsteams der E-Science-Tage 2025

Teil II

Wissenschaftliche Beiträge

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

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

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

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

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

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

Keywords: Research Data Management, Collaborative Research

1 Introduction

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

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

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

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

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

2 Role of Information Infrastructure Projects in CRC

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

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

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

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

3 Key Questions

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

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

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

3.1 Individual Challenges of INF Projects

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 1: Data Management Approaches of CRCs.

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

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

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

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

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

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

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

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

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

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

3.4 Training and Capacity Building

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

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

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

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

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

3.5 Integration of AI and innovative Data Analysis Methods

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

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

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

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

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

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

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

3.6 Sustainability and Long-term Archiving

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

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

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

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

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

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

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

4 Discussion

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

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

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

5 Conclusion and Outlook

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

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

Acknowledgements

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

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

Authorship Contributions

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

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

MaRDMO: Future Gateway to FAIR Mathematical Data

Marco Reidelbach [1]

[1] Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany

Mathematical research data plays a crucial role across scientific disciplines, yet its documentation and dissemination remain challenging due to the lack of standardized research data management practices. The MaRDMO Plugin addresses these challenges by integrating mathematical models, algorithms, and interdisciplinary workflows into the established framework of the Research Data Management Organiser (RDMO). Built on FAIR principles, MaRDMO enables structured documentation and retrieval of mathematical research data through guided questionnaires. It connects to multiple knowledge graphs, including MathModDB, MathAlgoDB, and the MaRDI Portal. Users can document and search for models, algorithms, and workflows via dynamic selection interfaces that also leverage other sources such as Wikidata. The plugin facilitates the export to the individual MaRDI services, ensuring data quality through automated validation. By embedding mathematical research data management into the widely adopted RDMO platform, MaRDMO represents a significant step toward making mathematical research data more findable, accessible, and reusable.

Keywords: Mathematical Research Data, FAIR Data, Knowledge Graph

1 Introduction

Mathematics is often perceived as an abstract discipline, detached from the concept of research data and rarely producing it (Boege et al. 2023). Many mathematicians still assert, "I have no data, I'm just doing math." This perspective poses a challenge for research data management (RDM) at universities and research institutions. Anecdotal evidence also suggests that those responsible for supporting researchers in their RDM efforts often

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23913 (CC BY-SA 4.0).

struggle with how to assist mathematicians in managing and documenting their research data. Unlike in other scientific disciplines, where standards are well established[1], mathematics lacks these for its research data. In some cases, no concrete data management is expected at all[2].

To address these needs, the Mathematical Research Data Initiative, MaRDI (The MaRDI Consortium 2022), as part of the German National Research Data Infrastructure, NFDI (Hartl, Wössner, and Sure-Vetter 2021), has taken significant steps towards defining RDM standards in mathematics. Recognizing the increasing complexity and interdisciplinary potential of mathematical research data, MaRDI has developed services[3], recommendations (The MaRDI Consortium 2023, e. g., on what constitutes mathematical research data, what services, tools, and repositories are available for mathematical research data, and how to write a data management plan in mathematics), and outreach material (Bacher et al. 2023) to support the mathematical research community. At the core of these efforts is the application of the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al. 2016), ensuring that mathematical research data can be systematically documented, shared, and reused. These efforts aim to establish RDM as an integral component of mathematical research, aligning with funding agency requirements[4]. In this context, the German Research Foundation (DFG) published RDM recommendations for mathematics[5] in 2023, explicitly referring to the information provided by the MaRDI consortium.

Mathematical research data is not limited to mathematics but is widely used across scientific disciplines (The MaRDI Consortium 2022). Mathematical models are essential for simulations in engineering and the natural sciences (e.g. Alobaid et al. 2022), algorithms drive computational advances in fields such as medicine and economics (e.g. Albuquerque, Henriques, and Castelli 2025), and interdisciplinary workflows support applications in the humanities (e.g. Kostré et al. 2022). The increasing reliance on data-driven methods (Alexander et al. 2023) across all areas of research highlights the need for structured and standardized approaches to handling mathematical research data. Ensuring that mathematical research data adheres to FAIR principles is not only crucial for mathematics itself but also essential for achieving FAIR research data management in all other disciplines that rely on mathematical research data.

Algorithms, mathematical models, and interdisciplinary workflows are fundamental in most research disciplines. To manage these data types, MaRDI developed dedicated templates and ontologies. For example, the Mathematical Model Database (MathModDB) ontology (Schembera et al. 2023; Schembera et al. 2024; Schembera, Wübbeling, Kleikamp, et al. 2025; Schembera, Shehu, et al. 2025) enables the standardized documentation of

1 https://obofoundry.org/principles/fp-000-summary.html; *Visited on February 25, 2025.*
2 https://www.math.harvard.edu/media/DataManagement.pdf; *Visited on February 25, 2025*
3 https://portal.mardi4nfdi.de/wiki/MaRDI_Services; *Visited on February 25, 2025.*
4 https://rea.ec.europa.eu/open-science_en; *Visited on February 25, 2025.*
5 https://www.dfg.de/resource/blob/176136/42e4fda0ee48534306f2c603780beeae/handreichung-fachkollegium-mathematik-forschungsdaten-data.pdf; *Visited on May 12, 2025.*

mathematical models, including their formulations, involved quantities, related tasks, and modeled problems across various fields. Similarly, the Mathematical Algorithm Database (MathAlgoDB) ontology (MaRDI Task Area 2 – Scientific Computing @ University of Münster 2022) enables the standardized documentation of algorithms, their corresponding problems, implementing software, and relevant benchmarks. At present, the corresponding knowledge graphs[6] contain 123 documented models and 204 documented algorithms. For interdisciplinary workflows, MaRDI developed a standardized documentation template that captures mathematical models, process steps, methods, software, hardware, instruments, and datasets associated with specific research objectives (Boege et al. 2023). These templates are integrated into the MaRDI Portal (Schubotz et al. 2023), a wikibase envisioned as a future one-stop shop for mathematical research data.

Providing researchers from all disciplines with easy access to the aforementioned MaRDI services is essential for fostering the adoption of structured mathematical research data management. The MaRDMO Plugin[7] serves as a crucial gateway by integrating these services into the Research Data Management Organiser, RDMO (Neuroth et al. 2018), an established data management plan software widely used across disciplines (Enke et al. 2023). Through guided interviews, the MaRDMO Plugin enables researchers to document their algorithms, mathematical models (Reidelbach, Schembera, and Weber 2024), and interdisciplinary workflows (Reidelbach, Ferrer, and Weber 2023) in a standardized form and publish them directly in the respective MaRDI services, ensuring efficient dissemination. Additionally, researchers can query these services to re-use existing mathematical models, algorithms, workflows, or parts thereof for their own research. By embedding MaRDMO within RDMO, institutions can seamlessly incorporate it into their existing infrastructure, allowing RDM staff – who are often already familiar with RDMO – to support mathematicians and researchers from other disciplines working with mathematical research data effectively. To further extend accessibility, MaRDI will also provide its own RDMO instance, provided through the basic services DMP4NFDI (Diederichs et al. 2024), ensuring that researchers without institutional access to RDMO can still benefit from these services.

The remaining paper introduces the key functionalities of the MaRDMO Plugin, followed by a discussion of its future potential and development directions, with an emphasis on its role in advancing research data management in mathematics.

2 The MaRDMO Plugin

In the following section, the MaRDMO Plugin is introduced as a tool for documenting and querying mathematical models, algorithms, and interdisciplinary workflows through

6 https://mtsr2024.m1.mardi.ovh/ and https://mathalgodb.mardi4nfdi.de/; *Visited on February 27, 2025.*

7 https://github.com/MarcoReidelbach/MaRDMO-Plugin; *Visited on March 30, 2025.*

distinct questionnaires. To assist researchers in their documentation efforts, the individual questionnaires are semi-automated and linked to several knowledge graphs. All this follows an underlying data model, which is introduced first.

2.1 Data Model

Interdisciplinary research follows a structured process when addressing real-world problems, defining research objectives and employing various approaches to derive answers. These workflows can vary significantly depending on the research context. Some may be purely computational, where a problem is simulated based on a mathematical model and propagated using algorithms. Others may involve experimental components, where empirical data is collected, fed into a model, and subsequently analyzed through computational methods and some may follow data-driven or rule-based methodologies, such as statistical inference or heuristic optimization, without relying on an explicit mathematical model. Overall, however, data, methods (computational or experimental), mathematical models, instruments, software, and hardware are key aspects of these workflows but are not necessarily present in all cases.

MaRDI's interdisciplinary workflow documentation template structures these workflows into distinct processing steps, a concept closely related to Metadata4Ing (Iglezakis et al. 2023), where input data is transformed into output data through computational or experimental methods. Each step is further associated with relevant tools, such as software or instruments, and is assigned to specific research fields or mathematical areas. This structured approach enables precise documentation of how individual data sets are modified and how they flow across disciplines.

To integrate interdisciplinary workflows into the MaRDI Portal, the MaRDMO Plugin leverages, where possible, existing items from both the MaRDI Portal and Wikidata (Vrandečić and Krötzsch 2014). For example, existing items such as *Research Workflow*[8] or *Mathematical Model*[9] serve as classes in the data model for interdisciplinary workflows. Notably, an item describing a general *Processing Step*[10] did not exist and was subsequently added to the MaRDI Portal. Instances of these classes are related by properties such as *uses*[11] or *platform*[12]. Figure 1 (green boxes) provides a schematic representation of the integration of interdisciplinary workflows. Note that, for clarity, only the main aspects are shown. A complete list of employed items and properties can be found in the MaRDMO documentation[13].

8 https://www.wikidata.org/wiki/Q115682347; *Visited on March 7, 2025.*
9 https://www.wikidata.org/wiki/Q486902; *Visited on March 7th, 2025.*
10 https://portal.mardi4nfdi.de/wiki/Item:Q6486603; *Visited on March 20, 2025.*
11 https://www.wikidata.org/wiki/Property:P2283; *Visited on March 7, 2025.*
12 https://www.wikidata.org/wiki/Property:P400; *Visited on March 7, 2025.*
13 https://portal.mardi4nfdi.de/wiki/MaRDMO; *Visited on March 20, 2025.*

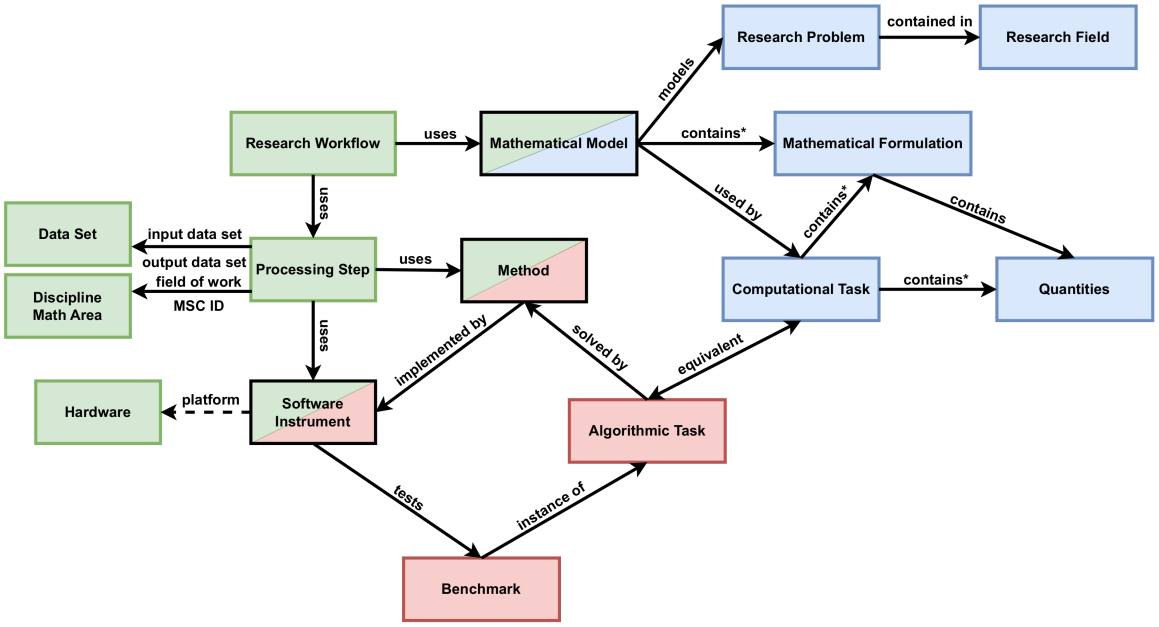


Figure 1: Data model employed by the MaRDMO Plugin. Nodes represent classes, while labeled, directed edges indicate relations between them. Classes relevant to algorithms, mathematical models, and interdisciplinary workflows are depicted in red, blue, and green, respectively. Two-colored nodes signify relevance to multiple data types. The *Method* class is shared between interdisciplinary workflows and algorithms, but only methods classified as algorithms are considered part of the algorithmic domain. *Software* and *Instruments* are distinct classes, combined here to parallel the *Tool* class in Metadata4Ing; however, only *Software* belongs to the algorithmic domain and is linked to the *Hardware* class. An asterisk in the edge label denotes the existence of multiple distinct relations. The publication class relevant for algorithms, mathematical models and interdisciplinary workflows is omitted for clarity.

As discussed earlier, mathematical models form a crucial component of many interdisciplinary workflows, providing the theoretical foundation for computational analyses and simulations. To ensure proper documentation of these workflows, the MaRDMO Plugin integrates the MathModDB ontology. This enables a standardized description of models, including their formulations, quantities, related tasks, and modeled research problems across various fields (c.f. Figure 1, blue boxes).

Both workflows and models are connected to research fields and problems. While this might seem like a duplication, the associated fields and problems can differ. For instance, a workflow could focus on testing different computational methods for a particular model, making the distinction necessary for accurate documentation.

MaRDMO further expanded to include the MathAlgoDB ontology, which provides a standardized description of algorithms (c.f. Figure 1, red boxes). By linking algorithms to their implementing software, the algorithmic problems they address, and related bench-

marks, MathAlgoDB enhances the overall documentation of the computational methods employed. Connecting these algorithms to the broader workflow context, MaRDMO now offers a more complete picture of the applied methods and their underlying computations. Furthermore, the computational tasks of MathModDB and the algorithmic problems of MathAlgoDB are interconnected, and MaRDMO can leverage this connection.

The MathModDB and MathAlgoDB ontology also support relations between items of the same class, not shown in Figure 1 but listed in the MaRDMO documentation, such as linking similar models or subclasses of algorithms. All these relations enhance the findability of algorithms, mathematical models and interdisciplinary workflows.

By the time of this writing, interdisciplinary workflows, mathematical models, and algorithms are documented in distinct knowledge graphs. Federated SPARQL queries enable the implementation of the data model across the individual knowledge graphs. This setup, however, is likely to evolve in the future.

2.2 Questionnaires and Functionalities in MaRDMO

To facilitate the structured documentation of mathematical models, algorithms, and interdisciplinary workflows, the MaRDMO Plugin provides distinct questionnaires. These questionnaires guide researchers through the documentation process by collecting relevant metadata in a standardized format. Each questionnaire aligns closely with the corresponding ontology, ensuring consistency with MathModDB, MathAlgoDB, and the interdisciplinary workflow documentation template. The questionnaires vary in complexity, with different numbers of sections and questions depending on the type of research data being documented. Despite these differences, all questionnaires share a common structure that facilitates a streamlined and user-friendly documentation process.

Each section of a questionnaire consists of one or more pages, which follow a structured format. The first step on either page involves selecting an appropriate item from an external source, e. g. for the documentation of a mathematical model, such an item could be the model itself, a research field, a research problem, a computational task, a quantity, or a mathematical formulation. For models, the preferred external source is MathModDB, while for algorithms, the primary reference is MathAlgoDB. If an item is not available in these sources, users can search alternative databases, including the MaRDI Portal or Wikidata. When an item from a lower-priority source is selected, MaRDMO automatically verifies whether it exists in a higher-priority source, ensuring consistency and minimizing redundancy. Item selection is supported by *Dynamic Optionset Providers*, which query the respective sources via API or SPARQL, presenting results in an autocomplete fashion (c. f. Figure 2). Currently, most providers support a subset of four sources: MathModDB, MathAlgoDB, the MaRDI Portal, and Wikidata, but additional sources can be incorporated as needed.

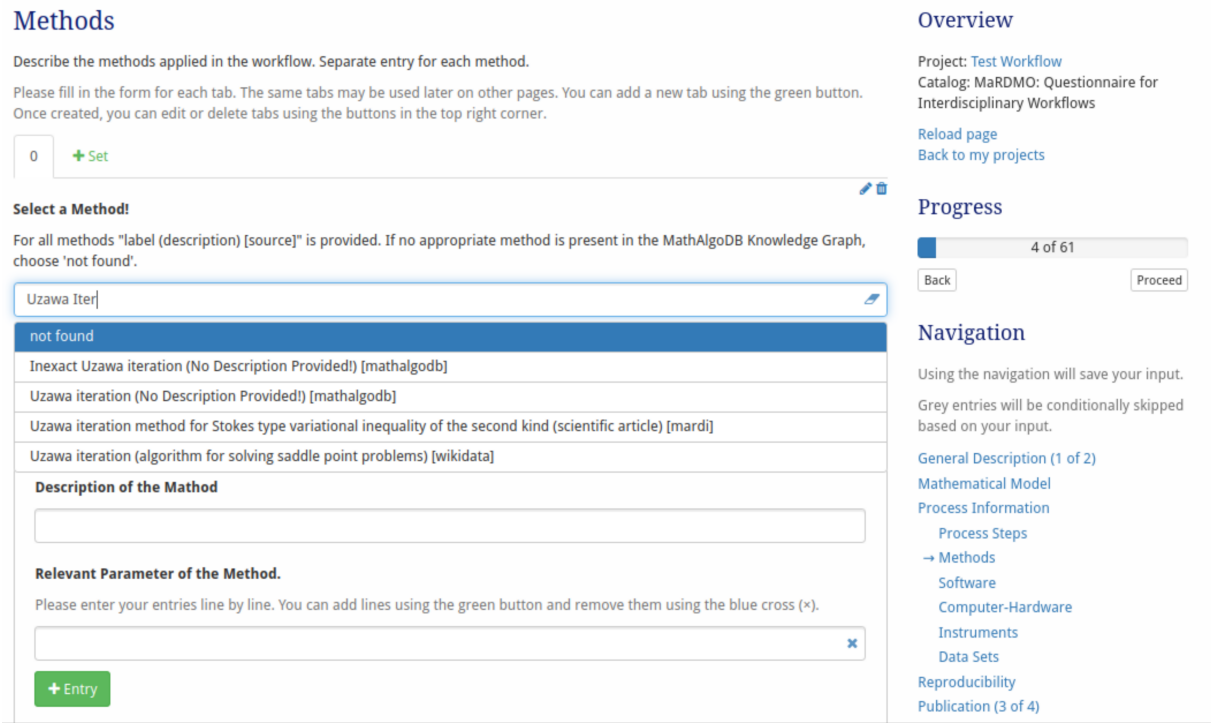


Figure 2: RDMO user interface with the *Methods* page of the interdisciplinary workflow catalog. *Uzawa Iteration* is searched, corresponding items are found in MathAlgoDB, the MaRDI Portal and Wikidata.

Following the selection, the next step on the page presents an information box containing key details about the selected item. The content of this box depends on the item's class. For instance, a research field entry may include only the name and description, whereas a mathematical formulation entry contains additional details such as properties, formulas, symbols, and encoded quantities. If an existing item is chosen, MaRDMO queries the respective source for further details and automatically populates the information box. If no suitable item is found, users must manually provide the necessary information.

In addition to describing the selected item, the information box also displays connections to items of other classes, e.g a *Research Problem* is contained in a *Research Field*. As for the pure item information, described before, MaRDMO adds existing relations automatically to the questionnaire. If a relation to an item from another class is established, MaRDMO checks whether the corresponding item already exists in the respective section of the questionnaire, e.g. if the *Research Field* containing the *Research Problem* is already part of the *Research Field* section. If it does, no further action is taken; otherwise, the related item is added to the appropriate section, and all available metadata is incorporated into the questionnaire. Dedicated automation schemes (c.f. Figure 2) define which data is added, how, and when, ensuring consistency and minimizing redundancy. This automation enhances efficiency and supports researchers by reducing manual input.

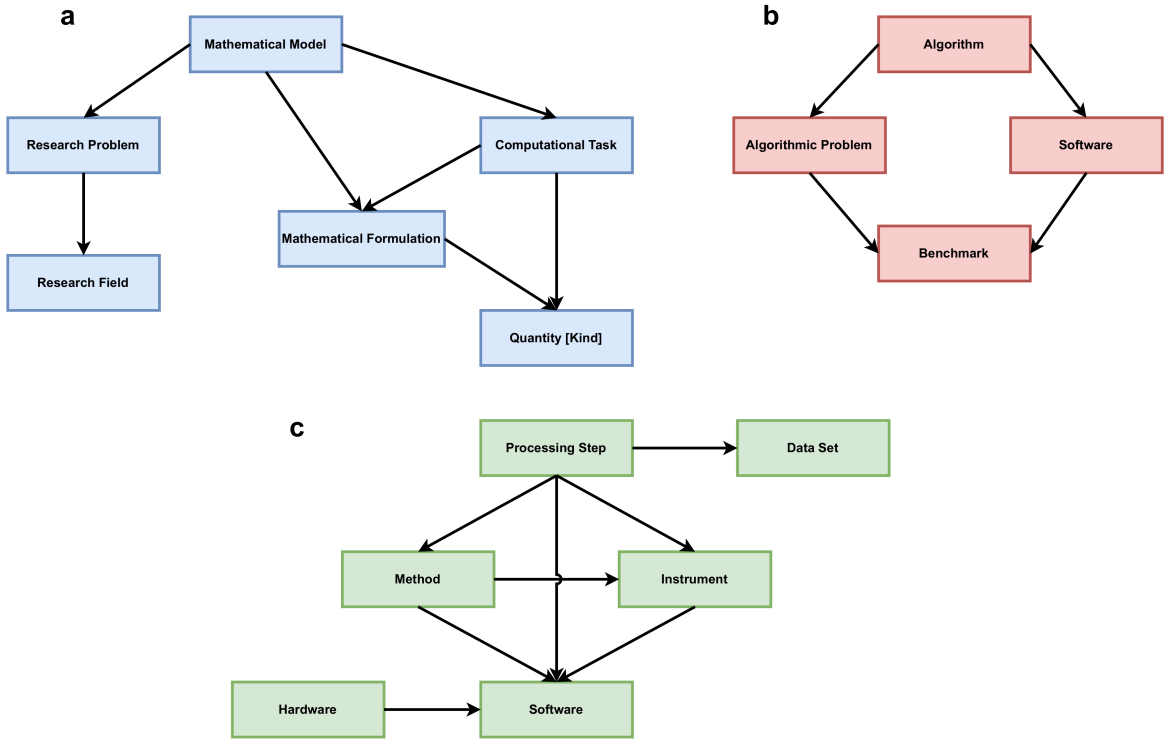


Figure 3: Automating schemes within the catalogs for (a) mathematical models, (b) algorithms (b) and (c) interdisciplinary workflows displaying the individual classes of the corresponding ontologies. The selection of items from either class leads to an automatic insertion of all downstream items. For clarity, the publication class, which is a downstream class of all other classes, has been omitted.

At the end of the pages, additional questions allow users to specify relationships between items of the same class. These include connections such as stating that one item *generalizes*, *approximates*, or *is similar to* another. Again, MaRDMO pre-fills these fields where possible based on the information from the external sources. Since defining such relationships can be challenging for individual researchers these questions are optional and it is envisioned that domain experts will contribute to refining these connections through later curation efforts.

It is important to note that all three documentation catalogs share a common publication section. The main difference lies in how individual publications are linked to other items[14]. To accommodate this, MaRDMO provides a dedicated page in the questionnaire, allowing users to select existing publications from MathAlgoDB, MathModDB, the MaRDI Portal, or Wikidata. If an appropriate publication is found, its DOI or URL is added to the questionnaire. If no suitable publication is available in these sources, users can manually provide a DOI or URL. When a DOI is entered, MaRDMO automatically checks the four

14 https://portal.mardi4nfdi.de/wiki/MaRDMO; *Visited on March 19, 2025.*

aforementioned sources, as well as the Crossref API[15], the DataCite API[16], the DOI API[17], the zbMath API[18], and the ORCiD API[19] for the complete metadata of the publication and its authors.

Next to the documentation catalogs, MaRDMO provides a dedicated search catalog that allows users to explore existing entities. Users can search for software, methods, or mathematical models and discover interdisciplinary workflows that utilize these entities. Similarly, mathematical models can be retrieved based on their application in specific tasks or their role in modeling research problems. Algorithms can be found based on the problems they solve or the software in which they are implemented. Additionally, keyword-based searches enable users to explore interdisciplinary workflows via research objectives, models via research problems, and algorithms via algorithmic problems. This search functionality provides an accessible interface for querying the individual MaRDI services while abstracting away the complexity of SPARQL.

2.3 MaRDMO Export

The MaRDMO Plugin integrates into the RDMO interface by adding a "MaRDMO Button" to the export section. This button enables researchers to export their documentation but is only functional if one of the MaRDMO catalogs has been used.

The export process follows a structured workflow. Once a documentation process is completed – or search parameters are selected – the export function first provides a preview. This preview presents the collected information in the standardized documentation template for interdisciplinary workflows, mathematical models, or algorithms, depending on the type of documentation. In the case of the search catalog, the SPARQL query generated by MaRDMO is displayed, allowing users to inspect the exact query executed on their behalf. While generating the preview, relevant publication information is retrieved if DOIs are provided. The thus obtained citation information is stored in the background for the final export.

If the documentation is considered to be correct, the user can proceed by clicking "Export to MaRDI Service," which triggers the final export. The export mechanism differs based on the type of research data. For mathematical models and algorithms, MaRDMO generates SPARQL INSERT queries to add the new information into the MathModDB or MathAlgoDB knowledge graphs. For interdisciplinary workflows, the export is handled via the REST API of the MaRDI Portal. This process follows an OAuth2 authentication protocol, allowing users to log in to the MaRDI Portal using their institutional credentials.

15 https://api.crossref.org/works/; *Visited on March 19, 2025.*
16 https://api.datacite.org/dois/; *Visited on March 19, 2025.*
17 https://citation.doi.org/metadata?doi=; *Visited on March 19, 2025.*
18 https://api.zbmath.org/v1/document/; *Visited on March 19, 2025.*
19 https://pub.orcid.org/v3.0; *Visited on March 19, 2025.*

The workflow documentation is then added to the MaRDI Portal under the user's name. While MaRDMO adds machine-readable information[20] to the knowledge graphs, human-readable representations[21] are automatically generated through structured templates[22] on each platform.

To maintain a certain level of data quality, a two-step validation process is envisioned. First, MaRDMO ensures automated ontological validation before export, verifying that all relationships follow the structure of the respective ontologies (MathModDB, MathAlgoDB, and interdisciplinary workflow) and checking whether the documentation is complete and connected. A documentation is considered complete and connected if it establishes all necessary relationships. However, users are not required to provide all related information themselves. For example, if a user only wants to add a new computational task, it must be ensured that the task is connected to a mathematical model, mathematical formulations and quantities as indicated in the automating scheme (c. f. Figure 2). However, connected and complete does not mean that relations within the individual classes need to be established. The second step of validation is a review by domain experts, which is carried out by the individual knowledge graph providers to ensure the accuracy and relevance of the content.

Once the export is completed, MaRDMO presents a landing page with direct links to the newly created items in the respective knowledge graphs. Additionally, MaRDMO updates the filled-out questionnaire. For newly created items in either knowledge graph, MaRDMO automatically appends the persistent identifiers (PIDs) to the questionnaire for future reference. For search queries, MaRDMO provides a summary of the SPARQL results, allowing a direct access of the relevant workflows, models, or algorithms.

3 Future Potential and Developments

The MaRDMO Plugin has the potential to significantly enhance the dissemination of mathematical research data across disciplines. Traditionally, the transfer of mathematical models and algorithms into new fields relies heavily on personal connections and interdisciplinary collaborations. While these collaborations are valuable, they also introduce limitations – especially when researchers are unaware of potential applications of their work in other domains.

For example, consider a mathematician who develops a model and an algorithm to extract logical rules from a dataset by transforming (data) objects into Boolean polynomials and computing the Gröbner basis of the corresponding ideal. If this work is only published

20 https://portal.mardi4nfdi.de/wiki/Item:Q6032641; *Visited on March 25, 2025.*
21 https://portal.mardi4nfdi.de/wiki/Workflow:6032641; *Visited on March 25, 2025.*
22 https://portal.mardi4nfdi.de/wiki/Template:Workflow; *Visited on March 25, 2025.*

in a mathematics journal, it is unlikely to reach researchers in fields such as Egyptology, where it might be applicable for analyzing destruction patterns in ancient egyptian statues. Without an accessible and structured way to describe and search for such models, their interdisciplinary use remains restricted. MaRDMO addresses this challenge by enabling researchers to search for models and algorithms based on their functionality – rather than requiring detailed knowledge of mathematical theory in the first place. By enabling structured search, MaRDMO helps researchers efficiently locate concise descriptions, relevant metadata, and implementation resources – provided by individual MaRDI services – thereby lowering the barrier for non-mathematicians to explore and adopt mathematical research data. Additionally, researchers from non-mathematical disciplines can document their own interdisciplinary workflows – such as applying logical data analysis to ancient egyptian objects – allowing these workflows to be discovered and reused in new contexts, such as medical research.

Beyond facilitating discoverability, MaRDMO also enhances research data management through PIDs. When researchers document their models, algorithms, or workflows in MaRDI services, they receive a unique PID, which can be referenced in publications. This improves the findability of research outputs and provides a quality assurance mechanism, as data integrated into MaRDI services follows structured documentation guidelines and validation. Such PIDs can also be added to DMPs to enhance the data provenance section.

Looking ahead, MaRDMO will expand to support additional MaRDI services, such as MaRDIFlow (Veluvali, Heiland, and Benner 2023). A key focus will also be on integrating additional sources to further enrich interdisciplinary documentation. For example, when documenting a workflow which includes experimental measurements through a specific measurement device, MaRDMO will retrieve relevant metadata from databases maintained by other NFDI consortia or domain-specific repositories[23]. This integration will be facilitated in collaboration with the DMP4NFDI basic service, allowing researchers to seamlessly incorporate external data sources into their documentation.

Another ongoing initiative is the development of a dedicated MaRDI RDMO instance, in collaboration with DMP4NFDI. This will provide researchers from all disciplines with direct access to MaRDMO, even if their institution does not offer an RDMO instance.

Finally, MaRDMO is exploring the potential use of Artificial Intelligence (AI) to assist in the documentation process. One potential approach is to use AI to analyze research papers and map their content onto existing ontologies, automatically identifying relevant aspects of mathematical models, algorithms, and interdisciplinary workflows. This could significantly reduce the manual effort required for documentation and improve the integration of research outputs into MaRDI services. AI-driven techniques could also help establish new connections between mathematical models and potential application areas, further supporting the interdisciplinary diffusion of mathematical research data.

23 https://o2a-data.de/; *Visited on March 20, 2025.*

Acknowledgements

M. Reidelbach is supported by MaRDI, funded by the Deutsche Forschungsgemeinschaft (DFG), project number 460135501, NFDI 29/1 “MaRDI – Mathematische Forschungsdateninitiative”.

Authorship Contributions

M. Reidelbach contributed to the conceptualization, tool development, and writing (first draft, review and editing).

Conflict of Interest

The author declares that there are no conflicts of interest.

Bibliography

Albuquerque, Carina, Roberto Henriques, and Mauro Castelli. 2025. “Deep learning-based object detection algorithms in medical imaging: Systematic review”. *Heliyon* 11 (1). https://doi.org/10.1016/j.heliyon.2024.e41137.

Alexander, Francis J., Meifeng Lin, Xiaoning Qian, and Byung-Jun Yoon. 2023. “Accelerating scientific discoveries through data-driven innovations”. *Patterns (NY)* 4 (11). https://doi.org/10.1016/j.patter.2023.100876.

Alobaid, Falah, Naser Almohammed, Massoud Massoudi Farid, Jan May, Philip Rößger, Andreas Richter, and Bernd Epple. 2022. “Progress in CFD Simulations of Fluidized Beds for Chemical and Energy Process Engineering”. *Progress in Energy and Combustion Science* 91. https://doi.org/10.1016/j.pecs.2021.100930.

Bacher, Tabea, Christiane Görgen, Tabea Krause, Andreas Matt, Daniel Ramos, and Bianca Violet. 2023. “Spreading the Love for Mathematical Research Data”. In *Proceedings of the Conference on Research Data Infrastructure,* volume 1. 1. TIB Open Publishing. https://doi.org/10.52825/cordi.v1i.369.

Boege, Tobias, René Fritze, Christiane Görgen, Jeroen Hanselman, Dorothea Iglezakis, Lars Kastner, Thomas Koprucki, et al. 2023. “Research-data management planning in the German mathematical community”. *European Mathematical Society Magazine,* number 130, 40–47. https://doi.org/10.4171/MAG/152.

Diederichs, Katja, Konrad Förstner, Daniela Hausen, Gerald Jagusch, Jochen Johannsen, Birte Lindstädt, Dominik Schmitz, Sabine Schönau, Thomas Stäcker, and Jürgen Windeck. 2024. *DMP4NFDI – NFDI Basic Service for Data Management Plans.* Zenodo. https://doi.org/10.5281/zenodo.13445355.

Enke, Harry, Daniela Hausen, Christin Henzen, Gerald Jagusch, Celia Krause, Sabine Schönau, Lukas Weimer, and Jürgen Windeck. 2023. *Data management planning: Concept for Setting up a Working Group in the NFDI Section Common Infrastructures.* Zenodo. https://doi.org/10.5281/zenodo.7540682.

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

Iglezakis, Dorothea, Marc Fuhrmans, Giacomo Lanza, Džulia Terzijska, Susanne Arndt, Sophia Leimer, and Johanna Hickmann. 2023. “Modelling Scientific Processes With the m4i Ontology”. In *Proceedings of the Conference on Research Data Infrastructure,* volume 1. https://doi.org/10.5281/ZENODO.8348784.

Kostré, Margarita, Vikram Sunkara, Christof Schütte, and Nataša Djurdjevac Conrad. 2022. “Understanding the romanization spreading on historical interregional networks in Northern Tunisia”. *Applied Network Science* 7 (53). https://doi.org/10.1007/s41109-022-00492-w.

MaRDI Task Area 2 – Scientific Computing @ University of Münster. 2022. *MathAlgoDB – Algorithm Knowledge Graph – Ontology (Version 0.1).* Visited on July 2, 2025. https://mathalgodb.mardi4nfdi.de/static/widoco/v1/index-en.html.

Neuroth, Heike, Claudia Engelhardt, Jochen Klar, Jens Ludwig, and Harry Enke. 2018. “Aktives Forschungsdatenmanagement”. *ABI Technik* 38 (1): 55–64. https://doi.org/10.1515/abitech-2018-0008.

Reidelbach, Marco, Eloi Ferrer, and Marcus Weber. 2023. “MaRDMO Plugin – Document and Retrieve Workflows Using the MaRDI Portal”. In *Proceedings of the Conference on Research Data Infrastructure,* volume 1. TIB Open Publishing. https://doi.org/10.52825/cordi.v1i.254.

Reidelbach, Marco, Björn Schembera, and Marcus Weber. 2024. “Towards a FAIR Documentation of Workflows and Models in Applied Mathematics”. *Lecture Notes in Computer Science* 14749:254–262. https://doi.org/10.1007/978-3-031-64529-7_27.

Schembera, Björn, Aurela Shehu, Burkhard Schmidt, Christine Biedinger, Jochen Fiedler, Marco Reidelbach, and Thomas Koprucki. 2025. *MathModDB – An Ontology and Knowledge Graph for Mathematical Models.* Zenodo. https://doi.org/10.5281/zenodo.14887915.

Schembera, Björn, Frank Wübbeling, Hendrik Kleikamp, Christine Biedinger, Jochen Fiedler, Marco Reidelbach, Aurela Shehu, et al. 2024. "Ontologies for Models and Algorithms in Applied Mathematics and Related Disciplines". In *Metadata and Semantic Research,* 2048:161–168. Springer Nature Switzerland. ISBN: 9783031659904. https://doi.org/10.1007/978-3-031-65990-4_14.

Schembera, Björn, Frank Wübbeling, Hendrik Kleikamp, Burkhard Schmidt, Aurela Shehu, Marco Reidelbach, Christine Biedinger, et al. 2025. "Towards a Knowledge Graph for Models and Algorithms in Applied Mathematics". *Communications in Computer and Information Science* 2331:95–109. https://doi.org/10.1007/978-3-031-81974-2_8.

Schembera, Björn, Frank Wübbeling, Thomas Koprucki, Christine Biedinger, Marco Reidelbach, Burkhard Schmidt, Dominik Göddeke, and Jochen Fiedler. 2023. "Building Ontologies and Knowledge Graphs for Mathematics and its Applications". In *Proceedings of the Conference on Research Data Infrastructure,* volume 1. TIB Open Publishing. https://doi.org/10.52825/cordi.v1i.255.

Schubotz, Moritz, Eloi Ferrer, Johannes Stegmüller, Daniel Mietchen, Olaf Teschke, Larissa Pusch, and Tim OF Conrad. 2023. *Bravo MaRDI: A Wikibase Powered Knowledge Graph on Mathematics.* https://doi.org/10.48550/arXiv.2309.11484.

The MaRDI Consortium. 2022. *MaRDI: Mathematical Research Data Initiative proposal.* Zenodo. https://doi.org/10.5281/zenodo.6552436.

———. 2023. *Research Data Management Planning in Mathematics.* Zenodo. https://doi.org/10.5281/zenodo.10018246.

Veluvali, Pavan L., Jan Heiland, and Peter Benner. 2023. "MaRDIFlow: A Workflow Framework for Documentation and Integration of FAIR Computational Experiments". In *Proceedings of the Conference on Research Data Infrastructure,* volume 1. TIB Open Publishing. https://doi.org/10.52825/cordi.v1i.323.

Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledgebase". *Communications of the ACM* 57 (10): 78–85. https://doi.org/10.1145/2629489.

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. https://doi.org/10.1038/sdata.2016.18.

Data-Driven Community Standards for Interdisciplinary Heterogeneous Information Networks

Mattis thor Straten [1,*], Steffen Strohm [1], Florian Thiery [2], Matthias Renz [1]

[1] Department of Computer Science, Kiel University;
[2] Leibniz-Zentrum für Archäologie & Research Squirrel Engineers;
[*] Corresponding author: mts@cs.uni-kiel

Creating a unified, interoperable representation by integrating diverse datasets is a key challenge in interdisciplinary research. Heterogeneous information networks (HINs) offer a graph-based approach to linking datasets while preserving their semantic structure. This study examines data-driven community standards, with a particular focus on ontologies, to ensure semantic interoperability in cultural heritage HINs.

Using the CIDOC Conceptual Reference Model (CIDOC-CRM) alongside other commonly used or domain-specific ontologies, this study develops the hybrid ArNO ontology for archaeo-natural research data based on the experience gained in NFDI4Objects Task Area 3. The data integration follows Linked Open Data principles to ensure schema-level consistency through ontology-based modelling, while enforcing data-level consistency through terminologies. This work contributes to the development of FAIR-compliant research infrastructures by establishing standards at schema and data level, thereby enhancing the reuse of data across the humanities and natural sciences.

This approach is demonstrated by integrating ancient DNA datasets from the Poseidon framework to enable cross-dataset analysis by linking natural, archaeological, and contextual data. The study also addresses challenges in CIDOC-CRM-based modelling, such as its event-centric nature.

Keywords: Data Integration, Ontologies, Linked Open Data, Heterogeneous Information Networks, Knowledge Graphs

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23914 (CC BY-SA 4.0).

1 Introduction

Interdisciplinary research requires disparate datasets to be integrated into a unified representation, enabling holistic, cross-dataset analyses to uncover hidden patterns and dependencies. This provides a foundation of combined data on which network analysis methods can be applied to derive new insights and solve problems – the second generation of semantic cultural heritage (CH) systems (Hyvönen 2019). Heterogeneous information networks (HINs) provide a flexible, graph-based framework for integrating diverse datasets (e.g. Patel, Paraskevopoulos, and Renz 2018a). They are particularly effective at preserving both the complex relationships of the data and the contextual information in an intuitively understandable graph structure consisting of typed nodes and edges.

Interdisciplinary initiatives related to research data management, such as the German National Research Data Infrastructure (NFDI, see Hartl, Wössner, and Sure-Vetter 2021), are creating knowledge graphs using ontologies, reference models and standards such as the Resource Description Framework to promote FAIR (Wilkinson et al. 2016) research data management and the NFDI vision of applying open science principles. Several domain-focused consortia within the NFDI are working on the interdisciplinary task of FAIRifying research data related to the material remains of human history. All resulting domain-specific knowledge graphs must be interoperable using the Base4NFDI Basic Services[1], such as TS4NFDI and KGI4NFDI. The NFDI4Objects consortium (N4O, see Rummel, Keller, and Fricke 2025; Thiery et al. 2023) addresses these challenges by fostering interdisciplinary collaboration across the humanities, cultural and natural sciences. The initiative covers data spanning approximately 2.6 million years and encompasses various types of objects, including human-made artifacts, biological specimens, and geological specimens. These objects evolve through various transformations, resulting in a so-called object biography, with Task Area 3 "Analytics and Experiments" (TA3) taking the lead in natural science research data. A key challenge in this process is the interdisciplinarity within the CH and archaeological domains (e.g., Strohm, Buenning, and Renz 2023), and the challenges of joint semantic modelling with the humanities, natural and life sciences, and geosciences.

This work presents the Archaeo-Natural Ontology (ArNO), a hybrid ontology for archaeo-natural research data that is developed within TA3. Archaeo-natural data refers to research data produced at the intersection of archaeology and the natural sciences, including anthropology, archaeogenetics, archaeozoology and archaeobotany, among others. The term highlights the necessity of integrative modelling approaches that can represent both cultural-historical context and scientific measurements within a unified semantic framework. ArNO is an application ontology that incorporates classes and properties from multiple foundational (Steller et al. 2025, p. 11) and domain ontologies (Guarino 1998, p. 8). It provides a self-contained schema that is suitable for representing interdisciplinary archaeo-natural research data in an HIN. It is also designed to integrate with existing in-

1 https://base4nfdi.de/projects; *Visited on May 20, 2025.*

frastructures modelled using the contained ontologies, such as the CIDOC-CRM-based NFDI4Objects Knowledge Graph (e.g., Voß et al. 2024; Voß and Heers 2024).

The aim of this work is to integrate anthropological ancient DNA (aDNA) data from the Poseidon framework (Schmid et al. 2024) into a generalisable format that allows it to be combined with other archaeo-natural and CH data. Therefore, the data representation must transcend the boundaries of the dataset, relating the information it contains to a wider context by matching different instances of a common object (Doerr 2005), such as spatial and temporal information or involved agents.

Using HINs for data integration enables cross-dataset analysis (e.g., Patel, Paraskevopoulos, and Renz 2018b) through common concepts, such as spatial and temporal information, which act as connectors between objects of different origins. Incorrect data integration results in invalid HINs, which may consequently lead to incorrect analysis results. This creates the need for standards at schema and data level to ensure semantically correct data integration. The proposed interoperable ontology provides these standards at the schema level. At the data level, terminologies – from keywords to enriched thesauri – are employed (Thiery and Engel 2016, pp. 259-261).

2 Preliminaries

2.1 Data Integration

In the context of this work, data integration refers to the fusion of datasets at the record level. This involves synchronising different representational instances of the same object in different datasets to create a consistent, interconnected data foundation under a mediated schema. The main challenges are finding corresponding schema elements (Legler and Naumann 2007) and equivalent object descriptions (duplicate detection) in different datasets, in order to create a single, consistent representation of the integrated data from both datasets (Bleiholder and Naumann 2008, p. 2).

2.2 Information Networks

An information network is a directed graph $G = (V, E, \phi, \psi)$, where the set of nodes V represents information objects – also known as entities – and the set of edges E represents relationships between pairs of these information objects – also known as relationships or properties. The functions ϕ and ψ map nodes and edges to their corresponding types, i.e., $\phi : V \longrightarrow \mathcal{A}$ and $\psi : E \longrightarrow \mathcal{R}$, where the set $\mathcal{A}$ contains all object types and the set $\mathcal{R}$ contains all relationship types. When more than one node or edge type is considered,

i.e., $|\mathcal{A}| > 1$ or $|\mathcal{R}| > 1$, this information network is called a heterogeneous information network (HIN, see Sun et al. 2011). HINs are represented using task-specific formats, e.g., by using adjacency matrices for efficient querying.

The network schema of an HIN G, denoted as $T_G = (\mathcal{A}, \mathcal{R})$, is a directed graph containing the object types $\mathcal{A}$ as labelled nodes (Sun et al. 2011). Two nodes are connected by a labelled edge for each relationship type in $\mathcal{R}$ that connects them. An example of a simple network schema containing nodes of the types *Sample*, *Site*, *Country* and *Time Period*, as well as the relationship types *found in*, *in country* and *dated as*, can be seen in Figure 1.

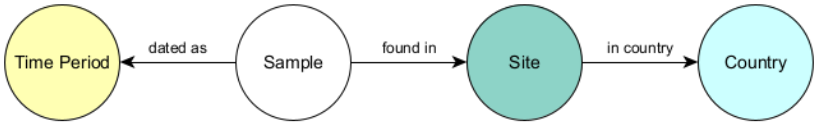


Figure 1: Network schema illustrating relationships among archaeological entities.

HINs are closely related to knowledge graphs (KGs), which are networks designed to store and convey knowledge about the real world. Their nodes represent entities of interest, while their edges capture the relationships between these entities. The underlying network conforms to a graph-based model, such as a directed edge-labelled graph, e.g., RDF (Cyganiak, Hyland-Wood, and Lanthaler 2014), or property graphs (Rodriguez and Neubauer 2010), e.g., Neo4J. Knowledge, which is defined as what is known, can be derived from external sources or inferred from the graph itself. It can consist of simple factual statements or more complex quantified assertions. To represent and reason about such structured knowledge, more expressive formalisms, such as ontologies or rules, can be employed (Hogan et al. 2021). Sun et al. (2022) state that KGs are special cases of HINs with a much richer schema than most HINs considered, and that they can add a hierarchical structure to the flat structure of an HIN at the schema or object instance level. Shi et al. (2017) state that KGs can be considered as HINs. They also note that KGs are too complex to be modelled as an HIN with a simple network schema because they are schema-rich networks that cannot be represented by a simple network schema. Similarly, Zheng, Qu, and Yang (2024) describe KGs as large-scale HINs.

2.3 Data Modelling

The Resource Description Framework (RDF, see Cyganiak, Hyland-Wood, and Lanthaler 2014) represents graph data in the form of subject-predicate-object triples, where the predicate represents a relationship linking the subject to the object. Internationalised Resource Identifiers (IRIs)[2] are the standard way for uniquely identifying entities in RDF.

2 https://datatracker.ietf.org/doc/html/rfc3987; *Visited on May 20, 2025.*

An ontology is a knowledge organisation system (Zeng 2008) that is often defined from the perspective of a specific domain and specifies terms by describing their relationships with other terms (Hitzler et al. 2012). It is a structured model of the scientific research process that defines general concepts, their internal structure and the relationships between them. Unlike a taxonomy, which organises entities hierarchically, an ontology explicitly represents relationships, enabling semantic networks for exploratory analysis and inference (Hughes, Constantopoulos, and Dallas 2015, p. 15). In computer and information science, an ontology is a technical artefact that describes a particular reality by defining a case-specific vocabulary consisting of explicit assumptions about the intended meaning of words (Guarino 1998, p. 4).

The Web Ontology Language (OWL) builds on RDF by providing a more expressive formalism based on Description Logic, which enables reasoning capabilities. It is used to formally represent ontologies, allowing the definition of object types (classes), relationships between object types (object properties) and associations between objects and literals (data properties) in a machine-interpretable way (Hitzler et al. 2012; Motik, Patel-Schneider, and Parsia 2012). In the context of Linked Open Data (LOD), which aims to provide openly available and interlinked online datasets that are ready for cross-querying (Schmidt, Thiery, and Trognitz 2022), Berners-Lee's (2006) linked data principles model are a well-known guideline.

2.4 Introducing the Used Ontologies

The schema-level ontology for archaeo-natural data, presented in Figure 2, contains seven top-level concept domains: (A) Cultural Heritage as the central domain within N4O, (B) the Application Domain (here, Biology is the first use case), the trinity of (C) Time, (D) Space and (E) Agent, as well as the metadata-focused entities (F) Bibliography and (G) Provenance. Note that the concept domains are intertwined. In bibliographic information, for example, agents are involved as authors, while time is represented by the year of publication.

(A): The CIDOC Conceptual Reference Model (CIDOC-CRM, Version 7.1.3) is "a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous CH information and similar information from other domains" (Bekiari et al. 2021, p. 9). CIDOC-CRM is used to represent, for example, aDNA samples as biological object individuals.

(B): To define biological features as part of the application domain for the Poseidon data, such as tissue, contamination or genetic sex, the Uber-anatomy ontology (UBERON, Version 2025-01-15)[3], the Phenotype And Trait Ontology (PATO, Version v2025-02-

3 https://www.ebi.ac.uk/ols4/ontologies/uberon; *Visited on May 20, 2025.*

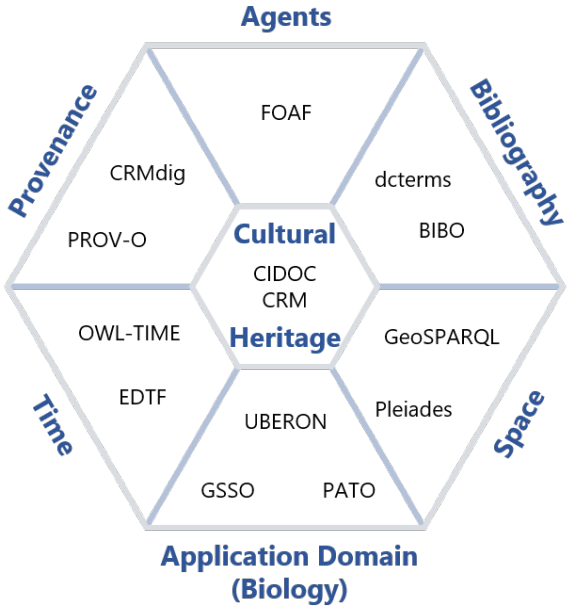


Figure 2: The ontologies grouped by the concept domain they are used for.

01)[4] and the Gender, Sex, and Sexual Orientation ontology (GSSO, Version 2.0.5)[5] are used respectively.

(C): Temporal concepts, such as time positions, durations, and temporal relations (including the ordering of intervals and instants), are represented using the OWL-Time ontology (Cox and Little 2022). The standardised definition of temporal features, such as time periods or years, is achieved using the Extended Date/Time Format (EDTF, Version from February 4, 2019)[6].

(D): Standardised representation of spatial data types (Simple Features) together with spatial query functions are provided by the GeoSPARQL ontology (Car et al. 2024; Car and Homburg 2022, Version 1.1) and used to represent concepts such as (discovery) sites. This enables effective use in geographic information systems and linked data contexts (e.g. Thiery et al. 2021). The humanities' perspective on geolocations, in addition to GeoSPARQL, is incorporated by including the concepts of Pleiades *Place* and *Location* (Elliott 2017, Version 4.0.1).

(E): People and other types of agents, such as organisations, are modelled by the Friend of a Friend (FOAF) Vocabulary Specification 0.99 (Brickley and Miller 2014).

(F): Bibliographic information, such as documents and journals, is represented using the Bibliographic Ontology (BIBO), which facilitates interoperability between data sources (Giasson and D'Arcus 2009, Version 1.3). A standardised set of metadata elements is defined by DCMI Metadata Terms (dcterms) and used to specify the concept of an author in accordance with BIBO (Board 2020, Version from January 20, 2020).

4 https://www.ebi.ac.uk/ols4/ontologies/pato; *Visited on May 20, 2025.*
5 https://www.ebi.ac.uk/ols4/ontologies/gsso; *Visited on May 20, 2025.*
6 https://www.loc.gov/standards/datetime/; *Visited on May 20, 2025.*

(G): The PROV Ontology (PROV-O) is a framework for representing relationships between entities, activities and agents (Lebo, Sahoo, and McGuinness 2013, W3C Recommendation 30 April 2013). It will mainly be used to describe the process of transforming data into RDF triples. CRM_{dig} (Doerr and Theodoridou 2011, Version 4.0) is an extension of CIDOC-CRM to represent the provenance of digital data generated from scientific observations. It is used to represent software executions for analysis.

The ontologies selected for the schema were chosen based on a balance of domain coverage, semantic interoperability, maturity, and uptake within the relevant research communities. While the chosen ontologies are well-established, there are issues with their use, and there are alternatives that warrant consideration. For example, the FOAF vocabulary, which is used to model agents, has not seen significant content updates since 2014 and is therefore considered outdated for modern Semantic Web applications. However, it is still included because many existing datasets are modelled using FOAF, and the aim is to achieve interoperability with these datasets. In the context of biological data, the selected ontologies (UBERON, PATO and GSSO) are all part of the OBO Foundry, a collaborative initiative that aims to create interoperable ontologies related to the life sciences. While integrating these ontologies provides a strong semantic basis for representing anatomical and phenotypic features, further aligning them with OBO or other upper-level categories could enhance cross-domain interoperability. However, such extensions were deemed out of scope for the current use case, which focuses on representing aDNA-related attributes. Regarding provenance modelling, PROV-O and CRM_{dig} were chosen due to their formal W3C endorsement (in the case of PROV-O) and compatibility with CIDOC-CRM-based N4O modelling approach.

Overall, the selected ontologies provide a coherent foundation for the integration of heterogeneous archaeo-natural data. However, ontology evolution must be continuously monitored, particularly with regard to maintenance and adoption. Future iterations of the schema may involve replacing or augmenting inactive vocabularies, such as FOAF, with robust alternatives that are actively maintained.

3 Related Work

The presented approach aims to extend CIDOC-CRM to be applicable to archaeo-natural data, allowing integration with data not based on CIDOC-CRM. There is prior work about combining ontologies, which also include CIDOC-CRM. Relevant approaches in the scope of this work are presented below.

Binding, May, and Tudhope (2008) demonstrate how CIDOC-CRM and its English Heritage extension (CRM-EH) can facilitate the integration of heterogeneous archaeological datasets by transforming relational databases into RDF. They highlight the challenges

and benefits of ontology-based data integration, particularly when modelling archaeological processes, ensuring semantic interoperability and improving dataset accessibility. They also emphasise the role of semi-automatic mapping and extraction tools in streamlining the transformation process.

Deicke (2016) explores the application of the CIDOC-CRM for modelling archaeological catalogues, focusing specifically on Late Bronze Age elite burials. The study shows that structured semantic modelling improves data reuse, provides a better understanding of archaeological contexts and enables integration into LOD initiatives The paper emphasises the advantages of using ontologies to structure research for greater transparency and interoperability, particularly in the fields of digital humanities and archaeology.

Nys, Ruymbeke, and Billen (2018) propose a hybrid ontology that integrates GeoSPARQL and OWL-Time with CIDOC-CRM, enhancing spatio-temporal reasoning. Although CIDOC-CRM includes spatial and temporal elements, a lack of standardisation limits its computational reasoning capabilities. To address this issue, the authors link CRM_{geo}, an extension of CIDOC-CRM designed for geospatial data, with GeoSPARQL for spatial representation and OWL-Time for temporal relationships. This enables complex spatio-temporal queries and demonstrates how semantic, spatial and temporal reasoning can facilitate knowledge discovery in archaeological datasets. This integration enhances the interoperability and analytical potential of CIDOC-CRM-based systems and is also considered in this work.

Hyvönen (2019) examines the evolution of semantic portals in CH, tracing their development from first-generation systems focused on data aggregation and search to second-generation systems that incorporate tools for data analysis and knowledge discovery. Third-generation systems, which integrate automated knowledge discovery and explainable AI, are then introduced. The paper introduces the Sampo model: a framework for building linked data infrastructures that support digital humanities research by providing interactive tools for filtering, visualising and analysing large datasets.

Koho et al. (2020) extend CIDOC-CRM with military history ontologies to create the WarSampo knowledge graph, which semantically represents spatio-temporal events related to Finland during the Second World War. They introduce CIDOC-CRM subclasses and SKOS[7] vocabularies to ensure interoperability while leveraging the event-based structure of CIDOC-CRM for data harmonisation. The WarSampo KG integrates heterogeneous historical datasets, enabling linked data queries and providing a foundation for semantic applications in digital humanities and military history research.

Tzitzikas et al. (2022) explore the integration of machine learning (ML) into CIDOC-CRM processes, addressing challenges posed by its event-centric structure and complex hierarchy, such as the manual effort required for schema alignment. They analyse key tasks in which ML could enhance CIDOC-CRM-based workflows, such as information

7 https://www.w3.org/2004/02/skos/; *Visited on May 20, 2025.*

extraction, data transformation and instance matching. The paper reviews successful applications of ML, outlines future research directions and highlights its potential to improve semantic interoperability, particularly when processing unstructured data. It also provides a curated list of publicly available CIDOC-CRM datasets that can be used to train and evaluate ML models.

4 From Data to Standards

4.1 Use Case: Poseidon Framework

The Poseidon Framework (Schmid et al. 2024) aims to enable FAIR and open management of ancient DNA (aDNA) data. It enriches genotype data with metadata and archaeological context data in a structured yet flexible format. The Poseidon Framework consists of three different archives, one of which is the Poseidon Community Archive[8]. It contains publication-wise genotype data to ensure reproducibility and is maintained by the Max Planck Institute for Evolutionary Anthropology and researchers worldwide.

In addition to other information, the Janno file contains standardised contextual information and metadata for genotyped individuals. This information – consisting of over 18,000 rows, each representing a contextualised genetic sample – originates from 194 datasets and their associated bibliographic metadata files. It is incorporated into the constructed HIN.

4.2 Motivating Community Standards

Data linkage is difficult to achieve due the variety of schemas and formats resulting from different requirements and modelling choices (Mountantonakis and Tzitzikas 2019). A naively created Poseidon network schema that only considers the attributes and their descriptions cannot be integrated with records containing similar information represented using different modelling decisions. For example, suppose the object type *Place* is called *Region* or *Location* in an HIN created from other data. There is no way to automatically infer that these node types represent the same object type and should therefore be unified. This highlights the need for standards at the schema level to ensure the interoperability of networks representing the same node types. Adherence to such standards, achieved by creating a common ontology, in well-structured data with the potential for analysis beyond the context of a specific dataset (Deicke 2016). Formal ontologies and terminologies are essential for achieving the precision of human-mediated knowledge (Doerr 2005).

8 https://github.com/poseidon-framework/community-archive/tree/master; *Visited on May 20, 2025.*

Standards are also needed at the data level to match and map differently represented instances of the same object (Thiery and Mees 2023).

Schema-Level Standards

One of the main barriers to working with linked data is the lack of a semantically consistent representation of object types (Middle 2024). Matching datasets to a global, dataset-overarching network schema provides unified representations of concepts and enables the semantic interoperability of data from different sources. Using ontologies as modelling standards addresses this by providing a unified conceptual framework that enables the semantic integration of network data at schema level. Well-known foundational ontologies or reference models can be used as the basis for such a global network schema, for example, CIDOC-CRM (e.g., Bekiari et al. 2021) for the CH domain or BFO (Otte, Beverley, and Ruttenberg 2022) for NFDI4Culture (Steller et al. 2025, p. 11). CIDOC-CRM offers a formal structure of object and relationship types that supports cross-dataset integration and exploration within the cultural heritage domain (Doerr 2005; Doerr, Light, and Hiebel 2020). It is also employed as the reference model for the application ontology (Guarino 1998, p. 8) of the N4O KG. The Archaeo-Natural Ontology (ArNO), which is presented in this work, is thus based on CIDOC-CRM to ensure the HIN's schema-level consistency with other CH datasets.

However, CIDOC-CRM alone is insufficient for representing the anthropological, archaeobotanical and archaeozoological data considered in TA3, no for ensuring interoperability with datasets from domains other than CH. Using CIDOC-CRM alone for the semantic modelling of spatial information (e.g., Padfield et al. 2019; Stein and Balandi 2019; Doerr 2003) raises conceptual issues in the context of cultural heritage (see Voß 2025). The CIDOC-CRM class *E53 Place* covers extents in the natural space in a purely physical sense, independent of temporal phenomena and matter. In contrast, *E27 Site* covers pieces of land or sea floor represented by photographs, paintings, and maps. This does not allign the Pleiades logic of *Place* – a descriptive and conceptual context for geographic or toponymic information – and *Location* – a record information about measurable points on the Earth's surface. To be interoperable with the CH domain and to be able to model the presented view of spatial concepts, this work uses CIDOC-CRM as a reference model, integrating domain-specific extensions, such as archaeology.link's LADO (e.g. Thiery and Mees 2025). This intertwines CIDOC-CRM classes with the Pleiades Place and Location concepts and GeoSPARQL to represent spatial objects.

Therefore, modelling all archaeo-natural data directly with CIDOC-CRM alone does not address the domain-specific and interdisciplinary requirements of TA3. To bridge this gap, the hybrid ArNO ontology (Nys, Ruymbeke, and Billen 2018) is being developed, combining classes and properties from the ontologies introduced above (see Section 2.4). This allows for a fine-grained representation of the considered archaeo-natural data and enables integration with datasets from CH and other domains that use some of the same

ontologies. ArNO is being developed on GitHub[9] (Straten and Thiery 2025). Subclasses and custom properties are used to accurately represent the Poseidon data (see Figure 3) while ensuring interoperability with other systems based on the used ontologies (Koho et al. 2020). All object types shown in the network schema Figure 3 are subclasses of classes from the ontologies incorporated in ArNO, particularly the CIDOC-CRM classes.

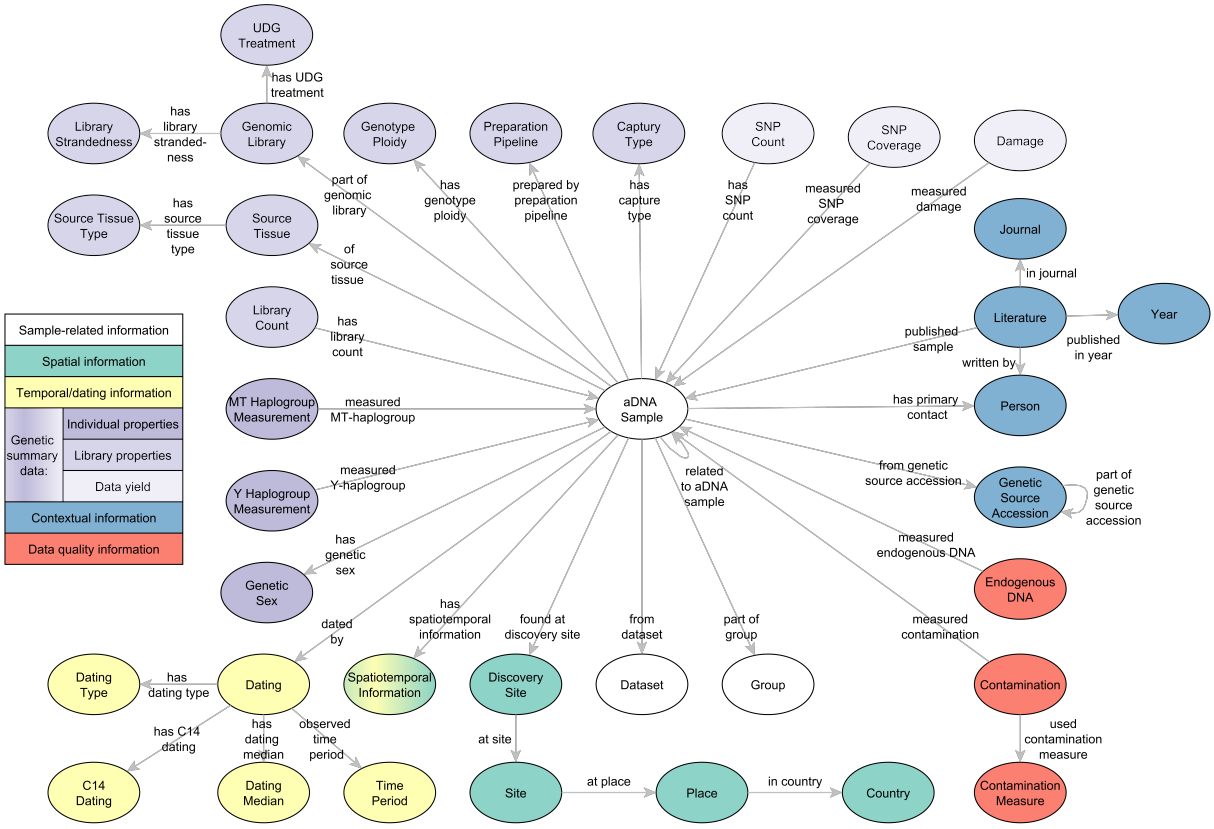


Figure 3: The network schema of the Poseidon network without data properties for better readability.

However, not all relationships are subproperties of properties from the used ontologies, particularly not from CIDOC-CRM. This is because CIDOC-CRM properties are not suitable for subtyping to create the relationship types of the Poseidon HIN, as they are designed to be generally applicable in the CH domain and do not specify the approach used for analysis in the natural and life sciences. As an event-centric approach, CIDOC-CRM typically models a series of events in which an object has been involved, rather than placing the individual object at the centre of the schema (Middle 2024; Bruseker, Carboni, and Guillem 2017; Bekiari et al. 2021). CRM-EH creates "virtual" events that are not part of the considered dataset in order to correctly connect such entities in their network. Currently, ArNO adopts an object-oriented approach by considering CIDOC-CRM as a reference model rather than applying it holistically to model the data. This results in the creation of new types of relationships linking *aDNA Samples* to their respective measure-

9 https://archaeonatural-cloud.github.io/archaeonatural-ontology/; *Visited on May 20, 2025.*

ments. This avoids definiting an event that contains no additional information, which would artificially increase the complexity of the network and hinder network analysis, resulting in slow and overly complex queries (Koho et al. 2020). This is particularly important for meta path-based HIN analysis (e.g., Sun et al. 2011), where the complexity of the network schema complicates the query definition. The hybrid ontology introduced in this work thus introduces new relationships that are not inherited from the CIDOC-CRM properties. These relationships directly connect the *aDNA Samples* to their respective measurements, without the need to create an event as a connector.

HIN data is stored in a task-specific format, enabling attribute values for nodes and relationships to be stored directly when modelling them similar to property graphs. However, when OWL and RDF are used to represent the network, it becomes necessary to use data properties to represent attributes. It is not possible to represent data properties of object properties (i.e. attributes of relationships). Therefore, an additional class, *aDNA Sample Relation*, had to be introduced in ArNO to enable attributes to be added to the relationship between *aDNA Samples*. An alternative way of representing this information is to use quintuples, which add both attributes to the relation using blank nodes, in a manner similar to the use of quadruples in AMT modelling (Thiery et al. 2022).

Data-Level Standards

In order to successfully integrate diverse datasets into a single HIN, it is crucial to map multiple instances of the same object to the same object representation. The same object may be referred to in different languages, by various identifiers or by conflicting modelling choices, which complicates entity identification and affects entity resolution. Gazetteers, controlled vocabularies and authority files – such as GeoNames[10] or the Integrated Authority File (GND)[11] – provide standardised identifiers and labels that support the harmonisation of different representations of the same object. In contrast, community-driven vocabularies, which are often created during domain-specific database projects (e.g., Thiery and Mees 2024, pp. 68-69), are less rigidly standardised and offer greater flexibility and adaptability to emerging needs. However, they may have limited potential for generalisation.

Different instances that refer to the same object can be resolved by *canonicalisation*, i.e., by transforming both objects into the same representation, or by *binding*, i.e., by establishing an additional relationship between these objects (Mountantonakis and Tzitzikas 2019). As one of the goals of the integrated network presented in this work is to perform holistic network analysis, canonicalisation is used to ensure that all of an object's semantics are captured in a single representative instance for use in network analysis. Binding would result in multiple linked instances of the same object, which would pre-

10 https://www.geonames.org/; *Visited on May 20, 2025.*
11 https://explore.gnd.network/en/; *Visited on May 20, 2025.*

vent the semantic integration and incorporation of all modelled information about the object into a single representative instance. If binding is chosen instead of canonicalisation, not all information – e.g., attributes and relationships to other objects – of all instances of the query object can be considered when using a single representative as the query object in network analysis tasks. However, canonicalisation might introduce limitations. By reducing multiple instances to a single representation, important nuances may be lost, e.g., various provenance trails, uncertainty in the original data sources or obfuscating contextual ambiguity. For example, if different datasets refer to an object with varying spatial precision or conflicting temporal attributions, canonicalisation may obscure these discrepancies in favour of a unified representation. Therefore, the trade-off between analytical clarity and the preservation of interpretative richness must therefore be carefully considered. The presented work represents spatial and temporal information as separate classes, enabling such conflicting information to be represented using multiple relationships.

Consistency at the data level is achieved through the use of IRIs where available. Rather than retrieving all relevant bibliographic information directly from the data files themselves, Digital Object Identifiers (DOIs) are used to consistently retrieve it via the Crossref API[12]. This prevents different representations of bibliographic objects from resulting in multiple instances of the same object being created. ISO 3166 country codes are used to ensure the consistent representation of *Countries* in the source datasets. Data-level consistency can be further ensured by providing IRIs for objects obtained using appropriate gazetteers for spatial data. However, the name of a *Site* alone, as it is found in most existing datasets, is not sufficient to correctly map it to an existing IRI using a gazetteer, since several *Sites* may have the same name. Therefore, geographic coordinate information must also be (created and) taken into account to make the mapping more likely to be correct.

The presented framework aims to create a semantically consistent integrated HIN that is interoperable with future datasets from the archaeo-natural domain and related fields. This is achieved by combining well-known ontologies and terminologies. The resulting LOD provides an opportunity to address new research questions and discover previously unknown relationships between objects (Middle 2024) by applying network analysis methods to a shared digital heritage rather than isolated datasets.

5 Discussion

Although CIDOC-CRM enables the representation of detailed and well-structured data, mapping data to it is a time-consuming and manual process that requires expertise, as the quality of the mapping depends on familiarity with both the dataset and the complex

12 https://api.crossref.org/; *Visited on May 20, 2025.*

ontology (Deicke 2016; Tzitzikas et al. 2022; Middle 2024). Several projects have avoided using CIDOC-CRM due to the high resources required for its correct implementation (Middle 2024). Nevertheless, this study uses CIDOC-CRM as a reference model to ensure interoperability in the CH domain and to facilitate the data integration into the N4O KG. Despite the fact that the data considered includes well-documented metadata and comes in well-structured formats, extensive and repeated consultations with domain experts were required to accurately map it to CIDOC-CRM. Without close collaboration between modellers and data owners, models that are semantically correct but misrepresent the underlying data may be created. Therefore, continuous communication is required for interdisciplinary collaboration to ensure that the structures accurately reflect the data content, enable future integration and meet the needs of researchers.

Representing the network data using RDF enables the use of blank nodes for missing data, which act as connectors. For example, this could be used for missing *Discovery Site* where *Site* information is available (see Figure 3). In traditional HINs, it would be difficult to associate the *Site* information with the *aDNA Sample* if there were no details about the connecting *Discovery Site* (see Figure 3). Therefore, HINs may be based on a star network schema (Sun, Yu, and Han 2009) to prevent the loss of link information. However, this limits the analytical depth obtained through meta-path-based analysis, in which objects are connected by a semantic sequence of edge types. This is because these paths always pass through the central node type of the star network schema when connecting a pair of object types.

In conventional RDF-based KGs, purely spatial or temporal information, e.g., *Discovery Site* or *Time Period*, is usually defined as a literal (attribute) using data properties. However, in order to discover patterns of spatial or temporal similarity in HINs, it is crucial to have objects representing such information. Therefore, the object types of *Discovery Site* and *Time Period* are included in ArNO.

The ArNO ontology is a work in progress, with the Poseidon datasets currently forming its foundation. Other archaeo-natural datasets from TA3, such as the archaeozoological data from the Ossobook framework (Kriegel et al. 2009) or the archaeobotanical data from ArboDat (Kreuz, Elger, and Frenzel 2022), will be integrated into the unified HIN based on ArNO and will be followed by other archaeo-natural data. The WarSampo KG (Koho et al. 2020) has extended its class structure iteratively when considering new datasets. The class structure of ArNO may also change when integrating such datasets in a community-driven manner. However, ArNO's modular structure, the use of CIDOC-CRM as a reference model and its extension using standard ontologies – e.g., for spatial and temporal information – allow for an integrated foundation that only needs to be extended when previously unseen classes or properties are added, for example, from a new application domain. Ontologies such as ArNO are introduced to the community through discussions within the NFDI consortium. These discussions take place in N4O groups such as Temporary Working Groups and Community Clusters, as well as groups

that transcend the N4O consortium, such as Memorandum of Understanding groups and NFDI sections.

The mapping of ArNO's classes and properties to top-level ontologies (Guarino 1998, p. 7) and other existing ontologies will be discussed later, with the aim of producing a consistent mapping of the data to other systems, e.g., by including their concepts to ArNO. For example, one could incorporate ETS[13], SOSA[14] or OBOE[15] in ArNO, allowing interoperability with natural data modelled in NFDI4Biodiversity. One could also incorporate BFO to enable interoperability with NFDIcore-based systems[16], aiming to ensure data interoperability across NFDI consortia. After the successful extension and integration into the N4O KG, extended documentation and exemplary data modelling will be provided.

6 Conclusion & Future Work

This work promotes community standards for research data that enhance the interoperability of heterogeneous datasets in an interdisciplinary context. Adherence to schema- and data-level standards ensures alignment with the FAIR principles, guaranteeing that integrated HINs are semantically consistent and interoperable with other datasets. This creates a rich source of reusable data for future research. Using standardised ontologies to enforce schema-level standards results in a unified representation of broad, multidisciplinary resources. This connects disparate datasets and makes connections to the wider Linked Data ecosystem (Middle 2024). By sharing this approach, this work aims to contribute to wider efforts to promote the interoperability and semantic consistency of heterogeneous datasets within CH and other related research fields.

KGs typically focus on structured knowledge representation and reasoning, often leveraging ontologies. In contrast, HINs are typically used for applying network analysis methods to discover new knowledge (e.g., Hogan et al. 2021; Shi et al. 2017). Sun et al. (2022) distinguish KGs and HINs by the fact that KGs may contain a hierarchical structure, whereas HINs are not defined to consider a hierarchical network schema. However, the use of ontologies presented here to standardise HINs at the schema level incorporates a taxonomic hierarchy into the network schema, where the node types are subclasses of ontology classes. Conventional HIN analysis methods also only consider flat, non-hierarchical network schemas and therefore fail to adequately represent hierarchically structured information modelled by ontologies. When ontologies are used to define the HIN network schema and IRIs to identify object instances, the distinction between HINs and KGs becomes less clear, particularly when data models that can accommodate both

13 https://biodivportal.gfbio.org/ontologies/ETS; *Visited on May 20, 2025.*
14 https://biodivportal.gfbio.org/ontologies/SOSA; *Visited on May 20, 2025.*
15 https://biodivportal.gfbio.org/ontologies/OBOE; *Visited on May 20, 2025.*
16 https://ise-fizkarlsruhe.github.io/nfdicore/; *Visited on May 20, 2025.*

heterogeneous networks and semantic representations are employed. Therefore, the presented standards will lead to greater interoperability between HINs and KGs, especially if standardised formats such as RDF are used for HIN representation. The achieved interoperability could enable the integration of separate HINs into a large, unified KG, and could also facilitate the application of HIN analysis methods to large-scale KGs. However, meta path-based methods require a manageable network schema in order to define relevant paths. Future work could therefore focus on discovering how HIN analysis methods can be applied to KGs and RDF data. Current HIN analysis methods typically use a flat network schema for knowledge discovery. Another promising area for future research will therefore be to extend HIN analysis methods (e.g., Sun et al. 2011; Rossi et al. 2019) to consider hierarchical ontology structures (e.g., Huang 2021).

Future work will also involve integrating the Arbodat and Ossobook datasets into ArNO, primarily through mapping them to existing object and relationship types. If necessary, new classes will be added to ArNO, particularly within the application domain (see Figure 2). The average of the *Discovery Site* coordinates will be computed to obtain representative spatial information for mapping the *Site* and *Location* label to IRIs. Different spatial granularities and inexact spatial matches can also be considered when mapping *Sites* to gazetteers. Future work will also involve incorporating top-level ontologies and ontologies from related domains into ArNO. This will facilitate ontology alignment, enable semantic mapping across domains, and improve compatibility with existing knowledge graph infrastructures.

Acknowledgements

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) PNr. 501836407 (NFDI4Objects), by the state of Schleswig-Holstein within the DatenCampus project (22021016), and MARISPACE-X – Smart MARItime Sensor Data SPACE X (Grant FKZ 68GX21002E) funded by the German Federal Ministry for Economic Affairs and Climate Action.

Authorship Contributions

- Conceptualization: M.t.S., F.T. and S.S.
- Methodology: M.t.S., F.T. and S.S.
- Formal analysis: M.t.S. and F.T
- Investigation: M.t.S. and F.T
- Writing (original draft preparation): M.t.S. and F.T
- Writing (review and editing): M.t.S., F.T. and S.S
- Supervision: M.R.

- Project administration: M.R.
- Funding acquisition: M.R.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

The authors declare no conflicts of interest.

Bibliography

Bekiari, Chryssoula, George Bruseker, Martin Doerr, Christian-Emil Ore, Stephen Stead, and Athanasios Velios. 2021. *Definition of the CIDOC Conceptual Reference Model v7.1.1.* Technical report. The CIDOC Conceptual Reference Model Special Interest Group. https://doi.org/10.26225/FDZH-X261.

Berners-Lee, Tim. 2006. *Linked Data – Design Issues.* Visited on March 24, 2025. https://www.w3.org/DesignIssues/LinkedData.html.

Binding, Ceri, Keith May, and Douglas Tudhope. 2008. "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction Via the CIDOC CRM", 280–290. ISBN: 978-3-540-87598-7. https://doi.org/10.1007/978-3-540-87599-4_30.

Bleiholder, Jens, and Felix Naumann. 2008. "Data Fusion". *ACM Computing Surveys* 41 (1): 1–41. https://doi.org/10.1145/1456650.1456651.

Board, DCMI Usage. 2020. *DCMI Metadata Terms.* Visited on March 24, 2025. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

Brickley, Dan, and Libby Miller. 2014. *FOAF Vocabulary Specification.* Visited on March 24, 2025. http://xmlns.com/foaf/spec/.

Bruseker, George, Nicola Carboni, and Anaïs Guillem. 2017. "Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM". In *Heritage and Archaeology in the Digital Age: Acquisition, Curation, and Dissemination of Spatial Cultural Heritage Data,* edited by Matthew L. Vincent, Víctor Manuel López-Menchero Bendicho, Marinos Ioannides, and Thomas E. Levy, 93–131. Cham: Springer International Publishing. ISBN: 978-3-319-65370-9. https://doi.org/10.1007/978-3-319-65370-9_6.

Car, Nicholas J., and Timo Homburg. 2022. “GeoSPARQL 1.1: Motivations, Details and Applications of the Decadal Update to the Most Important Geospatial LOD Standard”. *ISPRS International Journal of Geo-Information* 11 (2): 117. https://doi.org/10.3390/ijgi11020117.

Car, Nicholas J., Timo Homburg, Matthew Perry, Frans Knibbe, Simon J.D. Cox, Joseph Abhayaratna, Mathias Bonduel, Paul J. Cripps, and Krzysztof Janowicz. 2024. *OGC GeoSPARQL – A Geographic Query Language for RDF Data.* OGC Standard 22-047r1. Open Geospatial Consortium. Visited on March 24, 2025. https://docs.ogc.org/is/22-047r1/22-047r1.html.

Cox, Simon, and Chris Little. 2022. *Time Ontology in OWL.* W3C Candidate Recommendation Draft CRD-owl-time-20221115. World Wide Web Consortium (W3C). Visited on March 24, 2025. https://www.w3.org/TR/2022/CRD-owl-time-20221115/.

Cyganiak, Richard, David Hyland-Wood, and Markus Lanthaler. 2014. “RDF 1.1 Concepts and Abstract Syntax”. *W3C Proposed Recommendation,* visited on March 24, 2025. https://www.w3.org/TR/rdf11-concepts/.

Deicke, Aline Julia Elisabeth. 2016. “CIDOC CRM-based modeling of archaeological catalogue data.” In *DHC@ MTSR.* Visited on March 24, 2025. https://ceur-ws.org/Vol-1764/4.pdf.

Doerr, Martin. 2003. “The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata”. *AI Magazine* 24 (3): 75–75. https://doi.org/10.1609/aimag.v24i3.1720.

———. 2005. “The CIDOC CRM, an Ontological Approach to Schema Heterogeneity”. In *Semantic Interoperability and Integration,* edited by Y. Kalfoglou, M. Schorlemmer, A. Sheth, S. Staab, and M. Uschold, 4391:1–5. Dagstuhl Seminar Proceedings (DagSemProc). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/DagSemProc.04391.22.

Doerr, Martin, Richard Light, and Gerald Hiebel. 2020. *Implementing the CIDOC Conceptual Reference Model in RDF.* Technical report. CIDOC CRM Special Interest Group. Visited on March 24, 2025. https://cidoc-crm.org/Resources/implementing-the-cidoc-conceptual-reference-model-in-rdf.

Doerr, Martin, and Maria Theodoridou. 2011. “CRMdig: A Generic Digital Provenance Model for Scientific Observation”. In *3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP 11).* Visited on March 24, 2025. https://www.usenix.org/legacy/event/tapp11/tech/final_files/Doerr.pdf.

Elliott, Tom. 2017. *Pleiades Data Model.* Visited on March 24, 2025. https://pleiades.stoa.org/help/pleiades-data-model.

Giasson, Frédérick, and Bruce D’Arcus. 2009. *Bibliographic Ontology Specification.* Version 1.3, published on November 4, 2009. Visited on March 24, 2025. http://bibliontology.com.

Guarino, Nicola. 1998. “Formal Ontologies and Information Systems”. Visited on March 24, 2025. https://www.loa.istc.cnr.it/old/Papers/FOIS98.pdf.

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

Hitzler, Pascal, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. 2012. *OWL 2 Web Ontology Language Primer.* W3C Recommendation. World Wide Web Consortium (W3C). Visited on March 24, 2025. http://www.w3.org/TR/owl2-primer/.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia D’amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, et al. 2021. “Knowledge Graphs”. *ACM Computing Surveys* 54 (4): 1–37. https://doi.org/10.1145/3447772.

Huang, Yue. 2021. “Incorporating domain ontology information into clustering in heterogeneous networks”. *WIREs Data Mining and Knowledge Discovery* 11 (4): e1413. https://doi.org/10.1002/widm.1413.

Hughes, Lorna, Panos Constantopoulos, and Costis Dallas. 2015. “Digital Methods in the Humanities”. In *A New Companion to Digital Humanities,* edited by Susan Schreibmann, Ray Siemens, and John Unsworth, 150–170. John Wiley / Sons, Ltd. ISBN: 9781118680599. https://doi.org/10.1002/9781118680605.ch11.

Hyvönen, Eero. 2019. “Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery”. *Semantic Web* 11:1–7. https://doi.org/10.3233/SW-190386.

Koho, Mikko, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Eero Hyvönen. 2020. “WarSampo knowledge graph: Finland in the Second World War as Linked Open Data”. *Semantic Web* 12:1–14. https://doi.org/10.3233/SW-200392.

Kreuz, Angela, Kirsten Elger, and Simone Frenzel. 2022. *Digitising and sharing archaeobotanical data–the ArboDat 2016 Datacentre project.* Visited on March 24, 2025. https://gfzpublic.gfz-potsdam.de/rest/items/item_5017643_2/component/file_5018422/content.

Kriegel, Hans-Peter, Peer Kröger, Henriette Obermaier, Joris Peters, Matthias Renz, and Christiaan Hendrikus van der Meijden. 2009. “OSSOBOOK: database and knowledge-management techniques for archaeozoology”. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management,* 2091–2092. CIKM ’09. Hong Kong, China: Association for Computing Machinery. ISBN: 9781605585123. https://doi.org/10.1145/1645953.1646318.

Lebo, Timothy, Satya Sahoo, and Deborah McGuinness. 2013. *PROV-O: The PROV Ontology.* World Wide Web Consortium (W3C) Recommendation. Visited on March 24, 2025. https://www.w3.org/TR/prov-o/.

Legler, Frank, and Felix Naumann. 2007. “A Classification of Schema Mappings and Analysis of Mapping Tools”. In *Datenbanksysteme in Business, Technologie und Web (BTW 2007), 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Proceedings, 7.-9. März 2007, Aachen, Germany,* volume P-103, 449–463. LNI. GI. ISBN: 978-3-88579-197-3, visited on March 24, 2025. https://dl.gi.de/handle/20.500.12116/31815.

Middle, Sarah. 2024. “Linked Ancient World Data: Implementation, Advantages, and Barriers”. *Digital Classics Online* 10 (1): 16–49. https://doi.org/10.11588/dco.2024.10.104105.

Motik, Boris, Peter F. Patel-Schneider, and Bijan Parsia. 2012. *OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition).* Visited on March 24, 2025. https://www.w3.org/TR/owl2-syntax/.

Mountantonakis, Michalis, and Yannis Tzitzikas. 2019. “Large-scale Semantic Integration of Linked Data: A Survey”. *ACM Computing Surveys* 52:1–40. https://doi.org/10.1145/3345551.

Nys, Gilles-Antoine, Muriel Ruymbeke, and Roland Billen. 2018. “Spatio-Temporal Reasoning in CIDOC CRM: An Hybrid Ontology with GeoSPARQL and OWL-Time”. Visited on March 24, 2025. https://ceur-ws.org/Vol-2230/paper_04.pdf.

Otte, J Neil, John Beverley, and Alan Ruttenberg. 2022. “BFO: Basic formal ontology”. *Applied ontology* 17 (1): 17–43. https://doi.org/10.3233/AO-220262.

Padfield, Joseph, Kalliopi Kontiza, Antonis Bikakis, and Andreas Vlachidis. 2019. “Semantic Representation and Location Provenance of Cultural Heritage Information: the National Gallery Collection in London”. *Heritage* 2 (1): 648–665. https://doi.org/10.3390/heritage2010042.

Patel, Hardik, Pavlos Paraskevopoulos, and Matthias Renz. 2018a. “Data Fusion of Diverse Data Sources: Enrich Spatial Data Knowledge Using HINs”. In *Proceedings of the Fifth International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data,* 13–18. GeoRich’18. Houston, TX, USA: Association for Computing Machinery. ISBN: 9781450358323. https://doi.org/10.1145/3210272.3210275.

———. 2018b. “GeoTeGra: A System for the Creation of Knowledge Graph Based on Social Network Data with Geographical and Temporal Information”. In *2018 IEEE/-ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* 617–620. https://doi.org/10.1109/ASONAM.2018.8508674.

Rodriguez, Marko A., and Peter Neubauer. 2010. “Constructions from dots and lines”. *Bulletin of the American Society for Information Science and Technology* 36 (6): 35–41. https://doi.org/10.1002/bult.2010.1720360610.

Rossi, Ryan A., Nesreen K. Ahmed, Aldo Carranza, David Arbour, Anup Rao, Sungchul Kim, and Eunyee Koh. 2019. *Heterogeneous Network Motifs.* https://doi.org/10.48550/arXiv.1901.10026.

Rummel, Philipp Von, Christin Keller, and Fabian Fricke. 2025. “NFDI4Objects: Warum brauchen wir eine gemeinsame Forschungsdateninfrastruktur für die materiellen Hinterlassenschaften der Menschheitsgeschichte?” *Archäologische Informationen* 47 (NWDVA 2024 Bochum: Digitale Archäologie): Bd. 47 (2024): Archäologische Informationen. https://doi.org/10.11588/AI.2024.1.110390.

Schmid, Clemens, Ayshin Ghalichi, Thiseas C. Lamnidis, Dhananjaya B. A. Mudiyanselage, Wolfgang Haak, and Stephan Schiffels. 2024. “Poseidon – A framework for archaeogenetic human genotype data management”. *eLife* 13. https://doi.org/10.7554/elife.98317.1.

Schmidt, Sophie C., Florian Thiery, and Martina Trognitz. 2022. “Practices of Linked Open Data in Archaeology and Their Realisation in Wikidata”. *Digital* 2 (3): 333–364. https://doi.org/10.3390/digital2030019.

Shi, Chuan, Yitong Li, Jiawei Zhang, Yizhou Sun, and Philip S. Yu. 2017. “A survey of heterogeneous information network analysis”. *IEEE Transactions on Knowledge and Data Engineering* 29 (1): 17–37. https://doi.org/10.1109/TKDE.2016.2598561.

Stein, Regine, and Oguzhan Balandi. 2019. “Using LIDO for Evolving Object Documentation into CIDOC CRM”. *Heritage* 2 (1): 1023–1031. https://doi.org/10.3390/heritage2010066.

Steller, Jonatan Jalle, Tabea Tietz, Linnaea Charlotte Söhn, Harald Sack, Heike Fliegl, Torsten Schrade, Alexandra Büttner, Etienne Posthumus, and Oleksandra Bruns. 2025. *Knowledge Graph-based Research Data Integration for NFDI4Culture and Beyond.* Poster (E-Science-Tage 2025). https://doi.org/10.5281/zenodo.14989010.

Straten, Mattis thor, and Florian Thiery. 2025. “Archaeo-Natural Ontology (ArNO) – v0.1”. *Squirrel Papers* 7 (2): L4. https://doi.org/10.5281/zenodo.15095413.

Strohm, Steffen, Hartwig Buenning, and Matthias Renz. 2023. “Implementing a FAIR Information System for Archaeology-Related Interdisciplinary Research”. In *2023 IEEE 19th International Conference on e-Science (e-Science),* 1–2. https://doi.org/10.1109/e-Science58273.2023.10254840.

Sun, Yizhou, Jiawei Han, Xifeng Yan, Philip Yu, and Tianyi Wu. 2022. “Heterogeneous information networks: the past, the present, and the future”. *Proceedings of the VLDB Endowment* 15:3807–3811. https://doi.org/10.14778/3554821.3554901.

Sun, Yizhou, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. 2011. “PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks”. *Proceedings of the VLDB Endowment* 4 (11): 992–1003. https://doi.org/10.14778/3402707.3402736.

Sun, Yizhou, Yintao Yu, and Jiawei Han. 2009. “Ranking-based clustering of heterogeneous information networks with star network schema”. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 797–806. KDD ’09. Paris, France: Association for Computing Machinery. ISBN: 9781605584959. https://doi.org/10.1145/1557019.1557107.

Thiery, Florian, and Thomas Engel. 2016. “The Labeling System: A Bottom-up Approach for Enriched Vocabularies in the Humanities”. In *CAA2015. Keep the Revolution Going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology.* Edited by Stefano Campana, Roberto Scopigno, Gabriella Carpentiero, and Marianna Cirillo, 259–268. Oxford: Archaeopress. ISBN: 978-1-78491-337-3. https://doi.org/10.5281/zenodo.3741958.

Thiery, Florian, Timo Homburg, Sophie C. Schmidt, Jakob Voß, and Martina Trognitz. 2021. “SPARQLing Geodesy for Cultural Heritage – New Opportunities for Publishing and Analysing Volunteered Linked (geo-)data”. *FIG Peer Review Journal* FIG e-Working Week 2021 – Virtually in the Netherlands 21-25 June 2021. https://doi.org/10.5281/zenodo.5639381.

Thiery, Florian, and Allard W. Mees. 2023. “Taming Ambiguity – Dealing with doubts in archaeological datasets using LOD”. *Proceedings of the Computer Applications and Quantitative Methods in Archeology* CAA 2018: Human History and Digital Future (2018). https://doi.org/10.15496/PUBLIKATION-87762.

———. 2024. “Sharing Linked Open Data with domain-specific data-driven community hubs – archaeology.link in NFDI4Objects”. *Archeologia e Calcolatori* 35 (2): 63–74. https://doi.org/10.19282/ac.35.2.2024.08.

———. 2025. “Linked Archaeological Data Ontology (LADO) – v1.1”. *Squirrel Papers* 7 (2): L3. https://doi.org/10.5281/zenodo.15477358.

Thiery, Florian, Allard W. Mees, Bernhard Weisser, Felix F. Schäfer, Stefanie Baars, Sonja Nolte, Henriette Senst, and Philipp Von Rummel. 2023. “Object-Related Research Data Workflows Within NFDI4Objects and Beyond”. In *Proceedings of the Conference on Research Data Infrastructure,* edited by York Sure-Vetter and Carole Goble, volume 1, CoRDI2023–46. Hannover: TIB Open Publishing. https://doi.org/10.52825/cordi.v1i.326.

Thiery, Florian, Karsten Tolle, Allard Mees, and David Wigg-Wolf. 2022. “How to handle vagueness and uncertainty in graph-based LOD knowledge modelling? Dealing with archaeological numismatic and ceramological real world data”. In *Graphs and Networks in the Humanities 2022 (Graphum2022).* Squirrel Papers. https://doi.org/10.5281/zenodo.7184524.

Tzitzikas, Yannis, Michalis Mountantonakis, Pavlos Fafalios, and Yannis Marketakis. 2022. “CIDOC-CRM and Machine Learning: A Survey and Future Research”. *Heritage* 5 (3): 1612–1636. https://doi.org/10.3390/heritage5030084.

Voß, Jakob. 2025. *CIDOC-CRM in RDF Application Profile.* Visited on March 24, 2025. https://nfdi4objects.github.io/crm-rdf-ap/.

Voß, Jakob, and Josef Heers. 2024. *Integration von Forschungsdaten im NFDI4Objects Knowledge Graph.* Zenodo. https://doi.org/10.5281/zenodo.13744338.

Voß, Jakob, Josef Heers, Gerald Steilen, and Anja Gerber. 2024. *N4O Graph: The Knowledge Graph of NFDI4Objects.* Zenodo. https://doi.org/10.5281/zenodo.13946053.

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. https://doi.org/10.1038/sdata.2016.18.

Zeng, Marcia. 2008. "Knowledge Organization Systems (KOS)". *Knowledge Organization* 35:160–182. https://doi.org/10.5771/0943-7444-2008-2-3-160.

Zheng, Yuyan, Jianhua Qu, and Jiajia Yang. 2024. "StructSim: Meta-Structure-Based Similarity Measure in Heterogeneous Information Networks". *Applied Sciences* 14 (2). https://doi.org/10.3390/app14020935.

Systematic Assessment of the Prerequisites for Institutional RDM Services at a German University of Applied Sciences

Paulina Anna Dąbrowska [1], Esther Schneidenbach [2], Eske Heister [3], Lena Simon[2], Wojtek Dabrowski [1]

[1] Faculty 4: School of Computing, Communication and Business, HTW Berlin – University of Applied Sciences, Germany;
[2] Cooperation Center for Applied Sciences, HTW Berlin – University of Applied Sciences, Germany;
[3] University Library, HTW Berlin – University of Applied Sciences, Germany
Contact: fdm@htw-berlin.de

The project "FitForFDM" performed a comprehensive review of the research data management (RDM) landscape at the University of Applied Sciences (HTW) Berlin. Based on broadly accepted recommendations (e. g. FAIR Principles) and requirements of funding bodies, a questionnaire aimed at research practitioners was created and applied in semi-structured interviews with five cooperating research projects, covering all faculties of the university. In parallel, central data services were examined in terms of technical properties, intended use and practical usability for scientific activities. To augment the emerging superficial picture of RDM practices, actors indirectly involved in the research process were inquired (members of IT support and the Central Office for Research and Transfer, laboratory engineers, data- and cybersecurity officers) and a working group of representatives of the central university units and single faculties was formed.

The range of identified needs was broad. While the existing infrastructure was generally rated as good, it was deemed insufficient. The key challenges regarded practical implementation of the FAIR Principles in the absence of specialized counseling and technical support in RDM. Misunderstandings between numerous experts were amplified by confusion over vital terms (e. g. data protection vs. data security, storage vs. archiving). The lack of standards for project structure and documentation hampered systematic RDM, and already made efforts were not formalized. Nevertheless, a strong interest in learning materials and aids was expressed.

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23915 (CC BY-SA 4.0).

Methods employed and adjusted in our project shall serve as a reusable example for other Universities of Applied Sciences wishing to improve their RDM services.

Keywords: FitForFDM, research data management, semi-structured interviews, infrastructure, University of Applied Sciences

1 Introduction

In the era of data-intensive science (see e. g. Hey et al. 2009) systematic management of digital research data is recognized as crucial for future scientific endeavors. In its *Guidelines for Safeguarding Good Research Practice* (Deutsche Forschungsgemeinschaft 2019), the German Research Foundation refers directly to the FAIR Principles (Wilkinson et al. 2016) as a means of transparent and reusable research. Furthermore, the Federal Ministry of Education and Research (BMBF) defines the establishment of a data culture and the promotion of data literacy among researchers as central objectives in the coming years (Bundesministerium für Bildung und Forschung 2020). The Project "FitForFDM – Participatory development of a sustainable framework for FAIR cross-disciplinary research data management" ("Partizipative Entwicklung eines nachhaltigen Handlungsrahmens zum FAIRen fachgebiets-übergreifenden Forschungsdatenmanagement") strives to address these recommendations and requirements at the University of Applied Sciences Berlin (HTW Berlin). The main aims of the project are:

i) to document the demands and challenges concerning the status quo of research data management (RDM) practices at the university, and
ii) to formulate recommendations and provide support materials for the university-wide organization of RDM services.

In this work, the means taken to achieve objective *i* along with the results obtained thus far are outlined. We focus on the methodology developed and its strong participatory aspect. A comprehensive overview of the current state of RDM-relevant provisions and practices is an important first step before formulating universal recommendations. The methods employed and adjusted in our project shall serve as a reusable example for other Universities of Applied Sciences wishing to improve their RDM services.

2 Materials and Methods

The HTW Berlin does not have a central service unit dedicated to RDM yet. To systematically assess the current state of RDM at the university, five research projects – one per faculty – were contacted. During a kick-off meeting, all projects were briefly introduced,

followed by a presentation of basic RDM-related terms and assumptions. Ultimately, first impressions with regard to RDM in the context of participating projects were collected. In the initial phase of the cooperation, semi-structured interviews with expert research practitioners from the projects were conducted. The researchers were then shadowed by an RDM specialist from FitForFDM ("data scout") and assisted in activities related to data management, publication, or archiving. Throughout this support process, the data scout documented her observations concerning RDM as implemented in the project.

In parallel to the interviews, central IT services were examined in terms of technical properties (e.g. data versioning), intended use, and practical usability of data storage solutions for scientific activities. Additional information on top-down constraints on the services related to RDM was provided by regular discussions within a working group concerned with objective *ii* of FitForFDM: formulation of a Research Data Policy and Strategy for the university. The group started with an assessment of the current provision and the definition of the target scope of institutional support in RDM, guided by the RISE-DE Model (Hartmann, Jacob, and Weiß 2019). To ensure that the perspectives of all relevant stakeholders were taken into account, FitForFDM invited approximately 20 representatives of: the university board (Vice-President for Research, Transfer & Science Communications), central units (Information Technology Center, University Library and Central Office for Research and Transfer), all 5 faculties: researchers (professors and research associates) in national and international research projects, departmental leaders, research managers, open access and open science experts.

The initial picture of RDM practices and limitations that emerged from the combination of those information sources proved insufficient for the formulation of concrete recommendations that researchers could directly apply in their daily work. To augment this superficial description, numerous actors indirectly involved in the research process were additionally inquired about specific subtopics: grant support advisors from the central research service team, laboratory engineers, data- and cybersecurity officers, and IT support members.

2.1 Cooperating Research Projects

The cooperation with the chosen research projects was established already during the planning phase of the FitForFDM project. The partner projects were supposed to cover as broad as possible a variety of all kinds of research conducted at the university. The main selection criteria included project stage, funding agency, and different scientific disciplines.

The final set comprised projects funded by several funding agencies:

- European Commission (EC),

- German Research Foundation (DFG),
- Senate Department for Culture and Social Cohesion of the Land Berlin within the framework of the European Regional Development Fund (EU-EFRE),
- Institute for Applied Research in Berlin (IFAF-Berlin),

differing in size with local or remote working groups of 2-10 people and grant applications of four to more than 20 (one of NFDI[1] consortia) cooperating institutions. All project stages were represented shortly before or during the timeline of FitForFDM: planning/application, first/second half of the funding period, finished. The disciplinary variety was provided in the first line by involving one project from each faculty at the HTW Berlin. Each of the projects was chosen for its differentiating qualities in the hope of gaining unique insights:

1. School of Engineering – Energy & Information: A project gathering and modeling high-resolution real-time numerical data from a large array of power plant sensors;
2. School of Engineering – Technology & Life: A project collecting diverse types of data, from material characteristics through fluid dynamics to images, that needed to be stored and analyzed in context;
3. School of Business & Law: A project where a comparatively low volume of text data was generated, but the legal context of each document was of critical importance;
4. School of Computing, Communication & Business: A project working with massive data, for which a real-time preprocessing workflow was implemented to immediately delete large portions of the recordings, as saving all generated data would be technically infeasible;
5. School of Design & Culture: A project that yielded an amount of image and text data untypical for the discipline and in which archiving of the data was not yet finalized.

The methods used in the projects included wet laboratory experiments, digital modeling and test plant implementations, software development, digitization, and legal regulations investigation. Moreover, the projects identified themselves with keywords illustrated in Figure 1, from the HTW Berlin subjects catalog, which is based on the classification of the German Federal Office of Statistics (for current version see Statistisches Bundesamt 2024).

1 https://www.nfdi.de/; *Visited on June 4, 2025.*

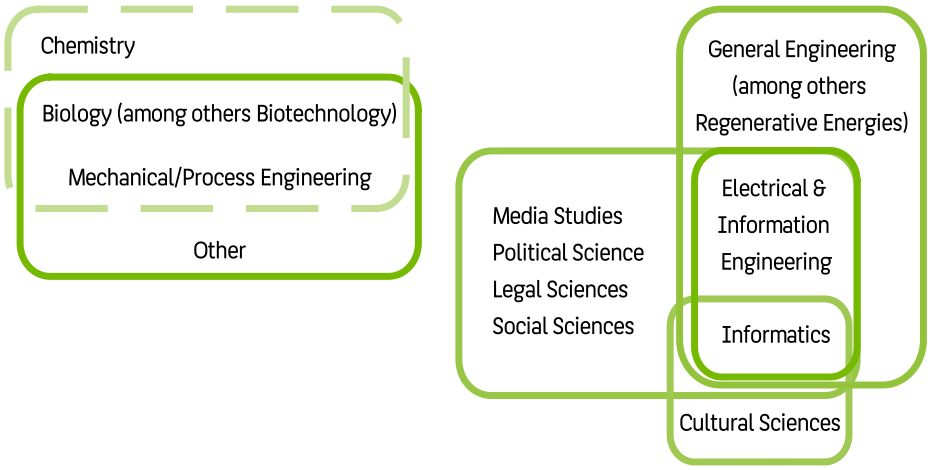


Figure 1: Keywords describing the five research projects cooperating with FitForFDM. Each frame corresponds to a single project, overlaps reflecting shared keywords. Dashed line indicates a follow up project to an originally chosen one, which started during the time span of FitForFDM.

Although the different characteristics of the projects presented a chance to gain insight into different requirements and ways of working with data, they should only be treated as examples; in order to derive general statements and identify opinions that represent outliers, a large number of similar projects would have to be examined. However, since this was not possible given the resources of the current project, a decision was made to favor breadth over depth of analysis.

2.2 Preparation of Interview Materials

The design of the interview outline began with a collection of broadly accepted recommendations and requirements for RDM in research projects. Among the resources considered were FAIR (Wilkinson et al. 2016) and CARE (Carroll et al. 2020) Principles, Good Scientific Practice Statute (Deutsche Forschungsgemeinschaft 2019; Hochschule für Technik und Wirtschaft Berlin 2022), requirements of research funding agencies (e. g. EC, DFG, BMBF), and the notion of a Research Data Life Cycle (see for example Hesse, Baier, and Schmidtke 2019, Figure 2; Cioffi, Goldman, and Marchese 2023). Another important aspect to consider was the specific situation of the HTW Berlin with its own organization of central and decentralized services, just starting on its path to institutional research data management. The questions of awareness, understanding, terminology, demands, and challenges stood in the foreground during the search for relevant topics. Using the SPSS method (Helfferich 2011), a hierarchical set of keywords was constructed. Based on them, questions were formulated, sorted, and grouped to form a systematic catalog, consequently integrated into an interview manual. Although the catalog was thematically

structured, the choice of exact questions and their order was to be determined during each semi-structured interview, according to its course and insights of every interviewee.

Since the interviews were to be audio-recorded, an appropriate consent declaration to the processing of personal data was prepared in consultation with the data protection officer, based on a template from Spiecker et al. (2022). In addition, an information sheet and a pre-interview form were created. The latter allowed for potential correction and addition of administrative information about the given project, further enabling custom adjustments to the interview manual adapted prior to each interview.

Once all necessary materials were ready, two dummy interviews were conducted. During the first test, the equipment and usability of the manual were proven. The second, full-scale, test interview involved questioning the FitForFDM member who did not participate in the development of the questions. This allowed to double check the technical set-up and if the questions were understandable and elicited informative answers. The time allotted for each interview appointment was one hour.

All documents and aids developed and used in the context of interviews, in particular, lists of questions and keywords, are available in Dąbrowska et al. (2025).

2.3 Assessment of the Interviews

Audio recordings in `.wma` format were transcribed with "f4x"[2] software, with customizable transcription rules (cf. Kuckartz and Rädiker 2019, pp. 41–42) chosen as follows:

- language: German,
- time marks: at paragraph end,
- pauses: single second resolution with points (.), (..), (...), etc.,
- word list: Forschungsdatenmanagement, FAIR, Prinzipien, CARE Prinzipien, Langzeitarchivierung, Datenintegrität, HTW Berlin, fachgebietsübergreifenden, große Fächerdiversität, Leseerkennungssoftware.

Manual corrections included amendments of clearly wrongly recognized words or expressions, false assignments of speakers, as well as additions of interjections (exclamations, sentimental words, e. g. "mhmhm" as expression of affirmation). The resulting transcripts were exported in `.docx` format and fed to "MaxQDA 2022"[3] software for analysis.

The categories used to code the data came from the keywords tree and the questions catalog developed in the course of the interview preparation phase, and comprised:

2 https://www.audiotranskription.de/f4x/; *Visited by the editors on July 8, 2025.*
3 https://www.maxqda.com/de/hilfe-mx22/willkommen; *Visited by the editors on July 8, 2025.*

- predefined requirements for research projects,
- FAIR & CARE Principles,
- data documentation including metadata,
- data handling,
- data access & publication,
- long-time preservation,
- competence building,
- needs & wishes.

The analysis was an iterative process, starting with an assignment of top-level categories to batches of text and formulation of exact definitions of all codes, adjusted to reflect the actual contents of the collected material. Additionally, all interviews were checked for any remaining transcription errors. Then a granular search was performed for segments related to particular subcodes. At the end of the analysis, codes applied only once or twice across all datasets were, if possible, replaced with their higher-level counterparts. In the next step, extracted text segments were translated into compressed statements directly related to predefined keywords and topics. Eventually, the conclusions of all interviews were collected and subsumed, taking the form of pseudonimized data presented in this work.

3 Results

The interviews were conducted from March to July 2023. The participating projects ran between 2013 and 2026 (latest planned end date: September 30, 2026), lasting 8 months to 5 years. The time distribution of all considered projects is illustrated in Figure 2. As shown there, some of the originally invited projects finished before the start of FitForFDM, which hindered the possibility of shadowing by the data scout. The shortest project was interrupted and closed prematurely, yet it granted an experience nobody could plan for. In the earliest project, on the other hand, the archival of a massive dataset was still to be performed, providing valuable insights. To compensate for the difficulty, the group leader of the third project offered their next proposal (dashed line in Figure 2) for shadowing during FitForFDM.

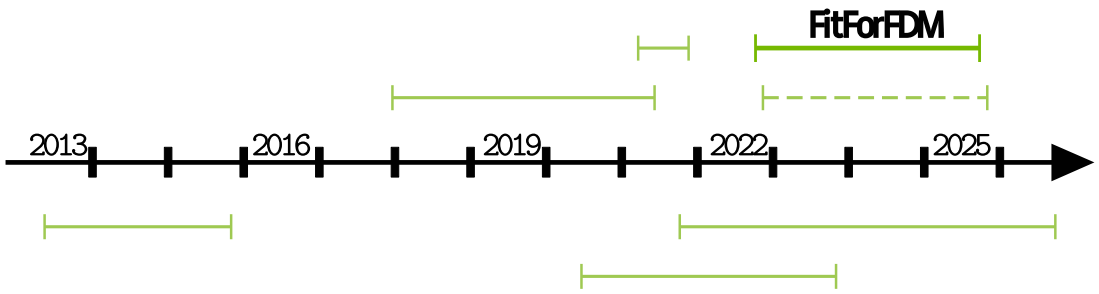


Figure 2: Timeline of involved research projects. Each green line corresponds to a single project, with FitForFDM indicated by an inscription. Dashed line denotes a project included later for shadowing.

A total of four professors and two research associates participated in the interviews (in one case two project members joined the interview). This differentiation exposed one of the contrasts between these two groups of researchers relevant to the topic of RDM: experienced project leaders knew a lot about the formalities around project planning, acquiring resources, reporting, and following various legal regulations, while young scientists could provide more information about data processing and management, technical setup, and practicalities of project realization.

Before insights from the semi-structured interviews are introduced, centrally available data services will be characterized. Next, compiled results of the analysis of all five interviews, grouped and ordered according to the initial keywords tree and the questions catalog, will be presented. Finally, the findings from talks with research support staff will be described.

3.1 Central Data Services

The Information Technology Center of the HTW Berlin offers several data storage solutions dedicated to various use cases. The best known among university members and intended for data sharing is HTW-Cloud. HTW-GitLab, appropriate for software development, is mostly utilized by IT experts among researchers. Since teaching often constitutes the main task for them, typically for a University of Applied Sciences, HTW-Moodle – a service used for management of teaching materials – was also included in this comparison. The last provision: HTW-Space, offers secure private storage for every university employee.

After consultations with both scientists and IT specialists, the criteria to describe data services, relevant for the former group and unambiguous for the latter, were formulated. Table 1 sums up the resulting delineation.

Table 1: Central Data Storage Services at the HTW Berlin.

Storage system	Appropriate for backup Automatic version control	Appropriate for personal data Space limit	Sharing
HTW-Cloud	× ✓ Up to 14 days	✓ Encrypted × Fair-Use Principle	✓ HTW-internally ✓ HTW-externally
HTW-GitLab	✓ ✓	× 100 MiB / File 10 GiB / GitLab-Project	✓ HTW-internally ✓ HTW-externally
HTW-Moodle	✓ ×	× 100 MB / File	✓ HTW-internally
HTW-Space	✓ ×	✓ 20 GB / User	× Not possible

In the case of GitLab, one feature additional to the listed in the table was repeatedly mentioned as requested by its users: code publishing, unfortunately not possible in the local instance.

3.2 Predefined Requirements for Research Projects

None of the approached projects used a data management plan (DMP). Only EC demanded inclusion of a separate DMP in a project proposal, but the document was static. None of the projects had applied for RDM funds, as these were either not available or not considered by the project leaders. Overall, the comprehensiveness of the RDM organization in the project planning stage was directly proportional to the complexity of the cooperation defined by the number of people or institutions involved and the possible presence of commercial partners. Software projects were a special case, since software development usually requires externally imposed data structures. Still, all interviewees admitted that discipline-specific good and common practices, also those unrelated to RDM, were not formalized within the projects and often implicit. Further, existing guidelines to specific aspects of research (e. g. digitization, FAIR Principles) were typically seen as superficial, lacking hints towards practical implementation. At the same time, existing HTW Berlin services (legal counseling, IT support) were assessed as functioning well.

3.3 FAIR & CARE Principles

Asked specifically about the FAIR Principles for scientific data management (Wilkinson et al. 2016), all participants demonstrated awareness and partial understanding of the term. Raw measurement data were typically considered unusable outside of the original context, and finding datasets from already closed projects (FA) was usually said to require direct contact with their creators (frequently already gone from the working group). However, several interviewees stressed the reuse of chosen experimental data (R) in student projects. In general, preprocessed data, models, and simulation codes were designated as potentially useful externally. Although steps fostering interoperability and reusability (IR) of particular research outputs were sometimes routinely taken, especially in the case of large cooperations, inadequate technical and content-wise support for implementation of FAIR was repeatedly reported as problematic. Yet, in two of the cooperating projects, provision of their products – digitized materials or new software – was among the main aims.

The same cannot be said about the CARE Principles, which were not heard of by any of the approached researchers. This result was not very surprising, though, since these principles were not applicable in the considered projects. However, the HTW Berlin regularly works with geographical, geological or socio-cultural datasets – all potentially relevant from the perspective of CARE Principles, so a bare awareness of their existence was not to be ruled out.

In the scope of data sharing and publication, legal sciences stood out as an exception. The Open Access movement, strong within natural sciences, did not stop the majority of digital resources in law (e.g. databases) from being available only upon (typically very costly) subscription yet (Fischer 2022).[4] Hampering the dissemination of any project results, some publishers in this discipline have not adapted the notion of digital object identifiers (DOIs) and continue to primarily distribute their periodicals and books only in paper form.

3.4 Data Documentation including Metadata

As already mentioned, inquired projects did not actively use DMPs, but the usual project management documentation was in place (work packages, tasks, deliverables...). Furthermore, some of the projects belonged to research domains in which specific research practices already include basic RDM elements. For example, automatically generated metadata, e. g. from measuring devices, were deliberately preserved, even though the term

4 For a comparison see https://open-access.network/en/information/subject-specific-open-access, and specifically for Legal Sciences https://open-access.network/en/information/subject-specific-open-access/law; *Visited on April 2, 2025.*

"metadata" in the context of RDM was by itself unfamiliar. People experienced with software often routinely wrote README files to describe their data and data structures, and everybody enumerated the information of *who* and *when* performed or produced something as particularly worth storing. Nonetheless, the desired data documentation framework was commonly determined not earlier then at the stage of publication, indirectly leading to detailed information often being available only in articles or dissertations. A brief introduction to the idea of systematic management and documentation of research activities brought out the impression that properly documenting all procedures would require far too much time and that nobody would read the resulting records anyway.

3.5 Data Handling

A strong tendency of almost all interviewees was to equate *data* solely with measurement data (including numerical, audio, video or graphics). After clarification that in the context of RDM all digital material used for research and produced by it (e. g. software and all kinds of text documents) is to be considered, a zoo of data formats and storage services within each single project was revealed. Commonly used proprietary tools gave rise to files in specific formats, sometimes additionally converted to open formats and assisted by numerous scripts, reports, and intermediate versions of partially processed results. Many researchers used storage services custom-built within the projects or purchased from external providers. Still, HTW-Cloud was consistently recognized as very convenient for data sharing and easy to use. On top of the already mentioned, local storage resources assigned to faculties, laboratories or degree programs were employed, although low availability of technical support turned them unusable to the less IT-adept. Repositories of project partners also played a role.

As already suggested in earlier sections, RDM was not explicitly defined, with the exception of data-heavy and software projects or their parts. Vital information was often shared only via direct communication. The responsibilities for various aspects of data processing and curation were dispersed and not formalized. An additional difficulty, common in research, but recognized only by one of the participating projects, was the duality of physical objects and digital data associated with them.

3.6 Data Access & Publication

The constellation of storage services in use was usually nowhere documented; only current project members had an overview of relevant digital locations and people responsible for access management. While services internal to the HTW Berlin allowed easy access regulation by users and were regularly employed both in research and teaching, external

storage was dedicated to scientific data and access management rested with single working group members, current or past.

Projects aiming to openly provide certain outputs faced the problem of lacking resources for long-term maintenance of the materials. Others rarely considered a data publication; customarily, processed data were put out in the form of graphs and summary statistics in articles, with actual numerical values never made publicly available. One of the researchers, more experienced in publishing data, expressed the opinion that good tools are already available, but standards regarding folder structures and similar details would ease the whole process.

3.7 Long-time Preservation

Similarly to the structure of data documentation, the question of long-term preservation (LTP) of research data was typically tackled at the end of a project. Again, in the case of long-term provision of the results being among the main goals, LTP was taken into account already at the stage of planning and the documentation was adjusted to the requirements imposed by the envisioned LTP solution. Nonetheless, at the time of the interviews, none of the participating projects had successfully deposited their data in accordance with the ten-year preservation period prescribed by the Deutsche Forschungsgemeinschaft (2019).

3.8 Competence Building

Despite the fact that the HTW Berlin offers an introductory course in research data management yearly, none of the interviewed scientists has ever attended a similar event. However, everyone expressed a strong wish for learning materials in various forms: introductory and domain-specific courses, always available resources for self-learning as well as ready components implementable in the teaching of students. Beyond these, a need for supporting materials in the form of templates or check-lists was articulated, along with elaborate counseling. In all these contexts, examples – both general and discipline-specific – were considered essential to make the provided resources usable.

3.9 Needs & Wishes

In the final part of the interviews, researchers were encouraged to describe their visions, opinions, and experiences loosely related to the topic of RDM. In addition to the earlier indicated need for more technical support in managing the existing infrastructure,

interviewees emphasized the importance of clearly distinguishing between RDM and cybersecurity. The importance of coordinating RDM efforts at the federal state level was pointed out, particularly with regard to the organization of archiving services for research. At the level of the HTW Berlin, the wish for a defined standard for research and student projects structure was mentioned. Asked about the perfect scenario for the local RDM framework, the interviewees expressed contradictory ideas: from completely centralized (storage) services, through centralized, but not hierarchical structure, to faculty- or degree-program-specific organization. One of the participants cited the experience that colleagues often preferred to build and adjust infrastructure themselves, customizing it freely for every scientific endeavor.

3.10 Further Insights

Conversations with other members of the university, in particular laboratory engineers, exposed a broader picture of the research organization. Firstly, members of every faculty used a plethora of data types and formats, only a few being specific to one discipline or group. Secondly, not surprisingly, the structure of university IT services proved to be hierarchical, with the top-level IT Center, followed by faculty IT services, and infrastructures belonging to degree programs and single laboratories. What turned out to be problematic was the absence of a single information hub on all of those resources or of a standard structure for scattered information. This situation typically led researchers to consult technical professionals in their local environment, potentially missing helpful solutions available in other parts of the university or faculty.

Another observation, which emerged as more and more stakeholders were consulted, was the significant differences in perspective. While researchers' requests were guided by practical needs and codes such as the FAIR Principles and data protection regulations, IT specialists were usually not aware of the implicit expectations stemming from Good Scientific Practice guides and focused only on the purely technical requirements and cybersecurity. The confusion grew if the scientist or both the scientist and the laboratory engineer approached was not a specialist in IT. In other words, a problem arose when non-IT- and non-RDM-specialists gathered to set up IT for RDM.

One more incongruence in the terminology used by different parties was the topic of long-term preservation and archiving of data. For digital preservation specialists (e. g. some librarians), the latter is inseparable from active curation of the records and clear separation of master and derivative copies. Researchers, on the other hand, repeating the formulations from the new guidelines, usually meant some service to deposit their project data for the prescribed 10 years period, which, incidentally, is not considered particularly *long-term* by the first group.

4 Summary & Conclusions

The FitForFDM Project placed a strong emphasis on the participatory aspect of its endeavor. Involvement of numerous stakeholders through individual inquiries about specific subtopics allowed to realistically determine the actual landscape and the requirements for institutional RDM.

As could be expected, the spread of needs proved to be broad. Although some of the services existing at the HTW Berlin were repeatedly praised, the discussions revealed challenges in securing and exploiting appropriate data storage solutions, both during and after a research project. Unavoidable misunderstandings in communication between experts from dissimilar disciplines (scientists, IT-support staff, data protection officers, librarians) hindered optimal organization of data management. In particular, the confusion between data protection and data security, and data storage, LTP and archiving call for clarification and unification in the university landscape. The frequently signaled intertwining of research and teaching activities characteristic at German Universities of Applied Sciences put additional pressure on the provided services and infrastructure to ensure safe and flexible environment equally adapted to both.

The existing recommendations and guidelines were described as too generic and could not be implemented effectively due to the lack of both expertise and specialized RDM counseling. In addition, insufficient technical and personal resources restricted the efforts towards long-term provision and maintenance of particular types of research outputs. Furthermore, systematic documentation of research activities in the digital space has not yet become habitual for all scientists and often appeared intimidating, labor-intensive, and time-consuming beyond reason. Nevertheless, academics expressed strong interest in aids for RDM in the form of templates or check-lists, along with training materials for current and future researchers, personal consultation, and dedicated RDM assistance. Just the introduction of some terms, notions and already existing services, e. g. CARE Principles or European Open Science Cloud[5], sparked awareness and interest in learning more among some interviewees.

The picture emerging from the discussions described above contains clear hints of most pressing needs and necessary changes for the arrangement of RDM services at the HTW Berlin, but also at any University of Applied Sciences. However, the authors recommend a few improvements of the presented methodology to anyone interested in reusing it in their local setting.

Firstly, the sets of projects to interview and to shadow should be separated. In the first group, the variety of project stages plays a vital role, while in the second the possibility to participate in ongoing scientific activities is the most important. Secondly, as can be seen

5 https://open-science-cloud.ec.europa.eu/services and https://www.openaire.eu/openaire-and-eosc; *Visited by the editors on June 4, 2025.*

in Figure 1, majority of the projects cooperating with FitForFDM identified themselves with Informatics, which was not the intention of the project initiators. While talking about storage services and technical setup for digital data with IT-literate interlocutors might be easier, such a choice can potentially bias the resulting impressions, concealing the extent of technical support needed on average. Worth stressing is the openness and flexibility in contacts with all stakeholders, as they often talk about the same issues using slightly different languages or vice versa. Respect and acknowledgement of the steps already taken and promoting good RDM reinforce good atmosphere and subsequent participation of the approached individuals.

Acknowledgements

The authors thank Prof. Stephanie Molthagen-Schnöring for support, supervision, and active contributions to the FitForFDM Project, among others the conceptualization and acquisition of funding.

The Project "Partizipative Entwicklung eines nachhaltigen Handlungsrahmens zum FAIR-en fachgebiets-übergreifenden Forschungsdatenmanagement (FitForFDM)" received funding from the German Federal Ministry of Education and Research, grant number 16FDF-H114, and the European Union within the program "NextGenerationEU".

Authorship Contributions

Contributions of each author according to the CRediT Contributor Role Taxonomy:

- Paulina Anna Dąbrowska: Methodology, Investigation, Data curation, Formal analysis, Visualization, Writing – original draft;
- Esther Schneidenbach: Methodology, Investigation, Data curation, Project administration, Writing – review & editing;
- Eske Heister: Conceptualization, Methodology, Investigation, Data curation, Writing – review & editing;
- Lena Simon: Conceptualization, Funding acquisition, Methodology, Project administration, Writing – review & editing;
- Wojtek Dabrowski: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Conflict of Interest

The authors declare no conflicts of interest.

Bibliography

Bundesministerium für Bildung und Forschung. 2020. *Aktionsplan Forschungsdaten.* Bundesministerium für Bildung und Forschung. Visited on March 5, 2025. https://www.bildung-forschung.digital/digitalezukunft/shareddocs/Downloads/files/163_20_faktenblatt_aktionsplan_3_final.pdf?___blob=publicationFile&v=2.

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, et al. 2020. "The CARE Principles for Indigenous Data Governance". *Data Science Journal* 19. https://doi.org/10.5334/dsj-2020-043.

Cioffi, Matt, Julie Goldman, and Sarah Marchese. 2023. *Harvard Biomedical Research Data Lifecycle.* Zenodo. https://doi.org/10.5281/ZENODO.8076168.

Dąbrowska, Paulina Anna, Eske Carmen Heister, Esther Schneidenbach, Lena Simon, and Piotr Wojtek Dabrowski. 2025. *Projekt "FitForFDM": Materialien zur Durchführung und Auswertung von Interviews zu den Begleitprojekten.* Zenodo. Dataset. Version v1. https://doi.org/10.5281/zenodo.15482294.

Deutsche Forschungsgemeinschaft. 2019. *Guidelines for Safeguarding Good Research Practice. Code of Conduct.* https://doi.org/10.5281/ZENODO.14281892.

Fischer, Georg. 2022. "Im Ringen um Erkenntnis und Anerkennung: Wie Rechtswissenschaftler*innen das eigene akademische Publizieren im Zuge von Open Access sehen". *RuZ – Recht und Zugang* 3 (1): 19–49. https://doi.org/10.5771/2699-1284-2022-1-19.

Hartmann, Niklas K., Boris Jacob, and Nadin Weiß. 2019. *RISE-DE – Referenzmodell für Strategieprozesse im institutionellen Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.3585556.

Helfferich, Cornelia. 2011. *Die Qualität qualitativer Daten.* 4th Edition. 182–189. Springer Fachmedien Wiesbaden: VS Verlag für Sozialwissenschaften. ISBN: 978-3-531-17382-5. https://doi.org/10.1007/978-3-531-92076-4.

Hesse, Elfi, Juliane Baier, and Knut Schmidtke. 2019. *FDM-Handbuch für HAW – Handlungshilfe für aktives Forschungsdatenmanagement an Hochschulen für angewandte Wissenschaften.* Dresden. Visited on June 4, 2025. http://nbn-resolving.de/urn:nbn:de:bsz:520-qucosa2-365568.

Hey, Tony, Stewart Tansley, Kristin Tolle, and Jim Gray. 2009. "Jim Gray on eScience: A Transformed Scientific Method". In *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research. ISBN: 978-0-9825442-0-4, visited on June 4, 2025. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

Hochschule für Technik und Wirtschaft Berlin. 2022. "Satzung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten an der HTW Berlin". In *Amtliches Mitteilungsblatt 14/22.*

Kuckartz, Udo, and Stefan Rädiker. 2019. *Analyzing Qualitative Data with MAXQDA.* 1st edition. Springer Cham. ISBN: 978-3-030-15670-1. https://doi.org/https://doi.org/10.1007/978-3-030-15671-8.

Spiecker, Claus, Janina Richter, Paul Walter, Maria Chlastak, Christine Burkart, Esther Schneidenbach, Reinhard Messerschmidt, Kirsten Endres, Christian Djeffal, and Eva Nürnberger. 2022. *EcoDM – Ökosystem Datenmanagement: Analysen – Empfehlungen – FAIRifizierung.* Zenodo. https://doi.org/10.5281/ZENODO.6256398.

Statistisches Bundesamt. 2024. *Systematik der Fächergruppen, Lehr- und Forschungsbereiche und Fachgebiete.* Visited on March 17, 2025. https://www.destatis.de/DE/Methoden/Klassifikationen/_inhalt.html.

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. https://doi.org/10.1038/sdata.2016.18.

Institutional Data Science Services at FDZ UB Mannheim: Enhancing Research Data Management

Renat Shigapov [1], Irene Schumm [1], Thomas Schmidt [1], Jan Kamlah [1], Larissa Will [1]

[1] University Library, University of Mannheim

Data science services have rapidly emerged as essential support mechanisms for research data management (RDM) at universities worldwide. In alignment with these developments, the research data center (FDZ) at the Mannheim University Library (UB Mannheim) has established institutional data science services at the University of Mannheim. Our goal is to enhance RDM, promote open science, and contribute to research reproducibility. FDZ supports researchers across the complete data science pipeline, emphasizing advanced data collection methods (OCR, audio/video transcription, API integration, web scraping) and chatbot and large language model (LLM) applications. Our services include data-science-as-a-service, education and training, consultancy, community engagement, and integration of data science methods within internal library workflows. In this paper we introduce our services, describe the challenges we faced, and analyze the impact these services have had on RDM at the University of Mannheim.

Keywords: data science services, RDM services, AI, research support

1 Introduction

Data Science Services have rapidly emerged as essential support mechanisms for research data management (RDM) at universities worldwide. Notable examples include services

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23916 (CC BY-SA 4.0).

at Harvard University[1], the University of Utah[2], Purdue University[3], NC State University[4], the University of Arizona[5] (Oliver et al. 2019), and the University of Groningen[6]. In Germany, several data science initiatives have recently been established. These include the Data Science Center at the University of Bremen (Steinmann, Thöricht, et al. 2023), the Bielefeld Center for Data Science (Weiß and Decker 2023), the consortium NFDI4DataScience (Schimmler et al. 2023) within the national research data infrastructure (Hartl, Wössner, and Sure-Vetter 2021) and some others listed in Steinmann, Nowotka, et al. (2023).

The latter describes the Mannheim Center for Data Science (MCDS)[7], established in 2020. The MCDS serves as a platform to connect and enhance research in and through data science at the University of Mannheim. It organizes courses, networking events, and lectures, contributing to science communication by regularly sharing information on new projects and developments in data science. Due to limited human resources, the MCDS does not offer data-science-as-a-service, data science consultancy, education and training on data science including research data management (RDM), or IT infrastructure. To address these gaps at least partially, the Research Data Center (FDZ) at the University Library Mannheim (UB Mannheim) created institutional data science services in 2024.

In this paper, we introduce the FDZ and its data science services, describe the challenges we have faced, and analyze the impact of these services on research data management at our institution.

2 FDZ UB Mannheim

The FDZ serves as the primary contact point and support hub for all research data management (RDM) needs at the University of Mannheim. Given the university's leading role in economics and social sciences, the FDZ focuses particularly on addressing the spe-

1 Data Science Services at Harvard University, https://www.iq.harvard.edu/data-science-services; *Visited on March 30, 2025.*

2 Data Science Services at University of Utah, https://cores.utah.edu/data-science-service; *Visited on March 30, 2025.*

3 Data Science Consulting Service at Purdue University, https://www.purdue.edu/data-science/dscs; *Visited on March 30, 2025.*

4 Data Science Consulting Team at NC State University, https://datascienceacademy.ncsu.edu/data-science-consulting/data-science-consulting-team; *Visited on March 30, 2025.*

5 Data Science & Visualization services at University Libraries, https://data.library.arizona.edu/data-science; *Visited on March 30, 2025.*

6 Data Science Team in the Center for Information Technology at University of Groningen, https://www.rug.nl/society-business/centre-for-information-technology/research/services/data-science; *Visited on March 30, 2025.*

7 Mannheim Center for Data Science, https://www.uni-mannheim.de/datascience; *Visited on March 30, 2025.*

cific needs and challenges of these fields, while also supporting interdisciplinary research efforts.

The FDZ provides a comprehensive suite of core RDM services, including the Mannheim data repository (MADATA), RDM consultancy and educational programs, support for creating data management plans (MADataPlan), and a secure data-sharing environment. Additionally, the FDZ plays an active role in multiple interdisciplinary projects including data literacy (MaDaLi^2)[8], chatbots for RDM (Schmidt et al. 2025; Shigapov and Schumm 2024), optical character recognition (OCR-BW[9]), knowledge graphs (Rossenova et al. 2023; Limani et al. 2023), BERD@BW (Gehrlein, Schumm, and Shigapov 2022), and BERD@NFDI (Gehrlein et al. 2022).

Participation in these diverse initiatives has significantly boosted the FDZ's internal expertise, enabling the center to attract highly qualified personnel specialized in data science and advanced RDM practices. This has led to the establishment and expansion of consultancy services focusing on OCR techniques, AI-driven methodologies, and advanced research data management.

3 Data Science Services

The Data Science Services at the FDZ include four core offerings: data-science-as-a-service, education and training, consultancy, and community engagement and networking.

Data-science-as-a-service. Our core offerings include automatic data collection methods such as Optical Character Recognition (OCR), audio/video transcription, web scraping, data acquisition via APIs, and transforming unstructured data into structured formats. Recently, we have expanded our scope by implementing data science pipelines, developing specialized chatbots, and collaboratively authoring data papers.

Education and training. Our education and training services include regular courses and hands-on workshops covering topics such as chatbots for research, AI tools for academic studies, open-source AI assistants for research data management (RDM), OCR applications in research, and automatic audio/video transcription. These workshops are integrated into our established training series, including "Research Skills" and specialized "RDM Seminars."

8 https://www.uni-mannheim.de/universitaet/lehre/lehrprojekte/data-literacy; *Visited on March 30, 2025.*

9 https://ocr-bw.bib.uni-mannheim.de; *Visited on March 30, 2025.*

Consultancy. We offer consultancy through multiple channels, including official FDZ email, direct personal contact with FDZ staff, subject librarian referrals, and participation in related courses and workshops. Consultations are provided through email, virtual meetings, and in-person sessions, complemented by regular consultancy hours dedicated to topics such as OCR and knowledge graphs.

Community engagement and networking. We actively participate in existing data science networking activities, particularly those organized by the MCDS. This includes contributing to the MCDS academic speed-dating events each semester, which promote collaboration among university researchers. Additionally, we engage in university-wide events such as the GESS Research Day, presenting our services directly to the research community.

Additionally, establishing these data science services enabled us to optimize internal library workflows through data science methods. This includes developing chatbots for general library use and specialized chatbots for the FDZ addressing research data management (RDM) and open science inquiries. We also work on automatic subject indexing tasks using Dewey Decimal Classification (DDC) and Regensburger Verbundklassifikation (RVK) systems, streamline RDM processes within FDZ (such as metadata review and creating data documentation), and represent the library in the university's strategic AI initiatives, including the development of the institutional AI strategy.

4 Challenges

Establishing institutional data science services at the FDZ involved several challenges. A key issue concerning the emerging field of AI applications within the data science services of the FDZ was the initial lack of an overarching institutional AI strategy at the University of Mannheim. We started building our data science services independently, which led to uncertainty regarding alignment with broader institutional goals. However, the FDZ has since actively contributed to the development of the university-wide AI strategy, ensuring our services align closely with institutional priorities and standards.

Another considerable hurdle is the high infrastructural costs, particularly related to the need for computational resources such as GPUs, essential for delivering advanced AI-based services frequently requested by researchers. Nevertheless, the university IT plans to expand GPU infrastructure, promising support for future projects.

Personnel costs also presented significant challenges. Operating and maintaining sophisticated data science services demand highly skilled staff, which increases operational costs in the long term. To ensure sustainability, the FDZ is actively seeking more AI, data sci-

ence, and RDM-related projects, along with internal and external funding opportunities to ensure consistent support and development of these services.

Finally, scalability represents a critical challenge due to the potentially high costs associated with expanding data-science-as-a-service. To mitigate scalability issues, the FDZ is exploring multi-tier support models and strategic project partnerships. By partnering directly in projects, the FDZ can split costs, optimize resources, and sustainably grow its data science service offerings.

5 Impact on RDM

Since our data science services were established only recently, we currently provide a qualitative analysis of their initial impact on RDM at the University of Mannheim.

The primary beneficiaries of our data-science-as-a-service offerings are researchers who lack the resources or expertise to manage complex data science tasks independently. Our data collection services, including OCR, audio/video transcription, web scraping, and data structuring, have increased the availability and accessibility of FAIR-compliant research data within the university. By actively engaging in data collection, we also directly support and advise researchers on best practices in FAIR data sharing through data repositories and data papers, enhancing overall transparency and reproducibility of research projects.

Recognizing higher participation rates in AI-focused workshops compared to traditional RDM sessions, we incorporated essential RDM concepts into our AI-oriented training programs. This integrated approach broadens our outreach, ensuring that a wider group of researchers develops competencies in both AI methods and RDM practices.

Moreover, automating RDM workflows (such as metadata validation and the creation of data documentation) has improved our internal processes at the FDZ of the UB Mannheim. This automation reduces the administrative burden, resulting in faster metadata checks and smoother data submission and publication.

6 Discussion

Currently, the University of Mannheim hosts two distinct structures supporting data science activities: the MCDS and the FDZ. While the MCDS operates as a researcher-driven initiative, the FDZ serves as an infrastructural department. Despite their separate organizational setups, these entities effectively complement one another. For instance,

data scientists from FDZ actively participate in the academic speed dating events on data science organized each semester by the MCDS.

Although Putzke et al. (2023) suggest that "data science roles should be housed in newly created, independent data science centers," our experience at the University of Mannheim demonstrates that parallel structures, such as those embodied by the MCDS and the FDZ, can also yield substantial benefits, provided their respective services are clearly defined and mutually complementary.

The integration between data science services and RDM services occurs naturally within our framework, as our data scientists also function as data stewards (RDM consultants). This dual role is highly beneficial, ensuring that every data science consultation inherently addresses RDM principles and practices, and vice versa.

7 Conclusions

Establishing institutional data science services at the FDZ has had a positive impact on research data management (RDM) practices at the University of Mannheim. Despite initial challenges (including infrastructure costs, staffing requirements, and the lack of a comprehensive institutional AI strategy), our services have quickly demonstrated benefits. By providing practical data science solutions, RDM-included AI training, and consultancy, we have expanded the amount of available research data, improved researcher competencies in data stewardship, and streamlined internal RDM workflows.

The dual role of our data scientists as RDM consultants has ensured the integration of data science and RDM practices, enhancing researcher awareness of FAIR principles and open science standards. Going forward, continued collaboration with parallel structures such as the MCDS, strategic infrastructure investment, and sustained resource planning will be critical. Further quantitative evaluation will also be essential to fully measure long-term impact and refine our service offering.

Acknowledgements

The authors thank the whole FDZ and UB Mannheim team for their contributions and support.

Authorship Contributions

- Renat Shigapov: Writing – original draft
- Irene Schumm: Writing – review & editing
- Thomas Schmidt: Writing – review & editing
- Jan Kamlah: Writing – review & editing
- Larissa Will: Writing – review & editing

Conflict of Interest

The authors are not aware of any potential conflicts of interest.

Bibliography

Gehrlein, Sabine, Annette Klein, Irene Schumm, and Klaus Tochtermann. 2022. "BERD@-NFDI für unstrukturierte Daten in den Wirtschafts- und Sozialwissenschaften". *Zeitschrift für Bibliothekswesen und Bibliographie* 69 (1–2): 59–67. https://doi.org/10.3196/1864295020691280.

Gehrlein, Sabine, Irene Schumm, and Renat Shigapov. 2022. "BERD@BW – A Science Data Center to foster Open Science in Business, Economics and Social Sciences". In *E-Science-Tage 2021: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 314–319. heiBOOKS. ISBN: 978-3-948083-55-7. https://doi.org/10.11588/HEIBOOKS.979.C13743.

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

Limani, Fidan, Renat Shigapov, Auriol Degbelo, Sonja Schimmler, Benjamin Zapilko, Lozana Rossenova, Stefan Dietze, Konrad U. Förstner, Saurav Karmakar, and Andreas Czerniak. 2023. *Who is using Knowledge Graphs in NFDI? An overview by the Working Group "Knowledge Graphs".* Zenodo. https://doi.org/10.5281/ZENODO.8332776.

Oliver, Jeffrey C., Christine Kollen, Benjamin Hickson, and Fernando Rios. 2019. "Data Science Support at the Academic Library". *Journal of Library Administration* 59 (3): 241–257. https://doi.org/10.1080/01930826.2019.1583015.

Putzke, Johannes, Thomas Schimmer, Manuela Richter, and Lucas Hamel. 2023. “Towards the Establishment of Data Science Centers at Higher Education Institutions”. In *INFORMATIK 2023 – Designing Futures: Zukünfte gestalten,* 117–121. Gesellschaft für Informatik e.V. https://doi.org/10.18420/INF2023_10.

Rossenova, Lozana, Renat Shigapov, Moritz Schubotz, and Daniel Mietchen. 2023. *KGI4-NFDI: Knowledge Graph Infrastructure for the German National Research Data Infrastructure.* Zenodo. https://doi.org/10.5281/ZENODO.8337432.

Schimmler, Sonja, Bianca Wentzel, Arnim Bleier, Stefan Dietze, Saurav Karmakar, Peter Mutschke, Angelie Kraft, et al. 2023. “NFDI4DS Infrastructure and Services”. In *INFORMATIK 2023 – Designing Futures: Zukünfte gestalten.* Gesellschaft für Informatik e.V. https://doi.org/10.18420/INF2023_103.

Schmidt, Thomas, Renat Shigapov, Irene Schumm, and Jan Kamlah. 2025. *Perplexity-inspired metasearch-based alternatives to FAIR GPT: Open-source AI consultants for RDM.* Zenodo. https://doi.org/10.5281/ZENODO.15038486.

Shigapov, Renat, and Irene Schumm. 2024. “FAIR GPT: A virtual consultant for research data management in ChatGPT”. *arXiv,* https://doi.org/10.48550/ARXIV.2410.07108.

Steinmann, Lena, Dirk Nowotka, Lea Oberländer, Helen Pfuhl, Heiner Stuckenschmidt, and Rolf Drechsler. 2023. “Workshop: „Aktuelle Entwicklungen und Perspektiven (an Hochschulen) im Bereich Data Science“”. In *INFORMATIK 2023 – Designing Futures: Zukünfte gestalten.* Gesellschaft für Informatik e.V. https://doi.org/10.18420/INF2023_05.

Steinmann, Lena, Heike Thöricht, Sandra Zänkert, and Rolf Drechsler. 2023. “Das Data Science Center an der Universität Bremen: Interdisziplinärer Knotenpunkt und Service-Infrastruktur für die datenintensive Forschung”. In *E-Science-Tage 2023: Empower Your Research – Preserver Your Data,* edited by Vincent Heuveline, Nina Bisheh, and Philipp Kling, 246–252. heiBOOKS. https://doi.org/10.11588/HEIBOOKS.1288.C18083.

Weiß, Katharina, and Reinhold Decker. 2023. “Data Science und Forschungsdaten als Treiber für wissenschaftliche Interdisziplinarität”. In *INFORMATIK 2023 – Designing Futures: Zukünfte gestalten.* Gesellschaft für Informatik e.V. https://doi.org/10.18420/INF2023_09.

Forschungsdatenmanagement in Nordrhein-Westfalen: Das Landeskonzept FDM und die Verstetigung der Landesinitiative fdm.nrw

Magdalene Cyra [1], Matthias Fingerhuth [1], Lioba Schreyer [1], Christina Schröder [1]

[1] Landesinitiative fdm.nrw

Die Herausforderungen im Forschungsdatenmanagement sind zahlreich und reichen von infrastrukturellen und technischen bis hin zu strategischen Themen. Um diesen Herausforderungen zu begegnen, werden in Nordrhein-Westfalen hochschulübergreifende Kooperationen gefördert, die zuletzt in einem moderierten Strategieprozess zusammengeführt wurden. Als Ergebnis dieses Strategieprozesses wurde ein Landeskonzept Forschungsdatenmanagement verabschiedet. In diesem sind die Leitplanken für die Entwicklung des Forschungsdatenmanagements an den Hochschulen in Nordrhein-Westfalen festgehalten. Neben Aufbau und Betrieb von hochschulübergreifenden, generischen Infrastrukturen werden auch die Anbindung an übergeordnete Entwicklungen insbesondere der Nationalen Forschungsdateninfrastruktur (NFDI) und die konkrete Umsetzung des Landeskonzepts Forschungsdatenmanagement beschrieben. Um die zahlreichen Handlungsstränge zusammenzuführen, wurde die bereits etablierte Landesinitiative für Forschungsdatenmanagement – fdm.nrw verstetigt. Das Landeskonzept FDM formuliert außerdem neue Aufgaben und Verantwortlichkeiten für die Landesinitiative fdm.nrw. Die Entstehung des Landeskonzepts und die Verstetigung der Landesinitiative in Nordrhein-Westfalen sind auf spezifische Rahmenbedingungen zurückzuführen. Dennoch lassen sich die abschließenden Erkenntnisse auf andere Rahmenbedingungen übertragen.

Keywords: Landesinitiative, Landesstrategie, Verstetigung, Kooperation, Synergie

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23917 (CC BY-SA 4.0).

1 Einleitung: Herausforderungen im FDM

Das Management von Forschungsdaten geht mit zahlreichen Herausforderungen einher. Zu diesen zählen zunächst der Aufbau und der Betrieb sowohl technischer als auch nichttechnischer Infrastrukturen zur Unterstützung der Forschenden. Es ist ebenso herausfordernd wie notwendig, übergeordnete Entwicklungen wie die Nationale Forschungsdateninfrastruktur (NFDI) voranzutreiben und gleichzeitig einzelne Personen und Institutionen zu beteiligen. Der Umstand der oftmals projektfinanzierten Fortentwicklung des Forschungsdatenmanagements (FDM) trägt seinen Anteil durch wechselndes Personal und zeitlich begrenzte Umsetzungszeiträume zur herausfordernden Gesamtlage bei. Die Aufzählung der Herausforderungen ist an dieser Stelle nicht abschließend, sondern lediglich exemplarisch, verdeutlicht aber, dass Lösungen dringend benötigt werden. Einen Lösungsansatz bieten hochschulübergreifende Kooperationen, wie sie in Nordrhein-Westfalen (NRW) gefördert werden. In diesem Beitrag geht es um zwei zentrale Elemente zur Umsetzung dieses Lösungsansatzes: das Landeskonzept FDM und die verstetigte Landesinitiative fdm.nrw. Im folgenden Abschnitt geben wir zunächst einen kurzen Überblick über die Hochschullandschaft in NRW sowie die Anfänge der Landesinitiative fdm.nrw. Dies ist der Hintergrund, der zur Darstellung des landesweiten Strategieprozesses und zur Verfassung eines Landeskonzepts FDM sowie zur Verstetigung der Landesinitiative für Forschungsdatenmanagement – fdm.nrw als Landesgeschäftsstelle, ohne eigene Rechtsform und weiterhin angesiedelt an der Universität Duisburg-Essen, geführt hat. Abschließend reflektieren wir über Erkenntnisse aus diesem Prozess.

2 Ausgangslage in Nordrhein-Westfalen und Projektzeit der Landesinitiative fdm.nrw

Die Hochschullandschaft in Nordrhein-Westfalen bedarf aufgrund ihrer Größe und ihrer Diversität einer besonderen Struktur zur standortübergreifenden Zusammenarbeit. Es gibt 14 öffentlich-rechtliche Universitäten, 15 öffentlich-rechtliche Hochschulen für angewandte Wissenschaften und sieben Kunst- und Musikhochschulen sowie weitere private und kirchliche Hochschulen mit Hauptsitz in NRW. Laut Statistik.NRW studierten im Wintersemester 2023/2024 über 750.000 Personen in NRW[1], davon über 41.000 Promovierende[2] und über 60.000 Personen waren hauptberuflich wissenschaftlich oder künstlerisch tätig[3]. Die vielen Hochschulen weisen dabei zusammengenommen ein enormes fachliches Spektrum auf. Um zu gewährleisten, dass den Hochschulen unabhängig von Größe oder

1 https://statistik.nrw/gesellschaft-und-staat/bildung-und-kultur/hochschulen/studierende-nach-hochschularten; *Besucht am 18. März 2025.*

2 https://statistik.nrw/gesellschaft-und-staat/bildung-und-kultur/hochschulen/promovierende-nach-faechergruppen-und-geschlecht; *Besucht am 18. März 2025.*

3 https://statistik.nrw/gesellschaft-und-staat/bildung-und-kultur/hochschulen/personal-an-hochschulen; *Besucht am 18. März 2025.*

Ausrichtung qualitativ hochwertige und sichere digitale Dienste und Infrastrukturen zur Verfügung stehen, fördert das Ministerium für Kultur und Wissenschaft des Landes NRW (MKW) die Digitalisierung mit der sog. Digitalisierungsoffensive. Für dieses Vorhaben wurde die Digitale Hochschule NRW (DH.NRW) eingerichtet, eine Kooperationsgemeinschaft des MKW und 41 Mitgliedshochschulen. Im Rahmen der DH.NRW wird das Ziel verfolgt, in hochschulübergreifenden Kooperationen Synergie- und Skaleneffekte zu fördern. Unter anderem wird der Aufbau und der Betrieb landesweiter IT-Dienste verfolgt, die allen Hochschulen in NRW den Zugang zu gemeinsamen Services und Infrastrukturen ermöglichen. Hochschulübergreifenden Kooperationen kommt damit in Zeiten knapper Ressourcen und gleichzeitig steigender Anforderungen eine besondere Bedeutung zu.

Die Anfänge der Landesinitiative fdm.nrw sind aus diesem Kontext erwachsen und gehen bis in das Jahr 2015 zurück. Bereits zu dieser Zeit wurde in NRW der Mehrwert hochschulübergreifender Kooperation im FDM erkannt. Eine für das Thema eingerichtete landesweite Arbeitsgruppe erhielt den Auftrag, Handlungsempfehlungen für den Aufbau von FDM-Strukturen für die Übertragbarkeit auf das Land NRW zu evaluieren und ein zentrales Informationsangebot über die Aktivitäten zum FDM an den Hochschulen in NRW zu schaffen. Darauf aufbauend formulierte diese Arbeitsgruppe die Empfehlung, eine eigens für das Thema FDM zuständige Landesinitiative zu gründen, die die Hochschulen auf die geplante NFDI vorbereiten sollte. Die „Landesinitiative NFDI“ verfolgte das Ziel, Stakeholder:innen in NRW zu diesem noch recht jungen Handlungsfeld zu informieren und die Beteiligungen von NRW-Hochschulen an Anträgen zu NFDI-Konsortien zu befördern.

Bereits zu diesem Zeitpunkt zeichnete sich eine Ausrichtung ab, die die Landesinitiative in NRW bis heute von Entwicklungen in anderen Bundesländern unterscheidet. Während die Landesinitiativen für FDM in vielen anderen Bundesländern in direktem Kontakt mit Forschenden stehen und diese schulen und beraten, sind die Angebote von fdm.nrw nicht primär an Forschende gerichtet. Sie richten sich vielmehr an hochschulpolitische Akteur:innen und an das FDM-Personal in den Hochschulen und zielen darauf ab, Synergien zwischen den Hochschulen zu schaffen und als Bindeglied zu wirken, sowohl zwischen den Hochschulen in NRW als auch zwischen Entwicklungen auf Landes- und Bundesebene.

Durch Vernetzung der Akteur:innen und Unterstützung zum Austausch von Expertise wurden die Hochschulen auf die NFDI vorbereitet und die Herausbildung von Konsortien unterstützt. Die zu diesem Zeitpunkt aufgebauten Strukturen existieren teils noch heute und können als Grundsteine der gegenwärtigen FDM-Community in NRW betrachtet werden. Dazu gehören die Einrichtung einer Mailingliste, die der Kommunikation dient, die Gründung der Arbeitsgruppe Awareness, die gemeinsame Informationsmaterialien für die Hochschulen erarbeitet hat, sowie die Ausrichtung des monatlichen Jour Fixe FDM, in dessen Rahmen über Entwicklungen aus allen Bereichen des FDM informiert wird.

Diese Rolle wurde ab 2019 in der zweiten Förderphase der „Landesinitiative fdm.nrw“ ausgebaut. Mit einem Lenkungskreis, der Vertretungen aus Hochschulleitung, zentralen

Einrichtungen, Forschung, DH.NRW und MKW zusammenbrachte, wurde eine Governancestruktur etabliert, die eine kontinuierliche Kommunikation mit den und in die landesweiten Hochschulgremien ermöglichte. Die bereits bestehenden Handlungsfelder der Vernetzung und Kommunikation, der Begleitung des Aufbaus der NFDI und der Initiierung und Begleitung thematischer Gruppen zur standortübergreifenden Zusammenarbeit wurden im Zuge der zweiten Projektphase um die Bereiche Weiterbildung und den Aufbau von FDM an HAW erweitert.

Mit den Weiterbildungsangeboten für Mitarbeitende aus den Infrastruktureinrichtungen wurde der Herausforderung der steigenden Nachfrage nach gut ausgebildetem Personal begegnet. Das Weiterbildungsangebot besteht aus themenspezifischen Workshops und dem ersten berufsbegleitenden Zertifikatskurs zum Forschungsdatenmanagement. In Zusammenarbeit mit der TH Köln, der ZBIW und der ZB MED wurde der Zertifikatskurs entwickelt und seit 2021 jährlich angeboten (Slowig u. a. 2022).

Zum Ende der zweiten Projektphase fand in der DH.NRW ein koordinierter Strategieprozess statt, der auch das Handlungsfeld FDM und damit die Landesinitiative fdm.nrw betraf.

3 Landeskonzept Forschungsdatenmanagement und Verstetigung der Landesinitiative fdm.nrw

Wie eingangs dargestellt bietet die DH.NRW Raum für eine Bandbreite von Digitalisierungsvorhaben, die von der Administration über die Lehre bis hin zu Forschung und Infrastrukturen reichen. Unter diesem Dach werden im Bereich Forschung neben FDM auch Themen wie High Performance Computing (HPC)[4] oder Cloud & Sourcing behandelt. Um diese Themen strategisch aufeinander abzustimmen und Synergien zu schaffen, wurde 2021 die Zielvision eines Digitalen Ökosystems formuliert. Für jedes dieser Themen sollte aus der DH.NRW heraus eine Landesstrategie erarbeitet werden. Ausgangspunkt für diese Strategieprozesse war eine Begehung der DH.NRW und ihrer Projekte durch NRW-externe Expert:innen, die den Status Quo in NRW begutachteten und basierend darauf Empfehlungen für die Weiterentwicklung aussprachen. Um das Landeskonzept FDM zu erarbeiten, wurde eine „Arbeitsgruppe FDM“ (AG FDM) innerhalb der DH.NRW eingerichtet. Die Zusammenstellung der AG FDM erfolgte durch die DH.NRW und war repräsentativ, so dass sowohl Leitungsebenen aller drei Hochschultypen als auch der Infrastruktureinrichtungen und fachliche Expert:innen beteiligt wurden. So wurde eine Rückkopplung in die Gremien der Hochschulen sichergestellt. Auch die Landesinitiative fdm.nrw war Teil der AG FDM. Die Erarbeitung des Landeskonzepts FDM erstreckte sich von 2022 bis zur Verabschiedung im Frühjahr 2024 und wurde durch eine externe Moderation unterstützt.

4 https://www.mkw.nrw/system/files/media/document/file/dh.nrw_landeskonzept_hpc_2024.pdf; *Besucht am 24. März 2025.*

Sowohl die Landesrektorenkonferenzen als auch die Kanzlerkonferenzen aller drei Hochschultypen sowie das MKW haben das Landeskonzept FDM (DH.NRW I AG FDM 2024) verabschiedet. Damit hat es für die DH.NRW einen verbindlichen Charakter.

Mit dem Ziel eine zukunftsorientierte FDM-Landschaft in NRW zu etablieren, die den Forschenden den Zugang zu zentralen FDM-Diensten ermöglicht, wurden fünf Handlungsfelder identifiziert, in denen die Hochschulen und das Land tätig werden. Diese Handlungsfelder werden im Landeskonzept FDM zunächst erläutert und im anschließenden Teil „Konzeptumsetzung“ spezifiziert. In Handlungsfeld I „Strategische Ausrichtung der Standorte“ bekennen sich die Hochschulen dazu, ihre Forschenden im FDM zu unterstützen und sich als Forschungsstandort zu FDM zu positionieren. Das Handlungsfeld II bezieht sich auf hochschulübergreifende Kooperation in Form von zentralen technischen und nicht-technischen Diensten. Im Handlungsfeld III wird beschrieben, wie die Anbindung an wissenschaftliche Bedarfe und die Einbindung an die NFDI in die nordrhein-westfälischen Strukturen gestaltet werden soll. Handlungsfeld IV widmet sich dem Bereich der Kompetenzentwicklung. Das Handlungsfeld V beleuchtet schließlich die Institutionalisierung des FDM in NRW. Es schließt mit der Bündelung der Prozesse bei der Landesinitiative fdm.nrw, die im Zuge des Landeskonzepts FDM verstetigt wurde, um ihrer Aufgabe der Begleitung der Umsetzung und der Fortentwicklung des Landeskonzepts nachkommen zu können.

In diesem Beitrag besonders hervorzuheben ist das Handlungsfeld II „Föderierte Informationsinfrastrukturen und Landesdienste“, das das Angebot eines zentralen Aufbaus und Betriebs grundlegender FDM-Dienste skizziert. Durch die Inanspruchnahme generischer Landesdienste werden an den Hochschulen Ressourcen für eine spezifische lokale Ausgestaltung der FDM-Dienste frei, die eine maßgeschneiderte Berücksichtigung der Bedarfe der Forschenden ermöglicht. Ausgehend vom Datenlebenszyklus werden gegenwärtig technische FDM-Landesdienste in unterschiedlichen Entwicklungsstadien projektiert bzw. angeboten. Unter Konsortialführerschaft der RWTH Aachen University in Zusammenarbeit mit der Universität Duisburg-Essen, der Universität Münster, der TH Köln und der Universität zu Köln wird ein landesweites GitLab-Angebot *(git.nrw)* aufgebaut, das Forschende bei der Versionierung von Forschungsdaten und kollaborativen Zusammenarbeit in Forschungsprojekten unterstützen wird. Der georedundante Forschungsdatenspeicher *DataStorage.nrw* (RWTH Aachen University, Universität Duisburg-Essen, Universität zu Köln und Universität Paderborn) steht den Forschenden für die Speicherung ihrer Forschungsdaten für zehn Jahre nach guter wissenschaftlicher Praxis über die Forschungsdatenmanagement-Plattform Coscine zur Verfügung. *Coscine.nrw* (RWTH Aachen University) unterstützt Forschende dabei, die FAIR-Prinzipien umzusetzen, bspw. ist die Vergabe von Metadaten notwendig. Sowohl DataStorage.nrw als auch Coscine.nrw wurden bereits deutlich vor der Erarbeitung des Landeskonzepts FDM initiiert. Sie erfüllen langfristige Aufgaben und sind als verstetigte Dienste für Forschende vertrauenswürdig und verlässlich. Unter Konsortialführerschaft der Universität Münster wird momentan von der Ruhr-Universität Bochum, der Hochschule Bonn-Rhein-Sieg und der TU Dortmund im Projekt *JupyterHub.nrw* ein landesweites JupyterHub-Angebot aufgebaut, das Forschen-

den Rechenkapazitäten niedrigschwellig zur Verfügung stellt. Der Sync-and-Share-Service *sciebo.nrw* (Universität Münster) dient der Kollaboration der Forschenden in Forschungsprojekten innerhalb NRWs und durch Gastzugänge auch darüber hinaus. Auch dieser Dienst existiert bereits seit 2015 und ist ebenfalls verstetigt. Die Bereiche der Datenpublikation und der Langzeitverfügbarkeit (über zehn Jahre hinaus) sind Themen für die gegenwärtig Konzepte innerhalb NRWs erarbeitet werden, die aber auch als generische FDM-Landesdienste verortet sind.

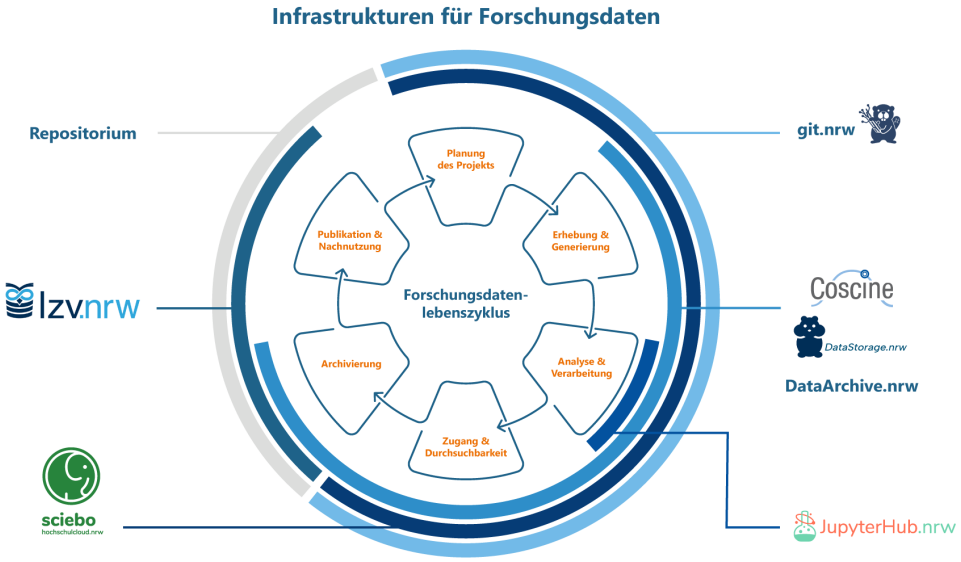


Abbildung 1: Die gegenwärtig schon betriebenen sowie zukünftig geplanten FDM-Landesdienste sind im Kontext des Forschungsdatenlebenszyklus eingeordnet. Die Dienste git.nrw und Jupyter.Hub.nrw befinden sich in Pilotphasen. Die Dienste Coscine und DataStorage.nrw sowie sciebo werden den Hochschulen bereits angeboten. Eine Lösung für die Langzeitverfügbarkeit wird aktuell in Zusammenarbeit mit der Landesinitiative LZV.nrw erarbeitet. Ein landesweiter Repositoriendienst wird zum jetzigen Zeitpunkt (Mai 2025) konzeptioniert.

Doch werden nicht nur technische Landesdienste aufgebaut und angeboten, sondern auch nicht-technische Landesdienste sollen in hochschulübergreifender Kooperation Forschende unterstützen. Im Schulungs-, Beratungs- und Kuratierungsnetzwerk werden ausgehend von den Hochschulbibliotheken in NRW Potentiale für die hochschulübergreifende Zusammenarbeit realisiert. Ziel der Arbeit in diesem Netzwerk ist die kooperative Zusammenarbeit, um Forschende flächendeckend und ressourcenschonend im Umgang mit Forschungsdaten zu unterstützen. Als eine Weiterentwicklung der Förderlinie FDMScouts.nrw wurde im Landeskonzept zudem ein Netzwerk der HAW beschrieben. In diesem sollen die Synergien in hochschulübergreifenden Kooperationen gehoben werden, um den strukturellen Besonderheiten der HAW zu begegnen. Für alle Arbeitsbereiche des Netzwerks ist die Verbindung zu Entwicklungen in der NFDI ein wichtiges Ziel und die Strukturen der Netzwerke bieten hier einen direkten Anknüpfungspunkt. Zukünftig soll das Zusammenspiel der technischen und der nicht-technischen Landesdienste fokussiert werden, um die

Forschenden in NRW flächendeckend in ihrem Umgang mit Forschungsdaten und der Umsetzung der FAIR-Prinzipien zu unterstützen.

Der Landesinitiative fdm.nrw kommt bei diesen zahlreichen Entwicklungen die Aufgabe zu, einen Informationsfluss zwischen den beteiligten Stakeholdergruppen zu unterstützen. In einem eigenen Netzwerk tauschen sich die Vertreter:innen der FDM-Landesdienste mehrmals jährlich zu aktuellen Entwicklungen und Anknüpfungspunkten untereinander aus. Zweimal im Jahr informieren die FDM-Landesdienste Interessierte der Hochschulen der DH.NRW in Veranstaltungen über ihren Status Quo sowie aktuelle Entwicklungen und Pläne und stehen für Rückfragen und weiteren Austausch zur Verfügung. Die Governancestruktur der DH.NRW bietet dabei einen Rahmen, um die Entwicklungen FDM-spezifischer Dienste mit anderen Themenfeldern der DH.NRW zu verbinden und so den Diskurs zu FDM gemeinsam mit anderen Digitalisierungsaktivitäten zu führen.

Allein dieser Ausschnitt aus dem Landeskonzept macht deutlich, welche umfangreichen Entwicklungen zum FDM gegenwärtig in NRW stattfinden. Ausgehend von der zuvor bereits in der Projektzeit etablierten Rolle als Ankerpunkt für die Vernetzung und Abstimmung der Stakeholdergruppen in NRW und der Funktion als Informationsdrehkreuz, wurde der Landesinitiative fdm.nrw in Handlungsfeld V die Aufgabe übertragen, die Umsetzung des Landeskonzepts zu begleiten. So werden im Kontext der Umsetzung durch fdm.nrw das Ineinandergreifen der Handlungsfelder und ihrer einzelnen Aktivitäten sichergestellt und die Bedarfe im Land und daraus resultierende Entwicklungen zusammengeführt.

Die AG FDM wurde auch über den Prozess hinaus damit beauftragt, die Umsetzung des Landeskonzepts FDM zu begleiten. Damit ist zum einen sichergestellt, dass die fachliche Expertise erhalten bleibt und weiterhin in die strategische Ausrichtung des Landes fließt, zum anderen ist die Rückkopplung mit den Hochschulgremien ebenfalls fortwährend gesichert. Um den Informationsfluss stärker zu untermauern, wurde 2024 das Kommunikationsnetzwerk fdm.nrw gegründet. Für diese Struktur wurden die Hochschulleitungen aller DH.NRW-Mitgliedshochschulen angefragt, eine Ansprechperson zu benennen, die als Bindeglied zwischen den Hochschulen und der Landesinitiative fdm.nrw agiert. Damit wurde ein zuverlässiger Kanal für die Weitergabe von Informationen zwischen den Hochschulen und der Landesinitiative fdm.nrw aufgebaut. Zum einen erhalten sie über eine Mailingliste Informationen und sind dann aufgefordert, diese innerhalb ihrer Hochschule an relevante Personen weiterzuleiten und ggf. zielgruppengerecht zu kontextualisieren. Die Mitglieder des Kommunikationsnetzwerks sind gleichzeitig aber auch in einer Position, um gebündelte Rückmeldung aus den Standorten an die Landesinitiative fdm.nrw zu geben. Als ein neues Format des Austauschs wurde das sog. Feedback-Forum eingerichtet, zu dem die Landesinitiative fdm.nrw bedarfsbezogen einlädt, um Konzeptideen in ihrer Entstehung mit der FDM-Community in NRW zu diskutieren. Ziel dieser Aktivitäten ist es, die Einbindung aller Hochschulen und ihrer Stakeholder:innen zu intensivieren und so gemeinsam die Umsetzung und die Weiterentwicklung des Landeskonzepts FDM zu verfolgen.

Die Verabschiedung des Landeskonzepts FDM liegt zum gegenwärtigen Zeitpunkt (Mai 2025) ein gutes Jahr zurück. Mit diesem Abstand und einem Blick auf die Umsetzung des Landeskonzepts FDM wollen wir abschließend einige Erkenntnisse aus der Durchführung des Strategieprozesses teilen.

4 Erkenntnisse aus diesem Prozess

Die eingangs beschriebenen Herausforderungen im FDM – der Aufbau dauerhafter Strukturen unter sich stetig wandelnden Bedingungen und die Gewährleistung einer flächendeckenden Beteiligung an den Entwicklungen – sind keineswegs spezifisch für das Land NRW. Auch wenn der Umgang mit ihnen in NRW maßgeblich von den dargestellten spezifischen Rahmenbedingungen beeinflusst ist, lassen sich Erkenntnisse daraus auf andere Kontexte transferieren.

Die Ausarbeitung des Landeskonzepts FDM, der Aufbau landesweiter FDM-Dienste und die dauerhafte Einrichtung der Landesinitiative fdm.nrw lassen sich als ein Erfolg für das FDM betrachten. Entscheidend war die Einbettung in politische Prozesse und Strukturen, um die Entwicklungen vorantreiben zu können. In NRW wird durch die DH.NRW eine enge Anbindung an die Hochschulleitungen und die Fördergeber gewährleistet. Sie stellte einen wesentlichen Grundstein für die bereits frühe Behandlung der Thematik und den Entschluss dar, Strukturen für das FDM in NRW nachhaltig zu gestalten, also zu verstetigen.

Doch die vorhandene Dachstruktur der DH.NRW ist nur ein Aspekt des Strategieprozesses und des Übergangs in die Umsetzung der angedachten Konzepte. Zu Beginn des Strategieprozesses blickte die Landesinitiative fdm.nrw bereits auf eine mehrjährige Projektförderung zurück, die es ihr erlaubt hat, die Entwicklungen auf diesen Prozess hin zu begleiten und zu unterstützen. Ihre Rolle in der Umsetzung des Landeskonzepts FDM ist damit nicht vollends neu, sondern setzt die bisherigen Entwicklungen fort, was sich gut daran ablesen lässt, dass die Handlungsfelder ihrer Projektzeit in weiten Teilen in der Aufgabe, die Umsetzung des Landeskonzepts FDM zu begleiten, aufgehen.

Dass hochschulübergreifende Kooperationen ressourcenschonend und damit zunehmend relevant sind, liegt auf der Hand. Wichtig ist jedoch die Gestaltung des Prozesses, um sich auf gemeinsame Ziele zu verständigen und den Weg zur Erreichung dieser Ziele zu erschließen. Die Erarbeitung einer gemeinsamen Landesstrategie, die durch die Gemeinschaft der Hochschulen und des MKW getragen wurde, ist ein wesentlicher Meilenstein, um die Synergien aus hochschulübergreifender Zusammenarbeit zu heben. Zentral ist dabei auch, es nicht bei einem Konzept zu belassen, sondern möglichst früh bereits die Umsetzung zu initiieren. Wie oben ausgeführt beinhaltet das Landeskonzept FDM den Abschnitt „Konzeptumsetzung“. Darin ausgeführt ist die konkrete Umsetzung der zuvor beschriebenen FDM-Landesdienste sowie weiterer hochschulübergreifender Dienste und

die Verstetigung der Landesinitiative fdm.nrw. Der Abschnitt umfasst damit die Benennung von Verantwortlichkeiten, die einen zeitnahen und realistischen Plan zur Umsetzung des Landeskonzepts beschreiben. Auch wurden einzelne Aspekte, neben der Landesinitiative auch der Aufbau von Speicherinfrastrukturen und Vernetzung des FDM-Personals, vor der Veröffentlichung des Landeskonzepts initiiert, was den Übergang von Konzept in Umsetzung erleichtert. Zusätzlich sei an dieser Stelle betont, dass der Erfolg der Umsetzung des Landeskonzepts FDM wesentlich vom Engagement der Hochschulen abhängt.

Abschließend verdeutlicht die Betrachtung des beschriebenen Themenkomplexes der Maßnahmen im FDM in NRW dazu, dass zahlreiche Stakeholder:innen in verschiedene Entwicklungen zeitgleich involviert sind. Die Einrichtung einer zentralen Stelle erlaubt es, diese Prozesse zusammenlaufen zu lassen, miteinander zu verknüpfen und strategisch auf kurzfristige Entwicklungen reagieren zu können. Gut in der Hochschullandschaft verankerte Landesinitiativen für FDM stellen eine solche zentrale Informations- und Koordinationsplattform dar, die langfristig die strategische Entwicklung dieses dynamischen Handlungsfeldes entscheidend mitgestalten.

Die langfristige Verankerung des FDM in NRW durch das Landeskonzept FDM und die Landesinitiative fdm.nrw kann als ein gelungenes Beispiel dafür dienen, wie mit den vielen Herausforderungen im FDM umgegangen werden kann. Der Aufbau einer grundlegenden Infrastruktur ermöglicht es, sich auf zukünftige innovative Lösungskonzepte im FDM zu fokussieren und so die Verfügbarkeit und Nachnutzung von Forschungsdaten weiter auszubauen. Die Zusammenarbeit der Hochschulen wird durch die Abstimmung gemeinsamer Ziele und die Umsetzung dieser gestärkt, denn die gemeinsame Vision, allen Forschenden Zugang zu zentralen FDM-Diensten und die bestmögliche Unterstützung beim FDM zu bieten, verbindet.

Danksagungen

Wir bedanken uns bei unseren Fördergebern, dem Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen und der Digitalen Hochschule NRW.

Die Landesinitiative fdm.nrw ist angesiedelt an der Universitätsbibliothek der Universität Duisburg-Essen. Wir möchten für die fortwährende Unterstützung, speziell auch im Kontext der Verstetigung, herzlich danken. Besonders hervorheben möchten wir Dr. Stephanie Rehwald, Leiterin der Landesinitiative fdm.nrw von 2020 bis 2023, die durch ihr Engagement das FDM in NRW maßgeblich mitgestaltet.

Der AG FDM möchten wir für ihren fortlaufenden Einsatz für die Entwicklung des FDM in NRW danken. Die Mitwirkung jedes einzelnen Mitglieds in der AG FDM hat dazu beigetragen, dass das Landeskonzept FDM die Breite der NRW-Hochschullandschaft abbildet.

Nicht zuletzt möchten wir unserer FDM-Community für die aktive, unterstützende und positive Zusammenarbeit danken.

Beiträge

Erster Entwurf: MC
Review & Überarbeitung: MF, CS, LS

Interessenkonflikte

Kein Interessenkonflikt bekannt.

Literaturverzeichnis

DH.NRW I AG FDM. 2024. *Digitales Ökosystem DH.NRW – Landeskonzept FDM.* Zenodo. https://doi.org/10.5281/ZENODO.11242683.

Slowig, Benjamin, Mirjam Blümm, Konrad U. Förstner, Marvin Lanczek, Birte Lindstädt, Rabea Müller, Ulrike Nickenig, Stephanie Rehwald und Lioba Schreyer. 2022. „Zertifikatskurs Forschungsdatenmanagement in NRW“. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB* 9 (3/2022). https://doi.org/10.5282/o-bib/5833.

The Triathlon Relay of Storage Space Provisioning

Ilona Lang [1*], Marcel Nellesen [1], Marius Politze [1]

[1] IT Center, RWTH Aachen Universiy;
[*] Corresponding author: lang@itc.rwth-aachen.de

The increasing demand for storage space in data-intensive sciences necessitates the establishment of cooperative information infrastructures among universities. This paper introduces the provisioning processes around DataStorage.nrw, a redundantly distributed cloud storage solution across four universities in North Rhine-Westphalia (NRW). Access to DataStorage.nrw is free for eligible researchers who meet specific criteria: they must be employed at a DH.NRW university, their storage needs must arise from a research project, and the data must adhere to FAIR principles. The allocation process resembles a triathlon relay, integrating various software disciplines to facilitate efficient resource provisioning. The workflow comprises three stages: Identity and Access Management (IAM) ensures secure user authentication; the RDM platform Coscine manages research metadata and project structures while streamlining compliance with FAIR principles; JARDS serves as the science-led review mechanism for storage applications, ensuring that resources align with funding objectives. By fostering cross-service interoperability and embracing agile processes, we aim to avoid isolated solutions and enhance collaboration among universities. Ultimately, our approach not only addresses immediate data management challenges but also provides a scalable framework that can serve as a blueprint for other universities seeking to establish similar cooperative infrastructures.

Keywords: Storage, IAM, Coscine, JARDS, FAIR Principles

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23918 (CC BY-SA 4.0).

1 Introduction

Data-intensive sciences require more and more storage space[1], which universities must provide. At the same time, the buffer calculations of individual universities lead to a sum of unused capacities in the many separate storage infrastructures. Part of the solution and national political demand is therefore to establish cooperative information infrastructures (Deutsche Forschungsgemeinschaft. Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme 2025) between universities. One of these information infrastructures is a redundantly distributed cloud storage for cooperative use – the DataStorage.nrw[2]. DataStorage.nrw comprises around 42 PB of usable storage space and is distributed across four locations at the Universities in Aachen, Cologne, Duisburg-Essen and Paderborn (see Figure 1). Storage space on the DataStorage is expected to be available from June 2025. The predecessor of DataStorage.nrw was the Research Data Storage (RDS), which was also redundantly distributed – but had a smaller storage capacity and was available to a smaller number of universities in NRW.

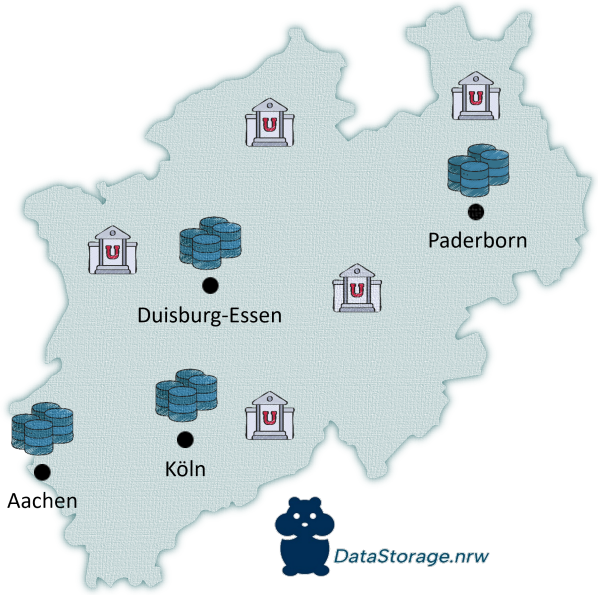


Figure 1: DataStorage.nrw – Map of the Locations.

Storage space on DataStorage.nrw is free of charge for researchers and a suitable storage location both for warm data during the active project phase and for cold data up to a maximum of 10 years after the end of the project. However, in addition to a cooperative use, the following criteria must be met: The researcher must be employed at a DH.NRW

1 https://www.kim.uni-konstanz.de/kim-news/news-detailseite/forschungsdateninventur-zeigt-grossen-speicherbedarf/; *Visited on March 31, 2025.*

2 https://www.land.nrw/pressemitteilung/riesen-wolke-macht-forschungsdaten-sicher; *Visited on March 31, 2025.*

university, their storage space requirement must have arisen within a research project, the storage be in compliance with the FAIR principles (Wilkinson et al. 2016). These three conditions must be checked as part of a scientific allocation procedure and require different tools to ensure that they are met. The review process is similar to a relay race in which different software disciplines come together to achieve the goal of allocating storage space – a triathlon relay of storage space provisioning. Table 1 gives an overview on all stages of the triathlon relay which will be described in the following chapters in more detail.

Table 1: Overview: Triathlon Relay of Storage Space Provisioning.

Stage	**Result**
1: Identity and Access Management	Cooperative use by all universities
2: Coscine – RDM Platform	Project management in a FAIR environment
3: JARDS – Software for Resource Allocation	Science-led review process
Finish line: DataStorage.nrw	Storage space allocation

2 First Stage: Identity and Access Management (IAM)

The first stage of the relay is the Identity and Access Management (IAM), it is important to know who the users are and what they can access. This is especially challenging with services that are available to a large variety of different institutions, and which supports cooperative use between universities. Additionally, an individual researcher might also change his university, go abroad or leave the university and go to the industry. In our platform we address this offering different login methods to the researchers. Firstly, they can use the DFN-AAI to login with their local identity provider, secondly social logins, such as ORCID are available for researchers that left the university, industry partners or international cooperation partners.

The Base4NFDI project IAM4NFDI[3] enables an easy connection of a service to one or more of the provided Community Authentication and Authorization Infrastructures (CAAI), e.g. the RegApp. Since these architectures are often complex, IAM4NFDI has regular calls for incubator projects. Within these technical questions and the connection of a service to the CAAIs are realized within a regular direct exchange between the service and the CAAI providers.

3 https://base4nfdi.de/projects/iam4nfdi; *Visited on March 31, 2025.*

One big benefit of the CAAIs is an easy connection of one service to many home organizations, without the need of a 1 to 1 communication between the different parties beforehand. The CAAI acts as an identity proxy in between the identity providers and the services. Therefore, identity provides only need to configure one identity proxy instead of multiple services and the services only need to configure one identity proxy instead of many identity providers. The utilized protocols SAML and OIDC are established and widely supported. This makes it also easier to transfer required information, which could include among others: Membership of a university, membership in a virtual organization, unified set of attributes and roles through IDM.nrw

This approach is that it is extremely scalable and represents the current state of art of identity and access management. As such it is also adopted by several federal and national projects such as the NFDI and NHR.

3 Second Stage: RDM Platform Coscine

The first relay handover is carried out from NFDI-Login to the RDM platform Coscine (re3data.org 2024). Coscine has been developed as open source software at RWTH Aachen University since 2018[4], has been in regular operation since 2023 and is used for the management of research (meta)data as well as for the allocation and provisioning of storage resources for research data. Coscine is developed according to the FAIR principles (Lang, Jansen, and Politze 2025) and implements interfaces for the so-called FAIR Digital Objects (Smedt, Koureas, and Wittenburg 2020). As part of Coscine.nrw, Coscine was established as a permanent service and thus made available to all DH.NRW universities. The relay transfer via the NFDI-Login provides Coscine with information on whether a person is an employee at a DH.NRW university. In addition, the login enables the registration of all universities within the DFN-AAI and, thanks to the ORCID option, also international researchers. Coscine itself has two main functions in the relay: To structure data storage in projects and ensure the implementation of a large part of the FAIR principles (Lang, Jansen, and Politze 2025).

Figure 2 shows the main workflow for researchers using Coscine. As described above, Coscine offers researchers a low-threshold login. After login, researchers create their projects using the web interface or via the Coscine API. Created projects and sub-projects are mapped in Coscine as a tree structure, enabling simple and hierarchical visualization. Project creators can invite all other project partners easily via e-mail and regulate their corresponding access via the project-related rights and roles concept. Coscine offers three different roles inside of projects: owners, members and guests. Based on the low-threshold login, external project partners can be invited to Coscine project, fostering collaborative co-operation across institutes and universities.

4 https://coscine.rwth-aachen.de/; *Visited on March 31, 2025.*

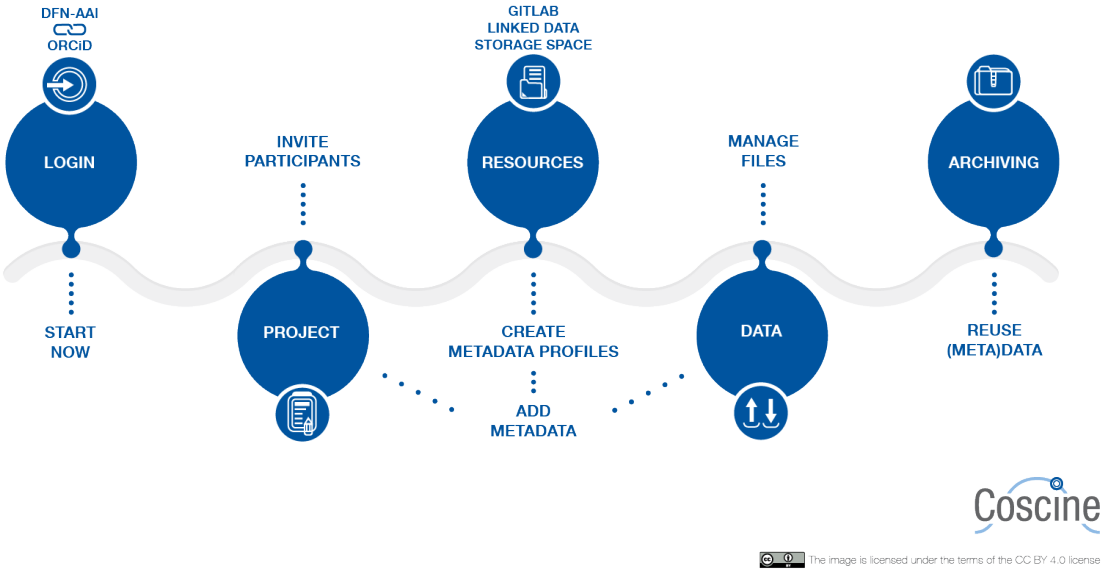


Figure 2: Coscine Workflow: Five Main Steps from Login to Archiving.

Research data is uploaded in Coscine at the resource level, where different resource types are available for each project depending on access rights and the selected login method of the respective project owner. For the DataStorage.nrw, the available storage space resource types are web resources (accessible only via the browser/API), S3 resources[5] (accessible via browser/API/S3) and WORM (write once, read many) resources (accessible via browser/API/S3). Apart from the DataStorage.nrw Coscine allows researchers to add other data sources of a project, for example via the GitLab resource type.

Coscine supports researchers by providing a rich metadata management environment that gives access to discipline specific metadata profiles. The metadata is automatically checked for completeness based on the refinements of each metadata profile. The metadata profiles are machine accessible based on their technical representation and validation via W3C standards RDF and SHACL as proposed by the AIMS Project (Grönewald et al. 2022). Moreover, Coscine implements the FAIR Digital Object Concept to represent stored data in concise digital entities (Politze et al. 2025; Heinrichs, Hunke, and Politze 2025). Coscine assigns for each project and resource (including data and metadata) a handle-based ePIC-PID (Kálmán, Kurzawe, and Schwardmann 2012). Individual files are described with metadata profiles, of which each RDF-triple includes PIDs leads to the described data by extending the handle URL. In summary, Coscine provides access to DataStorage.nrw while setting a project structure and enabling collaborative work in a FAIR environment.

As an intermediate stage, Coscine automatically allocates employees of DH.NRW universities 100 GB on DataStorage.nrw as web resource – a resource type with upload function

5 S3 (Simple Storage Service) is a simple object storage interface for data that supports the rapid retrieval of object data.

via the web interface and forced use of Coscine's metadata profiles. However, if researchers need more storage space and/or access via S3 resource, the relay needs to be passed on to JARDS[6], since Coscine can not guarantee the scientific character of a project and the linking to rich metadata in accordance with the FAIR principles (Lang, Jansen, and Politze 2025).

4 Third Stage: Science-led allocation procedure with JARDS

The third stage of the relay is the science-led review of the storage space applications. This stage is represented by the platform JARDS (Janetzko 2019). This abbreviation stands for Joint Application Review and Dispatch Service and was originally developed by the Jülich Super Computing Center for applications for computation time on the super computer. The RWTH Aachen University started an own instance in 2014 for its own high performance computer. Since then the platform was adapted by various universities and projects, e.g. the National High-performance Computing (NHR) and WestAI.

In 2022 this platform was adopted to support applications for storage space in Coscine. Multiple application types are supported based on the resource types that is requested. While less configurable DataStorage.nrw resources such as the *web* resource only require a minimal application more specialized resources such as *S3* and *Write Once Read Many (WORM)*, require a detailed description of the metadata management, the stored data and the data flows.

Figure 3 shows the main workflow from the DataStorage.nrw application to review using JARDS. After the researcher has submitted the application 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 (Figure 3). 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 (see *Criteria for Technical Evaluation* below). 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

6 https://jards.coscine.de/; *Visited on March 31, 2025.*

plan review. Usually, this step takes about one week. After a successful first technical review by the home institution, a second technical review is carried out by an independent university from DH.NRW, to ensure a fair review process. This step takes about another week.

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 DataStorage.nrw applications for *web* resources and up to three months for applications for *S3* and *WORM* resources.

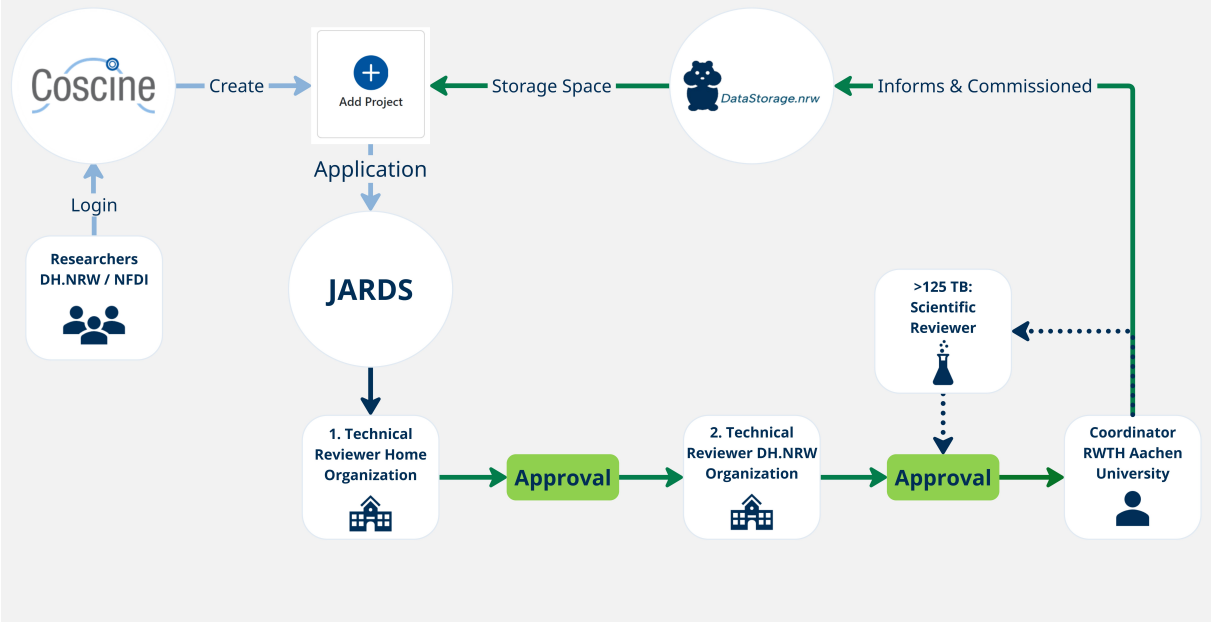


Figure 3: JARDS Workflow: The Main Steps Application, Review Process and Storage Allocation.

The technical evaluation of storage applications has the main objective to sensitize researchers for the FAIR principles and research data management in general. There are three main criteria that are part of the evaluation:

- **Project Context:** Is the application based on a research project?
- **Amount of Requested Storage:** Is the requested storage space suitable for the described project?
- **Metadata:** Is the data annotated with detailed metadata? Is metadata added in a consistent and reliable way? Is the structure and the findability of the metadata ensured?

The first question is aimed at ensuring the project structure. For example, it is not permitted to store photos from the last company trip on DataStorage.nrw (background: no research project and no research data). Likewise, it is not permitted to store research data without a specific project reference, as otherwise no data life cycle can be ensured (for example, there is no clear project end date). The second criterion is to ensure that the scope of the requested storage roughly matches the described project – a 125 TB application from a project with pure literature research, for example, would be questioned. The third criterion is aimed at the area of FAIR principles, which cannot be automatically ensured by Coscine if using S3 resources. This concerns the researchers' handling of metadata. In most cases, this is also the question area that generates the greatest consulting effort. The reviewers often provide advice on subject-specific metadata profiles – combined with the recommendation to use the metadata management tools provided by Coscine.

5 Finish Line and Outlook

After successfully completing all three relay stages, the researchers reach the finish line and receive their storage space on DataStorage.nrw. They can now manage this storage space independently in their Coscine projects.

With the new storage provision process presented here, we are moving from unstructured data volumes to clearly defined research project units. This transition and the focus on the FAIR principles is necessary to enable a regulated data life cycle and at the same time provide a large number of different universities with storage space. The first stage of IAM is crucial for collaboration – if not every potential project partner can log in, just a new isolated solution for data storage is created. The merging of different accounts via different login options has so far generated a lot of effort in the development of Coscine – thanks to the help of IAM4NFDI, we could remedy this situation.

The second stage in Coscine is crucial for the clearly defined research project units in a FAIR environment. Our aim is always to make as little work as possible for the researchers by automatically fulfilling most of the requirements of the FAIR principles through Coscine.

The third stage by JARDS covers a key feature of the DataStorage.nrw operating concept which is the demand for a science-led management. JARDS helps to ensure that DataStorage.nrw resources are used in line with the funding objectives and offers different features for a collaborative review process. Here, RDM personnel from NRW has to take on an essential role as "Gate Keepers" for storage space. A critical aspect of this gate keeping is the joint revision of technical evaluation criteria for storage applications to ensure that our review processes remain relevant and effective. Along with an update to the content, we are planning closer technical integration between Coscine and JARDS (e.g. automated reviewer and storage assignment). This should increase the technical

efficiency and sustainability of the processes. Another challenge is, that we only perform an initial curation during deposition, with no regular review of compliance with the criteria. In the future, JARDS could also play an essential role in a regular curation via its functions for projects.

Overall, we try to avoid isolated solutions. It is crucial that we consider cross-service interoperability in our processes. This approach provides us with greater flexibility concerning the underlying hardware while ensuring seamless integration across RDM services. The developments of the open source platforms Coscine and JARDS relay heavily on agile processes that facilitate regular adjustments based on RDM personal and researcher feedback and their changing needs. This requires significant training and communication efforts within the service management and development teams. One part of our solution is to focus on scalability and transparency across all participating processes in the 'Triathlon Relay of Storage Space Provisioning' to ensure smooth operation and adaptation over time.

Taken together, we aim for the regular further development of all components by engaging all parties involved. Our strategy involves linking services from NFDI, HPC, and storage infrastructure into a cohesive process. The advantages include scalability, cost reduction, and increased acceptance among researchers. However, we must also be mindful of potential disadvantages such as dependence on multiple services and the associated synchronization efforts. Another pressing issue of DataStorage.nrw is balancing storage duration with the lifespan of hardware. It is essential to recognize that cooperative infrastructures do not automatically equate to sustainability; ongoing financing is crucial for maintaining hardware. In this respect, we remain dependent on a binding political commitment to ensure the storage of data for at least ten years.

In conclusion, by embracing these challenges and fostering collaboration among all stakeholders, we are not only enhancing our data storage solutions but also paving the way for a sustainable and efficient research environment that aligns with the evolving needs of academia. Our commitment to the FAIR principles and cross-service interoperability help researchers to focus on their projects while benefiting from a reliable and well-structured data management system. The processes described here can serve as a template for the implementation of other cooperative information infrastructures between universities.

Acknowledgements

DataStorage.nrw is funded by Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (MKW: 214-76.01.09-7-7937 DFG: INST 222/1530-1). RDS and RDS.NRW are funded by Ministerium für Kultur und Wissenschaft des Landes Nordrhein Westfalen (MKW:124-4.06.05.08-139057, DFG: INST222/1261-1). The IAM4NFDI project is partially funded by DFG as part of NFDI (grant number 521466146). The work

on Coscine was partially supported with resources granted by DH.NRW for the project Coscine.nrw by Ministerium für Kultur und Wissenschaft des Landes Nordrhein Westfalen, by NFDI4Ing, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 442146713 and NFDI-MatWerk, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 460247524.

Authorship Contributions

Please specify the contributions of each author to the manuscript.

- Ilona Lang: Conceptualization, Methodology, Writing
- Marcel Nellesen: Conceptualization, Methodology, Writing
- Marius Politze: Conceptualization

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

Deutsche Forschungsgemeinschaft. Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme. 2025. *Digitale Forschungspraxis und kooperative Informationsinfrastrukturen. Ein Diskussionspapier der Deutschen Forschungsgemeinschaft zu Förderung und Finanzierung wissenschaftlicher Informationsinfrastrukturen.* Zenodo. https://doi.org/10.5281/zenodo.14621979.

Grönewald, Matthias, Patrick Mund, Matthias Bodenbrenner, Marc Fuhrmans, Benedikt Heinrichs, Matthias S. Müller, Peter F. Pelz, et al. 2022. "Mit AIMS zu einem Metadatenmanagement 4.0: FAIRe Forschungsdaten benötigen interoperable Metadaten". In *E-Science-Tage 2021: Share Your Research Data: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 91–104. Heidelberg: heiBOOKS. ISBN: 978-3-948083-54-0. https://doi.org/10.11588/heibooks.979.c13721.

Heinrichs, Benedikt, Sirieam Hunke, and Marius Politze. 2025. "Combining FAIR Digital Object Implementation Concepts in a Real-Life Application". *Open Conference Proceedings* 5. https://doi.org/10.52825/ocp.v5i.1173.

Janetzko, Florian. 2019. “JARDS Ein Softwarewerkzeug zur Handhabung von Ressourcenvergabeprozessen”. In *ZKI-AK Supercomputing Herbsttagung,* Berlin, Germany. 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.

Lang, Ilona, Katja Jansen, and Marius Politze. 2025. *Implementing the FAIR Principles with Coscine.* Zenodo. https://doi.org/10.5281/zenodo.14886335.

Politze, Marius, Benedikt Heinrichs, Sirieam Hunke, Ilona Lang, and Thomas Eifert. 2025. “FAIR Digital Objects: FAIRtilizer for the Digital Harvest”. In *EPiC Series in Computing,* 105:284–274. EasyChair. https://doi.org/10.29007/hfzk.

re3data.org. 2024. *Coscine.* re3data.org – Registry of Research Data Repositories. https://doi.org/10.17616/R31NJNJZ.

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. 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. https://doi.org/10.1038/sdata.2016.18.

Staying Flexible and Transparent in a Changing World: Agile RDM Service Development

Petar Hristov [1], Katja Jansen [1], Ilona Lang [1]

[1] IT Center, RWTH Aachen University

No matter which area you consider, the world is constantly changing, so requirements must be continuously adapted. This is especially true for Research Data Management (RDM), and particularly for RDM platforms like Coscine. The increasing number of users and external stakeholders, as well as the balancing act between maintenance of existing and the development of new features, represent just some of the challenges faced. In this paper, we illustrate how these challenges can be managed in practice to keep pace with current RDM trends and not miss important developments such as FAIR Digital Objects (Politze et al. 2025) or advancements in querying technologies like SPARQL. We also examine how our anchors – FAIR principles, organization and operational processes – help maintain focus, and what adapting methods to real-world circumstances can look like.

Keywords: Research Data Management, Coscine, FAIR Principles, Software Development, Scrum

1 Introduction

Coscine[1] is a generic platform for research data management (RDM) that has been under development at the IT Center of the RWTH Aachen University since 2018. Within just a few years, the platform has evolved from a start-up service for the RWTH to fulfilling

1 re3data.org: Coscine; editing status 2024-03-15; re3data.org – Registry of Research Data Repositories. http://doi.org/10.17616/R31NJNJZ; *Visited on March 28, 2025.*

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23919 (CC BY-SA 4.0).

a mandate for nationwide service, thus highlighting the importance of adaptability to changing challenges. In terms of Coscine, it is necessary to consider not only challenges common to other software solutions, such as balancing maintenance and new features, but also emerging RDM trends.

The mission of Coscine is to provide a generic platform for German researchers and their collaboration partners to access storage resources, jointly work on datasets online in research projects, and retain datasets for up to ten years after the end of the project. We promote high quality research by fostering the FAIR-Principles (Wilkinson et al. 2016) for discovering and accessing data during the active project phase through the implementation of the FAIR Digital Object concept (for more details see our mission statement[2]). The maintenance of core functionalities includes, for example, user login, which supports authentication via DFN-AAI and via ORCID, ensuring easy and barrier-free access. After logging in, projects and sub-projects can be created in a tree structure in Coscine. Access management is ensured via a regulated role structure. Resources within projects can either provide access to storage space from the Research Data Storage (RDS-Web, -S3 or -WORM) or be linked to external storage systems (e.g. GitLab, LinkedData). In future, the RDS storage will be replaced by DataStorage.nrw.

Metadata management plays a central role in Coscine; therefore, entering metadata is mandatory when creating projects and resources. For the use of RDS-Web resources via the UI or the API, it is also essential to enter metadata before successfully uploading a file. To ensure that metadata is entered when using S3 resources, the storage space allocation for this resource type is carried out exclusively via an application procedure. Metadata is entered at the data level using metadata profiles. Metadata profiles can be selected from a list or created individually through the AIMS Metadata Profile Service (Grönewald et al. 2022). All created metadata profiles can be reused by researchers within Coscine. In addition, Coscine assigns Persistent Identifiers (PIDs) based on the ePIC handle system to ensure unique and permanent identification at both project and resource level.

Below we present the challenges that we have encountered during the development of Coscine in recent years, starting from 2018 till today with almost 3,300 registered users. In addition, we provide an insight into how we have effectively overcome these challenges and what our anchors are for flexibility in an ever-changing world.

2 Our Anchors in a Changing World

Anchors are crucial for addressing the challenges of a constantly evolving world. Therefore, we have identified three main anchors that enable us to remain true to our Coscine mission and effectively respond to the changing challenges:

2 https://about.coscine.de/en/about/mission-statement/; *Visited on March 28, 2025.*

- We aim to provide a FAIR environment for research data.
- We operate at the IT Center of RWTH Aachen University.
- We apply agile methods for maintenance and ongoing development.

2.1 Our Stakeholders – A Changing World

If we look at the various stages of Coscine's development, we recognize three characteristic phases: *Development Phase*, *Pilot Phase* and *Regular Operation* (see Figure 1). At the beginning of the Coscine development phase in 2018, there was only one stakeholder determining the requirements for the RDM platform – the IT Center of RWTH Aachen University. With the transition to the pilot phase in 2020, the number of stakeholders increased significantly, leading to close collaborations, for example, with NFDI-MatWerk or NFDI4ING. Since entering regular operation in April 2023, the number of stakeholders has continued to grow steadily, along with increasing requirements.

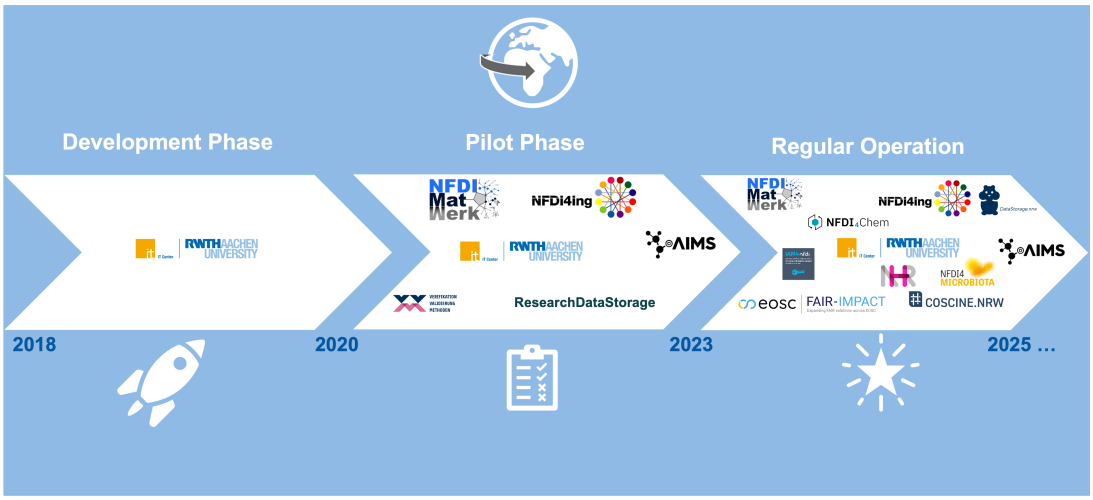


Figure 1: Coscine Stakeholder development over the years.

2.2 Our Development Phases in a Changing World

For an overview, we describe the three phases in more detail: the *Development Phase*, *Pilot Phase* and *Regular Operation*.

Development Phase (2018–2020)

The development of Coscine started in 2018 and continued for two years, characterized by a start-up mentality in the development team. The primary aim was to develop a

platform that focuses on RDM and supports the implementation of the FAIR principles (Lang, Jansen, and Politze 2025) in everyday life. In addition, it was important that the RDM platform not only covers a single section of the research data lifecycle, but can, at best, accompany researchers throughout the entire research data lifecycle.

This started with a single stakeholder – RWTH Aachen University. The primary requirement was to ensure that the RDM needs of RWTH researchers were met. As a result, the first decisions for the Coscine platform were made very low-threshold and the foundation for a generic RDM platform could be laid quickly.

The Coscine team, which was initially very small, chose aspects of Scrum and Kanban that were tailored to individual needs to organize development. Initially, however, the focus in the development phase was on the development team mastering the use of GitLab, for example, and building up and further optimizing the planning of the Scrum sprints step by step.

Pilot Phase (2020–2023)

The pilot phase began in 2020 at RWTH Aachen, transitioning the start-up mentality into a testing mentality. The purpose of the phase was to find out where the strengths and weaknesses of the platform lie and to incorporate improvements based on feedback from pilot users.

During this phase, many new stakeholders were added who were involved in the development of Coscine (see Figure 1) through funding and some of whom are still involved today (e.g., NFDI4ING and NFDI-MatWerk). This increased the need to collect ideas and feature requests, requiring clear prioritization for the implementation. We will go over how exactly this was achieved and what has changed in our prioritization process in the next chapter (see Section 3).

As the number of stakeholders increased, the Coscine development team gradually grew larger, as did the number of users.

Regular Operation (since 2023)

Coscine has been in regular operation since March 2023. This marked a milestone in development. As part of Coscine.nrw, the platform is made available to all DH.NRW universities as Software as a Service. As a result, the number of stakeholders and their needs continues to rise, and a balance needs to be found between maintenance of existing and the release of new features. During regular operation, the prioritization of new feature requests was finely adjusted to reflect real-world conditions.

To further support the increasing number of users, in addition to an extensive range of training courses (e.g., workshops on the basic functions and creating metadata profiles), tutorial videos have also been published to help users get started[3]. Furthermore, RDM staff at the DH.NRW level are trained to advise researchers within their own organization and introduce them to using Coscine. A peer review process is employed to allocate storage space, ensuring that researchers receive storage allocations after a successful and independent review. This ensures that all universities and technical colleges are provided with storage space in accordance with funding providers' specifications. To manage this increased support effort, a dedicated support model was developed for Coscine, and both the development team and the service management teams have grown since the transition to regular operation.

3 Challenges & Approaches

Below we share some challenges we have encountered in recent years, along with our experiences and proposed solutions.

3.1 Scalability

Increasing user numbers naturally, with around 3,300 users now registered in Coscine, result in a rise in feature requests, requiring a robust concept to ensure scalability. For us, the key has been the prioritization of requirements to best accommodate stakeholder wishes according to their importance.

During the *Development Phase*, prioritization was done internally, i.e. the decision about which features to be developed next was made by one or two people. External stakeholder influence was minimal, enabling a strong focus on creating a generic RDM platform to support FAIR principles.

To effectively handle the growing number of stakeholders during the *Pilot Phase*, we applied a prioritization model from the business sector: the WSJF (Weighted Shortest Job First) formula, originating from the Scaled Agile framework (SAFe)[4]. WSJF is commonly used by project owners to prioritize backlogs and determine next steps.

However, during *Regular Operation*, it became clear that the WSJF method in its original form was unsuitable for Coscine. We therefore modified the method by removing "job size" from the calculation. This adjustment reflects the necessity to fulfill stakeholder

3 https://docs.coscine.de/de/videos/; *Visited on March 28, 2025.*
4 https://framework.scaledagile.com/wsjf; *Visited on March 28, 2025.*

requirements regardless of job size, ensuring continued funding. Additionally, certain adjustments were necessary to keep up with evolving RDM trends.

Table 1 summarizes our prioritization process across the different development phases.

Table 1: Prioritization Approach of Coscine Development.

Phase	**Changing World**	**Style**	**Decision Maker**	**Process**
Development	One Stakeholder	Internal	Product Owner, Group Lead Coscine	Discussion
Pilot	Stakeholders with individual RDM requirements	Industry	Steering Board (Product Owner, Group Leads: Coscine, Consulting, HPC, Service Management, Department Head)	Voting via Weighted Shortest Job First (WSJF)
Regular Operation	Many Stakeholders; focus on scalable features	RDM	Steering Board (see above)	WSJF Adaption: no division by job size anymore

Figure 2 shows the process from the creation of issues and epics to prioritization and development. The creation is undertaken by the product owner, people in service management (May 2025: two full-time positions), the Scrum Team itself or external stakeholders. Issues, i.e. small work tasks, are prioritized by the product owner. Epics comprise larger work packages and are a collection of issues on one topic. These are prioritized by the Steering Board. The Scrum Team (May 2025: two full-time and three part-time developers) implements the specific work tasks within a two-week sprint with the typical Scrum events (planning, daily, review, retro).

3.2 Transparency

We aim to develop Coscine transparently, clearly communicating our current projects and methods. A key step toward transparency was developing Coscine as an open-source platform, enabling other platforms to reuse our features. A notable example is the NFDI Login, added in late 2024 through the IAM4NFDI project. Coscine contributed to the first incubator phase, enabling exchanges about technical implementation, i.e. NFDI login and account merging.

We also communicate clearly how Coscine supports FAIR principles (Jansen et al. 2024). For instance, we published insights at Zenodo through the FAIR-IMPACT Program Work-

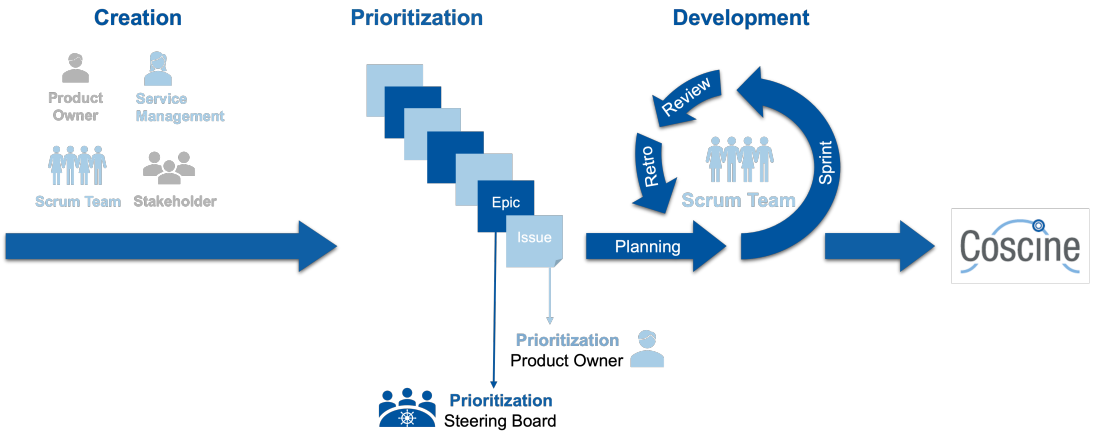


Figure 2: Agile Development Process for Coscine.

shop (Lang, Jansen, and Politze 2025). Additionally, a mission statement detailing our goals, mission, and values was published on our website[5].

Transparency extends to organizational practices. Our roadmap for larger work packages (EPICs) is publicly accessible, allowing users and stakeholders to view and directly submit feature requests[6]. At the development team level, work is tracked privately on a GitLab issue board, which reflects progress within each sprint. These issues represent smaller work packages that may collectively form an EPIC or address individual hotfixes and bug fixes. This level of development is not public to give the team space to work freely and to maintain a protected environment.

Transparency also involves reporting key performance indicators (KPIs). These metrics have been defined and are stored automatically. We have developed a visualization for external presentation, which is currently available only to selected individuals (e.g., stakeholders, DH.NRW staff). However, in coordination with the Coscine.nrw and DataStorage.nrw advisory board, we plan to make these publicly accessible over the course of 2025. This will provide insights into user numbers, quota usage, and login methods used.

3.3 RDM Streams

RDM is gaining significance, and trends in the field are constantly evolving. Coscine as an RDM platform is intended to grow with these trends and we therefore strive to keep an eye on the latest requirements and move forward as pioneers. For example, FAIR Digital Objects have been implemented (Politze et al. 2025) and continually adapted as part of the FAIR-IMPACT program.

5 https://about.coscine.de/en/about/mission-statement/; *Visited on March 28, 2025.*
6 https://git.rwth-aachen.de/groups/coscine/-/epic_boards/539; *Visited on March 28, 2025.*

The use of SPARQL queries for managing and retrieving metadata is becoming increasingly important for researchers. Recognizing this as a significant trend in RDM, we have identified SPARQL as a valuable addition to Coscine's capabilities. Offering researchers the possibility to use SPARQL enables more precise, efficient, and flexible querying of metadata, giving them finer control over their data. As part of our ongoing efforts to stay aligned with current and emerging RDM streams, we are actively exploring ways to integrate SPARQL query support into Coscine, empowering researchers to utilize advanced metadata management techniques in their everyday research activities.

Since excellent research depends on sufficient storage space, Coscine serves as a gateway to storage systems. Initially, we offered RDS, which will be replaced in 2025 by a new storage system called DataStorage.nrw. While this update introduces state-of-the-art technology for researchers, the transition poses a challenge: storage systems must typically be replaced approximately every five years, requiring repeated migrations. We are therefore currently in the migration phase from RDS to DataStorage.nrw and are establishing long-term migration plans for future cycles.

Coscine allows research data to be stored for up to 10 years after the end of a project. However, there are use cases where this period is not sufficient, and researchers require storage that guarantees long-term availability. To address this, we are currently collaborating with Long-Term Availability NRW (LZV.NRW) to explore integration options for long-term storage within Coscine.

3.4 Quality Management

To provide a functional and customizable platform for a large and diverse user base, a well-designed quality management approach is essential. Agile processes allow us to respond flexibly and confidently to changes, which is critical in the development area. We aim for a balanced focus between maintaining existing features and introducing new ones.

Development is inseparable from thorough testing, and the Coscine team has made several adjustments in this area in recent years. During the *Pilot Phase* (2018–2020), the priority was speed – many features were developed in a short time to deliver a basic RDM platform. Today, the focus has shifted to ensuring the quality of delivered features, even if this occasionally comes at the cost of development speed. Looking ahead, the team is placing greater emphasis on end-to-end (E2E) testing and adopting a test-driven development (TDD) approach.

We are also actively preparing for ISO/IEC 27001 certification to ensure security compliance – a major milestone in Coscine's development.

If users experience issues, our support team is available via a ticketing system, reachable by phone, chat, or email during service hours. Whether an on-call service will be needed in the future is currently under evaluation.

To continually improve the platform, we conduct an annual user survey to assess satisfaction and gather valuable feedback. We promote the survey through the Coscine newsletter, a banner on the platform, and the IT Center blog. For both the development and service management teams, these results are an important tool for understanding user needs and tracking changes compared to previous surveys.

4 Key Aspects

Mastering the challenges of a constantly changing world is not easy and requires continuous adjustments to methods and processes.

In essence, the following provides a summary of the biggest challenges faced by the Coscine RDM platform since its initial development until today:

1. **Scalability:** The platform should be usable by numerous stakeholders and therefore must fulfill a wide range of requirements and be quickly adaptable.

2. **Transparency:** We, as a team, clearly understand our mission and communicate it to the outside world. Additionally, we clearly define our processes and present them as openly and transparently as possible externally.

3. **RDM Streams:** The field of research data management is constantly evolving, and new changes are continually emerging. As an RDM platform, we aim to comply with these developments as effectively as possible and proactively move forward with the currents in the field.

4. **Quality Management:** The transition from a start-up to a permanent service requires constant attention to quality. We therefore do everything we can to deliver consistent quality while continuously improving.

We remain flexible and continuously adapt our workflows to changing requirements – like for example the adaptation of our prioritization method. We strive for active networking within the RDM sector, regularly participating in nationwide events, conferences, and similar activities (e.g. via fdm.nrw, NFDI, FAIR-Impact). Finally, we always remain true to our anchors: the FAIR principles, the IT Center of RWTH Aachen University, and agile methods.

Acknowledgements

The work on Coscine was partially supported with resources granted by DH.NRW for the project Coscine.nrw by Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen, by NFDI4Ing, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 442146713 and NFDI-MatWerk, funded by Deutsche Forschungsgemeinschaft (DFG) under project number 460247524.

Authorship Contributions

- Katja Jansen: Conceptualization, Methodology, Writing
- Petar Hristov: Conceptualization, Methodology, Writing
- Ilona Lang: Conceptualization, Methodology, Writing

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

Grönewald, Matthias, Patrick Mund, Matthias Bodenbrenner, Marc Fuhrmans, Benedikt Heinrichs, Matthias S. Müller, Peter F. Pelz, et al. 2022. "Mit AIMS zu einem Metadatenmanagement 4.0: FAIRe Forschungsdaten benötigen interoperable Metadaten". In *E-Science-Tage 2021: Share Your Research Data: Share Your Research Data,* edited by Vincent Heuveline and Nina Bisheh, 91–104. Heidelberg: heiBOOKS. ISBN: 978-3-948083-54-0. https://doi.org/10.11588/heibooks.979.c13721.

Jansen, Katja, Lukas C. Bossert, Amirreza Daei Rezaei Moghaddam, and Marius Politze. 2024. "Coscine – Makes your research data FAIR". In *From Data to Diamonds – Empowering Research with AI and RDM, Aachen (Germany), 14 Nov 2023 – 14 Nov 2023.* https://doi.org/10.18154/RWTH-2023-10428.

Lang, Ilona, Katja Jansen, and Marius Politze. 2025. *Implementing the FAIR Principles with Coscine.* Zenodo. https://doi.org/10.5281/zenodo.14886335.

Politze, Marius, Benedikt Heinrichs, Sirieam Hunke, Ilona Lang, and Thomas Eifert. 2025. "FAIR Digital Objects: FAIRtilizer for the Digital Harvest". In *EPiC Series in Computing,* 105:284–274. EasyChair. https://doi.org/10.29007/hfzk.

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. https://doi.org/10.1038/sdata.2016.18.

Ein standardisierter Datenmanagementplan als Hilfsmittel für niederschwelliges Forschungsdatenmanagement. Die Lösung für alle Herausforderungen?

Harald Kaluza [1]

[1] Deutsches Institut für Erwachsenenbildung – Leibniz-Zentrum für Lebenslanges Lernen e.V. (DIE), Bonn, Deutschland

Mit der seit 2017 geltenden Policy zum Forschungsdatenmanagement empfiehlt das Deutsche Institut für Erwachsenenbildung (DIE) seinen Forschenden die Erstellung FAIRer Daten. In den vergangenen Jahren konnten so erste (Forschungs-)Datensätze einer Nachnutzung durch Dritte zugeführt werden. Es wurde jedoch auch beobachtet, dass durch die Vielzahl an Maßnahmen, einzuhaltende (gesetzliche) Regelungen sowie einer zusätzlichen Dokumentation der Prozesse und Daten FDM oft als zusätzliche Last empfunden wird. Dieses Spannungsfeld hat zu einer erweiterten, zentralen Unterstützungsstruktur in Form von begleitender, persönlicher Beratung und des Angebots von Hilfsmitteln geführt. Seit 2024 wird am DIE für das Forschungsdatenmanagement der von ihm in einem Kooperationsprojekt großer Institutionen der Bildungsforschung mitentwickelte „Standardisierte Datenmanagementplan für die Bildungsforschung“ (Stamp) eingesetzt. Dieses Werkzeug hat den Anspruch, als „One Stop Shop“ einem Forschungsprojekt in allen Phasen niederschwellig sämtliche nötigen Maßnahmen und Prozesse projektspezifisch aufzuzeigen, vertiefende Informationen sowie weiterführende Materialien an die Hand zu geben. Der Text geht zunächst auf das Spannungsfeld zwischen den Interessen der Forschenden und den Anforderungen des Forschungsdatenmanagements ein, beleuchtet das institutionelle Beratungsangebot des DIE und stellt den Stamp als disziplinspezifische Lösung zum FDM vor. In einem weiteren Teil wird erörtert, ob der Stamp den formulierten Ansprüchen an ein niederschwelliges FDM gerecht werden kann. Es zeigt sich, dass der Stamp ein starkes Hilfsmittel für das FDM in der Bildungsforschung sein kann und das Verständnis für einzelne Maßnahmen deutlich verbessert.

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23920 (CC BY-SA 4.0).

Keywords: Forschungsdaten, Datenmanagementplan, Forschungsdatenmanagement, Standardisierung, Bildungsforschung

1 Ausgangslage und Herausforderungen

Das Deutsche Institut für Erwachsenenbildung – Leibniz Zentrum für Lebenslanges Lernen (DIE) mit Sitz in Bonn ist die zentrale Einrichtung für Weiterbildungsforschung in Deutschland. Es forscht zu Lernprozessen bei Erwachsenen, Weiterbildungsangeboten und deren Programmen, Lehrkräften in der Weiterbildung, der Organisation von Weiterbildungseinrichtungen sowie dem Weiterbildungssystem in Deutschland und Europa, betreibt Wissenstransfer und stellt Infrastrukturen für Forschung und Praxis bereit. Seine empirische (Weiter-)Bildungsforschung spiegelt sich in einem interdisziplinären Umfeld an Forschungsmethoden, Datenerhebungsverfahren und damit auch in einer Vielzahl an entstehenden qualitativen und quantitativen Daten unterschiedlichster typologischer Ausprägung wider. Um die Umsetzung eines geregelten und nachhaltigen Forschungsdatenmanagements (FDM) am DIE zu gewährleisten, wurde im Jahr 2017 eine Policy zum Forschungsdatenmanagement verabschiedet, die in einer zweiten Version 2024 aktualisiert wurde (Deutsches Institut für Erwachsenenbildung Leibniz-Zentrum für Lebenslanges Lernen 2024). Bereits hierbei hatte sich nach einer längeren Planungsphase gezeigt, dass die Einführung von Maßnahmen zum FDM eng mit den strategischen Vorgaben der Leitungsebene eines Instituts verknüpft, und auch von der Leitungsebene eines Instituts kommuniziert werden sollten (Kaluza 2019).

Das Ziel der Policy ist, dass Daten, die in Forschungsprozessen des DIE gewonnen werden, sowohl intern als auch extern nachnutzbar und zitierbar gemacht werden sollen, sofern es die rechtlichen und ethischen Rahmenbedingungen erlauben. Um dieses Ziel zu erreichen, müssen sich Forschende am DIE an die definierten Grundsätze halten, die im Folgenden kurz zusammengefasst seien: Jedes Forschungsvorhaben bzw. Forschungsprojekt muss das Management der zu generierenden Daten schon ab der Planungsphase berücksichtigen und in den ggf. zu beantragenden Mitteln einplanen. Diese und alle weiteren Maßnahmen werden in einem Datenmanagementplan projektspezifisch festgelegt. Unter besonderer Beachtung aller rechtlichen und ethischen Konventionen werden Handlungen durchgeführt, die die institutionelle Archivierung von mindestens 10 Jahren gewährleisten, im besten Fall aber eine Nachnutzung durch Dritte in einem fachspezifischen und akkreditierten Repositorium ermöglichen.

Klar war und ist, dass die Fülle an Anforderungen, die mit dem FDM einhergehen, sowohl die Forschenden als auch die Kuratierenden der Daten vor (neue) Herausforderungen stellen. So konstatierten Borgmann et al. bereits im Jahr 2012 (Borgman 2011), dass die Idee, Forschungsprozesse und -abläufe planmäßig auf die spätere Veröffentlichung der Daten abzustimmen für Forschende neu, ungewohnt und vor Allem mit einem gefühlten unnötigen Mehraufwand verbunden sei. Dieses aufgezeigte Spannungsfeld ist seitdem ein fester Be-

standteil des Diskurses im FDM und wird auch unter den Aspekten eines Kulturwandels hin zum FDM weiter thematisiert und disziplinübergreifend als eine der großen Herausforderungen bei dessen Umsetzung betrachtet (vgl. z. B. Ashiq, Usmani und Naeem 2020; Milewska u. a. 2021). Auch für die Umsetzung der FDM-Policy am DIE wurde deutlich, dass diesem Phänomen Rechnung getragen werden musste. Bereits in der ersten Fassung der Policy wurde festgelegt, dass Forschenden am DIE eine zentrale Unterstützungsstruktur zur Seite gestellt wird, die den kompletten Projekt- bzw. Forschungszyklus beim Managen der Daten begleitet. Diese besteht aus einem zentralen Ansprechpartner, der die Forschenden in allen organisatorischen, methodischen und technischen Fragen berät und ggf. zu weiteren betroffenen Organisationseinheiten des Instituts (z. B. Ethikkommission, Datenschutzbeauftragte, IT) vermittelt. Zudem werden Hilfsmittel zur Verfügung gestellt, die die Umsetzung der Maßnahmen des FDM erleichtern sollen, seien es beispielsweise Hilfen zur Beantragung von Drittmitteln, Vorlagen für informierte Einwilligungserklärungen oder vor Allem auch die Vorlage für einen Datenmanagementplan. Mit der Neufassung der Policy im Jahr 2024 wird die Nutzung des Stamp als zentrales Hilfsmittel festgelegt. Dieser wird im weiteren Verlauf dieser Publikation detaillierter vorgestellt,

Seit 2018 wurden am DIE auf diese Weise etwa 35 verschiedene Forschungsprojekte in etwa 150 Beratungen begleitet. Nach einem verhaltenen Beginn konnte insbesondere in den vergangenen zwei Jahren ein stark steigendes Beratungsaufkommen festgestellt werden. Die dabei abgedeckten Themenbereiche umfassen neben der allgemeinen Einführung in das FDM und dem Nutzen von Datenmanagementplänen, die Organisation und die Dokumentation von Daten und deren Archivierung und vor Allem die Themen Recht und Ethik, die in der empirischen Bildungsforschung eine zentrale Rolle einnehmen. Trotz einer wachsenden Akzeptanz und einer Vielzahl an abgedeckten Themen konnten jedoch immer wieder Beobachtungen gemacht werden, die sich auf den Wunsch nach spezifischeren, auf den eigenen Projektkontext angepasste und niederschwelliger Hilfsmittel und Vorlagen bezogen.

Das FDM in der empirischen Bildungsforschung setzt sich aus einer Vielzahl an einzelnen Prozessen und Maßnahmen zusammen, die den Forschungsprozess begleiten. Immer wieder kam es zu Fragen, warum bestimmte Maßnahmen im Projekt überhaupt durchgeführt werden sollten, das FDM schien an manchen Stellen zu reinem Selbstzweck zu gereichen. Der Bedarf an Handreichungen und Vorlagen für das FDM ist groß (vgl. z. B. Hausen u. a. 2022). Es wird jedoch immer wieder festgestellt, dass vorhandene Hilfsmittel meist nicht alle Aspekte des FDM abdecken, sie mithin zu kurz sind. Vorlagen für Datenmanagementpläne wurden zudem oft als reine Fragenkataloge wahrgenommen, bei denen Nutzende immer noch individuell und selbstständig Antworten finden mussten, für die sie wiederum weitere externe Quellen recherchieren mussten. Weitere Hilfsmittel waren zudem oft generisch und nicht auf die Herausforderungen der eigenen Disziplin bezogen (Netscher u. a. 2025; Bishop u. a. 2023).

Diese Beobachtungen am DIE deckten sich im Großteil mit denen weiterer Akteure in der Bildungsforschung. „Es fehlte bisher ein Rahmenmodel, das den Anspruch hat, die

Breite an Forschungsprojekten der empirischen Bildungsforschung abzudecken und dabei sowohl auf technische als auch datenschutz- und urheberrechtliche Aspekte einzugehen“ (Künstler-Sment u. a. 2024, S. 1120).

2 Der Standardisierte Datenmanagementplan für die Bildungsforschung (Stamp)

Die oben beschriebenen Herausforderungen und Ansprüche an ein niederschwelliges Datenmanagement im Bereich der empirischen Bildungsforschung führten zur Entwicklung des Standardisierten Datenmanagementplans für die Bildungsforschung (Stamp). Er wurde im Verbundprojekt „Domain Data Protokolle für die Empirische Bildungsforschung – ddp Bildung“[1], gefördert vom BMBF mit einer Laufzeit von drei Jahren zwischen 2019 und 2022 entwickelt. Der Stamp adaptiert hierbei das Konzept der Domain Data Protokolle (ddp) von Science Europe (Science Europe 2018), welches einen „Modell-DMP“ definiert, der sechs wichtige Themen des FDM *spezifisch für eine Disziplin* zusammenfasst und Orientierung bei den entsprechenden Prozessen bieten soll. Diese Themenbereiche (Module) sollen jeweils strukturiert Einzelbereiche und Hilfen enthalten, die bei der Orientierung unterstützen sollen. Diese bestehen aus *Mindestanforderungen*, *gesetzlichen Vorgaben*, die zu berücksichtigen sind, *Beispielen*, die das Verständnis der Anforderungen und gesetzlichen Rahmenbedingungen erhöhen sowie geltende *disziplinspezifische Standards* und *weiterführende Ressourcen*. Der Stamp als praktisch für die Bildungsforschung umgesetztes ddp bedeutet also bestenfalls einen niedereschwelligen „One Stop Shop“ für das Datenmanagement, der die Anforderungen und Charakteristika der Disziplin berücksichtigt. Er beinhaltet konkrete Handlungsanweisungen für das FDM in allen Phasen eines Projekts, die sich in seinen Modulen widerspiegeln, bettet diese in den Gesamtkontext des FDM ein und soll ein einfaches und zeitsparendes Generieren von nachnutzbaren, „fairen“ (Wilkinson u. a. 2016) Forschungsdaten ermöglichen. Forschende erhalten eine konkrete Auflistung von Maßnahmen mit Erklärungen, die sie dafür durchführen sollten und können bereits durchgeführte dokumentieren.

Der Stamp ist als Online-(Beta-)Version im Research Data Management Organizer (RDMO)[2] oder als Offline-Variante im Word- oder PDF-Format verfügbar[3]. Er wird seit dem Ende des Projekts vom Verbund Forschungsdaten Bildung (Verbund FDB[4]) gehostet, redaktionell betreut und weiterentwickelt.

1 https://www.forschungsdaten.org/index.php/DDP-Bildung; *Besucht am 31. März 2025.*
2 https://stamp-rdmo.aip.de/; *Besucht am 30. Mai 2025.*
3 https://www.forschungsdaten-bildung.de/stamp; *Besucht am 31. März 2025.*
4 https://www.forschungsdaten-bildung.de; *Besucht am 31. März 2025.*

2.1 Aufbau und Nutzung des Stamp

Der Stamp ist in seinen Inhalten modular aufgebaut und sollte auch im Rahmen eines Projektes modular, basierend auf den Kenntnissen des Nutzenden und den Spezifika eines Projektes genutzt werden. Inhaltlich besteht der Stamp aus den folgenden sieben *Modulen*: (1) Forschungsethik, (2) Datenschutz, (3) Urheberrecht, (4) Datenorganisation, (5) Nachvollziehbarkeit, (6) Daten teilen und (7) Langfristsicherung. Sie umfassen also den kompletten Zyklus eines Projektes bzw. der dort entstehenden Daten. Die von Science Europe definierten „Themen“ für das FDM wurden an dieser Stelle erweitert und in *Modul* umbenannt.

In jedes *Modul* wird zunächst mittels eines *Standards* (im ddp-Konzept noch „Mindestanforderung“) eingeführt, welcher das jeweilige Modul (also FDM-Thema) zusammenfasst, es in den Gesamtkontext des FDM in der empirischen Bildungsforschung setzt und die wichtigsten untergeordneten Themen und zu treffende Maßnahmen aufführt (vgl. Abbildung 1).

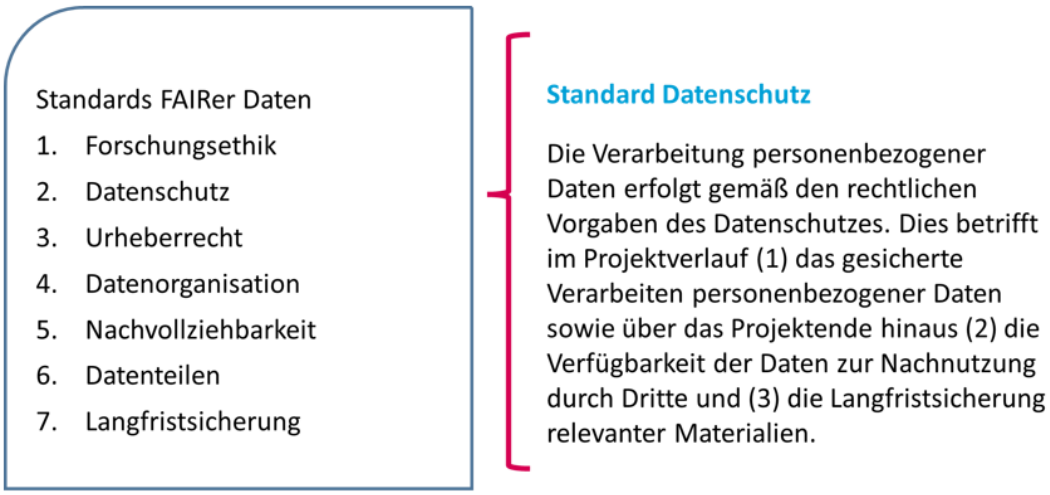


Abbildung 1: Übersicht Module und Auszug Mindestanforderung.

Zentrales Element eines jeden Moduls sind die *Checklisten* zu seinen einzelnen untergeordneten Themen. Diese umfassen eine Auflistung der konkreten Maßnahmen und Handlungsanweisungen zur Erfüllung des jeweiligen Standards. Darüber hinaus beinhalten die Listen neben den einzelnen Einträgen die zum jeweiligen Zeitpunkt relevanten Hilfestellungen. Das können *Querverweise* zu anderen, verwandten Modulen sein, aber auch Hinweise zu geltenden *rechtlichen Vorgaben*, *Handlungsempfehlungen* zu deren Umsetzung, konkrete *Fallbeispiele* aus der empirischen Bildungsforschung oder der Verweis auf das integrierte *Glossar* zur Erläuterung fachspezifischer Begriffe.

5.2.2. Checkliste Informierte Einwilligung

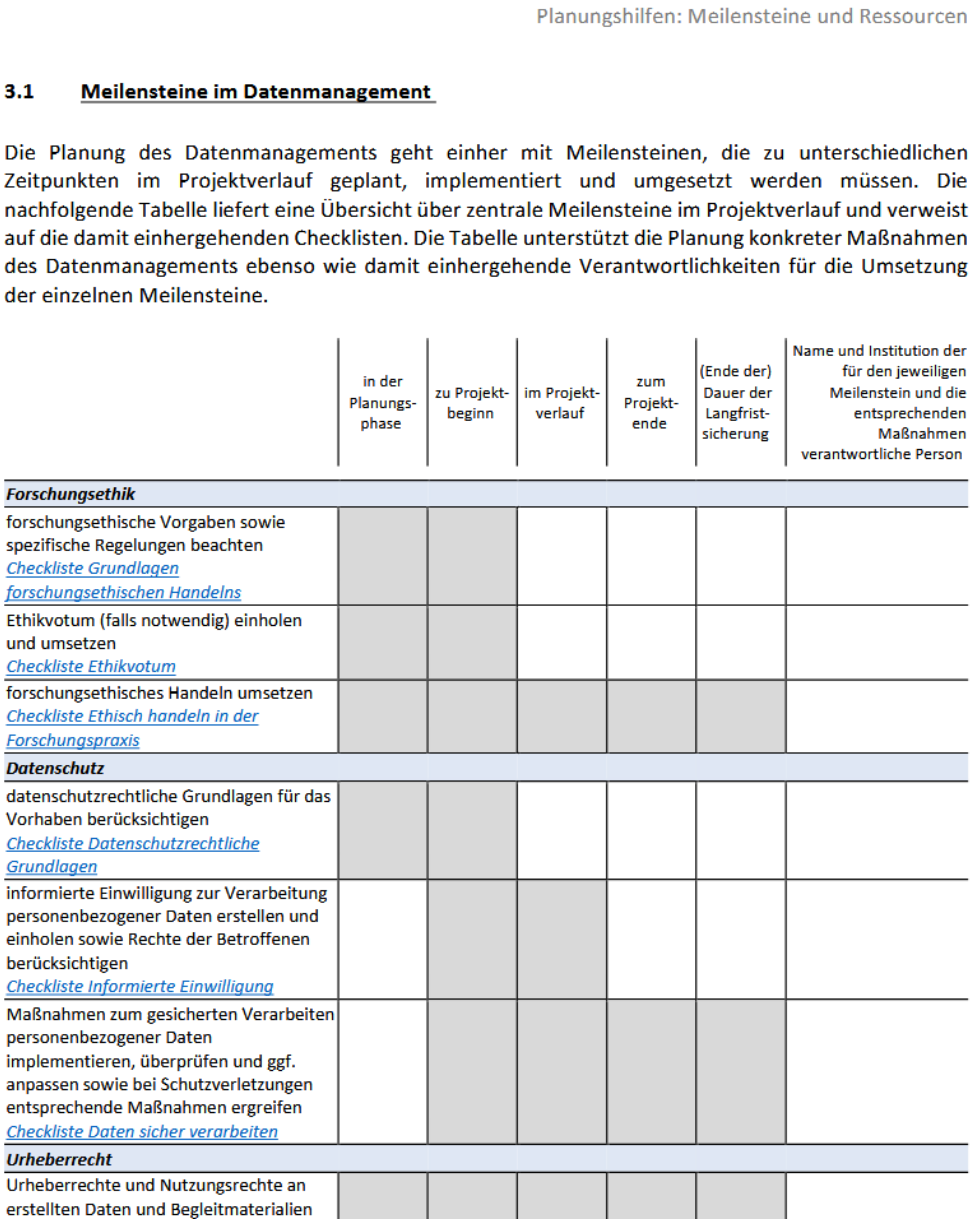
Task	zu bearbeiten*	ist erledigt	Hilfestellungen
Grundlagen der informierten Einwilligung klären. Das beinhaltet:	☐	☐	*Hinweis Informierte Einwilligung*
Institutionelle Vorgaben	☐		
Rechtsgrundlage und die sich daraus ergebenden Anforderungen an die Einwilligung	☒		*Fallbeispiel Betroffene aufklären*
Zwecke der geplanten Datenverarbeitungen insb. für das Vorhaben, für das Datenteilen, für die Langfristsicherung	☒		*Checklisten Grundlagen des Datenteilens* *Checklisten Grundlagen der Langfristsicherung*
Bei wem die Einwilligung einzuholen ist d. h. Personen, deren personenbezogene Daten verarbeitet werden, ggf. Sorgeberechtigte, Ankerpersonen zur Kontaktaufnahme, Dritte aufgrund von Aussagen der Studienteilnehmenden, Dritte aufgrund von Datenverknüpfungen	☒		*Hinweis Altersgrenzen* *Fallbeispiel Einholen von Kontaktdaten über Dritte* *Fallbeispiel Einwilligung Dritter (Datenschutz)*
Wie die Einwilligung einzuholen ist z. B. schriftlich, mündlich, online	☒		*Fallbeispiel Einwilligung bei verschiedenen Erhebungsmethoden*
Zu welchem Zeitpunkt die Einwilligung einzuholen ist	☒		*Fallbeispiel Einwilligung nach Datenerhebung*
Besonderheiten bei Foto- bzw. Videographien	☐		
Text für die informierte Einwilligung erstellen. Dabei ist zu beachten:	☐	☐	
Ausreichende Informationen geben, so dass die betroffene Person informiert einwilligt	☒		
einfache Form und **zielgruppengerechte Sprache**	☒		
wie das Votum einzuholen ist z. B. Verfahrensablauf, einzureichende Unterlagen	☒		*Fallbeispiel Einwilligung bei verschiedenen Erhebungsmethoden*

Abbildung 2: Auszug Checkliste Informierte Einwilligung.

Im Beispiel in Abbildung 2, Checkliste Informierte Einwilligung, wird neben den zu berücksichtigenden Schritten bei der Erstellung auf Fallbeispiele zur Aufklärung Betroffener Personen oder das Einholen der Einwilligung bei verschiedenen Erhebungsmethoden (schriftlich, mündlich, online) verwiesen.

Die Nutzung des Stamp erfolgt modular und unterstützt Forschende sowohl in der Orientierungs-, Planungs- und Umsetzungsphase mit den entsprechenden Modulen und weiteren Hilfsmitteln, die die Handhabung erleichtern und steuern sollen.

Zur *Orientierung* des Nutzenden dient als Einstieg der *Wegweiser,* der ermittelt, welche Module und Checklisten für das jeweilige Projekt relevant sind. Anhand von neun Fragen, die sich z. B. auf die Art der Datenerhebung und der daraus resultierenden Daten (qualitativ oder quantitativ) oder die untersuchten Personengruppen (Minderjährige? Erwachsene?) bezieht, werden die Module und untergeordneten Themen vorab eingegrenzt.



Planungshilfen: Meilensteine und Ressourcen

3.1 Meilensteine im Datenmanagement

Die Planung des Datenmanagements geht einher mit Meilensteinen, die zu unterschiedlichen Zeitpunkten im Projektverlauf geplant, implementiert und umgesetzt werden müssen. Die nachfolgende Tabelle liefert eine Übersicht über zentrale Meilensteine im Projektverlauf und verweist auf die damit einhergehenden Checklisten. Die Tabelle unterstützt die Planung konkreter Maßnahmen des Datenmanagements ebenso wie damit einhergehende Verantwortlichkeiten für die Umsetzung der einzelnen Meilensteine.

	in der Planungs-phase	zu Projekt-beginn	im Projekt-verlauf	zum Projekt-ende	(Ende der) Dauer der Langfrist-sicherung	Name und Institution der für den jeweiligen Meilenstein und die entsprechenden Maßnahmen verantwortliche Person
Forschungsethik						
forschungsethische Vorgaben sowie spezifische Regelungen beachten *Checkliste Grundlagen forschungsethischen Handelns*						
Ethikvotum (falls notwendig) einholen und umsetzen *Checkliste Ethikvotum*						
forschungsethisches Handeln umsetzen *Checkliste Ethisch handeln in der Forschungspraxis*						
Datenschutz						
datenschutzrechtliche Grundlagen für das Vorhaben berücksichtigen *Checkliste Datenschutzrechtliche Grundlagen*						
informierte Einwilligung zur Verarbeitung personenbezogener Daten erstellen und einholen sowie Rechte der Betroffenen berücksichtigen *Checkliste Informierte Einwilligung*						
Maßnahmen zum gesicherten Verarbeiten personenbezogener Daten implementieren, überprüfen und ggf. anpassen sowie bei Schutzverletzungen entsprechende Maßnahmen ergreifen *Checkliste Daten sicher verarbeiten*						
Urheberrecht						
Urheberrechte und Nutzungsrechte an erstellten Daten und Begleitmaterialien						

Abbildung 3: Auszug Meilensteinplan Datenmanagement.

Anhand der für das Projekt ermittelten relevanten Checklisten kann dann anhand des *Meilensteinplans* ermittelt werden, zu welchem Zeitpunkt eine bestimmte Checkliste im Projekt genutzt werden sollte (vgl. Abbildung 3). Weiterhin wird zur Unterstützung bei

der Erstellung von Drittmittelanträgen für das FDM eine anpassbare Textvorlage für einen Antrag zur Verfügung gestellt, die nach Eintragung der Projektspezifika einen vorformulierten Text mit den entsprechenden Maßnahmen des FDM passend zum Projekt erstellt. Zudem enthält der Stamp eine Kalkulationstabelle, die bei der Berechnung der etwaigen Sach- und Personenmittel anhand der relevanten Tätigkeiten aufführt.

Der Stamp greift also die Forderungen nach kontextabhängigen und niederschwelligen Hilfsmitteln auf, setzt diese durch einen modularen Aufbau, der es ermöglicht relevante Themen von unwichtigen zu trennen und diese in einen Zeitplan zu setzen, um.

3 Die Nutzungsstudie zum Stamp

Der Stamp wurde im Februar 2024 in der Version 1.0 auf den Seiten des Verbund FDB veröffentlicht (nach der Beta-Version im Jahr 2023) und ist seitdem auch bei den Beratungen zum FDM im DIE im Einsatz. Nach knapp einem Jahr in der Nutzung stellt sich die Frage, ob der Stamp den oben beschriebenen Ansprüchen und Herausforderungen gerecht werden kann. Hat der Stamp das FDM für die Forschenden niederschwellig zugänglich und leichter verständlich gemacht? Können die einzelnen Prozesse leichter und strukturierter durchgeführt werden und effizient gestaltet werden? Ist der Stamp gut zu handhaben? Hat sich aus Sicht der Datenkuratierenden die Qualität der entstandenen Forschungsdatensätze erhöht?

Im Januar 2025 startete die Stamp-Redaktionsgruppe im Verbund FDB eine Nutzungsstudie zu der mehr als 500 Bildungsforschende aus deutschen Hochschulen und außeruniversitären Bildungseinrichtungen eingeladen wurden. Die Studie besteht aus drei Phasen, in der Forschende den Stamp in einem ihrer Projekte benutzen und bewerten. Die Stamp-Redaktion bewertet zudem den Output der Nutzung. In Phase 1 der Studie werden in einem Vorab-Fragebogen die Projektspezifika der Teilnehmenden erfasst. Zudem werden die Vorkenntnisse und Erfahrungen im Forschungsdatenmanagement ermittelt. In Phase 2 bearbeiten die Teilnehmenden einen Teil ihres Projekts mit Hilfe eines passenden Moduls des Stamp. Hierbei können die Module zur Informierten Einwilligung, der Dokumentation der Daten, die Vorlage für den Drittmittelantrag oder das Wegweiserdokument genutzt werden. Die Teilnehmenden protokollieren die Nutzung des Stamp anhand eines Leitfadens, der sowohl die inhaltlichen (Verständlichkeit) als auch die formalen Aspekte (Aufbau, Usability) des Stamp bewerten soll. Das Protokoll wird zusammen mit dem entstandenen FDM-relevanten Projektoutput von der Stamp-Redaktion begutachtet, die Teilnehmenden bekommen ein Feedback. In Phase 3 werden in einem Interview einzelne Aspekte der Auswertung und des Feedbacks diskutiert.

Die Nutzungsstudie ist zum Zeitpunkt dieser Publikation noch nicht abgeschlossen. Neben ersten Tendenzen, die auf einem Rücklauf von fünf ausgewerteten Studien Teilnehmender

beruhen, fließen auch Erfahrungen aus der seit einem Jahr andauernden Nutzung des Stamp am DIE in die folgende Zusammenfassung der Ergebnisse mit ein.

Die grundsätzliche Rezeption des Stamp seitens der Nutzenden stellt sich sowohl in der Studie als auch im Rahmen der Nutzung am DIE positiv dar. Der Stamp wird insgesamt und auch innerhalb der Checklisten als gut strukturiert und (weitestgehend) verständlich wahrgenommen. Gelobt wurde insbesondere die strukturierte Aufbereitung bei den rechtlichen Modulen, die trotz der mitunter anspruchsvollen Inhalte eine gute Orientierung und Sicherheit vermitteln. Die Fallbeispiele mit konkretem Bezug zur Bildungsforschung, die gesetzlichen Rahmenbedingungen und Handlungsempfehlungen wurden häufig genutzt, obgleich eine direktere technische Einbindung in den Stamp (bisher Verlinkung zu teilweise externen Dokumenten auf dem Server des Verbund FDB) gewünscht wurde. Trotz der Menge an Modulen und der Breite der Themen gelang es bei der Gruppe der Befragten, einen passenden Einstieg in die Module und strukturierte Abläufe für das FDM zu vermitteln. Insbesondere der Meilensteinplan konnte hierbei unterstützen und den – durchaus vorhandenen – Respekt vor der Fülle an Maßnahmen reduzieren (Erfahrung aus DIE-Beratungen). Aus Sicht der Daten kuratierenden und Beratenden konnte nach Begutachtungen der Protokolle und erstellten FDM-Materialien und Daten festgestellt werden, dass deren Qualität deutlich erhöht wurde. Waren bei Begutachtungen von informierten Einwilligungen vor dem Stamp oft noch weitere Beratungen und Anpassungen nötig, konnten bei den, mit Hilfe des Stamp erstellten, Dokumenten, gut strukturierte und vollständige Einwilligungserklärungen bewertet werden.

Kritische Bewertungen einzelner Punkte beziehen sich auf einzelne Hilfsmittel, die teils zu abstrakt bzw. noch immer zu theoretisch formuliert sind. Auch ist der Bedarf nach weiteren konkreten Formulierungsbeispielen und Vorlagen groß und wird laut der Nutzenden nicht vollkommen abgedeckt. Wurden einerseits die Checklisten aufgrund ihrer Ausführlichkeit positiv bewertet, erscheinen sie anderen Teilnehmenden teils als zu kleinteilig und zu lang. Zudem wurde nicht bei jeder Checkliste deutlich, welches konkrete Ziel nach der Bearbeitung erreicht werden oder welches Produkt entstehen sollte, was insbesondere zum Tragen kam, wenn Teilnehmende nicht über den Wegweiser oder den Standard auf die Nutzung der Liste stießen. Hier müssten die einzelnen Inhalte noch stärker kontextualisiert werden. Zudem ist der Stamp strukturell auf ein in der Theorie „idealtypisches“ deutsches Forschungsprojekt in der Bildungsforschung zugeschnitten. Bei großen internationalen Kooperationsprojekten kann er zwar Orientierung bieten, auf Grund seiner Sprache und des enthaltenen Rechtsrahmens kann er internationalen Anforderungen aber nicht genügen.

Hinsichtlich der Usability lässt sich feststellen, dass die Nutzung eines Word- oder PDF-Dokumentes nicht als antiquiert oder einschränkend empfunden wurde. Trotz vorhandener RDMO-Instanz wird diese nicht unbedingt bevorzugt. Zu arbeiten ist an einer besseren und übersichtlicheren Verlinkung und Verzahnung der Inhalte und Hilfsmaterialien.

4 Weiterentwicklung

Der Stamp wird inhaltlich von einer Redaktion im Verbund FDB betreut, damit aktuelle Entwicklungen, z. B. im Bereich KI berücksichtigt werden können. Zudem soll der Stamp im Rahmen des NFDI-Konsortiums KonsortSWD – nfdi4society (der Verbund FDB ist dort eingebunden) inhaltlich, technisch und organisatorisch weiterentwickelt werden. Inhaltlich wird eine Erweiterung auf die Disziplinen Sozial- und Wirtschaftswissenschaften angestrebt, deren Methoden und Herausforderungen bei der Erhebung und Verarbeitung von Daten zu denen der Bildungsforschung einen großen Bezug aufweisen. Technisch soll der Fokus auf eine breitere und bessere Nutzung für maschinenlesbaren Austausch gelegt werden, in denen neben dem vorhandenen RDMO-Template auch das Programm DataWiz bedient werden kann. Zudem soll eine XML-Version eine universelle, plattformunabhängie Nachnutzung ermöglichen. Organisatorisch soll der Stamp in den RDM-Compass, der Informations- und Trainingsplattform von KonsortSWD eingebunden werden. Ebenfalls soll eine Anknüpfung an weitere NFDI-Initiativen und Konsortien erfolgen (z. B. base4nfdi). Ein Antrag für diese Maßnahmen im Rahmen der NFDI befindet sich zurzeit in Begutachtung.

5 Fazit

Der Stamp kann in seiner jetzigen Form ein starkes Hilfsmittel für das FDM in der Bildungsforschung sein. Erste Bewertungen zeigen, dass der Zugang zu und das Verständnis für einzelne Maßnahmen deutlich verbessert werden. Auch wurde eine Zunahme der Qualität der erstellten Materialien und Daten im FDM festgestellt. Der Stamp bleibt durch die Redaktion beim Verbund FDB auf aktuellem Stand und verstetigt sich durch inhaltliche, technische und organisatorische Erweiterung über KonsortSWD. Der Stamp kann zum immer noch aktuellen Diskurs und der Herausforderung der Awareness für das FDM einen positiven Beitrag leisten, in dem er durch seinen niederschwelligen und modularen Ansatz bestehende Hürden und Vorbehalte abbauen kann.

Danksagungen

Der Autor bedankt sich bei den Mitgliedern der Stamp-Redaktionsgruppe, die die Evaluation des Stamp durchführen. Weiterhin wird dem Verbund Forschungsdaten Bildung für die Pflege des Stamp und die gute Beratung gedankt.

Beiträge und Interessenkonflikte

Hiermit erkläre ich, dass kein Interessenkonflikt im Zusammenhang mit meiner Autorenschaft an der vorliegenden wissenschaftlichen Arbeit besteht. Es bestehen weder finanzielle noch persönliche Beziehungen zu Dritten, die die inhaltliche Darstellung oder die Ergebnisse der Publikation beeinflusst haben könnten.

Literaturverzeichnis

Ashiq, Murtaza, Muhammad Haroon Usmani und Muhammad Naeem. 2020. „A systematic literature review on research data management practices and services“. *Global Knowledge, Memory and Communication* 71 (8/9): 649–671. https://doi.org/10.1108/gkmc-07-2020-0103.

Bishop, Bradley Wade, Peter Neish, Ji Hyun Kim, Raphaëlle Bats, A. J. Million, Jake Carlson, Heather Moulaison-Sandy und Minh T. Pham. 2023. „Data Management Plan Implementation, Assessments, and Evaluations: Implications and Recommendations“. *Data Science Journal* 22. https://doi.org/10.5334/dsj-2023-027.

Borgman, Christine L. 2011. „The Conundrum of Sharing Research Data“. *SSRN Electronic Journal,* https://doi.org/10.2139/ssrn.1869155.

Deutsches Institut für Erwachsenenbildung Leibniz-Zentrum für Lebenslanges Lernen. 2024. „DIE-Forschungsdaten-Policy: Leitlinie zum Umgang mit Forschungsdaten des Deutschen Instituts für Erwachsenenbildung – Leibniz-Zentrum für Lebenslanges Lernen“. *Leibniz Gemeinschaft,* https://doi.org/10.58000/CMMD-7N86.

Hausen, Daniela, Gianpiero Favella, Matthias Fingerhuth, Kerstin Helbig, Sonja Hendriks, Alexia Meyermann und Volker Soßna. 2022. „Datenmanagementpläne in der Forschung – von Grundlagen zu Grundfragen“. *Bausteine Forschungsdatenmanagement,* https://doi.org/10.17192/bfdm.2022.1.8366.

Kaluza, Holger. 2019. „Forschungsdatenmanagement in einer wissenschaftlichen Spezialbibliothek – Chancen und Herausforderungen in einem interdisziplinären Forschungsinstitut“. In *Forschungsdaten – Sammeln, sichern, strukturieren. 8. Konferenz der Zentralbibliothek, Forschungszentrum Jülich, 4. – 6. Juni 2019,* herausgegeben von Bernhard Mittermeier. Forschungszentrum Jülich. Besucht am 2. Juli 2025. http://hdl.handle.net/2128/22275.

Künstler-Sment, Julia, Simon Eckert, Gianpiero Favella, Martin Kerwer und Malte Jansen. 2024. „Standardisierter Datenmanagementplan für die Bildungsforschung (Stamp) – Ein projektbegleitendes Tool für das Forschungsdatenmanagement“. *Zeitschrift für Erziehungswissenschaft* 27 (4): 1117–1140. https://doi.org/10.1007/s11618-024-01246-3.

Milewska, Agnieszka, Natalia Wiśniewska, Paulina Cimoszko und Jakub Rusakow. 2021. „A survey of medical researchers indicates poor awareness of research data management processes and a role for data librarians“. *Health Information & Libraries Journal* 39 (2): 132–141. https://doi.org/10.1111/hir.12403.

Netscher, Sebastian, Harald Kaluza, Reiner Mauer, Kati Mozygemba und Karsten Stephan. 2025. „The Standardized Data Management Plan for Educational Research: an Approach to Foster Tailored Data Management“. *International Journal of Digital Curation* 19 (1): 10. https://doi.org/10.2218/ijdc.v19i1.910.

Science Europe. 2018. *Guidance Document Presenting a Framework for Discipline-specific Research Data Management.* Zenodo. https://doi.org/10.5281/ZENODO.4925907.

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. https://doi.org/10.1038/sdata.2016.18.

How Are Research Data Referenced? The Use Case of the Research Data Repository RADAR

Dorothea Strecker [1,*], Kerstin Soltau [2], Felix Bach [2]

[1] Berlin School of Library and Information Science, Humboldt-Universität zu Berlin;
[2] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure;
[*] Corresponding author: dorothea.strecker@hu-berlin.de

Publishing research data aims to improve the transparency of research results and facilitate the reuse of datasets. In both cases, referencing the datasets that were used is recommended. Research data repositories can support data referencing through various measures and also benefit from it, for example using this information to demonstrate their impact. However, the literature shows that the practice of formally citing research data is not widespread, data metrics are not yet established, and effective incentive structures are lacking. This article examines how often and in what form datasets published via the research data repository RADAR are referenced. For this purpose, the data sources Google Scholar, DataCite Event Data and the Data Citation Corpus were analyzed.

The analysis shows that 27.9 % of the datasets in the repository were referenced at least once. 21.4 % of these references were (also) present in the reference lists and are therefore considered *data citations*. Datasets were referenced often in data availability statements. A comparison of the three data sources showed that there was little overlap in the coverage of references. In most cases (75.8 %), data and referencing objects were published in the same year. Two definition approaches were considered to investigate *data reuse*. 118 RADAR datasets were referenced more than once. Only 21 references had no overlaps in the authorship information – these datasets were referenced by researchers that were not involved in data collection.

Keywords: research data, data citation, research data repository, data reuse, data referencing

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23921 (CC BY-SA 4.0).

1 Introduction

The practice of citing datasets in research publications has been promoted for many years now. The Force 11 Joint Declaration of data citation principles state that "[...] data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse" (Data Citation Synthesis Group 2014). Data citation is ascribed many positive effects, including the attribution of data creators, the positioning of datasets within the scholarly record, or the assessment of the impact of datasets (Silvello 2018).

Research data repositories are important actors in realizing the vision of habitual data citation. They are facilitators of data citation and in turn might be interested in data metrics to communicate their impact (Lowenberg et al. 2019; Puebla et al. 2024). Data sharing, access and reuse are common themes in literature on "long-tail" research data – "[v]ast amounts of distributed, heterogeneous, often smaller scientific datasets of various types generated by individual researchers or small research groups; these data do not receive the same level of attention, funding, or infrastructure support as larger, well managed datasets but nevertheless contain a wealth of valuable specialized information that can potentially contribute to scientific knowledge" (Stahlman and Kouper 2024, p. 3).

This paper discusses the concepts *data use* and *data referencing*, outlines the measures a research data repository specialized on long-tail research data, RADAR[1], has implemented to support data referencing, and investigates how the collection of RADAR is referenced.

1.1 Data use

Researchers are advised to cite data when reusing them (Data Citation Synthesis Group 2014). This recommendation makes the process and motives of data referencing sound simple, and suggests a direct connection between using and referencing data. However, previous research has shown that research practices are complex, and the concepts referencing and using data need to be differentiated further to gain a better understanding of what *data citation* means.

Previous research has shown that data are used for varying purposes; some types of data use do not result in new findings (*background use*), and researchers therefore might not consider indicating data use in publications (Wynholds et al. 2012; Banaeefar et al. 2022). Data reuse is often distinguished from data use as an essential variant. However, upon

1 RADAR: https://doi.org/10.17616/R3ZX96; *Visited on May 13, 2025.*

closer inspection, the concept *data reuse* is not clearly defined. The literature discusses several approaches that could be used to identify data reuse:

- A dataset is used by someone other than the original data creator (data reuse defined by authorship, see He and Nahar 2016; Pasquetto, Randles, and Borgman 2017)
- A dataset is used multiple times (data reuse defined by number of uses, see Peters et al. 2016)
- A dataset is used to answer new research questions (data reuse defined by purpose, see Sandt et al. 2019)

Sandt et al. (2019) argue that all of these approaches have shortcomings, and research practices do not always fall clearly into one of the two categories *data use* and *data reuse*.

1.2 Data referencing

Researchers choose different methods to refer to data. Gregory et al. (2023) categorized referencing practices into *data citation* (data are referenced in the reference list), *data mention* (data are referenced in other parts of the publication), and *indirect data citation* (a related publication is referenced in the reference list). Previous research showed that data citation is rare, and that authors often choose other methods of referencing data (Belter 2014; Park and Wolfram 2017; Park, You, and Wolfram 2018; Quarati and Raffaghelli 2022). Bibliographic databases also indicate that data citation remains rare, with only a small proportion of datasets being cited overall (Ninkov et al. 2021; Peters et al. 2016; Robinson-García, Jiménez-Contreras, and Torres-Salinas 2016). However, there is some evidence that data citation increased over time (Peters et al. 2016; He and Nahar 2016). In comparison to evidence from bibliographic databases, authors also report higher rates of referencing data: In a survey, 58.3 % of participants reported that they often or always cite or mention data they use (Gregory et al. 2023). Studies that consider data citations as well as data mentions reveal higher rates of data referencing (Belter 2014; Gerasimov et al. 2024). There is also evidence that indirect data citation is common. 38 % of data papers – documents describing datasets and their generation – were cited (Stuart 2017). The analysis of a dataset from cancer research showed that indirect citation was more common than citing the dataset directly (Yoon et al. 2019). This suggests that bibliographic databases do not capture all forms of referencing data.

Characteristics of datasets impact the number of citations they receive. The number of citations Dryad datasets accrue in Scopus varies by discipline, and less aggregated forms were generally more likely to be cited (He and Nahar 2016). An analysis of the Data Citation Index (DCI) resulted in similar findings (Peters et al. 2016). This highlights the need to contextualize citation analyses by considering the subject of a dataset; however, metadata describing research data do not always include that information.

There are many challenges to detect and collect references to data, starting with methods authors choose to refer to datasets. Research has shown that when researchers refer to datasets, they sometimes refer to research data repositories instead (Peters et al. 2016; Yang and Colavizza 2025), or refer to datasets by name (Mathiak and Boland 2015). This makes it difficult to detect references and allocate them to specific datasets. Authors do not always use persistent forms of referencing datasets, such as DOIs (Yoon et al. 2019). There are many more opportunities for data references to get lost, even if authors include them in documents, for example if they are not submitted to indexing services by publishers (Borda 2023; Lowenberg et al. 2019).

Bibliographic resources vary significantly in the coverage they provide for data citations (Gerasimov et al. 2024). Open citation data initiatives are perceived to have great potential in tracing data (re)use, but they face challenges such as reference mining in full texts or the reliance on publishers providing open references (Taşkın 2025). The Data Citation Corpus aggregates references from several sources, including those that are not using DOIs (Puebla and Lowenberg 2024).

Several studies have analyzed self-citations, which can provide insights into the prevalence of data reuse. Bibliographic resources provide varying results when analyzed for rates of self-citation – the phenomenon is quite common in the DCI, but relatively low for datasets in OpenAlex (Krause et al. 2023; Park and Wolfram 2017). This might be a result of different approaches to indexing of the two sources. Self-citation rates are very high for generalist repositories: 85 % of citations of Dryad datasets in Scopus and 98.5 % of citations of Zenodo datasets were self-citations (He and Nahar 2016; Sandt et al. 2019). These findings could indicate that generalist repositories are frequently used by authors to deposit their data as a means to satisfy data policies or guidelines by journals and other stakeholders.

Repositories can facilitate data use. A study at the Inter-university Consortium for Political and Social Research (ICPSR), a social sciences data archive from the USA, has shown that an increase in the number of curatorial actions and subject terms assigned to data resulted in more users downloading data (Hemphill et al. 2022; Hemphill et al. 2024). They might also be interested in using evidence of data use themselves to demonstrate their value to stakeholders, since data citation "signals the added value of data repositories" (Puebla et al. 2024, p. 1).

1.3 RADAR – measures to support data use and referencing

The cross-disciplinary RADAR repository from FIZ Karlsruhe – Leibniz Institute for Information Infrastructure was developed as part of a DFG project. It has been available to academic institutions since 2017 as a comprehensive cloud service for archiving, publishing and optionally reviewing research data. Since 2021, "RADAR Local" has also

been available as an operating variant in which the software is run by FIZ Karlsruhe on institutional infrastructure. Furthermore, as a participant in the National Research Data Infrastructure (NFDI), FIZ Karlsruhe operates RADAR4Culture, RADAR4Chem, and RADAR4Memory – community-tailored publication services designed to meet the specific needs of their respective research communities. All RADAR offers and variants use the same software code. RADAR has already introduced a number of measures to promote the use and referencing of research data and to ensure that data remains usable in the long term.

Each published dataset receives a DOI based on the RADAR metadata schema[2] with ten mandatory and 13 optional fields based on the DataCite Kernel 4.4. Links to ORCID, ROR or GND promote attribution or linking to standard data. At present, the integration of terminologies via the TS4NFDI[3] is under development. Digital resources or software can be related via the metadata which can also be updated after publication. The license for the data set is selected from a variety of different license options; metadata are published under CC0. On the landing page of each dataset, RADAR offers citation recommendations in seven citation styles as well as export formats such as BibTeX, RIS and EndNote.

Particular attention is paid to the adherence to the FAIR Principles (Wilkinson et al. 2016). RADAR actively supports this through a variety of technical measures and even participated in the EU project FAIR-IMPACT. During this assessment process, the scores generated by the evaluator tool F-UJI (Devaraju et al. 2022), an indicator for the adherence of data publications to the FAIR Principles, was significantly increased (up to 91 %). RADAR metadata are widely available and machine-readable. Open interfaces such as REST API, OAI-PMH[4] for harvesting and FAIR Signposting[5] are offered. Integration into a knowledge graph and access via a SPARQL endpoint also enable semantic integration of the RADAR datasets into higher-level research data ecosystems.

Internal curation and an optional peer review step are provided for quality assurance during the publication workflow. Long-term archiving is implemented geo-redundantly in three copies at two locations. Published data is available for at least 25 years to guarantee potential for reuse over time.

Statistics on the landing pages indicate the number of page views and data downloads. From the perspective of the repository operators, it would be desirable if this section could be supplemented with information on data citation and data reuse. An integrated comment function, the sharing of the dataset via social media or the embedding of landing pages on other websites could additionally promote the dissemination of RADAR datasets in the future.

2 RADAR Metadata Schema: https://doi.org/10.25504/FAIRsharing.e26f92; *Visited on May 13, 2025.*
3 TS4NFDI: https://ts4nfdi.github.io/homepage/; *Visited on May 13, 2025.*
4 RADAR OAI-PMH: https://www.radar-service.eu/oai/; *Visited on May 13, 2025.*
5 FAIR Signposting: https://signposting.org/FAIR/; *Visited on May 13, 2025.*

2 Method

This study addresses the following questions:

1. Do researchers reference RADAR datasets?
2. Which methods do they choose when referencing RADAR datasets?
3. How do data sources compare in terms of the coverage of references?
4. How common is the reuse of RADAR datasets?

2.1 Data collection

DOIs assigned to RADAR datasets were retrieved from the RADAR API 2025-01-27. In a second step, references attributable to these DOIs were collected from three data sources: Google Scholar, the Data Citation Corpus, and DataCite Event Data.

Google Scholar indexes metadata and full texts of research outputs and was shown to offer comprehensive coverage of data references in previous research (Gerasimov et al. 2024). The Data Citation Corpus was developed by DataCite and the Chan Zuckerberg Initiative (Vierkant 2022). It identifies data citations in various sources, including references in full texts (Puebla and Lowenberg 2024). DataCite Event Data captures relationships between research data and other objects that are manifested in metadata, using the property *relatedIdentifier* of the DataCite Metadata Schema.[6] These data sources were selected to cover different approaches of capturing references to research data. Gerasimov et al. (2024) compiled an overview of the scope and availability of additional data sources covering references to research data.

References in Google Scholar were identified by searching for the dataset DOI in full texts between 2025-02-10 and 2025-02-12. All results were checked manually. Version 3.0 of the Data Citation Corpus was used, which was released 2025-02-01 (DataCite and Make Data Count 2025). Event data was retrieved from the DataCite API 2025-01-27. The references obtained from the three sources were deduplicated.

Similar to the approach used by Khan, Thelwall, and Kousha (2021), each full text referencing a RADAR dataset was accessed to determine where the reference occurred in the text (in the reference list, data availability statement, footnotes, full text; multiple methods of referencing were possible).

Metadata (author names, publication date) for datasets and referencing objects were added from OpenAlex and DataCite on 2025-02-10. Figure 1 gives an overview of the

6 DataCite Event Data: https://support.datacite.org/docs/consuming-citations-and-references; *Visited on May 13, 2025.*

data collection process. The resulting dataset includes 604 references between a RADAR dataset and a referencing object (DOI – DOI pairs). The data are published and can be accessed at RADAR (Strecker 2025).

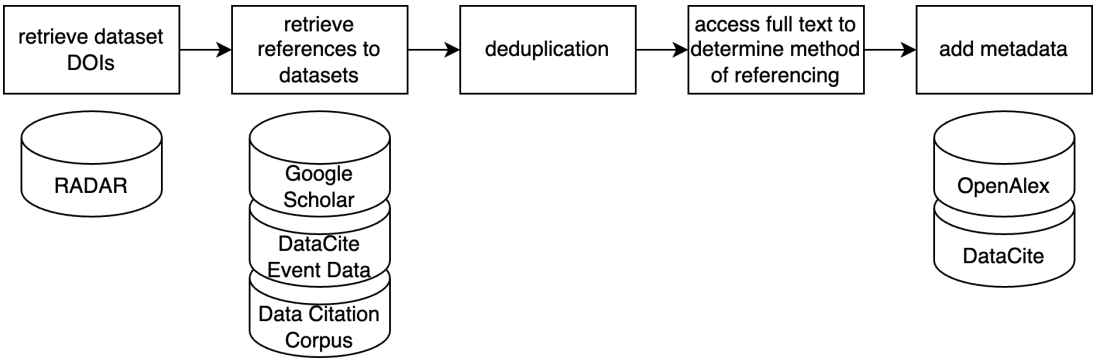


Figure 1: Outline of the data collection process.

3 Results

3.1 References to RADAR datasets

Of the 1.605 RADAR datasets, 27.9 % (448) were referenced at least once. The objects referencing the datasets were published between 2013 and 2025. The number of publications increases from 2017, reaches a peak in 2023, and tapers off in the following years (see Figure 2). In most cases (458; 75.8 %), datasets and referencing objects were published

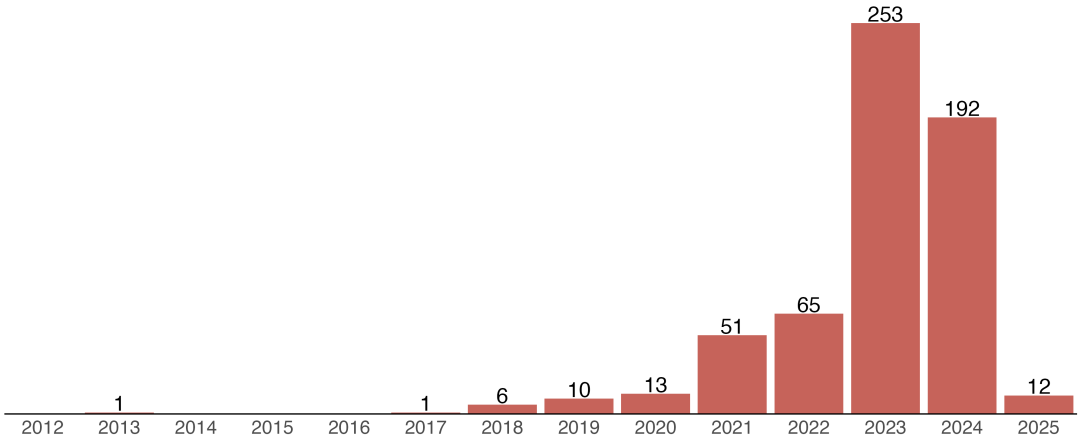


Figure 2: Distribution of the publication years of referencing objects.

in the same year. For the remaining cases, it was more common for data to be published before the referencing object (87; 14.4 %) than in the reverse order (49; 8.1 %).

3.2 Referencing methods

The analysis of the full text of the referencing object revealed that data citations (data are referenced in the reference list) represent 21.4 % of the references in the sample, 78.6 % are data mentions (data are referenced in other parts of the publication; see Figure 3, A). This categorization is based on Gregory et al. (2023), and is described in more detail above.

The most common position of data mentions (276) are in data availability statements, followed by other parts of the text, such as methods sections (60), and in footnotes (12; see Figure 3, B). Authors use more than one of these referencing methods in 388 cases.

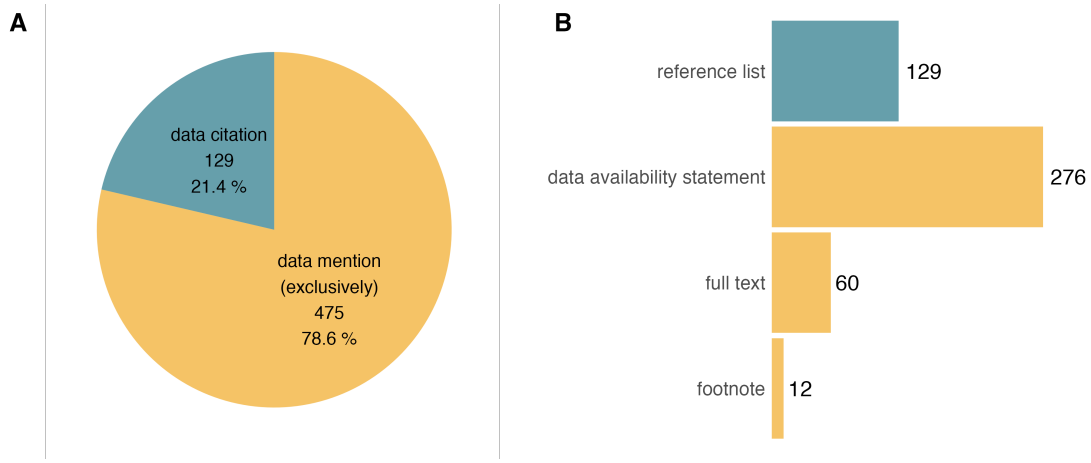


Figure 3: A referencing methods (based on Gregory et al. 2023), n = 604 references; B position of the reference in the referencing object (multiple options can apply).

In 220 cases, no referencing method could be determined, either because the referencing object was not a text (for example datasets), or because the DOI was not found in the full text (for example because the citation style did not display the DOI).

3.3 Coverage of data sources

369 references to RADAR datasets were found in Google Scholar, 37 in the Data Citation Corpus and 292 in DataCite Event Data. Some references were found in more than one data source, but overall, there is little overlap (see Figure 4). This applies to Google Scholar and DataCite Event Data in particular: Most references in the sample were identified in only one of these sources. Google Scholar covers most cases where datasets are referenced in the full text, whereas DataCite Event Data covers most cases where references are established in metadata only.

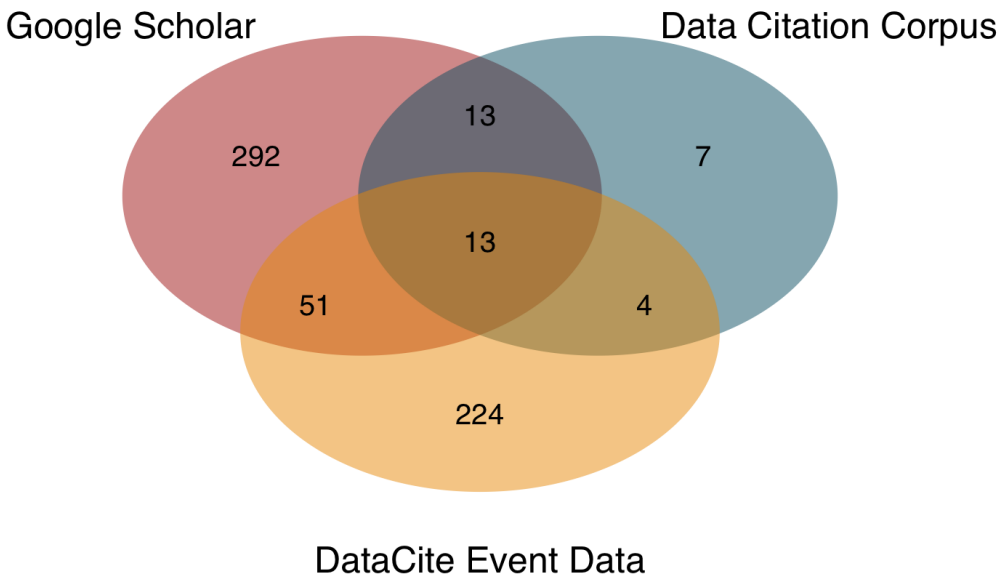


Figure 4: Number of references in the data sources.

3.4 Reuse of RADAR datasets

Two definitions were used to determine whether a RADAR dataset was reused: 1) the dataset was referenced more than once, and 2) there is no overlap in authorship between the dataset and the referencing object. 118 datasets in the sample (7.4 % of all RADAR datasets) were referenced more than once and could be considered "reused". However, it is important to note that there are preprints and corrections among these referencing objects that skew the results. 21 datasets (1.3 % of all RADAR datasets) have no overlap in authorship, which means they are referenced by researchers that were not involved in data collection.

4 Discussion

The analysis shows that RADAR datasets are referenced – a considerable share of the datasets published in RADAR were referenced at least once. The results also indicate that the practice of publishing and referencing datasets might have become more common over time.

"Data citation" can be a useful metaphor, for example to communicate to researchers the relevance of publishing and referencing their data for ensuring transparency of results. However, referencing data differs from referencing literature in some key aspects. For example, the analysis has shown that researchers often use different methods for referencing data, and self-citation is very common. This could indicate that researchers attribute

different meanings to referencing data. More research is needed to investigate what these references mean in the scholarly ecosysten; in the meantime, the metaphor "data citation" should be used carefully. Future research could also investigate whether references to research data vary depending on the type of repository they are published in. The self-citation rate at RADAR is very high (98.7 %). This observation is consistent with studies of other generalist repositories that are outlined above. It remains to be determined if this is the case for all types of research data repositories, or generalist repositories specifically.

The analysis also revealed that data sources vary significantly in terms of the references they cover. This is likely a result of researchers' referencing practices and the workflows data sources have established for recording connections between data and text publications. It is particularly noteworthy that in the case of RADAR, data sources tend to either capture references made in the full text or references made in metadata. To get a more complete representation of references to data, analyses should use multiple sources.

Although researchers' motives for referencing data were not the primary focus of this study, the results indicate that researchers reference data to adhere to policies, for example of journals or funders. This assumption seems plausible, because the most prevalent method researchers chose when referencing data was data availability statements, which are sometimes mandated or recommended to them. Unfortunately, adhering to these policies does currently not lead to the promised incentives for data sharing in the form of "impact" actually being realized, because the recommended referencing methods (data availability statements) are not recorded by databases that are easily machine readable. To make data referencing rewarding for researchers, policies and workflows of sources for reference data should be aligned more closely.

Repositories that want to use reference data to communicate the impact of datasets in their collection to stakeholders need a complete, accurate, and convenient source for this type of data. At the moment, reference data is dispersed and has to be reviewed and cleaned manually before it can be used. A truly useful data source would also require the cooperation of multiple stakeholders to ensure that references are counted as completely as possible. However, reference data alone do not convey impact. Repositories need data metrics – careful interpretations of reference data that give meaningful insights. This includes clear definitions of concepts like data use and data citation, and a consistent system for counting references. This is one of the goals of a new FORCE 11 group, "Data usage typologies"[7]. Overall, data metrics must be developed carefully to avoid harmful effects and misuse, as seen with metrics like the Journal Impact Factor.

7 FORCE 11 group Data usage typologies: https://force11.org/group/data-usage-typologies/; *Visited on May 13, 2025.*

5 Conclusion

This study demonstrated that datasets published in RADAR are referenced in journal articles and other research outputs. High self-citation rates and references in data availability statements suggest that researchers primarily reference data to make results of their own research more transparent, not because they reused existing data. Although authors might be asked to reference data in data availability statements, these references are often missing from sources of reference data. Policies and workflows of sources for reference data should be aligned more closely to reward data referencing.

The process also showed that analyses of data citations currently require a considerable amount of effort and care. Sources for reference data vary significantly in terms of their coverage and often require manual cleaning. Reference counts should be interpreted with caution – in order to meaningfully communicate their impact, repositories need data metrics, which are still under development.

Authorship Contributions

- Dorothea Strecker: conceptualization, writing (original draft), formal analysis, data curation.
- Kerstin Soltau: conceptualization, writing (original draft).
- Felix Bach: conceptualization, writing (original draft).

Conflict of Interest

The authors have no conflicts of interest to declare.

Bibliography

Banaeefar, Homeyra, Sarah Burchart, Elizabeth Moss, and Eszter Palvolgyi-Polyak. 2022. *Best Practice May Not Be Enough: Variation in Data Citation Using DOIs.* Poster presented at the annual meeting of the International Association for Social Science Information Service and Technology, June 9, 2022. https://doi.org/10.7302/4809.

Belter, Christopher W. 2014. "Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets". Edited by Howard I. Browman. *PLoS ONE* 9 (3): e92590. https://doi.org/10.1371/journal.pone.0092590.

Borda, Susan. 2023. “If Data is Used in the Forest and No-one is Around to Hear it, Did it Happen? a Citation Count Investigation”. *International Journal of Digital Curation* 17 (1): 14. https://doi.org/10.2218/ijdc.v17i1.830.

Data Citation Synthesis Group. 2014. “Joint Declaration of Data Citation Principles”. *FORCE11,* https://doi.org/10.25490/A97F-EGYK.

DataCite and Make Data Count. 2025. *Data Citation Corpus Data File.* Zenodo. https://doi.org/10.5281/zenodo.14897662.

Devaraju, Anusuriya, Robert Huber, Mustapha Mokrane, Patricia Herterich, Linas Cepinskas, Jerry de Vries, Herve L’Hours, Joy Davidson, and Angus White. 2022. *FAIRsFAIR Data Object Assessment Metrics.* Zenodo. https://doi.org/10.5281/ZENODO.6461229.

Gerasimov, Irina, Binita KC, Armin Mehrabian, James Acker, and Michael P. McGuire. 2024. “Comparison of datasets citation coverage in Google Scholar, Web of Science, Scopus, Crossref, and DataCite”. *Scientometrics* 129 (7): 3681–3704. https://doi.org/10.1007/s11192-024-05073-5.

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.* Zenodo. https://doi.org/10.5281/ZENODO.7555266.

He, Lin, and Vinita Nahar. 2016. “Reuse of scientific data in academic publications: An investigation of Dryad Digital Repository”. *Aslib Journal of Information Management* 68 (4): 478–494. https://doi.org/10.1108/AJIM-01-2016-0008.

Hemphill, Libby, Amy Pienta, Sara Lafia, Dharma Akmon, and David A. Bleckley. 2022. “How do properties of data, their curation, and their funding relate to reuse?” *Journal of the Association for Information Science and Technology* 73 (10): 1432–1444. https://doi.org/10.1002/asi.24646.

Hemphill, Libby, Andrea Thomer, Sara Lafia, Lizhou Fan, David Bleckley, and Elizabeth Moss. 2024. “A dataset for measuring the impact of research data and their curation”. *Scientific Data* 11 (1): 442. https://doi.org/10.1038/s41597-024-03303-2.

Khan, Nushrat, Mike Thelwall, and Kayvan Kousha. 2021. “Measuring the impact of biodiversity datasets: data reuse, citations and altmetrics”. *Scientometrics* 126 (4): 3621–3639. https://doi.org/10.1007/s11192-021-03890-6.

Krause, Geoff, Madelaine Hare, Mike Smit, and Philippe Mongeon. 2023. *Who Re-Uses Data? A Bibliometric Analysis of Dataset Citations.* https://doi.org/10.48550/ARXIV.2308.04379.

Lowenberg, Daniella, John Chodacki, Martin Fenner, Jennifer Kemp, and Matthew B. Jones. 2019. *Open Data Metrics: Lighting the Fire.* Zenodo. https://doi.org/10.5281/zenodo.3525349.

Mathiak, Brigitte, and Katarina Boland. 2015. “Challenges in Matching Dataset Citation Strings to Datasets in Social Science”. *D-Lib Magazine* 21 (1/2). https://doi.org/10.1045/january2015-mathiak.

Ninkov, Anton Boudreau, Kathleen Gregory, Isabella Peters, and Stefanie Haustein. 2021. “Datasets on DataCite – an Initial Bibliometric Investigation”. In *International Conference on Scientometrics and Informetrics (ISSI 2021).* Leuven, Belgium (Virtual): Zenodo. https://doi.org/10.5281/zenodo.4730857.

Park, Hyoungjoo, and Dietmar Wolfram. 2017. “An examination of research data sharing and re-use: implications for data citation practice”. *Scientometrics* 111 (1): 443–461. https://doi.org/10.1007/s11192-017-2240-2.

Park, Hyoungjoo, Sukjin You, and Dietmar Wolfram. 2018. “Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields”. *Journal of the Association for Information Science and Technology* 69 (11): 1346–1354. https://doi.org/10.1002/asi.24049.

Pasquetto, Irene V., Bernadette M. Randles, and Christine L. Borgman. 2017. “On the Reuse of Scientific Data”. *Data Science Journal* 16 (0): 8. https://doi.org/10.5334/dsj-2017-008.

Peters, Isabella, Peter Kraker, Elisabeth Lex, Christian Gumpenberger, and Juan Gorraiz. 2016. “Research data explored: an extended analysis of citations and altmetrics”. *Scientometrics* 107 (2): 723–744. https://doi.org/10.1007/s11192-016-1887-4.

Puebla, Iratxe, and Daniella Lowenberg. 2024. “Building Trust: Data Metrics as a Focal Point for Responsible Data Stewardship”. *Harvard Data Science Review,* number Special Issue 4, https://doi.org/10.1162/99608f92.e1f349c2.

Puebla, Iratxe, Stan Neumann, Kelly Stathis, Julian Gauthier, Audrey Hamelers, Lars Holm Nielsen, Luca Belletti, Fitz Elliott, and Mark Hahnel. 2024. *GREI Data citation best practices for repositories.* Technical report. https://doi.org/10.5281/ZENODO.10562429.

Quarati, Alfonso, and Juliana E Raffaghelli. 2022. “Do researchers use open research data? Exploring the relationships between usage trends and metadata quality across scientific disciplines from the Figshare case”. *Journal of Information Science* 48 (4): 423–448. https://doi.org/10.1177/0165551520961048.

Robinson-García, Nicolas, Evaristo Jiménez-Contreras, and Daniel Torres-Salinas. 2016. “Analyzing data citation practices using the data citation index”. *Journal of the Association for Information Science and Technology* 67 (12): 2964–2975. https://doi.org/10.1002/asi.23529.

Sandt, Stephanie van de, Sünje Dallmeier-Tiessen, Artemis Lavasa, and Vivien Petras. 2019. “The Definition of Reuse”. *Data Science Journal* 18 (1): 22. https://doi.org/10.5334/dsj-2019-022.

Silvello, Gianmaria. 2018. “Theory and practice of data citation”. *Journal of the Association for Information Science and Technology* 69 (1): 6–20. https://doi.org/10.1002/asi.23917.

Stahlman, Gretchen R., and Inna Kouper. 2024. “Evolution of the “long-tail” concept for scientific data”. *Journal of the Association for Information Science and Technology,* https://doi.org/10.1002/asi.24967.

Strecker, Dorothea. 2025. *Referenzierung von Forschungsdatenpublikationen in RADAR.* https://doi.org/10.22000/FBHFGZY8D43R3TJW.

Stuart, David. 2017. “Data bibliometrics: metrics before norms”. *Online Information Review* 41 (3): 428–435. https://doi.org/10.1108/OIR-01-2017-0008.

Taşkın, Zehra. 2025. “Sustaining the “frozen footprints” of scholarly communication through open citations”. *Journal of the Association for Information Science and Technology,* https://doi.org/10.1002/asi.24982.

Vierkant, Paul. 2022. “Wellcome Trust and the Chan Zuckerberg Initiative Partners with DataCite to Build the Open Global Data Citation Corpus”. *DataCite Blog,* https://doi.org/10.5438/VJZ9-KX84.

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. https://doi.org/10.1038/sdata.2016.18.

Wynholds, Laura A., Jillian C. Wallis, Christine L. Borgman, Ashley Sands, and Sharon Traweek. 2012. “Data, data use, and scientific inquiry: two case studies of data practices”. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries,* 19–22. JCDL ’12. ACM. https://doi.org/10.1145/2232817.2232822.

Yang, Puyu, and Giovanni Colavizza. 2025. *Research Data in Scientific Publications: A Cross-Field Analysis.* arXiv. https://doi.org/10.48550/arXiv.2502.01407.

Yoon, JungWon, EunKyung Chung, Jae Yun Lee, and Jihyun Kim. 2019. “How research data is cited in scholarly literature: A case study of HINTS: How HINTS data is cited in scholarly literature”. *Learned Publishing* 32 (3): 199–206. https://doi.org/10.1002/leap.1213.

Auslotung des Potenzials von Fachreferent:innen als FDM-affine Unterstützungsgruppe in einer forschungsintensiven Hochschullandschaft

Roberto Cozatl [1]

[1] Universitäts- und Landesbibliothek Sachsen-Anhalt, Martin-Luther-Universität Halle-Wittenberg

Der Vortrag stellte vor der Open-Science-Abteilung und das Fachreferat, die forschungsnahen Dienste der Universitäts- und Landesbibliothek Sachsen-Anhalt, die in engem Kontakt mit Forschern stehen und diese in einer Reihe von Fragen zum Forschungsdatenmanagement beraten. Das Ziel der Präsentation bestand darin, einen Einblick in die Prozesse zu geben, die es den Fachreferent:innen ermöglichten, eine effektivere und FDM-affine Arbeitsgruppe zu bilden.

Die Ergebnisse, die im ersten Jahr seit der Einführung dieser Veränderungen in der Bibliothek erzielt wurden, wurden im Detail vorgestellt. Die Einzelheiten der Ausbildung und der verwendeten Materialien wurden ebenso besprochen wie unser FDM-Curriculum, das sich dynamisch entwickelt, um den spezifischen Bedürfnissen unserer Gemeinschaften gerecht zu werden. Unsere Ansichten und Pläne für die Fortsetzung dieser FDM-Task Force im Jahr 2025 wurden ebenfalls diskutiert. Im Rahmen der Vorstellung der Idee wurde insbesondere darauf eingegangen, die Fachreferent:innen in die FDM-Support-Landschaft der akademischen Einrichtung zu integrieren. Dies soll gewährleisten, dass das Potenzial und die Arbeit der Fachreferent:innen als Ergänzung und Unterstützung für alle an der FDM-Unterstützung beteiligten Gruppen dienen. Zu den Akteuren zählen Data Stewards und Data Scientists.

Keywords: Fachreferat, Data Stewards, Data Scientists, FDM-Kompetenzen, FDM-Beratungen, Personalentwicklung, Forschungsdatenmanagement

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23922 (CC BY-SA 4.0).

1 Ausgangssituation

Die Martin-Luther-Universität (MLU) bietet als größte Universität in Sachsen-Anhalt ein breites Spektrum an Studienrichtungen an. Sie ist zudem Teil eines komplexen Forschungsumfelds, denn die wichtigsten deutschen Forschungsorganisationen sind mit eigenen Instituten in der Stadt vertreten. Daher sind die Bedürfnisse der verschiedenen wissenschaftlichen Communities vielfältig und sehr spezifisch.

Die Universitäts- und Landesbibliothek als größte Bibliothek des Landes Sachsen-Anhalt betreut einen großen Nutzerkreis, darunter rund 20.000 Studierende und etwa 4.000 Mitarbeiter:innen, von denen viele in verschiedenen Disziplinen und Forschungsprojekten auf regionaler, nationaler und internationaler Ebene tätig sind.

Über ihr physisches Netzwerk von Zweigbibliotheken bietet die Universität sowohl zentral als auch dezentral Unterstützung und Dienstleistungen für ihre verschiedenen Forschungsgemeinschaften in den Natur-, Sozial- und Geisteswissenschaften an.

Die sog. forschungsnahen Dienste der Bibliothek wurden so zusammengestellt, dass eine effiziente Unterstützung für unsere verschiedenen Communities gewährleistet ist. Die Bibliothek verfügt über engagierte Mitarbeiter:innen, die Open-Science-Angelegenheiten unterstützen. Aus dem im Jahr 2017 aufgestellten Team entwickelte sich seit dem Jahr 2022 eine eigenständige Abteilung.[1]

Seit 2018 bietet die Universität Unterstützung für Open Science und Forschungsdatenmanagement an. Diese Unterstützung wird in Form von drei Hauptlinien gewährt. Die Abteilung Open Science fungiert als Informationsbroker, der aktuelle und maßgeschneiderte Informationen über Open-Science-Aktivitäten, Schulungsangebote, Workshops usw. bereitstellt. Sie berät außerdem zu verschiedenen Fragen des Forschungsdatenmanagements, beispielsweise zu Datenformaten, zur Kuratierung und Aufbereitung von Daten für das Online-Publizieren und die Langzeitarchivierung, zur Erstellung von Forschungsdatenmanagementpläne. Auch zu rechtlichen Fragen, wie der Verwendung von Lizenzen für die Veröffentlichung, dem Urheberrecht, dem gewerblichen Schutzrecht und dem Datenschutz, gibt sie Auskunft. Die Bibliothek organisiert und führt Seminare, Informationsveranstaltungen, Workshops und Schulungen zu einer Reihe von Themen durch.[2]

Zu den forschungsnahen Diensten der Bibliothek gehören auch die Fachreferent:innen der Bibliothek, die ein breites fachdisziplinarisches Spektrum abdecken. Das Fachreferat spielt eine wichtige Rolle bei der fachbezogenen Unterstützung im Zusammenhang mit Open Science, Open-Access-Publikation und FDM-Angelegenheiten. In den vergangenen

1 Siehe: https://bibliothek.uni-halle.de/wp-content/uploads/2024/09/Organigramm_ULB_04-2023.pdf; *Besucht am 28. März 2025.*

2 https://bibliothek.uni-halle.de/schulen-beraten/forschungsdaten/; *Besucht am 21. März 2025.*

Jahren ist die Zahl dieser Kolleg:innengruppe gewachsen. Im Jahr 2024 bestand diese Gruppe aus 14 Mitarbeiter:innen.

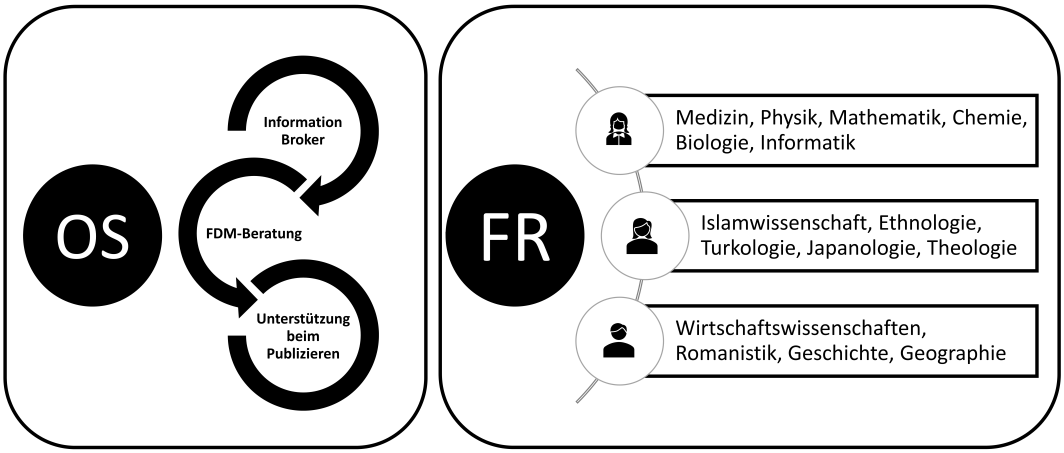


Abbildung 1: Der Open Science Abteilung (OS) und das Fachreferat (FR): Forschungsnahe Dienste an der Universitäts- und Landesbibliothek Sachsen-Anhalt im Fokus des Vortrags.

2 Motivation

Die Motivation für diese Arbeit entstand aus der Erkenntnis, dass unsere Fachreferent:innen bis vor Kurzem nicht immer in der Lage waren, ihr volles Potenzial auszuschöpfen, wenn es um FDM-Unterstützungsaufgaben ging. Dies lag unter anderem daran, dass sie in viele andere Bereiche der Bibliothek eingebunden waren und daher oft wenig Zeit für FDM-Aufgaben hatten. Eine Analyse dieser Situation ergab, dass bestimmte administrative Aufgaben, für die diese Gruppe in ihren verschiedenen Zweigbibliotheken zuständig ist, sie daran hinderte, mehr Ressourcen in anderen Bereichen einzusetzen.

Angesichts der wachsenden Bedeutung von Forschungsdaten und der damit verbundenen Verantwortung von wissenschaftlichen Einrichtungen und ihren Mitarbeiter:innen bei der Unterstützung forschungsbezogener Aufgaben und Aktivitäten (Helling u. a. 2022), war es für die Bibliothek unabdingbar, eine Reihe von Änderungen an ihrer Struktur und ihren Arbeitsabläufen vorzunehmen, um mehr Zeit für diese Gruppe aufwenden zu können.

Vor diesem Hintergrund und in dem Bestreben, auch in Zukunft eine geeignete Einrichtung für die im Zusammenhang mit Forschungsunterstützung und Digitalisierungsoffensive ste-

henden Herausforderungen zu sein, hat die Bibliothek im Jahr 2021 einen Umstrukturierungsprozess angestoßen.[3]

Der Umstrukturierungsprozess erstreckte sich über zwei Jahre von 2021 bis 2023 (zeitweise verhindert oder verzögert während der Corona-Pandemie) und berührte viele Aspekte der Bibliotheksorganisation. Für die Fachreferenten brachte die Umstrukturierung jedoch drei wichtige Ergebnisse: die Konsolidierung der Arbeitsabläufe und Aufgaben in allen Zweigbibliotheken, die Entstehung von Teams (koordiniert durch die Abteilungen Medienbearbeitung und Benutzung). Infolgedessen könnte die Zeit der Fachreferent:innen weiter optimiert werden, sodass mehr Zeit zur Verfügung stünde, um auf die immer anspruchsvolleren Bedürfnisse bestimmter wissenschaftlicher Gemeinschaften einzugehen. Dadurch wurde eine besser als zuvor aufgestellte FDM-Arbeitsgruppe geschaffen, die auf verschiedenen Ebenen innerhalb der Universität tätig werden kann. Im Jahr 2024 haben die Abteilung Open Science und diese neue FDM-Task Force ein neues Arbeitsprogramm und einen neuen Arbeitsplan aufgestellt.

3 Ergebnisse im Jahr 2024

Die Aktivitäten im Jahr 2024 konzentrierten sich auf die Organisation der Arbeit der beiden oben genannten forschungsnahen Dienste, der Abteilung Open Science und des Fachreferats.

Dies geschah über eine Reihe von Treffen sowie dem Aufbau einer Wissensbasis, in der wir unsere Arbeitsmaterialien sammeln konnten. Darüber hinaus umfasste die Arbeit im Jahr 2024 auch die Umstrukturierung unserer Lehrmaterialien und -kurse, um ein neues FDM-Curriculum zu schaffen, das dynamisch an die Bedürfnisse unserer verschiedenen Forschungsgemeinschaften angepasst werden kann.

Strategisch war es uns wichtig, eine verbesserte Interaktion mit unseren wissenschaftlichen Communities zu erreichen, was wir durch gezielte Platzierung von Informationen zu den Angeboten der Bibliothek in Seminaren, Forschungskolloquien und Arbeitstreffen versuchten umzusetzen. Dabei konnten wir auf die bereits etablierten Verbindungen zurückgreifen, die unsere Fachreferent:innen im Laufe der Jahre zu Wissenschaftler:innen im Zusammenhang mit anderen bibliothekarischen Aufgaben aufgebaut hatten.

Nicht zuletzt wurden die Kolleg:innen in die Beratungs- und FDM-Orientierungsrunden, die routinemäßig von der Abteilung Open Science durchgeführt werden, eingebunden und zur Teilnahme eingeladen.

3 Weitere Details zu diesem Umstrukturierungsprozess können dem Jahresbericht der Universitäts- und Landesbibliothek Sachsen-Anhalt 2021 entnommen werden (Berghaus-Sprengel u. a. 2022).

Drei Aspekte wurden in der Präsentation näher erläutert:

Informationsblätter: Die Erstellung dieser Vorlagen hat uns die Möglichkeit gegeben, die sehr unterschiedlichen Informationen, die von den Fachreferent:innen gesammelt werden, in einer Vorlage zu vereinheitlichen, die im Quervergleich überprüft werden kann, um spezifische Informationen über die Entwicklungen in den verschiedenen Disziplinen zu erhalten. Die Erstellung dieser Vorlagen wird fortgesetzt; derzeit liegen zwölf Vorlagen vor, die gepflegt und mit relevanten Informationen ergänzt werden. Wir planen, die Anzahl der Modelle in diesem Jahr zu erhöhen.

FDM-Curriculum: Durch die Umstrukturierung unserer Kurse, Seminare und Informationsveranstaltungen in ein einziges Curriculum sind wir nun in der Lage, maßgeschneiderte Veranstaltungen für bestimmte Forschergruppen oder Communities sehr kurzfristig und in der gewünschten Tiefe und Sprache (Englisch/Deutsch) anzubieten. Diese Kurse können an verschiedene Formate angepasst werden und bilden die Grundlage für die Zusammenarbeit mit spezialisierten Gruppen, wie z. B. FDM-Trainer:innen aus bestimmten NFDI Konsortien oder spezialisierten Institutionen. Ein erfolgreicher Proof of Concept, das diesen Ansatz bestätigte, war die Organisation des fachspezifischen Workshops mit den Kollegen von FAIRAgro im März letzten Jahres.

Entwicklung von FDM-Kompetenzen: Ein weiteres zentrales Ergebnis des vergangenen Jahres war eine Schulung, die wir für unsere Fachreferent:innen zur Durchführung von sog. *data talks* oder „Datengesprächen“ durchgeführt haben. Sowohl das Feedback, das wir Anfang letzten Jahres von unseren Kolleg:innen erhalten haben, als auch unsere eigenen Erfahrungen zeigen, dass die Fähigkeit, ein solches Datengespräch mit „Datenproduzent:innen zu den Besonderheiten und Anforderungen eines bestimmten Forschungsprojekts“ zu führen, gut vermittelt und gestärkt werden muss.

Verständlicherweise ist dies der Aspekt, in dem unsere Kolleg:innen am wenigsten Erfahrung hatten. Um diese Wissens- und Erfahrungslücke zu schließen, haben wir unsere Fachreferent:innen exemplarisch und praxisnah zu den wichtigsten Aspekte eines solchen Datengesprächs geschult. Die gewonnen Erkenntnisse konnten die Kolleg:innen durch die hospitierende Teilnahme an echten Datengesprächen, die von der OS-Abteilung im Jahr 2024 durchgeführt wurden, vertiefen.

4 Pläne und Aktivitäten für 2025

Aufgrund unserer Erfahrungen und Ergebnisse im Jahr 2024 haben wir festgestellt, dass wir einige Anpassungen und Optimierungen vornehmen sollten.

Die erste Erkenntnis ist, dass die von unserem Fachreferent:innen zusammengestellte Task Force sehr heterogen ist. Das liegt natürlich an den unterschiedlichen Profilen, Ausbildungswegen und fachlichen Hintergründen, aber auch an weiteren Faktoren wie der tatsächlich verfügbaren zusätzlichen Zeit in den Zweigbibliotheken. Dies war letztlich zu erwarten, da der Grad der Fachkenntnisse sowie die Bereitschaft, sich vertieft in das Feld des Forschungsdatenmanagements einzuarbeiten, nicht nur von persönlichen Variablen abhängen, sondern auch vom antizipierten Interesse potenzieller Datenproduzent:innen, das je nach fachdisziplinärer Community sehr unterschiedlich ausgeprägt sein kann: Einige Forschungsgemeinschaften sind weitaus aktiver als andere, und dies wirkt sich auch auf das Interesse und die Aufgeschlossenheit gegenüber FDM-Themen aus. Vor diesem Hintergrund haben wir beschlossen, unsere Task-Force zu optimieren und sie in zwei Gruppen aufzuteilen.

Die eine Gruppe ist für allgemeine Anfragen zuständig und bietet Einsteigerkurse oder Informationsveranstaltungen an, um die FDM-Unterstützung und -Dienstleistungen der Bibliothek vorzustellen. Die andere Gruppe ist so konfiguriert, dass sie spezialisierten Communities, die eine fachliche Anleitung mit spezifischen Werkzeugen benötigen, eine tiefergehende Unterstützung bietet oder ein disziplinspezifisches Seminar durchführt; auch eine Zusammenarbeit mit relevanten NFDI- oder anderen Forschungseinrichtungen in Deutschland ist möglich. Mit der letztgenannten Gruppe werden wir in diesem Jahr einige Weiterbildungsaktivitäten durchführen.

4.1 Bewertung der Kompetenz

Wir sind sehr daran interessiert, einen Einblick in das Niveau der FDM-Expertise und -Kompetenz unserer Kollegen in dieser zweiten Gruppe zu erhalten. Zu diesem Zweck werden wir mit Hilfe der verfügbaren Instrumente eine Erhebungs- und Bewertungsphase einleiten, in der wir Daten sammeln werden, die einige quantitative Antworten auf diese Frage liefern können.

Anhand der neuen Version 3.0 der Lernzielmatrix-Materialien (Petersen u. a. 2022) werden wir versuchen, die Kompetenz unserer Kollegen mit den Erwartungen abzugleichen, die an Gruppen wie Doktoranden, aber vor allem an Data Stewards gestellt werden. Wir gehen davon aus, dass wir Ergebnisse erhalten werden, bei denen unsere Fachreferent:innen im Allgemeinen etwas schlechter abschneiden als die Data Stewards, dass aber in einigen spezifischen Bereichen die Kompetenz unserer Kolleg:innen höher sein könnte. Dies könnte z. B. in Bereichen wie Metadatenstandards und persistente Identifikatoren der Fall sein.

Letztlich geht es darum, die Wissenslücken zwischen diesen Gruppen zu ermitteln und sicherzustellen, dass diese Lücken durch gegenseitigen Austausch und Schulungen geschlossen werden können. Die Identifizierung und angemessene Einbindung verschiedener Akteure, die Aspekte des Forschungsdatenmanagements an den Universitäten unterstützen, ist ein entscheidender Prozess, um sicherzustellen, dass die Universitäten ihre Humanressourcen effizient verwalten (Meyer-Doerpinghaus und Neuroth 2015).

4.2 Erstellung geeigneter Definitionen für FDM-Akteure

Die folgenden Definitionen von Daten Stewards, Data Scientists und Fachreferent:innen wurden während des Vortrags vorgestellt:

Data Stewards: Data Stewards sind Personen innerhalb einer Institution, die für die Qualität der Daten und Datenquellen verantwortlich sind. In einer wissenschaftlichen Einrichtung unterstützen sie das Forschungsdatenmanagement und reproduzierbare Datenworkflows in Forschungsgruppen, Instituten oder Abteilungen.

Data Scientists: Data Scientists durchforsten und interpretieren große Datenmengen, um daraus nützliche Informationen als Grundlage für strategische Entscheidungen abzuleiten. Zu den Hauptaufgaben von Data Scientists gehören die Datenanalyse, die Modellentwicklung, die Datenbereinigung, die Datenvisualisierung und die Dateninterpretation.

Fachreferat: Diese Gruppe von Informationsspezialisten verfügt über gute disziplinspezifische Kenntnisse und steht im ständigen Dialog mit dem wissenschaftlichen Personal der Fakultäten, was sie in die Lage versetzt, als eine Art Schnittstelle zwischen den Dienststellen und den Wissenschaftler:innen einer Organisation zu spielen.

Es wurde betont, dass es sich bei den Definitionen lediglich um Vorschläge handelt, die sich aus unseren Diskussionen in der Bibliothek und an unserer Universität ergeben haben. Diese Definitionen müssen natürlich verbessert, optimiert und mit anderen in der Literatur verfügbaren Erkenntnissen und Definitionen abgeglichen werden (siehe z. B. Verhulst u. a. 2020; Shutsko und Lindstädt 2020).

Als Feedback während der Konferenz wurde die Vorstellung von Data Stewards, die für die Qualität der Daten verantwortlich sind, mit dem Hinweis in Frage gestellt, dass die Verantwortung eher bei den Wissenschaftlern liegen sollte. Es muss also eindeutig dafür gesorgt werden, dass geeignete Definitionen verfügbar sind. Es wurde auch darauf hingewiesen, dass die Definitionen möglicherweise auf die speziellen Bedürfnisse und Gegebenheiten der einzelnen akademischen Einrichtungen zugeschnitten werden müssen.

Nichtsdestotrotz ist die Gewährleistung einer guten Zusammenarbeit zwischen Data Stewards, Data Scientists und Fachreferent:innen von entscheidender Bedeutung für unsere Einrichtung, um sicherzustellen, dass diese wertvollen Humanressourcen optimal genutzt und eingesetzt werden können. Um dieses Ziel zu erreichen, werden wir uns auf die Ergebnisse von Projekten wie DateStew (Seidlmayer u. a. 2023) stützen.

Was die MLU betrifft, so bemühen wir uns zunächst darum, das Profil aller verschiedenen FDM-Akteure in unserer Einrichtung zu ermitteln und zu verstehen, um dann mögliche Lücken in der Abdeckung bestimmter Kompetenzbereiche zu identifizieren, die dann angegangen werden sollten. Auf diese Weise wollen wir sicherstellen, dass unsere Fachreferent:innen in den Bereichen eingesetzt werden, in denen sie am meisten gebraucht werden, und so dafür sorgen, dass es keine Doppelarbeit mit anderen Gruppen wie Data Stewards und Data Scientists an der Universität gibt.

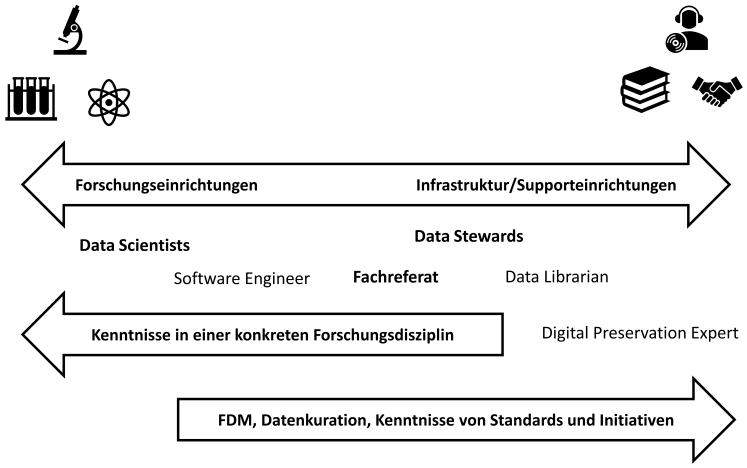


Abbildung 2: Mögliche Einordnung der Fachreferent:innen in die Forschungslandschaft der MLU: Die angemessene Platzierung von Data Stewards, Data Scientists und Fachreferent:innen wird je nach den Gegebenheiten der einzelnen Einrichtungen unterschiedlich aussehen. Doch sollte diese Frage gründlich bedacht werden.

Unser Ziel ist es, eine fruchtbare Zusammenarbeit mit diesen Gruppen zu erreichen, aber auch gute Diskussionen und Planungen mit den Universitätsbehörden und den Principal Investigators (PIs) unserer laufenden und zukünftigen SFBs und Forschungsprojekte an der Universität zu führen, um eine gute Platzierung und Koordination dieser Akteure in der Universitätslandschaft sicherzustellen.

Wir hoffen, dass diese Diskussionen zu einem nützlichen Verständnis und einer Einigung über den Bereich der Aktivitäten, Aufgaben und Verantwortlichkeiten der drei in diesem Dokument besprochenen Gruppen (Fachreferent:innen, Data Stewards und Data Scientists) führen werden. Darüber hinaus erwarten wir auch positive Auswirkungen auf die Arbeitsabläufe innerhalb der Universität und auf die Zusammenarbeit mit regionalen, nationalen und internationalen Partnern sowie gute Synergien mit einer Reihe von NFDI-

Initiativen, an denen einige wissenschaftliche Gemeinschaften unserer Einrichtung bereits aktiv beteiligt sind.

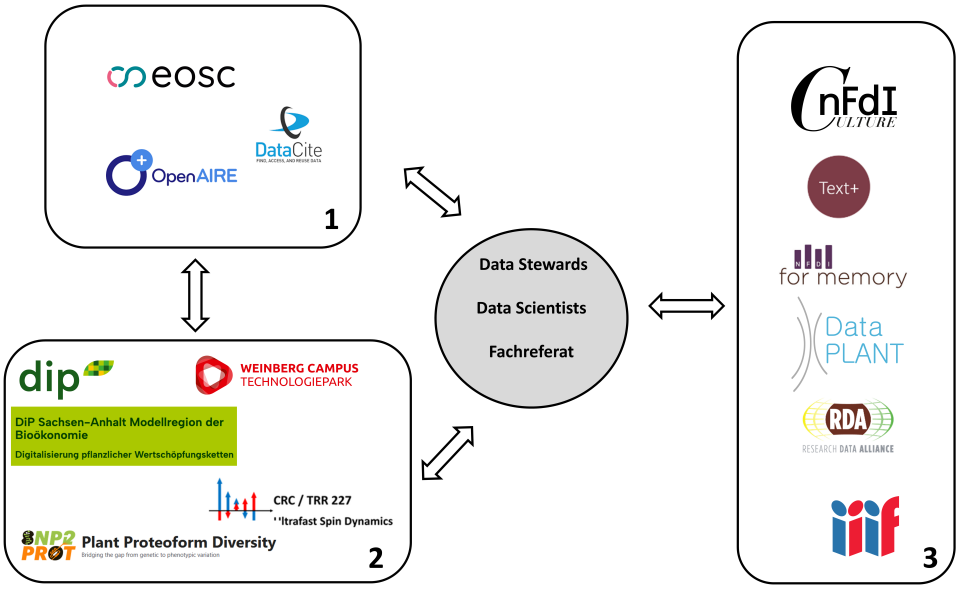


Abbildung 3: Kooperationen mit lokalen und nationalen Projekten und Initiativen: Die drei in diesem Dokument behandelten Berufsgruppen sollen, sobald ihre Rollen und Aufgabenbereiche richtig festgelegt sind, einen positiven Beitrag zur Unterstützung und Förderung der Zusammenarbeit und der Synergien zwischen lokalen Projekten sowie regionalen und nationalen Infrastrukturen leisten.

5 Fazit

Der Umstrukturierungsprozess an der ULB Sachsen-Anhalt war ein komplexer Prozess, der einige Jahre benötigte, der aber erste Ergebnisse hervorgebracht hat, von denen einige Bereiche bereits profitieren können. Dies gilt für die forschungsnahen Bereiche der Bibliothek, in denen für einige Fachreferent:innen mehr Zeit zur Verfügung steht, um sich den FDM-Unterstützungsaufgaben und Schulungen zu widmen. Ob sich diese zusätzliche Zeit bereits auszahlt, hängt natürlich sehr stark von der individuellen Situation unserer Kolleg:innen ab; das Fachgebiet, der persönliche Lebenslauf und die Bildungsbiographie spielen eine entscheidende Rolle für die individuelle Fähigkeit zur Wahrnehmung von FDM-Aufgaben. Der bloße Zeitgewinn an sich reicht natürlich nicht aus, wenn er nicht durch andere Maßnahmen ergänzt wird.

Der Schwerpunkt für die weitere Arbeit mit der Gruppe wird darin liegen, sicherzustellen, dass angemessene und spezifische Schulungsangebote für sie verfügbar sind. Auf organisatorischer Ebene wird es von grundlegender Bedeutung sein, dass ein ausreichender Austausch mit den Universitätsbehörden, den PIs und anderen Akteuren, die für die

Rekrutierung und Vermittlung von FDM-Support-Personal verantwortlich sind, gewährleistet ist.

Danksagungen

Dr. Silke Berndsen hat einen ersten Entwurf des Manuskripts gelesen und wesentlich verbessert. Der Autor profitierte von früheren Diskussionen über das Thema mit anderen Kolleg:innen in der Bibliothek, darunter Dr. Susann Özüyaman und Dr. Volker Adam.

Beiträge

Dr. Roberto Cozatl ist für die Konzeption der in diesem Beitrag vorgestellten Arbeit sowie für die Vorbereitung und Erstellung des Manuskripts verantwortlich.

Interessenkonflikte

Es sind keine Interessenkonflikte bekannt.

Literaturverzeichnis

Berghaus-Sprengel, Anke, Susann Özüyaman, Ellen Reihl, Daniel Horn, Roberto Cozatl, Thomas Hoffmann, Julia Knödler u. a. 2022. *Jahresbericht 2021.* https://doi.org/10.25673/92261.

Helling, Patrick, Felix Rau, Monika Linne, Jens Dierkes, Mijke Jetten, Anna Walek und Magdalena Szuflita-Żurawska. 2022. *The Importance of Demand and Environment for Defining and Establishing the Role of Data Stewards.* Zenodo. https://doi.org/10.5281/ZENODO.6511185.

Meyer-Doerpinghaus, Ulrich und Heike Neuroth. 2015. „Die Stärkung von Informationskompetenz im Kontext des Forschungsdatenmanagements: eine Herausforderung für Hochschulen und Politik“. *Zeitschrift für Bibliothekswesen und Bibliographie* 62 (2): 80–84. https://doi.org/10.3196/186429501562237.

Petersen, Britta, Claudia Engelhardt, Tanja Hörner, Juliane Jacob, Tatiana Kvetnaya, Andreas Mühlichen, Hermann Schranzhofer u. a. 2022. *Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards.* Zenodo. https://doi.org/10.5281/ZENODO.7034478.

Seidlmayer, Eva, Fabian Hoffmann, Jens Dierkes, Birte Lindstädt, Ralf Depping und Konrad Ulrich Förstner. 2023. „Forschung unterstützen – Empfehlungen für Data Stewardship an akademischen Forschungsinstitutionen". In *Ergebnisse des Projektes DataStew.* PUBLISSO. https://doi.org/10.4126/FRL01-006441397.

Shutsko, Aliaksandra und Birte Lindstädt. 2020. „Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten – NFDI4Health: Pilotprojekt zu Bibliotheken und Forschungsdatenkompetenzzentren als Multiplikatoren („Data Steward")". *GMS Medizin – Bibliothek – Information; 20(3):Doc27,* https://doi.org/10.3205/MBI000484.

Verhulst, Stefan G., Andrew J. Zahuranec, Andrew Young und Michelle Winowatan. 2020. *Wanted: Data Stewards – (Re-)defining the Roles and Responsibilities of Data Stewards for an Age of Data Collaboration.* Besucht am 28. Mai 2025. https://www.thegovlab.org/static/files/publications/wanted-data-stewards.pdf.

ROADS to Reusable Datasets: Potenziale und Herausforderungen bei der Datennachnutzung in der Hochschullehre

Anna Daudrich [1]

[1] ZHAW Zürcher Hochschule für Angewandte Wissenschaften, Schweiz

Die Wiederverwendung von Forschungsdaten bietet die Möglichkeit, wissenschaftliche Ressourcen langfristig effizienter und nachhaltiger zu nutzen. In der Praxis stehen diesem Potenzial jedoch zahlreiche Hürden entgegen. Der vorliegende Beitrag untersucht diese Herausforderungen und formuliert die Hypothese: Um die Datennachnutzung erfolgreich zu fördern, sollte deren Integration bereits im Studium gezielt unterstützt werden. Frühzeitige Erfahrungen mit der Nachnutzung von Forschungsdaten ermöglichen es Studierenden, zentrale Kompetenzen in den Bereichen Data Literacy und Forschungsdatenmanagement zu entwickeln – und schaffen somit die Grundlage für eine nachhaltige und effektive Nutzung von Daten in der späteren wissenschaftlichen Praxis.

Diese Hypothese stützt sich auf Erkenntnisse aus dem Projekt ROADS (Reusing Openly Accessible Research Data for Student Theses) an der ZHAW Zürcher Hochschule für Angewandte Wissenschaften. Eines der Projektziele ist es, Chancen und Herausforderungen der Datennachnutzung zu identifizieren – basierend auf Anwendungsfällen, in denen Studierende aus verschiedenen Disziplinen bestehende Forschungsdaten nachnutzten. Eine der herausfordernden Aufgaben für Studierende war die Datenrecherche und -interpretation – was auf die mangelnde Datenkompetenz zurückzuführen ist. Die Vielzahl an Datenquellen, die heterogene Datenlandschaft sowie die häufig unzureichende Dokumentation veröffentlichter Datensätze erschweren den Prozess der Datensuche. Die Bewältigung dieser Herausforderungen in den studentischen Anwendungsfällen erforderte einen hohen Zeit- und Arbeitsaufwand.

Trotz dieses Aufwands zeigt sich: Die Datennachnutzung bietet einen erheblichen Mehrwert. Wird sie systematisch in die Lehre integriert, können Studierende (und teilweise auch ihre Betreuenden) essenzielle Kompetenzen im Umgang mit Daten erwerben und

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23923 (CC BY-SA 4.0).

ein Bewusstsein für effizientes Forschungsdatenmanagement entwickeln. Langfristig kann dies zu ressourcenschonenderen und effizienteren Forschungspraktiken beitragen.

Keywords: Datennachnutzung, Datenkompetenz, Datenqualität, Forschungsdaten, FAIR

1 Einleitend: Ressourcenschonung durch Datennachnutzung – Chance oder Herausforderung?

Nachhaltigkeit und ein effizienter Umgang mit Ressourcen gewinnen zunehmend an Bedeutung. Auch im Forschungskontext ist dieses Thema präsent. Zahlreiche Hochschulen und Forschungsinstitutionen haben Nachhaltigkeitsstrategien inzwischen fest verankert.[1] Ein sorgfältig geplantes Forschungsdatenmanagement (FDM) sowie Open Research Data (ORD) können maßgeblich dazu beitragen, ressourcenschonende Forschungspraktiken zu fördern – ein Aspekt, der auch in den Richtlinien der Forschungsförderer häufig hervorgehoben wird.[2]

Insbesondere die Nachnutzung bereits vorhandener Daten birgt großes Potenzial: Anstatt zeit- und kostenintensiv neue Daten zu erheben, können Forschende auf bestehende Datensätze zurückgreifen. Dies soll helfen, Doppelerhebungen von Daten sowie redundante wissenschaftliche Untersuchungen zu vermeiden (Hollstein und Strübing 2018). Gerade bei Langzeitstudien oder bei schwer erreichbaren Personengruppen erweist sich diese Praxis als besonders vorteilhaft. Die Nachnutzung von Daten aus quantitativen Befragungen oder qualitativen Interviews minimiert zudem datenschutzrechtliche Herausforderungen, da bereits anonymisierte Daten verwendet werden können und keine neuen Einwilligungen von Befragten erforderlich sind (Netscher und Trixa 2019).[3] Darüber hinaus entfallen Auf-

1 Siehe dazu beispielsweise die Nachhaltigkeitsstrategien der folgenden Hochschulen: die ZHAW Zürcher Hochschule für Angewandte Wissenschaften (https://www.zhaw.ch/de/fokusthemen/zhaw-sustainable/nachhaltigkeitsstrategie; *Besucht am 30. März 2025*), die Universität Zürich (https://www.sustainability.uzh.ch/de/policy-reporting.html; *Besucht am 30. März 2025*) sowie die Friedrich-Schiller-Universität Jena (https://www.uni-jena.de/207933/nachhaltigkeitsstrategie; *Besucht am 30. März 2025*).

2 Siehe dazu beispielsweise die Nationale Schweizer Strategie für Open Research Data (https://www.swissuniversities.ch/themen/open-science/open-research-data/nationale-strategie; *Besucht am 30. März 2025*) sowie Grundlagen zur nachhaltigen Forschung bei der Deutschen Forschungsgemeinschaft (https://www.dfg.de/de/grundlagen-themen/entwicklungen-im-wissenschaftssystem/nachhaltigkeit-im-forschungsprozess; *Besucht am 30. März 2025*)

3 Anonymisierte Daten gelten nicht mehr als personenbezogene Daten. Sie unterliegen daher nicht den Bestimmungen des Datenschutzrechts. Das bedeutet, dass sie grundsätzlich für andere Zwecke verwendet werden dürfen – vorausgesetzt, eine Re-Identifizierung ist ausgeschlossen, und in der informierten Einwilligung (Informed Consent) wurde über eine mögliche Weiternutzung und/oder Weitergabe der Daten informiert. Zu beachten ist weiter, dass einer Weiternutzung und / oder Weitergabe die Ver-

gaben wie die Konzeption von Interviewleitfäden oder die Akquise von befragten Personen (Bendel 2021).

Obgleich die Nachnutzung von Forschungsdaten Chancen für eine ressourcenschonende Forschung bietet, stehen ihrer umfassenden Ausschöpfung zahlreiche Hürden entgegen. Im Folgenden werden diese Herausforderungen näher beleuchtet. Die Darstellung stützt sich auf Ergebnisse des Projekts ROADS („Reusing Openly Accessible Research Data for Student Theses“) an der ZHAW Zürcher Hochschule für Angewandte Wissenschaften.[4] Zudem werden zentrale Begrifflichkeiten im Zusammenhang mit der Datennachnutzung erläutert – insbesondere im Hinblick auf das ‚R‘ (Reusable) der FAIR-Prinzipien[5]. Hierbei wird der Beitrag aufzeigen, unter welchen Bedingungen Daten als nachnutzbar gelten. Es ist dabei wichtig zu betonen, dass sich die Schlussfolgerungen dieses Beitrags auf Fallbeispiele stützen und daher nicht als allgemeingültige Aussagen, sondern als empirisch fundierte Hypothesen verstanden werden sollten.

2 Fallbeispiel: Projekt ROADS

2.1 Ausgangslage

Das Projekt ROADS wurde an der ZHAW Zürcher Hochschule für Angewandte Wissenschaften ins Leben gerufen, vor dem Hintergrund, dass in der Forschung inzwischen eine hohe Vielzahl an Datensätzen generiert wurde – sei es in Form von Statistiken, Laborexperimenten, Bildern, Transkripten von Interviews usw. Zahlreiche dieser Datensätze stehen bereits den Forschungscommunities und teils auch der breiten Öffentlichkeit zur Verfügung, da sie etwa in generischen, fachspezifischen oder institutionellen Repositorien als ORD veröffentlicht sind. Damit das Potenzial offener Daten voll ausgeschöpft werden kann, ist es entscheidend, dass sie die FAIR-Kriterien erfüllen, also auffindbar, zugänglich, interoperabel und nachnutzbar sind.[6]

Trotz zahlreicher Initiativen zur Förderung von ORD werden weiterhin nicht alle Datensätze veröffentlicht und verbleiben oft innerhalb der Institutionen, in denen sie generiert wurden.[7] Diese Datensätze erfüllen zwar nicht alle Kriterien der FAIR-Prinzipien – sie

traulichkeit der Daten oder vertraglichen Regelungen entgegenstehen können (siehe dazu Art. 2 Abs. 1 DSG sowie Art. 5 lit. a DSG).

4 Projektlaufzeit: Dezember 2023 – Juni 2025, Co-Finanzierung: swissuniversities. Nähere Informationen zum Projekt: https://www.zhaw.ch/de/forschung/projekt/74318; *Besucht am 30. März 2025.*

5 Nähere Informationen zu den FAIR-Prinzipien: https://www.go-fair.org/fair-principles/; *Besucht am 30. März 2025.*

6 https://www.unesco.org/en/open-solutions/open-data; *Besucht am 30. März 2025.*

7 Dies bestätigen unter anderem die Ergebnisse der Umfrage des Schweizerischen Nationalfonds (SNF): https://www.snf.ch/de/iyEnUsdV0AiHk9wF/news/open-research-data-ein-erster-blick-auf-die-aktuelle-praxis; *Besucht am 30. März 2025.*

sind beispielsweise weder durch Suchmaschinen auffindbar noch für die (breite) Öffentlichkeit zugänglich –, dennoch besitzen sie weiterhin Potenzial zur Nachnutzung. Insbesondere die Datenautor:innen und andere Beteiligte an den Projekten, in deren Rahmen die Daten ursprünglich erhoben wurden, haben grundsätzlich die Möglichkeit, auf solche unveröffentlichten Datensätze zurückzugreifen und sie in neuen Forschungskontexten wiederzuverwenden.

Daraus ergibt sich die zentrale Frage: Sind all diese (öffentlich) verfügbaren Datensätze tatsächlich nachnutzbar, und was bedeutet „nachnutzbar“ überhaupt in diesem Kontext? Diese Überlegungen bildeten die Ausgangslage für das Projekt ROADS. Die Nachnutzbarkeit von Daten sowie die damit verbundenen Chancen und Herausforderungen wurden anhand konkreter Anwendungsfälle untersucht. Studierende der ZHAW auf verschiedenen Stufen (Bachelor, Master, Doktorat) und aus unterschiedlichen Disziplinen nutzten dabei bereits vorhandene Datensätze für ihre Abschlussarbeiten, Seminararbeiten und weitere studentische Projekte. Die Datennachnutzung erfolgte unter der Betreuung von Dozierenden aus verschiedenen ZHAW-Departementen. Zusätzlich wurde dieser Prozess durch Data Stewards aus der Hochschulbibliothek und den ZHAW Services Research Data unterstützt.

In jedem Anwendungsfall konnten die Studierenden und ihre Betreuenden selbst entscheiden, welche Daten genau und in welchem Kontext nachzunutzen sind. Zur Auswahl standen dabei vor allem offene Forschungsdaten. Aber auch unveröffentlichte Daten, etwa aus abgeschlossenen oder laufenden Forschungsprojekten der Betreuenden, boten sich grundsätzlich für eine Nachnutzung an. Da die Betreuenden in der Regel Zugang zu solchen Daten haben, konnten sie diese ihren Studierenden zu einer weiteren Auswertung zur Verfügung stellen – vorausgesetzt, dies war rechtlich zulässig. Tabelle 1 auf Seite 175 zeigt eine Übersicht der Daten, die zur Nachnutzung ausgewählt wurden.

Wie die Tabelle 1 zeigt, handelt es sich in den meisten Anwendungsfällen um nicht publizierte quantitative Daten. Die Entscheidung zugunsten quantitativer statt qualitativer Daten ist in diesem Fall wenig überraschend. Qualitative Daten sind in der Regel sehr spezifisch. Ihre Erhebung wird gezielt an den jeweiligen Forschungsgegenstand angepasst, so dass sie mit ihrem sozialen sowie kontextuellen Entstehungsrahmen eng verbunden sind (Gebel und Meyermann 2021). Die Entscheidung für nicht veröffentlichte Daten scheint jedoch im Widerspruch zu den Zielen von ORD und den FAIR-Prinzipien zu stehen: Wenn offene Forschungsdaten zugänglich und nachnutzbar sein sollen, warum werden dann stattdessen nicht veröffentlichte Daten wiederverwendet?

Diese und weitere Erkenntnisse aus den Anwendungsfällen im Rahmen von ROADS werden im Folgenden dargestellt. Dabei werden diese sowohl aus der Perspektive der Studierenden als auch der Betreuenden erläutert. Diese Erkenntnisse wurden auf unterschiedlichen Wegen erfasst: Die Studierenden reflektierten die Nachnutzung im jeweiligen Anwendungsfall durch schriftliche Ausführungen, die Teilnahme an Online-Umfragen sowie an

qualitativen Interviews. Die Betreuenden trugen ihre Reflexionen zur Datennachnutzung im Rahmen regelmäßiger Austauschgespräche und Projekt-Workshops bei.

Tabelle 1: Datennachnutzung in studentischen Arbeiten im Rahmen des Projekts ROADS.

Forschungsbereich	**Datenarten**	**Verfügbarkeit**	**Kontext der Nachnutzung**
Physiotherapie	Quantitative Bewegungsdaten, Qualitative Interviews	unveröffentlicht, veröffentlicht	Bachelor- & Masterarbeiten
Datenjournalismus	Quantitative Umfragen	unveröffentlicht, veröffentlicht	Seminararbeiten
Psychologie	Quantitative Umfragen	unveröffentlicht	Bachelor- & Masterarbeiten
Computergestützte Lebenswissen.	Quantitative Lebensmittel- & Gesundheitsdaten	unveröffentlicht, veröffentlicht	Seminararbeiten
Materialwissen.	Quantitative Daten zu Perowskit-Solarzellen	veröffentlicht	wissenschaftliche Veröffentlichung

2.2 Perspektive der Studierenden

Die Studierenden berichteten mehrheitlich, entweder keine oder nur geringe Erfahrung in der Nachnutzung von Daten zu haben. Ebenso war vielen – insbesondere auf der Bachelor-Stufe – die Erhebung eigener Daten noch nicht vertraut, wobei diese insgesamt häufiger praktiziert wurde als die Datennachnutzung. Zudem zeigten sich bei einigen Studierenden, vor allem zu Beginn des Bachelorstudiums, nicht nur Defizite im Umgang mit Daten, sondern auch Lücken in grundlegenden wissenschaftlichen Arbeitsmethoden. Aufgrund dieser mangelnden Erfahrung traten bereits zu Beginn der Anwendungsfälle zahlreiche Herausforderungen auf.

Ein zentrales Problemfeld stellte die Recherche und das Finden eines geeigneten Datensatzes dar. Viele Studierende verfügten über keine oder nur geringe Kenntnisse darüber, wo und wie sie nach relevanten Datensätzen suchen sollten. Diese Schwierigkeiten resultierten nicht nur aus mangelnder Erfahrung, sondern auch aus dem Fehlen spezifischer Schulungen zur Datenrecherche. Während in zahlreichen Studiengängen Kompetenzen in der Literaturrecherche – häufig vermittelt durch Bibliothekspersonal[8] – gefördert werden,

8 Siehe beispielsweise das Angebot der ZHAW Hochschulbibliothek zur Rechercheberatung: https://www.zhaw.ch/de/hochschulbibliothek/recherchehilfe-kurse#c5338, *Besucht am 31.03.2025.*

existieren vergleichbare Angebote für die Datenrecherche in der Regel nicht. Dies führte dazu, dass Datensätze oft zufällig, unsystematisch oder mit wenig zielführenden Suchstrategien identifiziert wurden, was die Eignung der Daten für die jeweilige Forschungsfrage erheblich einschränkte.

Eine zusätzliche Herausforderung ergab sich aus der häufig fehlenden Kompetenz im Umgang mit Software-Programmen, die für die Bearbeitung und Analyse der Daten erforderlich sind. Selbst bei gängiger Software wie Excel mussten die Studierenden oft zusätzliche Zeit investieren, um sich die notwendigen Kenntnisse anzueignen.

Neben unzureichenden Kompetenzen in der Datenrecherche und in der Bedienung von Software stellte die Qualität der veröffentlichten Datensätze eine weitere erhebliche Herausforderung dar. Häufig fehlten essenzielle Informationen in der Dokumentation, beispielsweise Erläuterungen zu zentralen Variablen oder eine detaillierte Beschreibung der Laborbedingungen, die für das Verständnis und die Interpretation der Daten erforderlich sind. Diese Dokumentationslücken erschwerten es den Studierenden erheblich, die Daten korrekt zu kontextualisieren und ihre Relevanz für die jeweilige Forschungsfrage adäquat zu bewerten.

Ein weiteres Hindernis war die mangelnde Standardisierung in der Datenerhebung. So werden Messdaten oft unter unterschiedlichen Laborbedingungen oder ohne die Anwendung standardisierter Messverfahren erfasst, was die Zusammenführung und den Vergleich verschiedener Datensätze erschwert. Ohne ein solides Verständnis wissenschaftlicher Standards fiel es vielen Studierenden schwer, diese Mängel zu erkennen und entsprechend kritisch zu bewerten.

Schließlich ist der Zugang zu publizierten Daten nicht immer uneingeschränkt gewährleistet. In einigen Fällen werden lediglich die Metadaten veröffentlicht, während der vollständige Zugriff auf die eigentlichen Daten erst nach der Kontaktaufnahme mit den Datenautor:innen gewährt wird – ein Prozess, der zu Verzögerungen führen kann. Gerade bei zeitlich beschränkten Vorhaben – etwa bei studentischen Arbeiten – kann dies problematisch sein, wenn eine zeitnahe Analyse der Daten erforderlich ist.

Angesichts dieser Herausforderungen sahen sich die Studierenden häufig gezwungen, nach alternativen Datensätzen zu suchen, selbst wenn der ursprünglich gefundene Datensatz auf den ersten Blick geeignet erschien. In einigen Fällen war es sogar notwendig, die ursprüngliche Forschungsfrage anzupassen oder sogar vollständig zu ändern, da keine geeigneten Datensätze für die ursprünglich formulierte Frage gefunden werden konnten. Aus diesen Gründen wurde die Recherche nach geeigneten Datensätzen von den Studierenden als eine der größten und zeitaufwendigsten Herausforderungen im Kontext der Datennachnutzung wahrgenommen.

2.3 Perspektive der Betreuenden

Da die Recherche nach geeigneten offenen Datensätzen für die Studierenden eine anspruchsvolle und zeitaufwändige Aufgabe darstellte, mussten die Betreuenden entsprechend reagieren, insbesondere um den vorgegebenen zeitlichen Rahmen für die studentischen Arbeiten einzuhalten. Eine pragmatische Lösung bestand darin, den Studierenden nicht veröffentlichte Daten aus ihren laufenden oder abgeschlossenen Forschungsprojekten für die Anwendungsfälle bereitzustellen. In diesem Fall waren die Daten bereits verfügbar, sodass weder zusätzlicher Aufwand für die Suche noch für die Interpretation erforderlich war. Ein weiterer Vorteil war, dass die Betreuenden in der Regel selbst in die Datenerhebung involviert waren und somit als Datenautor:innen bei Unklarheiten oder fehlender Information in der Dokumentation unmittelbar zur Verfügung standen. Dies führte zwar zu einer Zeitersparnis für die Studierenden, brachte jedoch zugleich neue Herausforderungen für die Betreuenden mit sich.

Ein zentraler Aspekt bestand in der Prüfung der rechtlichen Rahmenbedingungen bezüglich der Datennutzung. Insbesondere musste bei den personenbezogenen Daten sichergestellt werden, dass die Einwilligungen der Proband:innen klare und rechtlich gültige Regelungen dazu enthalten. In der Praxis zeigte sich jedoch, dass viele Einwilligungserklärungen unpräzise oder mehrdeutig formuliert sind. Beispielsweise wurde in einigen Fällen lediglich die Nachnutzung für „weitere Forschungszwecke“ erlaubt, ohne eine klare Definition dieses Begriffs. Dies war problematisch, da unklar blieb, ob Forschungszwecke mit Lehr- und Lernzwecken gleichzusetzen sind oder ob eine Masterarbeit als Forschungsarbeit gewertet werden kann. Die Klärung solcher Fragen erforderte zusätzlichen Aufwand, der nicht nur von den Betreuenden, sondern auch durch den Rechtsdienst der ZHAW getragen wurde, der in diesen Fällen konsultiert werden musste.

Zusätzlich stellte die technische Komplexität vieler Datensätze eine Herausforderung dar. Oft erforderten die bereitgestellten Daten spezifische Softwarekenntnisse für ihre Bearbeitung und Analyse, über die die Studierenden nicht immer verfügen. Da die Aneignung dieser Kenntnisse innerhalb des begrenzten Zeitrahmens einer studentischen Arbeit nicht realistisch war, bestand eine weitere Aufgabe der Betreuenden darin, die Daten in einfachere und zugänglichere Formate – beispielsweise als Excel-Tabellendokumente – zu überführen, um deren Nachnutzung zu erleichtern. Dieser Transformationsprozess erwies sich in einigen Fällen als besonders anspruchsvoll und erforderte zusätzlichen Ressourcenaufwand vonseiten der Betreuenden.

Auch wenn die rechtliche Nachnutzung von Daten letztendlich geklärt und die Daten in geeigneter Form den Studierenden bereitgestellt wurden, ergab sich ein weiteres Problem: die Bewertung der studentischen Leistung im Kontext der Datennachnutzung. Die Nachnutzung von nicht publizierten Daten aus Forschungsprojekten der Betreuenden führte zwar zu einer Reduktion des Zeit- und Arbeitsaufwands aufseiten der Studierenden im Vergleich zu einer eigenständigen Datenrecherche, da die Daten bereits vorlagen und deren Qualität sichergestellt war. Dennoch blieb unklar, welche Anforderungen die Betreuenden

in diesem Zusammenhang an die Studierenden stellen sollten – insbesondere im Hinblick auf die ECTS-Leistungspunkte. Für eine studentische Arbeit ist eine bestimmte Anzahl an ECTS vorgesehen, die einen entsprechenden Arbeitsaufwand widerspiegelt. In Fällen der Datennachnutzung entfiel jedoch teilweise ein erheblicher Teil dieses Aufwands, was zu einer potenziellen Ungleichbehandlung gegenüber Studierenden führte, die die benötigten Daten eigenständig recherchieren oder erheben mussten.

Nicht zuletzt stellte sich für die betreuenden Lehrpersonen die Frage, ab welchem Qualifikationsniveau – etwa bereits zu Beginn des Bachelorstudiums, in dessen fortgeschrittenem Verlauf oder erst auf Master- bzw. Promotionsniveau – die Datennachnutzung sinnvoll erfolgen kann. Ebenso war zu klären, unter welchem Maß an Komplexität die dafür vorgesehenen Daten sowie der jeweilige Kontext der Arbeiten ausgestaltet sein sollten, um eine effektive Wiederverwendung zu ermöglichen. Denn insbesondere in den frühen Phasen des Bachelor-Studiums mangelt es Studierenden sowohl an grundlegenden Datenkompetenzen als auch an Kenntnissen wissenschaftlicher Arbeitsmethoden.

Um den unterschiedlichen Kompetenzniveaus der Studierenden angemessen zu begegnen, wurden seitens der betreuenden Lehrpersonen unterschiedliche didaktische Ansätze verfolgt. Während die Nachnutzung von Daten bei fortgeschrittenen Studierenden primär im Rahmen von Abschlussarbeiten stattfand – wobei sowohl die Komplexität der Analyseverfahren als auch die Tiefe der Forschungsfragen insbesondere auf Master-Niveau gesteigert wurden – erfolgte sie in den frühen Phasen des Bachelorstudiums in einem vereinfachten Kontext, etwa im Rahmen einer Aufgabe mit dem Ziel, einen wissenschaftlich fundierten Blogbeitrag zu verfassen.

Solche komplexreduziertere Aufgabenformate zur Datennachnutzung ermöglichten zudem Gruppenarbeiten, in denen Studierende ihre unterschiedlichen Vorkenntnisse einbringen und etwaige Defizite im Bereich der Softwareanwendung oder methodischen Kompetenz wechselseitig ausgleichen konnten. Auch die eigenständige Recherche geeigneter Datensätze gestaltete sich im Rahmen solcher kooperativer Arbeitsformen häufig effizienter. Allerdings stellt sich in solchen Fällen die Frage, ob die von den Studierenden selbst recherchierten Datensätze einer zusätzlichen Überprüfung durch die betreuenden Lehrpersonen hinsichtlich ihrer Eignung für den jeweiligen Untersuchungskontext bedürfen. Dies erscheint insbesondere deshalb relevant, weil häufig unklar ist, inwieweit Studierende – insbesondere in den frühen Studienphasen – in der Lage sind, Datenquellen kritisch zu hinterfragen und die Qualität der Daten angemessen einzuschätzen.

Die genannten Herausforderungen unterstreichen die Notwendigkeit, standardisierte Leitlinien für die Nachnutzung von Daten im akademischen Lehr- und Lernkontext zu entwickeln und zu implementieren. Ohne derartige Vorgaben fehlen klar definierte Prozesse, die eine zielgruppenspezifische Ansprache sowie eine verlässliche Einschätzung des notwendigen Ressourcen- und Arbeitsaufwands ermöglichen.

3 Begriffsdefinitionen aus der Fallstudie: Nachnutzbarkeit, Datenqualität, Data Literacy

Die im vorherigen Abschnitt dargelegten Reflexionen von Studierenden und Betreuenden führen zu der Schlussfolgerung, dass die Nachnutzbarkeit von Daten im Wesentlichen von zwei Faktoren abhängt: der Datenqualität und der Datenkompetenz (siehe Abbildung 1). Diese beiden Faktoren sind eng miteinander verknüpft – ohne das Eine bleibt das Andere wirkungslos. Dies wurde unter anderem am Beispiel des „F“ in den FAIR-Prinzipien deutlich vor Augen geführt: Selbst wenn Daten durch Befolgen der guten FDM-Praxis bei der Dokumentation und Veröffentlichung technisch auffindbar sind, bleiben sie oft praktisch nicht auffindbar, wenn das Wissen über geeignete Suchstrategien und Quellen fehlt. Sind die Daten unauffindbar, können sie nicht nachgenutzt werden – auch wenn sie das Potenzial zur Wiederverwendung haben. Die Nachnutzbarkeit oder das „R“ in den FAIR-Prinzipien kann also nur gegeben sein, wenn die vorhergehenden drei Komponenten F-A-I allesamt erfüllt sind.

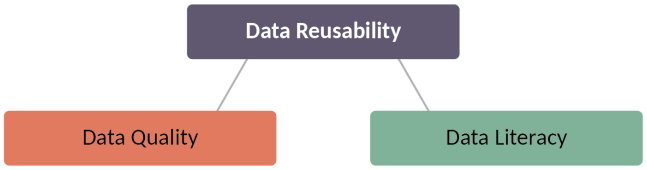


Abbildung 1: Was bedeutet die Nachnutzbarkeit von Daten?

Die Analyse der Anwendungsfälle im Projekt ROADS legt nahe, dass Datenqualität aus der Perspektive der Datennachnutzung ebenfalls als eine Zusammensetzung mehrerer Komponenten zu verstehen ist, nämlich aus inhaltlichen, technischen und rechtlichen Aspekten (Abbildung 2).[9] Inhaltliche Aspekte beziehen sich auf die gewährleistete Interpretierbarkeit der Daten, wobei der vorhandenen Dokumentation eine zentrale Bedeutung zukommt. Diese muss sämtliche relevanten Kontextinformationen enthalten, die erforderlich sind, um die Daten verstehen und interpretieren zu können. Technische Aspekte umfassen die Bereitstellung der Daten in Formaten, die idealerweise das Öffnen und Bearbeiten mit Open-Source-Software ermöglichen. Rechtliche Aspekte beziehen sich auf die explizite Festlegung der Bedingungen, unter denen eine Nachnutzung der Daten gestattet ist.

Basierend auf den Erkenntnissen aus dem Projekt RODS ist in Bezug auf Data Literacy im Kontext der Datennachnutzung im akademischen Lehr- und Lernumfeld zwischen

9 Detaillierte Ausführung zur Qualitätsprüfung der Daten bei der Datennachnutzung in der Checkliste: Mindestkriterien zur Nachnutzung von Forschungsdaten: https://doi.org/10.5281/zenodo.11385458 Diese Checkliste wurde im Rahmen des ROADS-Projekts erarbeitet.

Abbildung 2: Was bedeutet die Datenqualität im Kontext der Nachnutzung?

zwei Arten von Kompetenzen zu unterscheiden: Einerseits die Kompetenzen, die Studierende für die Wiederverwendung von Daten benötigen, und andererseits die Kenntnisse, die Lehrende zur Vermittlung und Förderung dieses Prozesses erfordern (Abbildung 3). Dabei sind die Kompetenzen in beiden Fällen eng mit den Merkmalen der Datenqualität verknüpft.

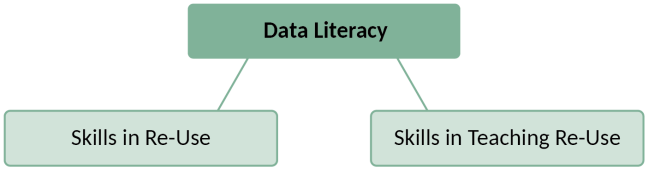


Abbildung 3: Was bedeutet Data Literacy im Kontext der Datennachnutzung im akademischen Lehr- und Lernumfeld?

Für Studierende umfasst Data Literacy unter anderem Kenntnisse über relevante Datenquellen sowie die Fähigkeit, diese gezielt zu recherchieren. Ebenso entscheidend ist der kritische Umgang mit Daten, insbesondere die Bewertung ihrer Zuverlässigkeit und der Angemessenheit ihrer Dokumentation. Zudem sollten Studierende mit einschlägiger Software vertraut sein und über rechtliche Kompetenzen im Umgang mit Daten verfügen, etwa hinsichtlich der Nachnutzungsbedingungen.

Für Lehrende liegt der Fokus auf methodisch-didaktischen Kompetenzen. Dazu gehört das Verständnis, welche Arten von Daten – in Abhängigkeit von Zielgruppe und Lehrkontext – für eine Wiederverwendung im akademischen Lehr- und Lernumfeld eignen und wie diese mit vertretbarem Aufwand in die Lehre integriert werden können. Ebenso relevant ist die Frage, wie die Datenkompetenz im Hinblick auf Datennachnutzung vermittelt werden sollte – ob als Bestandteil eines Curriculums oder als zusätzliches Lehrangebot.

Die beiden zentralen Komponenten – Datenqualität und Datenkompetenz – sind essenziell, um die Nachnutzbarkeit von Daten nachhaltig zu gewährleisten. Erst wenn beide Aspekte berücksichtigt werden, kann eine effektive Nachnutzung sichergestellt werden.

4 Ausblickend: Datennachnutzung als ein Schritt zur ressourcenschonenden Forschung

Nach der Darstellung der vielfältigen Herausforderungen im Zusammenhang mit der Datennachnutzung stellt sich die zentrale Frage, ob die Integration dieser Praxis in die akademische Lehre tatsächlich einen Mehrwert bietet oder ob die bestehenden Hürden die potenziellen Vorteile überwiegen. Der vorliegende Beitrag formuliert die Hypothese, dass Datennachnutzung im akademischen Lehrkontext weniger als Herausforderung, sondern vielmehr als Investition in eine nachhaltige, ressourcenschonende Forschung sowie in die langfristige Sicherung der Nachnutzbarkeit von Daten zu verstehen ist.[10]

Durch die gezielte Auseinandersetzung mit konkreten Problemstellungen werden Studierende frühzeitig für ein effizientes Forschungsdatenmanagement (FDM) sowie für die Relevanz der FAIR-Prinzipien sensibilisiert. Praktische Erfahrungen mit Herausforderungen wie unvollständigen Einwilligungserklärungen, die eine rechtssichere Nachnutzung verhindern, oder mit lückenhaften Dokumentationen, die die Interpretation der Daten erschweren, vermitteln Studierenden ein tiefgreifendes Verständnis dafür, welche Faktoren die Nachnutzbarkeit von Forschungsdaten wesentlich beeinflussen. Die FAIR-Prinzipien werden so nicht nur theoretisch erlernt, sondern im Sinne einer forschungspraktischen Kompetenzentwicklung verinnerlicht. Auf diese Weise entwickeln die Studierenden frühzeitig ein Bewusstsein für die Bedeutung nachhaltiger Datenpraktiken und wenden diese Prinzipien bei der Erhebung und Aufbereitung eigener Forschungsdaten – als künftige Generation von Open Scientists – selbstverständlich an.

Den Grundstein auf dem Weg zu einer nachhaltigen Forschung durch Datennachnutzung können unter anderem folgende Schritte bilden:

- **Schulungen zur Datenrecherche:** Spezialist:innen für Forschungsdaten sowie wissenschaftliche Bibliothekar:innen können gezielte Schulungen zur Datenrecherche anbieten – analog zu den bereits etablierten Kursen zur Literaturrecherche.
- **Praktische Anwendungsmöglichkeiten:** Lehrende und Forschende können den Studierenden vermehrt die Gelegenheit geben, bestehende Forschungsdaten wiederzuverwenden. Dies fördert nicht nur die Kompetenzentwicklung der Studierenden, sondern unterstützt auch eine Reflexion über nachhaltige Datenpraktiken innerhalb der wissenschaftlichen Gemeinschaft.
- **Entwicklung von Workflows und Richtlinien:** Es ist essenziell, klare Prozesse und Richtlinien für die Nachnutzung von Forschungsdaten im akademischen Lehr- und Lernkontext zu schaffen und zu etablieren.

10 Auf Bedeutung der curricularen Integration von Open-Science-Praktiken wird unter anderem auch in diesem Beitrag eingegangen: https://doi.org/10.7554/eLife.89736.

Die erfolgreiche Umsetzung dieser Schritte erfordert eine enge Zusammenarbeit zwischen Forschung, Lehre und forschungsunterstützenden Einrichtungen, wie beispielsweise Hochschulbibliotheken (Kohrs u. a. 2023).

Danksagungen

Ein besonderer Dank gilt swissuniversities für die Mitfinanzierung des Projekts. Ein ausdrücklicher Dank geht zudem an das ROADS-Projektteam: https://www.zhaw.ch/de/forschung/projekt/74318[11].

Interessenkonflikte

Die Autorin ist Koordinatorin des Projekts ROADS, dessen Ergebnisse die Grundlage des vorliegenden Beitrags bilden.

Literaturverzeichnis

Bendel, Alexander. 2021. „Von der Forschungsfrage zum Ergebnis: Eine Rekonstruktion der sekundäranalytischen Auswertung qualitativer Daten am Beispiel einer Masterarbeit“. Kap. Von der Forschungsfrage zum Ergebnis: Eine Rekonstruktion der sekundäranalytischen Auswertung qualitativer Daten am Beispiel einer Masterarbeit in *Qualitative Sekundäranalysen: Daten der Sozialforschung aufbereiten und nachnutzen,* herausgegeben von Caroline Richter und Katharina Mojescik, 37–54. Springer Fachmedien Wiesbaden. ISBN: 9783658328511. https://doi.org/10.1007/978-3-658-32851-1.

Gebel, Tobias und Alexia Meyermann. 2021. „Sekundäranalysen qualitativer Interviews: Eine Metaanalyse zur Praxis sekundäranalytischer Forschung zu Arbeitsorganisationen“. In *Qualitative Sekundäranalysen: Daten der Sozialforschung aufbereiten und nachnutzen,* herausgegeben von Caroline Richter und Katharina Mojescik, 17–36. Springer Fachmedien Wiesbaden. ISBN: 9783658328511. https://doi.org/10.1007/978-3-658-32851-1.

11 *Besucht am 1. Juli 2025 durch die Herausgeber*innen.*

Hollstein, Betina und Jörg Strübing. 2018. „Archivierung und Zugang zu Qualitativen Daten“. In *Archivierung und Zugang zu Qualitativen Daten. Working Paper Series des Rates für Sozial- und Wirtschaftsdaten (RatSWD),* herausgegeben von Regina T. Riphahn, 1–13. German Data Forum (RatSWD).

Kohrs, Friederike E., Susann Auer, Alexandra Bannach-Brown, Susann Fiedler, Tamarinde Laura Haven, Verena Heise, Constance Holman u. a. 2023. „Eleven strategies for making reproducible research and open science training the norm at research institutions“. *eLife* 12. https://doi.org/10.7554/elife.89736.

Netscher, Sebastian und Jessica Trixa. 2019. „Forschungsdatenmanagement in der Sekundäranalyse“. In *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten,* herausgegeben von Uwe Jensen, Sebastian Netscher und Kathrin Weller, 135–150. Verlag Barbara Budrich. https://doi.org/10.3224/84742233.09.

Challenge Accepted? Möglichkeiten und Grenzen eines NFDI-Serviceverzeichnisses am Beispiel des NFDI-N(HAW)igators

Eva Eilert [1], Martin Reiter [2]

[1] Hochschulbibliothek der Hochschule Düsseldorf, Düsseldorf, Deutschland
[2] Bibliothek der Hochschule Rhein-Waal, Kleve, Deutschland

Die Konsortien der Nationale Forschungsdateninfrastruktur (NFDI) bieten eine Vielzahl von Services für das Forschungsdatenmanagement (FDM), die für Forschung, forschungsorientierte Lehre und Transfer an Hochschulen für angewandte Wissenschaften (HAW) von großer Bedeutung sind. Aufgrund institutioneller Besonderheiten stehen HAW jedoch vor spezifischen Herausforderungen bei der Implementierung von FDM-Strukturen. Ein zentraler Lösungsansatz ist die Entwicklung eines Serviceverzeichnisses, das eine gezielte Zuordnung der NFDI-Services zu den Fachbereichen und Bedarfen der Forschenden an HAW ermöglicht – der NFDI-N(HAW)igator kann hierfür als beispielhafter Lösungsansatz dienen. Der Beitrag beleuchtet die Notwendigkeit einer klaren Service-Definition, die Bedeutung der Community-Beteiligung und die Herausforderungen bei der Aktualisierung und Pflege eines solchen Verzeichnisses. Es wird aufgezeigt, wie ein umfassendes Serviceverzeichnis zur effektiven Nutzung der NFDI-Angebote beitragen und die FDM-Praxis an HAW stärken kann.

Keywords: Nationale Forschungsdateninfrastruktur (NFDI), Hochschulen für angewandte Wissenschaften (HAW), FDM-Kontaktstelle, Serviceverzeichnis

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23924 (CC BY-SA 4.0).

1 Relevanz und Herausforderung der NFDI für Hochschulen für angewandte Wissenschaften

Die Fachkonsortien der Nationalen Forschungsdateninfrastruktur (NFDI) arbeiten auf die Bereitstellung von Services wie Helpdesks, technische Infrastrukturen oder Standards hin, die einem wissenschaftsgeleiteten und disziplinspezifisch ausgerichteten Umgang mit Forschungsdaten den Weg bereiten. Diese Services sind für Forschung, forschungsorientierte Lehre und Wissenstransfer von enormer Relevanz. In allen drei Bereichen spielen Hochschulen für angewandte Wissenschaften (HAW) bzw. Fachhochschulen (FH)[1] eine wichtige Rolle. „Insbesondere aufgrund ihres Praxisbezugs und ihrer Vernetzung mit regionalen Partnern“, so der Wissenschaftsrat , „sind HAW/FH wichtige Akteure im Wissenschafts- und Innovationssystem und können signifikant zum Transfer von Wissen beitragen“ (Wissenschaftsrat 2023, S. 31). Gleichzeitig stehen HAW aufgrund ihrer institutionellen Charakteristika vor der Aufgabe, adäquate Strukturen, Freiräume und Profile zu schaffen und/oder weiterzuentwickeln, um Forschung, forschungsorientierte Lehre und Transfer als Leistungsdimension zu stärken. Vor diesem Hintergrund ist die Vermittlung der Services der NFDI-Konsortien zum Forschungsdatenmanagement (FDM) ein wichtiger Baustein wissenschaftsunterstützender Strukturen an HAW.

Aus Hochschulperspektive bringt der Kosmos NFDI ganz eigene Herausforderungen mit sich. 26 Fachkonsortien bedeuten, dass potentiell 26 Mal Angebote für Support, Infrastrukturen, Standards etc. entwickelt oder zusammengeführt werden. Um diese Angebote zielgerichtet vermitteln zu können (und auf diese Weise zur Erhöhung der Nutzendenzahlen beizutragen), ist für die individuelle und zielgerichtete Beratung Forschender – neben allgemeiner Awarenesssteigerung für die NFDI – eine entsprechende Übersicht, d. h. ein Serviceverzeichnis, unerlässlich.

Insbesondere kleinere und/oder generalistisch ausgerichtete Servicestellen zum FDM, wie sie sich häufig an HAW finden, stellt die Orientierung in der NFDI und die Vermittlung der konsortialen FDM-Services allerdings vor Herausforderungen. Explizite Übersichten der wachsenden Angebote der einzelnen Konsortien liegen bislang nur sporadisch und nicht an zentraler Stelle vor. Lösungsansätze hierfür bieten unterschiedliche Übersichten und Handreichungen, wie z. B. eine Auflistung der FDM-Helpdesks der NFDI auf der Webseite der NFDI[2] (Arbeitsgruppe FDM-Helpdesk Netzwerk 2024), das Survival Kit des Projekts FDM-ndsHAW (Schmidt, Agniashvili und Mau 2024) sowie der NFDI-N(HAW)igator (Eilert und Reiter 2024), der prototypisch skizziert, wie ein Serviceverzeichnis aufgebaut sein kann.

1 Nachfolgend wird einheitlich die Bezeichnung HAW verwendet.

2 FDM-Helpdesks der NFDI: https://www.nfdi.de/helpdesks/; *Besucht am 1. Juli 2025 durch die Herausgeber*innen.*

2 Der NFDI-N(HAW)igator: Ein Lösungsansatz

Der NFDI-N(HAW)igator beruht auf Überlegungen, wie die erforderliche Kursbestimmung zwischen den Angeboten der Konsortien, den Bedarfen der Forschenden und FDM-Beratungsprozessen insbesondere an HAW ausgestaltet werden kann. Ausgangspunkt war die Frage, welches Konsortium welche Leistungen für welche Fachbereiche und welche Datentypen anbietet. Im Rahmen der Weiterentwicklung des NFDI-N(HAW)igators, seiner Vorstellung und Diskussion in verschiedenen Kreisen, aber auch ausgehend von Erfahrungen der Beratungspraxis kam die Frage hinzu, wie die FDM-Services Forschende in den Hochschulen erreichen. Schließlich ist von Relevanz, wie die Vermittlung der konsortialen FDM-Services einen Beitrag zur wechselseitigen, wirksamen Umsetzung strategischer Prozesse sowohl auf Seiten der Hochschulen (z. B. hinsichtlich Hochschulentwicklungsplänen) als auch der NFDI (z. B. hinsichtlich Strukturbericht und Vereinsstrategie) leisten kann. Im Unterschied zur Liste der FDM-Helpdesks und zum Survival Kit lag der Fokus von Anfang an dezidiert auf der Einbindung und Vermittlung konsortialer Services.

Der NFDI-N(HAW)igator wurde im Rahmen des Verbundprojektes der Hochschulen Rhein-Waal und Düsseldorf innerhalb der Förderlinie FDMScouts.nrw[3] (2020 – 2023) entwickelt. Zu diesem Zweck wurde die Gesamtmenge der Fächerspektren beider Hochschulen gebildet und so ein komplementäres Forschungsprofil generiert, das exemplarisch für die Schnittmengen und Unterschiede der Forschungsaktivitäten an HAW (in NRW) stehen kann. Der Zuschnitt auf die beiden Hochschulen bringt es trotz aller Beispielhaftigkeit mit sich, dass eine auf eine spezifische Spannbreite von Forschungsaktivitäten begrenzte *Auswahl* von NFDI-Konsortien vorgenommen wurde. Abbildung 1 gibt einen Überblick in den Aufbau des NFDI-N(HAW)igators. Im oberen Teil des NFDI-N(HAW)igators ist das Fächerspektrum der Hochschule Rhein-Waal und der Hochschule Düsseldorf – exemplarisch für das Fächerspektrum der HAW in NRW – anhand der Systematik der Fächergruppen, Lehr- und Forschungsbereiche und Fachgebiete des Statistischen Bundesamts (Destatis-Systematik, vgl. Statistisches Bundesamt 2023) aufgeschlüsselt und näherungsweise den Wissenschaftsbereichen der DFG zugeordnet. Der untere Teil der Grafik gibt in Form von Steckbriefen einen Überblick über die konsortialen Services innerhalb der NFDI. Die Linien innerhalb der Grafik verbinden die Steckbriefe mit den jeweiligen Fächern der Destatis-Systematik.

3 https://www.forschungsdaten.org/index.php/FDMScouts.nrw; *Besucht am 23. März 2025.*

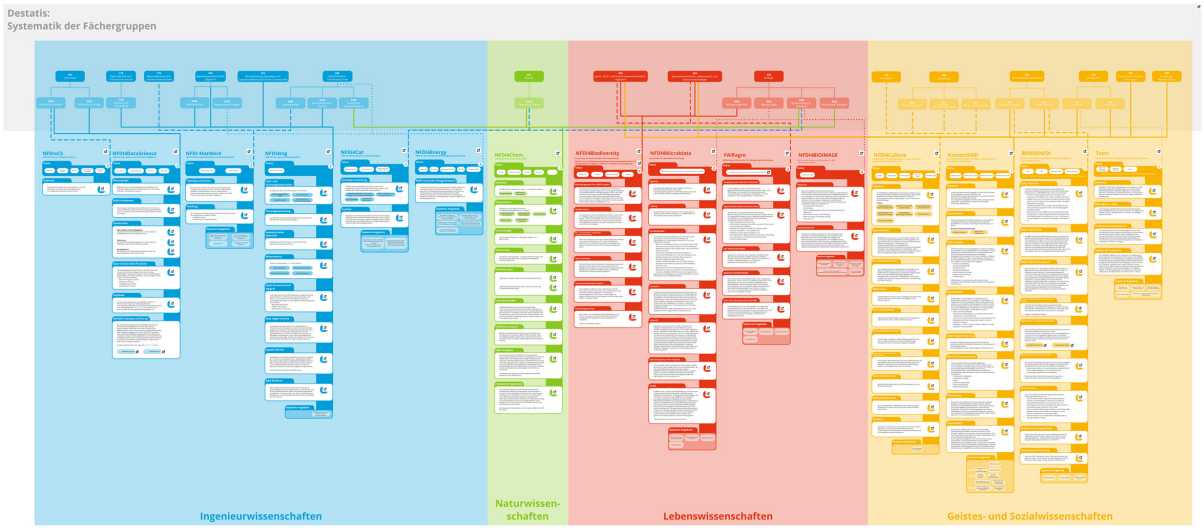


Abbildung 1: Der NFDI-N(HAW)igator in der Gesamtansicht.

3 Von der Kontaktaufnahme zur Servicevermittlung: Anwendung des NFDI-N(HAW)igators im Beratungsprozess

Ausgehend von einer allgemeinen Darstellung des Beratungsprozesses wird im Folgenden erläutert, wie FDM-Kontaktstellen einige der oben in Kapitel 1 genannten Herausforderungen mit Hilfe des NFDI-N(HAW)igators bewältigen können. Der Gesamtprozess ist in Abbildung 2 dargestellt.

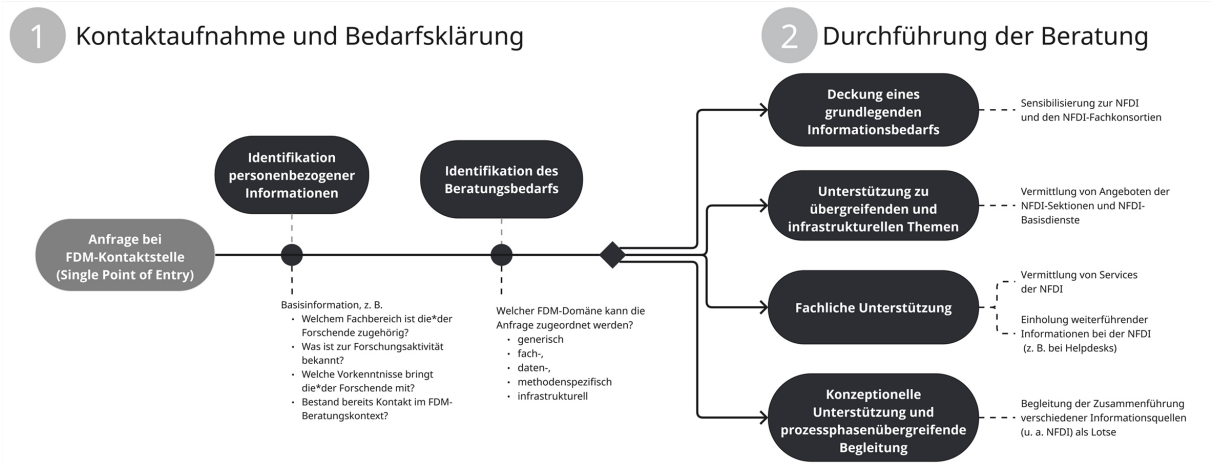


Abbildung 2: Flussdiagramm eines zweistufigen Beratungsprozess der FDM-Kontaktstelle.

Der musterhafte Beratungsprozess beginnt mit der Kontaktaufnahme der Forschenden über eine zentrale FDM-Kontaktstelle (Single Point of Entry). Organisatorisch sind diese Anlaufpunkte an den HAW in NRW (aber auch darüber hinaus) meist in der Forschungsförderung, weniger häufig auch in der Bibliothek angesiedelt und in der Regel mit geringen Ressourcen[4] ausgestattet. Typischerweise wird die Kontaktstelle von Einzelpersonen mit geringem bis mittlerem Stellenanteil betreut. Eine flächendeckende Etablierung von FDM-Kontaktstellen ist bei den HAW bislang noch nicht erfolgt, in zahlreichen Fällen befinden sie sich in teils drittmittelgeförderten Projektphasen.

Im nächsten Schritt werden die für die zielgerichtete Beratung relevanten Informationen ermittelt. Dazu gehören die Erfassung der persönlichen Daten sowie die Ermittlung des konkreten Beratungsbedarfs. Dafür werden Basisinformationen in Erfahrung gebracht wie z. B. das Fachgebiet der*des Forschenden, Details zu den Forschungsaktivitäten, bisherige Beratungsanfragen im FDM-Bereich und Vorkenntnisse der*des Forschenden. Darüber hinaus wird die Anfrage einer FDM-Domäne (generisch, fachspezifisch, daten- oder methodenspezifisch, infrastrukturell) zugeordnet.

Die weitere Durchführung der Beratung ist abhängig vom Beratungsbedarf. Es wird weitestgehend zwischen folgenden Bedarfen unterschieden: Grundlegende Informationen, Unterstützung zu übergreifenden und infrastrukturellen Themen, fachliche Unterstützung sowie konzeptionelle Unterstützung und Prozessbegleitung.

Die Deckung des grundlegenden Informationsbedarfs im Sinne einer Sensibilisierung für die NFDI und die NFDI-Fachkonsortien ist, wie verschiedene Quellen zeigen, an den HAW eine anzugehende Aufgabe. Lediglich rund 22 % der im Rahmen einer bundesweiten Studie befragten Personen an HAW haben von der NFDI überhaupt schon einmal gehört (vgl. Klocke u. a. 2023, S. 7). Dies wird auch durch die noch geringe aktive Nutzung der Services durch die staatlichen HAW (17,2 %) im Verhältnis zu den staatlichen Universitäten (80,5 %) bestätigt (vgl. Burk und Hetze 2023, S. 42). Gleichzeitig sieht die DFG für die zweite Förderphase der NFDI-Konsortien die „Erhöhung der Nutzung von Angeboten über den Kreis der Konsortialpartner hinaus“ (NFDI-Expertengremium 2023, S. 2) vor, die NFDI verfolgt laut Vereinsstrategie 2025/26 bei der Erstellung eines Dienstportfolios das Ziel, „die Nutzung der Dienste in der Wissenschaft zu steigern“.[5] Die Forschenden an den HAW werden daher für die NFDI und ihre Fachkonsortien im Beratungsprozess, wo immer sinnvoll, grundlegend sensibilisiert.

Bei der Unterstützung zu übergreifenden Themen und der fachlichen Unterstützung ist die Vermittlung von Services der NFDI-Konsortien oder der NFDI-Basisdienste, neben

4 Einen entsprechenden Titel trug der im Rahmen der Themenwoche „Digitale Kompetenzen in der Wissenschaft“ von der Volkswagenstiftung geförderte Workshop *Forschungsdatenmanagement mit geringen Ressourcen? Kompetenzaufbau durch Vernetzung von Hochschulen für angewandte Wissenschaften.*, s. https://www.hs-mainz.de/forschung/forschung-transfer/projekte/fdmhawrlp/forschungsdatenmanagement-mit-geringen-ressourcen-kompetenzaufbau-durch-vernetzung-von-hochschulen-fuer-angewandte-wissenschaften/; *Besucht am 22. Mai 2025.*

5 https://www.nfdi.de/verein/; *Besucht am 30. März 2025.*

dem Einholen von spezifischen Informationen zu diesen Services, relevant. Hier schließt sich ggf. eine längere Begleitung der Forschenden an. Eine solche Begleitung kann sowohl konzeptionelle Aspekte umfassen als auch eine Lotsenfunktion bei der Zusammenführung und Bereitstellung relevanter Informationen und Tools darstellen. In all diesen Phasen werden die Services der NFDI-Konsortien konkret einbezogen, wobei die damit verbundenen Herausforderungen im Folgenden beleuchtet werden.

4 Herausforderungen bei der Servicezuordnung und -vermittlung

Eine Herausforderung stellt die Zuordnung der NFDI-Konsortien zu den Wissenschaftsbereichen der DFG dar, in denen die Fächer der HAW nur teilweise abgebildet sind. Zudem besteht weder auf Ebene der Services noch auf Ebene der Konsortien eine die DFG-Wissenschaftsbereiche spezifizierende Feingliederung zur fachlich-systematischen Zuteilung (vgl. Nationale Forschungsdateninfrastruktur e.V. 2023, S. 3). Die Möglichkeit der Zuordnung einer Beratungsanfrage zu den Fachkonsortien ist dadurch erschwert und nur annäherungsweise möglich. Die Berücksichtigung der weiter gefassten Destatis-Systematik erlaubt eine feinere Auflösung und konkretere Zuordnung. Ein weiterer Vorteil der Verwendung der Destatis-Systematik ist der Umstand, dass sie an Hochschulen im Rahmen der Zuordnung von Personen zu einem Fach zu Verwaltungs- und Statistikzwecken ohnehin im Einsatz ist. Sie bildet zudem die Grundlage für die Fachzuordnung gemäß Kerndatensatz Forschung (KDSF) – Standard für Forschungsinformationen in Deutschland; über den KDSF ist perspektivisch ein Zusammendenken von NFDI-Serviceverzeichnis und Forschungsinformationssystemen möglich.

Auf Basis der Destatis-Systematik wurde im NFDI-N(HAW)igator eine Feingliederung vorgenommen, die eine differenzierte Zuteilung zu den relevanten NFDI-Konsortien ermöglicht, deren Services in Kurzübersichten (‚Steckbriefe') erfasst und mit den Fächerspektren vernetzt wurden. Die (inter-)disziplinäre Zuordnung über die Destatis-Systematik der NFDI-Konsortien ermöglicht somit eine schnelle Vermittlung der relevanten Services im Beratungsprozess.

Neben der grundlegenden Sensibilisierung ist bei konkreteren Bedarfen Unterstützung zu übergreifenden und infrastrukturellen als auch fachlichen Aspekten gefragt. Aufgrund struktureller Gegebenheiten, wie einem hohen Lehrdeputat und fehlenden Mittelbau, sind die Ressourcen für Forschungsaktivitäten an HAW im Vergleich zu Universitäten limitierter. Dies gilt analog für die Forschungsunterstützung. Dazu zählen etwa die häufig weniger umfangreich ausgebaute IT-Infrastruktur für die Forschung, der häufig projektbasierte, d. h. temporäre Status von FDM-Kontaktstellen sowie deren generalistische Ausrichtung.

Im Gegensatz zu den Universitäten, an denen FDM-Kontaktstellen häufig in Bibliotheken verankert sind und ggf. von Fachreferent*innen unterstützt werden können, ist die Situation an den HAW deutlich anders. Hier sind die FDM-Kontaktstellen seltener in Bibliotheken angesiedelt und die Rolle der Fachreferent*innen existiert in der Regel nicht. Dies führt zu einer besonderen Herausforderung bei der fachlichen Betreuung von Forschenden. Um diese bei der Bewältigung ihrer z. B. durch die Vorgaben von fördermittelgebenden Institutionen zunehmenden Aufgaben im Handlungsfeld FDM zu unterstützen, agieren Kontaktstellen als prozessphasenübergreifende Begleitung. Sie übernehmen, soweit möglich und angefordert, die ganzheitliche Begleitung des Forschungsprozesses (z. B. mit Lotsenfunktion) durch die Integration heterogener Informations- und Supportquellen, darunter die direkte Vernetzung mit NFDI-Helpdesks und Serviceverantwortlichen der Konsortien. Dadurch wird ein effizienter Zugang zu spezialisiertem Wissen und Ressourcen der NFDI gewährleistet.

Diesbezüglich bieten im NFDI-N(HAW)igator die Steckbriefe der Fachkonsortien konkrete Orientierungshilfen, die einen Überblick über (ausgewählte) nutzbare und geplante Services geben. Diese werden bewusst in einer fachsprachereduzierten Form beschrieben, um den Zugang für das FDM-Kontaktstellenpersonal mit den unterschiedlichsten Fachhintergründen zu erleichtern und um Einbußen hinsichtlich der Vermittelbarkeit von spezifischen Services zu verhindern. Hierbei wurde bewusst die Perspektive der FDM-Kontaktstellen eingenommen. Services sollten zudem keinen Einschränkungen bei der Nutzung unterliegen (z. B. institutionslimitierter Login). Abbildung 3 gibt einen exemplarischen Einblick in den Aufbau dieser Steckbriefe. Es werden die nutzbaren und geplanten Services dargestellt und mit Links zu den entsprechenden Webseiten angereichert.

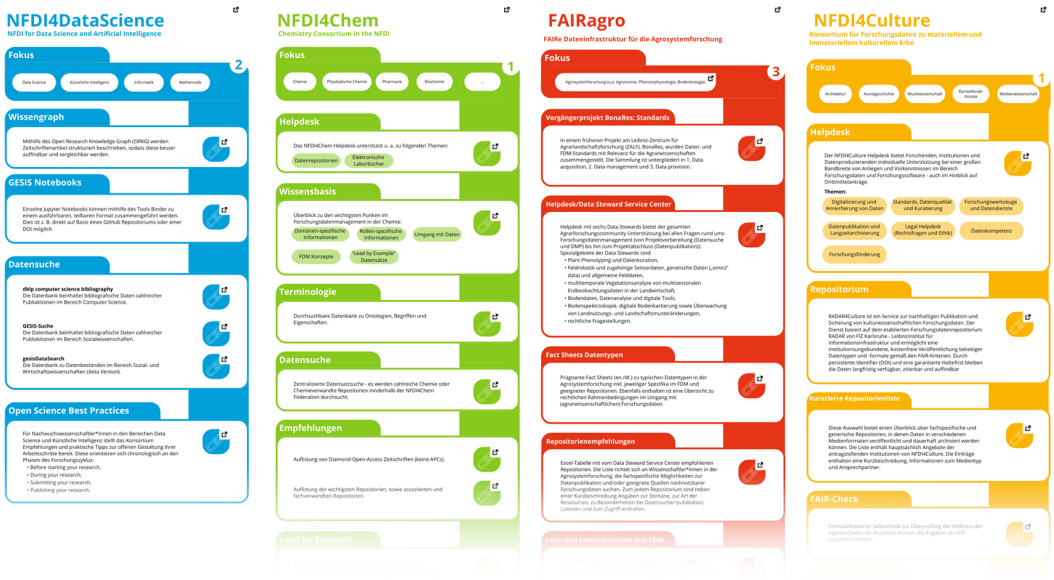


Abbildung 3: Kurzübersicht der einzelnen Konsortien (‚Steckbriefe‘) zur schnellen, inhaltlichen Orientierung.

Die Integration aller relevanten NFDI-Konsortien in den NFDI-N(HAW)igator ist von entscheidender Bedeutung, um ein vollständiges Bild der verfügbaren Angebote zu zeichnen und diesen zu einem hochschultypübergreifenden Serviceverzeichnis weiterzuentwickeln. Die dynamische Entwicklung der NFDI-Landschaft erfordert außerdem kontinuierliche Anpassungen und Aktualisierungen, um die Relevanz und Genauigkeit der Informationen zu gewährleisten. Eine der größten Herausforderungen liegt in der Sicherstellung der Aktualität der enthaltenen Informationen. Dementsprechend ist die Implementierung von Prozessen zur einfachen und schnellen Aktualisierung entscheidend.

Mit der wachsenden Anzahl an Angeboten steigt die Komplexität der Informationsarchitektur, die erforderlich ist, um einen Großteil der oder gar alle Services zu verzeichnen. Hinzu kommt das Erfordernis nach einer Steuerungsfunktion für ein solches Verzeichnis in Anbetracht der Vielzahl von Services. Im Rahmen einer SWOT-Analyse bringt der NFDI-Strukturbericht dies unter „Schwächen (intern)“ zum Ausdruck:

> „Dem Verein fehlen wichtige Steuerungsfunktionen in der Satzung: [....] Das NFDI-Dienste-Portfolio beinhaltet perspektivisch sowohl domänenspezifische Dienste (kommend aus Fach- und Methodenkonsortien) als auch übergreifende Dienste (kommend aus dem Basisdienstkonsortium). Die Zahl der aus den KgBLV (Konsortien gemäß Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018) und dem Basisdienstkonsortium kommenden Dienste beträgt derzeit über 600 Dienste und es nicht nicht (sic) klar, wie vollständig diese Zahl ist. Vor diesem Hintergrund bedarf es einer Steuerungsfunktion für ein übergreifendes NFDI-Dienste-Portfolio. Diese Steuerungsfunktion ist notwendig, um die Weiterentwicklung der Gesamtheit der Dienste aktiv zu gestalten. Flankierend ist zudem die Entwicklung einer übergreifenden Architektur notwendig, welche die Vielfalt der Dienste und der Dienste-Architekturen angemessen berücksichtigt“ (Sure-Vetter 2024, S. 121).

Ein strukturiertes Serviceverzeichnis oder Diensteportfolio kann dann wiederum einen gewichtigen Beitrag zur Erhöhung der Nutzendenzahlen leisten.

5 Ein interaktives Serviceverzeichnis: Proof of Concept

Der NFDI-N(HAW)igator hat zum Ziel, die Strukturierung eines solchen, im zitierten Bericht als Bedarf anerkannten Portfolios vorzunehmen und enthält aktuell 80 Services von 15 Konsortien (Stand Februar 2025).

Der statische Charakter der bisherigen grafischen Darstellung stößt naturgemäß an Grenzen der Nutzungsmöglichkeiten und Aktualisierbarkeit. Eine interaktive, intuitive Benutzeroberfläche und effektive Suchfunktionen versprechen demgegenüber eine deutliche Verbesserung der Auffindbarkeit der relevanten Services. Zuletzt wurde daher der Proof of Concept eines dynamischen, such- und filterbaren Serviceverzeichnisses entwickelt. An die Stelle des Conceptboards rückt eine (inhaltsidentische) Datenbank mit den entsprechenden Verknüpfungen von Fächerklassifikationen (Destatis-Systematik, erprobungsweise DFG-Systematik[6] und Dewey-Dezimalklassifikation der Deutschen Nationalbibliothek (DNB)[7]), NFDI-Konsortien und deren Fachspektren sowie den berücksichtigten Services.

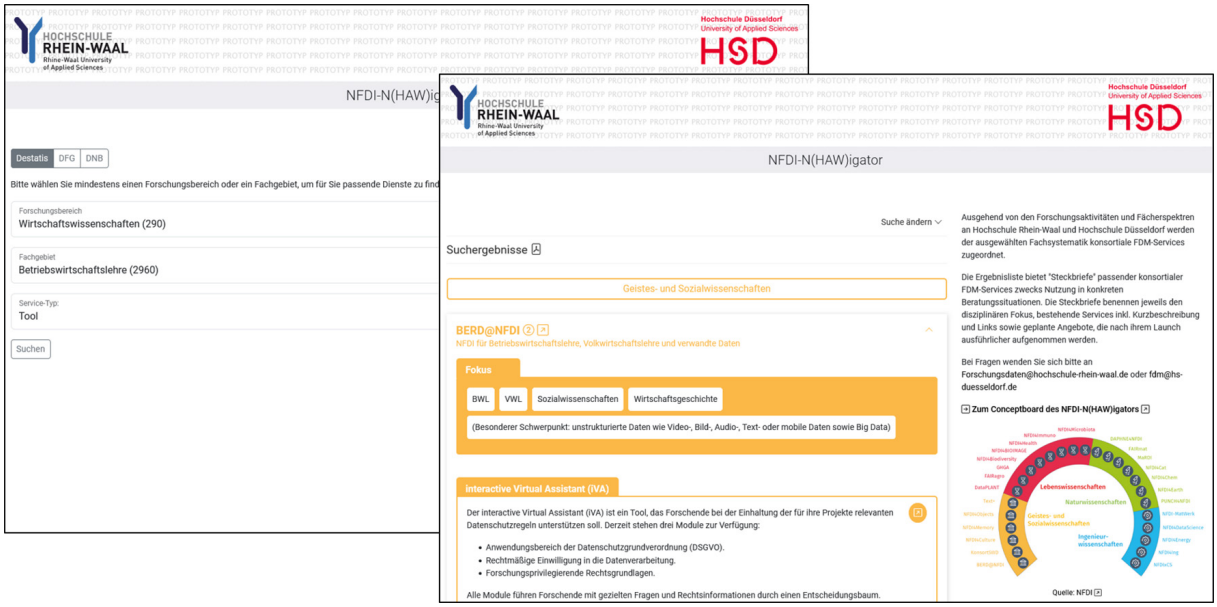


Abbildung 4: Forschungsbereich und Fachgebiet (basierend auf der Destatis-Systematik) wurden in der Suchmaske ausgewählt und ein Service-Typ (Tool) hinzugefügt (links). Der Ergebnisausschnitt (rechts) zeigt das verknüpfte Konsortium und angebotene Tools.

Im dynamischen NFDI-N(HAW)igator werden die Ergebnisse, wie in Abbildung 4 ersichtlich, auf Forschungsbereich, Fachgebiet und Service-Typ zugeschnitten ausgegeben. Die Darstellung wählt eine analoge Form zum Conceptboard, bietet aber neben der gesteigerten Übersichtlichkeit auch die Möglichkeit, Ergebnisse als PDF zu exportieren und anfragenden Personen direkt zur Verfügung zu stellen. Mit einer derartigen dynamischen Version ist aktuell ein Entwicklungsstand erreicht, der ein erhöhtes Maß an Nutzungsfreundlichkeit und Aktualisierbarkeit der Inhalte sicherstellt. Perspektivisch können auch weitere Klassifikationsschichten aktiv eingebunden werden.

6 https://www.dfg.de/resource/blob/331944/33422f091e941592cdc355038a865e03/fachsystematik-2024-2028-de-data.pdf; *Besucht am 22. Mai 2025.*

7 https://www.dnb.de/DE/Professionell/DDC-Deutsch/WasistdieDDC/wasistdieddc_node.html; *Besucht am 22. Mai 2025.*

Zum gegenwärtigen Zeitpunkt handelt es sich um einen nicht-öffentlich zugänglichen Arbeitsstand. In der dynamischen Version lassen sich Forschungsbereiche und Fachgebiete gemäß Destatis auswählen und mit einem Service-Typ kombinieren (vgl. Abbildung 4). Der Vorteil gegenüber dem Conceptboard liegt in der Möglichkeit der Eingrenzung der Suchergebnisse: Es werden nicht ganze Steckbriefe, sondern gezielte Ergebnisse erhalten, die wiederum direkt für den Beratungsprozess weiterverwendet werden können.

6 Definitorische Unschärfen im NFDI-Serviceverständnis

Während für die Felder ‚Forschungsbereich' und ‚Fachgebiet' aufgrund der Destatis-Systematik klare Kategorisierungen möglich sind, ist der Begriff ‚Service' im Kontext der NFDI durchaus komplexer. Die NFDI-Konsortialversammlung definiert ‚Service' als „eine technisch-organisatorische Lösung (...), die sowohl Speicher- und Rechenleistungen, Software, Prozesse und Workflows ebenso umfassen kann, wie die notwendige personelle Betreuung für unterschiedliche Service-Desks" (Konsortialversammlung des Vereins Nationale Forschungsdateninfrastruktur (NFDI) e.V. 2022, S. 2). Diese grundlegende Definition bildet die Basis für das Verständnis von NFDI-Services. Jedoch zeigt sich in der praktischen Umsetzung eine deutliche Vielfalt in der detaillierten Ausgestaltung von Kategorien. So unterscheiden sich die Konsortien, wie beispielsweise NFDI4Biodiversity und Text+, in der Art und Weise, wie sie ihre Services weiter untergliedern (z. B. „TOOL/APPLICATION", „SUPPORT", „DATA CURATION" auf Basis der de.NBI-Kategorien im erstgenannten Fall, „broadly defined" und „an offering developed within Text+ as well as other existing services from contributing partners" im zweitgenannten Fall. Vgl. Meister u. a. 2023, S. 10-11 und Buddenbohm u. a. 2024, S. 2). Für die Etablierung eines übergreifenden Serviceverzeichnisses müsste eine möglichst inklusive Definition für die Kategorienbildung erarbeitet werden.

7 Community-Beteiligung: Ein Schlüssel zur nachhaltigen Entwicklung eines Serviceverzeichnisses

Um Nachhaltigkeit, Ressourcenschonung und Bedarfsnähe bei der Entwicklung eines NFDI-Serviceverzeichnisses und damit in der FDM-Beratung zu gewährleisten, ist der aktive Einbezug der FDM-Community unerlässlich. Dies kann durch verschiedene Maßnahmen erreicht werden: Eine wichtige Strategie ist die Filterung und Reformulierung der konsortialen Servicebeschreibungen durch FDM-Kontaktstellenpersonal an den Hochschulen. Diese Anpassung soll hinsichtlich der Services die Verständlichkeit für die FDM-Kontaktstellen

und die Relevanz, Nutzbarkeit und Vermittlungsadäquanz für die Forschenden an den Hochschulen sichern. Ein weiterer entscheidender Aspekt ist die Informationsweitergabe durch und die Absprache mit Personen aus den NFDI-Konsortien, wie beispielsweise Helpdesk-Personal. Durch diesen direkten Austausch können aktuelle Informationen und spezifische Bedarfe schnell und effizient kommuniziert werden. Ein zentraler Vorteil dieser gemeinschaftlichen Herangehensweise ist die Reduzierung des Pflegeaufwands für einzelne Akteur*innen. Je mehr Personen bzw. Institutionen sich an der Aufbereitung und dem Austausch (basierend auf einem Konzept geregelter Zuständigkeiten und Verantwortlichkeiten) beteiligen, desto geringer ist der Aufwand für jede einzelne Person bzw. Institution.

8 Fazit: Ein Serviceverzeichnis als Mehrwert für Hochschulen und NFDI

Die Etablierung eines umfassenden Serviceverzeichnisses für die Angebote der NFDI stellt einen zentralen Baustein dar, um die vielfältigen Ressourcen und Dienstleistungen der NFDI-Konsortien für die Forschenden an den Hochschulen sichtbar und zugänglich zu machen. Es kann zum einen ein Instrument für die Awareness(-steigerung) für die Leistungen der NFDI von Forschenden und FDM-Kontaktstellenpersonal (an HAW und darüber hinaus), aber auch Entscheider*innen in Hochschulleitungen und Politik für die Leistungen der NFDI sein, wie es im NFDI-Strukturbericht formuliert wird: „Das angestrebte Dienste-Portfolio ist somit eng mit der weiterführenden Marketingstrategie von NFDI verknüpft. Es wird Forschenden zukünftig einen einzigartigen Überblick über die Dienste von NFDI bieten. Mit diesem Angebot werden Mehrwerte geschaffen, die Forschende nirgendwo anders finden können – in der Sprache des Marketings also ein Unique Selling Point“ (Sure-Vetter 2024, S. 87). Zum anderen kann der Aufbau eines solchen Verzeichnisses durch Mitwirkung der FDM-Community den Aufwand auf viele tragfähige Schultern verteilen. Der Austausch mit der FDM-Community zeigt, dass die Bereitschaft hierfür vorhanden ist. Durch die kooperative Herangehensweise – aus der auch der NFDI-N(HAW)igator hervorgegangen ist und weiterentwickelt wird – können Herausforderungen angegangen und ein nachhaltiger *Change* in der bundesweiten FDM-Landschaft bewirkt werden.

Hochschullogos sind aus der Lizenz ausgenommen. Es gilt das deutsche Urheberrecht.

Danksagungen

Die Entwicklung des dynamischen Proof of Concept wurde durchgeführt von Jörg Booth, Bibliothek der Hochschule Rhein-Waal Waal (https://orcid.org/0009-0002-8124-1772).

Die Autor*innen danken dem FDM-Personal der Hochschulen der Förderlinie FDMScouts.nrw, der IG FDM an HAW in NRW sowie der UAS7-Hochschulen für wertvolles Feedback zum NFDI-N(HAW)igator.

Die Förderlinie FDMScouts.nrw war ein Kooperationsvorhaben empfohlen durch die Digitale Hochschule NRW, gefördert durch das Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen. Sie wurde koordiniert durch die Landesinitiative für Forschungsdatenmanagement – fdm.nrw.

Beiträge

Beide Autor*innen haben gleichermaßen an der Erstellung des Manuskripts mitgewirkt.

Interessenkonflikte

Die Autor*innen erklären, dass kein Interessenkonflikt vorliegt.

Literaturverzeichnis

Buddenbohm, Stefan, Lukas Weimer, Laure Barbot, Edward J. Gray, Nanette Rissler-Pipka und Melina L. Jander. 2024. *Guidelines for adding Text+ Services to the SSH Open Marketplace.* Zenodo. https://doi.org/10.5281/zenodo.11067639.

Burk, Marian und Pascal Hetze. 2023. *Hochschul-Barometer: Lage und Entwicklung der Hochschulen aus Sicht ihrer Leitungen.* Herausgegeben von Stifterverband für die Deutsche Wissenschaft e. V. Besucht am 12. Juni 2025. https://www.hochschul-barometer.de/2023.

Eilert, Eva und Martin Reiter. 2024. *NFDI-N(HAW)igator.* Zenodo. https://doi.org/10.5281/zenodo.13973396.

Klocke, Andreas, Robert Werth, Arnela Balic und Meike Backes. 2023. *Entwicklung und Verbreitung von Forschungsdatenmanagement an Fachhochschulen und Hochschulen für Angewandte Wissenschaften (EVER_FDM): Schlussbericht zum Projekt EVER-_FDM.* Forschungszentrum Demografischer Wandel (FZDW), Frankfurt University of Applied Sciences, Frankfurt am Main. https://doi.org/10.2314/KXP:1877275336.

Konsortialversammlung des Vereins Nationale Forschungsdateninfrastruktur (NFDI) e.V. 2022. *Stellungnahme der NFDI-Konsortien zu Basisdiensten.* Zenodo. https://doi.org/10.5281/zenodo.6091656.

Meister, Maria, Manuel Feser, Lena Steilen, Birgitta König-Ries, Bernhard Seeger, Barbara Ebert und Frank Oliver Glöckner. 2023. *Service Portfolio Report 2023.* Zenodo. https://doi.org/10.5281/zenodo.8327366.

Nationale Forschungsdateninfrastruktur e.V. 2023. *Kurzinformation NFDI und Konsortien.* Besucht am 12. Juni 2025. https://www.nfdi.de/wp-content/uploads/2024/10/Einseiter-Konsortien_v2.pdf.

NFDI-Expertengremium. 2023. *Eckpunkte für die zweite Förderphase der NFDI-Konsortien.* Zenodo. Deutsche Forschungsgemeinschaft. https://doi.org/10.5281/zenodo.8269853.

Schmidt, Diana, Ana Agniashvili und Franziska Mau. 2024. *Survival Kit: Die Nationale Forschungsdateninfrastruktur Deutschland (NFDI).* Zenodo. https://doi.org/10.5281/zenodo.14441578.

Statistisches Bundesamt. 2023. *Systematik der Fächergruppen, Studienbereiche und Studienfächer: Stand Wintersemester 2023/2024.* Besucht am 12. Juni 2025. https://www.destatis.de/DE/Methoden/Klassifikationen/Bildung/studenten-pruefungsstatistik.html.

Sure-Vetter, York. 2024. *Erfahrungen mit der strukturellen Gestaltung von NFDI: Bericht des Direktors gemäß § 13 Abs. 2 der Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018.* Zenodo. https://doi.org/10.5281/zenodo.14867064.

Wissenschaftsrat. 2023. *Strukturen der Forschungsfinanzierung an deutschen Hochschulen: Positionspapier.* Positionspapier. Köln. https://doi.org/10.57674/pms3-pr05.

Pitfalls in the Field – Rechtliche Herausforderungen beim Einsatz von ML in der Agrosystemforschung

Constantin Breß [1,*], Lea Sophie Singson [1]

[1] FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Karlsruhe, Germany
[*] Kontakt: constantin.bress@fiz-karlsruhe.de

Beim Einsatz von Machine Learning in der Agrosystemforschung zeigen sich diverse rechtliche „Pitfalls“, welche hier vorgestellt werden. Zusätzlich wird eine Lösungsmöglichkeit präsentiert.

Die rechtlichen Probleme lassen sich auf verschiedene „Lebensphasen“ eines ML-Modells aufteilen: Training, die Existenz des Modells selbst und sein schlussendlicher Einsatz. Rechtliche Probleme ergeben sich vor allem im Zusammenhang mit dem Training: Daten können rechtlich geschützt sein. Gleichzeitig sieht das Recht gewisse Ausnahmen vor, welche fruchtbar gemacht werden können.

Ein zusätzliches und noch im Wachstum befindliches Feld ist der rechtliche Schutz der beim Training entstehenden Modelle. Bei deren Schaffung bedarf es nicht unerheblicher menschlicher Leistung, sodass die Beteiligten ein Interesse daran haben, ihre Urheberschaft anerkannt zu bekommen und eventuell auch daran, das weitere Schicksal der Modelle zu beeinflussen. Rechtlich bestehen dabei noch massive Unsicherheiten.

Den rechtlichen Problemen gemeinsam ist ihre hohe Komplexität, die sich teilweise erst im Laufe des Vorhabens zeigt. Daher können Rechtsfragen vor Projektbeginn nicht erschöpfend abgehandelt werden, sondern es ist eine rechtliche Begleitung notwendig, welche zudem auf hohem Niveau arbeitet – das auch aufgrund der Herausforderungen durch das neue EU-Datenrecht. FAIRagro setzt dabei auf das Modell des Data Steward Service Center (DSSC). Dieses unterstützt Forschende aus der Agrosystemcommunity bei auftretenden Problemen in allen Bereichen des Forschungsdatenmanagements – inklusive

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23925 (CC BY-SA 4.0).

rechtlichen und ethischen Themen. Dieses Modell kann auch Vorbild für andere Forschungsdatencommunities sein.

Keywords: Machine Learning, Data Law, Copyright Law, Data Protection Law, Artificial Intelligence Law

1 Einleitung

Künstliche Intelligenz (KI) sorgt nicht nur in Techniker:innenkreisen, sondern auch in der Rechtswissenschaft und -praxis für Diskussionen. Dies nicht zuletzt auch aufgrund der KI-Verordnung der EU[1], welche am 2. Februar 2025 auch schon teilweise Geltung erlangte (Art. 113). Diese Rechtsfragen stellen sich darüber hinaus nicht nur allgemein, sondern sind teilweise auch sehr domain-spezifisch und einzelfallbezogen.

Dieser Beitrag soll einen Überblick über die verschiedenen rechtlichen Probleme im Zusammenhang mit künstlicher Intelligenz innerhalb und außerhalb der KI-Verordnung im Kontext der Agrarforschung geben. Dafür wird teilweise vereinfachend zunächst das Verständnis von KI dargestellt, welches diesem Paper zu Grunde liegt. Da sich die rechtlichen Fragestellungen je nach Phase in Erstellung und Nutzung eines Modells unterscheiden, orientiert sich die Darstellung an diesen Phasen. Daran anschließend folgt ein Überblick über das Datenrecht, bevor auf die eigentlichen rechtlichen Problemstellungen im Zusammenhang mit künstlicher Intelligenz eingegangen wird. Der Untersuchung wird die Rechtslage in Deutschland zu Grunde gelegt und orientiert sich an den Fragestellungen aus dem NFDI-Projekt „FAIRagro“. Das Geschäftsgeheimnisrecht wird bei diesem Beitrag weitestgehend nicht behandelt.

2 Künstliche Intelligenz

KI oder Machine Learning wird hier in Anlehnung an die Definition der KI-Verordnung für ein KI-System verstanden. In Art. 3 Nr. 1 KI-VO ist ein *KI-System* definiert als

> ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Emp-

1 Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz [...] (Verordnung über künstliche Intelligenz), ABl. L v. 12. Jun. 2024.

> fehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können[.]

Indem die KI-Verordnung sich auf ein System bezieht, behandelt sie v.a. ein konkretes Modell, aber auch seine Einbettung in eine bestimmte Infrastruktur. Dieses Paper beschäftigt sich mit den einzelnen Modellen, sodass der Begriff der KI hier enger verstanden wird als in der KI-Verordnung. Dennoch erfasst der Begriff sowohl „einfache“ Modelle wie Klassifikatoren als auch *Large Language Models*. Ein solches Modell lässt sich wiederum (mindestens) in zwei rechtlich relevante Elemente aufspalten: Die grundsätzliche Architektur des Modells und die Weights, die innerhalb dieser Architektur eingebunden werden (so auch Schumacher 2023, S. 36; Übersicht bei Glauner 2024; synonym zu Architektur wird auch „Topologie“verwendet, etwa bei Schröler und Kuß 2022, Rn. 40–42).

Weights bezeichnet vereinfacht die Gewichtung der verschiedenen Elemente innerhalb der Architektur eines bestimmten Modells. Dadurch entsteht nach einem mehrschrittigen Prozess die Ausgabe des Modells. Die jeweiligen Weights hängen von den zum Training verwendeten Daten ab (weiterführend Schumacher 2023, S. 36 f.). Das Verhältnis von Architektur und Weights ist ungefähr so wie das Verhältnis zwischen einem Apfelkern und dem daraus wachsenden Baum: Der Kern gibt prinzipiell vor, wie der Baum aussehen soll; die konkrete Ausformung hängt aber von den Umgebungsfaktoren wie etwa der Bodenbeschaffenheit ab.

3 Lebensphasen einer KI

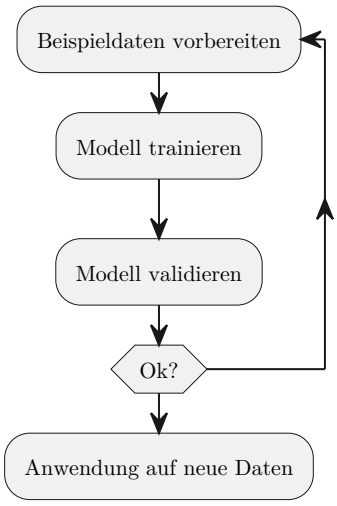


Abbildung 1: Lebenszyklus eines Machine Learning-Modells.

Abbildung 1 zeigt exemplarisch den Lebenszyklus eines Machine-Learning-Modells (in Anlehnung an Kleinkopf 2022, S. 48–62; Lee, Cooper und Grimmelmann 2024, S. 49–52). Dieser wird der weiteren rechtlichen Analyse zu Grunde gelegt.

4 Mögliche Schutzrechte an Daten

Als Rechte an Daten kommen vor allem das Urheberrecht und verwandte Rechte in Betracht. Je nach Zusammenstellung der Daten kann auch das Datenschutzrecht relevant sein.

4.1 Urheberrecht

Das Urheberrechtsgesetz (UrhG) schützt in erster Linie Werke, also konkrete Ausformungen einer Idee. Dieser Schutz setzt mindestens eine persönliche geistige Schöpfung voraus, das Werk muss die sogenannte *Schöpfungshöhe* erreichen (§ 2 Abs. 2 UrhG, für Datenbankwerke § 4 UrhG). Eine persönliche Schöpfung liegt vor, wenn das Werk auf einer menschlich-gestalterischen Tätigkeit beruht (Leistner 2020, Rn. 38; Schulze 2022, Rn.8). Soweit eine Maschine als Hilfsmittel herangezogen wird, kann eine solche Schöpfung gegeben sein, wenn ein Mensch darauf gestaltenden Einfluss hat (Leistner 2020, Rn. 40).

Zentrale Ergebnisse in der Agrarforschung sind Daten(-sammlungen). Zumindest Einzeldaten werden nach (fachspezifischen) Vorgaben erhoben, sodass der forschenden Person kein hinreichender Gestaltungsspielraum zukommt. Damit fehlt die nach § 2 Abs. 2 bzw. § 4 UrhG notwendige Schöpfungshöhe und die Daten sind folglich nicht urheberrechtlich geschützt (Bomhard und Siglmüller 2024, Rn. 11; Schulze 2022, Rn. 26). Für die verschiedenen Bestandteile eines Datensatzes kann sich aber eine abweichende Bewertung ergeben (ausführlich mit Beispielen siehe Hartmann 2022). Das kann auch etwa für eine vorgenommene Annotierung von Daten der Fall sein (Hacker 2020, S. 1027).

Dasselbe gilt in der Regel auch, wenn Daten aus der Verwendung von Computerprogrammen resultieren. Ein vom Schutz für Computerprogramme abgeleiteter Schutz für Daten ergibt sich nur, wenn die Daten im Programm selbst angelegt sind (Schulze 2022, Rn. 14).

Urheberrechtlicher Schutz kann sich in der Agrosystemforschung vor allem bei einer Zusammenstellung mehrerer Datensätze ergeben, sofern die Zusammenstellung Ergebnis eines kreativen, originellen Schaffens war. Dann würde sich ein Datenbankwerk nach § 4 Abs. 2 UrhG ergeben; allerdings sind dabei nicht die Daten als solche, sondern nur die Zusammenstellung als Datenbank geschützt (Weidert und Bug 2023, Rn. 34).

Ob die Zusammenstellung der Datensätze urheberrechtlichen Schutz erlangt, ist aber eine Frage des jeweiligen Einzelfalls.

4.2 Datenbankschutz

Erfolgt die Zusammenstellung des Datensatzes nicht in kreativer, origineller Weise, kann der Datensatz dennoch als Datenbank nach §§ 87a–87e UrhG geschützt sein. Im Unterschied zum Datenbank*werk* handelt es sich dabei um ein Leistungsschutzrecht, da es nicht an die schöpferische Tätigkeit einer Person anknüpft, sondern an investierter Leistung. Daher steht dieses Recht auch nicht der:dem Schöpfer:in des Werkes zu, sondern der:dem Datenbankhersteller:in. Da dieses Recht im Vergleich zum sonstigen Urheberrecht andere Voraussetzungen aufweist, handelt es sich dabei um ein Recht eigener Art, es wird daher auch als *sui-generis*-Recht bezeichnet. Ein weiterer Unterschied zum Urheberrecht besteht darin, dass dem:der Datenbankhersteller:in nur das ausschließliche Recht zur Vervielfältigung *wesentlicher* Teile der Datenbank zukommt. Im mkehrschluss kann damit jede Person unwesentliche Teile vervielfältigen, zusätzlich zu den Ausnahmen in § 87c UrhG.

Ein Datensatz ist als Datenbank i.S. des § 87a UrhG geschützt, wenn er sechs Bedingungen erfüllt: Es muss sich um eine Sammlung von Werken, Daten oder anderen unabhängigen Elementen handeln, diese müssen berücksichtigungsfähig für das Recht aus § 87a UrhG sein, systematisch oder methodisch angeordnet, einzeln mit Hilfe elektronischer Mittel oder auf andere Weise zugänglich sein und schließlich muss die Beschaffung, Überprüfung oder Darstellung eine nach Art oder Umfang wesentliche Investition erfordern.

Sammlung iS des § 87a UrhG ist die Zusammenstellung einer Vielzahl von Elementen (Vohwinkel 2024, Rn. 28 mwN).

Unabhängige Elemente iS des § 87a UrhG liegen vor, wenn sie aus der Datenbank herausgelöst werden können und ihnen trotzdem ein hinreichender Informationswert verbleibt.[2] Ob ein solcher Informationswert verbleibt, ist aus Perspektive eines interessierten Dritten zu beurteilen.[3] Damit reicht es aus, wenn irgendwer etwas mit den herausgelösten Elementen anfangen kann. Zwar muss dies nicht unbedingt auf ein einzelnes Datum zutreffen (wenn es nicht schon den Begriff des Datums iS des UrhG erfüllt), vielmehr kann das auch bei Datenabschnitten hinreichender Länge der Fall sein (Hacker 2020, S. 1029 f.).

Die Elemente müssen überhaupt berücksichtigungsfähig sein, sie dürfen damit keiner Bereichsausnahme unterliegen. Eine solche ist Art. 43 der Daten-Verordnung (D-VO)[4] (Richter 2024, Rn. 19 f.). Danach sind Daten vom *sui-generis*-Schutz nach § 87a ff. UrhG ausgenommen, wenn sie der Datenverordnung unterfallen und durch vernetzte Produkte

2 *EuGH*, Urt. v. 29. Okt. 2015 (C-490/14) – Verlag Esterbauer, Rn. 17; *EuGH*, Urt. v. 9. Nov. 2004 (C-444/02) – Fixtures Marketing, Rn. 29–32.

3 *EuGH*, Urt. v. 29. Okt. 2015 (C-490/14) – Verlag Esterbauer, Rn. 27.

4 Verordnung (EU) 2023/2854 des Europäischen Parlaments und des Rates vom 13. Dezember 2023 über harmonisierte Vorschriften für einen fairen Datenzugang und eine faire Datennutzung sowie zur Änderung der Verordnung (EU) 2017/2394 und der Richtlinie (EU) 2020/1828 (Datenverordnung), ABl. L v. 22. Dez. 2024.

(Art. 2 Nr. 5 D-VO) oder verbundene Dienstleistungen (Art. 2 Nr. 6 D-VO) erlangt oder erzeugt wurden. Aus ErwG 15 S. 8–10 D-VO ergibt sich, dass Sensordaten, auch in aufbereiteter Form, nicht nach §§ 87a ff. UrhG als Teil einer Datenbank geschützt sein können (Richter 2024, Rn. 17; Antoine 2024, Rn. 25 f.). Geschützt werden können aber Daten, welche aus solchen „Rohdaten" abgeleitet wurden, etwa durch weitere Auswertung oder „Sensorfusion" (ErwG 15 S. 13 f. D-VO). Daran werden die meisten Daten in Kontext der Agrarforschung wohl scheitern, sodass hier das *sui-generis*-Recht nicht greift.

Die Anordnung ist systematisch, wenn sich die Sammlung selbst an einem System, einer Klassifizierung oder einem Ordnungsschema orientiert (Dreier 2025b, Rn. 17). Sie ist methodisch, wenn die Sammlung selber einer bestimmten ordnenden Handlungsanweisung oder Plan folgt (Dreier 2025b, Rn. 17). Bei elektronischen Datenbanken kommt es vor allem auf die Durchsuchbarkeit an, wofür eine Aufbereitung der Daten mit Ordnungskriterien notwendig ist (Dreier 2025d, Rn. 7).

Ein Zugang zu den einzelnen Elementen wird dann als möglich angesehen, wenn die Sammlung ein geeignetes Zugangssystem umfasst, etwa ein Abfragesystem oder einen Index, welcher einen Einzelzugriff ermöglicht (Leistner 2020, Rn. 52). Ein solcher Index kann z. B. ein Inhaltsverzeichnis sein.

Die wesentliche Investition i.S. des § 87a UrhG muss sich erstens auf Beschaffung, Überprüfung oder Darstellung der Elemente beziehen. Damit kommt nicht jegliche Investition im Zusammenhang mit der Datenbank in Betracht. Nicht erfasst ist insbesondere die Generierung von Daten, sondern erst die nachgeordneten Schritte, da eben die Datenbank und nicht die Daten selbst geschützt wird (Zech 2015, S. 142 f.).[5]

Zu diesen nachgeordneten Schritten zählt etwa die Sammlung bereits anderweitig vorhandener Daten.[6] In Bezug auf die Aufzeichnung von Umweltdaten kann auf Grundlage der vom BGH angeführten Differenzierung sowohl für als auch gegen eine Berücksichtigungsfähigkeit argumentiert werden. Denn diese können einerseits als bereits vorhanden angesehen werden (Wiebe 2017, S. 341), andererseits auch als abgeleitet von natürlichen Phänomenen und damit erst neu erzeugt. Im Ergebnis ist das aber aufgrund der Bereichsausnahme in Art. 43 D-VO meist bedeutungslos.

Zweitens muss eine Investition vorliegen und diese wesentlich sein. Investition sind nicht nur finanzielle Aufwendungen, sondern jeder Einsatz von Zeit, Arbeit und Energie.[7] Die Investition ist wesentlich, wenn sie nicht ganz unbedeutend ist und nicht von jedermann leicht erbracht werden könnte.[8] Realistisch kann das allerdings nur im Einzelfall bewertet werden.

5 *BGH*, Urt. v. 1. Dez. 2010 (I ZR 196/08) – Zweite Zahnarztmeinung II, Rn. 19 mwN auch zur Rechtsprechung des EuGH.
6 *BGH*, Urt. v. 25. Mär. 2010 (I ZR 47/08) – Autobahnmaut, Rn. 19.
7 *EuGH*, Urt. v. 9. Nov. 2004 (C-444/02) – Fixtures Marketing, Rn. 44.
8 *BGH*, Urt. v. 1. Dez. 2010 (I ZR 196/08) – Zweite Zahnarztmeinung II, Rn. 23.

4.3 Lichtbildschutz

Auch ohne die für den Werkschutz nach UrhG eigentlich notwendige Schöpfungshöhe (dazu Kapitel 4.1) können Lichtbilder nach § 72 UrhG geschützt sein. Voraussetzung dafür ist einerseits die Schaffung eines „Urbilds“ durch Abbildung von etwas. Anders formuliert: Nicht ausreichend ist die bloße Reproduktion, etwa durch Kopie (Vogel 2020, Rn. 33; Lauber-Rönsberg 2024, Rn. 15-17). Andererseits muss die Erzeugung eines Mindestmaßes an Fertigkeiten bedürfen, beispielsweise durch die Bedienung der jeweiligen Vorrichtung (Vogel 2020, Rn. 34). Automatisierte Aufnahmen – beispielsweise durch Feldroboter – fallen damit nicht per se aus dem Schutzbereich heraus (Lauber-Rönsberg 2024, Rn. 34). Der Schutzumfang entspricht dann weitestgehend dem „normalen“ Urheberrecht (Lauber-Rönsberg 2024, Rn. 27).

In der Agrosystemforschung kann sich ein solcher Schutz v.a. bei Ablichtungen zwecks Bestimmung von Phänotypen ergeben. Ähnliches gilt für Mittel der Fernerkundung, seien es Aufnahmen per Satellit, Flugzeug oder Drohne (Katzenberger 1989, S. 118 f.).[9] Dabei kommt es nicht darauf an, ob ein Lichtbild im engeren Sinne entsteht, sondern nur, ob die „Aufnahme“ durch Aufzeichnung strahlender Energie erfolgt (Vogel 2020, Rn. 25 f.), sodass z. B. auch Radaraufnahmen erfasst sind.

4.4 Datenschutzrecht

Das Datenschutzrecht greift, wenn ein personenbezogenes Datum i.S. von Art. 4 Nr. 1 Datenschutz-Grundverordnung (DSGVO) vorliegt. Dafür braucht es eine Information, die sich auf eine natürliche Person bezieht. Entscheidend ist dabei, ob diese Person identifizierbar ist. Das ist dann der Fall, wenn sich eine Person zuverlässig aus einer Gruppe absondern lässt (Art.-29-Gruppe 2007, S. 14–24; Europäischer Datenschutzausschuss 2025, Rn. 16–25). Dafür sind Informationen zu berücksichtigen, welche die verarbeitende Stelle wahrscheinlich und vernünftigerweise nutzen würde (und könnte), um eine solche Absonderung vorzunehmen.[10]

Das Datenschutzrecht wird im Kontext von FAIRagro neben der Durchführung von Interviews vor allem im Falle von georeferenzierten Daten relevant, oder wenn Personen bei abbildenden Verfahren als „Beifang“ aufgenommen werden. Dafür kommt es aber wiederum auf den jeweiligen Einzelfall an. Auch ist zu betonen, dass auch bei einer Einordnung als personenbezogenes Datum die Verarbeitung der Daten möglich bleibt und dass die DSGVO die Verarbeitung für Zwecke der Forschung (erheblich) erleichtert.

9 *LG Berlin*, Urt. v. 30. Mai 1989 (16 O 33/89) – Satellitenfoto, GRUR 1990, S. 270.

10 *EuGH*, Urt. v. 9. Nov. 2023 (C-319/22) – Gesamtverband Autoteile-Handel, Rn. 45; *EuGH*, Urt. v. 19. Okt. 2016 (C-582/14) – Breyer, Rn. 44–48.

5 Rechtliche Fallstricke im Lebenszyklus

Im Folgenden werden die verschiedenen Fallstricke im Lebenszyklus eines Machine-Learning-Modells dargestellt. Dabei werden Urheberrecht, Datenbank- und Lichtbildschutz gemeinsam abgehandelt, da sie sich im Wesentlichen gleichen (§ 87b UrhG einerseits, § 72 Abs. 1 UrhG andererseits). Sofern sich Abweichungen ergeben sollten, werden diese kenntlich gemacht. Ein eigener Abschnitt ist der Frage gewidmet, ob das resultierende Modell eine Kopie der verwendeten Daten aus rechtlicher Sicht darstellt. Gleiches gilt für die Frage, ob sonst irgendwelche Rechte an dem Modell bestehen.

Da die verschiedenen Fallstricke an den Umgang mit den verschiedenen Datenarten anknüpfen, ergibt sich kein Unterschied aufgrund der Weise, auf die auf das Modell zugegriffen wird. Die Ausführungen gelten also sowohl für lokale als auch gehostete Modelle und auch für das Finetuning von bestehenden Modellen.

5.1 Urheberrecht

Für das Urheberrecht sind nur bestimmte Handlungen relevant, welche mit den Phasen im Lebenszyklus einhergehen. Diese Handlungen sind einerseits die Vervielfältigung, also die Kopie, möglicherweise urheberrechtlich geschützter Werke bei Datensammlung, Training und ggf. Validierung. Andererseits die Verbreitung, falls der verwendete Korpus weitergegeben werden sollte. Mit der Verbreitung verwandt ist die öffentliche Zugänglichmachung; diese betrifft grob umschrieben die Weitergabe über „neue“ Medien. Sollte eine dieser Handlungen vorgenommen werden, kann die:der Rechteinhaber:in Unterlassung (§ 97 Abs. 1 UrhG) und unter Umständen auch Schadensersatz (§ 97 Abs. 2 UrhG) verlangen. Dabei ist aber zu beachten, dass an den meisten Daten kein Urheber- oder verwandtes Recht besteht (s. Kapitel 4).

Vervielfältigung

Eine Vervielfältigung kann auf drei Weisen gerechtfertigt sein: Erstens durch eine Lizenz (etwa CC-BY 4.0), zweitens durch eine (konkludente) Einwilligung oder drittens durch gesetzliche Schranken. Bei der Lizenz kommt es auf die jeweiligen Bedingungen an.

Eine konkludente Einwilligung liegt dann vor, wenn die Zustimmung zu einer bestimmten Handlung nicht ausdrücklich erklärt wird, sich aus dem Verhalten der zustimmenden Person aber eindeutig und unmissverständlich ergibt, dass der Handlung zugestimmt wird. Dies kann beispielweise durch durch Zugänglichmachung des Werkes der:des Rechtein-

haber:in für bestimmte Nutzungen angenommen werden.[11] Auf eine Einwilligung wird für Text- und Datamining (TDM) aber meist nicht zurückgegriffen werden können. Denn Voraussetzung für die Einwilligung ist u. a., dass ein bestimmter Umgang mit einem Werk zu erwarten ist. Das ist für das Training von Modellen (bisher) nicht der Fall (Pesch und Böhme 2023a, S. 1007).

Falls diese beiden Möglichkeiten nicht in Betracht kommen, verbleiben die gesetzlichen Schranken. Gesetzliche Schranken ermöglichen eine Nutzung der geschützten Güter auch gegen den Willen der Rechteinhaber:innen. Sie haben zur Folge, dass bestimmte Nutzungen auch ohne Einwilligung des:der Urhebers:in erfolgen dürfen. Damit dies nicht zu einer extensiven Entwertung der Verwertungsrechte des:der Urhebers:in führt, sind den Schranken wiederum materielle Grenzen gesetzt (Kraetzig 2024, Rn. 46). Diese Grenzen ergeben sich etwa durch den *Dreistufentest*.

Danach sind Ausnahmen von den Verwertungsrechten der:des Urhebers:in nur gestattet, wenn sie (von Welser 2022, Rn. 22):

1. nur in bestimmten Sonderfällen angewandt werden,
2. die normale Verwertung des Werkes nicht beeinträchtigt wird und
3. die berechtigten Interessen des:der Urhebers:in nicht unzumutbar verletzt werden.

Dieser Dreistufentest wurde, zuletzt durch Art. 5 Abs. 5 der InfoSoc-RL[12], mehrfach verbindlich vorgeschrieben. Dennoch wurde er nicht explizit in nationales Recht aufgenommen, da die gesetzlich vorgeschriebenen Schranken an sich bereits die Voraussetzungen des Dreistufentests erfüllen (Kraetzig 2024, Rn. 45).

Im Folgenden sollen zwei der gesetzlichen Schranken hervorgehoben werden: § 44b und § 60d UrhG. Beide ermöglichen unter teilweise abweichenden Voraussetzungen TDM. Das Gesetz versteht darunter die Gewinnung von Informationen über einzelne Daten hinaus (Hofmann 2024, S. 170 f.). Damit ist auch das Training von KI-Modellen inkl. vorbereitender Handlungen erfasst, was sich auch mittelbar aus Art. 53 Abs. 1 lit. c KI-VO und deren ErwG 105 S. 3–7 ergibt (de la Durantaye 2024, Rn. 28; Pesch und Böhme 2023a, S. 1006).[13]

Dem Dreistufentest entsprechend ist TDM nicht allgemein zulässig, sondern an besondere Voraussetzungen geknüpft. Bei § 44b UrhG ist diese Voraussetzung die Flüchtigkeit der Vervielfältigung; bei § 60d UrhG ist die zusätzliche Voraussetzung die Beschränkung auf wissenschaftliche Forschungszwecke. Gemeinsame Voraussetzung beider Schranken ist

11 *BGH*, Urt. v. 11. Sept. 2024 (I ZR 140/23) – *Fototapete*, Rn. 36; *BGH*, Urt. v. 29. Apr. 2010 (I ZR 69/08) – *Vorschaubilder*, Rn. 86; *LG Hamburg*, Urt. v. 27. Sept. 2024 (310 O 227/23) – *LAION*, Rn. 86.

12 Richtlinie 2001/29/EG des Europäischen Parlaments und des Rates vom 22. Mai 2001 zur Harmonisierung bestimmter Aspekte des Urheberrechts und der verwandten Schutzrechte in der Informationsgesellschaft, ABl. L 167/10 vom 22. Juni 2001.

13 So auch *LG Hamburg*, Urt. v. 27. Sept. 2024 (310 O 227/23) – *LAION*, Rn. 79.

aber zunächst, dass der Zugang zu den Daten rechtmäßig erfolgt (Bomhard 2024b, Rn. 15). Dieser kann etwa auf einer entsprechenden Lizenz beruhen (welche aber etwa kein TDM zulässt) oder auch auf einer Einwilligung der:des Berechtigten. Eine solche kann sich (konkludent) durch öffentliche Zugänglichmachung im Internet ergeben.[14] Sofern ein Datensatz öffentlich abrufbar ist, kann also meist von einer entsprechenden Einwilligung ausgegangen werden.

§ 44b UrhG Daneben hat § 44b UrhG zwei weitere Voraussetzungen: Für die Daten darf kein *Nutzungsvorbehalt* existieren und die Vervielfältigungen sind zu *löschen*, wenn sie nicht mehr für das TDM benötigt werden.

Rechteinhaber:innen können Nutzungsvorbehalte für Daten bzw. Werke festlegen. Sofern die Daten online zugänglich sind, muss dieser Vorbehalt „maschinenlesbar" sein, § 44b Abs. 3 UrhG. Welche Anforderungen im Detail daran zu stellen sind, ist umstritten. Vorgeschlagen werden etwa ein Eintrag in der robots.txt, eine eigene ai.txt oder etwa Metadaten-Erweiterungen auf Grundlage von ODRL[15](Hamann 2024). In der Rechtsprechung wurde bereits ein Vorbehalt in natürlicher Sprache als ausreichend maschinenlesbar angesehen.[16]

Ein einheitlicher Standard hat sich bisher nicht herausgebildet, weshalb etwa auch die Wartung und Aktualisierung entsprechender Vorbehalte praktische Probleme zeitigt (Hamann 2024, S. 164 f.). Möglicherweise erarbeitet aber das auf der KI-VO beruhende Amt für künstliche Intelligenz der EU einen Standard, der eine gewisse Verbreitung erfahren könnte (Hamann 2024, S. 166).

Die Vervielfältigungen sind zu löschen, wenn sie nicht mehr für den Zweck des TDM benötigt werden. Darunter kann zwar auch die Validierung des Ergebnisses fallen, eine beliebig lange Speicherung ist davon aber nicht gedeckt (Bomhard 2024a, Rn. 20–23).

§ 60d UrhG § 60d UrhG regelt die Vervielfältigungen für TDM für Forschungszwecke. Der Paragraf kennt im Gegensatz zu § 44b UrhG weder einen Nutzungsvorbehalt noch müssen die Kopien gelöscht werden, wenn sie nicht mehr benötigt werden. Dafür ist der Kreis der Berechtigten auf Forschungsorganisationen (§ 60d Abs. 2 UrhG), Kulturerbe-Einrichtungen (§ 60d Abs. 3 Nr. 1 UrhG) und einzelne Forschende (§ 60d Abs. 3 Nr. 2 UrhG) beschränkt. Wissenschaftliche Forschung i.S. des ist jede methodische und systematische Tätigkeit mit dem Ziel des Erkenntnisgewinns in nachprüfbarer Weise (Dreier 2025c, Rn. 4). Hintergrund ist der besondere Schutz der Wissenschaft in der Rechtsord-

14 *BGH*, Urt. v. 11. Sept. 2024 (I ZR 140/23) – *Fototapete*, Rn. 36; *BGH*, Urt. v. 29. Apr. 2010 (I ZR 69/08) – *Vorschaubilder*, Rn. 86; so auch *LG Hamburg*, Urt. v. 27. Sept. 2024 (310 O 227/23) – *LAION*, Rn. 86.

15 *https://w3c.github.io/tdm-reservation-protocol/spec/*; *Besucht am 27. Mai 2025.*

16 *LG Hamburg*, Urt. v. 27. Sept. 2024 (310 O 227/23) – *LAION.*

nung, etwa in Art. 5 Abs. 3 S. 1 GG oder Art. 13 GRCh, weshalb der Begriff tendenziell weit ausgelegt wird.

Verbindendes Merkmal der verschiedenen Berechtigten ist eine nicht-kommerzielle Ausrichtung. Das Merkmal *nicht-kommerziell* wird im UrhG nicht genauer definiert. Die Frage, ob nicht kommerzielle Forschung vorliegt, wird stets anhand der konkreten wissenschaftlichen Tätigkeit beurteilt, nicht dagegen anhand der Organisation und Finanzierung der Einrichtung, in der die Forschung erfolgt (Bullinger 2022, Rn. 16).[17] Damit bleibt § 60d UrhG auch anwendbar, wenn kommerzielle Beteiligte Forschungshandlungen vornehmen oder von ihnen profitieren.[18] Die Einschränkung auf nicht-kommerzielle Forschung bedeutet auch nicht, dass keine Gewinnerzielung oder Kostendeckung mit den Forschungsergebnissen erfolgen darf. Erstere ist dann erlaubt, wenn die Gewinne in die Forschung reinvestiert werden (Bomhard 2024b, Rn. 17). Auch sind Kooperationen mit Angehörigen der Wirtschaft in sog. Public-Private-Partnerships möglich. Dies regelt explizit § 60d Abs. 2 S. 3 UrhG. Dieser erlaubt derartige Kooperationen, solange kein bestimmender Einfluss durch die Wirtschaftsakteure auf die Forschungsorganisation ausgeübt wird und der Wirtschaftspartner keine bevorzugten Zugang zu den Forschungsergebnissen bekommt (Bomhard 2024b, Rn. 22).

Über die Vervielfältigung hinaus ermöglicht § 60d Abs. 4 UrhG auch noch die *öffentliche Zugänglichmachung*. Sie ist sowohl hinsichtlich des Empfängerkreises als auch zeitlich begrenzt. Erlaubt ist die Zugänglichmachung entweder einem bestimmten abgegrenzten Personenkreis für die gemeinsame Forschung (Nr. 1) oder einzelnen Dritten zur Überprüfung der Forschungsqualität (Nr. 2). Die Zugänglichmachung ist zu beenden, sobald die gemeinsame wissenschaftliche Forschung, beziehungsweise die wissenschaftliche Überprüfung, abgeschlossen ist.

Verbreitung

Eine rechtlich relevante Verbreitung kann sich durch Weitergabe einer Datensammlung oder Verlinkung auf Daten ergeben (Linksammlung). Genauso wie die Vervielfältigung bedarf die Verbreitung einer Erlaubnis.

Weitergabe der Daten: Wie auch bei der Kopie der Daten kommen für die Rechtfertigung der Datenweitergabe entweder entsprechende Lizenzbestimmungen, eine Einwilligung oder gesetzliche Schranken in Betracht. Von den behandelten Schranken ermöglicht allerdings nur § 60d Abs. 4 UrhG die Weitergabe der Daten und dies auch nur unter engen Bedingungen.

17 *LG Hamburg*, Urt. v. 27. Sept. 2024 (310 O 227/23) – *LAION*, Rn. 118.
18 *LG Hamburg*, Urt. v. 27. Sept. 2024 (310 O 227/23) – *LAION*, Rn. 118.

So muss die Weitergabe entweder der gemeinsamen Forschung dienen oder aber der Überprüfung der Ergebnisse eines Forschungsvorhabens. Danach ist der Zugang zu beenden. Unter dem Vorzeichen der gemeinsamen Forschung kann ein Datensatz aber nicht allgemein öffentlich zugänglich gemacht werden. Vielmehr ist nur die Weitergabe innerhalb eines bestimmten und abgegrenzten Personenkreises zulässig (Bomhard 2024b, Rn. 28).

Verlinkung der Daten: Ein Korpus kann nicht nur durch Zusammenstellung der Daten geschaffen werden, sondern auch durch Verlinkung der zugrundeliegenden Daten. Darin liegt auch eine Möglichkeit, Datensätze, die auf Grundlage von § 44b UrhG genutzt werden (sollen), dauerhaft abrufbar zu halten.

Urheberrechtlich relevant (und damit potentiell erlaubnispflichtig) werden solche Linksammlungen, wenn durch die Verlinkung eine (weitere) öffentliche Zugänglichmachung i.S. des § 19a UrhG erfolgt. Sofern Inhalte nicht eingebettet, sondern nur verlinkt werden, handelt es sich meist nicht um eine erlaubnispflichtige Handlung. Denn durch eine solche Verlinkung wird nach Ansicht der Rechtsprechung kein neues Publikum erreicht. Etwas anderes gilt, wenn die Verlinkung auf ein rechtswidrig veröffentlichtes Werk zeigt. Eine (erneute) öffentliche Zugänglichmachung liegt nach Ansicht des EuGH in diese Falle vor, wenn der verlinkenden Person zumindest bewusst sein könnte, dass der Inhalt rechtswidrig eingestellt wurde. Ein solches Bewusstsein wird dabei vermutet, wenn die Verlinkung mit Gewinnerzielungsabsicht erfolgte (Dreier 2025a, Rn. 6b). Eine solche ist im Rahmen der Agrarforschung meist nicht anzunehmen.

Datenbanken

Besonderheiten ergeben sich für Datenbanken: Bei diesen sind die Vervielfältigung, Verbreitung oder öffentliche Wiedergabe nur dann erlaubnispflichtig, wenn sie einen wesentlichen Teil einer Datenbank betreffen. Ob ein Teil wesentlich ist, kann sich nach Umfang und Art des Teils bestimmen. Die Wesentlichkeit nach Umfang bestimmt sich nach der Größe der Datenbank und dem Verhältnis des entnommenen Teils dazu (Vohwinkel 2025, Rn. 13). Jedenfalls handelt es sich bei einer Entnahme von nur 10 % nicht um einen wesentlichen Teil hinsichtlich des Umfangs (Vohwinkel 2025, Rn. 13).

5.2 Datenschutzrecht

Wenn überhaupt das Datenschutzrecht in Bezug auf die Daten greift (dazu Kapitel 4.4), kann ein Training möglicherweise auf Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO, also ein berechtigtes Interesse gestützt werden (Europäischer Datenschutzausschuss 2024b, Rn. 59–135). Dafür kommt es aber auf den jeweiligen Einzelfall an (Übersicht zu den relevanten Kriterien bei Europäischer Datenschutzausschuss 2024a; Schantz 2025, Rn. 125). Dabei sollte

darauf geachtet werden, die Auswirkungen auf betroffene Personen möglichst zu reduzieren. Als eine solche Maßnahme schlägt der EDSA unter anderem Pseudonymisierung, die Berücksichtigung des urheberrechtichen Opt-Outs (dazu Kapitel 5.1) oder auch eine Riskoabschätzung dahingehend vor, ob Datenquellem möglicherweise personenbezogene Daten enthalten (Europäischer Datenschutzausschuss 2024b, Rn. 96–108).

Neben der Rechtmäßigkeit muss die Verarbeitung personenbezogener Daten weiteren Prinzipien gehorchen, welche in Art. 5 DSGVO näher beschrieben sind. Hervorzuheben ist dabei das Prinzip der Zweckbindung, welches sich auch auf das Training von Modellen auswirkt: Daten müssen für festgelegte, eindeutige und legitime Zwecke erhoben werden und dürfen nur dann weiterverarbeitet werden, wenn die Zwecke dieser Weiterverarbeitung kompatibel zu den Erhebungszwecken sind. Für die wissenschaftliche Forschung ist diese Zweckbindung gelockert (Art. 5 Abs. 1 lit. b HS. 2 Var. 1 DSGVO), wie auch andere Pflichten innerhalb der DSGVO. Allerdings gilt das nur, wenn die Zwecke der Verarbeitung nicht auch mit pseudonymisierten Daten erreicht werden können, außerdem müssen zusätzliche Maßnahmen getroffen werden, um die Daten zu schützen (Art. 89 DSGVO, ausführlich Weiss und Schuh 2024; Hüger 2024).

Ein kompatibler Zweck ersetzt dabei nach vorherrschender Auffassung nicht die Rechtsgrundlage für die Weiterverarbeitung; dabei kann es sich aber auch um die gleiche wie für die Erhebung handeln (Herbst 2024, Rn. 54). Idealerweise sollten allerdings keine personenbezogene Daten verarbeitet werden; sofern dies nicht ausgeschlossen ist, sollte das Risiko für die betroffenen Personen möglichst reduziert werden.

5.3 Modell als Kopie?

Sowohl aus urheber- als auch datenschutzrechtlicher Perspektive ist umstritten, ob ein Modell eine Kopie der zum Training verwendeten Daten darstellen kann. Urheberrechtlich ist das etwa bei Nutzung der Schranke in § 44b UrhG relevant (dazu Kapitel 5.1), da die Kopien nach Abschluss des TDM zu löschen sind, § 44b Abs. 2 S. 2 UrhG. Datenschutzrechtlich ist vor allem das Prinzip der Speicherbegrenzung zu beachten. Danach dürfen personenbezogene Daten nur so lange gespeichert werden, wie dies für die Verarbeitungszwecke notwendig ist, Art. 5 Abs. 1 lit. e DSGVO.

Verbindend ist dabei für beide Bereiche, wie weit die verwendeten Daten aus dem Modell extrahierbar sind. Entweder, weil so ein urheberrechtlich relevantes Werk extrahiert werden kann oder Informationen auf eine Person zurückgeführt werden können, diese also identifizierbar wird, wodurch die DSGVO anwendbar sein würde.

Dabei werden etwa Vergleiche mit (verlustbehafteter) Kompression angestellt, weshalb auch ein (generatives) Modell eine Kopie der verwendeten Daten darstellen würde (Pesch und Böhme 2023a, S. 1002-1005; ausführlich Dornis und Stober 2024, S. 74–80). Diese

Probleme ergeben sich nicht nur für generative Modelle, sondern auch für „klassische“ KI, da u. a. über sog. *Membership Inference Attacks* die verwendeten Trainingsdaten teilweise wiederhergestellt werden können (Käde und Maltzan 2020).

Gegen solche Ansätze wird zwar vorgebracht, dass Modelle nur Ableitungen und nicht die eigentlichen Daten speichern würden (de la Durantaye 2024, Rn. 19), sowohl Datenschutzrecht als auch Urheberrecht sind in dieser Hinsicht allerdings technikneutral: Solange ein Werk (irgendwie) rekonstruiert werden kann, liegt eine Vervielfältigung vor (Raue 2025, Rn. 18–21), solange eine Person aufgrund von Daten bzw. Rückgaben eines Modells identifiziert werden kann, ist es ein personenbezogenes Datum (ablehnend allerdings Hamburgische Beauftragter für Datenschutz und Informationsfreiheit 2024; differenzierend Europäischer Datenschutzausschuss 2024b).

Nicht nur aus rechtlicher Perspektive lässt sich noch nicht abschließend beurteilen, inwieweit Modelle als Kopien der verwendeten Daten betrachtet werden könnten. Auch Versuche, technische Benchmarks zu schaffen, stoßen aktuell noch an Grenzen (Guldimann u. a. 2025, S. 3 f., 8, 15). Sofern urheber- oder datenschutzrechtlich geschützte Daten zum Training verwendet wurden, sollte also im Zweifel davon ausgegangen werden, dass das Modell eine Kopie dieser Daten darstellen *könnte*, zumal die Entfernung von Informationen aus einem Modell (*machine unlearning*) quasi unmöglich ist (Pesch und Böhme 2023b, 919).

5.4 Rechte am trainierten Modell?

Neben Beschränkungen des faktischen Zugangs kann in der Wissenschaft wohl kein Schutz eines konkreten Modells (besser: der *Weights*) angenommen werden. Ein Schutz als Computerprogramm scheitert oftmals am fehlenden relevanten menschlichen Einfluss auf die konkrete Ausgestaltung (Schumacher 2023, S. 166–169), ein Schutz als Datenbank (§§ 87a–87e UrhG) scheitert an der fehlenden Unabhängigkeit der Elemente eines Modells; anders gesagt: Sie lassen sich nicht herauslösen, ohne, dass sie ihre Bedeutung verlieren (Kaulartz 2020, Rn. 14–18; 25; Schumacher 2023, S. 170). Vielmehr hängen sie von ihrem Kontext innerhalb des Modells ab. Damit verbleibt der Schutz als Geschäftsgeheimnis (Kaulartz 2020, Rn. 33 f. Kritisch dazu Schumacher 2023, S. 179–204). Dieser verträgt sich aber meist nicht mit der angestrebten möglichst hohen Transparenz innerhalb der Wissenschaft (Flöter und Cordes 2024, Rn. 10 f.).

6 Zusammenfassung und Ausblick

- An den meisten Daten besteht meist kein Urheberrecht.
- Einfache Messdaten sind auch in der Zusammenstellung als Datenbank nicht schutzfähig.

- Rechte können sich allerdings an Abbildungen ergeben.
- Datenschutzrecht greift meist nicht, kann aber in Einzelfällen relevant werden.
- Für die verwendeten Daten sollte eine Risikoabschätzung hinsichtlich Schutzrechten vorgenommen werden.
- Das Training von Machine-Learning-Modellen kann durch gesetzliche Ausnahmen („Schranken“) gerechtfertigt werden.
- Ob Modelle Kopien der verwendeten Daten darstellen, ist gegenwärtig stark umstritten und kann hier nicht abschließend beantwortet werden.
- An trainierten Modellen bestehen keine für die Forschung relevanten Schutzrechte.

Die rechtlichen Fallstricke sind auch in der Agrosystemforschung zahlreich. Den rechtlichen Sachverhalte inhärent ist auch, dass eine pauschale Antwort selten möglich ist und dass es stets auf den konkreten Einzelfall ankommt. Gleichzeitig erfordern rechtliche Beratungen einen signifikanten Kosten- und Zeitaufwand, der für Forschungsprojekte nicht stemmbar ist. Ein möglicher Lösungsansatz ist die Einrichtung eines Helpdesks, welcher mit Expertise in den jeweiligen Einzelfallbewertungen unterstützt. Daher umfasst der Helpdesk des FAIRagro-Konsortiums, das Data Steward Service Center (DSSC) auch eine Stelle, welche mit einer Volljuristin besetzt ist.

Danksagungen

Dieses Paper entstand im Rahmen des Projekts FAIRagro, welches von der Deutschen Forschungsgemeinschaft unter der Projektnummer 501899475 gefördert wird.

Beiträge

Constantin Breß und Lea Sophie Singson haben gleichermaßen zur Vorbereitung und Formulierung dieses Papers und des zugrundeliegenden Vortrags beigetragen.

Interessenkonflikte

Die Autor:innen erklären, dass in Bezug auf diese Arbeit keine potenziellen Interessenkonflikte bestehen.

Literaturverzeichnis

Antoine, Lucie. 2024. „Datenzugang im Spannungsfeld zwischen DSGVO, Geschäftsgeheimnisschutz und Datenbankherstellerrecht – Achillesferse des Data Act in der Praxis?“ *Computer und Recht,* 73–80. https://doi.org/10.9785/cr-2024-400204.

Art.-29-Gruppe. 2007. *Opinion 4/2007 on the concept of personal data,* Juni.

Bomhard, David. 2024a. „§ 44b UrhG“. In *BeckOK Urheberrecht,* 45. Aufl., herausgegeben von Horst-Peter Götting, Anne Lauber-Rönsberg und Nils Rauer.

———. 2024b. „§ 60d UrhG“. In *BeckOK Urheberrecht,* 44. Aufl., herausgegeben von Horst-Peter Götting, Anne Lauber-Rönsberg und Nils Rauer.

Bomhard, David und Jonas Siglmüller. 2024. „9.6 Datenlizenzvertrag“. In *Handbuch der IT-Verträge,* herausgegeben von Helmut Redeker. ISBN: 978-3-504-56008-9.

Bullinger, Winfried. 2022. „§ 60d UrhG“. In *Urheberrecht,* 6. Aufl., herausgegeben von Artur-Axel Wandtke und Winfried Bullinger. ISBN: 978-3-406-76200-0.

De la Durantaye, Katharina. 2024. „Nutzung urheberrechtlich geschützter Inhalte zum Training generativer künstlicher Intelligenz – ein Lagebericht“. *Zeitschrift für das gesamte Medienrecht,* 9–22. https://doi.org/10.9785/afp-2024-550103.

Dornis, Tim W. und Sebastian Stober. 2024. *Urheberrecht und Training generativer KI-Modelle: Technologische und juristische Grundlagen.* Nomos. ISBN: 978-3-7489-4955-8. https://doi.org/10.5771/9783748949558.

Dreier, Thomas. 2025a. „§ 19a UrhG“. In *Urheberrechtsgesetz,* 8. Aufl., herausgegeben von Thomas Dreier und Gernot Schulze. ISBN: 978-3-406-79302-8.

———. 2025b. „§ 4 UrhG“. In *Urheberrechtsgesetz,* 8. Aufl., herausgegeben von Thomas Dreier und Gernot Schulze. ISBN: 978-3-406-79302-8.

———. 2025c. „§ 60d UrhG“. In *Urheberrechtsgesetz,* 8. Aufl., herausgegeben von Thomas Dreier und Gernot Schulze. ISBN: 978-3-406-79302-8.

———. 2025d. „§ 87a UrhG“. In *Urheberrechtsgesetz,* 8. Aufl., herausgegeben von Thomas Dreier und Gernot Schulze. ISBN: 978-3-406-79302-8.

Europäischer Datenschutzausschuss. 2024a. *Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR.*

———. 2024b. *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models.*

———. 2025. *Guidelines 01/2025 on Pseudonymisation.*

Flöter, Benedikt und Christoph Cordes. 2024. „KI und Geistiges Eigentum: Neuronale Netze, Trainingsdaten und Output“. *Gewerblicher Rechtsschutz und Urheberrecht in der Praxis,* 668–670.

Glauner, Patrick. 2024. „Technical foundations of generative AI models“. *Legal-Tech-Zeitschrift,* 24–34.

Guldimann, Philipp, Alexander Spiridonov, Robin Staab, Nikola Jovanović, Mark Vero, Velko Vechev, Anna-Maria Gueorguieva u. a. 2025. „COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act“. Vorveröffentlichung, besucht am 16. Juni 2025. https://doi.org/10.48550/arXiv.2410.07959.

Hacker, Philipp. 2020. *Datenprivatrecht: Neue Technologien im Spannungsfeld von Datenschutzrecht und BGB.* Jus privatum, Band 244. Tübingen: Mohr Siebeck. ISBN: 978-3-16-159617-9.

Hamann, Hanjo. 2024. „Nutzungsvorbehalte für KI-Training in der Rechtsgeschäftslehre der Maschinenkommunikation“. *Zeitschrift für geistiges Eigentum* 16:113–168. https://doi.org/10.1628/zge-2024-0012.

Hamburgische Beauftragter für Datenschutz und Informationsfreiheit. 2024. *Diskussionspapier: Large Language Models und personenbezogene Daten.* Besucht am 17. Juni 2025. https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Diskussionspapier_HmbBfDI_KI_Modelle.pdf.

Hartmann, Thomas. 2022. „Forschungsdaten in den Naturwissenschaften: Eine urheberrechtliche Bestandsaufnahme mit ihren Implikationen für universitäres FDM“. In *E-Science-Tage 2021. Share your Research Data,* herausgegeben von Vincent Heuveline und Nina Bisheh. heiBOOKS. https://doi.org/10.11588/HEIBOOKS.979.C13728.

Herbst, Tobias. 2024. „Art. 5 DSGVO“. In *DS-GVO/BDSG,* 4. Aufl., herausgegeben von Jürgen Kühling und Benedikt Buchner. ISBN: 978-3-406-80263-8.

Hofmann, Franz. 2024. „Retten Schranken Geschäftsmodelle generativer KI-Systeme? Vortrag im Rahmen des Symposions »Generative K.I. und das Urheberrecht – Eine komplizierte Beziehung« des Instituts für Urheber- und Medienrecht am 10.11.2023 in München“. *Zeitschrift für Urheber- und Medienrecht,* 166–174.

Hüger, Jakob. 2024. „Die Rechtmäßigkeit von Datenverarbeitungen im Lebenszyklus von KI-Systemen“. *Zeitschrift für Digitalisierung und Recht,* 263–291.

Käde, Lisa und Stephanie von Maltzan. 2020. „Algorithmen, die nicht vergessen – Sicherheitslücken in Machine-Learning-Algorithmen und deren Bedeutung für den Schutz der Daten und der Urheberrechte“. *Zeitschrift zum Innovations- und Technikrecht,* 201–208.

Katzenberger, Paul. 1989. „Neue Urheberrechtsprobleme der Photographie – Reproduktionsphotographie, Luftbild- und Satellitenaufnahmen“. *Gewerblicher Rechtsschutz und Urheberrecht Internationaler Teil,* 116–119.

Kaulartz, Simon Apel Markus. 2020. „Rechtlicher Schutz von Machine Learning-Modellen“. *Recht Digital,* 24–34.

Kleinkopf, Felicitas. 2022. *Text- und Data-Mining: die Anforderungen digitaler Forschungsmethoden an ein innovations- und wissenschaftsfreundliches Urheberrecht.* 1. Auflage. Schriftenreihe des Archivs für Urheber- und Medienrecht, Band 300. Baden-Baden: Nomos. ISBN: 978-3-8487-7561-3.

Kraetzig, Viktoria. 2024. „vor § 44a UrhG". In *Urheberrecht,* 13. Aufl., herausgegeben von Axel Nordemann, Jan Bernd Nordemann und Christian Czychowski. ISBN: 978-3-17-043860-6.

Lauber-Rönsberg, Anne. 2024. „§ 72 UrhG". In *BeckOK Urheberrecht,* 44. Aufl., herausgegeben von Horst-Peter Götting, Anne Lauber-Rönsberg und Nils Rauer.

Lee, Katherine, A. Feder Cooper und James Grimmelmann. 2024. „Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain (The Short Version)". In *Proceedings of the Symposium on Computer Science and Law,* 48–63. CSLAW '24: Symposium on Computer Science and Law. Boston MA USA: ACM. Besucht am 25. Mai 2025. https://doi.org/10.1145/3614407.3643696.

Leistner, Matthias. 2020. „§ 4 UrhG". In *Urheberrecht,* 6. Aufl., herausgegeben von Ulrich Loewenheim, Matthias Leistner und Ansgar Ohly. ISBN: 978-3-406-72096-3.

Pesch, Paulina Jo und Rainer Böhme. 2023a. „Artpocalypse now? – Generative KI und die Vervielfältigung von Trainingsbildern". *Gewerblicher Rechtsschutz und Urheberrecht,* 997–1007.

———. 2023b. „Verarbeitung personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen". *Multimedia und Recht,* 917–923.

Raue. 2025. „§ 16 UrhG". In *Urheberrechtsgesetz,* 8. Aufl., herausgegeben von Thomas Dreier und Gernot Schulze. ISBN: 978-3-406-79302-8.

Richter, Heiko. 2024. „Art. 43 DA". In *BeckOK Datenschutzrecht,* 50. Aufl., herausgegeben von Heinrich Amadeus Wolff, Stefan Brink und Antje v. Ungern-Sternberg.

Schantz, Peter. 2025. „Art. 6 Abs. 1 DS-GVO". In *Datenschutzrecht,* 2. Aufl., herausgegeben von Spiros Simitis, Gerrit Hornung und Indra Spiecker gen. Döhmann. ISBN: 978-3-8487-8958-0.

Schröler, Roman und Christian Kuß. 2022. „E. Urheber-, Patent- und Markenrecht". In *Künstliche Intelligenz,* 1. Aufl., herausgegeben von Kuuya Chibanguza, Christian Kuß und Hans Steege. ISBN: 978-3-8487-7161-5.

Schulze, Gernot. 2022. „§ 2 UrhG". In *Urheberrechtsgesetz,* 7. Aufl., herausgegeben von Thomas Dreier und Gernot Schulze. ISBN: 978-3-406-77832-2.

Schumacher, Pablo. 2023. *Schutzfähigkeit Künstlicher Intelligenz: Die immaterialgüterrechtliche Schutzfähigkeit von KI-Technologien.* Trier University. ISBN: 978-3-758-43933-9.

Vogel, Martin. 2020. „§ 72 UrhG". In *Urheberrecht,* 6. Aufl., herausgegeben von Ulrich Loewenheim, Matthias Leistner und Ansgar Ohly. ISBN: 978-3-406-72096-3.

Vohwinkel, Moritz. 2024. „§ 87a UrhG“. In *BeckOK Urheberrecht,* 42. Aufl., herausgegeben von Horst-Peter Götting, Anne Lauber-Rönsberg und Nils Rauer.

———. 2025. „§ 87b UrhG“. In *BeckOK Urheberrecht,* 45. Aufl., herausgegeben von Horst-Peter Götting, Anne Lauber-Rönsberg und Nils Rauer.

Von Welser, Marcus. 2022. „§ 44a UrhG“. In *Urheberrecht,* 6. Aufl., herausgegeben von Artur-Axel Wandtke und Winfried Bullinger. ISBN: 978-3-406-76200-0.

Weidert, Stefan und Hannah Bug. 2023. „c) Rechte an Daten“. In *Handelsgesetzbuch,* 6. Aufl., herausgegeben von Volker Röhricht, Friedrich Graf von Westfalen, Ulrich Haas, Sebastian Mock und Heinz Wöstmann. ISBN: 978-3-504-45523-1.

Weiss, Mathias und Erik Schuh. 2024. „Die Zweckbestimmung und Zweckbindung als Weichenstellung für die DSGVO-konforme Nutzung von Daten für KI-Systeme“. *Zeitschrift für Digitalisierung und Recht,* 225–262.

Wiebe, Andreas. 2017. „Schutz von Maschinendaten durch das sui-generis-Schutzrecht für Datenbanken“. *Gewerblicher Rechtsschutz und Urheberrecht,* 338–345.

Zech, Herbert. 2015. „Daten als Wirtschaftsgut – Überlegungen zu einem „Recht des Datenerzeugers““. *Computer und Recht,* 137–146. https://doi.org/10.9785/9783504384579-002.

Resilient Hosting of Research Data Management Services

Halima Saker [1], Simon Pirkl [1], Jens Krüger [1], Holger Gauza [1], Suvasini Thangaraj [1], Ursula Eberhardt [1], Alexander Kirbis [1], Lucas Beuter [2]

[1] High Performance and Cloud Computing Group at IT Center, Eberhard Karls University Tübingen;
[2] IT Division, University Library, Eberhard Karls Universität Tübingen

High Availability (HA) environments achieve automatic failover to prevent data loss and minimize downtime while ensuring stability. The cost-effective and flexible HA setup avoids technology lock-in through its use of open-source components and virtual machine clusters. The system relies on two main components, which are the Pacemaker and Distributed Replicated Block Devices (DRBD). DRBD maintains real-time data replication, which results in backup servers obtaining exact copies of the primary server data. The Pacemaker system functions as a cluster manager to handle server cluster problems by performing automatic failover operations. The fencing mechanism of this system blocks simultaneous data access to prevent data corruption while maintaining data integrity. The HA setup provides scalability features that enable research environments to handle increasing data volumes, user numbers, and workflow demands through server additions or configuration modifications.

Keywords: Data integrity, Reproducibility and reuse, Digital sovereignty, Open source, High availability, RDM, RDM Service hosting

1 Introduction

The growing usage of digital platforms in the scientific research process (e.g., Navale, Kaeppler, and McAuliffe 2021; Gomes, Queiroz, and Ferreira 2020) requires the establishment of durable and available hosting systems for research data management services.

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23926 (CC BY-SA 4.0).

Users of online platforms for research data management consider access to research data to be a must-have feature (Reichenbach, Eberl, and Lindenmeier 2022). Therefore, institutions must ensure that data repositories, collaboration tools, and analytical platforms remain accessible at all times as research becomes more dependent on data. For research continuity and reproducibility purposes and to guarantee that data stays preserved over time, system integrity and availability serve as critical requirements rather than simple conveniences. Service disruptions can create severe research setbacks through data loss while also halting collaborations and damaging research results. Modern scientific workflows require an infrastructure that provides both reliability and resilience.

The continuity of RDM services at scientific research institutions, along with universities and data centers, encounters numerous challenges. The operational difficulties faced by scientific research institutions include system failures that occur without warning, together with network disruptions, hardware malfunctions, and routine maintenance activities, which all lead to service downtimes. Research fields including bioinformatics, climate science, and high-energy physics demand uninterrupted data processing; Thus, system downtime of even a short duration can critically affect experimental outcomes and data integrity. The growing implementation of open science practices requires research data to remain available for validation and replication studies as well as secondary analysis (Wilkinson et al. 2016). Research institutions that lack proper high-availability (HA) infrastructure may find themselves out of compliance with both funding agency expectations and established data stewardship standards, which refer to ongoing organizational and technical practices that ensure research data remains FAIR (Findable, Accessible, Interoperable, Reusable) throughout its lifecycle.

HA systems equip organizations with fault tolerance capabilities and data replication processes along with automated failover solutions, which work together to reduce downtime and prevent data loss. High-availability solutions play a critical role in maintaining the security and accessibility of research data across long durations. With proper implementation, an HA system automatically identifies failures and redistributes workloads to operational nodes without human assistance. Research organizations that face limited resources benefit significantly from this functionality because it eliminates unnecessary administrative tasks and preserves continuous operational performance.

Academic research environments encounter specific challenges because they handle multiple data types while dealing with changing computational requirements and restricted financial resources. These conditions require HA solutions which must be both flexible and affordable. The existing commercial HA systems prove too expensive and restrictive for public research applications because they do not meet long-term needs. This paper introduces an open-source HA architecture which addresses the requirements of research data management (RDM) at scale. Our tested solution maintains data integrity and availability for essential workflows through the implementation of Pacemaker and Distributed Replicated Block Device (DRBD). The selected tools meet academic requirements because they offer reliable performance alongside open-source licensing and affordable pric-

ing which supports digital sovereignty and reproducibility and sustainability in academic environments.

A promising approach to achieve HA in RDM services is to use Pacemaker and DRBD (LINBIT 2024), which are both open-source technologies that are used in enterprises and research networks. Pacemaker acts as a cluster resource manager that checks the health of the system and controls failover if there is a hardware or software failure (ClusterLabs 2023). DRBD offers data replication in real time between different machines, meaning that the research data will remain coherent and available in the case of a failure of the primary node. Altogether, these technologies create a strong infrastructure that prevents the risks of system failures and data corruption and, at the same time, allows for the necessary expansion of computational capacity to support growing research needs.

In addition to the availability of the data, its integrity is of great significance in scientific organizations (Lagoze 2014). Any alteration or degradation of the research data resulting from faulty failover mechanisms or unmonitored access can lead to wrong conclusions being made, thus compromising the validity of the scientific study. HA systems that employ fencing techniques such as quorum-based mechanisms and STONITH (Shoot The Other Node In The Head) prevent split-brain scenarios where two or more nodes try to write to the same portion of the data set at the same time. These guarantees are crucial to avoid the situation where there are two different versions of the data and to guarantee that there is only one latest version of the data at any point in time.

2 Related Work

Research data management systems receive multiple studies which examine their ability to maintain high availability (HA) and data integrity (Mesbahi, Rahmani, and Hosseinzadeh 2018). Research workflows require continuous platform access so researchers need robust infrastructure solutions to support this requirement (Navale and Bourne 2021; Wilkinson et al. 2016). Researchers have implemented multiple methods to achieve fault tolerance and minimize platform downtime in distributed platforms.

The distributed storage systems Ceph (Weil et al. 2006) and GlusterFS (Hedges, Hasan, and Powell 2013) implement data replication for HA but create performance issues that prevent their practical use in institutions with limited resources. DRBD (LINBIT 2024; Reisner and Ellenberg 2005) enables real-time block-level replication, which eliminates the requirement for distributed file systems.

The foundation of HA depends on automatic failover systems together with failure recovery functions. System resilience requires both load balancing and redundancy to function effectively (Ramy and Arafat 2019; Li, Zhang, and Wang 2020). The widely implemented enterprise solution Pacemaker (ClusterLabs 2023) lacks research-based exploration in aca-

demic environments. The research demonstrates Pacemaker's success as a scientific data hosting solution after its integration with DRBD.

Data integrity maintenance stands on the same level of importance as system availability. Node conflicts during split-brain events result in dangerous data version inconsistencies (Brown and Peters 2018; Santos and Lee 2021). STONITH fencing together with quorum-based decision-making provides protection against these risks (Zhao and Kumar 2017). The implementation of these techniques remains widespread in enterprise applications yet research IT infrastructure fails to adopt them. This work implements such safeguards to protect research data.

The current state of HA research for research data hosting shows limited development because most studies focus on general-purpose configurations. This research contribution fills this knowledge deficiency by creating an affordable and expandable HA system designed specifically for research institutions. A GitLab HA deployment case study within the de.NBI Cloud demonstrates both implementation difficulties and optimal implementation procedures.

Our open-source HA solution provides sustainable data availability with high data integrity in scientific environments.

3 Importance of Reliable Research Data Hosting

HA research data hosting functions as a solution to prevent interruptions in continuous research progress. Organizations that manage extensive datasets need to develop scalable redundant architectures to stop data loss and corruption, according to (Leitner, Huhn, and Scherer 2018).

The implementation of HA systems through failover mechanisms enables the system to recover from hardware, software, or network failures, which ensures continuous access to essential research tools, particularly in fields that require large-scale data processing, such as genomics and large-scale simulations (Gomes, Queiroz, and Ferreira 2020). Pacemaker, alongside redundancy functions, eliminates single points of failure by distributing the workload across multiple nodes while it automates failover procedures to minimize downtime (ClusterLabs 2023). The system also provides encryption and authentication features together with access restrictions to protect research data (Smith 2017). The strict data management policies from funding bodies and accreditation entities require institutions to implement robust HA strategies for maintaining their institutional credibility (Wilkinson et al. 2016). The research data availability and security needs of institutions are supported by HA solutions including Pacemaker and DRBD to maintain data integrity (LINBIT 2024).

4 Core Concepts of the HA System

The implementation of an HA system requires a clear framework of redundancy, failover, and data integrity (Endo et al. 2016). In scientific computing, irreproducible results are unpleasant, and data loss or corruption can lead to such results, costing time and damaging the credibility of research findings. To this end, an effective HA system design includes automated recovery strategies, real-time replication, and controlled access to prevent such risks.

4.1 High-Availability (HA) Clustering

HA clustering is the process of linking together several servers in order to make them work as one system such that if some components fail, the service will not stop. The main benefit of the clustering is failover for the purpose of redundancy to ensure that the workloads are redirected to the stand by nodes in case of failure. In an HA cluster, a monitoring system is supposed to be constantly running to check on the health of the active nodes and raise an alarm in case of any risk of failure and initiate the failover process accordingly.

Clusters can be classified as active-passive or active-active. The active-passive cluster is one where the role of the primary node is to manage the workload, and the standby nodes are passive and do not participate in the processing of the workload unless the primary node is unavailable. On the other hand, an active-active cluster shares the workload across multiple nodes at the same time for improved efficiency and load balancing. The difference between these two configurations is primarily governed by system needs, the existing infrastructure, and the nature of the research workflows that the system is expected to support.

4.2 Data Replication Strategies

HA systems require research data to be available and intact, which makes data replication critical. DRBD is a technology used to replicate data at the block level between multiple nodes, and DRBD is used in sync with Pacemaker to manage failover and to fail over control to a secondary node in case of a failure to avoid downtime. Geographic redundancy can enhance data resilience by spreading the replicas of the data across different locations to safeguard against failures that are localized. Asynchronous replication can be employed in low sensitive tasks to decrease the load, but synchronous replication with DRBD is preferable for research as it provides real time data consistency. In summary, DRBD and Pacemaker do a good job of replicating data to keep research data available, consistent, and secure.

4.3 Data Integrity Mechanisms

In addition to ensuring that HA systems remain available, strict data integrity controls must be implemented to prevent corruption and unauthorized modifications of data (Duggineni 2023). The split-brain scenario, where two nodes make different changes without coordination, calls for fencing to force a controlled, secure way of accessing critical data resources.

- **Quorum-based fencing**: This ensures that only a subset of nodes with sufficient voting power can change the data. This approach prevents two simultaneous write operations that can lead to inconsistencies if they happen at different times.
- **STONITH**: A last-resort mechanism that kills unresponsive nodes to prevent the cluster from becoming unstable. STONITH is most important for preventing bad processes from corrupting the synchronized datasets in some way.
- **Version Control and Checkpointing**: Some HA configurations also include version control systems that enable the researcher to go back to a previous version of the dataset in case of errors. This feature improves the reproducibility and the audit trail of scientific workflows.

4.4 Open-Source Architecture

A major strength of the proposed HA solution is that it is based on open source technologies. Proprietary systems bring in licensing expenses and possible vendor capture, whereas open-source frameworks are flexible, modular, and openly developed. Pacemaker, DRBD, and Corosync are integrated together to help research organizations develop their own HA configurations that suit their computing requirements. Thus, open source HA architectures can be further developed and improved, for example, in terms of development, security, and compatibility with other infrastructure components. The research computing environments must be flexible to grow with the dynamic nature of the scientific research, to support growing data sets, complex workflows and large collaboratory groups.

5 Core Technologies

Our setup includes DRBD for data replication, and Pacemaker for failover automation. The technologies were selected based on their maturity, open-source nature, and strong community support. Pacemaker and DRBD are widely adopted in high-availability clusters and integrate well into Linux environments. Corosync, used alongside Pacemaker, ensures reliable cluster communication and quorum decisions.

5.1 Pacemaker

Pacemaker functions as an open-source cluster resource manager which supports HA environments. The system tracks continuous health status and performs failovers through resource management and node failure detection and workload redistribution. Pacemaker functions as a core component to maintain service continuity by performing automatic application failovers to backup nodes which is essential for research environments that need accessible computational workflows and storage. The system operates with different cluster configurations including active-passive and active-active modes to maximize system reliability. Pacemaker implements complex decision-making through node weights and quorum calculations and fencing mechanisms to stop data corruption during failovers. The system uses DRBD integration to enforce strict data synchronization rules for maintaining data consistency while Corosync provides real-time cluster messaging to enhance failover efficiency.

5.2 Distributed Replicated Block Device (DRBD)

DRBD is a software-based networked storage replication solution that enables real-time data mirroring between nodes in an HA cluster. It operates at the block level, ensuring that changes made to one system are immediately replicated to another, reducing the risk of data loss during failures.

In research environments, DRBD is particularly valuable for preserving the integrity of scientific datasets. Unlike traditional backup solutions that rely on periodic snapshots, DRBD ensures that all write operations are immediately copied to a secondary node. This synchronization can be performed in synchronous, asynchronous, or dual-primary modes, each offering a balance between performance and consistency.

- **Synchronous Mode**: Guarantees that all data is written to both nodes before acknowledging a transaction. This mode ensures maximum data consistency but can introduce latency.
- **Asynchronous Mode**: Writes data to the primary node first and replicates it to the secondary node afterward. This approach reduces latency but carries a risk of minimal data loss in the event of failure.
- **Dual-Primary Mode**: Allows simultaneous read and write operations on both nodes, enabling load balancing. However, this mode requires advanced fencing mechanisms to prevent data corruption.

DRBD is crucial for maintaining uninterrupted access to research data, supporting applications that require real-time data consistency. When paired with Pacemaker, it enhances the overall resilience of scientific computing environments by ensuring seamless failover and disaster recovery capabilities.

5.3 Corosync

The cluster communication engine Corosync functions to synchronize messages between HA system nodes while maintaining node communication (Dake, Caulfield, and Beekhof 2008). The system effectively distributes state changes to maintain Pacemaker's current status. Corosync enables leader election and member management and failure detection in research computing environments. Pacemaker depends on reliable multicast messaging to achieve synchronized status updates because this functionality is essential for its failover process. The quorum-based consensus mechanism in Corosync prevents split-brain scenarios which prevents data inconsistencies. The system provides encryption for node-to-node communication which strengthens security in research environments with multiple nodes. The complete HA stack consists of Pacemaker, DRBD and Corosync which provides research data hosting with failover functionality and real-time replication and cluster coordination.

6 Architecture Overview & Failover Process

The architecture of an HA system is designed to prevent downtime and guarantee access to crucial research data. A good HA infrastructure is based on several parts, such as primary and secondary servers, network arrangements, and failover procedures. In this section, the elements of these architectures will be discussed in detail, and how they are interlinked to enhance the reliability and data integrity of the system in research environments.

6.1 System Components

A well designed HA system is a set of servers that are interconnected and work together in order to provide failover and data accessibility continuity. The main elements of the architecture are:

- **Primary Server**: The working node that acts as a point of contact for requests and data processing in real time. It acts as the primary channel through which users and other applications can access the research data.
- **Secondary Servers**: A hot standby node that keeps data in sync with the primary server using DRBD. In case of a failure, one of the secondary servers can easily take over control without any downtime.
- **Backup Server**: Extra sets of research data are stored on this server to increase the level of data redundancy in case of several failures. This server is a failback contingency plan in case of data loss.

- **Floating IP**: A method of using a single IP address that can be easily moved between servers in case of a failover. It enables users and applications to keep their connection intact without having to change settings manually.
- **Reverse Proxy**: It directs the network traffic and distributes the load among the available servers to increase the performance of the application.
- **Object Storage (S3 Ceph)**: Offered storage systems for big data sets employed in computational research and data mining purposes (MinIO, Inc 2023).

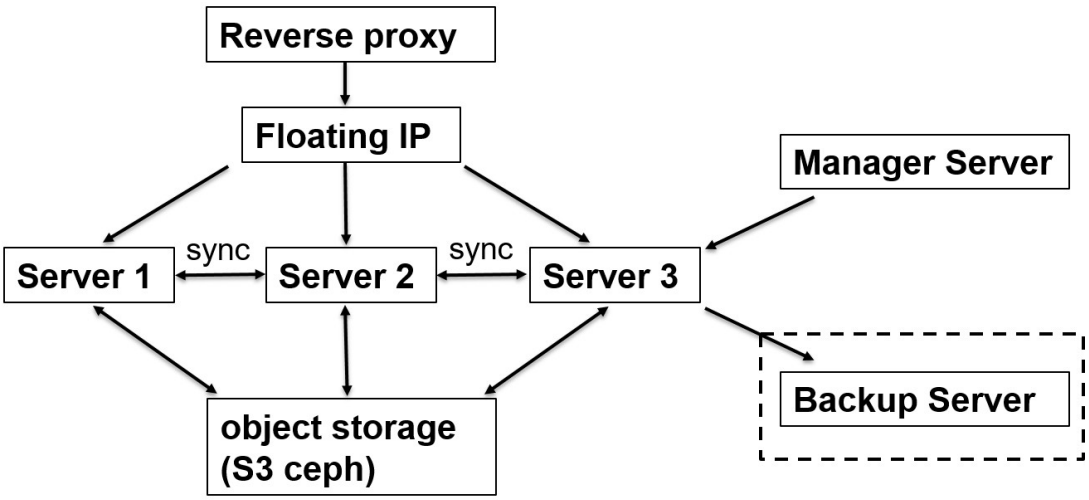


Figure 1: This diagram illustrates an HA system where multiple nodes are connected, ensuring redundancy and failover capabilities. A backup server is integrated to provide data protection and disaster recovery, enhancing system reliability and integrity.

6.2 Failover Mechanism

Failover is the crucial feature of an HA system that provides failure notification and subsequent failover of the primary server to a stand-by node without disrupting research. The failover process includes the following:

- **Failure Detection:** The Pacemaker is responsible for keeping a close eye on the primary server and its systems, including CPU load, memory consumption, and network connectivity. If the primary server is unavailable or dies, the Pacemaker initiates the failover process.
- **Resource Reallocation:** In case of failure, the Pacemaker shuts down the engaged node and makes the standby server the primary server. This ensures that all the data and applications are available to the users without any interruption.
- **Floating IP Reassignment:** The system reconfigures the floating IP address to the new primary server so that there is no break in communication for the researchers and other computational tools.

- **Data Synchronization:** First, DRBD checks that all the data on the new primary server is current and in sync with the last saved state to avoid data loss or corruption.
- **Fencing Activation:** To avoid the split-brain scenario, the Pacemaker starts the fencing mechanisms that deny the access to the critical storage resources to the active node.

Hence, the proposed model achieves a high level of availability by ensuring the continuous operation of the research environment through effective failover mechanisms.

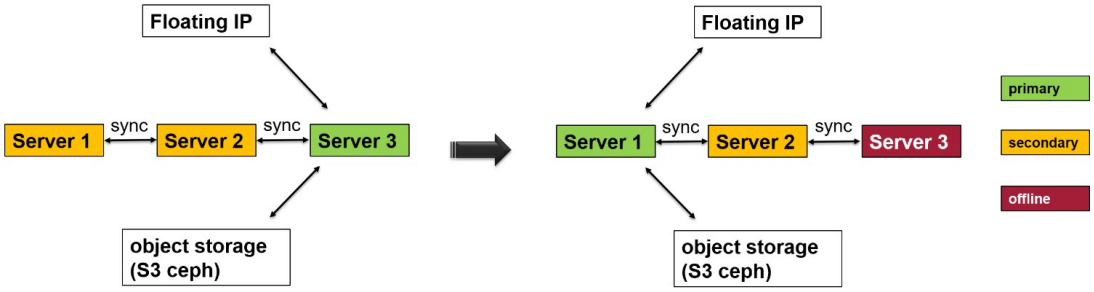


Figure 2: Failover Process.

6.3 Redundancy and Load Balancing

To increase the performance and robustness of the HA systems, the redundancy and load balancing mechanisms are incorporated. Data redundancy is ensured by DRBD, which is responsible for real-time replication of research data across multiple nodes to avoid a single point of failure. Load balancing is achieved as network requests are forwarded by the reverse proxy to active nodes, improving system performance and preventing the overload of a single server. Quorum-based decision-making is managed by Corosync, which makes failover decisions based on a majority consensus to prevent false positives and unnecessary transitions.

By integrating these architectural elements, research institutions can build a high-availability infrastructure capable of supporting intensive computational workloads while maintaining data integrity and security.

7 Network & Securing Integrity

Maintaining data integrity and network security is essential in HA research infrastructures. A good HA system not only provides failover but also protects research data from being corrupted, accessed inappropriately and cyber attacks. This section outlines the network

topology, fencing strategies, and security measures that are required to guarantee the data integrity of HA clusters.

Quorum-Based Fencing and Split-Brain Prevention: A major issue in HA designs is the avoidance of split-brain, where two nodes each take on the master role and try to push conflicting updates. Such incidents can lead to data errors, system failures, and possible degradation of research data. To address this issue, HA clusters use quorum-based fencing to prevent more than one node from accessing the shared storage at any given time (Jiménez-Peris et al. 2003).

Quorum mechanisms function by only allowing one node to take over control of the shared resources if it can get the support of the other nodes. When a node is isolated due to a network issue, quorum-based policies will stop it from making changes that could cause a mismatch. This ensures that only those nodes that are physically connected and actively participating can perform the read and write operations and thus avoid the situation of the cluster having two copies of data that are different.

Figure 3: Quorum-Based Fencing.

STONITH – Node Isolation: STONITH is a vital fencing mechanism that virtually shuts down failed or faulty nodes from the cluster to achieve data reliability (Robertson 2000). This approach entails powering cycle or forced shutdown of a node that is not responding properly within the cluster.

In this way, STONITH is used to prevent HA systems from being compromised by a node that is attempting to make changes to the shared storage while another node is operational. This ensures that even in case of software bugs, equipment failure, or network partitioning, the HA system remains functional and does not write data to the wrong location.

Secure Communication Among the Nodes: This paper highlights the importance of inter-node communication in HA clusters and the need to secure such communication against eavesdropping, tampering and unauthorized access. Today's HA systems encrypt their data traffic between the cluster nodes with TLS (Transport Layer Security).

Furthermore, role-based access control (RBAC) ensures that authentication is strictly enforced on who can change the configuration of the cluster. The use of multi-factor authentication (MFA) and public key infrastructure (PKI) improves the security by preventing

anyone from making changes to the HA environment without authenticating themselves properly.

Network Segmentation and Routine Auditing: HA clusters minimize attack risks because they operate through distinct networks which keep essential components in isolated subnets that block both external threats and lateral movement. Data integrity stands as an equally crucial factor in HA environments. The system uses checksum validation and automated integrity checks to synchronize data between primary and secondary nodes while scheduled integrity audits detect and correct potential errors or corruption.

When these security features are integrated into the RDM's HA infrastructure, it means that the infrastructure is secure from both internal and external threats that may cause damage to the integrity and confidentiality of research data.

8 Case Study: Implementation of GitLab HA Service in de.NBI Cloud

A highly available GitLab service by using HA clustering to achieve reliability, data integrity, and failover is offered on the de.NBI Cloud Tübingen. GitLab (GitLab B.V. 2024) is used extensively for research software development, version control, and collaboration, and so it becomes a critical platform in research data management that must be continuously available. To address these needs, the de.NBI GitLab HA setup incorporates several open-source technologies, such as Pacemaker, DRBD, Corosync, and S3-based storage replication.

High-Availability Architecture for GitLab: The GitLab HA cluster is running on several virtual machines (VMs), and the VMs are separated into different nodes of the architecture (Brandt et al. 2023). The main parts of the HA infrastructure are GitLab, DRBD, Corosync, and Pacemaker. GitLab provides repository hosting, CI/CD, and access control. DRBD guarantees the data synchronization between the primary and secondary nodes in real time to make sure that there is consistency. Corosync is used for failure detection and coordination between the cluster nodes and Pacemaker for managing the cluster resources, node failover, and fencing to achieve failover and availability in the system (e.g. GitLab B.V. 2023).

Data Replication and Backup Strategies: To prevent data loss and damage, the GitLab HA infrastructure has a strong multi-level backup and replication plan. The primary GitLab cluster is supposed to perform daily backups of the repositories, metadata,

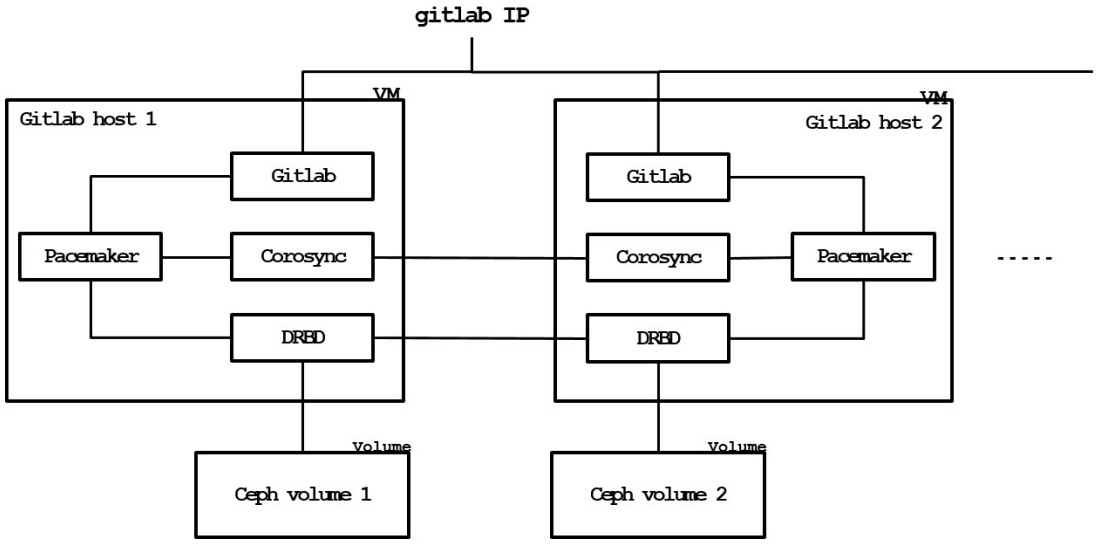


Figure 4: HA GitLab architecture designed with Pacemaker, Corosync, and DRBD for redundancy and failover, with Ceph storage for persistent data replication.

and application configurations. These backups are kept locally for a brief period of time to enable easy restoration. Furthermore, the backups are copied to an S3-based object storage system for long-term storage and a second S3 storage is used as a secondary copy. If the primary cluster fails, the last backup can easily be used to restore the secondary GitLab instance with minimal downtime and data loss.

Failover and Recovery Process: The failover and recovery process of the GitLab HA cluster is designed to avoid downtime and to maintain business process continuity. It is Corosync that keeps an eye on the health of each GitLab node, and in the event that the primary node becomes unresponsive, Pacemaker initiates the failover procedures. The GitLab secondary node is then able to take over the primary role, and the floating IP is moved to preserve the user's access. DRBD checks the data consistency and guarantees that the data is valid and ready to use on the new primary node. If required, the latest backup can be restored to the active node to include the most recent changes. When the failed node is returned online, it is first brought current with the active GitLab instance and then rejoined to the HA cluster.

Automation and Infrastructure-as-Code (IaC) Implementation: Manual configurations can lead to inconsistencies and involve a lot of administrative work. The use of Infrastructure-as-Code (IaC) tools like Ansible (Red Hat Inc. 2024) streamline the deployment processes, reproducibility, and reduce human error. Automated playbooks can be used to set up Pacemaker, Corosync, and DRBD based on predefined best practices. This enables the seamless configuration of GitLab HA instances and automatic integration of storage replication.

8.1 Real-World Application: GitLab and ARCmanager in NFDI4Plants

The resilient GitLab HA infrastructure described above is actively used to host the NFDI4Plants GitLab instance[1], which serves as the backend for the PLANTdataHUB. This instance enables researchers to store and manage their ARCs-structured, version-controlled collections of data and metadata.

In particular, the *ARCmanager* web service[2] connects seamlessly to this GitLab deployment. It allows users to create, browse, and edit ARCs directly in their browser while automatically syncing changes to the underlying Git repository. This setup offers a GUI-driven experience for researchers unfamiliar with Git workflows while maintaining the robustness and traceability of Git-based versioning. Importantly, thanks to the HA mechanisms (Pacemaker, DRBD, STONITH), the ARCmanager and its backing Git repositories remain continuously accessible and free from corruption even during server or network failures. As a result, plant science research projects that rely on these ARCs can proceed without disruption, ensuring transparency, traceability, and reproducibility.

9 Future Work & Next Steps

Although the application of HA systems with Pacemaker, DRBD, Corosync, and GitLab is very efficient in the preservation of service continuity in the de.NBI Cloud still has some possibilities for improvement and growth. As the research data management requirements are continuously developing, so does the infrastructure that supports them. This section highlights important aspects that can be improved further to increase the robustness, capacity, and security of HA deployments.

Expansion to Multi-Region High-Availability: Currently, the GitLab HA setup is currently running everything in a single cloud region. While intra-regional failover is good enough to cover most failures that are confined to a single region, adding multi-region replication would be another level of redundancy. This can be achieved by implementing geo-distributed DRBD clusters with asynchronous replication between regions, combined with multi-site Pacemaker clusters that include intelligent load balancing. Furthermore, deploying cross-region S3 storage replication will guarantee data consistency for backups, making the system even more reliable.

1 https://gitlab.nfdi4plants.de; *Visited by the editors on June 18, 2025.*
2 https://nfdi4plants.de/arcmanager/app/index.html; *Visited by the editors on June 18, 2025.*

Automated Deployment with Ansible Playbooks: To streamline the deployment and configuration of the HA cluster across various environments, we plan to develop and implement a comprehensive set of Ansible playbooks. This will reduce manual effort, ensure consistency, and enable rapid scaling and disaster recovery procedures.

Implementation of HA in Other Critical Services: Beyond GitLab, we aim to extend the high-availability setup to other essential research infrastructure services. This includes implementing HA for a reverse proxy (e.g., Nginx or HAProxy) to manage incoming traffic and distribute it across redundant service instances, as well as for identity and access management solutions like Keycloak. This will ensure end-to-end availability for the entire service stack.

Automated Scaling and Resource Optimization: Beyond failover, future work will investigate integrating dynamic scaling mechanisms with Pacemaker to automatically adjust computational resources based on real-time workload demands, particularly for bursty research computations. This would involve predictive analytics to anticipate resource needs and intelligent orchestration to optimize resource allocation, further enhancing efficiency and cost-effectiveness.

10 Conclusion

HA solutions are essential for ensuring the continuity of research data services. The growing dependence of research institutions on digital infrastructure makes protection against loss of data or corruption increasingly important. Pacemaker working with DRBD presents an economical and expandable open-source solution to address this problem. The HA setup provides research data accessibility by implementing node mirroring and fencing mechanisms to reduce corruption risks (LINBIT 2024; ClusterLabs 2023; Leitner, Huhn, and Scherer 2018). Research resources, including GitLab, maintain data integrity through this method, preventing simultaneous data access and ensuring data consistency (Wilkinson et al. 2016).

Research institutions can extend their infrastructure capacity by implementing this HA architecture through server additions or configuration modifications as their research needs expand. Extending this system is essential because it allows researchers to adapt their systems to the dynamic nature of their field. The utilization of open-source components eliminates vendor lock-in risks, which enables institutions to personalize their infrastructure needs while promoting collaborative development and community engagement (Sommestad and Karlsson 2014; Smith 2017).

The implementation of HA systems will likely spread throughout research institutions in the future. Research data management systems will gain increased resilience through automation tools for deployment and advanced monitoring and HA coverage extension to additional services (Bernstein and Smith 2019; Manerikar and Datta 2015). The continued growth of research data complexity requires sustained preservation of data integrity and availability to maintain scientific study reproducibility and trust in research outcomes. Research organizations can substantially benefit from HA solutions that use Pacemaker and DRBD to meet their present and future requirements while maintaining the dependability and accessibility of essential data (Navale and Bourne 2021; Gomes, Queiroz, and Ferreira 2020).

Acknowledgements

We acknowledge support for DataPLANT 442077441 through the German National Research Data Initiative (NFDI 7/1). We appreciate our co-fellows Dr.Adam Svahn, Mohamad Chehab, Amir Baleghi, and Fabian Paz for their valuable contributions to ensuring the High Availability of the Gitlab setup.

Authorship Contributions

Halima Saker drafted and wrote the manuscript. Simon Pirkl created and delivered the conference presentation. All authors reviewed and edited the manuscript.

Conflict of Interest

The authors declare that there are no competing interests or conflicts of interest related to the content or findings of this article.

Bibliography

Bernstein, David A., and Daniel J. Smith. 2019. "Automation in High-Availability Systems for Research Data". *Research Computing Systems Journal* 8:59–73. https://doi.org/10.1007/s00201-019-01168-7.

Brandt, Olaf, Holger Gauza, Jan Kaltenbach, Steve Kaminski, Jens Krüger, Fabian Paz, Saker Halima, Fabian Wannenmacher, Johannes Werner, and Thomas Zajac. 2023. *Distributed but Integrated: Forming a Science Gateway from Multiple Parts.* Zenodo. https://doi.org/10.5281/zenodo.7883191.

Brown, Alexander, and Daniel Peters. 2018. "Mitigating Split-Brain Scenarios in Distributed Systems: A Case Study". *IEEE Transactions on Distributed Systems* 29 (5): 1211–1223. https://doi.org/10.1109/TDS.2018.2814567.

ClusterLabs. 2023. *Pacemaker Documentation.* Visited on February 7, 2024. https://clusterlabs.org/pacemaker/doc/.

Dake, Steven C, Christine Caulfield, and Andrew Beekhof. 2008. "The corosync cluster engine". In *Linux Symposium,* 85:61–68. Citeseer.

Duggineni, Sasidhar. 2023. "Data integrity and risk". *Open Journal of Optimization* 12 (02). https://doi.org/10.4236/ojop.2023.122003.

Endo, Patricia T., Moisés Rodrigues, Glauco E. Gonçalves, Judith Kelner, Djamel H. Sadok, and Calin Curescu. 2016. "High availability in clouds: Systematic review and Research Challenges". *Journal of Cloud Computing* 5 (1). https://doi.org/10.1186/s13677-016-0066-8.

GitLab B.V.. 2023. *Reference architecture: up to 3,000 users | GitLab.* Https://docs.gitlab.com/ee/administration/reference_architectures/3k_users.html, Cited 22 Feb 2023.

———. 2024. *The most-comprehensive AI-powered DevSecOps platform | GitLab.* Visited on April 4, 2024. https://www.gitlab.com.

Gomes, Vitor C. F., Gilberto R. Queiroz, and Karine R. Ferreira. 2020. "An Overview of Platforms for Big Earth Observation Data Management and Analysis". *Remote Sensing* 12 (8). https://doi.org/10.3390/rs12081253.

Hedges, M. R., B. Hasan, and A. Powell. 2013. "A Performance Evaluation of the Gluster Parallel File System for Scientific Workloads". In *Proceedings of the International Conference on Data Management Technologies and Applications (DATA),* 82–92.

Jiménez-Peris, Ricardo, M. Patiño-Martínez, Gustavo Alonso, and Bettina Kemme. 2003. "Are quorums an alternative for data replication?" *ACM Transactions on Database Systems* (New York, NY, USA) 28 (3): 257–294. https://doi.org/10.1145/937598.937601.

Lagoze, Carl. 2014. "Big Data, data integrity, and the fracturing of the control zone". *Big Data & Society* 1 (2). https://doi.org/10.1177/2053951714558281.

Leitner, Philipp, Dietmar Huhn, and Klaus Scherer. 2018. "High-Availability Clusters in Cloud Computing Environments". *Journal of Cloud Computing: Advances, Systems and Applications* 7 (1): 1–16. https://doi.org/10.1186/s13677-018-0116-9.

Li, Yuan, Wei Zhang, and Hongbo Wang. 2020. “A Load Balancing Strategy for Cloud-Based High-Availability Systems”. *IEEE Transactions on Cloud Computing* 8 (3): 754–766. https://doi.org/10.1109/TCC.2019.2928371.

LINBIT. 2024. *DRBD.* Visited on February 7, 2024. https://linbit.com/drbd/.

Manerikar, Ashwin, and Rajesh Datta. 2015. “Improving Resilience in Research Data Management through Automation”. *Journal of Computational Research and Data Management* 6 (2): 45–60. https://doi.org/10.1145/2832578.

Mesbahi, Mohammad Reza, Amir Masoud Rahmani, and Mehdi Hosseinzadeh. 2018. “Reliability and high availability in cloud computing environments: a reference roadmap”. *Human-centric Computing and Information Sciences* 8 (1): 20. https://doi.org/10.1186/s13673-018-0143-8.

MinIO, Inc. 2023. *MinIO | High Performance, Kubernetes Native Object Storage.* Visited on March 19, 2023. https://min.io.

Navale, Vivek, and Philip E. Bourne. 2021. “Cloud computing applications for biomedical research”. *Methods in Molecular Biology* 2190:3–26. https://doi.org/10.1007/978-1-0716-0826-5_1.

Navale, Vivek, Denis von Kaeppler, and Matthew McAuliffe. 2021. “An overview of biomedical platforms for managing research data”. *Journal of Data, Information and Management* 3 (1): 21–27. https://doi.org/10.1007/s42488-020-00040-0.

Ramy, Mohamed, and Ahmed Arafat. 2019. “High-Availability and Load Balancing in Cloud Computing: Challenges and Solutions”. *International Journal of Cloud Computing* 9 (2): 112–127.

Red Hat Inc. 2024. *Ansible. Automation for everyone.* Visited on February 8, 2024. https://www.ansible.com/.

Reichenbach, Rebecca, Christoph Eberl, and Jörg Lindenmeier. 2022. “Online platforms for research data: A requirements and cost analysis”. *Science and Public Policy* 49 (4): 598–608. https://doi.org/10.1093/scipol/scac011.

Reisner, Philipp, and Lars Ellenberg. 2005. “DRBD: Distributed Replicated Block Device”. *Linux Journal* 2005 (145): 18–23.

Robertson, Alan. 2000. “Resource Fencing Using STONITH”. In *Proceedings of the High-Availability Linux Project.* Broomfield, Colorado: IBM Linux Technology Center.

Santos, Roberto, and Kevin Lee. 2021. “Ensuring Data Integrity in High-Availability Cloud Systems”. *Journal of Cloud Security* 6 (3): 221–237.

Smith, George R. 2017. “Benefits of Open-Source Software in High-Availability Research Data Management”. *Journal of Open Research Software* 5:1–7. https://doi.org/10.5334/jors.188.

Sommestad, Tomas, and Jonny Karlsson. 2014. “The Role of Open-Source Software in High Availability Systems”. *International Journal of Computer Science and Information Security* 12 (5): 18–25.

Weil, Sage A., Andrew W. Leung, Scott A. Brandt, and Ethan L. Miller. 2006. *Ceph: A Scalable, High-Performance Distributed File System.* 307–320. USENIX Association.

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. https://doi.org/10.1038/sdata.2016.18.

Zhao, Wei, and Rajesh Kumar. 2017. “Fencing Mechanisms in Cloud-Based HA Clusters: STONITH and Beyond”. In *Proceedings of the IEEE International Conference on Cloud Computing Technology,* 150–158.

Efficient Data Curation by Automating Workflows

Florian Fritze [1,2], Dorothea Iglezakis [1,2,*], Sarbani Roy [1,2], Björn Selent [1,3,*], Karoline Weinspach [2]

[1] Research Data Competence Center, University of Stuttgart;
[2] University Library, University of Stuttgart;
[3] Technical Information and Communication Services, University of Stuttgart;
[*] Corresponding author: fokus@izus.uni-stuttgart.de

Data curation results in high-quality data description, but takes time and effort. To still enable an intensive data curation process at the institutional data repository *DaRUS*, different approaches support and automate the process: to support correct and standardized input as early as the metadata entry stage, the researchers are assisted by interfaces to registries like ORCID for authors, ROR for institutions, to terminology services for a topic classification and to the institutional research information system for research projects. Automatic checks for keywords, URLs and funding information are embedded in a REST API named pubWorkflow to help the curation team and manage the publication workflow. An integration with easyReview is planned to also support a scientific quality check.

All integrations and tools build on the workflow engine and the controlled vocabulary support of Dataverse, the underlying repository platform.

Keywords: Data Curation, Metadata, Data Repository, Automation

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23927 (CC BY-SA 4.0).

1 DaRUS – The Data Repository of the University of Stuttgart

The University of Stuttgart's research data repository *DaRUS* is based on the open source software Dataverse[1]. Dataverse originates from the Institute for Quantitative Social Science (IQSS) of Harvard University which actively maintains the project and leads its ongoing development. A strong international community comprising 125 worldwide institutional installations and the Global Dataverse Community Consortium (GDCC)[2] vouches for Dataverse being a lasting solution in terms of research data management. *DaRUS* allows researchers of the University of Stuttgart to share their data with both the scientific community and public while keeping full control of the data. Dataverse's features include the possibility to organize and structure data and access to it, to enrich the data with descriptive metadata and eventually to publish it to the world. From the onset the project aimed to provide a tool for research data management that strongly adheres to the *FAIR* principles, i.e. to make the data *F*indable, *A*ccessible, *I*nteroperable and *R*eusable (Wilkinson et al. 2016).

To better meet the needs of the University of Stuttgart, individual adaptations to the upstream project have been implemented for *DaRUS*. Besides minor changes to the look and feel of the software and the handling of custom mime types of uploaded files, the changes make extensive use of Dataverse's inbuilt extension methods in order to add supplementary functionality. These methods are in particular the option to alter and add the schemata defining the metadata fields which are used to enrich the uploaded data, the possibility to hook external tools to Dataverse, the linking of external vocabularies which can be searched for suitable entries to fill the metadata fields and a workflow mechanism which can be used to call external actions prior to publishing and after successful publishing res.

In *DaRUS* external tools are used to preview a wide range of mime types, such as images, videos, tabular data or text data. Even zip-archives can be inspected by an external tool previewer. Individual discipline specific metadata blocks have been developed in Stuttgart and added in order to enter pinpoint information for data stemming from engineering or chemistry research. These metadata blocks basing on schemata and standards like *EngMeta* (Schembera and Iglezakis 2020; Seeland 2020), *EnzymeML* (Lauterbach et al. 2023) and *CodeMeta* (Jones et al. 2017; Iglezakis 2023), proved to be useful and thus have been adopted by other Dataverse installations, too. Furthermore, a block has been added which helps to describe settings and requirements to run simulations and experiments. The block containing relevant information for citation has been extended by project and funding information. These fields are already linked to an external vocabulary in order to assist with filling the data. More detailed information about this feature will be given

1 https://www.dataverse.org; *Visited on March 28, 2025.*
2 https://www.gdcc.io; *Visited on March 28, 2025.*

in section 2.1. Finally, *DaRUS* also uses an external workflow prior to publishing which will be described in section 2.2.

The combined strength from a solid upstream code base and a vibrant community makes Dataverse a sound tool for scientific RDM in general and a valued assistant to the scientists of the University of Stuttgart in particular. Since its introduction in 2019 *DaRUS* gained significantly in popularity as can be seen in Figure 1(a). A growing number of institutes of the University of Stuttgart own their Dataverse collection in *DaRUS* and keep creating an increasing number of datasets. Alongside with the increasing number of created datasets the number of published datasets grew constantly over the last years as shown in Figure 1(b). Thus, the repository plays a key role in both enhancing open science and accrediting the work involved with generating the data.

The researchers also notably recognize the added value of data containing rich and correct metadata and are ready to provide all necessary information. Nevertheless, it still means an additional effort to fill datasets with this metadata and not always is it possible to include dataset creation in procedural workflows. Numerous metadata also leads to an additional workload on the data curation level. Thus, in order to overcome this side effect of the increased use of *DaRUS*, several means have been introduced into *DaRUS* to ease and automate the workflows involved with dataset management for both the researchers and the curators.

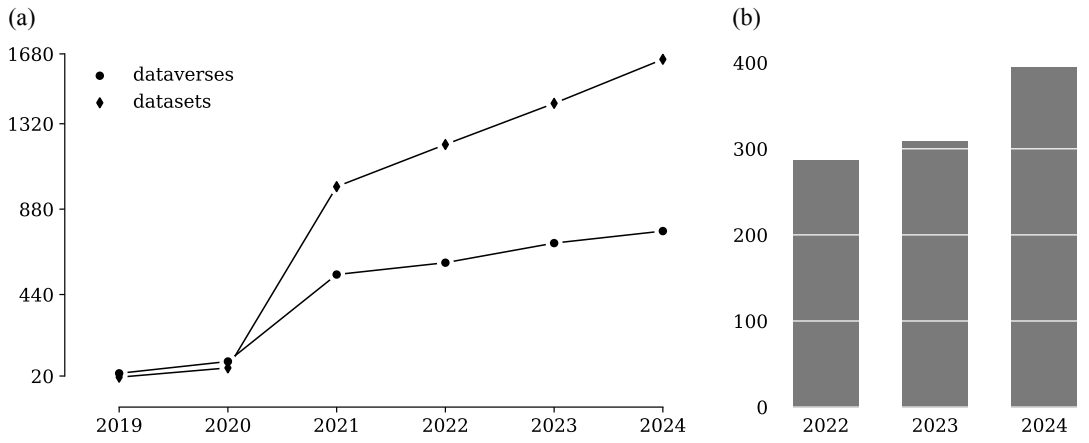


Figure 1: *DaRUS* usage (a) and number of published datasets (b).

1.1 The Publication Process

The publication process for datasets in *DaRUS* consists of several stages which involve different actors and actions. An overview is presented in Figure 2. The process starts with an author creating an initial dataset draft either by using the GUI or the API of *DaRUS*. At this stage only a limited number of metadata is required in order to do so successfully and the dataset is unconditionally editable. Once the author is inclined to publish the

dataset a Content Review process is triggered and a curator is notified. The curator is usually a peer of the author and the review at this stage is meant to contain a scientific discipline-specific survey of the data. The data is locked. If the data needs some form of alteration, the curator will unlock the dataset and return a review report with remarks to the author who will need to incorporate the changes. If the data passes the content review, the curator will start the publication process that involves the University's library staff for a formal and semantic review. The dataset is locked at this stage again until it either is returned to the author for major changes or unlocked for small changes by the curation team. If all parties involved are content with the review's outcome, the dataset is finally released and sent to the university bibliography, its PID is registered, and a fixed version of the dataset is published.

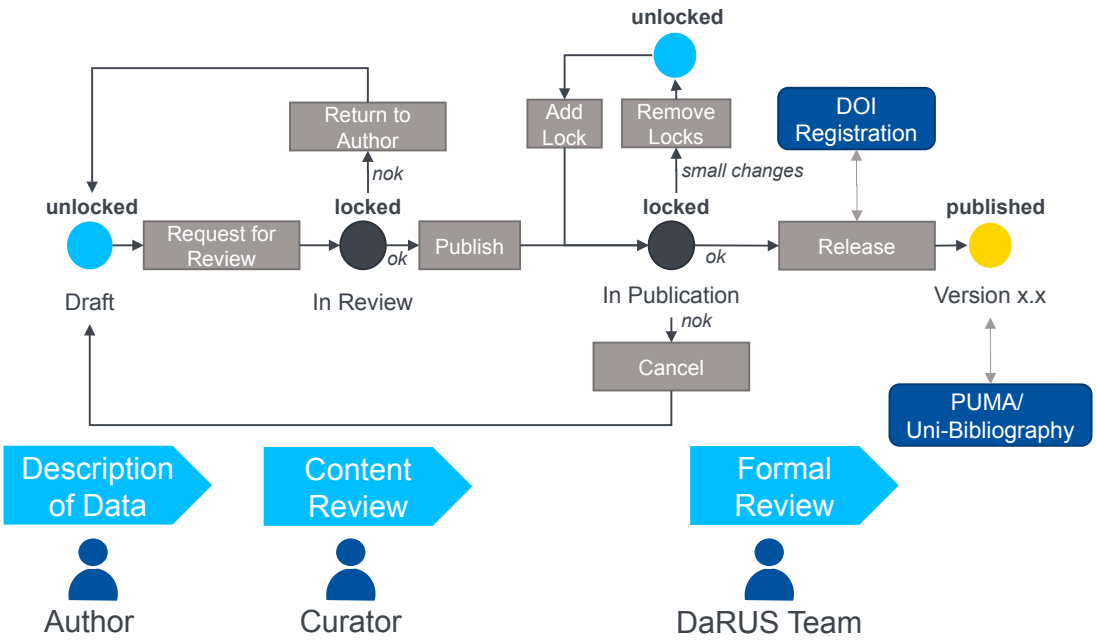


Figure 2: Publication workflow of DaRUS.

1.2 Data Curation at DaRUS

In this formal review, each dataset on *DaRUS* is checked for comprehensibility, reusability, interoperability and archivability[3]. To be *comprehensible*, the content and structure of the dataset should become clear without any further external information. The authors should describe the meaning of the files of the dataset, explain used abbreviations in file names, and make clear, how the data or code can be used. This information is expected in human-readable form in the description of the dataset and/or in a README file. The curators also look for and correct typos and enhance the readability. Figure 3 shows an example of a considered good dataset description.

3 https://www.izus.uni-stuttgart.de/en/fokus/darus/qualitycontrol/; *Visited on March 28, 2025.*

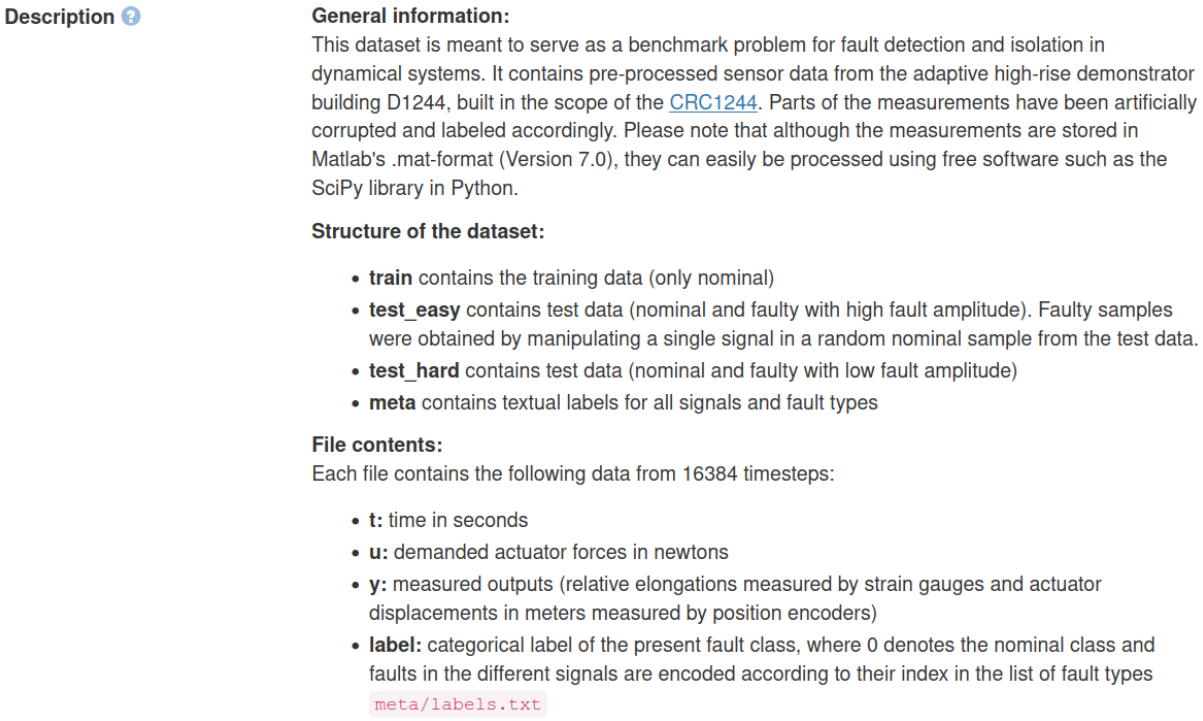


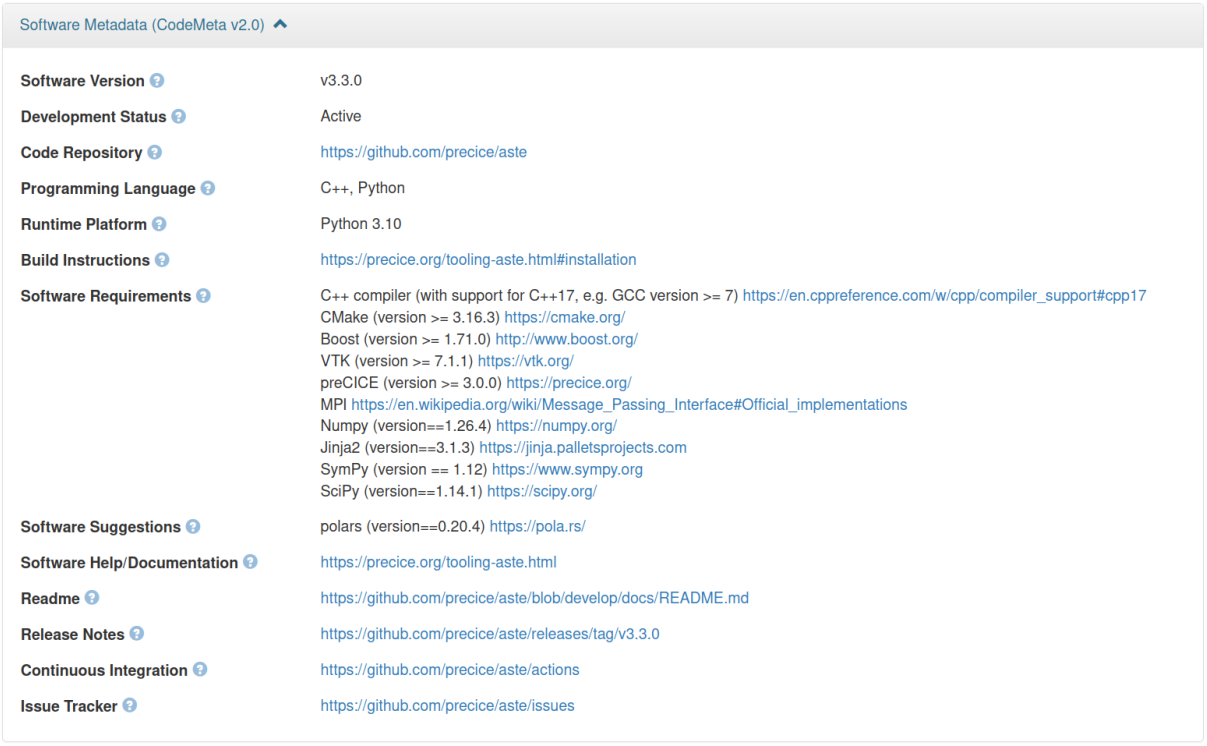
Figure 3: Exemplary dataset description.

To be *reusable*, a description of the tools and methods used to generate the data or necessary to open the data is essential. The latter is especially important for data in a proprietary data format that can only be read with specific software. The process metadata block basing on EngMeta provides the necessary metadata fields, for which completion is requested if applicable.

If the dataset contains primarily software or scripts rather than data, the authors are encouraged to provide software specific metadata via a CodeMeta[4] based metadata block. Metadata fields are the necessary dependencies with their version, CPU or memory requirements or links to documentation for example. This information is supposed to ease reuse of the software or to make the code run even after a few years have passed and the original users are unavailable (see Figure 4 for an example).

In addition, *DaRUS* users are assisted in converting their data in an open format, especially to transform tabular data in a specific csv format that enables Dataverse to transform the data in an archival format to make it available in different formats and offer the aforementioned preview functionalities for the data.

4 https://codemeta.github.io/index.html/; *Visited on March 30, 2025.*



Software Metadata (CodeMeta v2.0)

Software Version	v3.3.0
Development Status	Active
Code Repository	https://github.com/precice/aste
Programming Language	C++, Python
Runtime Platform	Python 3.10
Build Instructions	https://precice.org/tooling-aste.html#installation
Software Requirements	C++ compiler (with support for C++17, e.g. GCC version >= 7) https://en.cppreference.com/w/cpp/compiler_support#cpp17 CMake (version >= 3.16.3) https://cmake.org/ Boost (version >= 1.71.0) http://www.boost.org/ VTK (version >= 7.1.1) https://vtk.org/ preCICE (version >= 3.0.0) https://precice.org/ MPI https://en.wikipedia.org/wiki/Message_Passing_Interface#Official_implementations Numpy (version==1.26.4) https://numpy.org/ Jinja2 (version==3.1.3) https://jinja.palletsprojects.com SymPy (version == 1.12) https://www.sympy.org SciPy (version==1.14.1) https://scipy.org/
Software Suggestions	polars (version==0.20.4) https://pola.rs/
Software Help/Documentation	https://precice.org/tooling-aste.html
Readme	https://github.com/precice/aste/blob/develop/docs/README.md
Release Notes	https://github.com/precice/aste/releases/tag/v3.3.0
Continuous Integration	https://github.com/precice/aste/actions
Issue Tracker	https://github.com/precice/aste/issues

Figure 4: Exemplary metadata documentation of a research software set.

To be *interoperable*, the metadata should be linked to persistent identifiers (ORCID[5] for authors and ROR IDs[6] for organizations, DOIs for links to text publications and other datasets) and use controlled vocabularies for subject indexing like keywords and topic classification.

The curators also ensure that all used URLs resolve at least at the time of publication and that the contact information is long-term reachable.

Each dataset is curated by two different persons, one from the library publication team and one from the data management team. The curation effort for each dataset including all communication with the authors range between 10 minutes and several hours, depending on the state of completeness of the available metadata and the experience of the authors. The effort is rewarded by high-quality data sets and the authors' quick learning. Data curation serves as RDM consulting through the back door, so to speak. However, as the volume of publications increases, the workload on the curation team also rises steadily.

5 https://orcid.org; *Visited on March 30, 2025.*
6 https://ror.org; *Visited on March 30, 2025.*

2 Automating workflows

Two different approaches are used to keep the curation effort as low as possible and thus make the publication workflow scalable: On the one hand, support for authors through interfaces to external PID and vocabulary services, and on the other hand, automated checks to support curators and reviewers.

2.1 User Input Support by Interfaces

Efficient metadata management is crucial for ensuring the discoverability, interoperability, and consistency of research data. To achieve this, *DaRUS* integrates third-party vocabulary and persistent identifier (PID) services using Dataverse's flexible external vocabulary support mechanism (Myers and Tykhonov 2023)[7]. This framework allows customized scripts and field-specific JSON configurations to define how metadata fields interact with external services and vocabularies, thereby streamlining data entry and enhancing metadata quality.

Integration of ORCID and ROR for Researcher and Institution Identification: As part of this functionality, *DaRUS* connects with ORCID and ROR databases to enable an autofill feature for relevant metadata fields. ORCID (Open Researcher and Contributor ID) provides unique identifiers for researchers, ensuring precise attribution of research outputs and affiliations. Similarly, the Research Organization Registry (ROR) assigns persistent identifiers to research institutions, standardizing institutional affiliations in research metadata. Through this integration, users can search for an author's name to automatically retrieve and populate the corresponding ORCID ID, while searching for a university or organization suggests and autofills the appropriate ROR ID (cf. Figure 5). This feature enhances the accuracy and consistency of metadata related to researchers and institutions.

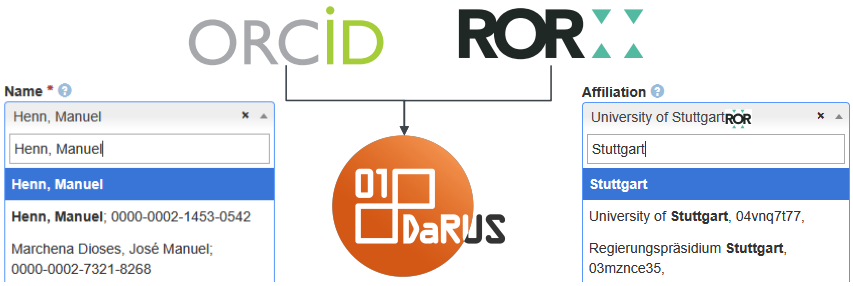


Figure 5: Interface to ORCID and ROR.

Topic Classification Using the TIB Terminology Service: To further standardize metadata, *DaRUS* incorporates topic classification based on the DFG Subject Classi-

7 https://github.com/gdcc/dataverse-external-vocab-support; *Visited on March 30, 2025.*

fication Ontology (2024–2028). Unlike broad subject classification (e.g., "Physics" vs. "Quantum Optics") or free-text keywords, which lack standardization, topic classification uses predefined, structured categories that enhance metadata consistency, discoverability, and interoperability across research repositories.

To implement this, *DaRUS* integrates the TIB Terminology Service[8], a REST API maintained by the German National Library of Science and Technology (TIB) that provides authoritative classifications and controlled vocabularies. Through this integration, users can search for a research topic, and the system suggests appropriate classification terms from the DFG Subject Classification Ontology. The controlled vocabulary name and the corresponding URI are then automatically populated in the metadata fields as depicted in Figure 6, ensuring consistency and standardization in subject classification.

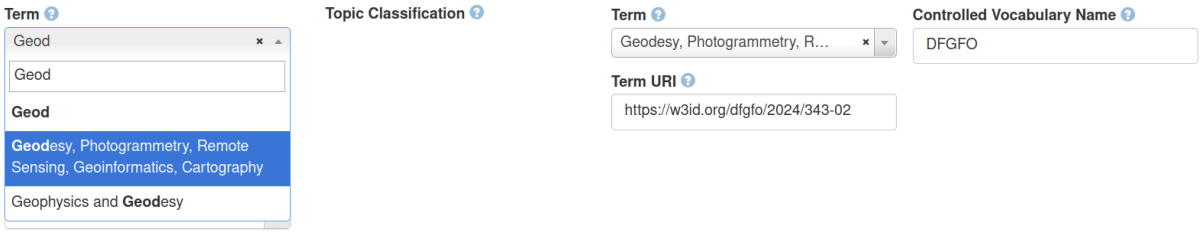


Figure 6: Interface to the terminology service to check for topic classification.

Standardizing Metadata for Project and Funding Information: Building upon these enhancements, metadata related to project details and associated funding information has also been standardized within *DaRUS*. A major challenge was the structural limitation of Dataverse, which only supports two hierarchical levels in its metadata framework. This constraint required an alternative approach to linking project and funding metadata across multiple fields simultaneously.

To address this, a functionality that allows users to search by project name or acronym has been implemented, which then autofills relevant details across multiple metadata fields. Specifically, when a project is selected, *DaRUS*:

- Fills the acronym or name in the Project metadata field.
- Simultaneously populates the Funding Information metadata field with details such as the funding agency, related project acronym, and project identifier.

This functionality is enabled through integration with the Research Information System (FIS) of the University of Stuttgart, a database-driven platform that collects, analyzes, and manages metadata on research activities, including researchers, publications, projects, and funding. By leveraging the FIS REST API, *DaRUS* retrieves structured metadata, ensuring accurate and consistent project and funding information.

8 https://terminology.tib.eu/ts; *Visited on March 31, 2025.*

A significant challenge in this implementation was handling complex scenarios where a dataset is linked to multiple projects, some with multiple funding sources. Our solution addresses this by using the project acronym as a linking key, enabling accurate autofill even in cases involving multiple grants per project. This approach ensures the structured representation of funding relationships, enhancing metadata completeness and accuracy as can be seen in Figure 7.

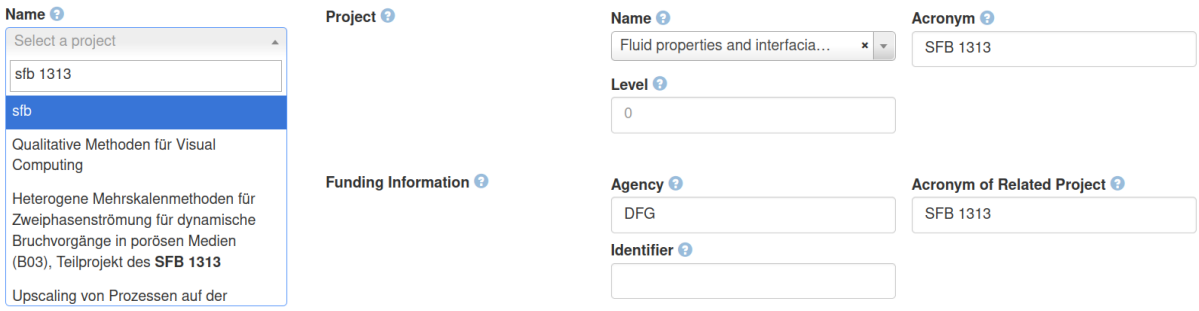


Figure 7: Interface to the research information system.

By implementing these automated interfaces, *DaRUS* significantly improves metadata standardization, reduces manual effort, and enhances the quality and interoperability of research data across repositories.

2.2 Formal Curation Support by pubWorkflow

Dataverse allows to integrate an external tool into the publication workflow as schematically shown in Figure 8(a). In our case the external tool is *pubWorkflow*, a Python REST API which can perform different tasks in the workflow. When a user clicks on "Publish" on a dataset page, the item will not be published instantaneously, but the publication will be stopped and *pubWorkflow* will be called to enable the curation workflow. On submission to our curation workflow, the dataset will be locked by default and can be unlocked by *pubWorkflow* to adapt the dataset for final publication. During the curation workflow, automatic checks will be triggered against various public APIs to facilitate the review. Table 1 gives an overview of the *DaRUS* metadata fields with validation logic implemented in *pubWorkflow*.

Existing metadata of a dataset will be sent to specific endpoints to get review recommendations and validations of a dataset's metadata.

Table 1: Metadata fields automatically checked by *pubWorkflow*.

metadata property	**DaRUS property name**	**example value**
author	authorName	Fritze, Florian
	authorIdentifier	https://orcid.org/0000-0002-9949-3815
grant_number	grantNumberValue	458524799
	grantNumberAgency	DFG
keyword	keywordTermURI	http://www.wikidata.org/entity/Q831774
	keywordValue	Density Matrix
ds_description	dsDescription	A linked README.md file in the description as HTML markup
publication	publicationIDNumber	https://doi.org/10.1002/ejic.202200709
topic_classification	topicClassValue	Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas
	topicClassVocabURI	https://w3id.org/dfgfo/2024/323

As shown in Table 2, we have integrated the ORCID API, Wikidata API, TIB Terminology Service API, Crossref API, Unpaywall API, DataCite API, Library of Congress Subject Headings API, OpenAire API and Loterre Chemistry Vocabulary API endpoints.

Table 2: API endpoints used for the automatic checks.

metadata property	**API endpoint**
author	https://pub.orcid.org/v3.0/expanded-search
grant_number	https://api.openaire.eu/search/projects
keyword	https://query.wikidata.org/sparql https://api.terminology.tib.eu/api/search https://sparql-endpoint.loterre.fr/LoterreVocabularies/sparql https://id.loc.gov/authorities/subjects/suggest
publication	https://api.crossref.org/works/ https://api.unpaywall.org/v2/ https://api.datacite.org/dois/
topic_classification	https://api.terminology.tib.eu/api/search

For instance, if there is a dataset which has a keyword with no Term URI the specific keyword will be sent to the aforementioned terminology services to automatically retrieve suggestions for Term URIs. Figure 8(b) shows an example for the keyword *Physics*. The reviewer of the dataset does not need to enter the keyword manually on various

terminology websites. If all validation checks are done, *pubWorkflow* will send a review mail to the ticket system so that the reviewers can see all the validation results. The next step will be to temporarily unlock the dataset in order to improve the metadata quality of the dataset and curate it as a *FAIR* dataset. If there is missing metadata, *pubWorkflow* will suggest it during the review. If there is already metadata in a specific metadata field, *pubWorkflow* will check if the field is matching our review standards. Let's say there is term value and a term URI for the topic classification section. PubWorkflow will validate the metadata. Another use case is the ORCID validation. If an author has no ORCID entry, *pubWorkflow* will send the author's name to the ORCID API to retrieve possible ORCID identifiers or the second case if there is an ORCID already, *pubWorkflow* will check if the author's name and ORCID identifier match with the entries in the ORCID database. So there will be suggestions for additional metadata entries and there will be also validations of current already entered metadata entries. These automated retrievals and checks can bring more efficiency to the publication workflow because there is no need anymore to do it manually piece by piece.

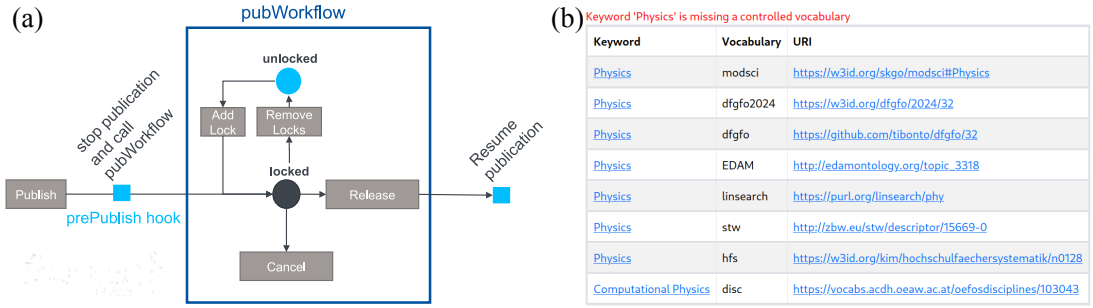


Figure 8: (a) Embedding of *pubWorkflow* in the publication workflow of Dataverse; (b) Term URI suggestion for given keyword.

3 Outlook

3.1 Scientific Curation Support

So far, the scientific review and the formal review are two separate steps in the publication workflow. To speed up the process and to allow the authors to get feedback from both scientific and formal viewpoint together, integration of *pubWorkflow* with *easyReview*[9] is planned, an online tool that allows to add comments to metadata and files and to interact with the authors in a chat-like interface.

Furthermore, an integration with the new overlay journal *JodaKISS*[10] is intended, which offers a scientific peer review for already published simulation data and software. Partic-

9 https://github.com/JR-1991/easyReview; *Visited on March 30, 2025.*
10 https://jodakiss.episciences.org/; *Visited on March 30, 2025.*

ularly high-quality data and software sets that are of distinct value to science can thus achieve additional visibility.

3.2 Automated Subject Indexing

To further improve the quality of metadata while also simplifying the process of entering said metadata for researchers, our ongoing project DA-FDM (digital assistant for research data management) focuses on enhancing the visibility and discoverability of research data within *DaRUS* through automated subject indexing. To this end, we are currently developing a tool called DA-FDM which can suggest terms from controlled vocabularies for the topic classification field in *DaRUS*. Whenever a researcher enters their research data in *DaRUS* and provides a related publication of the dataset or at least an abstract, this information will be passed to an AI model. The process then involves generating vector embeddings of the respective text, be it the related publication or the abstract, computing similarity measures against curated reference corpora, and inferring a subject classification based on nearest-neighbour relationships. The researcher will then be provided with a list of suggested terms from a controlled classification system, from which they can simply pick the terms best describing their dataset and confirm them. Thus, they do not have to think of keywords or know any subject indexing standards themselves, yet their data will be properly indexed and thereby easily findable and reusable. The information about their choices will also act as a feedback loop through which the model can be constantly improved to yield more refined results.

Acknowledgements

The development of the interfaces was partly funded by the Ministry of Science, Research and the Arts of the State of Baden-Württemberg (MWK) as part of the project "Digital Assistant for Research Data Management (DA-FDM)". The automation of the publication process was funded by the University of Stuttgart as part of the project ScalableRDM.

In addition to the authors, Anett Seeland (https://orcid.org/0000-0001-7979-8083) was involved in automating the data curation process conceptually and helped with implementing the interfaces and tools.

easyReview was written by Jan Range (https://orcid.org/0000-0001-6478-1051), the integration with JodaKISS is planned together with Sibylle Hermann (https://orcid.org/0000-0001-9239-8789).

Authorship Contributions

The presented manuscript results from joint efforts by the listed authors. In particular, section 2.2 (*Formal Curation Support by pubWorkflow*) has been written by Florian Fritze, section 2.1 (*User Input Support by Interfaces*) by Sarbani Roy, sections 1 (*DaRUS – The Data Repository of the University of Stuttgart*)and 1.1(*The publication process*) by Björn Selent and section 3.2 (*Automated Subject Indexing*) by Karoline Weinspach. All other text has been authored by Dorothea Iglezakis.

Conflict of Interest

The authors declare no conflicts of interest.

Bibliography

Iglezakis, Dorothea. 2023. *CodeMeta Metadata Block Configuration for Dataverse.* Software. Visited on March 5, 2025. https://doi.org/10.18419/darus-3291.

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, https://doi.org/10.5063/schema/codemeta-2.0.

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

Myers, James D., and Vyacheslav Tykhonov. 2023. *A Plug-in Approach to Controlled Vocabulary Support in Dataverse.* Zenodo. https://doi.org/10.5281/zenodo.8133723.

Schembera, Björn, and Dorothea Iglezakis. 2020. "EngMeta – Metadata for Computational Engineering". *International Journal of Metadata, Semantics and Ontologies* 14 (1): 26–38. https://doi.org/10.1504/IJMSO.2020.107792.

Seeland, Anett. 2020. *EngMeta Metadata Block Configuration for Dataverse.* Software. Visited on March 5, 2025. https://doi.org/10.18419/darus-508.

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. https://doi.org/10.1038/sdata.2016.18.

Bremen's Interdisciplinary Data Competence Center DataNord: An Integrative Approach for Data Literacy in Research

Lena Steinmann [1], Tanja Hörner [2], Uta Bohnebeck [3], Rolf Drechsler [1], Frank Oliver Glöckner [4], Iris Pigeot [5]

[1] Data Science Center, University of Bremen,
[2] U Bremen Research Alliance, University of Bremen,
[3] Hochschule Bremen, City University of Applied Sciences (HSB),
[4] Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI) / University of Bremen,
[5] Leibniz Institute for Prevention Research and Epidemiology – BIPS / University of Bremen

As data becomes an increasingly vital resource in science, society, and the economy, the need for strong data competencies continues to grow. The interdisciplinary data competence center *DataNord*, established in 2023, provides a scalable and integrative model for promoting responsible, data-driven research and strengthening data literacy among researchers. Anchored in the U Bremen Research Alliance, DataNord brings together universities, non-university research institutions, and infrastructure providers – connecting regional expertise with broader initiatives, such as the National Research Data Infrastructure.

DataNord operates across three interconnected pillars: learning, research, and networking. Based on comprehensive need assessments (2020–2023), its activities are tailored to researchers in Bremen, spanning all disciplines and skill levels. The structured training program Data Train offers basic training in research data management and data science. At the University of Bremen's Data Science Center, an interdisciplinary team of data scientists delivers flexible training modules and operates a centralized helpdesk for personalized consultation, closely connected with data steward services.

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23928 (CC BY-SA 4.0).

In the "Research Academy", interdisciplinary projects foster advanced data expertise in collaborative settings. Networking activities serve both individual and systemic levels: formats like the "Data Community Club" promote peer-to-peer exchange, while broader initiatives such as the "Research Data Day" and a Citizen Science project connect academia with society, industry, and policy stakeholders – broadening DataNord's reach.

This integrative approach strengthens data culture and creates a collaborative ecosystem in which knowledge transfer, innovation, and digital sovereignty can flourish. With its regional focus and national integration, DataNord serves as a best practice example for sustainable, future-proof data competence infrastructures in Germany and beyond.

Keywords: Data Train, Research Data Management, Data Science, Data Services

1 Background: The Relevance of Data Competencies in a Changing World

As digital transformation accelerates, the value of data continues to rise. Data processing and analysis methods – especially those involving Artificial Intelligence (AI) – have become essential tools for generating new insights and supporting a wide range of social and economic processes. A solid understanding of how these tools work, where they can be applied, and what their limitations are is crucial for responsible use. At the same time, critical reflection on the ethical, legal, and societal aspects of data technologies is necessary to mitigate risks and address emerging challenges. Political frameworks such as the European Union's AI Act[1] and the German government's Data Strategy[2] explicitly acknowledge and respond to these concerns.

Data skills are also becoming increasingly vital in the job market. According to the World Economic Forum's Future of Jobs Report 2025[3], AI and big data are among the fastest-growing areas of expertise. By 2030, 39 % of workplace skills are expected to change in this direction, and AI alone is projected to create 11 million new jobs.

In the research landscape, the reuse of data across disciplines – paired with transparent, innovative analysis methods – is key to tackling complex societal challenges. This was clearly demonstrated during the COVID-19 pandemic and remains central to fields like climate research (Conner et al. 2023). The FAIR principles provide a foundational framework for responsible data handling (Wilkinson et al. 2016). Since 2019, both national and international initiatives have been working to build infrastructures to support FAIR data

1 https://artificialintelligenceact.eu/; *Visited on March 24, 2025.*
2 https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/2023/datenstrategie.pdf?__blob=publicationFile&v=4; *Visited on March 24, 2025.*
3 https://www.weforum.org/publications/the-future-of-jobs-report-2025/; *Visited on March 24, 2025.*

practices – among them, Germany's National Research Data Infrastructure (NFDI), with 26 discipline-specific consortia, the European Open Science Cloud (EOSC) and Gaia-X. Moreover, major funding agencies, such as the German Research Foundation (DFG), now require detailed data management plans outlining how data will be handled throughout the research lifecycle[4].

Despite these developments, many researchers still lack the skills to handle data effectively. Mons (2020) reports that PhD students spend up to 80 % of their time on "data wrangling" – cleaning and organizing data – rather than on actual research. The European Commission's Expert Group on FAIR Data calls for the professionalization of data science and stewardship roles throughout the research lifecycle (European Commission. Directorate General for Research and Innovation 2018). The German Council for Information Infrastructures stresses the need for ongoing training (Rat für Informationsinfrastrukturen 2019), while the Stifterverband (2021) identifies data literacy as a critical skill for the future. In line with this, the German Science and Humanities Council encourages academic institutions to establish dedicated training and support services (German Science and Humanities Council 2022).

To ensure excellence and innovation in research, robust support structures are essential. In response, the German Federal Ministry of Education and Research (BMBF), as part of Germany's national Data Strategy and Action Plan for Research Data, launched a nationwide initiative in June 2022 to establish data competence centers at universities and research institutions[5] well aligned with the National Research Data Infrastructure – NFDI. The goal is strengthening data competencies in sciences and data-driven research in Germany – together.

2 DataNord: An Interdisciplinary Data Competence Center for the Bremen Region

The interdisciplinary data competence center DataNord[6] has been established for the Bremen region. It is one of eleven national centers funded by the BMBF since November 2023, as part of a nationwide initiative to run for three years. As the only center of its kind in Northern Germany, DataNord plays a key role in promoting interdisciplinary learning and collaborative research – driving innovation within Bremen and beyond.

4 https://www.dfg.de/en/news/news-topics/announcements-proposals/2022/info-wissenschaft-22-25; *Visited on March 24, 2025.*

5 https://www.bmbf.de/SharedDocs/Bekanntmachungen/DE/2022/06/2022-06-21-Bekanntmachung-Datenkompetenzzentren.html; *Visited on March 24, 2025.*

6 https://www.datanord-bremen.de; *Visited on March 24, 2025.*

DataNord is being developed under the umbrella of the U Bremen Research Alliance (UBRA), a regional collaboration network of the University of Bremen and twelve non-university research institutes – including all four major German research organizations (Helmholtz Association, Leibniz Association, Fraunhofer Society, and Max Planck Society) as well as the German Research Center for Artificial Intelligence (DFKI). The initiative cooperates with three additional regional partners including the Bremen University of Applied Sciences (see Figure 1). By bringing together these diverse partners, DataNord serves as a microcosm of the German research landscape and offers a scalable role model for nationwide developments in the academic system.

At the national level, DataNord collaborates closely with the twelve NFDI consortia UBRA members are currently involved, the ten other data competence centers, and broader networks such as state-level research data management (RDM) initiatives, NFDI e.V., and infrastructure providers like the German Network for Bioinformatics Infrastructure (de.NBI). Its collaborative model follows a structure similar to the NFDI, bringing together core institutional partners actively shaping the center's activities, alongside associated partners – both regional and national – who benefit from its services and contribute their expertise.

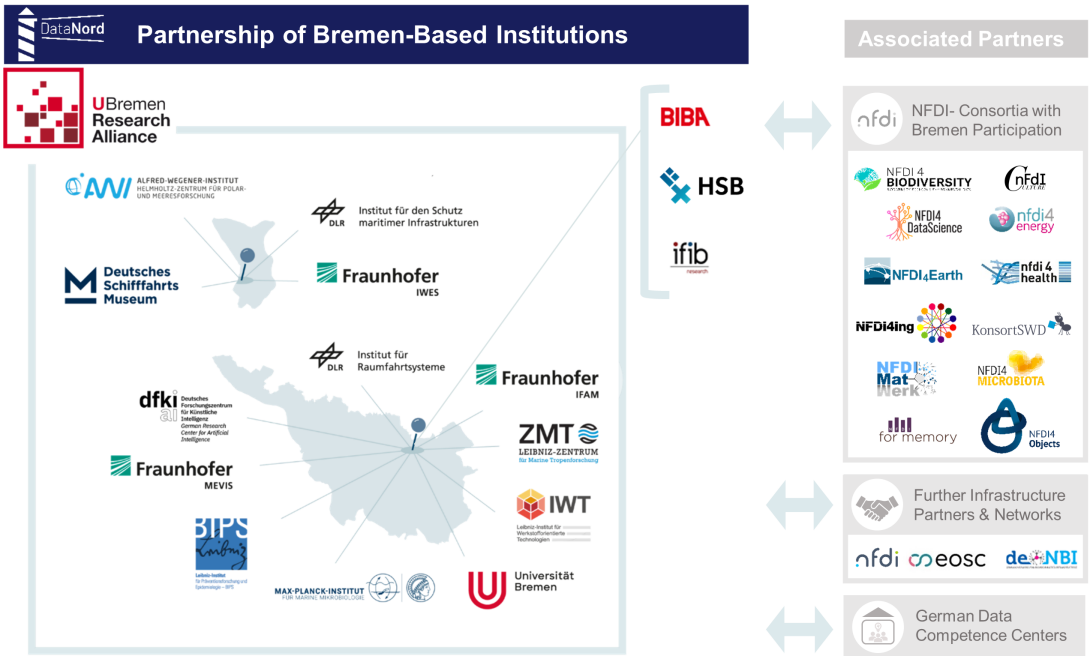


Figure 1: DataNord's core institutional partners (left), as well as associated partners such as the NFDI-consortia, data competence centers, infrastructure partners and networks (right).

2.1 Interdisciplinary Collaboration Concept: Building on Bremen's Strength

DataNord brings together interdisciplinary expertise from the core institutional partners, NFDI consortia, and regional infrastructure centers (see Figure 1). Based on the research priorities of its partners, DataNord focuses on five scientific profile areas: (1) Environmental and Marine Sciences, (2) Social Sciences, (3) Material and Engineering Sciences, (4) Health Sciences, and (5) Humanities.

By creating synergies across institutional boundaries, DataNord consolidates deep expertise in RDM and data science. Its interdisciplinary approach makes best use of Bremen's existing data-related structures and builds upon them. Shared methods and cross-cutting research questions help in bridging disciplinary boundaries, creating new opportunities for collaboration in research and education while breaking down disciplinary silos.

The organizational structure of DataNord is anchored in the UBRA, which provides strategic direction through governance structures at both the scientific and administrative level. Since 2019, UBRA's "Lead Project RDM and Data Science" has connected data initiatives across Bremen and, through its Working Group "Research Data" (AG Forschungsdaten), created a platform for exchange and joint action among local stakeholders (see joint white papter: Pigeot et al. 2021).

At the core of DataNord are two established initiatives: the University of Bremen's Data Science Center (Steinmann, Thöricht, et al. 2023), which offers customized training and consulting through an interdisciplinary team of data scientists, and the UBRA "Data Train" program (Hörner et al. 2021), which equips early-career researchers with basic skills in data science and RDM, while also offering structured networking opportunities.

In addition, DataNord benefits from the broad methodological and applied expertise of its partners. Highlights include the BIPS's strength in managing sensitive health data, the DFKI's leading role in developing and applying AI methods, Fraunhofer MEVIS's experience in designing tools for AI collaboration with industry, and the AWI's "Data Flow from Observations to Analysis & Archives" (O2A) framework, which enables seamless integration from data acquisition to archiving. In addition, DataNord complements the development of a comprehensive RDM service at Bremen University of Applied Sciences, supported by the BMBF as part of the "FDM@HSB" project[7]. Finally, DataNord draws on the expertise of renowned research data centers such as PANGAEA (Earth and Environmental Sciences) and Qualiservice (Social Sciences), which contribute deep knowledge in data archiving, reuse, data protection, and research ethics. The State and University Library Bremen complements these efforts with its expertise in open access, data publishing, long-term archiving, and metadata management.

7 https://www.hs-bremen.de/forschen/forschungs-und-transferprofil/forschungsprojekt/fdmhsb-forschungsdatenmanagement-der-hochschule-bremen-fdm-hsb/; *Visited on March 24, 2025.*

2.2 An Integrative Model for Connecting Regional and National Initiatives

DataNord brings together local actors that are actively involved in broader initiatives such as the NFDI or EOSC, positioning itself as a key interface for regional coordination and dynamic exchange between regional and national activities. In this role, DataNord positions itself as the "on-site" hub for NFDI activities, with a focus on the (interdisciplinary) application of data science methods and acts as the central point of contact for data-related activities in the Bremen region.

One outstanding example of how regional and national efforts are connected is the *Data Train* program, which has emerged as a model of good practice. From the beginning, Bremen-based NFDI consortia have been actively involved in shaping the program by contributing to both teaching and conceptual design. In turn, Data Train offers have been made accessible to NFDI communities across Germany wherever possible – with >36 % of participants coming from outside the region during the last three years. Under the umbrella of DataNord, this collaborative model is being expanded to include all partners as well as additional associated partners, including complementary data competence centers. This nationwide collaboration is emphasized by the joint publication of a training handbook with NFDI4Health (Dierkes et al. 2023) and by contributions to the development of the "RDM-Learning Objective Matrix" (Petersen et al. 2022, 2023; Petersen et al. 2025).

The University of Bremen acts as a Data Science Center and plays a key role in linking regional and national helpdesk structures. As the first point of contact for the DataNord network, its helpdesk handles many inquiries directly and, when needed, either reaches out to specialized services for additional expert input – such as those offered by NFDI consortia, PANGAEA, Qualiservice, or other data competence centers – or connects researchers with them. In doing so, it increases the visibility of these services within the region and helps in expanding the target groups. By maintaining close collaboration with other regional and national initiatives, it is ensured that support is efficient, seamless, and tailored – minimizing barriers and administrative overhead for researchers while enhancing the overall consultation experience.

Additionally, DataNord serves as a regional networking platform for contributors committed to national initiatives, thereby integrating their expertise. Examples include NFDI networking meetings, the UBRA Working Group "Research Data", and the Data Community Club (see Section 4.3).

In this way, DataNord is a catalyst for innovative, data-driven, and interdisciplinary research and training in the Bremen region – while also maintaining strong ties to national structures and amplifying its impact far beyond the regional level.

3 Approach, Target Group, and Topics

DataNord takes an interdisciplinary and holistic approach, offering tailored support for researchers at all career stages – from doctoral candidates and postdocs to senior researchers and professors – and across varying levels of data expertise. Common threads across the diverse partner institutions include a shared methodological knowledge in data science and a comprehensive focus on data handling throughout the entire research data lifecycle: from data collection and processing, to preparing data for automated analysis, applying AI methods, and ensuring data are made available according to the FAIR principles. This also involves promoting a broader cultural shift toward a "FAIR data culture", while addressing ethical, legal, and social considerations. Through targeted science communication and close collaboration with external partners, DataNord facilitates the transfer of knowledge into society, industry, public administration, and policy making.

All of DataNord's offers are aligned with the specific needs of the Bremen research community. These were identified through a series of assessments: over 70 guided interviews in 2020, a quantitative survey with more than 500 participants in 2021, and a follow-up network survey in early 2023 as part of the DataNord conceptual phase involving 82 selected experts from the partner institutions and disciplines.

The results highlighted the diversity of research data used within the network – reflected in the interdisciplinary composition of the center. Accordingly, DataNord addresses a wide range of data types. The most common data sources mentioned include observations, experiments, simulations, and archival data (such as texts or samples). In the social sciences, additional formats – such as survey and interview data, social media content, as well as image, video, and audio data – are in particular relevant. Frequently used file types include text-based formats (e.g. log files from behavioral tracking or sensor data) and image files (e.g. MRI scans or microscopy images).

The data analysis methods range from data visualization tools, statistical and mathematical techniques, as well as methods for cleaning, processing, to integrating datasets. Although AI methods are not yet widely adopted, interest in their use is high – highlighting a growing demand and a clear need to close the existing AI skills gap.

Based on these insights, DataNord's activities are designed to reflect the diversity of data and methods used in the region. Activities are offered in coordinated, interdisciplinary and discipline-specific formats, tailored to varying levels of expertise. Data science methods are particularly well suited for interdisciplinary training, as their technical foundations remain consistent across fields – even as their applications differ. This makes them an ideal vehicle for cross-disciplinary skills development. The impact of DataNord's activities on building data science competencies is illustrated in Figure 2.

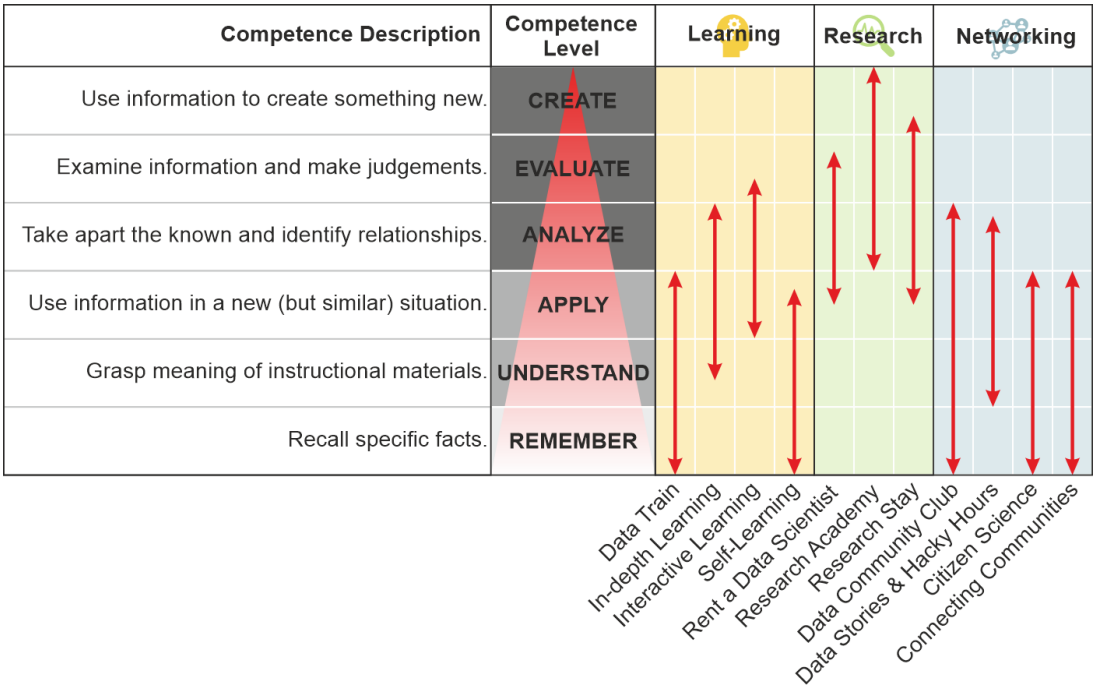


Figure 2: Overview of how DataNord's activities contribute to the development of data competencies, based on the revised Bloom's taxonomy (Bloom et al. 1956). Notably, the *Research Academy* projects are particularly significant, as they play a central role in enabling participants to reach the highest level of proficiency: *Creating*.

Professors and senior researchers were identified as a particularly hard-to-reach group, often due to limited time and differing priorities. To meet their needs – such as time efficiency, immediate research benefit, and strategic relevance – DataNord offers flexible, high-level formats like individual consultations, advanced networking, and research-focused projects (Figure 3). By involving them not only as learners but also as mentors and contributors (e.g., in the Research Academy or as Data Train lecturers), DataNord fosters leadership engagement and positions them as role models – reinforcing data literacy as a leadership priority across institutions.

4 Offers and Activities

4.1 Learning

Data Train: Launched in 2021, the structured training program *Data Train* forms the basis of DataNord's learning activities with more than 8,000 participations so far. Recognized as a "Good Practice in FAIR Competence Education" in Europe (Garbuglia, Saenen, and Gaillard 2022), the program focuses on providing cross-disciplinary funda-

mentals in RDM, data science, ELSA (ethical, legal, and social aspects), and critical thinking. It consists of three structured learning "tracks": the Starter Track and two hands-on Operator Tracks "Data Steward" and "Data Scientist", focusing on the specific skill sets researchers need to manage and analyze research data, respectively. The courses are delivered by volunteer experts from the DataNord network, who share their knowledge with the broader research community. From the beginning, the program has been closely associated with the NFDI (see Section 2.2).

In-Depth Learning: In addition to the basic training provided by Data Train, a team of five data scientists, covering the regionally anchored scientific profile areas (see Section 2.1), offers modular in-depth courses. These flexible "learning modules" follow a building-block approach and can be offered multiple times per year. The aim is to create a diverse portfolio of training options tailored to different skill levels. Emphasis is placed on hands-on learning that encourages participant interaction and practical data work. Each module is closely aligned with the research areas of DataNord's partner institutions. On-demand formats are also available, specifically designed to meet the needs of graduate schools, departments, or research institutes.

Interactive Learning: The aim is to spark enthusiasm for data science methods through accessible, engaging, and practice-oriented formats. Offerings range from low-threshold lunch-time sessions like "Data Snacks" to immersive Summer/Winter Schools, Data Hacks, and gamified formats that create an intensive and interactive learning environment. Participants work directly with real (often their own) data, gaining practical skills in data science methods as well as soft skills such as teamwork, communication, and peer-to-peer networking.

Self-Learning: As part of the Data Train program, a curated catalog of selected digital self-learning resources covering research data management and data science topcis – called the Digital Toolkit – is being developed and made openly accessible. It consists of both newly created and existing materials and enables flexible, self-paced learning tailored to individual needs. In addition, teaching materials from the advanced training modules are being prepared as Open Educational Resources (OER), making them reusable and adaptable for other institutions. These OERs can also be integrated into university teaching, addressing the growing demand in the DataNord network for training resources for university lecturers.

4.2 Research

"Rent a Data Scientist" – Data Consulting Services: To complement the region's existing RDM support services with targeted data science expertise, an interdisciplinary team of data scientists is establishing a centralized data science consulting service at the University of Bremen's Data Science Center. In close collaboration with local data

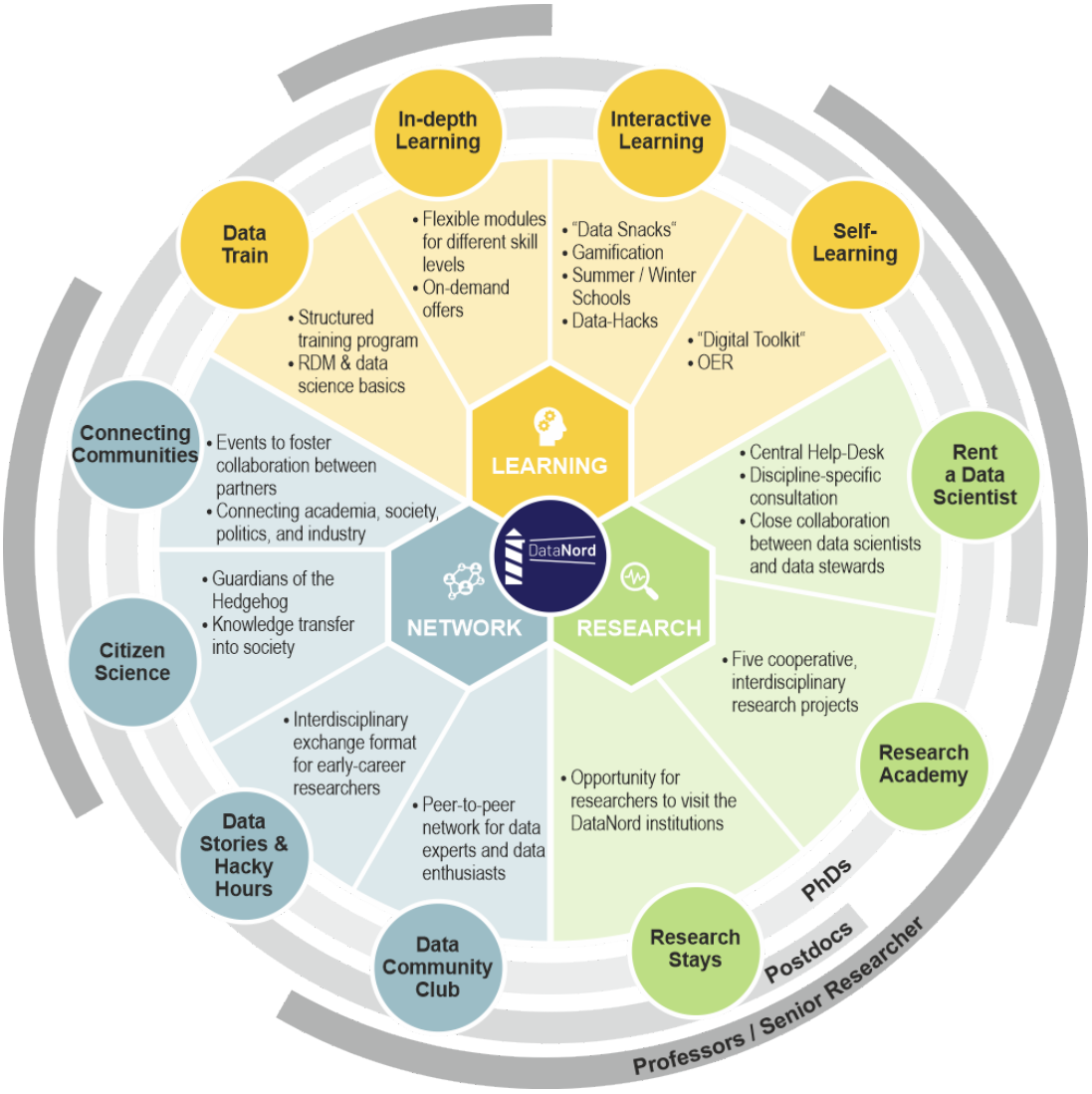


Figure 3: Overview of DataNord's activities across the three central areas – learning, research, and networking – and their alignment with key target groups (doctoral candidates, postdocs, and professors/senior researchers). All three groups also participate in citizen science initiatives, helping to facilitate knowledge transfer to society. While students are not the primary audience, selected learning activities are made accessible to them as well.

stewards, the helpdesk supports researchers throughout the entire data lifecycle – from the proposal stage to research practice. The helpdesk service is offered at three levels: (1) *Exploration* – Short consultations for proposal writing or research-related questions (e.g., creating DMPs, code fixes); (2) *Guidance* – In-depth support for specific data analysis or data management questions through regular meetings over weeks or months (e.g., assistance with implementing mixed-methods designs or analyzing geospatial data); (3) *Realization* – Long-term, project-based collaboration following the "Rent an Expert" concept, where a data scientist takes on key project tasks (e.g., responsibility for data

analysis as part of a publication). With this flexible and integrative consulting approach, DataNord ensures researchers receive tailored and efficient support.

"Research Academy" – Collaborative Research: As part of the DataNord Research Academy, five collaborative research projects tackle interdisciplinary questions in data science, drawing on diverse datasets. These projects bring together 34 researchers from ten DataNord partner institutions and are rooted in at least two of DataNord's thematic focus areas (see Chapter 2). This setup sparks new collaborations across disciplinary and institutional boundaries motivating researchers to think out of the box and to leave their comfort zones. Researchers at all career stages are intentionally involved, with both professors and postdocs taking on leadership roles. This cross-level collaboration fosters mutual learning and facilitates the exchange of expertise across generations. By embedding skill development directly within the research process, data science competencies are built in a deep and lasting way – going well beyond what traditional training formats typically achieve (see Figure 2). The knowledge and capabilities developed within these projects are shared with the wider research community through learning and networking opportunities. The Research Academy intentionally extends DataNord's service portfolio to prepare the next generation of data scientists and to shape a strong, future-oriented data culture in academia.

Research Stays: To foster international collaboration in data science, DataNord offers research stays – particularly for early-career researchers – at its partner institutions. These stays provide access to Bremen's data infrastructure, technical expertise, and methodological know-how. Just as importantly, they promote the exchange of data, perspectives, and ideas, paving the way for new research questions and innovative approaches. This two-way knowledge exchange strengthens long-term partnerships and supports the spread of data science skills and practices well beyond the region.

4.3 Networking

Data Community Club: Building on the existing data steward network of the UBRA, and inspired by the Data Champion Concepts from the University of Cambridge and TU Delft, a Data Community Club is being established in Bremen. This club brings together data experts – such as data scientists and data stewards – with data enthusiasts and data champions from various disciplines, creating a platform for exchange on RDM, data science, and related topics. In regular peer-to-peer meetings, practical approaches and solution strategies are discussed, with the aim of sustainably embedding data competencies into workflows across the entire data lifecycle. To incentivize and recognize outstanding contributions related to research data – as well as dedicated efforts to promote FAIR data practices – DataNord awards an annual "*Data Champion Award*" with a prize of €1,500.

Data Stories & Hacky Hours: *Data Stories* are part of the Data Train program and offer a platform for exchange between academia and industry. Inspiring talks from companies (e.g., Airbus) and researchers showcase real-world applications of RDM and data science. Informal networking fosters conversation and connection between participants and experts. *Hacky Hours* are regular, informal meetups between experts and users – open to all, with a focus on PhD students and postdocs. In a peer-to-peer format, early-career researchers discuss case-specific methods, exchange experiences, and receive targeted support for project-specific questions from experts. This interdisciplinary exchange bridges disciplinary and methodological gaps and enables innovative solutions.

Citizen Science: DataNord leverages Citizen Science as an innovative way to strengthen data literacy among both the public and researchers. Led by the Leibniz Institute for Prevention Research and Epidemiology – BIPS, and in collaboration with the Igelfreunde e.V. network and NFDI4Biodiversity, the project "Guardians of the Hedgehogs – Together for the Protection of Our Spiky Neighbors" engages citizens in tracking the hedgehog population in Bremen. Participants learn how to collect, document, and analyze data systematically. The collected data are made available according to the FAIR principles, and can be used in further scientific studies.

Connecting Communities: This activity strengthens collaboration within the DataNord consortium and promotes knowledge transfer across academia, society, policy, and industry. The *UBRA "Working Group on Research Data" serving as "DataNord Sounding Board*" plays a key role by bringing together representatives from all DataNord partner institutions. It coordinates regional activities in RDM and data science, connects Bremen-based NFDI consortia, and fosters synergies for joint initiatives. A key event in this effort is the *Research Data Day*, which offers researchers, data experts, and enthusiasts an annual platform to discuss best practices and current challenges. In addition, DataNord supports the regional networking of NFDI consortia through targeted events. The regional implementation of the *OneNFDI vision* of the NFDI Association was taken up by the "RDM and Data Science" lead project of the UBRA and has been further developed in the framework of DataNord. The Data Science Center is also driving the establishment of a network of the national data science community and, in this context, e.g., organized a nationwide workshop in 2024 that included a dedicated session for the data competence centers (Steinmann, Nowotka, et al. 2023; Pfuhl et al. 2024).

5 Quality Management

To ensure continuous improvement and future-oriented, adaptive development, the following quality assurance measures have been implemented: (1) *Establishment of an Expert Panel:* The UBRA Working Group "Research Data", serving as the DataNord Sounding Board, supports the strategic development of services and contributes to quality assurance through expert feedback and governance measures. (2) *Evaluation and Optimization*

of Training Programs: All training formats are regularly evaluated to enable ongoing improvements based on participant feedback. Involvement in and alignment with published competency frameworks (Demchenko et al. 2022; Engelhardt et al. 2022; Petersen et al. 2025), also help to ensure the quality of the offerings. (3) *Quality Control in Consulting Services*: For complex inquiries, the helpdesk team applies the four-eyes principle to ensure well-founded, high-quality advice. (4) *Regular Internal Feedback Sessions*: Within the consulting team, continuous reflection sessions are held to exchange experiences and identify best practices. (5) *Cross-regional Knowledge Exchange*: Ongoing collaboration with other data competence centers, NFDI consortia, and regional initiatives enables mutual knowledge transfer and systematic comparison of best practices in both consulting and training. All parties benefit by identifying successful concepts and emerging developments early, co-developing them, and integrating them into their own offerings. (6) *Staff Development*: Ongoing professional development for staff helps to maintain a high standard of expertise and service quality. (7) *Monitoring of Usage and Participation*: Systematic tracking of user engagement and participant numbers supports the evaluation and strategic planning of offerings.

An important upcoming step towards long-term quality assurance is the planned certification of the Data Train program. Through the Academy for Continuing Education at the University of Bremen, a structure is being established for regular quality reviews of the program's modules. In addition, assignments will be introduced to quantitatively assess participants' learning progress after completing a course. These evaluations will help measure training effectiveness and serve as key indicators of the program's impact.

6 Conclusion & Outlook

Data literacy is no longer a niche skill – it's a core requirement for innovation, interdisciplinary collaboration, and progress across science, society, and the economy. DataNord plays a key role in empowering researchers to use research data efficiently, responsibly, and effectively.

With its interdisciplinary approach, strategic focus on regional strengths, and close alignment with national initiatives, DataNord serves as a best-practice model for other regions in Germany. It is already putting into action at the regional level what is being pursued nationally: the widespread development of data skills across all scientific disciplines. Close collaboration with NFDI consortia, partner data competence centers, and existing networks among its member institutions ensures that developed structures and concepts can be scaled beyond Bremen – strengthening the national research data strategy. DataNord is already making a significant contribution to the networking and advancement of the data competence landscape at national level with interoperable and supra-regionally accessible services such as Data Train.

DataNord also creates a space for interdisciplinary collaboration and innovation, enabling excellent research while actively promoting societal and economic progress. Through its work in promoting critical engagement with AI, communicating data-driven insights clearly, and training a new generation of data-savvy professionals, DataNord strengthens digital sovereignty and fosters knowledge transfer into industry.

Looking ahead, expanding DataNord's services in line with evolving needs will be key. The demand is growing. For instance, it is already clear that there is a strong demand for university lecturers to be trained in integrating data science topics and skills into their teaching – concepts and courses are being developed accordingly. At the same time, DataNord is pursuing long-term sustainability, aiming to secure and further develop its structures and staff in a reliable and future-proof way. Given the long-term relevance of data competence centers, their strategic planning must also be long-term – supported by sustainable, yet flexible and agile governance models that allow service offerings to continuously adapt to changing demands.

In the long run, DataNord has the potential to become a key driver in advancing data science skills in research, helping to ensure that scientists remain capable and confident in an increasingly data-driven world. The future of research is data-driven – and DataNord is building the foundation to ensure that today's and tomorrow's scientists are ready for it.

Acknowledgements

The DataNord project is funded under the grant number 16DKZ2026A by the BMBF and financed by the European Union – NextGenerationEU. The authors would like to thank all partners, collaborators, and supporters of DataNord.

Authorship Contributions

Lena Steinmann and Tanja Hörner were primarily responsible for developing and writing the manuscript. Uta Bohnebeck, Rolf Drechsler, Frank Oliver Glöckner, and Iris Pigeot played a key role in shaping the conceptual foundations of DataNord, as presented in the manuscript, and contributed through critical feedback and revisions. All authors reviewed and approved the final version of the manuscript.

Conflict of Interest

The authors declare that there are no conflicts of interest related to this work.

Bibliography

Bloom, Benjamin S., Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of educational objectives: the classification of educational goals; Handbook I: Cognitive domain.* Edited by Benjamin S. Bloom. New York: David McKay Company.

Conner, Alden, Scott Hosking, James Lloyd, Achintya Rao, Gavin Shaddick, and Malvika Sharan. 2023. *Tackling climate change with data science and AI.* Zenodo. https://doi.org/10.5281/ZENODO.7712969.

Demchenko, Yuri, Tomasz Wiktorski, Adam Bellloum, Steve Brewer, Cuadrado Gallego Juan José, and Viktoriya Degeler. 2022. *EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS) Release 4.* Zenodo. https://doi.org/10.5281/ZENODO.7506445.

Dierkes, Jens, Julia Fürst, Tanja Hörner, Georg Sebastian Klammt, Birte Lindstädt, Iris Pigeot, Katja Restel, Carsten Schmidt, Dagmar Waltemath, and Atinkut Alamirrw Zeleke. 2023. "Training concepts in research data management and data science with the focus on health research". *PUBLISSO,* https://doi.org/10.4126/FRL01-006441348.

Engelhardt, Claudia, Katarzyna Biernacka, Aoife Coffey, Ronald Cornet, Alina Danciu, Yuri Demchenko, Stephen Downes, et al. 2022. *D7.4 How to be FAIR with your data. A teaching and training handbook for higher education institutions.* Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.6674301.

European Commission. Directorate General for Research and Innovation. 2018. *Turning FAIR into reality :final report and action plan from the European Commission expert group on FAIR data.* LU: Publications Office. https://doi.org/10.2777/54599.

Garbuglia, Federica, Bregt Saenen, and Vinciane Gaillard. 2022. *Data Train – Training in Research Data Management and Data Science: Good Practice #1.* Zenodo. https://doi.org/10.5281/ZENODO.6206717.

German Science and Humanities Council. 2022. "Probleme und Perspektiven des Hochschulbaus 2030". *Positionspapier,* https://doi.org/10.57674/Z38P-RH78.

Hörner, Tanja, Iris Pigeot, Frank Oliver Glöckner, and 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,* number 3, 56–69. https://doi.org/10.17192/bfdm.2021.3.8343.

Mons, Barend. 2020. "Invest 5 % of research funds in ensuring data are reusable". *Nature* 578 (7796): 491–491. https://doi.org/10.1038/d41586-020-00505-7.

Petersen, Britta, Franziska Altemeier, Sophie Boße, Maya Dalby, Nina Düvel, Claudia Engelhardt, Mark Fichtner, et al. 2025. *Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM).* Zenodo. https://doi.org/10.5281/ZENODO.15025246.

Petersen, Britta, Claudia Engelhardt, Tanja Hörner, Juliane Jacob, Tatiana Kvetnaya, Andreas Mühlichen, Hermann Schranzhofer, et al. 2022. *Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards.* Zenodo. https://doi.org/10.5281/ZENODO.7034478.

———. 2023. *Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards.* Zenodo. https://doi.org/10.5281/ZENODO.8010617.

Pfuhl, Helen, Lena Steinmann, Dirk Nowotka, and Rolf Drechsler. 2024. "Aufbau einer überregionalen Data-Science-Community". In *INFORMATIK 2024,* 2197–2203. Gesellschaft für Informatik e.V. https://doi.org/10.18420/INF2024_190.

Pigeot, Iris, Frank Oliver Glöckner, Rolf Drechsler, Tanja Hörner, Derk Hergen Schönfeld, Lena Steinmann, and Björn Oliver Schmidt. 2021. *Etablierung eines kooperativen Forschungsdatenmanagements in der U Bremen Research Alliance.* Technical report. U Bremen Research Alliance. https://doi.org/10.5281/zenodo.4775371.

Rat für Informationsinfrastrukturen. 2019. *Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, zweite Auflage.* Visited on June 25, 2025. https://rfii.de/?p=4043.

Steinmann, Lena, Dirk Nowotka, Lea Oberländer, Helen Pfuhl, Heiner Stuckenschmidt, and Rolf Drechsler. 2023. "Workshop: „Aktuelle Entwicklungen und Perspektiven (an Hochschulen) im Bereich Data Science"". In *INFORMATIK 2023 – Designing Futures: Zukünfte gestalten.* Gesellschaft für Informatik e.V. https://doi.org/10.18420/INF2023_05.

Steinmann, Lena, Heike Thöricht, Sandra Zänkert, and Rolf Drechsler. 2023. "Das Data Science Center an der Universität Bremen: Interdisziplinärer Knotenpunkt und Service-Infrastruktur für die datenintensive Forschung". In *E-Science-Tage 2023: Empower Your Research – Preserver Your Data,* edited by Vincent Heuveline, Nina Bisheh, and Philipp Kling, 246–252. heiBOOKS. https://doi.org/10.11588/HEIBOOKS.1288.C18083.

Stifterverband. 2021. *Data Literacy Charta.* Visited on June 25, 2025. https://www.stifterverband.org/charta-data-literacy.

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. https://doi.org/10.1038/sdata.2016.18.

Terminology Services in the DACH Region Landscape – What Are the Essential Requirements?

Roman Baum [1,*], Syphax Bouazouni [2], Jan Fillies [2], Naouel Karam [2], Oliver Koepler [3], Pooya Oladazimi [3], Julia Sasse [1]

[1] ZB MED – Information Centre for Life Sciences, Cologne, Germany
[2] InfAI – Institute for Applied Informatics association, Leipzig, Germany
[3] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
[*] Corresponding author: baum@zbmed.de

Terminologies and terminology services play a crucial role in the National Research Data Infrastructure (NFDI) and related fields, especially in the DACH region, to ensure the generation of FAIR research (meta)data. The primary objective of the Base4NFDI basic service Terminology Services 4 NFDI (TS4NFDI) is to standardize and harmonize terminology services, establishing an interoperable and sustainable solution. This solution will be integrated into the long-term NFDI infrastructure, ensuring consistency and reusability across domains, improving terminology management, and fostering cross-domain collaboration within the NFDI community.

A comprehensive requirement analysis was conducted during the initialization phase of TS4NFDI, providing insights for its future development. A survey comprising 69 participants has provided valuable insights into the current challenges and needs for terminology services. The analysis focused on three key groups: 1) developers (of web services), 2) managers of web services which want to interact with terminology services and 3) users working with terminologies.

TS4NFDI developed work packages for the integration phase in response to survey participant feedback. The following key objectives have been identified: 1) The creation of collections and entity sets of terminologies (WP1), 2) the establishment of a NFDI-wide mapping service (WP2), and 3) the provision of support for term requests (WP3).

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23929 (CC BY-SA 4.0).

Furthermore, user requirements will be addressed in WP6, focusing on community engagement, communication, and training. This includes incubator projects to integrate terminology services into user services.

Keywords: terminology services, terminology, ontology, FAIR, metadata

1 Introduction

Given the exponential increase in data volume, occurring at a rate of doubling every three years (Henke et al. 2016), FAIR machine-actionable (meta)data is needed to prevent the loss of meaningful insights. In order to create FAIR machine-actionable (meta)data, it is necessary to establish an overarching research data management framework. Such a framework should be built upon metadata that describes research data and their generation across all domains. To ensure that disciplines can create and share a common understanding of the concepts and relations used to describe data, terminologies must be employed as a formal representation of domain knowledge.

Terminologies have a long tradition. The term was first used in 1770 by Christian Gottfried Schütz. The term 'Terminology' is derived from Medieval Latin 'terminus', signifying 'word' or 'expression', with the latter element deriving from the Greek '-logia', signifying 'a dealing with' or 'a speaking of'.[1] It is widely acknowledged that Eugen Wüster is the founder of modern terminology theory (Humbley 2022). Trojar has authored a scientific article on Wüster's perspective on terminologies. Trojar (2017) implies that Wüster's discourse on synchrony is related to modern semantics. Wüster prepared six ISO recommendations and one ISO standard as part of a committee for the international standardization of terminology ISO/TC37 (Trojar 2017). In ISO 1087-1 (2000) a terminology is defined as a set of designations or terms belonging to one special language. Moreover, terminologies could include several types of relations such as hierarchical, associative, sequential associative, temporal sequential and causal associative relations (Zemmouchi-Ghomari and Ghomari 2012).

In this manuscript the term terminology is used as an umbrella term for the terms terminology, ontology, vocabulary and thesauri. For a comprehensive analysis of the distinctions between these terms, readers are directed to consult the work of Pieterse and Kourie (Pieterse and Kourie 2014). Similarly, in the context of EOSC you may read the term semantic artefact, which is used even a bit broader (Jonquet et al. 2023).

A terminology service (TS) can be defined as a web-based service which is used for the management and provision of access to terminologies, including vocabularies, terminologies, ontologies, and classifications or thesauri. These services are crucial in various do-

1 https://www.etymonline.com/word/terminology; *Visited on March 26, 2025.*

mains where standardized terminologies are essential for effective data interoperability. The functionality of a terminology service encompasses querying, browsing, visualisation, and mapping between terminologies (Koepler et al. 2024).

A comprehensive overview of terminology services is available at the Basic Register of Thesauri, Ontologies & Classification (BARTOC, see Ledl 2015). BARTOC lists comprehensive metadata about terminology services and links to their original source. As shown in BARTOC the most commonly used software stacks for individual terminology services are the Ontology Lookup Service (OLS), OntoPortal and Skosmos. In this context, a software stack is defined as the technology used to host terminologies. OLS was developed in the early 2000 at EMBL European Bioinformatics Institute (EMBL-EBI, see Jupp et al. 2015). The OntoPortal system is developed by the OntoPortal Alliance, a consortium of research and infrastructure teams, and is derived from the BioPortal system (Jonquet et al. 2023). BioPortal was developed in 2007 at Stanford University (Noy et al. 2009). OLS and BioPortal are repositories for biomedical ontologies, supporting the W3C standards Resource Description Framework (RDF)[2] and the Web Ontology Language (OWL)[3]. The development of Skosmos was initiated in 2014 at the National Library of Finland (Suominen et al. 2015). It is specified for the W3C standard Simple Knowledge Organization System (SKOS)[4] model. All of these solutions are open-source. Other well known terminology service software stacks are Ontobee (Xiang et al. 2011), Aber-OWL (Hoehndorf et al. 2015) and the commercial solution CENtree[5].

In addition to terminology services, a range of open-source tools is available to support terminology engineers in the creation of terminologies. Examples of such tools include Protégé (Musen 2015), VocBench (Stellato et al. 2020) or ROBOT (Jackson et al. 2019). Proprietary and commercial tools such as SDL Trados Multiterm[6], Topbraid Composer[7] and Congree TermTiger[8] are also available.

2 Terminology Services and their usage in the DACH region landscape

To manage the complexity of existing software stacks for terminology services, this paper focuses on services provided in the DACH region. However, there are multiple terminology services hosted in the DACH region which are based on OLS such as the TIB

2 https://www.w3.org/TR/rdf11-concepts/; *Visited on March 26, 2025.*
3 https://www.w3.org/TR/owl2-overview/; *Visited on March 26, 2025.*
4 https://www.w3.org/TR/skos-reference/; *Visited on March 26, 2025.*
5 https://scibite.com/solutions/ontology-management/centree/; *Visited on March 26, 2025.*
6 https://www.trados.com/product/multiterm/; *Visited on March 26, 2025.*
7 https://allegrograph.com/topbraid-composer/; *Visited on March 26, 2025.*
8 https://www.congree.com/en/product/congree-termtiger; *Visited on March 26, 2025.*

Terminology Service (Strömert et al. 2023), SemLookP (Baum 2024), the BERD Terminology Service (Shigapov, Koepler, and Baum 2024), the Plasma-MDS Ontology Lookup Service (Becker, Chaerony Siffa, and Baum 2024) or the ANNO Terminologieserver[9]. BiodivPortal (Karam et al. 2024), MatPortal[10], and TechnoPortal[11] are based on OntoPortal. Furthermore, the Skosmos technology has multiple instances in the DACH region, including ACDH-CH/DARIAH Vocabularies[12], BARTOC Skosmos[13], Thesaurus for the Social Sciences (TheSoz, see Zapilko et al. 2013), ZPID Vocabularies (Trillitzsch 2022) and ZBW classifications (Gastmeyer, Wannags, and Neubert 2016).

In addition to these instances, several other (self-developed) terminology services exist in the DACH region such as DAtendrehscheibe für Normdaten und TErminologien (DANTE, see Helms 2018) developed at the VZG in Göttingen. Federal institutes also host terminology services, such as the WFS Geographische Namen GN-DE[14] managed by the Federal Agency for Cartography and Geodesy, and the Checklist API[15], hosted by the German Red List Centre on behalf of the German Federal Agency for Nature Conservation (BfN). It is important to note that all three of these terminology services provide their own API, with their own specific schema. Additionally, there is the Central terminology server[16] for the healthcare sector, hosted by Federal Institute for Drugs and Medical Devices (BfArM). This BfArM – Terminology Server reuses the software stack from the Austrian e-Health Terminology Browser (ELGA)[17]. Alongside SemLookP and the BfArM – Terminology Server, the Service Unit Terminological Services (SU-Term-Serv Ingenerf et al. 2023) is a third terminology service for the medical domain in Germany and used by the Medical Informatics Initiative.

The BfArM Terminology Server, ELGA and the SU-TermServ facilitate the provision of terminologies with a HL7® FHIR® API which is widely used in hospitals and is the standard for the electronic patient record (ePA).[18] In Switzerland, the DCC Terminology Service (Krauss et al. 2021), hosted by the Swiss Personalised Health Network (SPHN), makes its data available to machines via RDF. While the Integrated Authority File (GND)[19] serves a broader purpose than a traditional terminology service, certain parts of the GND are typically used similarly to terms in terminology services. Finally,

9 https://ols.imise.uni-leipzig.de/; *Visited on March 26, 2025.*
10 https://matportal.org/; *Visited on March 26, 2025.*
11 https://technoportal.hevs.ch/; *Visited on March 26, 2025.*
12 https://vocabs.dariah.eu/en/; *Visited on March 26, 2025.*
13 https://skosmos.bartoc.org/en/; *Visited on March 26, 2025.*
14 https://gdz.bkg.bund.de/index.php/default/wfs-geographische-namen-gn-de-wfs-gnde.html; *Visited on March 26, 2025.*
15 https://checklisten.rotelistezentrum.de/api/public/swagger-ui; *Visited on March 26, 2025.*
16 https://www.bfarm.de/EN/Code-systems/Services/Terminologyserver/_node.html; *Visited on March 26, 2025.*
17 https://termgit.elga.gv.at/index.html; *Visited on March 26, 2025.*
18 https://www.bundesgesundheitsministerium.de/ministerium/ressortforschung/handlungsfelder/digitalisierung/sd-epa-fhir-cdr.html; *Visited on March 26, 2025.*
19 https://gnd.network/Webs/gnd/DE/Home/home_node.html; *Visited on March 26, 2025.*

the ORGANON[20] terminology toolbox, developed at the Free University of Berlin, also provides information to a term, but with a different scope than a traditional terminology service. A term in ORGANON must be used in several discursive contexts.

The use of terminology services in the DACH region is characterised by a greater breadth, as many widely used terminology services are not hosted within the region. Well-known examples include OLS (Jupp et al. 2015), BioPortal (Noy et al. 2009), AgroPortal (Jonquet et al. 2018), Iconclass (Couprie 1983), NERC Vocabulary Service (Leadbetter, Lowry, and Clements 2014), Getty Thesaurus of Geographic Names[21], and SNOMED CT Browser[22].

The National Research Data Infrastructure (NFDI), with its discipline-specific consortia, can be seen as the representative of the DACH region in terms of gathering requirements for terminology services. Despite the fact that the various consortia of the NFDI address distinct knowledge domains, there is a considerable degree of overlap in the topics they tackle. These cross-cutting topics are addressed in the five sections of the NFDI listed below: 1) Common Infrastructures, 2) Ethical, Legal and Social Aspects, 3) (Meta)data, Terminologies, Provenance, 4) Training & Education and 5) Industry Engagement. Specific working groups are established within each respective section to ensure comprehensive and systematic handling of the various topics.

As terminologies and terminology services are of significant importance across various NFDI consortia, two dedicated working groups have been convened to address these issues. The two working groups are: 1) Ontology Harmonization and Mapping and 2) Terminology Services. The working group Ontology Harmonisation and Mapping focuses on the consistent use of terminologies and the organisation of vocabulary items into structured ontologies (Anders, Arera-Rütenik, et al. 2022). These are prerequisites for FAIR data, in particular regarding cross-community reuse and machine-actionability of research data. Furthermore, the Terminology Services working group aims to foster joint developments, share best practices and reduce barriers to the use of terminologies and terminology services (Anders, Bailly, et al. 2022).

As demonstrated in the preceding paragraphs, the DACH region is characterised by a diverse and heterogeneous landscape of disparate terminology services. The NFDI consortia and sections are fostering multiple approaches to harmonise the different terminology services. The working group Terminology Services was the genesis of the basic service Terminology Services 4 NFDI (TS4NFDI), which is funded by Base4NFDI.

20 https://organon-lexicon.org/; *Visited on March 26, 2025.*
21 https://www.getty.edu/research/tools/vocabularies/tgn/; *Visited on March 26, 2025.*
22 https://browser.ihtsdotools.org/; *Visited on March 26, 2025.*

3 Terminology Services 4 NFDI

Terminology Services 4 NFDI (TS4NFDI) is a cross-domain basic service for the provision, curation, development, harmonization and mapping of terminologies. TS4NFDI is funded by Base4NFDI, which itself is a joint initiative of all NFDI consortia. The objectives of TS4NFDI are as follows: 1) to standardise and harmonise the management of terminology within the NFDI and 2) to facilitate consensus-building and the interoperability of services across disciplines in order to achieve a common representation of knowledge and engineering framework. TS4NFDI provides a way to integrate and converge individual solutions into a standardized, interoperable, sustainable service suite embedded in a long-term NFDI infrastructure, thereby improving the overall efficiency and effectiveness of terminology management and use, as well as cross-domain collaboration within the NFDI community.

In order to gain a structured understanding of how terminology services and terminologies are used in the different NFDI consortia and beyond, a comprehensive requirement analysis (Baum et al. 2024) was conducted by TS4NFDI. The creation of a survey was an essential step in the process of producing a comprehensive requirement analysis. The survey shall capture the requirements of the NFDI consortia and identify potential use cases to support them in integrating terminology services to their applications. The following sections present the key findings of the survey.

4 TS4NFDI Survey Methodology

An online survey (LimeSurvey platform[23], hosted by the TIB in Hannover) was conducted targeting researchers, developers of web services and data managers within the 26 NFDI consortia and Base4NFDI. Furthermore, the survey targeted users who work with terminologies. The survey consisted of multiple-choice, list-based, and open-text questions, structured around key user roles: developers, managers, and terminology users. Participants were recruited via NFDI mailing lists, consortium meetings, and community outreach efforts. The survey remained open for six weeks (from 6th May 2024 to 14th June 2024), and 69 individuals completed it. Nevertheless before the first question an overview of technical terms used in the survey was shown (see Table 1).

23 https://www.limesurvey.org; *Visited on March 26, 2025.*

Table 1: This table contains explanations of technical terms used in the survey.

Explanation of technical terms
We refer as a **terminology** to any terminological resource, also referred to as semantic artifacts. This can be a formal ontology, a taxonomy, or any useful source of Semantic Web-compliant collections of terms (e.g. locations available through a geographical database such as Geonames).
A **terminology service**, also referred to as an ontology repository, is a system for managing and providing access to terminologies. These services are crucial in various domains where standardized terminologies are essential for effective data interoperability. They offer functionalities such as browsing, searching, mapping, and annotation as well as APIs to access those programmatically.
Term mapping involves linking different terms describing the same concept, thereby enabling more effective search and information retrieval. In data management and integration, term mapping helps align terminologies used in different datasets or databases, facilitating data interoperability and integration across various systems or domains.
In knowledge representation and reasoning, a **knowledge graph** is a knowledge base that uses a graph-structured data model or topology to represent and operate on data. Knowledge graphs are often used to store interlinked descriptions of entities – objects, events, situations, or abstract concepts – while also encoding the semantics or relationships underlying those entities.
Annotation or tagging is the process of attaching relevant metadata about concepts (e.g., people, places, organizations, products, or topics) to a text document or other unstructured content. Unlike classic text annotations, which are intended for the reader's reference, semantic annotations can also be used by machines. Semantically tagged documents are easier to find, interpret, combine, and reuse.

The survey itself was structured in eight sections (see Figure 1). The first and the last section contained questions for all participants. The remaining sections were only presented to participants who had provided specific answers in the survey. The survey was further segmented into three distinct sections, each tailored to specific target groups. The three possible branches are: 1) Developer, 2) Manager and 3) User/Other. Prior to and following these three branches for the specific target groups, general questions were posed to participants who had previously encountered terminologies. The general questions were divided into two sections due to the presence of a discrepancy in their respective contexts. To clarify, the subsequent query was posed at the conclusion of the survey, following the administration of questions pertaining to specific features: 'Can you

think of new terminology service features that would improve your system/workflows?'. This approach allowed participants to propose any missing features that may have been overlooked during the initial conception of the survey. Nevertheless, concomitant with these five branches, a branch was present for participants, for the purpose of potential new users of terminologies.

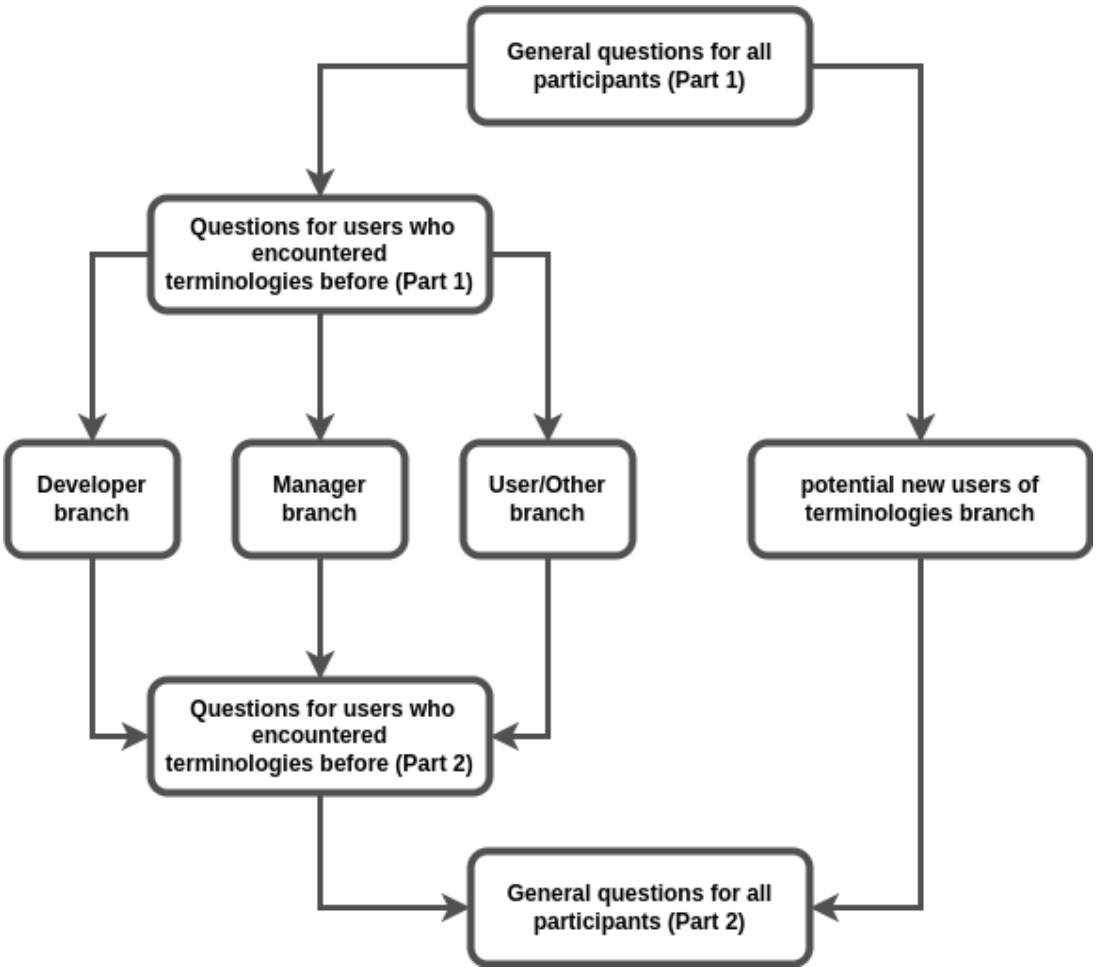


Figure 1: Structure of the survey.

The survey comprised a combination of three distinct question types: multiple choice, list ratio and free text. It was also possible to combine these types (see Figure 2). The majority of questions that incorporated a free text answer option were not required. However, the majority of questions incorporated multiple-choice or list-ratio types.

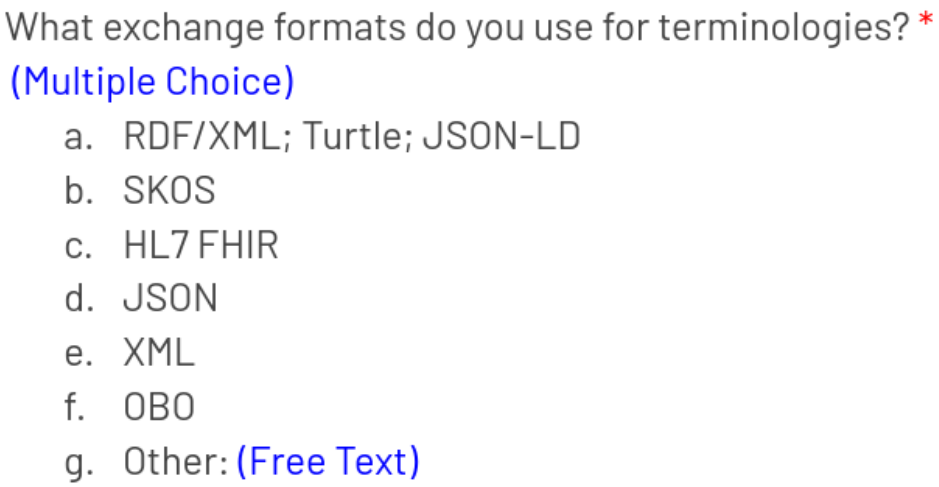


Figure 2: A question from the survey is depicted as an example. The red asterisk at the end of the question marks that an answer from the participant is required. If the participant selected the option 'Other:', a free text was provided to enter the used format(s).

5 Results & Discussion

The full questionnaire and all the multiple choice and list response results from the survey are published in the TS4NFDI Requirements Analysis Report (Baum et al. 2024). The free text responses could not be published for privacy reasons. In this paper, however, we want to focus on the key findings.

A total of 69 participants completed the survey. The survey successfully reached all 26 consortia and Base4NFDI, indicating a comprehensive representation of the target group. Of the 69 participants, at least one was associated with one of these consortia. It was possible for a participant to select more than one consortium in the survey, as an individual could be a member of different consortia. The TS4NFDI Requirements Analysis Report (Baum et al. 2024) also provides a more detailed overview of the number of participants related to each consortium. Only 2 participants didn't encounter terminology before this survey and are not part of the following presentation of the key findings. The remaining 67 participants were divided into the three specific target groups as follows: Developer (40.30 %, 27 participants); Manager (46.27 %, 31 participants); and User (4.48 %, 3 participants)/Other (8.96 %, 6 participants).

Among the 67 participants who responded to the question about creating or using custom collections of terminologies, 64 % (43 participants) expressed a clear need for such functionality (Figure 3). A smaller fraction (12 %, 8 participants) indicated no need, while 24 % (16 participants) were uncertain. This suggests that while most users value the ability to customize terminology subsets, a significant portion may lack the necessary

experience or knowledge to determine their need. Supporting these users with guidelines and use cases could increase terminology use.

The majority of participants (84 %, 56 out of 67 participants) indicated the necessity of mappings between terms, emphasizing the importance of interoperability across different terminologies (Figure 4). Only 5 participants (7 %) stated that they did not require mappings, while 6 (9 %) were unsure. These findings reinforce the idea that terminology services must prioritize robust mapping mechanisms to facilitate seamless data integration and interoperability across domains.

The need for a feature that allows users to request missing terms was assessed within the Manager branch of the survey (n=31). An overwhelming 94 % (29 participants) agreed that terminology services should support term requests, while only 2 participants (6 %) found it unnecessary (Figure 5). This strong consensus highlights the dynamic nature of terminologies and the necessity of continuous updates to reflect evolving research needs.

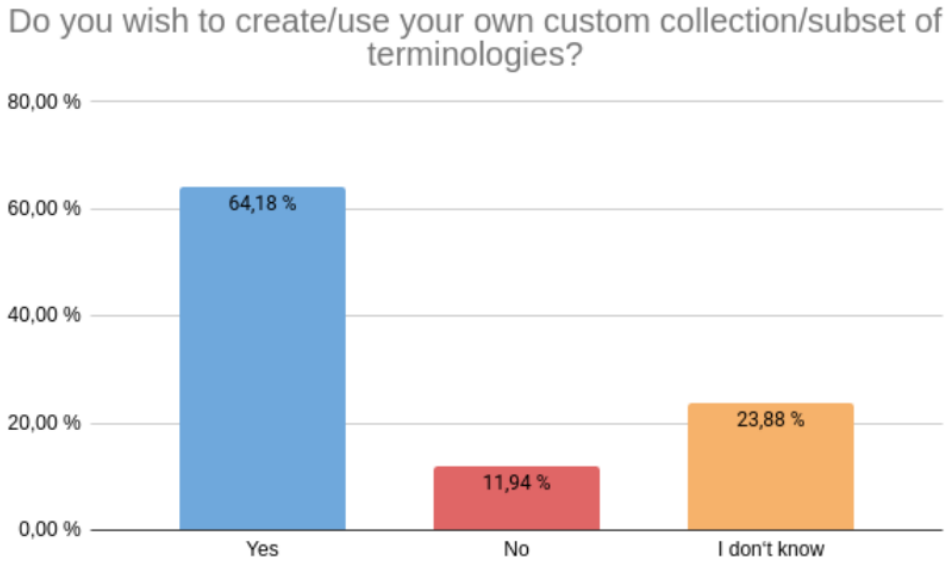


Figure 3: The figure presents three distinct potential responses for the x-axis, while the y-axis depicts the corresponding percentage values. The blue bar represents the answer "Yes", the red bar represents the answer "No", and the orange bar represents the answer "I don't know" (n=67).

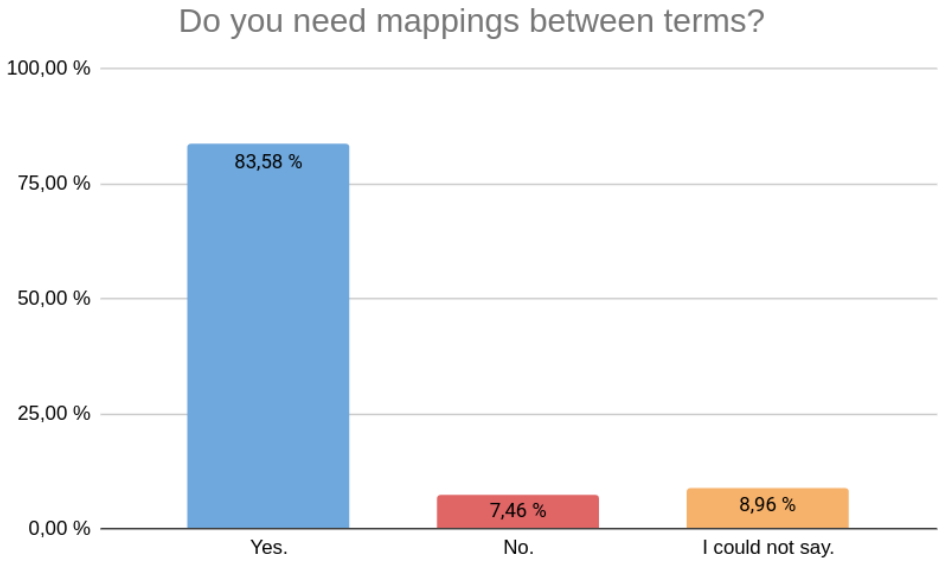


Figure 4: The figure presents three distinct potential responses for the x-axis, while the y-axis depicts the corresponding percentage values. The blue bar represents the answer "Yes", the red bar represents the answer "No", and the orange bar represents the answer "I could not say" (n=67).

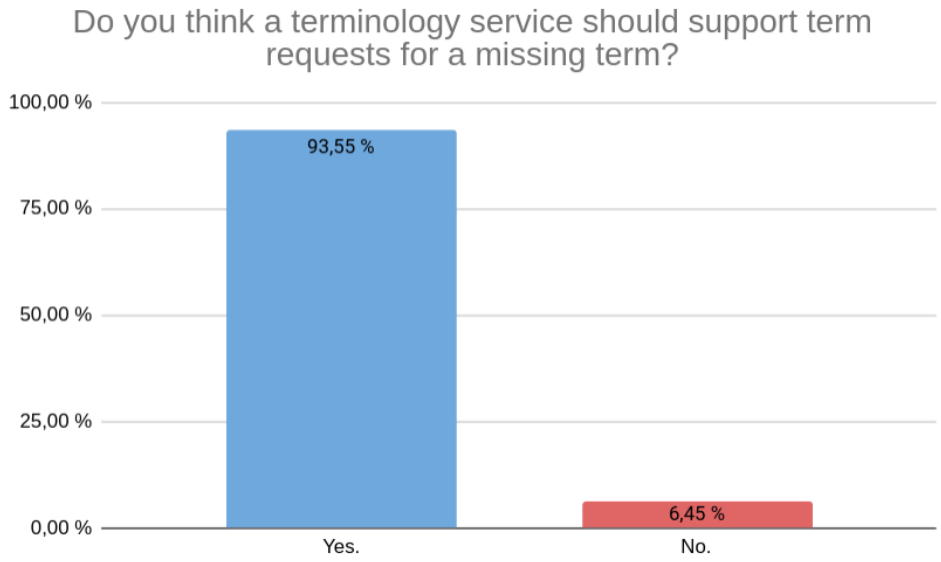


Figure 5: The figure presents two distinct potential responses for the x-axis, while the y-axis depicts the corresponding percentage values. The blue bar represents the answer "Yes", and the red bar represents the answer "No" (n=31).

Participants were asked about their expectations from TS4NFDI in its cross-consortium work. The results indicate high demand for multiple forms of support (see Figure 6). The

findings underline the necessity of a broad support offer that extends beyond technical implementations. Transfer of knowledge and best practices seems to be a critical factor in the successful harmonization of terminology services across the NFDI.

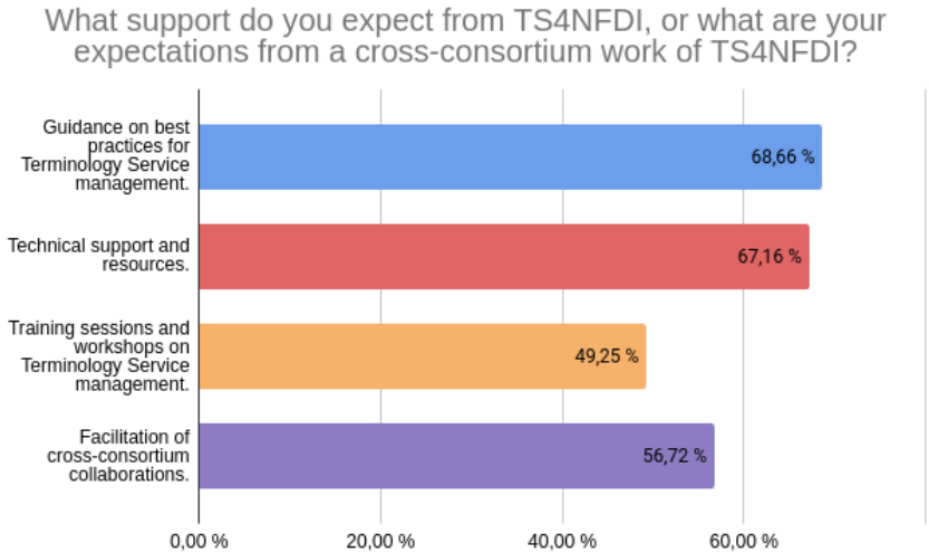


Figure 6: The figure presents four distinct potential responses for the y-axis, while the x-axis depicts the corresponding percentage values. The blue bar represents the answer "Guidance on best practices for Terminology Service management.", the red bar represents the answer "Technical support and resources.", the orange bar represents the answer "Training sessions and workshops on Terminology Service management.", and the purple bar represents the answer "Facilitation of cross-consortium collaborations" (n=67).

6 Conclusion and Outlook

The TS4NFDI team has aligned the requirements and expectations of the survey participants with the work packages (see Table 2) for the integration phase proposal. The survey results highlighted a strong demand for custom terminology collections (64 %), term mappings (84 %), and support for term requests (94 %). These findings directly shaped the objectives of TS4NFDI's integration phase, which focuses on: 1) creating collections and entity sets of terminologies (WP1), 2) Establishing an NFDI-wide mapping service (WP2) and 3) Providing support for term requests (WP3).

Table 2: Workpackages of TS4NFDI.

Work package Title	**Description of the Work Package**
WP1 TS4NFDI Service Portal	The TS4NFDI Service Portal is designed to facilitate the creation and management of entity sets and terminology collections. This enables domain experts to configure the response of the API Gateway or the TSS following their specific requirements.
WP2 Mapping Service	This work package aims to create and establish a mapping service. A mapping registry will be developed to store all relevant NFDI terminology mappings, accessible via APIs and open data formats (JSKOS, SSSOM, RDF, CSV) to adhere to FAIR principles.
WP3 Workflows for terminology curation	This work package will promote the possibility to request new terminologies or missing terms.
WP4 Extension of the API Gateway	The functionality of the TS4NFDI API Gateway will be extended with respect to the requirements of the TS4NFDI Service Portal (e.g. collection and entity set management) and additional user requirements (e.g. adding new terminology software stacks as backend resources) collected via incubator projects in WP6.
WP5 Extension of the Terminology Service Suite	The functionality of the Terminology Service Suite will be extended with respect to the requirements of the TS4NFDI Service Portal and the user requirements collected via incubator projects in WP6.
WP6 Community engagement, communication, and training	This work package will foster community interaction and integration with the TS4NFDI service, supporting user-driven development of new and existing features. Incubator projects will be conducted to integrate TS4NFDI into consortia services by exploring and testing new features with the TS4NFDI team.

Additionally, survey participants expressed a clear expectation for guidance, technical support, training, and cross-consortium collaboration, reinforcing the importance of Work Package 6 (Community engagement, communication, and training). This work package will provide support and a framework for incubator projects that integrate terminology services into user workflows. The incubator projects will enable hands-on collaboration between TS4NFDI and service providers, supporting activities such as: hosting of terminologies, establishment of a terminology service, addition of terminology services to the TS4NFDI API Gateway, integration of Terminology Service Suite widgets into other ser-

vices, and provision of collection(s) via the TS4NFDI API Gateway. Each incubator will run for six months, with dedicated resources (e.g. developer capacity) to support integration into another service. The first cycle, launched in Q1/2025 with five projects, will be completed in Q3/2025, followed by additional cycles. TS4NFDI will offer at least four of these cycles A minimum of four incubator cycles is planned during the integration phase, with progress tracked via an incubator dashboard on the TS4NFDI Service Portal.

Acknowledgements

The authors thank the participants of the E-Science-Tage conference for the comments and suggestions received in Heidelberg. They would also like to express their gratitude to Base4NFDI, and in particular to their service stewards André Giesler and Thomas Bauer, for their invaluable support in setting up the survey. Finally, we would like to express our gratitude for the support and contribution of the NFDI section working group Terminology Services members.

The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding and support. Work of the described TS4NFDI project is funded by DFG as part of Base4NFDI. Grant number: 521466146.

Authorship Contributions

- Roman Baum: conceptualization, supervision, writing first draft.
- All authors: writing & editing.

All authors read and approved the submitted version of the manuscript.

Conflict of Interest

The authors declare that there is no conflict of Interest.

Bibliography

Anders, Ivonne, Tobias Arera-Rütenik, Susanne Arndt, Roman Baum, Noemi Betancort, Ina Blümel, Christian Busse, et al. 2022. *Ontology Harmonization and Mapping* –

Working Group Charter (NFDI section-metadata). Zenodo. https://doi.org/10.5281/ZENODO.6726519.

Anders, Ivonne, Kolja Bailly, Roman Baum, Felix Engel, Felix Ernst, Luca M. Ghiringhelli, Angela Kailus, et al. 2022. *Terminology Services – Working Group Charter (NFDI section-metadata).* Zenodo. https://doi.org/10.5281/ZENODO.6759325.

Baum, Roman. 2024. *SemLookP widgets – Embed ontologies in service applications.* Zenodo. https://doi.org/10.5281/ZENODO.10529181.

Baum, Roman, Syphax Bouazzouni, Jan Fillies, Naouel Karam, Oliver Koepler, Pooya Oladazimi, and Julia Sasse. 2024. *TS4NFDI Requirements Analysis Report.* Zenodo. https://doi.org/10.5281/ZENODO.13862158.

Becker, Markus M., Ihda Chaerony Siffa, and Roman Baum. 2024. *Terminology service for research data management and knowledge discovery in low-temperature plasma physics.* Zenodo. https://doi.org/10.5281/ZENODO.14412030.

Couprie, Leendert D. 1983. "Iconclass: an iconographic classification system". *Art Libraries Journal* 8 (2): 32–49. https://doi.org/10.1017/s0307472200003436.

Gastmeyer, Manuela, Max-Michael Wannags, and Joachim Neubert. 2016. "Relaunch des Standard-Thesaurus Wirtschaft – Dynamik in der Wissensrepräsentation". *Information – Wissenschaft & Praxis* 67 (4): 217–240. https://doi.org/10.1515/iwp-2016-0039.

Helms, Tobias. 2018. "DANTE – DAtendrehscheibe für Normdaten und TErminologien". *AKMB-news: Informationen zu Kunst, Museum und Bibliothek* 24 (1): 14–15. https://doi.org/10.11588/akmb.2018.1.63475.

Henke, Nicolaus, Jacques Bughin, Michael Chui, James Manyika, Tamin Saleh, Bill Wiseman, and Guru Sethupathy. 2016. *The age of analytics: Competing in a data-driven world | McKinsey.* Technical report. Visited on March 31, 2025. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.

Hoehndorf, Robert, Luke Slater, Paul N. Schofield, and Georgios V. Gkoutos. 2015. "AberOWL: a framework for ontology-based data access in biology". *BMC Bioinformatics* 16 (1): 26. Visited on March 31, 2025. https://doi.org/10.1186/s12859-015-0456-9.

Humbley, John. 2022. "Chapter 1. The reception of Wüster's General Theory of Terminology". In *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge,* edited by Pamela Faber and Marie-Claude L'Homme, 15–36. Terminology and Lexicography Research and Practice. John Benjamins Publishing Company. ISBN: 9789027257789. https://doi.org/10.1075/tlrp.23.01hum.

Ingenerf, Josef, Joshua Wiedekopf, Andreas Beyer, Muhammad Adnan, Michael Marschollek, and Birger Haarbrandt. 2023. “Service Unit Terminological Services (SU-TermServ) der MII”. *MIRACUM DIFUTURE Journal,* visited on May 14, 2025. https://mii-termserv.de/assets/pdf/paper/20230505-MIRACUM_Journal-SU-TermServ.pdf.

Jackson, Rebecca C., James P. Balhoff, Eric Douglass, Nomi L. Harris, Christopher J. Mungall, and James A. Overton. 2019. “ROBOT: A Tool for Automating Ontology Workflows”. *BMC Bioinformatics* 20 (1): 407. https://doi.org/10.1186/s12859-019-3002-3.

Jonquet, Clément, John Graybeal, Syphax Bouazzouni, Michael Dorf, Nicola Fiore, Xeni Kechagioglou, Timothy Redmond, et al. 2023. “Ontology Repositories and Semantic Artefact Catalogues with the OntoPortal Technology”. In *The Semantic Web – ISWC 2023,* edited by Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, 38–58. Cham: Springer Nature Switzerland. ISBN: 9783031472435. https://doi.org/10.1007/978-3-031-47243-5_3.

Jonquet, Clément, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzalé Yeumo, Vincent Emonet, John Graybeal, et al. 2018. “AgroPortal: A vocabulary and ontology repository for agronomy”. *Computers and Electronics in Agriculture* 144:126–143. https://doi.org/10.1016/j.compag.2017.10.012.

Jupp, Simon, Tony Burdett, Catherine Leroy, and Helen E. Parkinson. 2015. “A new Ontology Lookup Service at EMBL-EBI”. In *Proceedings of the 8th International Conference on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2015),* volume 1546. Cambridge: CEUR Workshop Proceedings. Visited on March 31, 2025. https://www.semanticscholar.org/paper/A-new-Ontology-Lookup-Service-at-EMBL-EBI-Jupp-Burdett/b83bfbfc1f2f08e5b88af5ef65ef2a8687ac4112.

Karam, Naouel, Jan Fillies, Clement Jonquet, Syphax Bouazzouni, Felicitas Löffler, Franziska Zander, Birgitta König-Ries, Anton Güntsch, Michael Diepenbroek, and Adrian Paschke. 2024. “BiodivPortal: Enabling Semantic Services for Biodiversity within the German National Research Data Infrastructure”. *Datenbank-Spektrum* 24 (2): 129–137. https://doi.org/10.1007/s13222-024-00474-5.

Koepler, Oliver, Julia Sasse, Naouel Karam, and Roman Baum. 2024. *Terminology Services 4 NFDI (TS4NFDI) – Initialisation Phase Proposal.* Zenodo. https://doi.org/10.5281/zenodo.12188787.

Krauss, Philip, Vasundra Touré, Kristin Gnodtke, Katrin Crameri, and Sabine Österle. 2021. “DCC Terminology Service – An Automated CI/CD Pipeline for Converting Clinical and Biomedical Terminologies in Graph Format for the Swiss Personalized Health Network”. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute, *Applied Sciences* 11 (23): 11311. https://doi.org/10.3390/app112311311.

Leadbetter, Adam M., Roy K. Lowry, and D. Oliver Clements. 2014. "Putting meaning into NETMAR – the open service network for marine environmental data". *International Journal of Digital Earth* 7 (10): 811–828. https://doi.org/10.1080/17538947.2013.781243.

Ledl, Andreas. 2015. *Das Basel Register of Thesauri, Ontologies & Classifi cations (BARTOC).* Zenodo. https://doi.org/10.5281/zenodo.30751.

Musen, Mark A. 2015. "The protégé project: a look back and a look forward". *AI Matters* 1 (4): 4–12. https://doi.org/10.1145/2757001.2757003.

Noy, Natalya F., Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clément Jonquet, et al. 2009. "BioPortal: ontologies and integrated data resources at the click of a mouse". *Nucleic Acids Research* 37:W170–W173. https://doi.org/10.1093/nar/gkp440.

Pieterse, Vreda, and Derrick G. Kourie. 2014. "Lists, Taxonomies, Lattices, Thesauri and Ontologies: Paving a Pathway Through a Terminological Jungle". *KNOWLEDGE ORGANIZATION* 41 (3): 217–229. https://doi.org/10.5771/0943-7444-2014-3-217.

Shigapov, Renat, Oliver Koepler, and Roman Baum. 2024. *A terminology service via TS4NFDI for BERD@NFDI knowledge graph infrastructure.* Zenodo. https://doi.org/10.5281/zenodo.14230727.

Stellato, Armando, Manuel Fiorelli, Andrea Turbati, Tiziano Lorenzetti, Willem van Gemert, Denis Dechandon, Christine Laaboudi-Spoiden, et al. 2020. "VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons". Edited by Aidan Hogan. *Semantic Web* 11 (5): 855–881. https://doi.org/10.3233/sw-200370.

Strömert, Philip, Vatsal Limbachia, Pooya Oladazimi, Johannes Hunold, and Oliver Koepler. 2023. "Towards a Versatile Terminology Service for Empowering FAIR Research Data: Enabling Ontology Discovery, Design, Curation, and Utilization Across Scientific Communities". In *Knowledge Graphs: Semantics, Machine Learning, and Languages,* 53–69. IOS Press. ISBN: 9781643684253. https://doi.org/10.3233/ssw230005.

Suominen, Osma, Henri Ylikotila, Sini Pessala, Mikko Lappalainen, Matias Frosterus, Jouni Tuominen, Thomas Baker, Caterina Caracciolo, and Armin Retterath. 2015. *Publishing SKOS vocabularies with Skosmos.* Visited on June 25, 2025. https://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf.

Trillitzsch, Tina. 2022. *Ontologie und kontrollierte Vokabulare am ZPID.* Visited on March 31, 2025. Visited on March 31, 2025. https://doi.org/10.23668/psycharchives.8379.

Trojar, Mitja. 2017. "Wüster's View of Terminology". *Slovenski jezik / Slovene Linguistic Studies* 11. Visited on March 31, 2025. https://ojs.zrc-sazu.si/sjsls/article/view/7344.

Xiang, Zuoshuang, Chris Mungall, Alan Ruttenberg, and Yongqun He. 2011. “Ontobee: A Linked Data Server and Browser for Ontology Terms”. In *ICBO: International Conference on Biomedical Ontology.* Visited on June 25, 2025. https://ontobee.org/doc/Ontobee_ICBO-2011_Proceeding.pdf.

Zapilko, Benjamin, Johann Schaible, Philipp Mayr, and Brigitte Mathiak. 2013. “TheSoz: A SKOS representation of the thesaurus for the social sciences”. *Semantic Web* 4 (3): 257–263. https://doi.org/10.3233/sw-2012-0081.

Zemmouchi-Ghomari, Leila, and Abdessamed Réda Ghomari. 2012. “Ontology versus terminology, from the perspective of ontologists”. *International Journal of Web Science* 1 (4): 315. https://doi.org/10.1504/ijws.2012.052531.

Forschungsethik & Datenschutz – unterstützende Geschwister im Forschungsdatenmanagement. Eine Einordnung aus Perspektive der Pflege- und Sozialwissenschaften

Andrea Kuhn [1]

[1] Forschungsnetzwerk Gesundheit an der Hochschule für Wirtschaft und Gesellschaft Ludwigshafen (HWG LU), Germany
Kontakt: FDM Scout im Projekt fdh@haw.rlp

Im gesamten Forschungsprozess ist gute wissenschaftliche Praxis mit rechtlichen und ethischen Aspekten verwoben. Die Forschungsethik reflektiert den Forschungsprozess. Sie schützt die Autonomie der Forschungsteilnehmenden, bewahrt vor Schaden, beantwortet Fragen der Gerechtigkeit und klärt die Verantwortung. In vielen Fachdisziplinen ist ein forschungsethisches Clearing vorgesehen oder gar verbindlich vorgeschrieben. Damit Hand in Hand geht der Schutz der informationellen Selbstbestimmung. Sie ist im Grundgesetz verankert und bereits in den neunzehnhundertachtziger Jahren vom Bundesverfassungsgericht gestärkt worden. Seit 2018 ist die Datenschutzgrundverordnung verbindlich. Trotzdem finden beide ineinandergreifende Felder noch nicht ausreichend Beachtung.

Der rechtssichere und forschungsethisch korrekte Umgang mit Forschungsdaten von Anfang an bildet eine unverzichtbare Grundlage des Forschungsdatenmanagements. Die teilweise neuen normativen Anforderungen werden transparent, Wissen und Schritte des ethischen Clearings werden anhand biomedizinischer Prinzipien skizziert und diverse Unterstützungsoptionen aufgezeigt.

Nur wenn Forschungsethik und Datenschutz bei der Planung beachtet werden, stehen die Folgeschritte des Forschungsdatenlebenszyklus, Datenerhebung, Analyse, Publikation und Archivierung auf rechtssicheren Füßen. Insbesondere für die Datenpublikation als neues

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23930 (CC BY-SA 4.0).

Element der Forschungsaktivitäten ist es unerlässlich, von Beginn an ethische und rechtliche Sicherheit zu schaffen, damit eine Nachnutzung gelingt. Verdeutlicht wird die hohe Relevanz von Forschungsethik und Datenschutz für alle Forschenden. Die Nichtbeachtung der forschungsbezogenen Normen kann Forschungsprojekte gefährden, das Renommee und die Drittmittelstärke von Einrichtungen beschädigen und wissenschaftliche Karrieren einschränken, behindern oder gar vernichten.

Der Beitrag hilft, beim Umgang mit Forschungsdaten im gesamten Forschungsprozess fundierte Entscheidungen zu treffen und rechtlich relevanten Schritte zu beachten. Er bietet Unterstützung für ein zukunftsgerichtetes, sicheres Forschungsdatenmanagement.

Keywords: Ethisches Clearing, Informed Consent, Datenschutz, Forschungsdatenmanagementprozess

1 Forschungsethisches Clearing – Reflexion des Forschungsprozesses

Im Forschungsprozess ergeben sich mannigfaltige forschungsethische Fragen, so zum Beispiel: Welche ethischen Anforderungen sind bei der Planung und dem Design meines Forschungsvorhabens zu beachten? Welche Daten werden wie erhoben? Handelt es sich um personenbezogene Daten? Was muss ich bei der Aufbereitung und Analyse der Daten beachten, wie kann ich hierbei sowohl die Qualität der Daten als auch den Schutz der Forschungsteilnehmenden sicherstellen? Wem „gehören“ die erhobenen Daten? Wie publiziere ich die Forschungsergebnisse korrekt? Wem obliegt die Autorenschaft? Rechtfertigt mein inhaltlicher Beitrag die Autorenschaft? Darf ich vertragsrechtlich die Ergebnisse publizieren? Welche Daten kann ich unter welchen Voraussetzungen teilen und publizieren? Besteht ein Interessenkonflikt der Forschenden? Wer darf Zugriff auf archivierte Daten haben? Unter welchen Bedingungen dürfen Daten nachgenutzt werden? All diese Fragen und noch viele mehr müssen Forschende reflektieren und möglichst gut begründete Antworten finden. Die permanente ethische Reflexion aller Schritte des kompletten Forschungsprozesses wird forschungsethisches Clearing genannt. Es stellt den Schutz der Menschen als Forschungsteilnehmende in den Mittelpunkt forschenden Handelns.

Law is minimal ethics dient als erste Orientierung. Die Aussage meint, dass all die Aspekte, auf die sich eine Mehrheit in einer Gesellschaft oder eines bestimmten Settings einigen kann, in Gesetze, Standards, Leitlinien etc. gefasst werden. Diese Normsetzungen unterstützen die Forschenden. Übergreifend bindend sind beispielsweise die Datenschutzgrundverordnung (DSGVO) als EU-weit verbindliches Gesetz, das Grundgesetz als Verfassung der Bundesrepublik Deutschland, oder die Leitlinien guter wissenschaftlicher Praxis, die die Deutsche Forschungsgemeinschaft (DFG) verfasst hat. Abhängig von Forschungsfrage und Feld können darüber hinaus viele weitere Normsetzungen relevant sein, dies gilt

es jeweils zu prüfen. Unverzichtbar ist die forschungsethische, reflexive Selbstkontrolle zur Einordnung des eigenen Forschungsvorhabens. Werte, Normen, Prinzipien und ggfs. Qualitätsstandards der jeweiligen Disziplin können helfen.

Im Forschungsverlauf umfasst die forschungsethische Selbstkontrolle zudem diverse ethisch relevante Bausteine, die die disziplinäre Fachlichkeit der Forschenden auszeichnen (vgl. Monteverde und Bischofberger 2020): Erstens sind unter der Rubrik Wissenschaftlichkeit die Gütekriterien des jeweiligen Forschungsdesigns genauso wie die Ehrlichkeit und Integrität der Forschenden zu reflektieren. Zweitens soll sich die Forschung insgesamt durch ein positives Nutzen-Risiko-Verhältnis auszeichnen, wobei dieses Verhältnis im besten Fall zugunsten des einzelnen Teilnehmenden ausschlägt. Dies ist aber nicht immer möglich, deshalb ist auch der Nutzen für bestimmte Zielgruppen oder die Gesellschaft als Ganzes eine zu bilanzierende Kategorie. Eng damit verwoben ist die Fairness gegenüber potentiellen Teilnehmenden, die einen Bezug zur Fragestellung haben. Hier ist zu reflektieren, ob sie eine faire Chance zur Teilnahme haben und zu begründen, warum das ggfs. nicht möglich wäre. Der dritte Punkt ist die Erzeugung des Informed Consent der Teilnehmende (siehe Abschnitt 2.3), insbesondere ist darauf zu achten, dass die Teilnehmenden den Forschungszusammenhang verstanden haben und die Einwilligung zum Forschungsvorhaben ohne Druck erfolgt. Die Datenschutzgrundverordnung bildet seit 2018 die rechtliche Normsetzung im Umgang mit den zu erhebenden Daten (siehe Abschnitt 2.4). Viertens sind potentielle Vulnerabilitäten der Teilnehmenden zu reflektieren, um forschungsbedingte Risiken für sie aber ggfs. auch für die Forschenden zu minimieren. Bei sensiblen Forschungsfragen ist ein Ethikvotum erforderlich, das über das eigene ethische Clearing der Forschenden hinaus eine unabhängige externe ethische Begutachtung des Forschungsvorhabens vornimmt. Hierzu stehen in den unterschiedlichen Disziplinen meist Ethikkommissionen bereit. Bei medizinischen Fragestellungen ist das Ethikvotum durch eine definierte Ethikkommission gesetzlich verbindlich in den Heilberufsgesetzen der Länder geregelt. Aber auch bei vielen anderen Fragestellungen der empirischen Forschung wird unterdessen von Drittmittelgebern und Verlagen ein ethisches Gutachten verlangt. Liegt dieses nicht vor, kann die Publikation von Ergebnissen oder die Erteilung von Förderzusagen abgelehnt werden. Ziel sowohl des eigenständigen forschungsethischen Clearings als auch des externen Ethikvotums ist die Fürsorge und das Wohlergehen der Forschungsteilnehmenden. Ihnen ist ein adäquater Schutz vor Schaden und bei potentiell möglichen Vulnerabilitäten ggfs. eine Betreuung sicherzustellen. Das könnte beispielsweise bei hochsensiblen Fragestellungen zu Gewalterfahrungen oder existenziellen Verlusten erforderlich sein. In dieses Spektrum fällt auch, dass Niemand ausgenutzt werden soll. Das könnte beispielsweise dadurch geschehen, dass man den Teilnehmenden die Forschungsergebnisse vorenthält. Dies geht einher mit der Transparenz von möglichen Interessenkonflikten der Forschenden, die spätestens bei der Publikation offenzulegen sind.

Nach diesem überblicksartigen Einstieg vertieft Abschnitt 2 vier ethische und rechtliche Bausteine anhand forschungsethischer Prinzipien.

2 Forschungsethisches Clearing – ethische und rechtliche Bausteine

Abschnitt 2 entfaltet analog der biomedizinethischen Prinzipien Gerechtigkeit, Fürsorge und Schutz, Autonomie, nicht schaden nach Beauchamp und Childress (deutschsprachig siehe Beauchamp und Childress 2024) die forschungsethischen und inhärenten rechtlichen Bausteine des ethischen Clearings. Ausgehend von grundsätzlichen Überlegungen zu unterschiedlichen Dimensionen der Gerechtigkeit (Abschnitt 2.1) werden die ethischen Dimensionen von Fürsorge, Schutz und Wohlergehen der Menschen im Forschungsprozess aufgenommen und Vulnerabilitäten analysiert (Abschnitt 2.2). Abschnitt 2.3 schließt mit dem Informed Consent an, dem Instrument zur Stärkung der Autonomie von Forschungsteilnehmenden. Abschließend zeigt Abschnitt 2.4 den gesetzlich normierten Datenschutz als ein zentrales Instrument im ethischen Clearing, um die Rechte der Teilnehmenden an ihren Daten zu schützen und sie aus dieser Perspektive vor Schaden zu bewahren.

2.1 Grundsätzliche Überlegungen zur Gerechtigkeit

Der erste forschungsethische Baustein beinhaltet grundsätzliche Reflexionen zur Gerechtigkeit in der Planungsphase der Studie. Im ersten Schritt ist die Frage der Notwendigkeit der angedachten Forschung zu beantworten. Besteht tatsächlich eine Wissenslücke, die eine neue Erhebung rechtfertigt oder liegt bereits gesicherte Empirie zur Frage vor oder gibt es (ggfs. publizierte) Datensätze, die über eine Sekundäranalyse die Fragestellung hinreichend beantworten können?

Die zweite gerechtigkeitsethische Frage fokussiert die Notwendigkeit einer Drittmittelfinanzierung des geplanten Forschungsvorhabens. Hier ist aus gesamtgesellschaftlicher Perspektive die Relevanz der Forschung und der damit verbundene Mitteleinsatz aus Steuermitteln zu rechtfertigen. Dabei ist der aktuelle politische Wille zu reflektieren, es kann durchaus vorkommen, dass dieser der wissenschaftlich-fachlichen Einschätzung zur Relevanz des Forschungsvorhabens zuwiderläuft. An der Stelle befinden sich Forschende in grundsätzlichen Reflexionsprozessen zum Einsatz und zur inhärenten Steuerung von Forschung über die (ggfs. politikgetriebene) Vergabe von Drittmitteln. Forschende müssen eine Begründung gegenüber Drittmittelgebern liefern, warum ihre Forschung gerechtfertigter Weise zu finanzieren wäre.

Der dritte Schritt nimmt den angestrebten Erkenntnisgewinn in den Blick. Dazu ist die Passgenauigkeit des angedachten Designs zu reflektieren. In den Sozialwissenschaften wäre beispielsweise zu fragen, ob der erforderliche Erkenntnisgewinn methodisch besser mit einem quantitativen oder einem qualitativen Design zu erreichen ist, ob die Synthese beider Methoden zielführend wäre oder ob es eher eine partizipative Ausrichtung der Forschung braucht (u. a.). Ein weiterer Punkt betrifft die Neugier der Forschenden: Man muss sich

im Sinne der Datensparsamkeit fragen, ob wirklich alle erhobenen Daten nötig zur Beantwortung der konkreten Forschungsfrage sind. Dies betrifft insbesondere personenbezogene Daten.

Die grundsätzliche Frage zum angestrebten Erkenntnisgewinn ist auch, ob zur Beantwortung der Forschungsfrage eine theoriesynthetisierende Quellenarbeit als zweite, leider in vielen Disziplinen oft vernachlässigte Aufgabe wissenschaftlichen Arbeitens zielführend(er) wäre, um einen breiteren Erkenntnisgewinn abzusichern als eine einmalige oder wiederholte empirische Querschnittserhebungen zu einzelnen Facetten des Problems inkl. der damit verbundenen Belastung der Befragten.

Zudem wäre die Sekundärnutzung von bereits erhobenen Datensätzen zu prüfen. Dies können sowohl eigene Datensätze als auch publizierte Datensätze sein. Das in vielen Disziplinen neue Forschungsdatenmanagement zeigt den strukturierten Prozess zu diesem ressourcenschonenden Herangehen auf. Es ergeben sich neue Möglichkeiten, einzelne Datensätze zu größeren Datensätzen zusammenzuführen gemeinsam zu betrachten und zu einer neuen Art der „Langzeitbeobachtung“ zusammenzufügen, wobei allerdings die Aussagekraft abhängig von der Fragestellung zu überprüfen bleibt.

Allerdings gilt zu beachten, dass die Optionen der Sekundärnutzung von bei Menschen erhobenen Daten wiederum dem Informed Consent unterliegen, d. h. die Zustimmung der befragten Personen muss vorliegen, ansonsten ist die Zweitverwertung weder rechtssicher noch forschungsethisch vertretbar. Die personenbezogenen Aspekte von Menschen als Teilnehmende im Forschungsprozess vertiefen die folgenden Abschnitte.

2.2 Fürsorge, Schutz und Wohlergehen – Menschen im Forschungsprozess

Der zweite forschungsethische Baustein nimmt die Zielgruppe und ihre Vulnerabilitäten sowie das Forschungsdesign und die Gütekriterien der Forschung in den Blick. Die wissenschaftliche Qualität eines Forschungsprojekts gewinnt durch forschungsethisches Clearing.

Die grundsätzliche Frage ist die nach dem Einbezug von Menschen in die Forschung. Hierbei ist zu beachten, dass Menschen mittelbar und unmittelbar einbezogen werden können. Mittelbar betrifft Menschen die Forschung, wenn sie die Nutzenden sind, z. B. bei neuer Software, beim autonomen Fahren, bei neuen Bauwerken oder neuer Haustechnik und vielem mehr. Erforderlich sind in diesen technischen Anwendungsfeldern z. B. der Ingenieurwissenschaften und IT-Wissenschaften eine Technikfolgenabschätzung und daraus abgeleitete Sicherheitsauflagen. Diese können je nach Disziplin und Standard unterschiedlich sein.

Unmittelbar betroffen sind Menschen, wenn sie direkt an der Forschung teilnehmen, beispielsweise als Proband*in in der Medizin- und Arzneimittelforschung oder als Forschungsteilnehmende in der Sozial- und Wirtschaftsforschung. In diesen Designs braucht es immer die umfassende Information über die Forschung, die Zustimmung der Teilnehmenden, den sog. Informed Consent (siehe Abschnitt 2.3) und die Datenschutzerklärung (siehe Abschnitt 2.4). Regelhaft ist ein Ethikvotum vor Beginn der Forschung einzuholen (siehe Abschnitt 1), das die eigene forschungsethische Reflexion stützt und die Teilnehmenden vor Missbrauch schützt. Maßgeblich sind hier sowohl gesetzliche Normen als auch disziplinabhängige Vorgehensweisen. Informationen und Ansprechpartner*innen findet man bei den entsprechenden Gremien der Fachdisziplinen (siehe auch Abschnitt 4).

Bei der Konzeption von Forschungsdesigns an und mit Menschen ist deren Vulnerabilität das zentrale Kriterium. So sind die Zielgruppe der Forschung sowie die Ein- und Ausschlusskriterien in Bezug auf Vulnerabilitäten zu reflektieren und es muss entschieden werden, ob der Einschluss einer grundsätzlich vulnerablen Gruppe, wie z. B. Kinder oder demenziell erkrankte Menschen, erforderlich ist. Umfangreiche forschungsethische Hilfe bietet die Biomedizinkonvention des Europarates sowie ihre Zusatzprotokolle und Erläuterungen. Eine Übersicht bietet das Infoportal Bioethik-Konvention.de. Dort werden verschiedene potentielle Vulnerabilitäten aufgezeigt, die nicht direkt augenscheinlich sind, mit denen sich Forschende aber konstruktiv auseinandersetzen müssen, um gut damit umgehen zu können. Monteverde und Bischofberger (2020) haben die Vulnerabilitätstypen mit Beispielen hinterlegt:

- **kognitive Vulnerabilität**: urteilsunfähige Menschen: z. B. Neugeborene, Kleinkinder, Menschen mit Demenz, geistige Behinderung oder im Wachkoma.
- **situative Vulnerabilität**: Menschen, die aufgrund einer Notfallsituation oder Sprachbarrieren nicht einwilligen können.
- **institutionelle Vulnerabilität** urteilsfähige Menschen, die in Institutionen unter dem Einfluss Dritter stehen (z. B. Armeeangehörige, Gefangene).
- **autoritätsbezogene Vulnerabilität**: urteilsfähige Menschen, die unter dem Einfluss Dritter stehen (z. B. Angehörige, Fachperson).
- **medizinische Vulnerabilität**: Patient*innen, für die keine wirksame Therapien zur Verfügung stehen.
- **ökonomische Vulnerabilität**: Menschen, die aus einer finanziellen Notlage heraus an Forschungsprojekten teilnehmen.
- **soziale Vulnerabilität**: Angehörige von Randgruppen (z. B. Obdachlose, Sans-Papiers, siehe Monteverde und Bischofberger 2020, S. 190).

Die im forschungsethischen Clearing herausgearbeiteten Vulnerabilitätsaspekte wirken sich auch auf die Wahl des Designs aus. Je nachdem ob beispielsweise Messungen, Befragungen oder Beobachtungen durchgeführt werden, sind andere Punkte zu beachten. Die Messinstrumente sind einer kritischen Reflexion zu unterziehen. Sie müssen zielgruppenspezifisch sein, d. h. verständlich und zum Setting passend. Bei Laborversuchsanordnungen sollen möglichst ungefährlich sein, damit niemand zu Schaden kommt, weder Teilnehmen-

de noch Forschende oder Dritte. Darüber hinaus sind vertiefend viele fachspezifische Anforderung zu berücksichtigen, deren detaillierte Darlegung den Rahmen dieses Beitrages sprengen würde.

Ein weiteres Thema ist die wissenschaftliche Qualität der Forschung. In der qualitativen Forschung ist diese über die Kriterien Glaubwürdigkeit, Authentizität, Übertragbarkeit, Verlässlichkeit und die Bestätigung der Teilnehmenden nachvollziehen. In der quantitativen Forschung sind die Gütekriterien Validität, Reliabilität und Objektivität ausschlaggebend.

Eine Quellenarbeit kann in einer eigenen Studie hinderliche oder gar vom Forschenden verursachte Vulnerabilitäten zwar umgehen. Hier gilt es aber die in der jeweiligen in die Quellenanalyse eingeschlossenen Studie inhärenten Vulnerabilitäten zu reflektieren, auch um ggfs. einen *Bias* in den vorliegenden Ergebnissen zu erkennen und in den theoretischen Ableitungen sichtbar zu machen.

Alle angeleuchteten Aspekte sind im Forschungsdatenmanagement hochrelevant. Wenn gewonnene Daten durch Nichtbeachtung von Vulnerabilitäten beeinflusst oder gar verfälscht wurden, die Messinstrumente unverständlich waren oder durch Antwortzwang oder Unverständlichkeit ein inhärentes Bias erzeugten, Laboruntersuchungen Menschen schaden könnten oder die Qualität von Forschungsprojekten nicht nachvollziehbar ist, können solche Datensätze nicht gewinnbringend weitergenutzt werden. Vielmehr können schlecht synthetisierte Datensätze mehr Schaden verursachen und zu weitreichenden Fehlinterpretationen führen als einzelne kleine Querschnittsstudien das könnten. In diesem Kontext steigt die Verantwortung von Forschenden, die ihre Daten publizieren wollen, erheblich.

2.3 Informed Consent – partizipativer Prozess zur Stärkung der Autonomie

Die informierte Einwilligung oder Informed Consent umfasst grob drei Prozessschritte: Erstens die umfassende Information der potentiellen Forschungsteilnehmenden; zweitens die ausdrückliche Einwilligung der Forschungsteilnehmenden, die mitmachen wollen; drittens die Datenerhebung bei diesen Personen. Wenn keine Einwilligung besteht, ist die Erhebung und Nutzung der Daten in den allermeisten Fällen rechtswidrig. Nur in besonderen Einzelfällen ist es forschungsethisch zu rechtfertigen, den Informed Consent zumindest teilweise zu umgehen, dann braucht es aber sicher ein Ethikvotum und eine sehr gute Begründung.

Die Information zur Einwilligung muss mindestes folgende Informationen enthalten: Die Beschreibung des Forschungsvorhabens, des Forschungszweckes und des Ziels der Forschung; die Benennung der Zielgruppe bzw. der Forschungsteilnehmenden und des Settings, in dem die Forschung stattfindet, sowie die Kontaktdaten der Forschenden.

Das ethische Instrument des Informed Consent zur Stärkung der Autonomie der Forschungsteilnehmenden ist nach 1945 seit Jahrzehnten in den internationalen Forschungskonventionen hinterlegt. Seit 2018 erfährt er Stärkung und gesetzliche Verbindlichkeit in der DSGVO. Dort sind folgende rechtliche Grundlagen normiert: Die Erhebung (sensibler) personenbezogener Daten, dabei gilt zu beachten, dass es sich um mehr als nur den Namen handelt. Der Verwendungszweck, die Verarbeitung, die Aufbewahrung und der Zugriff der Daten muss transparent gemacht werden, dazu sind umsetzbare konkrete Aussagen zu treffen, z. B. zu Verfahren der Pseudonymisierung oder Anonymisierung, zur Zugriffssicherheit, zur Archivierungsform und Archivierungszeit. Auch über die Veröffentlichung der Forschungsergebnisse und ggfs. der aufbereiteten Datensätze ist der Forschungsteilnehmende zu informieren und seine Zustimmung einzuholen. Zu beachten ist in diesem Zusammenhang die integre Autor*innenschaft. Ebenso ist der Teilnehmende darüber zu informieren, wenn Daten und Forschungsergebnisse nur intern genutzt und nicht veröffentlicht werden sollen. Anhand all dieser Informationen werden Teilnehmende befähigt über ihre Teilnahme zu entscheiden.

Die DSGVO gewährt Forschenden eine Privilegierung der Datenverarbeitung zu Forschungszwecken, den sogenannten *Broad Consent*. In der Praxis bedeutet dies, dass beispielsweise die Option einer Sekundäranalyse benannt sein muss, aber noch nicht genau, was wann später analysiert werden wird. Diese Freiheit der Forschenden birgt eine große Verantwortung gegenüber den Teilnehmenden. Zudem scheinen sich Aufsichtsbehörden und Datenschutzbeauftragte teilweise uneinig zu sein, wie genau die Zweckbindung erläutert werden muss bzw. wie frei Forschende in der Nachverwertung sind. Es ist noch viel Bewegung in den juristischen und politischen Interpretationen festzustellen. Wichtig zu beachten ist an der Stelle, dass nur Forschende dieses Privileg innehaben. Sobald Befragungen durchgeführt werden, die nicht für sich den Anspruch einer Forschung erheben, sind die Möglichkeiten der Durchführenden bedeutend enger, gerade weil die „Laien-Durchführenden“ nicht den kontrollierenden Anforderungen guten wissenschaftlichen Arbeitens unterliegen.

Abschließend muss jedes Informed Consent-Formular Rechtshinweise für die Teilnehmenden enthalten. Sie sind darüber zu informieren, wer verantwortlich für die Forschung ist (meist die Instituts- oder Einrichtungsleitung, seltener die Forschungsleitung), wo sie Nachfragen stellen können (meist die Forschungsleitung) und wo sie sich ggfs. beschweren können, falls sie denken, dass gegen ihre Rechte verstoßen wurde. Hierzu sind die Datenschutzbeauftragten der Institution und des jeweiligen Bundeslandes Ansprechpartner. Sie sind im Formular aufzuführen.

Der dritte Schritt des Informed Consent ist die Erklärung. Sie muss freiwillig zustande kommen, es darf niemand aus welchen Gründen auch immer zur Teilnahme genötigt werden (vgl. Abschnitt 2.2 zu Vulnerabilität). Weiterhin muss Forschungsteilnehmenden ein Widerrufsrecht gewährt werden: Sie dürfen ohne Konsequenzen ihre Teilnahme am Forschungsprojekt zurückziehen, ihre ggfs. bereits erhobenen Daten sind in dem Fall zu löschen. Allerdings besteht an dem Punkt die Einschränkung der Rückverfolgbarkeit. Soll-

ten die Daten bereits anonymisiert worden sein und eine Rückverfolgbarkeit somit ausgeschlossen sein, wäre das Löschen der Daten nicht mehr möglich. Auch darüber ist der Teilnehmende in Kenntnis zu setzen.

Die Bestätigung des Informed Consent sollte möglichst schriftlich mit Datum und Unterschrift erfolgen. Davon abweichend können besondere Anforderungen in qualitativen Designs nötig sein, diese gilt es dann im forschungsethischen Clearing nachvollziehbar zu begründen. Alternative Anforderungen ergeben sich bei digitalen bzw. Online-Befragungen. Erforderlich ist hier eine vorgeschaltete Einwilligungs- und Datenschutzerklärung beispielsweise in der Einladungsmail zur Befragung. Die Befragung ist erst nach der Zustimmung, z. B. über das aktive Setzen eines Kreuzes oder Hakens, zu öffnen.

Selbst wenn die Teilnehmenden der Forschung informiert zugestimmt haben, ist ihre Autonomie in der Folge während der Durchführung der Studie zu wahren. So sollen sie beispielsweise bei einer Befragung nicht zur Beantwortung aller Fragen gezwungen werden. Vielmehr dürfen sie einzelne Fragen auslassen, wenn sie dazu keine Informationen haben oder nicht antworten möchten. Beachtet man dies in der Instrumentenkonstruktion nicht, brechen Befragte entweder die Befragung ab oder kreuzen einfach irgendeine Antwortoption an, um im Fragebogen fortfahren zu können. Dieses z. B. auf Neugier, Pflichtbewusstsein oder soziale Erwünschtheit gründende Verhalten verfälscht die Ergebnisse, ersteres führt zu hohen Abbruchraten. Aus all diesen Gründen müssen Fragebögen einen Pretest durchlaufen, bevor sie eingesetzt werden. Wer keine Zeit für einen Pretest hat, hat keine Zeit für Forschung.

Auch in vielen anderen Designs, wie z. B. Medikamententestreihen, teilnehmenden Beobachtungen, Fokusgruppendiskussionen, Techniktests und vielen anderen mehr können Teilnehmende jederzeit ihre Einwilligung zur Teilnahme zurückziehen.

Mit der Etablierung des Forschungsdatenmanagements zeigte sich die Notwendigkeit für die Unterstützung zu den Fragen des Informed Consent. Einige unterstützende Tools zeigt Abschnitt 4 auf. Eng mit dem Informed Consent hängt der Datenschutz zusammen, wie in Abschnitt 2.4 erläutert wird.

2.4 Datenschutz – Nicht Schaden und Schutz der Teilnehmenden-Rechte

Der vierte Baustein des forschungsethischen Clearings umfasst den Datenschutz. Zwar lassen sich heute Überschneidungen zum in der Wissenschaft nach den Erfahrungen vor 1945 immer stärker standardisierten ethisch fundierten Informed Consent zur Gewährung der Autonomie der Forschungsteilnehmenden erkennen (siehe Abschnitt 2.3), jedoch kommt der Datenschutz aus der gesetzlichen Normierung des Nicht-Schadens-Prinzips als negative Unterlassungspflicht. Heute befruchten sich die ethischen Prinzipien der Autonomie

und des Nicht Schadens. Sie wirken zusammen als Schutz der Forschungsteilnehmenden. Die Wurzeln liegen im Grundgesetz, welches seit 1949 das allgemeine Persönlichkeitsrecht in Art. 2 Abs. 1 in Verbindung mit der unantastbaren Menschenwürde in Art. 1 Abs. 1 sichert (Bundesrepublik Deutschland 1949). Diese Menschenrechte wurden 1983 vom Bundesverfassungsgericht aufgrund der rasant zunehmenden neuen Möglichkeiten moderner Datenverarbeitung um das Recht auf informationelle Selbstbestimmung erweitert, in den Leitsätzen heißt es:

> „*1. Unter den Bedingungen der modernen Datenverarbeitung wird der Schutz des Einzelnen gegen unbegrenzte Erhebung, Speicherung, Verwendung und Weitergabe seiner persönlichen Daten von dem allgemeinen Persönlichkeitsrecht des GG Art 2 Abs. 1 in Verbindung mit GG Art 1 Abs. 1 umfaßt. Das Grundrecht gewährleistet insoweit die Befugnis des Einzelnen, grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten zu bestimmen.*“ (Bundesverfassungsgericht 1983)

In 2016 verabschiedete die Europäische Union die General Data Protection Regulation (GDPR) zu Deutsch die Datenschutzgrundverordnung (DSGVO), die seit 2018 verbindlich ist (Europäische Union 2016). Eine für die Forschung zentrale Regelung umfasst den Schutz der personenbezogenen Daten. Nach Art. 4 Abs. 1 DSGVO handelt es sich dabei um alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. Es betrifft direkte Identifikatoren wie den Namen, die Adresse, die Stimme, das Bild aber auch indirekte Identifikatoren wie den Ort, das Geschlecht, die Ausbildung, die Schulart, Kennnummern (z. B. Matrikelnummer), IP-Adressen, das Geburtsdatum, den Jahrgang (u. a.). Diese Daten ermöglichen, zumindest in Korrelation mit anderen in der Befragung angegebenen Daten, die Identifizierung der Person. Weiterhin schützt die DSGVO sensible personenbezogene Daten in Art. 9 Abs. 1, dazu gehören die rassische und ethnische Herkunft, politische Meinungen sowie religiöse und weltanschauliche Überzeugungen, eine Gewerkschaftszugehörigkeit, genetische und biometrische Daten, Gesundheitsdaten, Daten zu Sexualleben und zur sexuellen Orientierung (u. a.). Falls die Erhebung einer dieser Datenkategorien vorgesehen ist oder nicht ausgeschlossen werden kann, muss eine Datenschutzerklärung beigefügt werden. Die Erhebung personenbezogener (sensibler) Daten darf nur mit Einwilligung der natürlichen Person erfolgen, sie sind in Art. 6 Abs. 1, Art. 9 Abs. 2a normiert. Ausnahmen von dieser Pflicht sind in Art. 9 Abs. 2b ff. und in weiteren Artikeln der DSGVO niedergelegt. Hierunter fallen u. a. Gesundheitsdaten, die im öffentlichen Interesse stehen, wie z. B. die Datenerhebung während der Coronapandemie oder Daten, die zum Schutze anderer Menschen notwendig sind. Die Erhebungszwecke der Daten müssen, wie beim Informed Consent, der Datenschutzerklärung zu entnehmen sein, ansonsten ist die Erhebung rechtswidrig. Vorsicht! Selbst bei Onlinebefragungen, die keine personenbezogenen Daten im Fragekatalog erheben, werden die IP-Adressen zumindest kurzfristig gespeichert, manchmal auch aus Sicherheitsgründen längerfristig, auch dies ist den potentiell Teilnehmenden mitzuteilen.

Bei der Umsetzung der datenschutzrechtlichen Anforderungen im Forschungsprozess von der Analyse, der Speicherung über die Publikation bis zur Archivierung der Daten helfen Zugangs- und Zugriffsbeschränkungen wie Pseudonymisierung und Anonymisierung, ein Passwortschutz, die Verschlüsselung von Dateien, das mehrfache Speichern auf sicheren Medien, die sichere Archivierung und der Zugang nur für einen definierten Personenkreis. Falls die Daten nicht mehr gebraucht werden, sind sie datenschutzsicher zu vernichten.

Unterstützung bieten die Datenschutzbeauftragten der Forschungseinrichtungen und Hochschulen, FDM-Stewards und Scouts an Hochschulen, sofern vorhanden, sowie diverse FDM-Informationsangebote im Internet wie beispielsweise die disziplinären Beratungsangebote der NFDI-Helpdesks (Nationale Forschungsdateninfrastruktur e.V. – NFDI – o.J.).

Im forschungsethischen Clearing gehen Information und Datenschutz Hand in Hand. In Abschnitt 3 werden die verschränkten Anforderungen anhand eines Beispiels, der Forschung mit Kindern, aufgezeigt.

3 Informierte Zustimmung, Einwilligungsfähigkeit & Datenschutz bei Kindern – Ein Beispiel

In Forschungsprojekten mit Kindern entfaltet sich die Bandbreite des zuvor theoretisch dargelegten Vorgehens im ethischen Clearing. *Kinder sind nicht gleich Kinder* lautet die erste Reflexionsstufe. Hochrelevant beim forschenden Vorgehen ist ihr Alter und ihre entwicklungsabhängige Einwilligungsfähigkeit. Zudem kommen den Eltern Rechte zu, die ebenfalls nicht statisch sind. Also haben es Forschende mit mindestens zwei Parteien zu tun. Das primäre Kategoriensystem der Reflexion der Anforderungen, was wann zu tun ist, ist das Alter (Nachfolgendes in Anlehnung an Verbund Forschungsdaten Bildung[1]). Für die Forschung an und mit unter 14-Jährigen ist die Einwilligung der Eltern erforderlich und aus rechtlicher Sicht auch allein ausreichend. Forschungsethisch ist jedoch zusätzlich die informelle Zustimmung der Kinder geboten, um ihre Instrumentalisierung zu vermeiden (siehe Abschnitt 2.2). Die Kinder sind in die Anliegen der Forschungsprojekte einzubeziehen, das kann beispielsweise mittels kindgerechter Erklärungen erfolgen. Forschende sollten nicht auf einer Teilnahme bestehen, wenn zwar die Eltern zugestimmt haben, das Kind aber augenscheinlich kein Interesse oder gar Angst vor der Teilnahme hat oder dieses nonverbal kommuniziert. Bei 14 bis 17-Jährigen wird es noch komplexer, weil Kinder unterschiedlich weit in ihrer Entwicklung sind. Unter 16 Jahren muss aus rechtlicher Sicht immer die Einwilligung der Eltern vorliegen. Zusätzlich muss die Einwilligung der Kinder vorliegen, falls diese in Bezug zu den Inhalten der Forschung einwilligungsfähig sind. Falls sie dazu nicht einwilligungsfähig sein sollten, soll trotzdem die informelle Zu-

1 https://www.forschungsdaten-bildung.de/datenmanagement/recht-ethik/datenschutz-forschung/; *Besucht am 19. Mai 2025.*

stimmung zur Teilnahme gegeben sein, damit kein Kind zur Teilnahme gezwungen wird, auch nicht durch die Eltern (vgl. Vulnerabilität in Abschnitt 2.2). Über 16-Jährige können eigenständig einwilligen, sofern sie zur Sache einwilligungsfähig sind. Hier können sich die Anforderungen drehen, denn es kann durchaus möglich sein, dass die Jugendlichen einer Forschung zustimmen, die Eltern aber dagegen sind. In solchen Situationen braucht es gute ethische Aushandlungsprozesse, bei denen das Wohl des Jugendlichen im Mittelpunkt steht. Bei 18-Jährigen könnte man denken, nun wären alle Probleme gelöst, denn volljährige Personen sind rechtlich eigenständig einwilligungsfähig. Allerdings kann es aufgrund von Behinderungen und Entwicklungsstörungen auch in der Altersgruppe und in weiteren Altersgruppen im Lebensverlauf vorkommen, dass die Einwilligung der Eltern einzuholen ist, sofern sie die Vorsorgevollmacht oder die Betreuungsverfügung innehaben. Die Betreuungsverfügung kann aber auch bei amtlich bestellten Betreuenden liegen, dann ist deren Einwilligung einzuholen. Nach wie vor müssen die Forschungsteilnehmenden mindestens informell zustimmen, niemand darf zur Teilnahme gezwungen werden. Diese Aspekte sind im ethischen Clearing des Einzelfalls zu klären.

In Forschungsprojekten mit Kindern und eingeschränkt einwilligungsfähigen Personen ist der Datenschutz über alle Altersgruppen hinweg im gesamten Forschungsprozess besonders sensibel zu beachten. Zudem ist fast immer ein Ethikvotum erforderlich. Kinder genießen, wie beeinträchtigte Personen auch, einen besonderen Schutz, der in vielen Gesetzeswerken normiert ist (z. B. Grundgesetz, Bürgerliches Gesetzbuch, Schulgesetze, Heilberufsgesetze der Länder, Betreuungsrecht, etc.) und von diversen Behörden (z. B. Schulaufsichtsbehörden, Heilberufekammern, Landesämter, etc.) überwacht wird.

Abschnitt 4 gibt einen exemplarischen Einblick in Unterstützungsinstrumente für Forschende, damit ethisch gute Forschung und Datenmanagement trotz der skizzierten hohen Anforderungen gelingen können.

4 Unterstützung für Forschende – eine kleine Auswahl

Unterstützung finden Forschende an vielen Stellen. Bezogen auf das Forschungsdatenmanagement wird hier eine kleine exemplarische Auswahl vorgestellt, die weder den Anspruch auf Vollständigkeit noch auf generelle Qualitätsbewertung erhebt.

Die Homepage des Rats für Sozial und WirtschaftsDaten (RatSWD)[2] bietet umfangreiche Informationen, die in langjähriger Arbeit zusammengetragen wurden. Dort finden sich

2 https://www.konsortswd.de/ueber-uns/ratswd/; *Besucht am 26. Juni 2025 durch die Herausgeber*innen.*

- Checklisten rund um das Thema Datenschutz[3]
- Das Thema Datenschutz wird nochmals aufgearbeitet[4]
- Zu Best Practice Forschungsethik sind diverse Materialien eingestellt[5]
- Weiterhin wird eine Übersicht zu Ethikkommissionen unterschiedlicher Disziplinen und Institutionen angeboten, die allerdings nicht den Anspruch auf Vollständigkeit erhebt[6]
- Zum Informed Consent sind ebenfalls Unterlagen hinterlegt[7]
- Ergänzend finden sich umfangreiche Publikationen zum gesamten Themenkomplex des forschungsethischen Clearings[8]

Der Ethics and Data Protection Decision Tree ist ein interaktives Tool, bereitgestellt von der Europäischen Union. Er unterstützt bei der Identifikation potentieller ethischer Risiken im Forschungsprozess, fördert das ethische Clearing und ermöglicht den sicheren Umgang mit den Vorgaben des Datenschutzes. Angelegt wurde er zum Schutz der fundamentalen Menschenrechte und Freiheiten der Forschungsteilnehmenden. Zu finden ist das Tool unter https://ec.europa.eu/assets/rtd/ethics-data-protection-decision-tree/index.html[9].

Darüber hinaus gibt es viele weitere Seiten zur Unterstützung unterschiedlichster Forschungsprozesse aus der Community des Forschungsdatenmanagements und bei den Fachdisziplinen. Diese sind in persönlicher Recherche zu entdecken. Abschnitt 5 schließt mit dem Fazit, warum es angeraten ist, diese Instrumente zu nutzen.

5 Fazit – Warum ist das forschungsethische Clearing unabdingbar?

Im Forschungsprozess und somit auch im Forschungsdatenmanagement steht das forschungsethische Clearing ganz am Anfang und es begleitet im Folgenden alle Schritte im Forschungsprozess und des inhärenten Datenmanagementprozesses. Wenn die Grundlagen nicht gleich zu Anfang in der Konzeption des Forschungsprojektes gelegt werden,

3 https://www.konsortswd.de/themen/best%20practice-forschungsethik/datenschutz%20forschungsethik/; *Besucht am 25. Juni 2025 durch die Herausgeber*innen.*

4 https://www.konsortswd.de/themen/datenschutz/; *Besucht am 25. Juni 2025 durch die Herausgeber*innen.*

5 https://www.konsortswd.de/themen/best%20practice-forschungsethik/; *Besucht am 25. Juni 2025 durch die Herausgeber*innen.*

6 https://www.konsortswd.de/themen/forschungsethik/ethikkommissionen/; *Besucht am 25. Juni 2025 durch die Herausgeber*innen.*

7 https://www.konsortswd.de/themen/best%20practice-forschungsethik/einwilligungsunterlagen/; *Besucht am 25. Juni 2025 durch die Herausgeber*innen.*

8 https://www.konsortswd.de/themen/forschungsethik/; *Besucht am 25. Juni 2025 durch die Herausgeber*innen.*

9 *Besucht am 25. Juni 2025 durch die Herausgeber*innen.*

steht der gesamte Prozess auf tönernen Füßen. Rechtliche und ethische Unterlassungen und Fehler lassen sich nicht im Nachgang heilen. Die Publikation der Forschungsergebnisse ist gefährdet, wenn beispielsweise nötige Ethikanträge bei Vulnerabilitäten nicht gestellt wurden. Die Sekundärnutzung und Publikation von Datensätzen zur Nachnutzung sind unzulässig, wenn nicht die passende Information und Zustimmung erfolgte. Eine weitere Gefahr wäre der Abbruch und das Scheitern des gesamten Forschungsprojektes bis hin zur Zurückzahlung bereits erhaltener Drittmittel, weil Daten unrechtmäßig erhoben wurden. Forschenden sind sicher diverse konkrete Beispiele zu diesen generisch geschilderten Konsequenzen bekannt. Die Darlegung spezieller, der Autorin bekannter Beispiele in einer Publikation verbietet sich aus forschungsethischen Gründen. Solche Vorkommnisse haben definitiv negative Auswirkungen auf die wissenschaftliche Karriere der Forschenden. Sie verursachen gleichzeitig Reputationsschäden für Forschungsinstitute und Forschungseinrichtungen an Hochschulen, die schwer wettzumachen sind. Dabei ist forschungsethisches Clearing keine „Raketenwissenschaft“, sondern ein permanenter Reflexions- und Begründungsprozess im Sinne guter wissenschaftlicher Praxis, der zu guten Forschungsprojekten, belastbaren Forschungsergebnissen und wertvollen Forschungsdaten führt. Viel Erfolg!

Danksagungen

Ich danke dem Projektteam von fdm@haw.rlp für die Unterstützung bei der Entwicklung der Inhalte, der gemeinsamen Reflexion mit vielen konstruktiven Hinweisen. Ohne euch hätte dieser Beitrag nicht auf den E-Science-Tagen präsentiert werden können.

Interessenkonflikte

Die Autorin gibt an, dass kein Interessenkonflikt besteht.

Literaturverzeichnis

Beauchamp, Tom L. und James F. Childress. 2024. *Prinzipien Der Bioethik.* 600. Verlag Karl Alber. ISBN: 9783495492437.

Bundesrepublik Deutschland. 1949. *Grundgesetz für die Bundesrepublik Deutschland.* Besucht am 27. März 2025. https://www.gesetze-im-internet.de/gg/GG.pdf.

Bundesverfassungsgericht. 1983. *Urteil vom 15. Dezember 1983.* Besucht am 25. Juni 2025. https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/1983/12/rs19831215_1bvr020983.html.

Europäische Union. 2016. *Verordnung (EU) 2026 / 679 des Europäischen Parlaments und des Rates – vom 27. April 2016 – zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95 / 46 EG (Datenschutz-Grundverordnung).* Besucht am 5. März 2025. https://www.dfg.de/resource/blob/173732/4166759430af8dc2256f0fa54e009f03/kodex-gwp-data.pdf.

Monteverde, Settimio und Iren Bischofberger. 2020. „Forschung, Pflege und Ethik“. In *Handbuch Pflegeethik. Ethisch denken und handeln in den Praxisfeldern der Pflege,* 2. Aufl., herausgegeben von Settimio Monteverde, 176–195. Verlag W. Kohlhammer. ISBN: 9783170359246.

FDM Meets HAW: Lessons Learned bei der Begleitung von Projekten der angewandten Forschung

Sarah Boelter [1,*], Romy Meyer[2,*], Wesley Preßler [1], Lucie Schmidt [1], Amélie Oberdorfer[2], Christoph Schmidt [2]

[1] Ernst-Abbe-Hochschule Jena, Jena, Deutschland;
[2] Hochschule Nordhausen, Nordhausen, Deutschland;
[*] Kontakt: Sarah Boelter (sarah.boelter@eah-jena.de) und Romy Meyer (romy.meyer@hs-nordhausen.de)

Die FAIR-Prinzipien haben sich als Standard für einen nachhaltigen Umgang mit Daten in der Forschung etabliert und gelten als eine Grundlage der guten wissenschaftlichen Praxis. Daher gewinnt Forschungsdatenmanagement, wie an vielen Forschungseinrichtungen, auch an Hochschulen für angewandte Wissenschaften (HAW) zunehmend an Bedeutung. Insbesondere strukturelle und personelle Rahmenbedingungen an HAW, wie höhere Lehrdeputate und eine Abhängigkeit von Drittmitteln für Forschungsvorhaben, stellen diese Einrichtungen vor besondere Herausforderungen beim Aufbau eines institutionellen FDM.

Dieser Artikel beleuchtet die spezifischen Herausforderungen und Potenziale des FDM an HAW und legt dabei die Erkenntnisse aus zwei begleiteten Forschungsprojekten in Thüringen zugrunde. Durch eine retrospektive Analyse der FDM-Begleitung dieser Projekte, einem sozial- und einem ingenieurwissenschaftlichen Projekt, zeigen sich gleichartige Problemfelder und auch Lösungsansätze. Die Forschenden wünschen sich praxisnahe und handlungsleitende Handreichungen zu FDM-Ressourcen, die einen vereinfachten Zugang zu FDM Themen versprechen. Die Unterstützung durch (über-) regionale FDM-Initiativen und die Verfügbarkeit von dediziertem FDM-Personal vor Ort an den HAW sind kritische Faktoren für den erfolgreichen Aufbau von FDM.

Abschließend wird die Notwendigkeit und Chance hervorgehoben, FDM als integralen Bestandteil der Forschungskultur an HAW zu begreifen und hochschulweit zu verankern.

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23931 (CC BY-SA 4.0).

Zudem werden Möglichkeiten der Synergien zwischen FDM und weiteren Themen sowie des Ausbaus (über-)regionaler Kooperationen aufgezeigt.

Keywords: Hochschule für angewandte Wissenschaften, Forschungsdatenmanagement, Bedarfserhebung, Retrospektive

1 Relevanz und Rahmenbedingungen von FDM an HAW

Die Veröffentlichung der FAIR-Prinzipien (findable, accessible, interoperable, reusable, siehe Wilkinson u. a. 2016) markiert einen wichtigen Meilenstein auf dem Weg zu einer nachhaltigen und reproduzierbaren Wissenschaft. Für die Nachnutzung vorhandener Forschungsdaten und die Nachvollziehbarkeit der aus ihnen gewonnenen Erkenntnisse sind FAIRe Daten elementar und zum Inbegriff guter wissenschaftlicher Praxis geworden (Hochschulrektorenkonferenz 2015; Deutsche Forschungsgemeinschaft 2022).

Auch an Hochschulen angewandter Wissenschaft (HAW) bzw. Fachhochschulen hat der Umgang mit Forschungsdaten einen enormen Bedeutungszuwachs erfahren. Um die Spezifika im Forschungsdatenmanagement (FDM) an HAW nachvollziehen zu können, ist ein Blick auf die Rahmenbedingungen von Forschung an HAW unabdingbar. Anders als an Universitäten verstehen sich nicht alle Bereiche einer HAW als Teil einer forschungsaktiven Institution. Gleichwohl haben sich HAW in den vergangenen Jahren als gewichtige Akteure in der drittmittelfinanzierten Forschungslandschaft etabliert. Insbesondere durch ihre praxisorientierte und anwendungsnahe Ausrichtung leisten sie inzwischen einen relevanten Beitrag zur Innovations- und Forschungsleistung in Deutschland (Statistisches Bundesamt 2023; Deutsche Forschungsgemeinschaft 2023).

Die historische Entwicklung einer HAW spiegelt sich bis heute vielfach in ihrem Profil wider. Entstanden aus technischen Fachschulen mit starkem Ausbildungsauftrag, haben sich deren Studiengänge seit Mitte des 20. Jahrhunderts akademisiert und um sozial-, gesundheits- und wirtschaftswissenschaftliche Fächer erweitert. Der Fokus auf die anwendungsbezogene Forschung bleibt prägend: Viele Forschungsprojekte entstehen in enger Kooperation mit Partnern aus Wirtschaft und Gesellschaft, und greifen gesellschaftlich relevante, aktuelle Fragestellungen auf. Dieser Transfergedanke wird zunehmend auch von Fördermittelgebern und Wissenschaftspolitik eingefordert.

Gleichzeitig bleibt der strukturelle Rahmen für Forschung an HAW herausfordernd. Professorinnen und Professoren erbringen in vielen Bundesländern in der Regel ein doppelt so hohes Lehrdeputat wie ihre Kolleginnen und Kollegen an Universitäten. So auch in

Thüringen. Die Thüringer Lehrverpflichtungsverordnung (ThürLVVO)[1] sieht vor, dass die Lehrverpflichtung für HAW-Professuren 18 Lehrveranstaltungsstunden pro Semester umfassen. An Universitäten sind es stattdessen die Hälfte. Haushaltsfinanzierte Stellen für forschendes Personal jenseits der Professur sind selten. Wissenschaftliche Mitarbeitende sind meist entweder als Lehrkräfte für besondere Aufgaben mit einem Lehrdeputat, welches noch höher ist als das von professoralen Lehrenden (vgl. § 4 (2) ThürLVVO), oder projektgebunden angestellt. Damit ist die Forschung an HAW nahezu vollständig von Drittmitteln abhängig (Geyer, Berger und Dudenbostel 2016; Wirarda 2024).

Besonders Förderlinien und -programme der Bundesministerien spielen an den HAW eine zentrale Rolle, da sie anwendungsorientierte und interdisziplinäre Projekte fördern – also genau die Schwerpunkte, die an HAW im Fokus stehen. Förderprogramme der EU, wie etwa der Europäische Sozialfonds, sind ebenfalls relevant. Dagegen bleibt der Zugang zu DFG-Mitteln gering, da diese in erster Linie grundlagenorientierte Forschung fördert, deren Formate die anwendungsorientierte Forschung an HAW – i. d. R. ohne ein allgemeines Promotionsrecht – weniger anspricht (Himmelrath und Olbrich 2022; Statistisches Bundesamt 2023; Deutsche Forschungsgemeinschaft 2023).

Die Anforderungen an FDM nehmen hingegen stetig zu: Vorgaben von DFG und EU-Programmen (Deutsche Forschungsgemeinschaft 2021; OpenAIRE 2024) setzen Standards, an denen sich inzwischen auch andere, für HAW wichtigere Förderer, wie Bundesministerien oder Stiftungen, orientieren. Um im Wettbewerb um qualitativ hochwertige Forschungsergebnisse und Fördermittel bestehen zu können, müssen auch HAW zunehmend, ihre Forschenden mit Services und technischen Infrastrukturen beim Umgang mit Forschungsdaten unterstützen (Hochschulrektorenkonferenz 2015; Blümm u. a. 2023).

2 Situation von FDM an den Thüringer HAW

In Thüringen gibt es mit dem Thüringer Kompetenznetzwerk Forschungsdatenmanagement (TKFDM) bereits seit 2018 eine FDM-Landesinitiative, die eine überregionale Anerkennung in der deutschen Community erreicht hat. Angesiedelt mit je einer Kontaktperson an den vier Universitäten ist sie Anlaufstelle für alle Forschenden in Thüringen (Assmann u. a. 2024). Seit Ende 2022 wird das Team um zwei Data Stewards ergänzt, die von Thüringer Forschungsgruppen zur Verbesserung ihrer FDM-Praktiken kostenfrei angefordert werden können (Kirsch und Lindt 2024). Wissenschaftlerinnen und Wissenschaftler an HAW sollten von der Landesinitiative ebenso beraten werden wie jene der Universitäten, jedoch verzeichnete das TKFDM bis 2022 kaum Anfragen von Forschenden der Thüringer HAW. Diese Zielgruppe nahm auch kaum an den TKFDM Angeboten, wie den Coffee

1 https://landesrecht.thueringen.de/bsth/document/jlr-LVerpflVTH2005rahmen; *Besucht am 29. Mai 2025.*

Lectures, den Thüringer FDM Tagen oder dem FAIRest Dataset Award teil, und dies trotz steigender Notwendigkeit und Relevanz von FDM in der angewandten Forschung.

Bis Herbst 2022 standen für die HAW weder Zuwendungen des Landes zum Aufbau eines institutionell verankerten FDM bereit, noch hatten vor-Ort-Aktivitäten der Landesinitiative stattgefunden. Die Situation ändert sich, seit sich auch die vier Thüringer HAW im Rahmen des BMFTR-geförderten Projektes FDM-HAW Kompetenzcluster Jena-Erfurt-Nordhausen-Schmalkalden (FDM-HAWK)[2] erstmals umfassend mit dem Thema FDM beschäftigen. Im Vorhaben stehen das Abbilden von Ist-Stand und Bedarfen sowie das Aufzeigen und Etablieren von FDM-Strukturen an den HAW-Standorten im Mittelpunkt. Dabei ist das zentrale Anliegen, Wissenschaftlerinnen und Wissenschaftler zu Themen des FDM zu sensibilisieren, und dabei Kräfte und Ressourcen über die beteiligten HAW hinaus auch zukünftig in und mit der Landesinitiative zu bündeln.

Als Teil der IST-Stand- und Bedarfsanalyse zum FDM an den beteiligten Thüringer Hochschulen wurden im Rahmen des FDM-HAWK-Projekts seit Anfang 2023 exemplarisch Forschungsprojekte an jedem Standort begleitet. In diesem Artikel werden die Erkenntnisse aus zwei dieser begleiteten Projekte der Ernst-Abbe-Hochschule Jena (EAH Jena) und der Hochschule Nordhausen (HS Nordhausen) zusammengeführt. Beide Projekte haben unterschiedliche Voraussetzungen, wie die fachliche Ausrichtung (sozial- vs. ingenieurwissenschaftlich), die Laufzeit (drei vs. fünf Jahre), Vorkenntnissen der Projektbeteiligten sowie Bearbeitungsstand zu Beginn der FDM-Begleitung. Dennoch stehen sie auch exemplarisch für typische Förderer, die an den Hochschulen Drittmittel für die Forschung bereitstellen.

3 Online-Retrospektive zur Reflexion der FDM-Begleitung

Im Rahmen der FDM-Begleitung und in Vorbereitung auf diesen Artikel haben die Autorinnen und Autoren eine strukturierte Reflexion durchgeführt, um gemeinsame *Lessons Learned* zu formulieren. Ziel war es, die Erfahrungen und Perspektiven der Forschenden und des FDM-Personals zu erfassen und in einen Dialog zu überführen, der sowohl den Prozess der Projektbegleitung dokumentiert als auch Anregungen für die Ausgestaltung des institutionellen FDM an den Einrichtungen liefert.

2 https://forschungsdaten-thueringen.de/fdm-hawk-de.html; *Besucht am 29. Mai 2025.*

Tabelle 1: Fragenkomplexe und Reflexionsfragen der gemeinsamen Retrospektive.

Fragenkomplex	**Reflexionsfragen**
Vor der Begleitung	Wie haben Sie ihre Daten im Projekt organisiert? Welche Erwartungen und Hoffnungen hatten Sie an die FDM-Begleitung? Mit welchen Fragen bzw. Themen hatten Sie sich bereits auseinandergesetzt?
FDM-Begleitung	Welche Handlungsfelder wurden identifiziert? Wie wurde an die Herausforderungen herangegangen?
Veränderungen	Retrospektiv, was hat sich durch die FDM-Begleitung im Projekt bzw. bei Ihnen verändert bzw. verbessert?
Wünsche	Was wünschen Sie sich zukünftig an FDM-Unterstützung durch die Serviceeinrichtungen der Hochschulen? Z. B. Bibliothek, Datenschutz, Forschung und Transfer, Rechtsamt/Justiziariat, IT/Medienzentrum, Thüringer Kompetenznetzwerk FDM

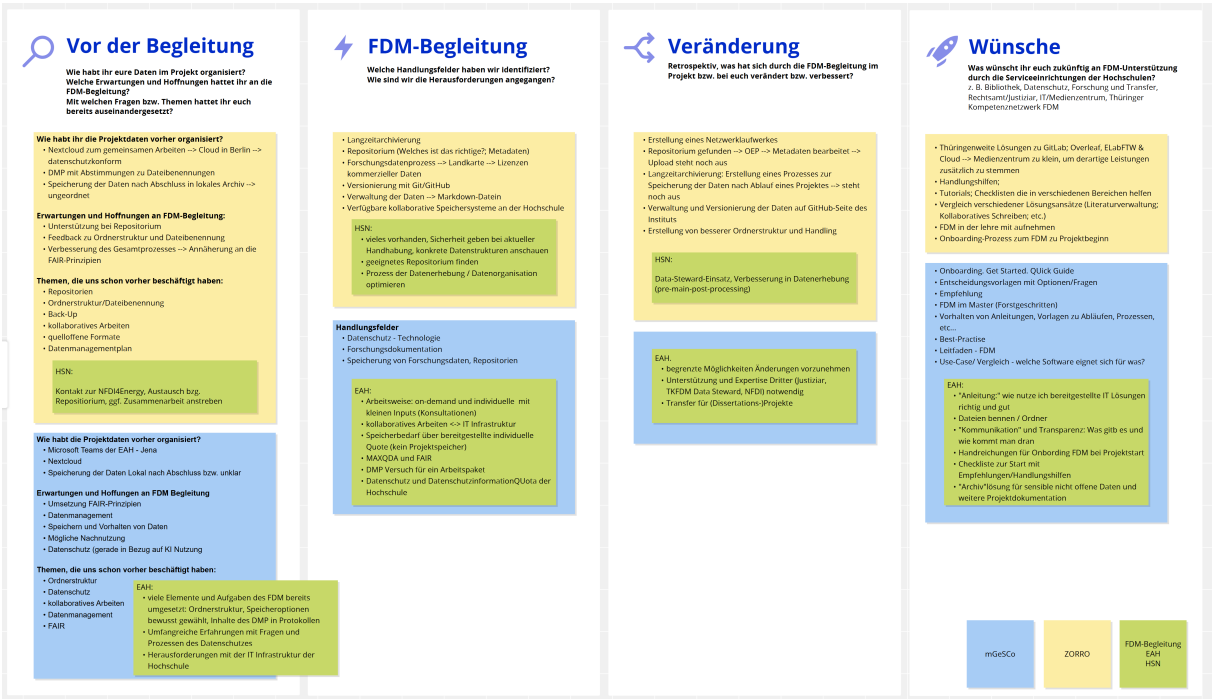


Abbildung 1: Ergebnisdarstellung der Online-Retrospektive.

Eine Online-Retrospektive[3], wie sie in der agilen Projektarbeit etabliert ist (Fuchs u. a. 2019), bot den methodischen Rahmen. Auf einem digitalen Whiteboard waren Reflexionsfragen zu den vier Komplexen *Vor der Begleitung*, *FDM-Begleitung*, *Veränderungen* und *Wünsche* (vgl. Tabelle 1) gestellt.

Die Fragenkomplexe wurden nacheinander besprochen. Dabei hatten die Teilnehmenden für jeden Komplex zunächst Gelegenheit ihre Gedanken auf virtuellen Kärtchen festzuhalten. Danach folgte eine Vorstellung der Notizen mit gemeinsamer Diskussion. Das so entstandene virtuelle Whiteboard (siehe Abbildung 1) bildet die Grundlage für die folgende Ergebnisdarstellung und wurde im Team der Autorinnen und Autoren abgestimmt.

4 Begleitete Projekte und Zusammenarbeit

Im folgenden Abschnitt werden der Ablauf und die Ergebnisse der Zusammenarbeit mit den zwei Forschungsprojekten vorgestellt. Die Auswahl der begleiteten Projekte erfolgte gezielt unter forschungsstarken Professorinnen und Professoren der Hochschule, die nicht nur eine hohe wissenschaftliche Expertise einbringen, sondern auch besonderes Engagement und die Bereitschaft, sich aktiv mit den Anforderungen und Potenzialen eines strukturierten Datenmanagements auseinanderzusetzen. Dies macht sie zu wichtigen Akteurinnen und Akteuren bei der Etablierung von FDM-Strukturen an den Hochschulen. Nachfolgend wird ein Überblick über die begleiteten Projekte sowie die dabei bearbeiteten Themen gegeben.

Multi-Generation Smart Community (mGeSCo) – Ernst-Abbe-Hochschule Jena

Das interdisziplinäre Forschungsprojekt mGeSCo[4] widmet sich den Herausforderungen der digitalen Transformation im Kontext eines smarten Quartiers in Jena Neulobeda. Im Zentrum stehen dabei die Bereiche Wohnen, Leben, Arbeiten und Pflegen – Themen, die im Rahmen der grundlegenden Sanierung und Modernisierung eines DDR-Plattenbaus neu gedacht werden. Das Projekt kooperiert mit vier regionalen Wirtschafts- und Sozialpartnern und wurde von der Carl-Zeiss-Stiftung von Juni 2021 bis Februar 2025 gefördert. Auffallend ist, dass seitens des Förderers seinerzeit noch keine spezifischen Anforderungen an das FDM oder an die FAIR-Kriterien gestellt wurden. Das Projektteam setzte sich aus zwei Professoren – aus den Fächern Informatik und Soziologie –, drei wissenschaftlichen Mitarbeitenden sowie einer studentischen Hilfskraft zusammen. Da es sich um eine reale Quartiersentwicklung handelte, musste die inhaltliche Ausrichtung fortlaufend an die Dynamiken im Reallabor und an die Entscheidungen der Praxispartner angepasst werden. Zusätzlich stellte die Fluktuation des wissenschaftlichen Personals das Projektteam wiederholt vor Herausforderungen.

3 https://agilescrumgroup.de/retrospektive-formen-mit-beispielen-und-ideen; *Besucht am 29. Mai 2025.*

4 https://www.eah-jena.de/mgesco; *Besucht am 29. Mai 2025.*

Die Begleitung des Projekts begann etwa zur Halbzeit des ursprünglich bewilligten Förderzeitraums. Zu diesem Zeitpunkt waren bereits wesentliche Entscheidungen zur Projektausrichtung getroffen worden, sodass die Beratung vor allem auf die Optimierung und Anpassung bestehender Prozesse ausgerichtet war. Der Beratungsprozess gliederte sich in mehrere Phasen, die jeweils unterschiedliche Schwerpunkte hatten (vgl. Tabelle 2). Im Dezember 2022 fand ein erstes Kennenlernen statt. Darauf aufbauend wurde im Januar 2023 ein FDM-Input in Form von zwei Workshops gegeben, um Grundlagen zu vermitteln und eine Bestandsaufnahme der Forschungsdaten (audiovisuelle Informationen, Surveydaten, Textdaten, Nutzungsdaten), methodischen Zugänge (Befragung, Interviews, Workshops/Beobachtung) und verwendeter Infrastrukturen (Speicheroptionen, Software, Lizenzen, kollaboratives Arbeiten und Wissensmanagement) des Projekts vorzunehmen. Bemerkenswert war zudem, dass dieses Projekt bereits umfangreiche Erfahrungen mit Themen des Datenschutzes gesammelt hatte: Für ein anstehendes Arbeitspaket wurde eine Datenschutzfolgeabschätzung durchgeführt und mit dem Landesdatenschutzbeauftragten abgestimmt. Auch eine dokumentierte Ordnerstruktur und Dateibenennung existierten bereits im Projekt, ohne dass diese unter dem Begriff Datenmanagementplan oder FDM gefasst wurden. Darüber hinaus wurden Themen für die weitere Begleitung identifiziert.

Die darauffolgende Phase der intensiven Begleitung erstreckte sich von Februar bis Juni 2023. In diesem Zeitraum fanden monatliche Konsultationen statt, die sich jeweils auf spezifische Themenschwerpunkte konzentrierten. Diese regelmäßigen Treffen ermöglichten es, den Fortschritt des Projekts kontinuierlich zu reflektieren und bei auftretenden Herausforderungen gezielt beratend einzugreifen. Dabei wurden u. a. die proprietäre Analysesoftware MAXQDA[5] unter den FAIR-Prinzipien (Wilkinson u. a. 2016) beleuchtet, die Möglichkeiten von Datenveröffentlichungen diskutiert, nach Alternativen für die kollaborative Zusammenarbeit mittels einer proprietären und zwischenzeitlich terminierten Landeslizenz gesucht, und das Fehlen eines dedizierten Hochschul-Dienstes für die Langzeitspeicherung thematisiert. Auch die Erarbeitung eines Datenmanagementplans für ein geplantes Arbeitspaket wurde in Angriff genommen. Ernüchternd für die Projektmitarbeitenden war jedoch die Erkenntnis, dass sich Best Practices des FDM zum fortgeschrittenen Projektstand kaum umsetzen ließen. Dies lag zum einen an den infrastrukturellen Gegebenheiten und fehlenden Alternativen an der Hochschule, zum anderen an früh getroffenen Entscheidungen im Projekt, deren nachträgliche Änderung mit erheblichen Mehraufwänden verbunden gewesen wäre und fachspezifischen Quasi-Standards widersprochen hätten. Das Arbeitspaket für welches ein Datenmanagementplan erarbeitet wurde, konnte aufgrund einer geänderten Ausrichtung in der Quartiersentwicklung nicht umgesetzt werden. Von April bis Juli 2024, wurde das Thema Datenschutz erneut aufgegriffen und ein weiteres Mal intensiv bearbeitet. Unter Einbeziehung des Justiziars der Hochschule wurde eine Datenschutzinformation, in der die Verarbeitung personenbezogener Daten auf die informierte Einwilligung gestützt wird, erarbeitet, um datenschutzrechtlichen Anforderungen einer Befragung der Bewohnenden des smarten Quartiers gerecht zu werden.

5 https://www.maxqda.com; *Besucht am 29. Mai 2025.*

Ergänzend zur intensiven Begleitung wurde der Austausch in unterschiedlichen Formaten, wie Kolloquien und anderen forschungsbezogenen Veranstaltungen des Projektes und der Hochschule fortgeführt. Dabei bestand der Mehrwert für das institutionelle FDM darin, dass die erlangten Erkenntnisse und identifizierten Herausforderungen in die Hochschule zurück gespiegelt werden konnten, um daraus zukünftig Handreichungen und Workshop-Themen abzuleiten.

Zero Carbon Cross Energy System II – Fallbeispiel Thüringen (ZO.RRO II KMU) – Hochschule Nordhausen

Das Energieforschungsprojekt ZO.RRO II[6] arbeitet nach einer ersten, abgeschlossenen Projektphase (2019 – 2022) weiterhin an der Dekarbonisierung der Energieversorgung im Freistaat Thüringen. Im Zentrum des Vorhabens steht die Berechnung von realistischen Zukunftsszenarien einer klimaneutralen Energieversorgung, vor allem mit Blick auf kleinere und mittlere Unternehmen. In der dabei entwickelten Energiesystemmodellierung entstehen vor allem Messdaten, Software und Modellcluster, aus denen Tools und Verfahren zur Optimierung einer klimaneutralen Energieversorgung entstehen sollen.

Tabelle 2: Zusammenarbeit mit den Projekten.

	mGeSCo (EAH Jena)	**ZO.RRO II (HS Nordhausen)**
Kontaktaufnahme/Zielsetzung	12/2022: Kennenlerntreffen 01/2023: FDM-Input und Workshop (2 Termine)	02/2023–05/2023: Kennenlerntreffen Austausch über Daten im Projekt sowie Kenntnisse/Erwartungen zu FDM (3 Termine)
Intensive Begleitung	02/2023–06/2023: Monatliche Konsultationen mit Themenschwerpunkten	05/2023–08/2023: Datenanalyse sowie Beantwortung verschiedener Fragen
Fortwährender Kontakt und Austausch	Austausch in anderen Formaten, u. a. Kolloquien, Veranstaltungen mit Forschungsbezug Rückspiegelung von Bedarfen und Erkenntnisse zur Erarbeitung von Handreichungen und Workshops	andauernder Kontakt mit TKFDM-Data-Stewards, Kontakt zur NFDI Austausch in erweiterten Formaten, v. a. FDM-Input in Dienstbesprechungen des Instituts
Retrospektive	02/2025 gemeinsame Retrospektive	

Das vom Thüringer Ministerium für Umwelt, Energie, Naturschutz und Forsten sowie vom Bundesministerium für Wirtschaft und Energie geförderte Projekt hat eine Laufzeit von

6 https://zorro-thueringen.de/projekte/zo-rro-ii-kmu; *Besucht am 29. Mai 2025.*

Anfang 2022 bis Juli 2026. Das Projektteam besteht aus einem Professor, drei bis vier wissenschaftlichen Mitarbeitenden und mehreren studentischen Hilfskräften.

Ein erstes Kennenlernen von Projektmitarbeitenden und FDM-Personal fand im Februar 2023 statt. Es folgten mehrere Termine bis zum Mai des gleichen Jahres, in denen ein Austausch über die im Projekt entstehenden Daten, die vorhandenen FDM-Kenntnisse sowie das Klären offener Fragen stattfand. Ein Datenmanagementplan sowie Konventionen zur Strukturierung von Daten lagen bereits vor. Fragen zum geeigneten Repositorium, zu Ontologien sowie zur Verwendung von Metadaten wurden aufgeworfen. Eine genauere Analyse von Datenerhebung, Abläufen in der Projektarbeit sowie Fragen zur späteren Datenveröffentlichung führten schließlich dazu, dass im Herbst 2023 ein Data-Steward-Einsatz bei der Thüringer Landesinitiative TKFDM beantragt wurde. Ab Januar 2024, noch vor Halbzeit des Projekts, startete die Unterstützung der Data-Stewards, was zum Teil vor Ort an der HS Nordhausen stattfand und vor allem die Erstellung einer Landkarte zur Datenerhebung, einen Git-Workshop zur Versionierung sowie die Entwicklung einer strukturierten Datendokumentation mit Hilfe von Markdown-Dateien beinhaltete. Weiterhin wurde ein projektbezogenes Netzlaufwerk eingerichtet, sowie die Struktur und Ordnung der gespeicherten Dateien verbessert. Gemeinsam mit Data-Stewards und FDM-Personal erfolgte eine intensive Beschäftigung mit dem ausgewählten Repositorium (Open-Energy-Platform[7]), wobei unter anderem Metadaten und Ontologien thematisiert wurden. Auch erfolgte die Kontaktaufnahme zum Fachkonsortium der Nationalen Forschungsdateninfrastruktur (NFDI) NFDI4Energy[8]. Hierbei wurden Perspektiven der Energieforschung sowie Möglichkeiten eines künftigen Austausches besprochen.

Der Kontakt zu den TKFDM Data-Stewards sowie zum Konsortium der NFDI dauert im März 2025 weiter an. Fragen zur Langzeitspeicherung, möglicherweise auch über die obligatorischen zehn Jahre hinaus, wurden ebenfalls angesprochen, sind aber bisher nicht abschließend geklärt. Darüber hinaus hat sich aus der FDM-Begleitung der Energiesystemmodellierung ein regelmäßiger, erweiterter Austausch zum Thema FDM über das Projektteam hinaus entwickelt. Monatlich findet ein FDM-Impuls im Institut für Regenerative Energietechnik statt, was die Sensibilisierung zum Thema über das Projekt hinaus wesentlich steigert.

Durch die enge Begleitung der beiden Projekte konnten zentrale Themen, Herausforderungen und Veränderungen im Umgang mit Forschungsdaten identifiziert werden. Die Zusammenarbeit brachte einen doppelten Gewinn: Die Projekte profitierten von der strukturierten Unterstützung, während die Hochschulen wertvolle Einblicke in die Bedarfe und Potenziale eines nachhaltigen FDM gewannen.

7 https://openenergyplatform.org; *Besucht am 31. März 2025.*
8 https://nfdi4energy.uol.de; *Besucht am 31. März 2025.*

5 Lessons Learned und zukünftige Herausforderungen

Als abschließende Frage der gemeinsamen Retrospektive wurde ein Austausch über Wünsche zur FDM-Unterstützung durch die Hochschule geführt – insbesondere zu Aspekten des institutionellen FDM und den Serviceangeboten für Forschende. Diese Diskussion bildete zugleich die Überleitung zu den zentralen Fragestellungen der Bedarfsermittlung und den abzuleitenden *Lessons Learned*.

Trotz unterschiedlicher fachlicher Ausrichtungen, abweichendem Beginn der Projektbegleitung und ungleicher Vorkenntnisse äußerten die Projektmitarbeitenden ähnliche Bedarfe bezüglich des institutionellen FDM. Sie wünschen sich praxisnahe, einfache Handreichungen und Leitfäden, die den Einstieg ins FDM erleichtern. Entscheidungshilfen zu Tools und Anwendungen sind ebenfalls gefordert, aber auch transparente Prozesse für die Möglichkeiten der Nutzung von (bereits vorhandenen) Hochschul-Services und technischen Infrastrukturen stehen im Fokus. Ein gemeinsamer Wunsch ist außerdem, FDM in der Lehre zu verankern – im Rahmen des wissenschaftlichen Arbeitens (Propädeutik) oder auch in den Curricula von Masterstudiengängen. Die Forschenden des sozialwissenschaftlich ausgerichteten Projekts an der EAH Jena wünschen sich insbesondere klare Regelungen und Handlungsleitfäden beim Umgang mit sensiblen, personenbezogenen Daten. Bei den Mitarbeitenden des ingenieurwissenschaftlichen Projektes an der HS Nordhausen steht der Bedarf an Schulungen zu konkreten Softwareangeboten sowie zur Datenanalyse stärker im Fokus. Auch die fehlende institutionelle Möglichkeit zur Langzeitspeicherung von Projektdaten wurden von beiden Projekten angesprochen.

Aus der Retrospektive haben die Autorinnen und Autoren fünf zentrale *Lessons Learned* abgeleitet. Diese werden in den folgenden Abschnitten unter den Stichworten *Versteckte Talente*, *Von Anfang an*, *Kreativität trifft Netzwerk*, *Kultur- und Strukturwandel* sowie *FDM-Personal vor Ort* vorgestellt.

Versteckte Talente: Auch, wenn nur in einem der beiden Beispielprojekte zu Beginn der Begleitung ein Datenmanagementplan vorlag, so wurden Entscheidungen und Konventionen zu Ordnerstrukturen sowie zur Anwendung von Tools bereits in beiden Projekten bewusst getroffen und dokumentiert. Vieles, was für das Datenmanagement wichtig und nützlich ist, wird bereits praktiziert, ohne die Bezeichnung FDM zu tragen. Gerade hier kommt es oft auf ganz konkrete Tipps an, die bei offenen Fragen helfen und den strukturierten Umgang mit Daten verbessern oder verfeinern. Eine Sensibilisierung, was FDM bedeutet, hilft dabei, tiefer in das Themenfeld einzusteigen.

Von Anfang an FDM mit Ressourcen planen: FDM ist keine punktuelle Aufgabe, sondern begleitet das gesamte Forschungsprojekt von Beginn bis zum Abschluss.

Beispielsweise zeigt sich am mGeSCO-Projekt der EAH Jena, dass die Entscheidung für die Nutzung einer Kollaborationsplattform mit integrierter Speicherlösung in den frühen Projektphasen getroffen wurde, später kaum noch geändert werden konnte. Spätere Änderungen kosten zusätzlichen Aufwand und überproportional viel Zeit und Produktivität. Für einen effizienten Umgang mit Daten unterstützt FDM, das von Forschenden durchgeführt wird, die Projektplanung und -durchführung. Dies sollte durch den FDM-Support der Serviceeinrichtungen der Hochschule gestärkt werden.

Kreativität trifft Netzwerk: Trotz der oft geringen Ressourcen gerade kleinerer HAW kann FDM gelingen, wenn Lösungen nachgenutzt oder Kräfte gebündelt werden können. Nicht alles muss an der eigenen Hochschule vorhanden sein oder neu entwickelt werden. Man kann voneinander profitieren. Dies zeigt die Zusammenarbeit mit TKFDM Data Steward, der NFDI4Energy aber auch die Konsultation hochschulinterner juristische Expertise. Nicht immer geht es um die perfekte Lösung, sondern darum, bedarfsorientiert das Praktikable einzusetzen. Beispielsweise wurde MAXQDA als fachspezifische proprietäre Analysesoftware genutzt, wohingegen die Sicherung der Daten nach Projektende in einem offenen Format erfolgte. Herausfordernd ist vielmehr, die Vielfalt und Unübersichtlichkeit der FDM-Angebote zu kanalisieren, um Forschenden aktuelle und passende Informationen bereitzustellen.

Kultur- und Strukturwandel: HAW neu denken für Forschung und FDM! Das bedeutet, dass FDM auch an HAW immer wichtiger wird. Doch die Rahmenbedingungen sind andere als bei Einrichtungen mit langjähriger Forschungstradition. Forschungsprojekte sind stärker anwendungsorientiert, haben tendenziell kürzere Laufzeiten und weniger Freiräume als grundlagenorientierte Forschungsvorhaben. Zudem ist die Serviceinfrastruktur an HAW (noch) nicht auf Forschung ausgerichtet. Erst im letzten Jahrzehnt haben sich HAW von den einstigen praxisnahen Ingenieursschulen und Ausbildungsstätten zu Forschungseinrichtungen entwickelt. Zusätzliche Bedarfe der Forschenden stehen häufig im unmittelbaren Ressourcenkonflikt mit Notwendigkeiten in der Lehre. Das spüren Forschende vor Ort, da Sie bei zentralen Entscheidungen, wie beispielsweise zum IT-Service Angebot nicht immer mitgedacht werden. Hier ist künftig ein Umdenken gefordert.

FDM-Personal vor Ort: Eine Ansprechperson macht den Unterschied! Forschungsdatenmanagement ist ein dynamisches Themenfeld – etwa hinsichtlich der Anforderungen von Forschungsförderern und Fachgesellschaften, oder der Verfügbarkeit daten- und fachspezifischer IT-Ressourcen. Es ist wichtig, stets den Überblick über aktuelle Möglichkeiten und Angebote zu behalten und dabei Kenntnisse zu den örtlichen Rahmenbedingungen zu berücksichtigen. Forschende an HAW haben oft zu wenig Zeit, um sich intensiv mit diesen Themen auseinanderzusetzen. Daher ist es besonders hilfreich, eine kompetente Ansprechperson vor Ort zu haben, die bei konkreten Fragestellungen unterstützt. So können auch wiederkehrende Bedarfe frühzeitig identifiziert und Prozesse gemeinsam mit

Kooperationspartnern gestaltet werden. FDM ist eine Querschnittsaufgabe, die sowohl Serviceeinrichtungen vor Ort als auch die (über-)regionale FDM-Community betrifft.

Die in diesem Artikel vorgestellten *Lessons Learned* lassen sich in vielen Punkten auf andere Projekte und HAW übertragen, da vergleichbare Herausforderungen bestehen und sich ähnliche Lösungsansätze ergeben. Bereits in der Retrospektive wurde durch den Austausch zweier unabhängiger FDM-Begleitungen ein Dialog geschaffen, der Lösungsstrategien sichtbar macht und die Einordnung in einen übergeordneten Kontext ermöglicht.

6 Schlussfolgerungen für den Aufbau von FDM an HAW

Die *Lessons Learned* aus den begleiteten Projekten sind nur ein Aspekt einer mehrstufigen Ist-Stand- und Bedarfsanalyse an den Hochschulen und im FDM-HAWK Projekt. Sie spiegeln vor allem die Perspektive der Forschenden wider. Im abschließenden Abschnitt soll der Blick auf institutionelle und strategische Überlegungen zum Aufbau von FDM an HAW erweitert werden.

HAW für FDM fit zu machen, bedeutet, FDM institutionell zu denken – über ein Forschungsprojekt hinaus. Gerade Mitarbeitende in der Verwaltung und in den Serviceeinrichtungen sind eine oft unterschätzte, aber zentrale Zielgruppe. Sie spielen eine Schlüsselrolle, wenn es darum geht, FDM strukturell zu verankern, hochschulspezifische (IT-)Ressourcen bereitzustellen und die gesamte Hochschule – von der Lehre bis zur Forschung – zu unterstützen. HAW haben hier die Chance trotz oder wegen ihres späteren Starts, aus den Erfahrungen anderer Akteure zu lernen und FDM von Anfang an ganzheitlich und strategisch anzugehen. Ein vielversprechender Ansatz hierfür ist die RISE-DE Selbstevaluation (Hartmann, Jacob und Weiß 2019), die an beiden Hochschulen in einem FDM-Gremium durchgeführt wurden. Dabei wurden auch die Erkenntnisse aus der FDM-Begleitung der Forschungsprojekte berücksichtigt.

Da an HAW das Thema FDM oft nur mit einem geringem Stellenumfang – im Thüringer Beispiel mit jeweils 0,5 Vollzeitäquivalenten an den Hochschulstandorten – bearbeitet werden kann, ist es sehr hilfreich, Aufgaben mit Synergiepotential zum Thema FDM mitzudenken und zu bündeln (IT-Sicherheit, Datenkompetenz in Lehre und Forschung, Open Science in allen Facetten, Forschungsförderung, Qualitätssicherung von Promotionen). Neben der Sensibilisierung über diese Partner- und Querschnittsaufgaben bei den beteiligten Serviceeinrichtungen, ist das Abbilden von Prozessen sinnvoll, um Zuständigkeiten und ein koordiniertes Handeln zu ermöglichen. Praxisnahe Handreichungen ergänzen das Serviceangebot zweckmäßig, da sie Forschenden Hilfestellungen anbieten und den Beratungsbedarf für wiederkehrende Fragen verringern können. All diese Strukturen mit aufzubauen sind zusätzlichen Aufgaben, die natürlich initial Aufwände kosten aber län-

gerfristig zielführend für ein ressourcenoptimiertes, institutionelles FDM an HAW sein können.

FDM gelingt insbesondere durch Zusammenarbeit über Einrichtungsgrenzen hinweg: HAW können durch Partnerschaften mit anderen Hochschulen, lokalen Forschungseinrichtungen oder einer FDM-Landesinitiative von bestehenden Schulungen, Services und Netzwerken profitieren. In Thüringen zeigt sich auf Grundlage interner Auswertungen[9], dass einige HAW inzwischen in ähnlichem Umfang Beratungsanfragen und TKFDM Data Steward Einsätze verzeichnen wie Universitäten von vergleichbarer Größe. Der Schlüssel liegt in der gemeinsamen Nutzung vorhandener Ressourcen, und einer starken HAW-Stimme in der Landesinitiative und in der Community. So kann FDM auch mit begrenzten Mitteln an HAW sichtbar werden und nachhaltig gelingen.

Danksagungen

Die Autorinnen und Autoren danken den Förderern und den Hochschulen der beteiligten Projekte für die finanzielle und ideelle Unterstützung. Das FDM-HAW Kompetenzcluster Jena-Erfurt-Nordhausen-Schmalkalden (FDM-HAWK, FKZ 16FDFH107A-D) wird gefördert durch das Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) und finanziert über NextGenerationEU. Multi-Generation Smart Community (mGeSCO) wird gefördert durch die Carl-Zeiss-Stiftung. Zero Carbon Energy System II – Fallbeispiel Thüringen (ZO.RRO II, FKZ O3EIO73A-F) wird gefördert durch das Thüringer Ministerium für Umwelt, Energie, Naturschutz und Forsten, sowie durch das Bundesministerium für Wirtschaft und Energie (BMWE).

Beiträge

- Sarah Boelter: Konzeption, Verschriftlichung – Erstellung des Originalentwurfs, Verschriftlichung – Überprüfung und Bearbeitung
- Romy Meyer: Konzeption, Verschriftlichung – Erstellung des Originalentwurfs, Verschriftlichung – Überprüfung und Bearbeitung
- Wesley Preßler: Verschriftlichung – Überprüfung und Bearbeitung
- Lucie Schmidt: Verschriftlichung – Überprüfung und Bearbeitung
- Amélie Oberdorfer: Verschriftlichung – Überprüfung und Bearbeitung
- Christoph Schmidt: Verschriftlichung – Überprüfung und Bearbeitung

9 Die zugrunde liegenden Daten basieren auf internen Erhebungen und Einsatzdokumentationen der beteiligten Hochschulen. Sie werden regelmäßig für Zwecke der Evaluation und Weiterentwicklung genutzt, sind jedoch nicht öffentlich zugänglich.

Teile dieses Textes wurden mit Unterstützung von ChatGPT-40 sprachlich überarbeitet. Die generierten Inhalte wurden kritisch von den Autorinnen und Autoren überprüft und angepasst. Die endgültige Fassung wird von den Autorinnen und Autoren verantwortet.

Interessenkonflikte

Die Autorinnen und Autoren erklären, dass keine Interessenkonflikte bestehen.

Literaturverzeichnis

Assmann, Cora, Sarah Boelter, Roman Gerlach, Kevin Lang, Kevin Lindt, Stefan Kirsch, Nadine Neute und Jessica Rex. 2024. *Poster: Thüringer Aktivitäten zum Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/ZENODO.13734215.

Blümm, Mirjam, Christine Burkart, Maria Chlastak, Katharina Fritsch, Heike Neuroth, Andreas Schieberle, Stefan Schmunk und Stephanie Werner. 2023. „Datenmanagementpläne an Fachhochschulen /Hochschulen für Angewandte Wissenschaften Eine Bestandsaufnahme und Empfehlungen". In *Schriften zur Informationswissenschaft,* Bd. 77. Hochschulverband Informationswissenschaft (HI) e.V. https://doi.org/10.5281/ZENODO.8275739.

Deutsche Forschungsgemeinschaft. 2021. *Umgang mit Forschungsdaten. Checkliste für Antragstellende zur Planung und zur Beschreibung des Umgangs mit Forschungsdaten in Forschungsvorhaben.* Besucht am 31. März 2025. https://www.dfg.de/resource/blob/174732/3c6343eed2054edc0d184edff9786044/forschungsdaten-checkliste-de-data.pdf.

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

———. 2023. *DFG Jahresbericht 2023.* Besucht am 29. Mai 2025. https://www.dfg.de/resource/blob/336226/3abb5950d9eee7540d13396e81b91b48/dfg-jb2023-data.pdf.

Fuchs, Christoph, Philipp Barthel, Katharina Winter und Thomas Hess. 2019. „Agile Methoden in der digitalen Transformation – mehr als ein Konzept für die Softwareentwicklung". *Wirtschaftsinformatik und Management* 11 (4): 196–207. https://doi.org/10.1365/s35764-019-00192-8.

Geyer, Anton, Florian Berger und Tobias Dudenbostel. 2016. *Empirische Begleitung des Programms „Forschung an Fachhochschulen".* Abschlussbericht. Technopolis Group. Besucht am 29. Mai 2025. https://www.forschung-haw.de/fachhochschulen/shareddocs/downloads/files/abschlussbericht_begleitmassnahme_fafh.pdf.

Hartmann, Niklas K., Boris Jacob und Nadin Weiß. 2019. *RISE-DE – Referenzmodell für Strategieprozesse im institutionellen Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.3585556.

Himmelrath, Armin und Max Olbrich. 2022. „Trotz gesetzlicher Vorgaben: Fachhochschulen erhalten zu wenig Forschungsgelder". *Spiegel,* besucht am 28. April 2024. https://www.spiegel.de/panorama/bildung/fachhochschulen-erhalten-zu-wenig-forschungsgelder-trotz-gesetzlicher-vorgabe-a-bc14bc91-9ad3-42e6-858b-858d23cceed0.

Hochschulrektorenkonferenz. 2015. *Wie Hochschulleitungen die Entwicklung des FDMs steuern können: Orientierungspfade, Handlungsoptionen, Szenarien.* Besucht am 31. März 2025. https://www.hrk.de/positionen/beschluss/detail/wie-hochschulleitungen-die-entwicklung-des-forschungsdatenmanagements-steuern-koennen-orientierungsp/.

Kirsch, Stefan und Kevin Lindt. 2024. *Rent a Data Steward – Hands-on support for Thuringian research groups: concept meeting first insights.* Zenodo. https://doi.org/10.5281/ZENODO.13908567.

OpenAIRE. 2024. *Guides for Researchers: RDM in Horizon Europe Proposals.* Besucht am 29. Mai 2025. https://www.openaire.eu/rdm-in-horizon-europe-proposals.

Statistisches Bundesamt. 2023. *Statistischer Bericht: Finanzen der Hochschulen 2021: Fachserie 11 Reihe 4.5.* Besucht am 31. März 2025. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Bildungsfinanzen-Ausbildungsfoerderung/Publikationen/Downloads-Bildungsfinanzen/statistischer-bericht-finanzen-hochschulen-2110450217005.xlsx.

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. https://doi.org/10.1038/sdata.2016.18.

Wirarda, Jan-Martin. 2024. „Ein Dauerstellen-Programm für die HAWs: Interview mit Jörg Bagdahn, Sprecher der Hochschulen für angewandte Wissenschaften in der Hochschulrektorenkonferenz". *Wiarda Blog,* besucht am 19. Mai 2025. https://www.jmwiarda.de/2024/07/01/ein-dauerstellen-programm-f%C3%BCr-die-haws/.

Secure Infrastructures and Protection of Personal Data: Approaches for the Development of the NFDIxCS Research Data Management Container

Patrick Brunner [1,*], Stefan Bavendiek [2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Karlsruhe, Germany
[2] University of Hamburg, Hamburg, Germany
[*] Corresponding author: patrick.brunner@fiz-karlsruhe.de

This paper proposes a concept for the sharing of personal data for scientific research purposes and a potential role of the NFDIxCS infrastructure in that process. The concept takes account of the so-called Research Data Management Container (RDMC) that is developed by the NFDIxCS consortium. The RDMC should incorporate research data together with the corresponding metadata, the research software and the required execution environment. With regard to the proposed concept, the paper discusses some legal and technical issues concerning the transmission of personal data for scientific research purposes. The legal part of the paper addresses the purpose limitation principle that is stipulated, inter alia, by the GDPR. It analyses two distinct legal interpretations for the notion of data collection and further processing under the GDPR: a role-based approach and a data-based approach. The role-based approach ties these legal terms to the controller of the personal data, the data-based approach to the data itself. As explained in the paper, both approaches may be based on the wording of the GDPR provisions and recitals. The approaches lead to different legal conditions for the transmission of personal data. Our decision in favour of the data-based approach rests on a teleological interpretation of the purpose limitation principle. This is followed by a technical perspective of the requirements and existing solutions to address some of the discussed issues and legal requirements, with a focus on secure data sharing in the context of research containers. After an overview of the challenges has been provided, several potential solutions based on existing technology are presented.

Keywords: data sharing, personal data, purpose limitation principle, data security

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23932 (CC BY-SA 4.0).

1 Introduction

The NFDIxCS consortium has the goal of being a central infrastructure for and together with the computer sciences in Germany to promote the (inter-)national availability and reusability of research data from this field. For this objective, the NFDIxCS consortium is developing a so-called Research Data Management Container (RDMC) to enable researchers in the computer sciences to bundle their research data together with the corresponding research software and the required execution environment in a "time capsule". The RDMC will enable a sustainable, subsequent use of the research data that it holds. The long-term availability of research data is ensured in compliance with the FAIR principles.

The research data in the field of computer science is diverse. For example, it includes data that concerns the interactions between human beings and machines (Saporito et al. 2024). This kind of research data – if it relates to identifiable human beings – is personal data under data protection law.[1] Furthermore, some legal scholars have argued, that computer code in certain instances is personal data as it may relate to natural persons (Purtova and Leenes 2023).

A research project does not end with the acquisition of research findings; the communication of research results and research data is an essential part of scientific research. In general, scientific research builds upon the knowledge, produced by previous research findings. In that regard, it is necessary to document research results and keep the underlying research data available; on the one hand to ensure the verifiability of the research results and on the other hand to promote future scientific research. In contrast, the availability of personaldata must be limited to what is necessary in order to protect data subjects. Furthermore, data subjects must be provided with information and the means to exert their rights in relation to the processing of their personal data. Under EU Data protection law, research data that is personal data needs to be anonymized as soon as the research purposes allow it (Art. 89 para. 1 General Data Protection Regulation, GDPR). Long-term availability, transfer and subsequent use of personal data without a data protection concept is not compatible with EU data protection laws. The following concept is a proposition for the NFDIxCS infrastructure to facilitate access to and transfer of personal data for research purposes within the computer sciences.

1 Keeping in mind that the highest competent authority for the interpretation of EU law, the Court of Justice for the European Union (CJEU), follows a broad understanding for the notion of personal data, cf. CJEU, judgment of 07.03.2024, OC, C-479/22, ECLI:EU:C:2024:215, para. 44 et seq.

2 Scope and limitation of this paper

This paper proposes a data transfer concept for the NFDIxCS infrastructure, with the aim of facilitating the lawful sharing and access to personal data in research (Section 3). There are also other viable concepts for the sharing of personal data for scientific research purposes.

The process explanation should give the reader an overview of our proposed concept that is based on a NFDIxCS platform and a NFDIxCS repository that allows for the findability of research data from the computer sciences. With a view to the proposed concept, we then address some principles for the processing of personal data under the GDPR (Art. 5 GDPR) that need specific attention. From a legal perspective (Section 4), we will discuss the conceptual transfer of personal data in light of the purpose limitation principle (Art. 5 (1)(b) GDPR). From a technical perspective (Section 5), we present different solutions for ensuring the confidentiality and integrity of personal data (Art. 5 (1)(f) GDPR) during transmission.

3 Process explanation for the proposed data transfer concept

At the first level, the RDMC provider (the data holder) provides metadata for their research data that is personal data (in the following also referred to as research data) to the NFDIxCS platform via the 'NFDIxCS Platform for Metadata' (NPM) (A, see Figure 1). At the same time, the NPM suggests encryption methods or specific tools for encrypting the research data. The research data remains with the RDMC provider (e.g. on the servers of the research institute) and is not uploaded to the NPM. Next, the metadata is retrieved from the NPM, stored on the NFDIxCS platform (B), and sent to the NFDIxCS repository, where it becomes publicly viewable (C).

At the second level, the RDMC user (a user registered with the NFDIxCS platform) searches the NFDIxCS repository for relevant research datasets and finds relevant metadata (1). After verifying their identity via the 'Authentication Platform' (AP), the RDMC-User can send a standardised access request to the RDMC-Provider (2). The RDMC-Provider then checks the access request to ensure it is legally admissible, before deciding whether to send the encrypted research data and the encryption key to the RDMC-User via a secure third-party service (3).

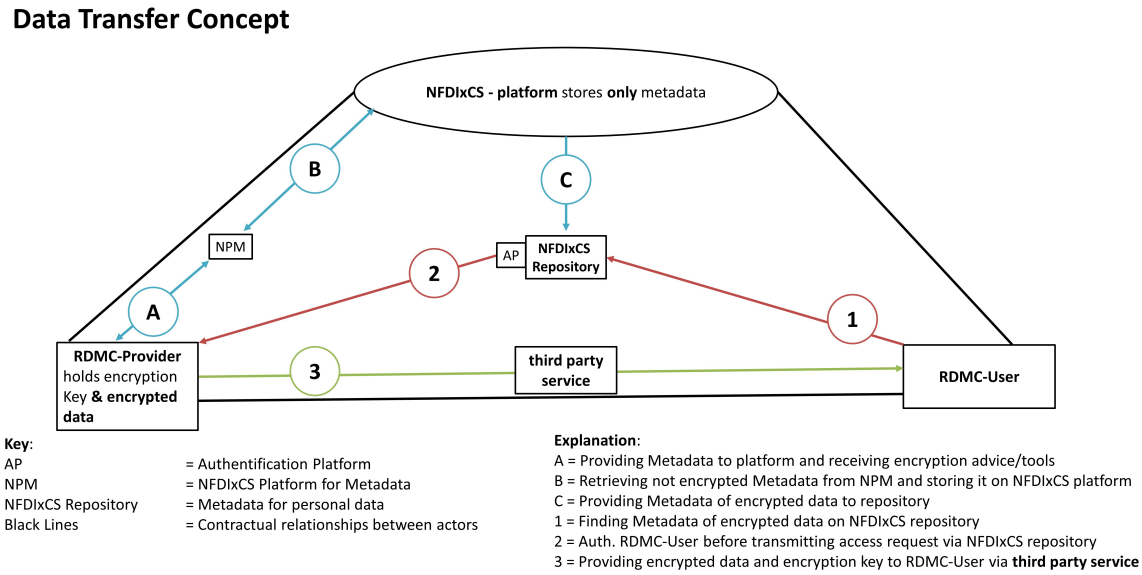


Figure 1: Graphic of the Data Transfer Concept.

4 Legal issues pertaining to the concept

This part of the paper addresses the principle of purpose limitation (Art. 5(1)(b) GDPR), which is a key legal issue in the conceptual transfer of personal data.

In the proposed concept, if the RDMC-Provider wishes to transmit personal data to an RDMC-User, the question arises as to whether this is permitted under data protection law. The following discussion is based on the premise that the RDMC-Provider initially collects personal data from data subjects for a specific research project, and later transmits it to an RDMC-User for another research project. This allows the research data to be reused in a scientific context.

4.1 The purpose limitation principle with regard to scientific research

Personal data is any information that relates to an identified or identifiably person, the data subject (Art. 4(1) GDPR). i.e., the usage, of personal data in the EU is regulated by multiple data protection laws, most importantly the General Data Protection Regulation (GDPR).

The subsequent use of personal data is particularly important under data protection law. The GDPR's purpose limitation principle stipulates that personal data shall be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes" (Art. 5(1)(b) GDPR). Data collection is an active activity by which the collecting entity obtains personal data for the first time (Roßnagel 2025a, DS-GVO Art. 4 Nr. 2, para. 15).

The purpose of the data collection establishes the basis for assessing the admissibility of data processing under data protection law and the associated obligations in a particular processing situation. Thus, the purpose specification serves, among other things, to assess the lawfulness of the processing within the meaning of Art. 6(1) GDPR.[2] Furthermore, the purpose specification creates transparency regarding the reason for the processing. It prevents data collection without cause.

The processing of personal data for scientific research purposes is broadly interpreted under the GDPR. In a non-exhaustive manner, the legislator explains that scientific research purposes should include "technological development and demonstration, fundamental research, applied research and privately funded research" (Recital 159). However, processing personal data for research purposes usually requires specification of a particular research project (European Data Protection Board 2018, para. 10).

4.2 Further processing personal data for secondary research purposes

Any processing activity after the initial data collection constitutes a further processing of personal data under the GDPR. If personal data is collected for one or more purpose(s) (hereinafter also referred to as: primary purposes) and is later processed for a different purpose, the latter constitutes a secondary use of the personal data. Under data protection law, this is referred to as a further processing of personal data for purposes, that were not specified at the time of the original data collection (hereinafter also referred to as: further processing for secondary purposes).

With regard to the purpose limitation principle, further processing of personal data for secondary purposes is permissible under the same conditions as the original processing of that data, if the data subjects to whom the data refer have consented to the processing of their personal data for the secondary purposes (Art. 6(4) GDPR). The technical implementation of dynamic consent models facilitates the ongoing inclusion of data subjects in processing activities concerning their personal data (Lay et al. 2024). These models can support the permissibility of further processing personal data for secondary purposes by enabling the data subjects to give and change their consent for a particular processing pur-

2 Cf. CJEU, 'SS' SIA, C-175/20, ECLI:EU:C:2022:124, para. 66.

pose via a dedicated platform. However, there is a risk that data subjects, after initially providing their data, will not respond to consent requests for secondary purposes.

Further processing of personal data for secondary purposes is also permissible if the processing is based on a legal basis in Union or Member State law that constitutes a necessary and proportionate measure in a democratic society to protect the objectives set out in Article 23(1) GDPR (Art. 6(4) GDPR). These objectives are primarily those of general public interest, such as public security and public health.

Finally, further processing is permitted if the purposes of the original data collection are not incompatible with the secondary purposes. This generally requires a compatibility test, taking into account, inter alia, any links between the primary and secondary purposes, the context of the initial data collection, the sensitivity of the personal data concerned, the potential consequences of further processing for the data subjects, and the existence of appropriate safeguards (Article 6 (4) GDPR).

With regard to scientific research, the further processing of personal data for secondary purposes is privileged. Further processing of personal data for secondary scientific or historical research purposes is not considered incompatible with the purposes of the initial data collection (Art. 5(1)(b) GDPR), if appropriate safeguards are in place to protect the data subjects (Art. 89(1) GDPR). If this is the case, the compatibility between the primary processing purpose and the secondary scientific or historical research purposes is assumed without the need for a compatibility test.[3]

4.3 Safeguards for the processing of personal data for scientific or historical research purposes

Safeguards mentioned by Art. 89(1) GDPR to protect the data subjects are primarily the anonymisation and the pseudonymisation of personal data. Personal data must always be kept in a pseudonymised form during the research process, i.e., "in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person" (Art. 4(5) GDPR). Anonymisation is the process of changing personal data so that the data subjects can no longer be identified with "means reasonably likely to be used, (...) either by the controller or by another person" (Recital 26). Personal data must be anonymised as soon as the research purposes, for which the data is being processed, allow it (Art. 89(1) GDPR).

Additional safeguarding measures (cf. §§ 27(1), 22(2) of the German Bundesdatenschutzgesetz) include but are not limited to:

3 Critical on the legal implications of Art. 5(1)(b) GDPR: Roßnagel 2025b, DS-GVO Art. 5, para. 109.

- encrypting personal data;
- controlling access to personal data (need-to-know principle);
- logging of access and processing activities;
- training for people involved in processing personal data;
- involving a data protection officer;
- securing the capabilities and the availability, integrity, confidentiality and resilience of systems used for data processing;
- establishing procedures to ensure compliance with data protection laws when personal data is further processed for other purposes (including by transmission of personal data);
- establishing a procedure for regular audits of the aforementioned measures.

4.4 A question of interpretation: The difference between data collection and further processing

Different interpretational approaches exist regarding the understanding of the interdependent concepts of data collection and further processing in data protection law. A distinction is made between a role-based and a data-based interpretation of the purpose limitation principle.

4.5 Role-based approach

The role-based approach of the purpose limitation principle (Becker et al. 2022), links the distinction between data collection and further processing to the controller. This approach is founded on the understanding that the GDPR's rules always relate to the processing of personal data and the purposes for which the data is being processed (Becker et al. 2022, p. 138). The controller decides the purposes and means by which personal data is being processed in a particular situation (Art. 4(7) GDPR). It is the controller's responsibility to comply with the GDPR's rules when processing personal data (Art. 5(2) and Art. 24(1) GDPR). Therefore, the notion of further processing personal data should be tied to the respective controller, who has previously collected the data for primary purposes (Becker et al. 2022, p. 138). Accordingly, if a controller obtains personal data for the first time, this is considered a data collection and not a further processing of personal data, regardless of whether the data has previously been processed by another controller.

For the proposed concept, this means that the transmission of personal data by the RDMC-Provider constitutes further processing of that data. This is because the reason for the transmission is to enable the subsequent research by the RDMC-User, which is a distinct purpose from the research purposes for which the data was originally collected by the RDMC-Provider. However, the data procurement by the RDMC-User represents

a new initial data collection for the purposes specified in the data access request to the RDMC-Provider.

Different provisions and a recital of the GDPR support this approach:

Art. 5(1)(b) GDPR: Becker et al. base their approach on the wording of Art. 5(1)(b) GDPR, which sets out the purpose limitation principle by requiring personal data to be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes", but does not specify from which source personal data shall be collected. This allows for the interpretation that a "primary" collection of personal data may also occur, where data has previously been processed by another controller (Becker et al. 2022, p. 38).[4]

Art. 6(4)(b) GDPR: If personal data is to be further processed for secondary purposes, a compatibility test is usually required to ensure that the collection purposes and the secondary purposes are not incompatible, in accordance with the purpose limitation principle. As previously mentioned, this compatibility test must take into account various aspects of the original data collection and the further processing, including „the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller" (Art. 6(4)(b) GDPR). This indicates a connection between the original collection of personal data and the controller who wishes to further process personal data for secondary purposes. This suggests that further processing of personal data for secondary purposes is preceded by the collection of the data by the same controller.

Recital 50 GDPR: The legislator, having regard to the compatibility test, explains that in "order to ascertain whether a purpose of further processing is compatible with the purpose for which the personal data are initially collected, the controller, *after having met all the requirements for the lawfulness of the original processing*, should take into account, inter alia (. . .)" (emphasis by the author) the various aspects mentioned by Art. 6 (4) GDPR. Consequently, the legislator holds that a controller who wishes to further process personal data has previously processed the data for primary purposes, and that the lawfulness of the original data processing is a prerequisite for the permissibility of further processing that data.

4 Cf. also Becker et al. 2022, The authors mention the "generic manner" of data collection, which is unspecific in relation to the data source, also with reference to the wording of Articles 17(1)(a) and 25(2) GDPR.

4.6 Data-based approach

The data-based approach of the purpose limitation principle, implicitly argued by the literature and data protection authorities (European Data Protection Board 2021, p. 6; Specht-Riemenschneider 2023, p. 651; Article 29 Working Party 2013, p. 26[5]), considers the data-lifecycle for the notion of further processing. According to this approach, any processing of personal data that has previously been collected by a controller constitutes further processing of that data, irrespective of the controller for the further processing operation. Thus, the procurement of personal data that has previously been processed by another controller qualifies as further processing and is subject to a compatibility assessment.

For the proposed concept, this would mean that the transmission of personal data by the RDMC-Provider to the RDMC-User constitutes a further processing operation. Likewise, the procurement and subsequent processing of personal data by the RDMC-User for their own research purposes and under their legal responsibility also constitutes further processing of personal data. It is therefore limited by the purposes for which the data was initially collected by the RDMC-Provider.

Under the GDPR, this approach may be based on the following provisions and the following recital:

Art. 5(1) GDPR: The principles for processing personal data refer explicitly to the personal data itself, allowing the interpretation that these principles pertain to the personal data itself. In this regard, Art. 5(1)(b) GDPR refers to the collection of personal data "for specified, legitimate and explicit purposes" as well as further processing without explicitly mentioning the respective controller as a decisive factor. The controller is, however, responsible for complying with these principles (Art. 5(2) GDPR).

Art. 17(2) GDPR: Under certain circumstances, data subjects have the right to demand that the data controller erases personal data related to them. This is the case, for example, if the personal data has been processed unlawfully. If the controller is obliged to erase personal data that they have published, they must take "reasonable steps" to "inform controllers which are processing the personal data that a data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data"

5 The WP29 Opinion 03/2013 has not been officially endorsed by the EDPB, see: https://www.edpb.europa.eu/our-work-tools/general-guidance/endorsed-wp29-guidelines_en; *Visited on March 26, 2025.* With a view to the "determination of 'purposes' ", the EDPB has referenced this WP29 opinion, see EDPB, 2020, Guidelines 05/2020 on consent under Regulation 2016/679, Version 1.1, page 13, footnote 28, available at: https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052020-consent-under-regulation-2016679_en; *Visited on March 26, 2025.*

(Art. 17(2) GDPR). This indicates that a change of the data controller does not constitute new and independent data processing.

Art. 19 GDPR: Referring to Art. 19 GDPR, Frenzel argues that the controller who collected the personal data is obligated to convey the collection purposes to data recipients (Frenzel 2021, para. 29). Art. 19 GDPR obliges the controller to "communicate any rectification or erasure of personal data or restriction of processing" to each recipient to whom the personal data have been disclosed.

Recital 50 GDPR: With regard to the data collection and the purpose limitation principle, the first sentence of recital 50 GDPR refers to purposes "for which the personal data were initially collected", allowing the interpretation that once personal data has been collected for the first time, any subsequent processing is considered further processing of that personal data, irrespective of the controller.

4.7 Conceptual relevance of the different approaches

Both approaches are supported by the GDPR's provisions and (non-binding) recitals. A distinction between the two approaches is also relevant for the concept.

Under the role-based approach, the procurement of personal data by the RDMC-User from the RDMC-Provider is considered an initial data collection. Any further processing by the RDMC-User is limited by the purposes specified by the RDMC-User at the time of the procurement.

Under the data-based approach, the procurement of personal data by the RDMC-User from the RDMC-Provider is considered a further processing of that data. Any further processing by the RDMC-User is limited by the purposes for which the data was initially collected by the RDMC-Provider. In order to assess the legal permissibility of any further processing, the RDMC-User must be aware of the purposes and context of the initial data collection by the RDMC-Provider.

The data-based approach requires the RDMC-Provider to communicate the original data collection purposes to the RDMC-User to ensure that the data is further processed in compliance with the purpose limitation principle.

4.8 Evaluating the approaches based on a teleological interpretation of the purpose limitation principle

As demonstrated, the wording of the GDPR's provisions and recitals can support both approaches. In addition, the teleological interpretation takes into account the aim of the purpose limitation principle stipulated by Art. 5(1)(b) GDPR.

By specifying the processing purpose, the controller sets out the reasons for why they wish to process the personal data. This purpose specification ensures transparent and predictable data processing (Article 29 Working Party 2013, p. 13). It also establishes the basis for assessing other processing principles and related obligations (Article 5 GDPR) in a specific processing context.

Based on the specified and communicated purposes at the time of the data collection, the data subjects can decide whether to provide their data to the controller and whether they wish to influence the processing of their data by asserting their data subject rights. The permissibility of further processing personal data for secondary purposes must take into account, among other things, the reasonable expectations of the data subject based on the initial data collection (Article 6 (4)(b) and Recital 50 GDPR). The purpose limitation principle is therefore fundamental to the data subject's control of their personal data (cf. Recital 7 GDPR).

In this respect, the role-based approach is problematic. It leads to a break in the collection purposes when data is transferred from one controller to another who intends to process the data for secondary purposes. Under this approach, the transmission of personal data for secondary purposes leads to a shift in the basis that limits further processing in relation to the purposes of the initial data collection. This poses the risk that data subjects lose sight of, and control over, the purpose(s) for which their data is being processed. This is because data subjects may not be actively involved in the data transmission and the specification of the new collection purpose(s). This could happen, for example, if data subjects do not respond to consent requests sent to them. This loss of control continues and intensifies with subsequent data transmissions that lead to different collection purposes. Consequently, there is a risk that personal data will be processed for reasons unrelated to the initial purpose(s) for which it was provided by the data subjects and that the data subjects could not have anticipated such a subsequent processing of their personal data. From the data subject's perspective, the role-based approach allows for an unpredictable processing of their personal data.

The data-based approach limits further processing of personal data by any controller with regard to the purposes specified at the time of the initial data collection. If the data is initially collected directly from data subjects, they can assess the reasons and extent to which their data is being processed based on the specified purposes and decide accordingly whether to provide their data. However, even if the data is not collected directly from the

data subjects, communicating the collection purposes enables them to assess any potential further processing of their data. Based on the specified purposes and related expectations by the data subjects, they can assert their rights.

The data-based approach does not prevent subsequent use of personal data from being efficient. Although the initial collection purposes are not inherent to the personal data itself, this can be addressed through the use of high-quality metadata sets that explain these purposes.

The data-based approach leads to a more effective protection of data subjects in relation to the processing of their personal data. It is consistent with the legislative aim of enabling data subjects' control of their personal data (Recital 7 GDPR).

For our concept we follow the data-based approach.

5 Technical issues pertaining to the concept

The exchange of private information between two parties over untrusted networks is a long-standing challenge in Computer Science. Regarding privacy, the primary security objective to achieve is confidentiality of data (Art. 5 (1)(f) GDPR), while integrity and authenticity are also part of the problem. Solutions involving applied cryptography have been developed, they also add additional challenges like key management and usability as the most prominent issues. In order to address this challenge within the NFDIxCS project, we are going to describe currently existing solutions that solve this objective. The key challenge that we face in NFDIxCS is that we as a platform provider do not want to have access to the plaintext data at any given time. This will include a significant risk by taking our infrastructure out of the chain of trust. If RDMC containers include sensitive data, we will provide the user with an option to request this data from the Data Owner. The protected data will however not be processed on the NFDIxCS platform. This leaves us with options to suggest for the Data Owner or third-party service providers.

The proposed options would allow the Data Owner to exchange the protected dataset with a privileged user securely, without that third part service gaining access to the plaintext data.

5.1 Potential solution: End-to-End Encrypted cloud storage

Several service providers offer solutions for end-to-end encrypted storage space. These solutions vary and come with different drawbacks that affect both the security and usability of the solutions. One possible solution, both for third party providers and for self-hosting,

is the end-to-end encrypted app for the Nextcloud platform (Nextcloud GmbH 2021). This approach requires the users that intend to exchange encrypted data, to both have an established account registered with the same Nextcloud server. Additionally, the feature only works through the Nextcloud client applications on users end devices, while web access is out of scope. Consequently, this approach also does not allow exchanging data with unregistered users. The data is encrypted by generated 2048-bit RSA keys within the user's native client applications. The public key pairs can then be found by other registered users on the Nextcloud server and be used to encrypt data intended for the owner of the private key. In case the access to the key material or the end device is lost, a mnemonic passphrase is generated as a recovery key and offered for download to the user during the setup process. This however requires the user to manage this mnemonic passphrase securely and possibly for a long period of time, which presents one of the most difficult challenges on their own. Consequently, many users will either forget this recovery phrase or not store them securely. The use of 2048-bit RSA keys is also problematic, considering that current recommendations for RSA key length is already at 3072-bit (Bundesamt für Sicherheit in der Informationstechnik 2025).

Alternative solutions offered by service providers can be found with "Proton Drive" (Proton AG 2022) or the open-source solution provided by "Filen" (Filen Cloud Dienste UG 2025). Both providers implement encrypted cloud storage independent of the used end device and work in web browser sessions as well. While their internal implementations differ, they both make use of the user's password for both authentication and symmetric encryption. This allows any user to log in to their encrypted cloud storage through any end device or web browser and decrypt their data on their own device, either through a client application or within a browser session. Another advantage provided by these solutions is the option to share encrypted data by generating a URL with an additional generated password that is used to decrypt the shared file. While the URL is generated by the server application, the creation of the password and re-encryption of the shared file takes place on the users end device. Consequently, the end-to-end encryption is maintained while this approach also allows sharing the encrypted files with users outside the provider's registered user base. Unlike the previously described solution this approach is much more flexible since it allows secure file sharing without requiring specific client applications and even with yet unknown recipients.

5.2 Potential solution: Browser side encryption

Part of the above-described solutions was the option to generate URLs and keys to enable the download of files by users outside the hosting service. This option can also be implemented directly into an upload service without user registration and management. One such implementation can be found with "Firefox Send" which is now developed as a community drive fork (Visée 2025) of the original Mozilla project. This service enables a file upload interface that can be used to upload files which are then encrypted on the

client within the browser session before the encrypted file is uploaded. A download URL is then generated with the decryption key as a parameter generated by the client. The hosting server only sees the encrypted data and is thereby removed from the chain of trust regarding the confidentiality of the contents. The advantage with this solution is the simplicity and lack of user management. A potential issue with this approach is the potential for abuse, which was also named as a reason for the shutdown of the original "Firefox Send" project (Mozilla Foundation 2023), which was faced with malicious users mounting malware and phishing attacks through the service. In the context of the NFDIxCS project, this issue could be addressed however by providing this service only to users that are authenticated by a research institution through the single-sign-on service like the academic cloud project (Gesellschaft für wissenschaftliche Datenverarbeitung 2025). For repeated download and persistent data management the previously discussed cloud storage solutions may provide a more suitable solution that integrated this client-side URL generation for e2ee download.

5.3 Unsolved problems

All proposed solutions include different drawbacks that need to be considered. The most pressing issue for solutions that require authentication is the use of user defined passwords for authentication and encryption. While the introduced solutions handle both parts appropriately in their implementation, the use of passwords is generally a significant security concern. In terms of authentication, it has long been known that passwords are a very weak method (Zviran and Haga 1999) that are commonly the target of attacks (Bundesamt für Sicherheit in der Informationstechnik 2019). The introduction of two-factor-authentication can help to mitigate some of these risks, but most 2fa methods will not address all these threats. A primary attack vector against authentication methods to this day is still the use of phishing (Hong 2012), which includes tricking the user into providing their password to a fake web site. Most second factor authentication methods are also vulnerable to phishing. One commonly available alternative that solves these issues entirely can be found with the Fido2 specification (Fido Alliance 2020), which allows for entirely phishing resistant authentication process. This method is already available in all commonly used browsers and operating systems, including mobile devices. The most widespread implementation of Fido2 can be found with "Passkeys", which has become popular on mobile platforms like Android and IOS.

5.4 Further Research

Future solutions need to address the problem of phishing resistant authentication while also providing an option for client-side encryption. Such a solution needs to be both convenient to use and secure to be implemented across common operating systems and

devices. The Fido2 specification has already shown great progress in terms of authentication and is actively being pushed by large vendors to have received widespread adoption. Future work will address potential solutions based on this method that involves client-side encryption in common web applications, which will enhance the existing solutions and address remaining issues.

6 Conclusion

The transmission of personal data by the RDMC-Provider to the RDMC-User is a sensitive issue that needs consideration of various legal, ethical and technical aspects.

With regard to the purpose limitation principle, the transmission of personal data for secondary purposes constitutes a further processing of the personal data. Such a transmission is only legally permissible, if the purposes of the intended further processing are not incompatible with the purposes of the initial data collection. Processing personal data for historical or scientific research purposes is not incompatible with the initial data collection purposes, provided that necessary safeguards exist to protect the data subjects. Following the data-based approach of the purpose limitation principle, the RDMC-Provider must communicate the purposes of the initial data collection to the RDMC-User before transmitting the personal data. The RDMC-User must communicate the further processing purposes as well as the existing safeguards to the RDMC-Provider, who needs to assess the legal admissibility of the further processing before transmitting the personal data. The RDMC-Provider and the RDMC-User should lay down the binding conditions for the further processing of the personal data via a written contract. Both, the RDMC-Provider and the RDMC-User must inform the data subjects of the intended further processing of their personal data with enough time in advance for the data subjects to potentially assert their rights prior to the further processing.

For the transmission of protected data between the RDMC-Provider and RDMC-User, several technical solutions are available to design this process in a secure and reliable manner without requiring technical knowledge or training of the users. The use of end-to-end encryption will ensure that the potential for compromised data is limited to a minimum and avoid the common issues with cloud-based services and data protection.

Acknowledgements

The authors would like to thank the anonymous reviewers and the editors for their helpful feedback and comments on the manuscript. Furthermore, the authors would like to thank Dr. Dara Hallinan for his suggestions on linguistic improvements to the manuscript.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – (NFDI-52/1) – 501930651.

Authorship Contributions

- Part 1-4 of the paper is authored by Patrick Brunner.
- Part 5 of the paper is authored by Stefan Bavendiek.
- The abstract and Part 6 are co-authored.

Conflict of Interest

The authors work on the NFDIxCS project.

Bibliography

Article 29 Working Party. 2013. *Opinion 03/2013 on purpose limitation. WP 203.* Visited on March 26, 2025. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

Becker, Regina, Davit Chokoshvili, Giovanni Comandé, Edward S. Dove, Alison Hall, Colin Mitchell, Fruzsina Molnár-Gábor, Pilar Nicolàs, Sini Tervo, and Adrian Thorogood. 2022. "Secondary Use of Personal Health Data: When Is It "Further Processing" Under the GDPR, and What Are the Implications for Data Controllers?" *European Journal of Health Law* 30 (2): 129–157. https://doi.org/10.1163/15718093-bja10094.

Bundesamt für Sicherheit in der Informationstechnik. 2019. *The state of IT security in Germany in 2019.* Technical report. Visited on June 27, 2025. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/Securitysituation/IT-Security-Situation-in-Germany-2019.pdf?___blob=publicationFile&v=3.

———. 2025. *Cryptographic Mechanisms: Recommendations and Key Lengths.* Technical report. Visited on June 27, 2025. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/TechGuidelines/TG02102/BSI-TR-02102-1.pdf.

European Data Protection Board. 2018. *Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak.* Visited on March 26, 2025. https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-032020-processing-data-concerning-health-purpose_en.

European Data Protection Board. 2021. *Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research.* Visited on March 25, 2025. https://www.edpb.europa.eu/our-work-tools/our-documents/other-guidance/edpb-document-response-request-european-commission_en.

Fido Alliance. 2020. *FIDO UAF Architectural Overview.* Visited on March 25, 2025. https://fidoalliance.org/specs/fido-uaf-v1.2-ps-20201020/fido-uaf-overview-v1.2-ps-20201020.pdf.

Filen Cloud Dienste UG. 2025. *Documentation: Cryptography.* Visited on March 27, 2025. https://docs.filen.io/docs/api/guides/cryptography/.

Frenzel, Eike Michael. 2021. "DS-GVO Art. 5 Grundsätze für die Verarbeitung personenbezogener Daten". In *Datenschutz-Grundverordnung Bundesdatenschutzgesetz: DS-GVO BDSG,* 3rd edition, edited by Boris P. Paal and Daniel A. Pauly. München: C. H. Beck. ISBN: 978-3-406-75374-9.

Gesellschaft für wissenschaftliche Datenverarbeitung. 2025. *Academic Cloud.* Visited on March 27, 2025. https://academiccloud.de.

Hong, Jason. 2012. "The state of phishing attacks". *Communications of the ACM* 55 (1): 74–81. https://doi.org/10.1145/2063176.2063197.

Lay, Winnie, Loretta Gasparini, William Siero, and Elizabeth K Hughes. 2024. "A rapid review of the benefits and challenges of dynamic consent". *Research Ethics* 21 (1): 180–202. https://doi.org/10.1177/17470161241278064.

Mozilla Foundation. 2023. *What happened to Firefox Send?* Visited on March 27, 2025. https://support.mozilla.org/en-US/kb/what-happened-firefox-send.

Nextcloud GmbH. 2021. *Nextcloud end-to-end encryption RFC.* Visited on March 27, 2025. https://github.com/nextcloud/end_to_end_encryption_rfc/blob/master/RFC.md.

Proton AG. 2022. *The Proton Drive security model.* Visited on March 27, 2025. https://proton.me/blog/protondrive-security.

Purtova, Nadezhda, and Ronald Leenes. 2023. "Code as personal data: implications for data protection law and regulation of algorithms". *International Data Privacy Law* 13 (4): 245–266. https://doi.org/10.1093/idpl/ipad019.

Roßnagel, Alexander. 2025a. "DS-GVO Art. 4 Nr. 2 Begriffsbestimmung "Verarbeitung"". In *Datenschutzrecht,* 2nd edition, edited by Spiro Simitis, Gerrit Hornung, and Indra Spiecker. Nomos Verlag. ISBN: 978-3-8487-8958-0, visited on June 27, 2025. https://beck-online.beck.de/Bcid/Y-400-W-SimHorSpiKoDatenSchR-G-EWG_DSGVO-A-4-N-2.

———. 2025b. “DS-GVO Art. 5 Grundsätze für die Verarbeitung personenbezogener Daten”. In *Datenschutzrecht,* 2nd edition, edited by Spiro Simitis, Gerrit Hornung, and Indra Spiecker. Nomos Verlag. ISBN: 978-3-8487-8958-0.

Saporito, Antonio, Parinaz Tabari, Mattia De Rosa, Vittorio Fuccella, and Gennaro Costagliola. 2024. “Exploring the Privacy Horizons: A Survey on HCI and HRI”. In *Computational Science and Its Applications – ICCSA 2024 Workshops,* 113–125. Springer Nature Switzerland. ISBN: 9783031653186. https://doi.org/10.1007/978-3-031-65318-6_8.

Specht-Riemenschneider, Louisa. 2023. “Datennutz und Datenschutz: Zum Verhältnis zwischen Datenwirtschaftsrecht und DSGVO”. *Zeitschrift für Europäisches Privatrecht,* number 3, 638–672.

Visée, Tim. 2025. *Send.* Visited on March 27, 2025. https://github.com/timvisee/send.

Zviran, Moshe, and William J. Haga. 1999. “Password security: an empirical study”. *Journal of Management Information Systems* 15 (4): 161–185. Visited on June 27, 2025. https://hdl.handle.net/10945/40319.

Kooperation zwischen Universität und Informationsinfrastruktur – das Forschungsdatenrepositorium RADAR für Brandenburg

Maria Büttner [1], Boris Jacob [1], Daniela Mertzen [1], Christian Riedel [1], Kerstin Soltau [2]

[1] Universität Potsdam, Potsdam, Deutschland;
[2] FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Karlsruhe, Deutschland

In diesem Beitrag liegt der Fokus auf den Fortschritten der Kooperation an der Universität Potsdam zwischen Forschung und FDM-Support im Bereich der technisch-wissenschaftlichen Dienste, insbesondere bei der überinstitutionellen Etablierung und dem dauerhaften Betrieb des Datenrepositoriums RADAR für die Forschungsdaten-Archivierung und -Publikation.

Die Repositoriumssoftware RADAR wird vom FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur entwickelt. FIZ Karlsruhe betreibt diese bislang zum einen als eigene Cloud-Instanz[1] (RADAR Cloud), zum anderen auch im Auftrag einer Einrichtung auf deren lokaler IT (RADAR Local), so an der Universität Potsdam. Diese lokale Instanz wird im Kontext des Projekts IN-FDM-BB, einem Verbund von acht staatlichen forschenden Hochschulen in Brandenburg, zu einer regionalen Cloud erweitert. RADAR wird so zu RADAR-BB und über die Grenzen der Universität Potsdam hinaus geöffnet: Von einem institutionellen Repositorium für die Universität Potsdam zu einem überinstitutionellen Shared Service für Brandenburg zur Archivierung und Publikation von Forschungsdaten.

RADAR-BB wird zukünftig auch durch den an der UP angesiedelte SFB 1294 – „Data Assimilation“ genutzt. Aus Sicht der Forschenden und der FDM-Beauftragten des Sonderforschungsbereichs (SFB) ist ein zentrales, von der Universität auf eigener IT betriebenes

1 https://www.radar-service.eu/; *Besucht am 3. Juli 2025 durch die Herausgeber*innen.*

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23933 (CC BY-SA 4.0).

Repositorium von großem Nutzen. Dies wird ebenso erläutert, wie der dazugehörige Workflow. Dieser kann, gemeinsam mit weiteren in Brandenburg erprobten Workflows sowie Erfahrungswerten, als Blue Print für Institutionen und Verbünde bei der Entwicklung eigener Lösungen dienen.

Vorgestellt werden die technischen Grundlagen, die organisatorische Umsetzung und die zukünftige Implementierung sowohl für den SFB als auch für die beteiligten öffentlichen Hochschulen Brandenburgs. Außerdem wird auf Herausforderungen in der Umsetzung insbesondere organisatorischer und finanzieller Art eingegangen.

Keywords: Forschungsdatenmanagement, Repositorium, Landesinitiative, FDM-BB, RADAR

1 Einleitung

An der Universität Potsdam setzen Forschung und FDM-Support ihre Zusammenarbeit im Bereich Forschungsdaten fort – es wurde bereits über die enge Kooperation von FDM-Team und Sonderforschungsbereich (SFB) 1294 *Data Assimilation* in den Bereichen Schulungen/Training, Öffentlichkeitsarbeit/Transfer und technisch-wissenschaftliche Dienste (Mertzen u. a. 2023)[2] berichtet. In diesem Beitrag liegt der Fokus auf den Fortschritten im Bereich der technisch-wissenschaftlichen Dienste, konkret auf dem Aufbau des Datenrepositoriums RADAR für die Forschungsdatenarchivierung und -publikation. Besonders ist, dass inzwischen eine weitere Dimension hinzugekommen ist: Im Rahmen des Projekts IN-FDM-BB soll dieser FDM-Dienst überinstitutionell aufgebaut und betrieben werden. Nachfolgend werden technische Bedingungen, organisatorischer Aufbau und der Vielfalt der Hochschullandschaft flexibel dienende Workflows vorgestellt.

2 Bedarfe

2.1 Universität Potsdam

An der Uni Potsdam wurde 2018/2019 ein Multi-Stakeholder-Strategieprozess zum Thema Forschungsdaten durchgeführt. Der Prozess wurde auf Basis des Research Infrastructure Self-Evaluation Frameworks (RISE v1.1) des Digital Curation Centres gestaltet, dieses

2 FDM-Capacity Building im Tandem – Kooperation von Zentraleinrichtung und Sonderforschungsbereich an der Universität Potsdam. Vortrag bei den E-Science-Tagen 2023 (Heuveline, Bisheh und Kling 2023). Siehe Beiträge: https://heibox.uni-heidelberg.de/d/37fcfbffd30340ca88ba/; *Besucht am 19. März 2025.*

wurde im Verlauf ins Deutsche übersetzt und an den deutschen Wissenschaftskontext angepasst. RISE-DE ist ein Referenzmodell für Strategieprozesse im institutionellen Forschungsdatenmanagement und bietet einen Bewertungsrahmen zur Selbstevaluation und Zielbestimmung, das sich als Werkzeug zur Gestaltung einer strukturierten, stakeholderorientierten Strategieentwicklung für das FDM an deutschsprachigen Hochschulen und Forschungseinrichtungen eignet (Hartmann, Jacob und Weiß 2019, S. 3).

Im Ergebnis des Strategieprozesses verabschiedete die Universität Potsdam 2020 ihre erste Forschungsdatenstrategie, in der folgende Handlungsziele in Bezug auf Archivierung und Publikation von Daten und Code an der UP festgelegt wurden:

> „[1] Ein Dienst zur institutionellen (Bitstream-)Archivierung von Daten an der UP in Übereinstimmung mit Empfehlung 7 der Richtlinien zur guten wissenschaftlichen Praxis der DFG ist nutzbar. [...] [2] Ein Dienst zur DOI-identifizierten Publikation von Daten an der UP ist verfügbar. Dieser soll jedoch nur genutzt werden, wenn keine geeigneten disziplin- bzw. datentypspezifischen Infrastrukturen vorhanden sind. Die Möglichkeit zur DOI-Vergabe für im Versionierungsdienst der UP gehaltenen Code besteht. [...] [3] Die organisatorischen und ggf. technischen Voraussetzungen für den Nachweis von Daten- und Softwareveröffentlichungen in der Bibliographie der Universität Potsdam sind geschaffen [...]“ (Universität Potsdam, Präsidium 2020, S. 17).

2.2 SFB 1294 – Anforderungen eines Sonderforschungsbereichs

Im Sonderforschungsbereich 1294 Data Assimilation[3] arbeiten Forschende aus sechs Forschungseinrichtungen zusammen: Der Humboldt-Universität zu Berlin, der Technischen Universität Berlin, dem Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS), dem GFZ Helmholtz-Zentrum für Geoforschung und der Technischen Universität Ilmenau. Das Forschungsdatenmanagement des SFB ist an der Universität Potsdam angesiedelt.

Im Mittelpunkt des Forschungsdatenmanagements des SFB stehen die Verbesserung der Reproduzierbarkeit, das kollaborative Arbeiten sowie die Umsetzung der DFG-Richtlinien zur guten wissenschaftlichen Praxis[4] (Deutsche Forschungsgemeinschaft 2025). Zu diesem Zweck nutzen die Forschenden eine Reihe von IT-Plattformen, die entweder der zentralen Infrastruktur der Universität Potsdam angehören, im SFB selbst betrieben werden oder extern angesiedelt sind.

Der Betrieb einer eigenen IT-Infrastruktur ist für einen SFB jedoch mit einigen Hürden verbunden. Der Betrieb benötigt zunächst eine*n SFB-Mitarbeiter*in, die*der über

3 https://www.sfb1294.de/; *Besucht am 20. März 2025.*
4 https://www.sfb1294.de/research/research-area-z/z03; *Besucht am 27. Mai 2025.*

IT-Administrationskenntnisse verfügt. Solche Kenntnisse sind zwingend erforderlich, um den reibungslosen Betrieb, das Einspielen sicherheitsrelevanter Updates, das Einpflegen SFB-spezifischer Anpassungen der Plattform sowie die notwendigen Absprachen mit der zentralen Universitäts-IT zu gewährleisten. Das hierfür notwendige Kompetenzprofil ist für wissenschaftliches Personal in Form eines Data Stewards nicht üblich und daher in Forschungsanträgen oft nicht vorgesehen.

Aus diesem Grund sowie aus IT-Sicherheitsgründen ist an der Universität Potsdam der Betrieb einer eigenen IT-Infrastruktur nur eingeschränkt möglich. Das hat zur Folge, dass der Zugriff auf das vom SFB 1294 selbst betriebene Datenarchiv DSpace nur aus dem Netz der Universität Potsdam oder mit Hilfe eines Accounts der Universität Potsdam über VPN möglich ist. Dieser Umstand schafft Barrieren für die Forschenden und erschwert die Archivierung der im SFB erhobenen Datensätze gemäß den Richtlinien der DFG. Die in SFBs übliche universitätsübergreifende Zusammenarbeit wird dadurch erheblich erschwert.

Die Nutzung zentraler hochschuleigener Infrastruktur mit geregeltem sicherem Zugang würde kollaboratives Arbeiten erleichtern. Außerdem könnte die Verantwortung für IT-Sicherheit, Updates und technischen Support an eine zentrale Stelle ausgelagert werden. Auch dauerhafte Speicherung wäre auf einer hochschuleigenen Infrastruktur gesichert. Die dauerhafte Sicherung auf DSpace ist über das Ende der Laufzeit des SFBs hinaus nicht gewährleistet.

2.3 Kooperative Bereitstellung eines FDM-Dienstes in Brandenburg

Die acht staatlich geförderten forschenden Hochschulen in Brandenburg arbeiten bereits seit Ende 2019 in dem vom brandenburgischen Ministerium für Wissenschaft, Forschung und Kultur (MWFK) geförderten Projekt „Forschungsdatenmanagement in Brandenburg“ (FDM-BB) eng zusammen, aus dem sowohl die Landesinitiative FDM-BB entstand als auch der Grundstein für das Projekt IN-FDM-BB gelegt wurde. Ziel des seit Oktober 2022 geförderten Projekts IN-FDM-BB ist die Institutionalisierung eines nachhaltigen Forschungsdatenmanagements in Brandenburg. Neben dem Aufbau lokaler FDM-Kompetenzen und -Dienstleistungen sowie dem Ausbau von Vernetzung und Wissenstransfer, sollen auch forschungsunterstützende IT-Dienste dauerhaft hochschulübergreifend als Shared Services angeboten werden (vgl. Forschungsdatenstrategie für das Land Brandenburg vom Ministerium für Wissenschaft, Forschung und Kultur des Landes Brandenburg 2022, S. 6f.).

Die brandenburgweite, kooperative Bereitstellung wissenschaftlich-technologischer IT-Dienste stellt dabei eine besondere Herausforderung aufgrund unterschiedlicher Bedarfe, Kompetenzen und Ressourcen je Hochschule dar. Nach sorgfältiger Prüfung konnten

der Research Data Management Organiser (RDMO) als Tool zur Erstellung von Datenmanagementplänen sowie RADAR als bundeslandweites disziplinübergreifendes Forschungsdaten-Repositorium identifiziert werden (Jacob u. a. 2024). Nachfolgend wird ausschließlich auf die kooperative Implementierung von RADAR eingegangen.

Um RADAR für Verbundzwecke nutzbar zu machen, wurde die an der Universität Potsdam betriebene RADAR Local-Instanz in Zusammenarbeit mit FIZ Karlsruhe ausgebaut, unter Berücksichtigung der individuellen Bedarfe der verschiedenen Hochschulen als auch der Bedarfe des Verbundes. Dabei konnte und kann auf Erfahrungswerte des lokalen IT- und FDM-Personals zurückgegriffen werden. Software, Speicherinfrastruktur und technisches Know-How sollen dabei zentral bereitgestellt werden, damit auch kleinere Hochschulen des Verbundes, die nicht über notwendiges Personal oder Infrastruktur verfügen, den Dienst nutzen können. Das Repositorium soll insbesondere den Fachrichtungen dienen, für die keine disziplinspezifische Infrastruktur existiert.

3 Realisation

3.1 Ausbau von RADAR Local zu einer regionalen Cloud für Brandenburg

Im Rahmen des Projekts IN-FDM-BB soll ein generisches Forschungsdatenrepositorium zur hochschulübergreifenden Nutzung bereitgestellt werden, das zentral gehostet und technisch betreut wird, gleichzeitig jedoch flexibel genug ist, um den unterschiedlichen Anforderungen der brandenburgischen Hochschulen gerecht zu werden.

Die Repositoriumssoftware RADAR wird an der Universität Potsdam in der Betriebsvariante RADAR Local eingesetzt. Die Pflege und (Weiter-)Entwicklung der Software erfolgt zentral durch FIZ Karlsruhe (Leibniz-Institut für Informationsinfrastruktur) – ohne personelle Bindung auf Seiten der Institution. Bei RADAR Local betreibt FIZ Karlsruhe die RADAR-Software auf der lokalen IT-Infrastruktur der nutzenden Einrichtung und unter ausschließlicher Verwendung institutionseigener Speichersysteme. Der Dienst wird unter einer lokalen Domain angeboten; die Anbieterrolle für den Dienst liegt dabei vollständig bei der nutzenden Einrichtung. Gleichzeitig werden maximale Datensouveränität und -sicherheit gewährleistet, da die Daten die Heimateinrichtung nicht verlassen.

Die gemeinsame Nutzung einer lokalen RADAR-Instanz durch mehrere Einrichtungen war bislang nicht vorgesehen, wurde durch die RADAR-Software nicht unterstützt und stellt somit eine Neuerung dar. Für das Verbundprojekt IN-FDM-BB wurde die an der Universität Potsdam betriebene RADAR-Instanz zu einer überinstitutionell nutzbaren Cloud-Plattform – RADAR-BB – weiterentwickelt. Für die Nutzung im Verbund erweiterte FIZ Karlsruhe die RADAR-Software entsprechend, sodass innerhalb einer lokalen

Instanz unterschiedliche Umgebungen für einzelne Institutionen eingerichtet und gepflegt werden können. Auf der RADAR-BB Startseite werden alle publizierten Forschungsdaten mit dem Logo der jeweiligen Hochschule ausgezeichnet. So werden Verbund-Design und Hochschul-Identität gleichermaßen gewährleistet. Das Repositorium trägt so zur Sichtbarkeit des Forschungsoutputs der brandenburgischen Hochschulen bei.

3.2 Funktionalitäten von RADAR-BB

RADAR-BB bietet für die Archivierung und Publikation von Forschungsdaten, als auch für deren optionalen Vorab-Review, ein breites Spektrum an Funktionen: Bei der Archivierung können Daten sicher über flexibel wählbare Haltefristen (z. B. 10 Jahre nach DFG-Kodex) aufbewahrt werden. Der Zugriff auf archivierte Daten kann über unterschiedliche Zugriffsrechte granular gesteuert werden. Durch die Publikation erhalten Forschungsdaten einen DOI (DataCite), und werden mindestens 25 Jahre aufbewahrt und öffentlich verfügbar gemacht. Optional sind befristete oder unbegrenzte Embargos möglich. Die Metadaten werden über einen OAI-Provider zum Harvesting angeboten. Darüber hinaus gibt es die Möglichkeit des Peer Review, bei dem Daten durch externe Personen im Vorfeld einer Publikation begutachtet werden können.

Der Upload von Forschungsdaten in das Repositorium kann auf verschiedenen Wegen erfolgen: über ein webbasiertes Formular, via RADAR API (für automatisierte Prozesse), WebDAV-unterstützt oder über Importe aus GitHub (ab Mitte 2025 auch aus GitLab).

Für die Metadatenannotation stehen ein formularbasierter Editor, eine XML-Uploadoption sowie die RADAR API zur Verfügung. Zudem können Metadaten auch aus RDMO nach RADAR-BB übertragen werden (Wedlich-Zachodin u. a. 2024). Das generische RADAR-Metadatenschemata orientiert sich am DataCite Metadata Kernel, bietet 10 Pflicht- und 13 optionale Felder, eine Vielzahl an kontrollierten Vokabularen, Terminologien und Normdaten (z. B. ORCID, ROR, GND sowie sämtliche über TS4NFDI[5] verfügbare Ontologien) und ist als Standard auf FAIRsharing[6] gelistet.

Zur Authentifizierung kann sowohl die institutionelle Nutzerkennung über Shibboleth (DFN-AAI) als auch ein manuell eingerichtetes Benutzerkonto genutzt werden.

5 Terminology Services 4 NFDI (TS4NFDI): https://terminology.services.base4nfdi.de/; *Besucht am 28. Mai 2025.*

6 FAIRsharing.org: RADAR Metadata Schema, DOI: https://doi.org/10.25504/FAIRsharing.e26f92; *Zuletzt editiert am 31. Januar 2023, 18:17 Uhr; Besucht am 25. März 2025.*

3.3 Anpassungsoptionen und institutionelle Integration bei RADAR-BB

Jede institutionelle RADAR-Umgebung lässt sich sowohl visuell als auch organisatorisch den individuellen Anforderungen anpassen und eigenständig verwalten. Zu den möglichen Anpassungen zählen unter anderem individuelles Branding (z. B. Logo, Schmuckfarbe), eigenes DOI-Präfix pro Institution, Einrichtung von Arbeitsbereichen (z. B. für Forschungsgruppen oder Projekte), Verwaltung der Speicherquoten, Festlegung von Standardwerten für Metadaten und Zugriff auf nutzungsspezifische Statistiken.

Im Rahmen der landesweiten Implementierung als regionale Cloud RADAR-BB kann zusätzlich das hochschulübergreifende Forschungsdatenportal individuell gestaltet werden. Es bietet unter Brandenburger Domain zentralen Zugriff auf Datenpublikationen der beteiligten Hochschulen.

Alle Forschungsdaten innerhalb von RADAR-BB werden zentral im Archivspeicher an der Universität Potsdam gesichert. So ist es auch kleineren Hochschulen möglich, DFG- und datenschutzkonform die eigenen Forschungsdaten als Bitstream zu archivieren und zu publizieren. Kosten für Speicher, Personal sowie Know-How werden vom Verbund geteilt.

3.4 Technische Anforderungen beim Betrieb von RADAR BB

Für den Betrieb von RADAR BB stellt die Universität Potsdam FIZ Karlsruhe als IT-Ressourcen zwei virtuelle Maschinen (jeweils mit Quad-Core-Prozessoren mit mind. 3GHz, 16 GB RAM und 200 GB HD) inkl. Betriebssystem zur Installation der RADAR-Software, ein Network File System (NFS) mit mindestens 100 GB als temporärer Speicher zur Organisation der Forschungsdaten sowie Archivspeicher (Festplatte, bedarfsorientierte Größe) für die dauerhafte Speicherung der archivierten und publizierten Daten bereit. Der Zugriff durch FIZ Karlsruhe erfolgt über Virtual Private Network (VPN). Die Sicherung aller Daten (Archiv, NFS, virtuelle Maschinen) wird durch die Universität Potsdam übernommen. Eine detaillierte Checkliste[7] zu den IT-Voraussetzungen für RADAR Local ist auf der Website von FIZ Karlsruhe verfügbar.

FIZ Karlsruhe gewährleistet im Gegenzug den stabilen Betrieb der RADAR-Software auf den bereitgestellten virtuellen Maschinen. Dazu zählen die Anbindung an institutionelle Speichersysteme auf Ebene der Software, die Wartung der Software sowie das Einspielen von Weiterentwicklungen und Updates. Um die parallele Installation und Pflege der RADAR-Software auf mehreren lokalen Instanzen effizient zu gestalten, wurde der

7 https://radar.products.fiz-karlsruhe.de/sites/default/files/radar/docs/info/RADAR_Local_Checkliste.pdf; *Besucht am 23. Mai 2025.*

Automatisierungsgrad in den Bereichen Softwareentwicklung, Systemkonfiguration und Deployment deutlich erhöht. Dies wird durch den Einsatz von Container-Virtualisierung (z. B. Docker, Docker Swarm) für die Vielzahl an Microservices sowie mittels Orchestrierungswerkzeugen wie Ansible realisiert (Landwehr u. a. 2022).

4 Rollen und organisatorische Umsetzung

4.1 Technisch-organisatorische Zuständigkeit

Die hochschulübergreifende Nutzung von RADAR-BB im Bundesland Brandenburg erfolgt nach klaren Zuständigkeiten und Workflows:

Die *Universität Potsdam* tritt als zentraler technischer und rechtlicher Anbieter auf. Sie betreibt das Repositorium, stellt die Speicherinfrastruktur bereit und verwaltet das überinstitutionelle Design des Forschungsdatenportals. Zudem agiert sie als Vertragspartner gegenüber den beteiligten Hochschulen und rechnet die Speichernutzung mit diesen ab. Darüber hinaus schult sie verantwortliches Personal der Hochschulen und bietet bundeslandweiten 1st-Level-Support.

Die *beteiligten Hochschulen in Brandenburg* benennen jeweils eigene Administrator*innen zur Verwaltung ihrer institutionellen RADAR-BB-Umgebungen. Sie richten darin Arbeitsbereiche ein und bilden institutionelle Workflows bei der Archivierung und Publikation von Forschungsdaten ab.

Die technische Betreuung der Plattform erfolgt in enger Zusammenarbeit mit FIZ Karlsruhe. *FIZ Karlsruhe* betreibt hierfür die RADAR-Software im Auftrag der Universität Potsdam komplett auf der lokalen IT-Infrastruktur und sorgt für die Anbindung an die universitätseigenen Speichersysteme. Zudem agiert es als technischer 2nd-Level-Support und verantwortet die Weiterentwicklung.

4.2 RADAR Rollen- und Rechtekonzept

RADAR arbeitet mit einem festen Rollen- und Rechtekonzept (Abbildung 1). Bei der lokalen Umsetzung von RADAR-BB an jeder Hochschule gibt es die drei zentralen Rollen Administrator*in, Kurator*in und Subkurator*in.

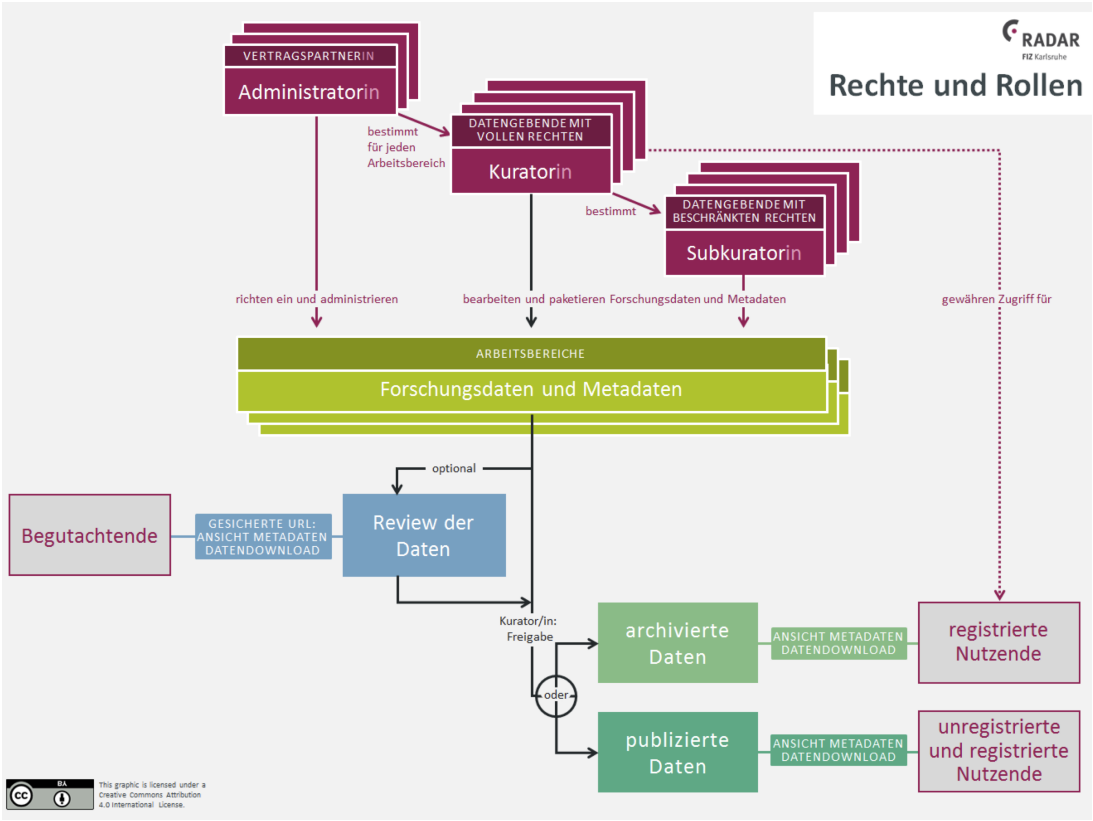


Abbildung 1: Rollen- und Rechtemodell RADAR[8].

*Administrator*innen* verwalten RADAR an der Institution, richten Arbeitsbereiche ein und benennen Datengeber*innen (Kurator*innen). *Kurator*innen* laden Daten hoch, annotieren Metadaten, kuratieren, publizieren und/oder archivieren, außerdem können sie Reviewer benennen und beauftragen. *Subkurator*innen* laden Daten hoch und annotieren Metadaten, können die Datenpakete jedoch nicht freigeben. Dies ist Kurator*innen und Administrator*innen vorbehalten. Diese Rollen sind festgelegt, aber flexibel anpassbar an interne Organisation, Verantwortlichkeiten und Workflows für Datenbereitstellung und Kuratierung, aufgrund individuell einsetzbarer Rollenzuordnung.

4.3 Verbund-Administration und Beispielworkflow „Scientist“

Im Zuge der Vorbereitung, des Trainings und der Implementierung von RADAR-BB wurden verschiedene Workflows für Brandenburg entwickelt, um den unterschiedlichen Betriebsszenarien gerecht zu werden. Zentral ist dabei die Rolle RADAR-BB-Verbund-Administrator*in, die*der für die Ersteinrichtung der Vertragsbereiche der einzelnen Hoch-

8 https://radar.products.fiz-karlsruhe.de/de/radarfeatures/rollen-und-rechtemodell; *Besucht am 28. März 2025.*

schulen, d. h. der hochschuleigenen Arbeitsumgebungen, für den technischen 1st-Level-Support und für die Speicherinfrastruktur des Gesamtsystems zuständig ist. In Brandenburg ist diese Rolle an der Universität Potsdam angesiedelt. und bildet eine Art Rahmen um die gesamte technisch-organisatorische Struktur (Abbildung 2).

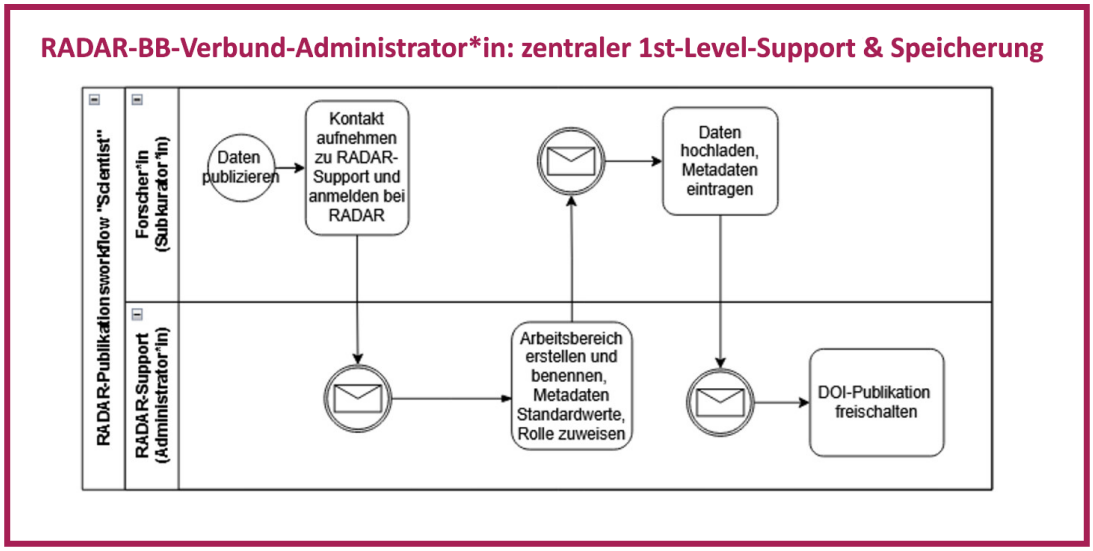


Abbildung 2: RADAR-BB-Verbund-Administrator*in und Workflow „Scientist“.

Innerhalb dieses Rahmens finden verschiedene Betriebsszenarien und Publikationsworkflows statt. Um auf die Vielfalt der Hochschullandschaft und Breite der Fachgebiete flexibel eingehen zu können, sind diese je nach vorhandenen Personal- und Zeitressourcen der jeweiligen Hochschule oder der Forschenden einsetzbar. In Abbildung 2 ist der Workflow abgebildet, der mit „Scientist“ bezeichnet wird. Dabei übernimmt der hochschuleigene RADAR-Support (z. B. FDM-Verantwortliche einer Hochschule) sowohl die Rollen Administrator*in als auch Kurator*in. Die datengebenden Forschenden sind als Subkurator*innen für die Bereitstellung der Daten und Metadaten zuständig. Dabei werden sie vom RADAR-Support unterstützt. Die Freigabe, im Sinne der Publikation, erfolgt durch den RADAR-Support. Dieses Modell eignet sich, wenn Datengebende wenig Erfahrung mit bzw. keine personellen und/oder zeitlichen Ressourcen für die Publikation in einem Repositorium haben.

4.4 Kooperativer Workflow: Strukturierte Bereiche und Projekte mit Data Stewards

Das FDM-Team des SFB 1294 Data Assimilation (WiMi/Data Steward & IT-Admin) unterstützt Forschende bei der Einhaltung der DFG-Richtlinien im Umgang mit Daten und Software. Bei der Verteilung der Verantwortlichkeiten für die Archivierung von Daten und

Software gibt es in Forschungsverbünden unterschiedliche Ansätze. Im SFB 1294 wird eine Philosophie der forschungsbegleitenden Vermittlung von FDM-Kompetenzen verfolgt. Das bedeutet, dass das FDM-Team des SFB für die Bereitstellung der IT-Infrastruktur, die Vermittlung von Kompetenzen und die Beratung der Forschenden verantwortlich ist. Die Forschenden wiederum sollen die FDM-Best-Practices als Bestandteil guter wissenschaftlicher Praxis verstehen und in ihre Forschungsarbeit integrieren. Daraus folgt, dass die Forschenden dafür verantwortlich sind, alle digitalen Artefakte, die im Rahmen einer Peer-Review-Publikation entstehen, im SFB-internen Archiv DSpace zu hinterlegen. Das FDM-Team des SFB prüft die Einträge auf Vollständigkeit, kommuniziert eventuelle Nachbesserungen mit den beteiligten Forschenden und gibt die Einträge schließlich für die Archivierung frei.

Dieser Workflow ist auch nach der Migration des DSpace-Archivs nach RADAR-BB möglich. In diesem Fall würde entsprechend dem Rollen- und Rechtekonzept von RADAR-BB den Forschenden die Rolle Subkurator*in und dem FDM-Team des SFB die Rolle Kurator*in zugewiesen. Forschende, die ihre digitalen Artefakte in RADAR-BB archivieren oder publizieren möchten, kontaktieren das FDM-Team des SFB. In Anlehnung an das Publikations- und Archivierungskonzept des SFB, erstellt der RADAR-Support der Universität Potsdam (Administrator*in) den entsprechenden Arbeitsbereich und weist dem FDM-Team die Rolle der Kurator*innen zu. Das FDM-Team richtet den Arbeitsbereich weiter ein, definiert beispielsweise Metadaten-Standardwerte und weist den Forschenden die Rolle der Subkurator*innen zu. Diese können dann ihre Artefakte in RADAR-BB hochladen, die nach einer Prüfung durch das FDM-Team des SFBs zur Veröffentlichung oder Archivierung freigegeben werden (Workflow „Data Steward“, Abbildung 3).

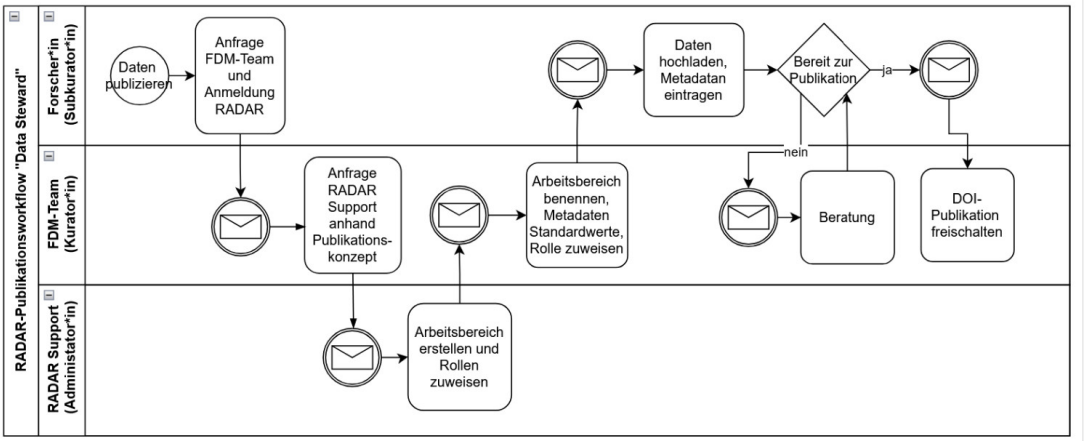


Abbildung 3: Workflow „Data Steward“.

4.5 Workflows und Blue-Prints für die über-institutionelle Nutzung

Sowohl der in Abbildung 2 gezeigte Workflow „Scientist“ als auch der vom SFB beschriebene Workflow „Data Steward“ (Abbildung 3) ermöglichen Feedback- und Korrekturschleifen zwischen allen Ebenen, die in der Darstellung nicht separat aufgenommen sind. Dies gilt ebenso für weitere mögliche Betriebsszenarien und Workflows. Je nach individueller Ressourcenlage der beteiligten Hochschule und Bedarfe sowie Möglichkeiten der Datenlieferanten bzw. Forschenden können die Aufgaben und Rollen verteilt werden. Beratungs- und Betreuungsdichte sind dementsprechend anzupassen. Angesiedelt sein kann die Betreuung der jeweiligen RADAR-Umgebung (d. h. der Radar-Support an der Hochschule) bei den FDM-Verantwortlichen einer Institution, der Bibliothek oder auch der IT-Stelle. Die Zuordnung kann flexibel je nach spezifischer universitätsinterner Struktur erfolgen, da es nur seitens FIZ Karlsruhe und der betreibenden Hochschule (hier Universität Potsdam) über IT-Know-How für den technischen Support bedarf. Zu beachten ist jedoch, dass sowohl die Archivierung für 10 Jahre nach DFG-Kodex als auch die Publikation für mindestens 25 Jahre es erfordern, dass die über diese langen Zeiträume hinweg genutzten Kontakt-E-Mail-Adressen gleichbleibend sind (beispielsweise durch Funktionsadressen). Die hohe Flexibilität des Systems bei gleichzeitig stabilem Rechte- und Rollenmanagement ermöglichen es, die hier gezeigten Workflows als Blue-Prints im Verbund einzusetzen und darüber hinaus für andere Institutionen zu übernehmen.

5 Herausforderung und Lessons Learned

Der Umsetzungsprozess ist weniger von technischen als von organisatorischen Herausforderungen begleitet. So muss die Datenschutzkonformität über mehrere Hochschulen hinweg garantiert sein. Und auch der Speicherbedarf kann nur geschätzt werden, ebenso wie der Betreuungsbedarf, der – wie dargelegt – individuell verschieden sein kann, aber stets zumindest einen betreuenden RADAR-Support per Hochschule erfordert. Eine weitere Herausforderung ist der Zeithorizont, da es sich hierbei nicht um kurzfristige Projekte handelt, sondern um Infrastrukturen, die langfristig vorhanden und betreut werden müssen. Sowohl technische Infrastruktur, als auch Finanzierung und Personal müssen dabei mitgedacht werden. In Brandenburg wird eine zentrale Finanzierung durch das Bundesland angestrebt. Die Bereitstellung der Software durch FIZ Karlsruhe erfolgt auf vertraglicher Grundlage und das System wird vom Zentrum für Informationstechnologie und Medienmanagement (ZIM) der Universität Potsdam 1st-Level technisch betreut. Die 2nd-Level-Betreuung übernimmt, ebenfalls vertraglich fixiert, FIZ Karlsruhe. Eine Zusammenarbeit oder zumindest ein fachlicher Austausch mit den Hochschularchiven wären anzudenken, da diese durch die Umstellung auf elektronische Verwaltung vor ähnlichen Herausforderungen stehen.

6 Resümee

Die acht staatlichen Hochschulen Brandenburgs und FIZ Karlsruhe haben sich zusammengetan, um eine gemeinsame, lokal betriebene Repositoriumsstruktur zu schaffen. Technischer 1st-Level-Support, IT- und Speicherinfrastruktur des brandenburgweiten Repositoriums RADAR-BB liegen zentral an der Universität Potsdam. Diese als Shared Service bezeichnete Struktur ermöglicht auch kleineren Hochschulen die Nutzung und den Betrieb technischer Dienste, die sonst die hochschuleigenen Ressourcen übersteigen würden. Wenngleich konkrete Betriebsszenarien und Workflows von individuellen Ressourcen abhängen, kann dennoch ein qualitativ weitestgehend gleichwertiger Dienst an den am Verbund beteiligten Hochschulen angeboten werden. Dieser erfüllt datenschutzkonform die Forderung der DFG nach 10-jähriger sicherer Speicherung der Forschungsdaten und bietet außerdem eine sichere Publikationsplattform, die Datenpakete und Metadaten mit DOI versehen und nachhaltig auffindbar sowie zugänglich halten.

Das Repositorium RADAR-BB präsentiert sich im gemeinsamen Design des Verbundprojekts. Zugleich werden alle Forschungsdatenpublikationen mit dem individuellen Logo der Hochschule ausgezeichnet und so deren individueller Beitrag zur Forschungslandschaft deutlich. Die Hoheit darüber, welche Forschungsdaten publiziert bzw. archiviert werden, bleibt dabei dank separater und individuell verwalteter Arbeitsumgebungen bei den einzelnen Hochschulen.

Ein im Verbund betriebenes Repositorium erfüllt die Wünsche der Forschenden und die Forderungen der DFG nach lokaler Archivierung auch nicht publizierter Rohdaten. Zugleich überzeugt es durch Verteilung der Ressourcen langfristig mit einer effektiven Kosten-Nutzen-Rechnung, da nicht jede Hochschule individuell die Verantwortung und den Aufwand zu schultern hat. Sorgfältig geplante Organisation und Workflows erleichtern dabei die Umsetzung.

Danksagungen

Wir danken den Teilnehmenden an Schulungen und Weiterbildungen für ihr Feedback. Das Projekt wird seit 2022 vom Bund (BMBF) und vom Land Brandenburg (MWFK) mit 2,44 Millionen Euro finanziert, um in den Auf- und Ausbau von nachhaltigen Strukturen im Forschungsdatenmanagement des Landes Brandenburg zu investieren. Christian Riedel wurde im Rahmen des DFG-geförderten Sonderforschungsbereichs 1294, Projekt Z03, gefördert.

Beiträge

Konzeptualisierung und Idee des Beitrags stammen von Maria Büttner, Kerstin Soltau und Boris Jacob. Vortragende im Panel waren Daniela Mertzen, Kerstin Soltau und Christian Riedel. Alle Autoren waren an der Vorbereitung, Erarbeitung und Überarbeitung des Textes beteiligt.

Interessenkonflikte

Die Autor*innen unterliegen keinen Interessenkonflikten.

Literaturverzeichnis

Deutsche Forschungsgemeinschaft. 2025. *Guidelines for Safeguarding Good Research Practice. Code of Conduct.* Zenodo. https://doi.org/10.5281/ZENODO.3923601.

Hartmann, Niklas K., Boris Jacob und Nadin Weiß. 2019. *RISE-DE – Referenzmodell für Strategieprozesse im institutionellen Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.3585556.

Heuveline, Vincent, Nina Bisheh und Philipp Kling, Hrsg. 2023. *E-Science-Tage 2023.* Heidelberg: heiBOOKS. ISBN: 978-3-948083-91-5. https://doi.org/10.11588/HEIBOOKS.1288.

Jacob, Boris, Janna Kienbaum, Kerstin Soltau, Jochen Klar und Janine Straka. 2024. *Landesweite Bereitstellung kooperativer Forschungsdaten-Dienste.* Poster. Besucht am 20. März 2025. https://nbn-resolving.org/urn:nbn:de:0290-opus4-188950.

Landwehr, Matthias, Gabriel Schneider, Stefan Hofmann, Matthias Razum und Kerstin Soltau. 2022. „Lokal betrieben, remote gepflegt – Software für ein Datenrepositorium in Kooperation implementieren". In *E-Science-Tage 2021: Share Your Research Data,* herausgegeben von Vincent Heuveline und Nina Bisheh, 44–53. heiBOOKS. https://doi.org/10.11588/HEIBOOKS.979.C13717.

Mertzen, Daniela, Heike Neuroth, Carsten Schneemann und Kathrin Woywod. 2023. „Zertifikatskurs „Forschungsdatenmanagement für Studierende" der Landesinitiative für Forschungsdatenmanagement in Brandenburg", https://doi.org/10.18420/INF2023_08.

Ministerium für Wissenschaft, Forschung und Kultur des Landes Brandenburg. 2022. *Forschungsdatenstrategie für das Land Brandenburg.* Besucht am 20. März 2025. https://mwfk.brandenburg.de/sixcms/media.php/9/MWFK_Forschungsdatenstrategie_Barrierefrei.pdf.

Universität Potsdam, Präsidium. 2020. „Forschungsdatenstrategie 2019-2022“. *Universitätsverlag Potsdam,* https://doi.org/10.25932/PUBLISHUP-44436.

Wedlich-Zachodin, Kerstin Vanessa, Kerstin Soltau, Elisa Rodenburg und Mari Elisa Kusniemi. 2024. *Research Data Management Working Group Webinar – Data Management Plans: Use and Reuse (Slides).* Zenodo. https://doi.org/10.5281/ZENODO.10649084.

Beitrag zur Leistungsmessung bei der Evaluation einer mit RISE-DE an einer Hochschule für angewandte Wissenschaften erstellten Forschungsdatenstrategie

Ian Wolff [1,*], Claus Spiecker [2], Daniela Mertzen [3]

[1] Hochschule für nachhaltige Entwicklung Eberswalde, Eberswalde, Deutschland
[2] Fachhochschule Potsdam, Potsdam, Deutschland
[3] Universität Potsdam, Potsdam, Deutschland
[*] Kontakt: ian.wolff@hnee.de

Der Artikel enthält einen Vorschlag zur Evaluation einer mit RISE-DE Referenzmodell erstellten Forschungsdatenstrategie. In RISE-DE ist das „Handlungsfeld 9 Leistungsmessung“ zwar angedacht, aber ohne eine weitere Darstellung, wie diese Evaluation in der Praxis durchgeführt werden könnte. Dieses Defizit soll nun durch einen konkreten Vorschlag zur Vorgehensweise ausgeglichen werden. Nach einer Einordnung von Evaluationen im Rahmen einer Theorie zum Management, wird darauf eingegangen, dass eine Leistungsmessung im öffentlichen Dienst andere Kriterien als in Wirtschaftsunternehmen erfordert, da keine Gewinnmaximierung im Mittelpunkt steht. Der Fokus der Evaluation wird eher darauf gelegt, dass öffentliche Leistungen möglichst wirtschaftlich, effizient und wirkungsvoll erbracht werden. Orientiert an der „Handreichung zur Entwicklung von Methoden für die Evaluation von Anpassungsstrategien“ von Kaiser und Kind (2019) zur Evaluation einer Strategie wird eine Vorgehensweise vorgestellt und anhand der Forschungsdatenstrategie der HNEE am Thema Schulungen beispielhaft erläutert. Die betrachtete Forschungsdatenstrategie wurde 2023 im Rahmen der Landesinitiative Forschungsdatenmanagement in Brandenburg (FDM-BB) erstellt. Durch die Darstellung wird deutlich gemacht, dass die Anwendung der vorgestellten Methode konkrete Erkenntnisse, insbesondere durch Einbeziehung einer Umfeldanalyse und unabhängig der relativ kurzen Laufzeit der Strategie, ermöglicht. Die vorgeschlagene Methode zur Evaluation ließe sich durch leichte Modifikationen bei den Evaluationsfragen und Bewertungsrubriken auch auf Forschungsdatenstrategien, die nicht mit RISE-DE erstellt wurden, anwenden.

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23934 (CC BY-SA 4.0).

Keywords: Forschungsdatenstrategie, RISE-DE, Wirkungsanalyse, Leistungsmessung, Hochschule für angewandte Wissenschaften

1 Einleitung

Ein wichtiger Teil bei der Institutionalisierung des Forschungsdatenmanagements (FDM) ist eine sorgfältige Planung, die in Form der Erstellung einer Forschungsdatenstrategie erfolgen kann. In dieser werden unter Berücksichtigung der Ressourcen der eigenen Institution Ziele zur Verwirklichung eines FDMs festgelegt. Aufgrund sich beständig verändernder äußerer Rahmenbedingungen muss diese interne institutionelle Planung eine regelmäßige Überprüfung und gegebenenfalls Anpassung erfahren.

Im Artikel wird ein Vorschlag zur Evaluation einer mit dem RISE-DE Referenzmodell erstellten Forschungsdatenstrategie vorgestellt und anhand der Forschungsdatenstrategie der HNEE am Themenfeld Schulungen beispielhaft erläutert. Hierzu findet zunächst eine Einordnung von Forschungsdatenstrategien vor dem Hintergrund zu einer Theorie zum Management statt. Nach einer kurzen Darstellung bisher in Deutschland verwendeter Evaluationsmodelle, erfolgt eine Einordnung einer Wirkungs- und Leistungsmessung bei der Evaluation einer Forschungsdatenstrategie. Als Beispiel dient die Forschungsdatenstrategie der HNEE, die 2023 im Rahmen eines Projekts der Landesinitiative Forschungsdatenmanagement in Brandenburg (FDM-BB) 2023 erstellt wurde. Zur Orientierung wird das RISE-DE Referenzmodell kurz vorgestellt, ehe im Hauptteil der Arbeit ein Vorschlag zur Ausgestaltung einer Wirkungsanalyse im Rahmen und auf Basis von RISE-DE vorgestellt wird, der an das Modell zur Evaluation von Kaiser und Kind (2019) angelehnt ist. Die Anwendung der so entwickelten Wirkungsanalyse auf die Forschungsdatenstrategie der HNEE wird am Thema Schulungen beispielhaft erläutert. Im abschließenden Fazit werden die Erkenntnisse zusammengefasst.

2 Einordnung

2.1 Strategien im FDM und Leistungsmessung im öffentlichen Dienst

In der Theorie zum Management wird zwischen drei Dimensionen unterschieden, dem normativen, strategischen und operativen Management. Das normative Management befasst sich mit den grundlegenden Werten, Zielen und Normen einer Institution. Es legt den ethischen und langfristigen Orientierungsrahmen fest, innerhalb dessen strategische und operative Entscheidungen getroffen werden (Kamis und Tribler 2022, S. 13-19). Dieser

Dimension kommen überinstitutionelle wissenschafts- und förderpolitische Empfehlungen und Positionspapiere am nächsten. Auf institutioneller Ebene sind es die Forschungsdatenpolicies, die den langfristigen Wert von Forschungsdaten bewahren und ein nachhaltiges Forschungsdatenmanagement (FDM) an Hochschulen verankern (Hiemenz und Kuberek 2019, S. 2). Die Entscheidungen auf strategischer und operativer Ebene müssen im Einklang mit den normativen Vorgaben stehen.

Eine Forschungsdatenstrategie dient der Umsetzung der Forschungsdatenpolicy, d. h. der festgelegten Richtlinien im Umgang mit Forschungsdaten. Sie legt einen langfristigen Handlungsplan fest, um klare Ziele mit konkreten Maßnahmen in einem bestimmten Zeitraum zu erreichen und vorhandene Ressourcen effizient einzusetzen. Sie schafft dabei Klarheit über die Möglichkeiten einer Institution, gibt eine Einordnung der gesamten Thematik und sichert ein gezieltes nachhaltiges Wachstum. Im Bereich des Forschungsdatenmanagements gibt es mehrere Evaluationsmodelle zur Erstellung von Strategien. Das Digital Curation Center (DDC) lieferte bereits 2017 das Research Infrastructure Self Evaluation Framework (RISE)[1], das 2019 eine Anpassung für das deutsche Wissenschaftssystem mit dem RISE-DE Referenzmodell erfuhr (Rans und Whyte 2017; Hartmann, Jacob und Weiß 2019). In Deutschland folgte 2020 das DIAMANT-Modell mit einer stärkeren Orientierung auf Prozessentwicklung (Gerhards u. a. 2020).

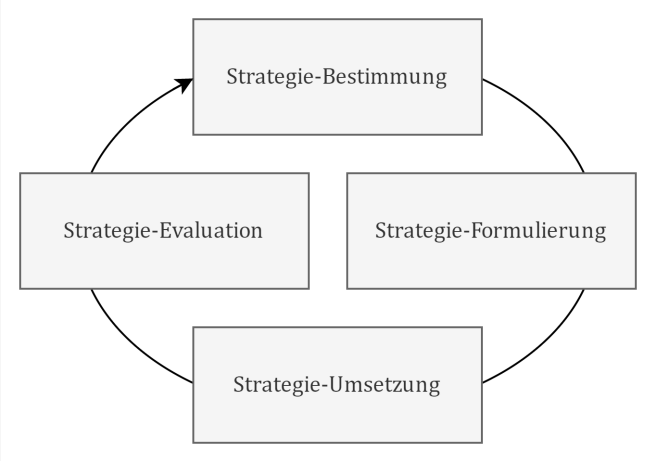


Abbildung 1: Schritte im Strategieprozess (Kamis und Tribler 2022, S. 175).

Sobald eine Strategie entwickelt und verabschiedet ist, folgt die Umsetzung verbunden mit einem entsprechenden Controlling und darauf dann die Evaluation (vgl. Abbildung 1). Im RISE-DE Referenzmodell wird die Leistungsmessung zwar bereits erwähnt, dafür aber noch keine Methode zur Verfügung gestellt. Es wurde von den Institutionen in Deutschland die eine Forschungsdatenstrategie besitzen bisher auch noch keine Evaluation durchgeführt (Ariza de Schellenberger u. a. 2023).[2] Grundsätzlich erfordert eine Leistungsmes-

1 https://www.dcc.ac.uk/guidance/how-guides/RISE; *Besucht am 17. März 2025.*
2 https://www.forschungsdaten.org/index.php/Forschungsdaten-Strategien; *Besucht am 17. März 2025.*

sung im öffentlichen Dienst andere Kriterien als in Wirtschaftsunternehmen, da hier nicht die Gewinnmaximierung im Mittelpunkt steht. Stattdessen kann der Fokus darauf liegen, öffentliche Leistungen möglichst wirtschaftlich, effizient und wirkungsvoll zu erbringen (Mettler und Rohner 2009). Beispiele finden sich hier im Bereich des Bibliothekswesens (Düren und Lipka 2019), in dem Leistungsmessung bereits länger eine Rolle spielt (Goodall 1988; Poll und Boekhorst 2007) oder auch dort, wo Konzepte aus der freien Wirtschaft auf bibliothekarische Rahmenbedingungen übertragen wurden (Schramm 2019).

2.2 Rahmenbedingungen

Die Landesinitiative zum Forschungsdatenmanagement im Bundesland Brandenburg (FDM-BB)[3] befasst sich im Drittmittelprojekt „Institutionalisiertes und nachhaltiges Forschungsdatenmanagement in Brandenburg“ (IN-FDM-BB) mit dem Aufbau eines landesweiten Forschungsdatenmanagements an allen staatlichen forschenden Hochschulen Brandenburgs. Eine wichtige Komponente in den Arbeitspaketen des Projekts bei der Institutionalisierung des FDMs ist die Entwicklung und Evaluation einer Forschungsdatenstrategie an jeder Hochschule (IN-FDM-BB/FHP und IN-FDM-BB/UP 2023). Die Universität Potsdam hatte bereits im Rahmen des Projekts FDMentor[4] im Zeitraum zwischen Juni 2018 und September 2019 eine Forschungsdatenstrategie für den Zeitraum 2019 bis 2022 erstellt. Aus diesem Prozess ging das RISE-DE Referenzmodell als theoretischer Grundrahmen zum Erstellen von Strategien hervor (Jacob, Weiß und Hartmann 2019). Daran orientierte sich die HNEE und erstellte von März 2023 bis September 2023 eine Strategie für den Zeitraum von 2023 bis 2025 (Wolff und Hochschule Für Nachhaltige Entwicklung Eberswalde 2023).

2.3 Das RISE-DE Referenzmodell

Das RISE-DE Referenzmodell ist ein konzeptionelles Framework, das Hochschulen und Forschungseinrichtungen bei der Entwicklung und Umsetzung von Strategien für das institutionelle Forschungsdatenmanagement (FDM) unterstützt (Hartmann, Jacob und Weiß 2019). Es stellt einen Bewertungsrahmen für die Selbsteinschätzung und Zieldefinition bereit und fungiert als Instrument zur strukturierten, stakeholderorientierten Strategieentwicklung im Bereich des FDM. Es werden 8 Themenfelder mit 25 Themen für eine Evaluation vorgegeben. Je Thema erfolgt dann eine Einschätzung des aktuellen bzw. angestrebten Zielzustandes in den Stufen 0-3 (0 = „keine (nennenswerte) Aktivität“ und 3 = „herausragende Aktivität“). Als Themenfeld 9, ist im Modell unter der Rubrik Ergebnisse

3 https://fdm-bb.de/ und https://www.forschungsdaten.org/index.php/FDM-BB#Arbeitspakete; *Besucht am 18. März 2025.*

4 https://www.forschungsdaten.org/index.php/FDMentor; *Besucht am 18. März 2025.*

eine Leistungsmessung optional vorgesehen. Diese wird jedoch weder definiert noch weiter ausgeführt und erscheint nur in einer Übersichtsgrafik.

3 Ein Modell zur Leistungs- und Wirkungsanalyse für RISE-DE

Ein bedeutender Teil des strategischen Managements ist, die Ergebnisse des eigenen Handelns sicherzustellen. Da Strategien eine grundlegende Richtung vorgeben und so den Erfolg einer Institution mitbestimmen, ist deren Überprüfung von Wichtigkeit, insbesondere in Hinblick auf mögliche notwendige Veränderungen der zukünftigen Strategien. So werden bei der Evaluation einer Strategie und der damit einhergehenden Leistungs- und Wirkungsanalyse insbesondere die folgenden Faktoren betrachtet (Kamis und Tribler 2022, S. 175-176; Umlauf 2022):

- **Zielüberprüfung**: das erfolgreiche Erreichen festgelegter Handlungsziele.
- **Leistungsindikatoren**: Festlegung identifizierbarer Kenngrößen.
- **Effizienz**: Umfang der Wirkung eingesetzten Ressourcen.
- **Externe Einflüsse**: Einflussfaktoren auf die behandelten Themen mit einem ganzheitlichen Blick.
- **Anpassungsfähigkeit und langfristige Auswirkung**: Flexibilität einer Strategie auf neue Herausforderungen zu reagieren bei gleichbleibenden längerfristigen Effekten.

Im RISE-DE Referenzmodell werden Ziele und die damit angestrebten Erfolge der Strategie sehr stark innerhalb eines Themas bzw. eines Themenfeldes definiert und festgelegt. Die Auswirkungen des Erreichens eines Ziels in einem Themenfeld auf andere Themenfelder bzw. auf eine ganze Zielgruppe werden dabei nur bedingt oder gar nicht berücksichtigt. Bei einer bereits bestehenden Strategie wird so die Wirksamkeit einzelner Themen auf das FDM in seiner Gesamtheit nicht vollständig erfasst. Im RISE-Modell ist zwar eine optionale Leistungsmessung vorgesehen, diese wird aber nicht weiter ausgeführt. Im Folgenden wird ein Vorschlag für ein Modell der Leistungs- und Wirkungsanalyse auf der Basis des RISE-DE Referenzmodells gemacht. Diese Ausgestaltung von „Themenfeld 9: Wirkungsanalyse“ orientiert sich an der „Handreichung zur Entwicklung von Methoden für die Evaluation von Anpassungsstrategien“ des Umweltbundesamts (Kaiser und Kind 2019).

3.1 Vorüberlegungen

Die folgenden Vorüberlegungen dienen dazu, einen passenden Rahmen für eine erfolgreiche Evaluation zu definieren und alle wichtigen Aspekte zu berücksichtigen. Dies umfasst das Festlegen von Zielen, die Berücksichtigung der vorhandenen Ressourcen, die Bestimmung von Evaluationsverantwortlichen, die einzubindenden Stakeholder*innen und die Festlegung, wie mit den Ergebnissen umgegangen werden soll.

Ziele festlegen: Bei der Evaluation einer Forschungsdatenstrategie kann von einem Szenario ausgegangen werden, in dem Prozesse und Maßnahmen einer bereits bestehenden Strategie evaluiert werden mit dem Ziel, die Ergebnisse für eine neue, vielleicht längerfristige Strategie zu nutzen bzw. den Erfolg der Strategie zu bestimmen. Die Anpassung von laufenden Prozessen wäre ein Ziel – verbunden mit dem Vorteil, aus Fehlern zu lernen und Abläufe zur Zielerfüllung zu optimieren. Eine Herausforderung könnte sein, dass bei einer Evaluation eventuelle Defizite nicht ausreichend dargestellt werden können, sodass Ergebnisse nur eingeschränkt für eine Verbesserung genutzt werden können.

Umfang der Evaluation passend zu den Ressourcen gestalten: Es muss darauf geachtet werden, dass personelle und zeitliche Ressourcen je nach Institution individuell zur Verfügung stehen und gemäß dem geplanten Umfang einer Evaluation einsetzbar sind.

Wer führt die Evaluation durch? Grundsätzlich bietet es sich an, dass die Evaluation der Strategie durch das an einer Einrichtung für das FDM zuständige Personal durchgeführt wird, da dieses über eine entsprechende fachliche Expertise verfügt. Eine Evaluation kann auch durch eine andere Einheit der Einrichtung oder auf einer übergeordneten Stufe in der Hierarchie durchgeführt werden. Die Vorteile bestehen in einer höheren Objektivität bzw. darin, dass Ergebnisse leichter umgesetzt werden können.

Welche Akteur*innen sollen eingebunden werden? Je nach Stufe des RISE-DE Referenzmodells sind bereits bei der Erstellung der Strategie unterschiedlich viele Akteure miteinbezogen worden, bzw. wurde die Wirksamkeit der Strategie durch entsprechende Gremien legitimiert (Wolff und Spiecker 2024). Somit ist zu überlegen, ob diese Personen bei der Evaluation auch miteinbezogen werden. Weiterhin ist je nach Thema eine Beteiligung der Hochschulleitung entsprechend zu berücksichtigen. Im Besonderen gilt das auch für die Personen, die an der Umsetzung der bisherigen Forschungsdatenstrategie aktiv beteiligt sind.

Umgang mit den Ergebnissen festlegen: Die Wirksamkeit des Ergebnisberichts ist u. a. von seiner Reichweite abhängig. So sind beim Umgang mit den Ergebnissen folgende Aspekte zu beachten: Adressatenkreis, interne Behandlung oder Veröffentlichung und Form der Veröffentlichung. Gerade bei der Form ist zwischen einem längeren wissenschaftlichen Bericht und einer kürzeren Zusammenfassung samt Handlungsempfehlungen für Entscheidungsträger zu unterscheiden.

3.2 Schritte der Evaluation

Die Handreichung von Kaiser und Kind (2019) definiert acht Schritte zur Evaluation einer Strategie. Unter Berücksichtigung des RISE-DE Referenzmodells sind Schritt 1 „Den Evaluationsgegenstand eingrenzen“ und Schritt 4 „Evaluationsfragen in Haupt- und Teilkriterien unterteilen und Bewertungsrubriken festlegen“ bereits festgelegt, da diese durch die RISE-DE Themen bereits im Detail vorgegeben werden. Im Folgenden werden die acht Schritte vorgestellt.

Schritt 1: Den Evaluationsgegenstand eingrenzen

Es ist sinnvoll, den Inhalt der Evaluation konkret einzugrenzen. In diesem Falle handelt es sich um die Evaluation einer mit Hilfe von RISE-DE erstellten Forschungsdatenstrategie.

Schritt 2: Ein Wirkungsmodell entwickeln

Ein Wirkungsmodell veranschaulicht Wirkungszusammenhänge und kann als Hilfestellung dienen, um die Wirkung von Aktivitäten und Prozessen zu eruieren. Es kann für ein gesamtes RISE-DE Themenfeld oder einzelnes Thema entwickelt werden. In Abbildung 2 werden die theoretischen Elemente eines Wirkungsmodells für ein RISE-DE Thema dargestellt.

Ressourceneinsatz und Umsetzung: Dieser Abschnitt umfasst alle erforderlichen Ressourcen zur Realisierung eines Projekts, einer Maßnahme oder einer Strategie, darunter Personal, finanzielle Mittel, Zeit und Material. Falls erforderlich, können Ressourceneinsatz und Umsetzung separat betrachtet werden. Zum Erreichen eines Outputs sind außerdem Maßnahmen oder Meilensteine notwendig, die ebenfalls evaluiert bzw. für die Kennzahlen entwickelt werden können.

Ergebnis (Output): Der Output bezeichnet die konkreten Resultate, erbrachten Leistungen oder produzierten Erzeugnisse einer Maßnahme, eines Projekts oder einer Strategie.

Wirkung (Outcome): Der Outcome beschreibt die kurz- bis mittelfristigen Effekte einer Maßnahme, eines Projekts oder einer Strategie, die in der Regel die Zielgruppe betreffen.

Langfristige Wirkung (Impact): Der Impact umfasst die langfristigen Veränderungen, die durch eine Maßnahme, ein Projekt oder eine Strategie auf Systemebene erzielt werden.

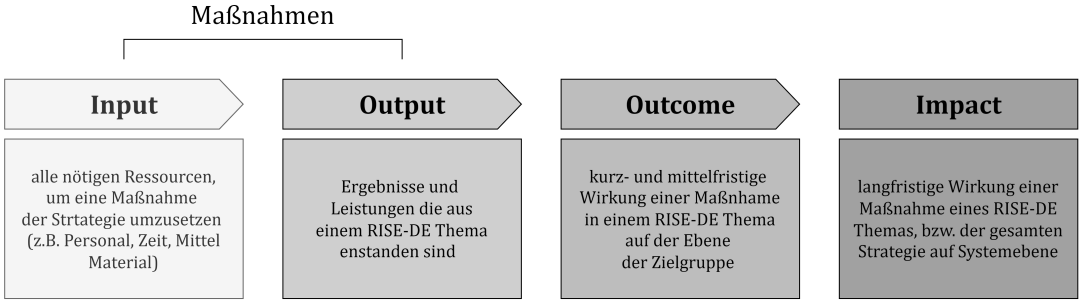


Abbildung 2: Elemente eines Wirkungsmodells eines RISE-DE Themas (angelehnt an Kaiser und Kind 2019).

Schritt 3: Evaluationsfragen entwickeln

Während das Wirkmodell den gesamten Prozess übersichtlich veranschaulicht, dienen Evaluationsfragen dazu, gezielte Schwerpunkte für die Evaluation festzulegen. Sie richten den Fokus auf spezifische Aspekte und ermöglichen eine gezielte Eingrenzung der Evaluation, um den größtmöglichen Nutzen zu erzielen. Das Ziel bei der Formulierung evaluativer Fragestellungen besteht darin, eine überschaubare Anzahl übergeordneter Fragen zu entwickeln, unter denen sich alle relevanten spezifischen Themen und Aspekte einordnen lassen. Dafür kann es erforderlich sein, dass die zentralen Fragen einen höheren Abstraktionsgrad aufweisen. Mitunter ist es auch notwendig, dass nicht nur eine Frage pro RISE-DE Thema, sondern auch für eine einzelne Stufe und bei komplexen Stufen sogar mehrere Fragen entwickelt werden müssen. Zu beachten ist, dass die Fragen keine deskriptive, sondern eine bewertende Antwort erfordern müssen.

Schritt 4: Evaluationsfragen in Haupt- und Teilkriterien unterteilen und Bewertungsrubriken festlegen

Da die Evaluation in erster Linie auf mit RISE-DE erstellte Forschungsdatenstrategien abzielt stellen die 25 Themen von RISE-DE die Basis für den Inhalt der Evaluationsfragen dar. Eine Bewertung erfolgt durch die jeweilige Zuordnung zu einer der klar definierten und beschriebenen Stufe von 0 bis 3. Falls die Forschungsdatenstrategie nicht mit RISE-DE erstellt wurde, müssten die Evaluationsfragen und damit auch die Haupt- und Teilkriterien, d. h. die Themen und Bewertungsrubriken entsprechend angepasst werden.

Schritt 5: Passende Indikatoren und Quellen auswählen

Die passenden Indikatoren werden jeweils für die einzelnen RISE-DE Themen oder feingranularer für eine jeweilige Stufe entwickelt. Die letztendlichen Quellen hängen von den jeweiligen Indikatoren ab und können beispielsweise bereits bestehende Erhebungen oder Berichte sein. Es können aber auch neue Daten erhoben und Personen mit spezifischem Wissen befragt werden. Aus Indikatoren und Quellen kann dann ein Analyseraster abgeleitet werden, das eine Struktur wie in Tabelle 1 aufweisen kann.

Tabelle 1: Mögliche Struktur für ein Analyseraster.

Fragestellung	**RISE-DE Thema**	**Indikator**	**Quelle**
Frage	Thema	Indikator A	Quelle w
			Quelle x
		Indikator B	Quelle y
	Thema	Indikator C	Quelle z

Schritt 6: Daten erheben und ggf. Umfeldanalyse durchführen

Unabhängig davon, ob vorhandene Daten genutzt oder völlig neue Daten erhoben werden, müssen diese zunächst gesichtet und zusammengestellt werden. Dies bildet die Grundlage für die anschließende Auswertung. Gemäß der in Schritt 5 gewählten Quellen werden passende Methoden zur Datenerhebung bestimmt. Soll die Wirkung eines Service hinterfragt werden, bieten sich Interviews oder bei einem kleinen Personenkreis die Verwendung von Fragebögen an. Bereits durchgeführte Bedarfserhebungen zum FDM können ebenfalls für die Evaluation verwendet werden.

Umfeldanalyse bei Serviceleistungen

Bei Serviceleistungen bietet es sich weiterhin an, die eigenen Vorhaben, die sich aus der Zieldefinition mit RISE-DE ergeben, mit einer Umfeldanalyse zu kontrastieren (Kaufmann 2021, S. 9-18). Dies ist gerade für die Anpassung einer Strategie sinnvoll, um den Einsatz der eigenen Ressourcen gezielter vorzunehmen und Handlungen abzuschätzen. Eine Umfeldanalyse verläuft in drei Schritten:

A. Themenfelder mappen: Im ersten Schritt werden die relevanten Themenfelder erfasst und strukturiert, die entscheidende Entwicklungen für den Service einer Einrichtung umfassen. Bei der Identifizierung relevanter Themen gilt es, aus einer Vielzahl unterschiedlicher Strömungen jene herauszufiltern, die das aktuelle Geschäftsmodell (d. h. öffentliche Leistungen möglichst wirtschaftlich, effizient und wirkungsvoll zu erbringen) maßgeblich beeinflussen. Einige Faktoren können sein: ökonomische Faktoren, technologische Faktoren und Trendentwicklungen, politische Faktoren, rechtliche Entwicklungen, institutionelle Entwicklungen und Konkurrenzentwicklungen.

B. Relevante Themen scannen: Im zweiten Schritt werden die als relevant identifizierten Themen vertieft analysiert. Dabei gilt es, folgende Kategorien gründlich zu durchdenken und präzise zu dokumentieren:

1. Was genau verstehen wir unter dem Begriff (konkrete Beschreibung des Themas)?
2. Welche Entwicklung wird erwartet?
3. Wie hoch ist die Wahrscheinlichkeit des Eintritts der erwarteten Entwicklung?
4. Welche Auswirkungen hat diese erwartete Entwicklung auf die Einrichtung/den Service?
5. Welche Chancen und Risiken ergeben sich daraus?
6. Welche Auswirkungen ergeben sich für ein Thema bzw. inwiefern werden sich die Spielregeln in der Branche verändern?
7. Welche strategischen Fragestellungen ergeben sich daraus für die Einrichtung/den Service?
8. Welche Schlüsselmaßnahmen sind zu ergreifen (sofort, mittel- und langfristig)?

C. Handlungsfelder ableiten: Nach einer gründlichen Analyse der als relevant eingestuften Themen stehen bereits die zentralen Maßnahmen zur Bewältigung der Herausforderung fest – sowohl kurzfristige Sofortmaßnahmen als auch mittel- bis langfristige Strategien. Diese müssen nun an die spezifischen Anforderungen angepasst und in die Umsetzung überführt werden.

Schritt 7: Auswerten und bewerten

Die Beurteilung der Daten stellt einen wesentlichen Bestandteil der Evaluation dar. Es sollten mithilfe der festgelegten Indikatoren Bewertungskriterien bzw. Messgrößen definiert werden, anhand derer dann bewertet wird, ab wann ein einzelnes Thema oder eine Maßnahme innerhalb einer Stufe als erfüllt anzusehen ist.

Schritt 8: Die Ergebnisse kommunizieren

Eine Evaluation folgt einem Nutzen und sollte eine zielgruppenspezifische Aufarbeitung erfahren. Weiterhin sollte der Zeitpunkt der Publikation bedacht werden, außerdem ob eventuell ein Workshop organisiert und die Möglichkeit zum Feedback gegeben werden sollte.

4 Beispiel einer Evaluation: Schulungsveranstaltungen an der HNEE

Die HNEE profitiert im Rahmen des Projekts IN-FDM-BB von verbundübergreifenden Leistungen im Bereich Schulungen. Daher soll im Folgenden aufgrund der guten Datenlage exemplarisch ein Blick auf die Schulungsveranstaltungen der HNEE geworfen werden. Im RISE-DE Referenzmodell wird dieser Punkt unter Themenfeld „4 Schulungen“ durch das Thema „4.2 Schulungsveranstaltungen“ repräsentiert (Hartmann, Jacob und Weiß 2019). Die HNEE besaß zum Zeitpunkt der Strategieentwicklung im Juni 2023 noch kein Schulungskonzept und bewertete sich selbst daher mit Stufe 0. Für die Zielfestlegung wurde die Stufe 2 festgehalten, die beinhaltet, dass regelmäßige generische FDM Grundlagenkurse gehalten und nach Bedarf auch fachspezifische Kurse gegeben werden.

Nach der Festlegung des Evaluationsgegenstandes (Schritt 1) auf den Bereich Schulungen, umfasst der zweite Schritt der Evaluation die Entwicklung eines Wirkungsmodells für die Schulungsveranstaltungen. Zum *Input* für diesen Bereich zählen die aufzubringenden Personalressourcen für die Lehrenden, physische verfügbare oder auch online nutzbare Räume für die Schulungen und ggf. Materialien. Der *Output* sind konkrete Schulungsangebote für die jeweiligen Zielgruppen. Der gewünschte *Outcome* wäre ein verbessertes Datenmanagement bei den Zielgruppen und der *Impact* auf Systemebene, dass mehr Forschungsdaten von den Forschenden der HNEE transparenter, nachvollziehbarer und reproduzierbarer zur Publikation kommen.

In den Schritten 3, 4 und 5 wurden eine entsprechende Fragestellung für die Stufe 2 entwickelt, Indikatoren definiert und entsprechende Quellen zur Evaluation ausgewählt.

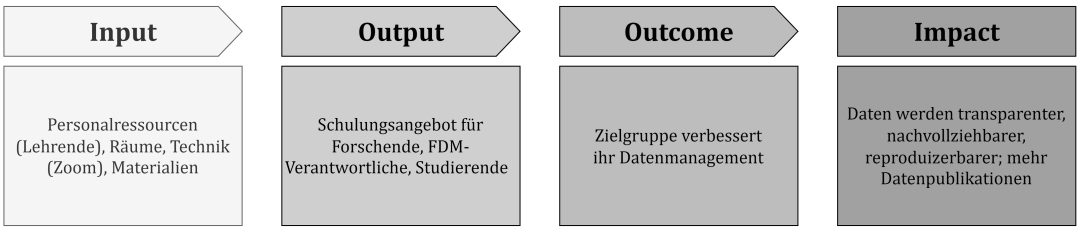


Abbildung 3: Wirkungsmodell für das RISE-DE Thema 4.2 Schulungsveranstaltungen (angelehnt an Kaiser und Kind 2019).

Als Entscheidungskriterium für die Zuordnung zur Stufe 2 ist die Frage zu beantworten: Wurden regelmäßig Kurse durch den FDM Kontaktpunkt gehalten? In Tabelle 2 wird das Analyseraster zum Thema mit den relevanten Indikatoren, Quellen und einigen ersten Ergebnissen gemäß Schritt 6 (Daten erheben) dargestellt.

Da es sich bei den Schulungsveranstaltungen um ein Serviceangebot handelt, bietet es sich weiterhin an, eine *Umfeldanalyse* (optionaler Teil von Schritt 6) durchzuführen, die zusätzlich zum internen Erheben von Kennzahlen, auch eine Einordnung in das Gesamtumfeld bietet. Schulungen befinden sich beispielhaft in einem Gefüge der folgenden Themen:

Politische Vorgaben: Hier liefert die Forschungsdatenstrategie für das Land Brandenburg (Ministerium für Wissenschaft, Forschung und Kultur des Landes Brandenburg 2022) eine Vorgabe, die gute wissenschaftliche Praxis (Deutsche Forschungsgemeinschaft 2019) ist zu beachten und auch die Inhalte der Data-Literacy-Charta sind relevant.[5]

Bedarfe der Teilnehmenden: Bedarfserhebungen können ein Bild liefern, was potenzielle Teilnehmenden an Wissen und Unterstützung benötigen (Wolff 2023). Für potenzielle Zielgruppen sind die Motivation zur Teilnahme an den Kursen oder auch die Anforderungen dafür relevant.

Organisatorische Aspekte: Es stellen sich hier verschiedene Fragen zu personellen Ressourcen, die für Lehre, Erstellung von Materialien (Barrierefreiheit) oder auch Werbemaßnahmen, die beachtet werden müssen. Dies ist besonders im Zusammenhang mit den an der HNEE zur Verfügung stehenden personellen Ressourcen für das FDM relevant, die in den nächsten Jahren nicht mehr als 50 % eines VZÄ betragen werden.

Konkurrenzentwicklung: Entwicklungen von Schulungsangeboten im Rahmen der NFDI oder durch andere Landesinitiativen.

5 https://www.stifterverband.org/charta-data-literacy; *Besucht am 18. März 2025.*

Tabelle 2: Analyseraster zum Thema Schulungsveranstaltungen zur Stufe 2 des RISE-DE Themas 4.2 Schulungsveranstaltungen für die HNEE.

Indikator	**Quelle**	**Ergebnisse HNEE**
Anzahl der regelmäßigen generischen Kurse	Reporting Zertifikatskurs	2
Anzahl der Anmeldungen und Teilnahmen	Reporting Zertifikatskurs	1. Anmeldungen: 10 Teilnahmen: 5 2. Anmeldungen: 10 Teilnahmen: 8
Anzahl der weiteren Veranstaltungen	Reporting der Veranstaltungen	4
Anzahl der Anmeldungen und Teilnahmen	Erhebung zu den Veranstaltungen	29
Wissensstand vor und nach der Veranstaltung (10-Punkte-Skala)	Erhebung zur Veranstaltung	(N=17) Mittelwert Wissensstand vor Veranstaltung: 4,94 Mittelwert Wissensstand nach Veranstaltung: 6,94
	Outcome: Prüfung zur Verbesserung des Datenmanagement von Forschenden	Es könnten einzelne Anfragen per Mail oder Interviews oder ein kleiner Survey durchgeführt werden
	Impact: mehr Datenpublikationen	Publikationsreporting (FIS)

Alleinstellungsmerkmale: Welchen Wert besitzt das eigene Angebot? Ist es bereits akkreditiert, zertifiziert oder in ein Curriculum eingebunden?

Bei der Verortung der Themen wurde bereits teilweise ein Mapping der Thematiken vorgenommen (Schritt 1 der Umfeldanalyse). Diese Themen sollen nun in einem zweiten Schritt gescannt werden. Dazu werden im Schritt 2 der Umfeldanalyse die Fragen des in Kapitel 3.2 zu Schritt 5 (Passende Indikatoren und Quellen auswählen) festgelegten Indikatoren in Tabelle 3 detaillierter betrachtet.

Tabelle 3: Ergebnisse der Umfeldanalyse zum Thema Schulungsveranstaltungen zur Stufe 2 des RISE-DE Themas 4.2 Schulungsveranstaltungen für die HNEE.

Thema	**RISE-DE 4.2 Schulungsveranstaltungen**
Erwartete Entwicklungen	Im Land Brandenburg ist von einem Fortbestehen der Landesinitiative zum FDM auszugehen. Eine Fortführung des Zertifikatsprogramms ist daher wahrscheinlich. Generische Kursangebote anderer Landesinitiativen bestehen bereits, jedoch nur mit partieller Öffnung für Dritte. Schulungsangebote der NFDI bestehen teilweise oder sind im Aufbau.
Eintrittswahrscheinlichkeit	Im Rahmen der Ende 2025 geplanten Fortführung der Landesinitiative FDM-BB: deutschlandweite Kursangebote mit Zertifikat voraussichtlich innerhalb von 2 Jahren oder später. Angebote der NFDI in den kommenden 2 Jahren.
Auswirkung auf unsere Einrichtung	**Chancen:** Die Nutzung der fachspezifischen Angebote der NFDI ist von Vorteil. Im Bereich der generischen Angebote profitiert die HNEE als kleine Hochschule von einem gemeinschaftlich organisierten Angebot im Rahmen der Landesinitiative FDM-BB. **Risiken:** Angebote könnten inhaltlich nicht zu den Anforderungen der HNEE passen. Das gemeinschaftliche Angebot der Landesinitiative könnte aufgrund schwindender Ressourcen eingeschränkt werden.
Auswirkungen auf die Spielregeln im Thema	Unter Umständen könnten mit dem Wegfallen eines brandenburgweiten Angebots und der Abhängigkeit von überregionalen Angeboten die bisher erhobenen Bedarfe der HNEE Forschenden nicht gedeckt werden. Eine überarbeitete Forschungsdatenstrategie für das Land Brandenburg könnte den Fokus auf andere FDM-Themen, weg von den Schulungen, lenken.
Strategische Fragestellung für uns	Die HNEE ist gerade aufgrund geringer Ressourcen im FDM auf externe Angebote angewiesen, bei zugleich weiterhin wachsenden Bedarfen im FDM.
Schlüsselmaßnahme (sofort, mittel-, langfristig)	Weitere Vernetzung innerhalb der Landesinitiative FDM-BB und Befürwortung eines gemeinschaftlichen Schulungsangebots. Weitere Vernetzung und Sichtbarkeit zu HNEE relevanten Themen im fachspezifischen Sektor, z. B. im Rahmen der NFDI.

Schritt 7: Auswerten und bewerten

Aus der Umfeldanalyse ergeben sich einige Handlungsfelder. Im Fall der Schulungsveranstaltungen müssen die politischen Gegebenheiten im Blick behalten werden, gerade hinsichtlich einer überarbeiteten Forschungsdatenstrategie für das Land Brandenburg. Vorerst steht der Erhalt der qualitativ hochwertigen Schulungsangebote des Landes im Vordergrund. Dies zeigen die Anmeldezahlen, aus denen sogar eine höhere Nachfrage hervorgeht, als durch die zur Verfügung stehende Kapazitäten abgedeckt werden kann. Eigene personelle Kapazitäten müssen im Blick behalten werden. Gerade was den Erfolg weiterer Veranstaltungen anbelangt. So wurden vier weitere Veranstaltungen gehalten, die einen Lernerfolg hatten, jedoch eher mäßig besucht wurden. Hier gilt es den Nutzen solcher Formate zu überdenken. Dabei könnte eine Messung des Outcomes und des Effekts auf die Teilnehmenden an diesen Formaten Aufschluss liefern.

Im achten und letzten Schritt werden dann die Ergebnisse der Evaluation kommuniziert. An der HNEE geschieht das in erster Linie intern, im Laufe des Projekts IN-FDM-BB ist aber eine öffentliche Publikation zur befristeten Strategie vorgesehen.[6]

5 Fazit

Dieser Artikel zeigt erste Ansätze auf, wie eine Leistungs- und Wirkungsanalyse anhand einer mit dem RISE-DE Referenzmodell erstellten Forschungsdatenstrategie durchgeführt werden kann. Das vorgeschlagene Modell für die Evaluation füllt so die bestehende Lücke für das in RISE-DE zwar angedachte, aber nicht weiter ausgeführte Handlungsfeld 9. Leistungsmessung. Dabei stand das Ziel im Fokus, eine bestehende Forschungsdatenstrategie zu evaluieren, um eine längerfristige Strategie zu entwickeln. Als Beispiel für die Anwendung der Wirkungsanalyse wurde das RISE-DE Thema „4.2 Schulungsveranstaltungen“ an der HNEE dargestellt. Am Beispiel wurde deutlich, dass eine solche Analyse den Vorteil bietet, den Einfluss von Maßnahmen auf die gewünschten Zielgruppen und auf systemischer Ebene sichtbar zu machen und die Ergebnisse der Evaluation in eine zukünftige Anpassung von Services anhand der Ist- und Zielbestimmung von RISE-DE einfließen zu lassen. Gerade in Verbindung mit einer Umfeldanalyse kann das Gesamtgefüge, in dem sich ein Service befindet, erschlossen und letztendlich Prozesse verdeutlicht werden. Unabhängig von der relativ kurzen Laufzeit von bisher ca. 19 Monaten seit der Verabschiedung der Forschungsdatenstrategie an der HNEE konnten dennoch Ergebnisse erzielt werden. Die vorgestellte Methode zur Evaluation ließe sich durch leichte Modifikationen bei den Evaluationsfragen und Bewertungsrubriken (Schritt 4) auch auf Forschungsdatenstrategien, die nicht mit RISE-DE erstellt wurden, anwenden.

6 https://fdm-bb.de/publikationen-2/; *Besucht am 18. März 2025.*

Danksagungen

Dank gilt dem Bundesministerium für Bildung und Forschung (BMBF) für die Förderung des Projekts IN-FDM-BB (16FDFH126A-E), der Europäische Union (NextGenerationEU), sowie dem Ministerium für Wissenschaft, Forschung und Kultur (MWFK) des Landes Brandenburg.

Beiträge

- Ian Wolff – Hauptautor
- Claus Spiecker – Hauptautor
- Daniela Mertzen – Konzeptionalisierung, Review

Interessenskonflike

Es liegt kein Konflikt vor.

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. 2023. „Anwendung des FDM-Referenzmodells RISE-DE im Verbund“. *Bausteine Forschungsdatenmanagement,* Nr. 2, https://doi.org/10.17192/bfdm.2023.1.8551.

Deutsche Forschungsgemeinschaft. 2019. *Guidelines for Safeguarding Good Research Practice. Code of Conduct.* https://doi.org/10.5281/ZENODO.14281892.

Düren, Petra und Regine Lipka. 2019. *Praxishandbuch Strategische Planung und Controlling in Bibliotheken.* De Gruyter. ISBN: 9783110528176. https://doi.org/10.1515/9783110528176.

IN-FDM-BB/FHP und IN-FDM-BB/UP. 2023. *IN-FDM-BB Förderantrag: Institutionalisiertes und nachhaltiges Forschungsdatenmanagement in Brandenburg.* Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.7737224.

Gerhards, Lea, Marina Lemaire, Stefan Kellendonk und André Förster. 2020. „Das DIAMANT-Modell 2.0“, https://doi.org/10.25353/UBTR-XXXX-F5D2-FFFB.

Goodall, Deborah L. 1988. „Performance measurement: a historical perspective“. *Journal of librarianship* 20 (2): 128–144. https://doi.org/10.1177/096100068802000204.

Hartmann, Niklas K., Boris Jacob und Nadin Weiß. 2019. *RISE-DE – Referenzmodell für Strategieprozesse im institutionellen Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.3585556.

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. https://doi.org/10.14279/DEPOSITONCE-8412. https://depositonce.tu-berlin.de/handle/11303/9354.

Jacob, Boris, Nadin Weiß und Niklas Hartmann. 2019. *FDMentor : Schlussbericht : Teilvorhaben: Strategieentwicklung.* https://doi.org/10.2314/KXP:1686154682.

Kaiser, Theresa und Christian Kind. 2019. *Handreichung zur Entwicklung von Methoden für die Evaluation von Anpassungsstrategien.* Besucht am 27. Juni 2025. https://www.umweltbundesamt.de/publikationen/handreichung-zur-entwicklung-von-methoden-fuer-die.

Kamis, Alcay und Henrik Tribler. 2022. *Workbook Strategisches Management: Die besten Werkzeuge für jede Phase des Strategieprozesses.* Schäffer-Poeschel. ISBN: 97837910-53912. https://doi.org/10.34156/9783791053912.

Kaufmann, Traute. 2021. *Strategiewerkzeuge aus der Praxis: Analyse und Beurteilung der strategischen Ausgangslage.* Springer Berlin Heidelberg. ISBN: 9783662631058. https://doi.org/10.1007/978-3-662-63105-8.

Mettler, Tobias und Peter Rohner. 2009. „Performance Management in Health Care: The Past, The Present, and The Future“. In *Wirtschaftsinformatik Proceedings 2009.* Besucht am 27. Juni 2025. https://aisel.aisnet.org/wi2009/149.

Ministerium für Wissenschaft, Forschung und Kultur des Landes Brandenburg. 2022. *Forschungsdatenstrategie für das Land Brandenburg.* Besucht am 20. März 2025. https://mwfk.brandenburg.de/sixcms/media.php/9/MWFK_Forschungsdatenstrategie_Barrierefrei.pdf.

Poll, Roswitha und Peter te Boekhorst. 2007. *Measuring Quality : Performance Measurement in Libraries 2nd revised edition.* Besucht am 27. Juni 2025. https://repository.ifla.org/handle/20.500.14598/956.

Rans, Jonathan und Angus Whyte. 2017. *Using RISE, the Research Infrastructure Self-Evaluation Framework.* Digital Curation Centre. Besucht am 23. März 2025. http://www.dcc.ac.uk/sites/default/files/documents/publications/UsingRISE_v1_1.pdf.

Schramm, Saskia. 2019. „Konzeptionierung einer Balanced Scorecard für die Commerzbibliothek der Handelskammer Hamburg“. Magisterarbeit, Hochschule für angewandte Wissenschaften Hamburg. Besucht am 27. Juni 2025. http://hdl.handle.net/20.500.12738/8679.

Umlauf, Conrad. 2022. „Strategisches Management“. In *Erfolgreiches Management von Bibliotheken und Informationseinrichtungen: Fachratgeber für die Bibliotheksleitung und Bibliothekare. Grundwerk,* herausgegeben von Cornelia Vonhof und Konrad Umlauf. Dashöfer GmbH. ISBN: 978-3931832469.

Wolff, Ian. 2023. „Auswertung Bedarfserhebung mit daraus folgenden Aktivitäten für die Hochschule für nachhaltige Entwicklung Eberswalde“. In *IN-FDM-BB Werkstattbericht: W 1.2.2 Auswertung Bedarfserhebung mit daraus folgenden Aktivitäten.* Besucht am 27. Juni 2025. https://nbn-resolving.org/urn:nbn:de:kobv:eb1-opus-8266.

Wolff, Ian und Hochschule Für Nachhaltige Entwicklung Eberswalde. 2023. *Forschungsdatenstrategie 2023 – 2025.* https://doi.org/10.57741/OPUS4-677.

Wolff, Ian und Claus Spiecker. 2024. „Entwicklung einer Strategie und Governance für Forschungsdaten“. *Bausteine Forschungsdatenmanagement,* Nr. 2, https://doi.org/10.17192/BFDM.2024.2.8702.

The Impact of Digitalisation on Science: Workflows, Outputs, and Trust

Felix Bach [1], Linda Nierling [2], Christian Bonatto Minella [1], Angelina Sophie Dähms [2]

[1] FIZ Karlsruhe, Leibniz Institute for Information Infrastructure, Karlsruhe, Germany
[2] Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute for Technology, Karlsruhe, Germany

The digital transformation of science is reshaping not only research workflows but also the integrity, openness, and societal trust in scientific knowledge. This paper investigates these developments through the lens of the Leibniz ScienceCampus DiTraRe and its interdisciplinary work on digital research infrastructures. Tracing the historical foundations of digital science – from Leibniz's binary logic to the AI-driven research environments of today – we highlight how shifts in data collection, knowledge organisation, and publication cultures redefine what constitutes scientific evidence and reproducibility.

We examine contemporary challenges and potentials through the use case of the Chemotion Electronic Lab Notebook, demonstrating how domain-specific digital tools can foster transparency, efficiency, and trustworthiness across research lifecycles. A particular focus is placed on the role of artificial intelligence (AI), including generative models and large language models (LLMs), which are increasingly integrated into scientific processes. We discuss emerging practices such as multi-agent LLM collaboration to mitigate hallucinations and the rise of autonomous AI research assistants like Sakana AI's "AI Scientist".

At the same time, the paper addresses the ethical and epistemic challenges posed by algorithmic processes, the impact of digitalisation on public trust, and the institutional responses through guidelines from DFG, UNESCO, and the European Commission. By connecting technical developments with cultural and normative reflections, we argue that building trusted digital research workflows requires a careful balance between innovation and responsibility, supported by interdisciplinary collaboration and continuous governance.

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23935 (CC BY-SA 4.0).

Keywords: Generative AI, Open Science, Research Data, Reliability, Research Integrity

1 Introduction

The Science Campus DiTraRe[1] investigates the digital transformation of research, focusing on how scientific workflows and research outputs have already evolved – and will continue to change – through digitalisation, and what this means for trust in science. DiTraRe examines the effects and potentials of digitalisation in areas such as data collection, knowledge organisation, the use of artificial intelligence (AI), handling of sensitive data, and changing publication cultures.

Research clusters address specific use cases, while cross-cutting dimensions explore overarching issues. This contribution highlights two of these dimensions – Reflections & Resonance and Tools & Processes – alongside the research cluster Smart Data Acquisition, using the "Chemotion Electronic Lab Notebook" as a use case to examine current challenges and opportunities in the digitalisation of science, particularly in light of recent developments in AI.

The article is structured as follows: Chapter 2 outlines a brief history of digitalisation in science. Chapter 3 discusses current AI-related challenges in science, with a focus on public trust and research integrity. Chapter 4 presents the "Smart Labs and AI" use case, exploring the digital transformation of research data practices. The conclusion summarises key developments, highlighting opportunities, challenges, and risks posed by AI in research.

2 History of digitalisation in science

The digitalisation of science is a multi-century journey that, in the context of the DiTraRe framework, is understood not only as a technological progression but also as a transformation of research infrastructures, knowledge practices, and the ways scientific trust and transparency are maintained. It has transformed how knowledge is generated, disseminated, and preserved. From early mechanical innovations to the rise of AI, each phase reflects evolving paradigms in research, communication, and epistemology. Today, with

1 The DiTraRe project, funded by the Leibniz Association and carried out by FIZ Karlsruhe and the Karlsruhe Institute of Technology from 2023 to 2027 (https://www.ditrare.de; *Visited on June 10, 2025*) deals with the influence of digital processes on research the influence of digital processes on research results, both in terms of research methods and in terms of with regard to communication in science and society.

the emergence of interdisciplinary initiatives such as the DiTraRe, this historical evolution is examined through the lens of changing research workflows, research infrastructure, scientific outputs, and the public's trust in science.

2.1 Foundations in Logic and Mechanisation (17th – 19th Century)

The roots of digital science can be traced to the work of Gottfried Wilhelm Leibniz, who introduced the binary number system in his 1703 publication *Explication de l'Arithmétique Binaire* and demonstrated how all numbers could be represented using only the digits 0 and 1 (O'Regan 2021). He also designed a logical calculating machine and envisioned a universal language of logic, which he termed the *calculus ratiocinator*, aiming to formalise human reasoning through symbolic logic (O'Regan 2021).

Born in 1646, Leibniz was not only a mathematician and inventor but also a philosopher who believed that human reasoning could be formalised through symbolic logic. His vision of the *calculus ratiocinator* aimed to express all rational thought in a formal language, anticipating later developments in computation and artificial intelligence (Smith 2007; O'Regan 2021). In this sense, the origins of AI lie not (only) in engineering but in a philosophical project concerned with modelling the structure of human cognition. This conceptual leap laid the groundwork for digital computation.

The 19th century saw further developments with Charles Babbage's Analytical Engine and Ada Lovelace, who is credited with writing the first algorithm, foreseeing a machine capable of manipulating symbols beyond arithmetic (Toole 1998). Meanwhile, Joseph Marie Jacquard's programmable loom (1801) demonstrated early mechanisation of control via punch cards.

2.2 Mathematical Formalism and the Birth of Computing (20th Century)

In 1936, Alan Turing formalised the concept of computation with the Turing machine, introducing a theoretical model of algorithmic processing in his seminal paper *On Computable Numbers, with an Application to the Entscheidungsproblem* (Turing 1936). Turing's work was deeply influenced by foundational developments in mathematical logic, particularly Kurt Gödel's incompleteness theorems (Gödel 1931), which had revealed the inherent limitations of formal axiomatic systems and intensified interest in the mechanisation of reasoning (Davis 2000). This line of thought also paralleled Alonzo Church's formulation of the lambda calculus, which arrived at equivalent conclusions concerning the limits of computation (Church 1936).

Around the same time, Claude Shannon's 1937 master's thesis applied Boolean algebra to switching circuits, laying the groundwork for digital logic – a theoretical insight that would later underpin modern digital circuit design (Shannon 1937, 1948). Together, these developments converged to establish the logical and physical basis of computing machinery, bridging abstract logic and engineering (O'Regan 2021; Ceruzzi 2000).

During World War II, programmable computers emerged, notably *Konrad Zuse's Z3* (1941), regarded as the first fully operational digital computer implementing binary arithmetic and programmability (Ceruzzi 2000; Randell 1973). A foundational advance followed with John von Neumann's *First Draft of a Report on the EDVAC* (1945), which introduced the concept of a *stored-program architecture*, where instructions and data share the same memory space – setting the standard for future computer systems (Neumann 1945; O'Regan 2021).

Networking and Personal Computing (1960s–1980s)

The ARPANET[2] project, initiated in 1969, connected multiple research institutions and later evolved into the modern Internet (Hafner and Lyon 1996). Meanwhile, the development of personal computers (e.g., Altair 8800, Apple I) in the 1970s began democratising computational access.

Digital tools found increasing application in research from the 1970s onward, notably in *early computational biology*, such as the first successful DNA sequencing efforts by Sanger and colleagues (Sanger, Nicklen, and Coulson 1977). At the same time, *expert systems* emerged in the field of artificial intelligence, enabling symbolic reasoning within narrowly defined domains – for example, MYCIN in medical diagnosis – marking a shift toward knowledge-based systems (Feigenbaum 1983; Buchanan and Shortliffe 1984). These developments laid the groundwork for rethinking how scientific data is collected, structured, and validated – an enduring concern for initiatives like *DiTraRe*, which interrogate the evolving dynamics of research in the digital age.

The World Wide Web and Open Science (1990s–2000s)

The launch of the World Wide Web by Tim Berners-Lee fundamentally transformed scientific communication and laid the groundwork for the development of integrated research

2 ARPANET stands for 'Advanced Research Projects Agency Network'. It was developed by the Advanced Research Projects Agency (ARPA) of the U.S. Department of Defense in the late 1960s and is widely considered the precursor to the modern internet. ARPANET was the first network to implement the TCP/IP protocol suite, and it enabled multiple computers to communicate on a single network – a revolutionary concept at the time.

infrastructures – such as those examined in DiTraRe – that enable new forms of collaboration, data sharing, and openness within the scientific community (Berners-Lee and Fischetti 1999). The emergence of online repositories like *arXiv.org*, which pioneered the open-access dissemination of preprints in physics and related fields (Ginsparg 1994), and open knowledge platforms such as *Wikipedia*, which crowdsources encyclopaedic content from global contributors (Lih 2009), further accelerated the global spread of information.

The digitisation of libraries – such as Google Books and the Internet Archive – and the rise of open access publishing have redefined how scientific texts are preserved, discovered, and accessed (Borgman 2007; Kahle 2007; Suber 2012). This shift has been accompanied by growing emphasis on data curation, metadata standards, and transparency in scientific outputs, as reflected in initiatives like the FAIR principles (Wilkinson et al. 2016). These are all central themes in DiTraRe's analysis of digital research workflows and infrastructure development.

2.3 Data-Driven Research and the AI Revolution (2000s–present)

The 21st century has seen a fundamental shift towards data-intensive science, actively explored by initiatives such as DiTraRe (Toole 1998). These initiatives examine both the technical and ethical dimensions of digital research workflows, focusing on how they can improve research quality and public trust. This shift has been driven by advances in cloud computing, big data infrastructures, and reproducibility tools such as Electronic Lab Notebooks (ELNs Toole 1998).

Large-scale projects like CERN's Large Hadron Collider illustrate the enormous scale of data modern science must handle (Evans and Bryant 2008; Bejar Alonso et al. 2020). Jim Gray's "The Fourth Paradigm" (2011) captured this transition, describing a new model of discovery where data exploration complements hypothesis-driven research.

AI and machine learning have since become integral to scientific workflows, reshaping how data is processed, interpreted, and applied across disciplines (Jordan and Mitchell 2015). Within DiTraRe, AI is applied to areas such as semantic data enrichment, metadata curation, automated annotation of research outputs, and predictive modelling (Jordan and Mitchell 2015). Through the use of Large language models like GPT-4 now practices in science are being reshaped with regard to knowledge processing and communication.

Emerging AI research assistants – such as Sakana AI's "AI Scientist" – are pushing the boundaries of autonomous scientific work. According to its creators, this system can generate novel research ideas, write and execute code, conduct simulations, summarise

results, and even draft scientific manuscripts, complete with automated peer review to assess the quality of generated research content (Sakana AI 2024).

In addition, platforms like Elicit, SciSpace, and ResearchRabbit support researchers in literature discovery and comprehension. These tools use large language models to filter relevant publications, summarise complex findings, and visualise citation networks, thereby improving access to and understanding of scientific knowledge (Academia Insider 2024; Elephas 2024).

To address the persistent issue of AI hallucinations, researchers have begun experimenting with multi-agent collaboration techniques. In such frameworks, multiple LLMs are used to evaluate and cross-check each other's outputs, identifying errors and inconsistencies. Key techniques include:

- Adversarial debates, where models challenge one another's responses to promote consensus or correction (Yang et al. 2025),
- Uncertainty-Aware Fusion, which integrates outputs based on model confidence to reduce hallucinated content (Dey, Merugu, and Kaveri 2025), and
- Ensemble models with shared weights, which optimise resource use while improving output reliability (Arteaga, Schön, and Pielawski 2024).

These developments are not merely technical optimisations – they are essential steps toward building trustworthy AI systems in scientific contexts.

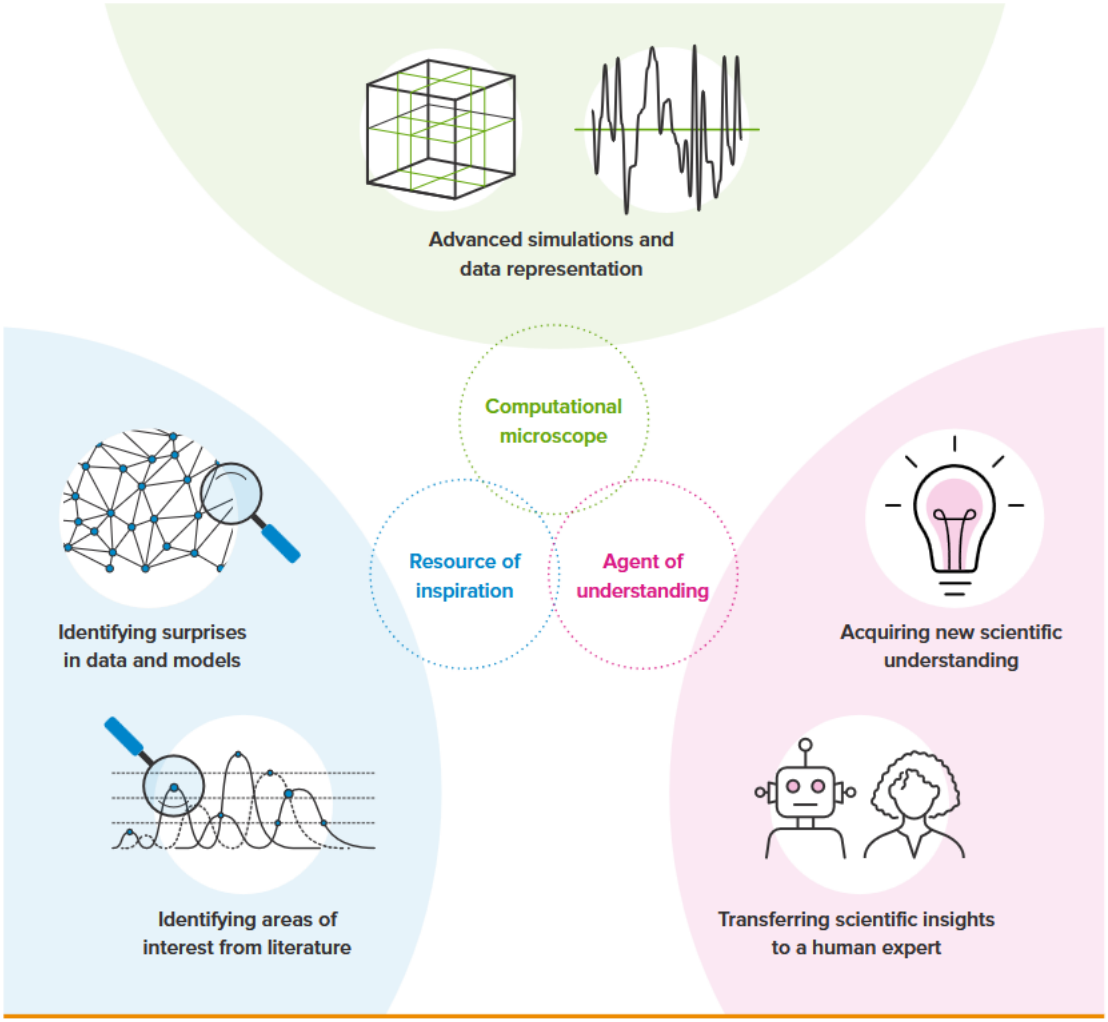
The DiTraRe initiative actively engages with these advances, assessing both their transformative potential and associated risks. Of particular interest are questions of research integrity, transparency, and public trust in a time when AI is increasingly involved in producing and interpreting scientific knowledge.

According to the Royal Society, AI can support scientific understanding in three key ways (The Royal Society 2024, p. 31), shown in Figure 1:

- Computational Microscope: Simulating and visualising data in ways experiments alone cannot.
- Resource of Inspiration: Providing novel, unexpected insights.
- Agent of Understanding: In theory, autonomously generating and communicating knowledge – though this remains a conceptual frontier.

While AI has shown surprising capabilities, it has not yet reached true autonomy in understanding.

Reproduction of a visualisation of the three general roles of AI for scientific research as either a computational microscope, resource of human inspiration, or an agent of understanding[124].



124 The diagram describes three possible ways in which AI can contribute to scientific understanding. The 'computational microscope' refers to the role of AI in providing information through advanced simulation and data representation that cannot be obtained through experimentation. 'Resource of inspiration' refers to scenarios in which AI provides information that expands the scope of human imagination or creativity. The 'agent of understanding' illustrates a scenario in which autonomous AI systems can share insights with human experts by translating observations into new knowledge. As of yet, there is no evidence to suggest that computers can act as true agents of scientific understanding. See: Krenn M. *et al.* 2022. On Scientific Understanding with Artificial Intelligence.

Figure 1: Three possibilities for AI supported scientific understanding (Source: The Royal Society 2024, p. 31).

Generative AI promises increased efficiency in idea generation, literature review, and writing (Albrecht 2023), but it also raises challenges around transparency, bias, and authorship (Fecher et al. 2023; Messeri and Crockett 2024). Multi-model strategies represent one way forward, ensuring that AI becomes not just faster, but also more reliable and responsible in supporting scientific work.

3 AI in science: public trust & research integrity

In light of AI, the digitalisation of research affects not only how science is conducted but also how it is perceived by the public. Factors such as transparency, data openness, and the potential for reproducibility can enhance public trust in scientific processes (Elliott 2017; Floridi and Cowls 2019). However, looking towards AI, the reliance on black-box AI systems and large-scale datasets raises concerns about the reliability and accessibility of scientific findings. Especially as recent development in generative AI challenges workflows, infrastructures and outputs in science (Schreiber and Ohly 2024). On the one hand, generative AI provides new chances for knowledge generation and processing, e.g. by an increased efficiency and productivity in scientific working practices with various possible applications. It further provides potentials for research data, since it offers potentials in training small data sets, and enables LLM-based knowledge extraction from data bases and provides easier programming tools (Digital Science et al. 2023). On the other hand, there are many open questions related to generative AI in science. These range from technical limits and touch the transparency and traceability of scientific results with regard to detection, authentication and the widely cited "hallucinations" (European Research Council 2023). With regard to competences and skills, AI-literacy of researchers comes into play. Further ethical questions of the use of generative AI come up and aspects like overreliance, spread of false information, biased knowledge or source misrepresentation (Jahnel et al. 2025).

The digitalisation of research not only affects how science is conducted but also how it is perceived by the public.

Focusing on generative AI and open data, challenges on two levels come up with regard to trust in science: (1) trust in science from the public as well as (2) research integrity within science. While in general, trust in science remains high, there are four components of trust in scientists which encompass competence, integrity, benevolence and openness (Cologna et al. 2025, p. 714). The components are defined (Mede and Cologna 2025): Competence encompasses the expertise, intelligence, and qualifications required to conduct high-quality research. Integrity is about how honest, ethical, and sincere scientists are. Benevolence addresses how much scientists care about the well-being of others, improving the lives of others, and considering the interests of others. Openness asks how open scientists are to feedback, transparency, and consideration for the view of others. When looking at generative AI, the aspect of integrity as well as openness are of particular interest – while recent developments in the use of generative AI in science may challenge research integrity due to epistemic risks and also other types of misuse, e.g., plagiarism, cheating, generate fictitious data and analyses (Meça and Shkëlzeni 2023). According to Cologna et al. the openness aspect is currently more challenging, since the public perception of the openness of scientific practices is lower than the perceived trust in research integrity. With respect to research data management practices to keep research integrity high are touched in this article. However, there could be further attempts to increase the public

perception of openness in science, e.g., by more communication of results to the public (see Figure 2, Cologna et al. 2025, p. 719). Measures proposed embrace being receptive to feedback, transparent about funding and data sources, communicating about science with the public by encouraging public participation in genuine dialogue, including insights and needs of societal actors.

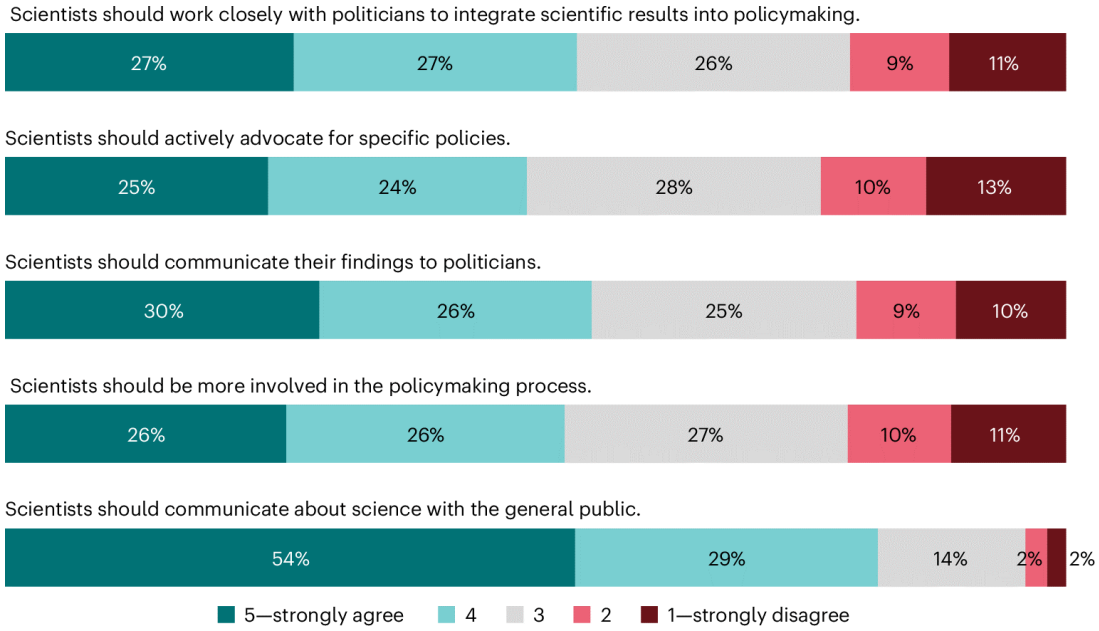


Figure 2: Normative perception of scientists in society and policy making (Cologna et al. 2025).

Open science and open data have traditionally been seen as key to building public trust, as highlighted by Rosman et al. (2022). In their study, participants were asked about the importance of open science practices and whether scientists who adopt these approaches are perceived as more trustworthy. The vast majority considered it important for researchers to make their findings publicly accessible and to follow open science principles. Most participants also reported greater trust in science when materials, data, and code were openly available.

To further strengthen public trust in open data, especially when using AI, several measures can be taken. First, ensuring the reproducibility of AI-based research is essential. This can be supported through reproducibility checklists, standardized data-sharing protocols, and field-specific reproducibility guidelines. Additionally, investing in open repositories is critical. This includes the sharing of datasets, software versions, and workflows, as well as developing context-aware documentation to help adapt AI models to local research settings (The Royal Society 2024, p. 13). Albeit these measures, still modes of engaging and communicating with the public have to developed and explored for the specific context of open data research and practice.

Moreover, in the upcoming use of generative AI in science there might be the risk that research integrity erodes, thus science-internal quality assurance needs to be built up in order to remain high. Up to now, there are several guidelines and statements published with regard to the use of generative AI in science. One widely cited statement is that of the German Research Association (Deutsche Forschungsgemeinschaft 2023). In this statement, the use of generative AI is supported, but binds to the principles of good scientific practice and the transparency and traceability of scientific results. However, the use of models should be thoroughly disclosed and a discursive process including experiences with generative AI is envisaged (Deutsche Forschungsgemeinschaft 2023). Other guidelines focus on the importance of the human factor, "[...] the usage processes should ensure humans' interactive engagement with GenAI and higher-order thinking, as well as human accountability for decisions related to the accuracy of AI-generated content, and their impact on human behaviours" (UNESCO 2023, p. 29) or point out the communicative process of all actors of the science system: "[...] publishers and researchers at all stages of their careers are essential in shaping the discussion on AI and how it can serve the public interest in research" (European Commission 2024, p. 4).

4 Creating Trusted Digital Research Workflows: A Cultural Challenge

The digitalisation of research is transforming not only the tools we use but also the way scientific knowledge is generated, validated, and shared. This transformation, however, is not purely technological – it is profoundly cultural (OECD 2020). In this chapter, we examine how trusted digital research workflows can be developed and sustained. We focus in particular on laboratory practices in chemistry, drawing on insights from the DiTraRe use case *Smart Data Acquisition* and the Electronic Lab Notebook (ELN) *Chemotion*[3]. Through this lens, we explore how digital infrastructures influence everyday scientific work, the nature of research outputs, and the reproducibility of experimental results.

4.1 From Infrastructure to Cultural Practice

The digitalisation of research entails more than the integration of new technologies or software platforms; it necessitates a fundamental transformation of the epistemic cultures that define individual scientific disciplines. These cultures have evolved over time through established methodologies, procedural routines, and disciplinary tools. Each field possesses distinct norms regarding what constitutes valid evidence, how experiments are conducted and recorded, and the frameworks through which data are interpreted. Such

3 http://www.chemotion.net; *Visited on June 10, 2025.*

entrenched disciplinary practices play a critical role in shaping both the adoption and perceived credibility of digital workflows. This influence is clearly illustrated in the field of chemistry, as demonstrated in Figure 3 which compares a laboratory notebook from 1927 with one from today (Herres-Pawlis, Liermann, and Koepler 2020). Despite a century of technological advancement, experimental documentation is still predominantly conducted with pen and paper (Butler 2005). This persistence of analogue practices highlights the extent to which digitalisation in science is entangled with cultural inertia.

Meanwhile, as described in Chapter 2, the evolution of computing technologies has led to powerful tools that enable standardised, computer-assisted workflows. These workflows hold enormous potential to enhance the reliability, efficiency, and reproducibility of research. Yet their implementation remains uneven, and their success depends as much on social acceptance and institutional support as on technical innovation (National Academies of Sciences, Engineering, and Medicine 2022).

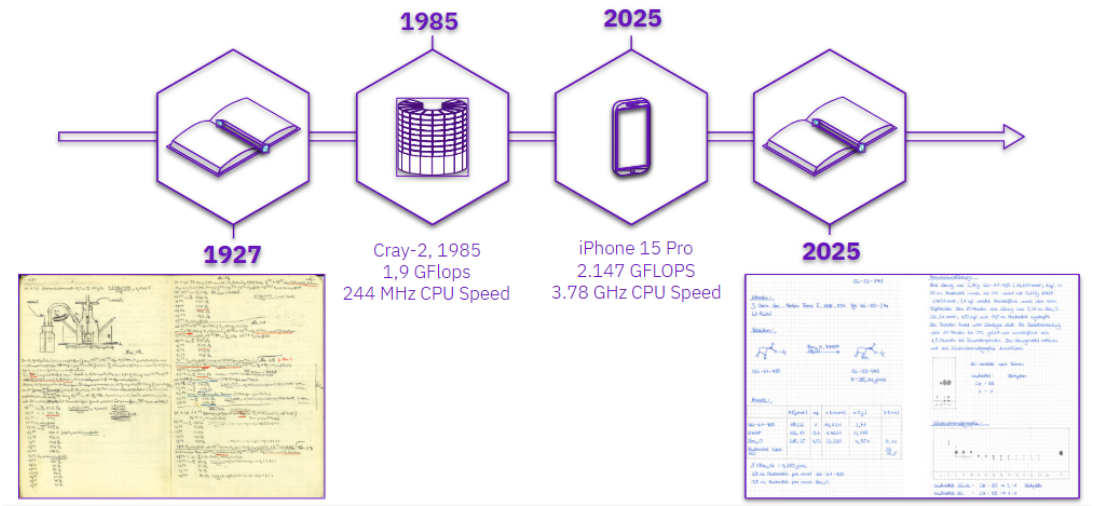


Figure 3: Comparison of the evolution of laboratory notebooks and computers: Despite decades of digitalisation, traditional paper lab notebooks remain widespread (Herres-Pawlis, Liermann, and Koepler 2020).

4.2 Creating Trustworthy Digital Workflows

The question, then, is not simply how to digitalise, but how to do so in a way that researchers can trust and adopt. Initiatives such as NFDI4Chem[4] provide important answers. They demonstrate how digital workflows can support researchers across the entire data life cycle – spanning data acquisition, processing, analysis, publication, and long-term preservation – by adhering to open, interoperable standards.

4 Chemistry Consortium in the National Research Data Infrastructure – http://www.nfdi4chem.de; *Visited on June 10, 2025.*

Central to this approach is the development of user-friendly, domain-specific tools such as Chemotion ELN. Designed specifically for the needs of chemists, Chemotion ELN enables structured and semantically rich documentation of experimental data, helping ensure both interpretability and reproducibility. Rather than replacing researchers' practices, it extends them into digital spaces where data can be preserved, shared, and reused.

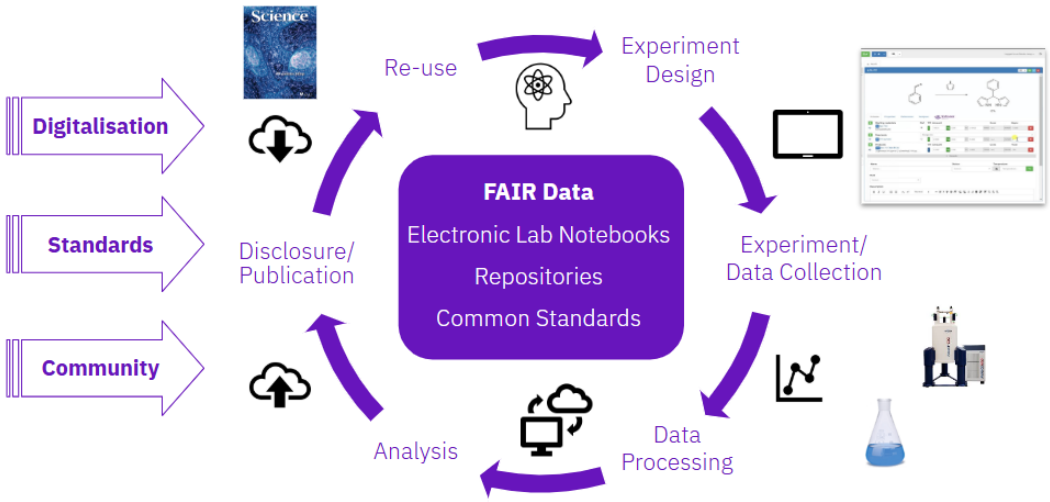


Figure 4: From analogue to FAIR: Digital workflows in chemistry as proposed by NFDI4Chem using Chemotion ELN and FAIR data repositories like RADAR4Chem (Source: NFDI4Chem).

With tools like Chemotion, chemists can record structured data in a semantically rich way, ensuring interpretability and reproducibility. The resulting FAIR data can then be stored in repositories like RADAR4Chem[5], guaranteeing long-term access and global reusability – fuelling a new cycle of data-driven science.

4.3 Chemotion: A Model for Trusted Research Digitalisation

The Chemotion platform exemplifies a comprehensive and integrated approach to digitalisation in the chemistry laboratory. It enables researchers to document experimental workflows, chemical reactions, and results in a structured, machine-readable format. This not only facilitates the systematic organisation of data but also substantially improves reproducibility by reducing the risk of human error and ensuring that experimental conditions are recorded with precision.

Standardisation plays a key role in Chemotion's design. By using consistent data formats and controlled vocabularies, the platform ensures compatibility with other digital research

5 https://radar4chem.radar-service.eu/radar/en/home; *Visited on June 10, 2025.*

tools and repositories, thereby supporting interoperability across systems and institutions. In practice, this means that data entered into Chemotion can be seamlessly exported, analysed, or shared – forming a coherent link in a broader digital research infrastructure.

Another important feature is the automation of data logging. Chemotion can directly capture outputs from laboratory instruments, streamlining the recording process and reducing the administrative burden on researchers. This automation also enhances data accuracy and consistency, since manual transcription steps – where errors frequently occur – are minimised.

Crucially, Chemotion supports collaborative work through cloud-based access. Researchers from different institutions or disciplines can work together in shared digital environments, enabling more efficient knowledge exchange and interdisciplinary experimentation. The platform's design is also aligned with the FAIR principles – ensuring that data is findable, accessible, interoperable, and reusable. This is particularly important for promoting sustainable open science, where datasets remain usable and meaningful over time.

From a compliance perspective, Chemotion offers tangible benefits as well. Its structured documentation makes it easier to fulfil regulatory and laboratory safety requirements (Tremouilhac et al. 2020). The ability to track, search, and verify digital records simplifies audit processes and helps demonstrate adherence to standards of good scientific practice. In sum, Chemotion represents a major step forward in the digitalisation of laboratory research – enhancing not only efficiency and reproducibility, but also transparency and trust.

4.4 Beyond Digitalisation: The Role of AI in Research Workflows

As digital infrastructures mature, AI is becoming an increasingly important actor in shaping how data is curated, interpreted, and reused. Within platforms such as ELN and research data repositories, AI technologies have the potential to augment laboratory workflows in numerous ways. In DiTraRe, we have identified several promising measures in this regard, that are described in the following and that we plan to implement and evaluate during the project.

One of the most immediate benefits lies from our perspective in the automatic extraction and standardisation of metadata. AI systems can analyse raw experimental data and generate relevant metadata entries, reducing the amount of manual input required and improving consistency across records. In tandem, machine learning models can classify and tag chemical reactions or compounds based on their content and contextual relationships, facilitating more efficient data management.

AI also enhances data integrity through predictive data cleaning and anomaly detection, as we believe. By identifying inconsistencies or missing information within datasets, it supports higher standards of data quality. Natural Language Processing (NLP) techniques extend these capabilities further by integrating information from research literature and patents, linking experimental data with existing knowledge and enhancing the contextual understanding of results.

The creation of chemistry-specific knowledge graphs through semantic analysis allows for more intuitive data discovery, helping researchers navigate complex data landscapes and uncover new relationships between concepts, substances, and reactions. AI can also support experimental design by analysing prior reaction data and recommending optimised conditions, thereby reducing reliance on trial-and-error methods.

On the user interface level, AI-powered search functions enable context-aware querying of experimental data, making it easier for researchers to locate relevant datasets or protocols. Finally, AI can assist in summarising results and generating structured reports, offering insights that might otherwise remain hidden in the data.

From our perspective, these capabilities mark a significant shift in how research is conducted and interpreted. AI moves from being a back-end tool to a semi-autonomous partner in the scientific process, suggesting hypotheses, identifying trends, and enabling new modes of data-driven discovery.

4.5 Reflection on the Use Case

Also, with respect to open data, generative AI has the potential to strengthen and transform open research data. It can improve small datasets and metadata quality, but can also bring interactive elements, e.g. by acting as a virtual collaborator (Digital Science et al. 2023).

It seems that also for open research data, new forms of Human-AI-Collaboration might emerge. However, this requires clearly defining the role of human verification in research. But how much should AI handle, and where do we draw the line for human verification? And given the pure amount of data sets, where are the limits of “human-in-the-loop” approaches?

This shift also changes how we organise competences in the field of open research data, since competences and skills change and a specific AI literacy is needed, or in other words, there is a change from human control knowledge to translational knowledge (National Academies of Sciences, Engineering, and Medicine 2022).

In order to govern this change, standards and guidelines are needed. Challenges include the rapid pace of AI developments, which often outstrip the ability of guidelines to keep up with emerging functionalities and applications. On the other hand, the development and implementation of the guidelines raises questions of responsibility. Should research communities lead the way, or should top-down regulations set the framework? And how can a discursive process between the two levels be organised? And lastly, how detailed should guidelines be to remain practical? In order to follow a smooth implementation for a scalable and effective AI integration in research, these key questions have to be tackled by the open data community in a joint approach, also including relevant actors of the science system.

5 Conclusion

The digital transformation of science reaches far beyond the mere adoption of new technologies. As we have shown throughout this contribution, it reshapes fundamental aspects of how research is conducted, how results are documented and shared, and how trust in scientific knowledge is created and maintained – both within science and in its relation to society.

At the heart of this transformation is a cultural challenge. While the technical means for digitalisation are often available, their integration into established scientific workflows requires a shift in epistemic practices and disciplinary habits (The Royal Society 2024). Our exploration of laboratory work in chemistry, through the DiTraRe use case Smart Data Acquisition and the Electronic Lab Notebook Chemotion, has highlighted this vividly: despite a century of technological advancement, paper-based documentation remains widespread, signalling the persistence of analogue routines.

However, tools like Chemotion illustrate how digital infrastructures – if designed in alignment with domain-specific needs – can enable not only standardisation and reproducibility but also a more collaborative and transparent research process. By facilitating FAIR data practices and integrating structured workflows, Chemotion supports both scientific integrity and open science. When combined with repositories such as RADAR4Chem[6] and other repositories for chemistry data recommended by NFDI4Chem[7], it contributes to a sustainable cycle of data-driven discovery.

The inclusion of AI adds a new dimension to this development. AI can assist with metadata curation, semantic annotation, error detection, and even hypothesis generation (Digital Science et al. 2023). These capabilities point toward a shift from manual documentation to machine-augmented understanding. Yet, this shift also raises new questions:

6 https://radar4chem.radar-service.eu/; *Visited on June 10, 2025.*
7 https://www.nfdi4chem.de/repos/; *Visited on June 10, 2025.*

What remains the role of human oversight? How transparent are the processes behind AI-generated outputs? And what standards must be in place to ensure that AI contributes to, rather than undermines, the integrity of research?

These questions also extend to public trust. On the one hand, we argue that the principles of open science – transparency, reproducibility, accessibility – are strengthened through digitalisation and AI. On the other hand, reliance on algorithms and vast datasets can alienate non-specialist audiences and challenge the transparency and integrity of the scientific process (European Research Council 2023). As survey data shows, trust in scientists is generally high, but it is conditional on perceived openness, competence, and accountability. Maintaining and enhancing that trust requires not only technical solutions but also deliberate strategies in science communication and governance. To address these issues, institutional frameworks must keep pace with technological innovation (Cologna et al. 2025). National and international bodies – such as Deutsche Forschungsgemeinschaft (2023), UNESCO (2023), and the European Commission (2024) – are beginning to articulate principles and recommendations for the responsible use of generative AI in research. These efforts are essential, but they must be complemented by bottom-up initiatives within research communities, where disciplinary standards, peer review practices, and ethical guidelines are continuously developed and discussed.

Ultimately, trusted digital research workflows emerge at the intersection of technical infrastructure, cultural change, and institutional reflection. They are not simply adopted; they are cultivated (Jacyszyn et al. 2025). The DiTraRe initiative contributes to this cultivation by fostering interdisciplinary dialogue and by developing exemplary use cases that demonstrate both the potential and the limitations of digitalisation in science.

As science enters the age of AI, we are witnessing not only new tools but new ways of knowledge generation and procession. This calls for shared responsibility of researchers as well as actors of the science system alike: to shape the future of research in a way that is open, transparent, and worthy of trust. Continued exchange and a dedicated guideline on the use of generative AI in research data management practice might be valuable start.

Acknowledgements

This work was supported by the following initiatives: *Leibniz ScienceCampus "Digital Transformation of Research" (DiTraRe)*, funded by the Leibniz Association as part of the Leibniz Competition and *NFDI4Chem*, funded by the Deutsche Forschungsgemeinschaft (DFG), Project Number 441958208.

Authorship Contributions

- Felix Bach: Conceptualisation; Methodology; Writing – Original Draft; Supervision; Data Curation; Review & Editing.
- Linda Nierling: Theoretical Framework; Writing – Original Draft; Writing – Review & Editing; Ethical Reflection; Integration of TA perspectives.
- Christian Bonatto Minella: Writing – Review & Editing; Interoperability Aspects; Proofreading.
- Angelina Sophie Dähms: Writing – Review & Editing, Literature Review; Proofreading.

Bibliography

Academia Insider. 2024. *AI Tools for Research Papers and Academic Research.* Visited on April 1, 2025. https://academiainsider.com.

Albrecht, Steffen. 2023. *ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen.* Technical report. https://doi.org/10.5445/IR/1000158070.

Arteaga, Gabriel Y., Thomas B. Schön, and Nicolas Pielawski. 2024. *Hallucination Detection in LLMs: Fast and Memory-Efficient Fine-Tuned Models.* https://doi.org/10.48550/ARXIV.2409.02976.

Bejar Alonso, Isabel, Oliver Brüning, Paolo Fessia, Mike Lamont, Lucio Rossi, Laurent Tavian, and Markus Zerlauth. 2020. *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report.* CERN Yellow Reports: Monographs 10 (2020). CERN. https://doi.org/10.23731/CYRM-2020-0010.

Berners-Lee, Tim, and Mark Fischetti. 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor.* Harper San Francisco. ISBN: 1402842937.

Borgman, Christine L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet.* The MIT Press. ISBN: 9780262255783. https://doi.org/10.7551/mitpress/7434.001.0001.

Buchanan, Bruce, and Edward H. Shortliffe. 1984. *Rule-based Expert System – The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley. Visited on June 27, 2025. http://www.shortliffe.net/Buchanan-Shortliffe-1984/MYCIN%20Book.htm.

Butler, Declan. 2005. "A new leaf". *Nature* 436 (7047): 20–21. https://doi.org/10.1038/436020a.

Ceruzzi, Paul. 2000. *History of Modern Computing.* 412. MIT Press. ISBN: 9780262531696.

Church, Alonzo. 1936. “An Unsolvable Problem of Elementary Number Theory”. *American Journal of Mathematics* 58 (2): 345. https://doi.org/10.2307/2371045.

Cologna, Viktoria, Niels G. Mede, Sebastian Berger, John Besley, Cameron Brick, Marina Joubert, Edward W. Maibach, et al. 2025. “Trust in scientists and their role in society across 68 countries”. *Nature Human Behaviour* 9 (4): 713–730. https://doi.org/10.1038/s41562-024-02090-5.

Davis, Martin. 2000. *Engines of Logic: Mathematicians and the Origin of the Computer.* 272. W. W. Norton / Company. ISBN: 9780393322293.

Deutsche Forschungsgemeinschaft. 2023. *Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaftenund das Förderhandeln der DFG.* Visited on April 1, 2025. https://www.dfg.de/resource/blob/289674/ff57cf46c5ca109cb18533b21fba49bd/230921-stellungnahme-praesidium-ki-ai-ai-data.pdf.

Dey, Prasenjit, Srujana Merugu, and Sivaramakrishnan Kaveri. 2025. “Uncertainty-Aware Fusion: An Ensemble Framework for Mitigating Hallucinations in Large Language Models”. In *Companion Proceedings of the ACM on Web Conference 2025,* 947–951. WWW ’25. ACM. https://doi.org/10.1145/3701716.3715523.

Digital Science, Mark Hahnel, Graham Smith, henning schoenenberger henning, Niki Scaplehorn, and Laura Day. 2023. *The State of Open Data 2023.* Technical report. https://doi.org/10.6084/M9.FIGSHARE.24428194.V1.

Elephas. 2024. *Best AI Research Assistants in 2024.* Visited on April 1, 2025. https://elephas.app/blog/best-ai-research-assistants.

Elliott, Kevin C. 2017. *Tapestry of Values An Introduction to Values in Science: An Introduction to Values in Science.* 224. Oxford University Press, Incorporated. ISBN: 9780190260811.

European Commission. 2024. *Living Guidelines on the Responsible Use of Generative AI in Research.* Visited on April 1, 2025. https://research-and-innovation.ec.europa.eu/document/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en.

European Research Council. 2023. *Use and Impact of Artificial Intelligence in the Scientific Process.* Visited on April 1, 2025. https://erc.europa.eu/sites/default/files/2023-12/AI_in_science.pdf.

Evans, Lyndon, and Philip Bryant. 2008. “LHC Machine”. *Journal of Instrumentation* 3 (08). https://doi.org/10.1088/1748-0221/3/08/s08001.

Fecher, Benedikt, Marcel Hebing, Melissa Laufer, Jörg Pohle, and Fabian Sofsky. 2023. “Friend or foe? Exploring the implications of large language models on the science system”. *AI and Society* 40 (2): 447–459. https://doi.org/10.1007/s00146-023-01791-1.

Feigenbaum, Edward A. 1983. *The fifth generation: Artificial intelligence and Japan's computer challenge to the world.* 275. Addison-Wesley. ISBN: 0201115190.

Floridi, Luciano, and Josh Cowls. 2019. "A Unified Framework of Five Principles for AI in Society". *Harvard Data Science Review,* https://doi.org/10.1162/99608f92.8cd550d1.

Ginsparg, Paul. 1994. "First Steps Towards Electronic Research Communication". *Computers in Physics* 8 (4): 390–396. https://doi.org/10.1063/1.4823313.

Gödel, Kurt. 1931. "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I". *Monatshefte für Mathematik und Physik* 38 (1): 173–198. https://doi.org/10.1007/bf01700692.

Hafner, Katie, and Matthew Lyon. 1996. *Where Wizards Stay Up Late The Origins Of The Internet: The Origins Of The Internet.* Simon / Schuster Books. ISBN: 9780684872162.

Herres-Pawlis, Sonja, Johannes C. Liermann, and Oliver Koepler. 2020. "Research Data in Chemistry – Results of the first NFDI4Chem Community Survey". *Zeitschrift für anorganische und allgemeine Chemie* 646 (21): 1748–1757. https://doi.org/10.1002/zaac.202000339.

Jacquard, Joseph Marie. 1801. *Programmable loom using punch cards.* Lyon: Science Museum.

Jacyszyn, Anna M., Felix Bach, Harald Sack, Christiane Noe, Marius Albiez, Sabine Barthlott, Silvia Becker, et al. 2025. *Interim Report of the Leibniz Science Campus "Digital Transformation of Research" (DiTraRe).* Zenodo. https://doi.org/10.5281/ZENODO.14941635.

Jahnel, Jutta, Michael Nentwich, Pauline Riousset, Christine Michram, Christoph Kehl, and Steffen Albrecht. 2025. *Generative AI in Policy-Oriented Technology Assessment.* Visited on June 27, 2025. https://publikationen.bibliothek.kit.edu/1000179858.

Jordan, Michael I., and Tom M. Mitchell. 2015. "Machine learning: Trends, perspectives, and prospects". *Science* 349 (6245): 255–260. https://doi.org/10.1126/science.aaa8415.

Kahle, Brewster. 2007. "Universal Access to All Knowledge". *American Archivist* 70 (1): 23–31. https://www.jstor.org/stable/40294448.

Leibniz, Gottfried Wilhelm. 1703. *Explication de l'Arithmétique Binaire.* Mémoires de l'Académie Royale des Sciences.

Lih, Andrew. 2009. *The Wikipedia revolution: how a bunch of nobodies created the world's greatest encyclopedia.* 246. Hyperion. ISBN: 9781401303716.

Meça, Alba, and Nirvana Shkëlzeni. 2023. "Academic Integrity in the Face of Generative Language Models". In *Emerging Technologies in Computing,* 58–70. Springer Nature Switzerland. ISBN: 9783031502156. https://doi.org/10.1007/978-3-031-50215-6_5.

Mede, Niels G., and Viktoria Cologna. 2025. *Trust in Science and Science-Related Populism: A Global Assessment.* https://doi.org/10.17605/OSF.IO/9KSRJ.

Messeri, Lisa, and Molly J. Crockett. 2024. "Artificial intelligence and illusions of understanding in scientific research". *Nature* 627 (8002): 49–58. https://doi.org/10.1038/s41586-024-07146-0.

National Academies of Sciences, Engineering, and Medicine. 2022. *Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop.* National Academies Press. ISBN: 9780309686525. https://doi.org/10.17226/26532.

Neumann, John von. 1945. *First Draft of a Report on the EDVAC.* Technical report. Moore School of Electrical Engineering. Visited on June 27, 2025. http://web.mit.edu/sts.035/www/PDFs/edvac.pdf.

O'Regan, Gerard. 2021. *A Brief History of Computing.* Springer International Publishing. ISBN: 9783030665999. https://doi.org/10.1007/978-3-030-66599-9.

Randell, Brian. 1973. *The Origins of Digital Computers.* Springer Berlin Heidelberg. ISBN: 9783642961458. https://doi.org/10.1007/978-3-642-96145-8.

Rosman, Tom, Michael Bosnjak, Henning Silber, Joanna Koßmann, and Tobias Heycke. 2022. "Open science and public trust in science: Results from two studies". *Public Understanding of Science* 31 (8): 1046–1062. https://doi.org/10.1177/09636625221100686.

Sakana AI. 2024. *AI Scientist.* Visited on April 1, 2025. https://sakana.ai/ai-scientist.

Sanger, Fred, Steve Nicklen, and Alan R. Coulson. 1977. "DNA sequencing with chain-terminating inhibitors". *Proceedings of the National Academy of Sciences* 74 (12): 5463–5467. https://doi.org/10.1073/pnas.74.12.5463.

Schreiber, Gerhard, and Lukas Ohly. 2024. *KI:Text: Diskurse über KI-Textgeneratoren.* De Gruyter. ISBN: 9783111351490. https://doi.org/10.1515/9783111351490.

Shannon, Claude E. 1937. "A Symbolic Analysis of Relay and Switching Circuits". Master's thesis, MIT. Visited on June 27, 2025. http://hdl.handle.net/1721.1/11173.

———. 1948. "A Mathematical Theory of Communication". *Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Smith, Roger. 2007. *Being Human: Historical Knowledge and the Creation of Human Nature.* 256. Columbia University Press. ISBN: 9780231141666.

Suber, Peter. 2012. *Open access.* 242. MIT Press. ISBN: 9780262517638.

The Royal Society. 2024. *Science in the Age of AI.* Visited on April 1, 2025. https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/.

Tolle, Kristin M., D. Stewart W. Tansley, and Anthony J. G. Hey. 2011. "The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]". *Proceedings of the IEEE* 99 (8): 1334–1337. https://doi.org/10.1109/jproc.2011.2155130.

Toole, Betty A. 1998. *Ada, the Enchantress of Numbers: Prophet of the Computer Age.* 323. Critical Connection. ISBN: 9780912647180.

Tremouilhac, Pierre, Chia-Lin Lin, Pei-Chi Huang, Yu-Chieh Huang, An Nguyen, Nicole Jung, Felix Bach, et al. 2020. "The Repository Chemotion: Infrastructure for Sustainable Research in Chemistry**". *Angewandte Chemie International Edition* 59 (50): 22771–22778. https://doi.org/10.1002/anie.202007702.

Turing, Alan M. 1936. "On Computable Numbers, with an Application to the Entscheidungsproblem". In *Proceedings of the London Mathematical Society,* 42:230–265. 3. https://doi.org/10.1016/b978-0-08-009217-1.50024-4.

UNESCO. 2023. *Guidance for Generative AI in Education and Research.* Visited on April 1, 2025. https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/.

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. https://doi.org/10.1038/sdata.2016.18.

Yang, Yi, Yitong Ma, Hao Feng, Yiming Cheng, and Zhu Han. 2025. "Minimizing Hallucinations and Communication Costs: Adversarial Debate and Voting Mechanisms in LLM-Based Multi-Agents". *Applied Sciences* 15 (7): 3676. https://doi.org/10.3390/app15073676.

Nachnutzbar, aber nicht reproduzierbar. Abgeleitete Textformate für die Arbeit mit urheberrechtsbewehrten Texten

Philippe Genêt [1], Lukas Weimer [2]

[1] Deutsche Nationalbibliothek, Frankfurt am Main
[2] Niedersächsische Staats- und Universitätsbibliothek Göttingen

Der vorliegende Bericht behandelt die Herausforderungen und Lösungsansätze im Umgang mit urheberrechtlich geschützten Texten in der wissenschaftlichen Forschung. Trotz der Einführung der Text und Data Mining-Schranke (§60d UrhG) im Jahr 2018 bleiben zahlreiche Einschränkungen bestehen, die die freie Nutzung dieser Ressourcen durch Forschende verhindern. Bibliotheken, die den gesetzlichen Auftrag haben, ihre Sammlungen der Allgemeinheit zugänglich zu machen, engagieren sich daher u. a. in Text+, einem Konsortium der Nationalen Forschungsdateninfrastruktur.

Text+ zielt darauf ab, sprach- und textbasierte Forschungsdaten dauerhaft zu bewahren und für die Forschung verfügbar zu machen. Ein zentraler Ansatz dabei ist die Nutzung abgeleiteter Textformate (ATFs), die mittels Methoden des Text- und Data Minings erstellt werden. Mit ATFs können Forschungsfragen beantworten werden, ohne die Rechte der Urheberrechtsinhaber zu beeinträchtigen. Die Herstellung von ATFs erfolgt durch Anreicherungen wie Part-of-Speech-Tagging und Lemmatisierung und durch anschließende gezielte Informationsreduktion, die eine Rekonstruktion des Ausgangstextes verhindert.

In Text+ arbeiten verschiedene Universitäten und außeruniversitäre Einrichtungen an der Standardisierung und rechtlichen Klärung von ATFs. Zudem werden Tools und Textkorpora bereitgestellt, die nur in Form von ATFs veröffentlicht werden dürfen. Das Projekt fördert auch die Forschung zur Eignung und Rekonstruierbarkeit von ATFs und betreibt Disseminationsmaßnahmen, um das Thema einer breiteren Öffentlichkeit zugänglich zu machen. Die DNB untersucht zudem im Projekt CORAL die Möglichkeit, Sprachmodelle mit ATFs zu trainieren, ohne Originaltexte reproduzieren zu können. Durch diese Aktivitäten möchten die beteiligten Bibliotheken ihre Relevanz für die Forschungsgemeinschaft unterstreichen und sich weiter als wertvolle Partner der Wissenschaft etablieren.

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23936 (CC BY-SA 4.0).

Keywords: Abgeleitete Textformate, Urheberrecht, NFDI, Text+

1 Herausforderung Urheberrecht

Forschung mit urheberrechtsbewehrten Texten unterliegt zahlreichen Einschränkungen. Daran ändert auch die Einführung der sogenannten Text und Data Mining-Schranke (§60d UrhG[1]) im Jahre 2018 nur wenig (Iacino, Kamocki und Leinen 2023). Zwar erlaubt sie die Analyse urheberrechtlich geschützten Materials mit maschinellen Methoden, wie sie zum Beispiel die Computational Literary Studies bzw. die Digital Humanities anwenden, zu wissenschaftlichen Zwecken grundsätzlich – von einer freien Nutzung dieser Ressourcen durch Forschende kann jedoch keine Rede sein. Vielmehr sind Zusammenstellung, öffentliche Zugänglichmachung und Archivierung von Forschungsdaten weiterhin an zahlreiche Bedingungen geknüpft. Der lange Wirksamkeitszeitraum des Urheberrechts – es greift bis 70 Jahre nach dem Tod der Autorin bzw. des Autors – bewirkt, dass sich etwa die digitale Literaturwissenschaft sehr intensiv mit Werken aus dem 19. Jahrhundert beschäftigt, kaum aber mit neueren Texten.

Bibliotheken haben ein genuines Interesse und häufig auch den gesetzlichen Auftrag[2], ihre Sammlungen für die Allgemeinheit nutzbar zu machen. Daraus erwächst der Wunsch, der Wissenschaft Zugang auch zu Werken zu ermöglichen, die dem Urheberrecht unterstehen. Ein Ausdruck dieses Wunsches ist das Engagement mehrerer Bibliotheken in Text+[3], einem der 26 Fach- und Methodenkonsortien der Nationalen Forschungsdateninfrastruktur (NFDI[4]).

Die Mission von Text+ ist es, sprach- und textbasierte Forschungsdaten dauerhaft zu bewahren und für die Forschung verfügbar zu machen. Text+ nutzt dazu ein Netzwerk aus Daten- und Kompetenzzentren, die ihre Ressourcen und Erfahrung zur Verfügung stellen. Die Daten und Dienste können zentral abgerufen und verarbeitet werden, wie beispielsweise über die Text+ Registry[5] oder die Federated Content Search[6]. Viele der zur Verfügung gestellten Ressourcen können frei genutzt werden, einige jedoch sind aufgrund von rechtlichen Beschränkungen nur begrenzt zugänglich. Sehr umfangreiche und überwiegend urheberrechtlich geschützte Bestände bringen hier insbesondere die Deutsche Nationalbibliothek (DNB) und die Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB Göttingen) ein.

1 https://www.gesetze-im-internet.de/urhg/___60d.html; *Besucht am 26. März 2025.*
2 Vgl. zum Beispiel das Gesetz über die Deutsche Nationalbibliothek, insbesondere §2 Abs. 1, https://www.gesetze-im-internet.de/dnbg/BJNR133800006.html; *Besucht am 26. März 2025.*
3 https://text-plus.org/; *Besucht am 26. März 2025.*
4 https://www.nfdi.de/; *Besucht am 26. März 2025.*
5 https://registry.text-plus.org/; *Besucht am 26. März 2025.*
6 https://fcs.text-plus.org/; *Besucht am 26. März 2025.*

2 Die Antwort: abgeleitete Textformate (ATF)

Mit dem Ziel, auch solche Werke für die Wissenschaft nutzbar zu machen, rücken in der gemeinsamen Arbeit in Text+ abgeleitete Textformate (ATF, siehe Schöch u. a. 2020) in den Fokus. Sie können – unter Verwendung von Methoden des Text- und Data Minings – so hergestellt werden, dass einerseits das Ergebnis noch die Beantwortung mindestens einer Forschungsfrage ermöglicht, der verbleibende Rest andererseits aber die Rechte der Urheberrechtsinhaber nicht mehr beeinträchtigt. Voraussetzung dafür ist u. a., dass keine urheberrechtlich geschützten Elemente übernommen werden und dass keine triviale Möglichkeit zur Rekonstruktion des Ausgangstextes besteht. Solche ATFs können daher frei veröffentlicht werden. Bei der Erzeugung von mehr als einem ATF von einem Dokument bzw. einem Korpus ist darauf zu achten, dass auch durch deren Kombination keine Rekonstruktion möglich ist. Nicht alle ATFs sind automatisch rechtefrei. Ist eine Rekonstruktion des Originaltextes ohne größeren Aufwand möglich, fällt das ATF weiterhin unter das Urheberrecht (vgl. Iacino u. a. 2024).

2.1 Herstellung von ATFs

ATFs werden auf der Basis des Ausgangstextes durch die Anwendung einer Reihe von Veränderungen erstellt. Hierzu gehören in einem ersten Schritt meist Anreicherungen wie Part-of-Speech-Tagging (POS), Lemmatisierung und/oder Verlinkungen von Named Entities zu Normdaten in Form von Annotationen. Ebenso sind statistische Analysen auf dem Ausgangstext als Vorverarbeitungsschritte sinnvoll, diese sind ebenfalls als Annotationen zu sehen. Danach erfolgt eine gezielte Informationsreduktion, die einerseits auf einer Reihe von Veränderungen basiert, die typischerweise automatisiert ablaufen, und ganz wesentlich auf der Entscheidung, welche Granularität diesen Operationen zu Grunde liegen.

Vier Operationen stehen für die Informationsreduktion zur Verfügung: Löschen, Behalten, Ersetzen und Vertauschen. Diese können auf verschiedenen Granularitätsebenen greifen (z. B. auf Token-, Satz- oder Absatzebene) und in Bezug auf unterschiedliche Größen (z. B. pro Dokument, pro Werk oder pro Korpus). Granularität und Bezugsgrößen sind für die Wiederherstellbarkeit des Ausgangstextes entscheidend. Ist beispielsweise die Wortreihenfolge in einem kurzen Satz randomisiert, ist die Reproduktion des ursprünglichen Wortlauts recht einfach. Ist die Reihenfolge der Wörter hingegen auf einer ganzen Seite oder gar in einem ganzen Dokument vertauscht, ist es – auch mit maschinellen Mitteln – in der Regel unmöglich, den exakten Originalwortlaut zu rekonstruieren.

Unterschieden werden vor allem tokenbasierte und vektorbasierte ATFs. Bei Letzteren werden einzelne Wörter durch hochdimensionale Vektoren mit numerischen Einträgen ersetzt. So entstehen zum Beispiel Word Embeddings, die unter anderem beim Training

von Sprachmodellen zum Einsatz kommen. Im Folgenden wenden wir uns jedoch lediglich tokenbasierten ATFs zu.

2.2 Beispiele für ATFs

Tabelle 1: Beispiel für bag of words, hier: Vergleich der Häufigkeit von 15 Tokens in drei Romanen von Charles Dickens.[7]

	Dickens_Barnaby	**Dickens_Cricket**	**Dickens_Dombey**
,	26044	3192	37907
the	12027	1443	15884
‘	10303	41	15587
and	10267	1173	13548
.	9629	1351	13442
to	6274	761	9314
of	6378	698	9147
a	5470	689	7125
in	4396	526	6164
his	4151	337	5067
I	3313	596	4841
that	3179	365	5038
he	3242	273	3896
was	2888	392	4088
it	2497	442	3836

Ein prominentes, vielseitig einsetzbares und darum häufig verwendetes ATF ist die einfache Dokument-Term-Matrix, auch bag of words genannt. Dabei handelt es sich um eine Liste der im Einzeltext vorkommenden Tokens und deren absoluter Häufigkeit. Sie kann erweitert werden, indem zum Beispiel die Wortform, die Wortart oder das Lemma der Tokens angegeben werden. Diese einfache Frequenzliste dient zahlreichen analytischen Zwecken, etwa der Klassifikation oder dem Clustering von Texten, und reicht sogar für die Autorschaftsattribution aus.

Ein ebenfalls sehr bekanntes Beispiel für ATFs sind n-Gramme, seit sie von Google LLC im großen Stil veröffentlicht wurden[8]. Ein n-Gramm ist eine Folge von n aufeinanderfolgenden Tokens, die üblicherweise als alphabetische oder nach Häufigkeit sortierte Listen dargestellt werden. Als Sonderform gilt die Ranking-Liste, bei der statt der Frequenz

7 https://github.com/dh-trier/tmr/blob/gh-pages/target/en/three/frq-lower.csv; *Besucht am 26. März 2025.*

8 https://storage.googleapis.com/books/ngrams/books/datasetsv3.html; *Besucht am 26. März 2025.*

nur der Rang in der nach Frequenz sortierten Liste angegeben wird. Mit diesem Format kann zum Beispiel Stilometrie betrieben oder der grammatische Wandel über Zeiträume hinweg betrachtet werden. Zudem werden sie in der automatischen Spracherkennung verwendet.

Tabelle 2: Beispiel für n-Gramme, hier: Trigramme aus einem Korpus von fünf Romanen Theodor Fontanes.[9]

n-Gramm	**Häufigkeit**
gott sei dank	43
ja gnädigste frau	17
auch heute wieder	13
doch auch wieder	11
ist doch auch	11
ist immer so	10
gnädigste frau ist	10
war so war	10
nein gnädigste frau	10
wird ja wohl	9
ist doch immer	9
das ist recht	9

Ein drittes Verfahren operiert mit der selektiven Maskierung oder Löschung von Tokens. Löscht man beispielsweise den Sprechtext aus Dramen, so ist das Ergebnis nicht mehr urheberrechtlich geschützt, dient aber noch verschiedenen wissenschaftlichen Zwecken wie etwa der Netzwerkanalyse.

3 Aktivitäten in Text+ und darüber hinaus

In Text+ arbeiten die SUB Göttingen und die DNB gemeinsam mit Partner:innen aus der Forschung – insbesondere von der Universität Trier und dem Leibniz-Institut für Deutsche Sprache – an verschiedenen Aspekten der abgeleiteten Textformate. Mit der Entwicklung eines DIN-Standards soll die Herstellung und der Gebrauch von ATFs vereinheitlicht werden. Daneben wird in Text+ auch die komplexe rechtliche Lage beleuchtet – so erscheint zum Beispiel die Publikation „Legal status of Derived Text Formats“ (Iacino u. a. 2024) in der kommenden Ausgabe 3/2024 der Zeitschrift Recht und Zugang.

9 https://github.com/dh-trier/tmr/blob/gh-pages/target/de/fontane/ngr-wordform-3/allngrs.tsv; *Besucht am 26. März 2025.*

10 https://textgridrep.org/project/TGPR-adf9b705-1533-b0ef-491b-674878350ecb; *Besucht am 26. März 2025.*

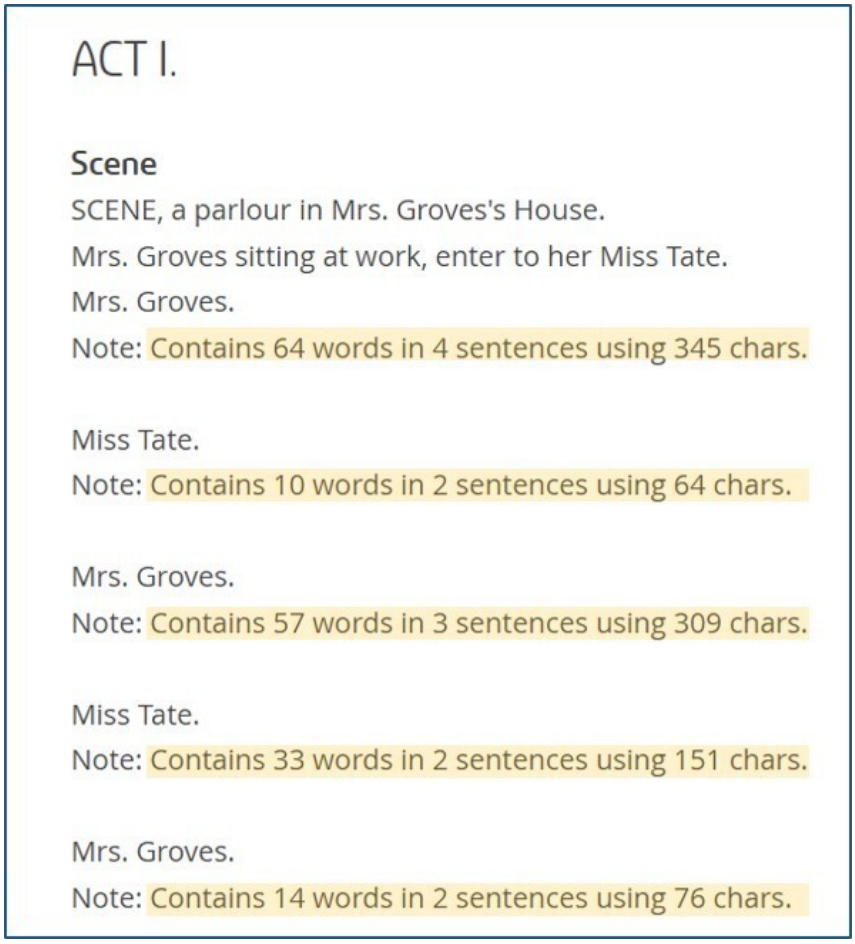


Abbildung 1: Beispiel für Maskierung/Löschung von Tokens, hier: Auszug aus dem American Drama Corpus im TextGrid Repository.[10]

Einige der von Text+ bereitgestellten Tools bieten sich für die Weiterverarbeitung von ATFs an, wie zum Beispiel die Natural Language Processing-Pipeline MONAPipe[11] der SUB Göttingen sowie das von der Universität Tübingen bereitgestellte Annotationstool WebLicht[12]. Gleichzeitig bietet Text+ über seine Datenzentren Textkorpora an, die nur in Form von ATFs veröffentlicht werden dürfen.

Um mehr Wissenschaftler:innen zu motivieren, sich mit dem Thema auseinanderzusetzen, fördert Text+ ebenfalls die einschlägige Forschung, zum Beispiel zur Eignung bestimmter ATFs für einzelne Forschungsfragen (Du und Schöch 2025; Du 2023) oder zur Rekonstruierbarkeit von Ausgangstexten (Kugler u. a. 2023). Für eine breitere Rezeption und Bekanntmachung des Themas dienen schließlich zahlreiche Disseminationsmaßnahmen wie Vorträge, Workshops und Informationsmaterial im Text+ Webportal[13].

11 https://pypi.org/project/monapipe/; *Besucht am 26. März 2025.*
12 https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/Main_Page.html; *Besucht am 26. März 2025.*
13 https://text-plus.org/themen-dokumentation/atf/; *Besucht am 26. März 2025.*

Die DNB lotet daneben noch im Forschungsprojekt CORAL[14] die Möglichkeit aus, ob und wie leistungsfähige Sprachmodelle mit ATFs trainiert werden können, die das Wissen aus urheberrechtlich geschützten Quellen in sich tragen, ohne Gefahr zu laufen, Originaltexte reproduzieren zu können.

Mit dem Engagement möchten die in Text+ aktiven Bibliotheken ihre Relevanz für die Forschungscommunitys unterstreichen und ihnen noch bessere Partnerinnen werden.

Danksagungen

Die Autoren danken den Mitgliedern der AG Abgeleitete Textformate in der Task Area Collections in Text+: Florian Barth, José Calvo Tello, Keli Du, Jörg Knappen, Daniel Kurzawe, Peter Leinen, Piroska Lendvai, Christof Schöch, Thorsten Trippel und Andreas Witt.

Beiträge

- Philippe Genêt: Verschriftlichung – Originalentwurf
- Lukas Weimer: Verschriftlichung – Überprüfung & Überarbeitung

Mittelgeber

Die vorliegende Publikation wurde im Rahmen des Konsortiums Text+ im Kontext der Arbeit des Vereins Nationale Forschungsdateninfrastruktur (NFDI) e.V. verfasst. NFDI wird von der Bundesrepublik Deutschland und den 16 Bundesländern finanziert, und das Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 460033370. Die Autoren bedanken sich für die Förderung sowie Unterstützung. Ein Dank geht außerdem an alle Einrichtungen und Akteur:innen, die sich für den Verein und dessen Ziele engagieren. Das diesem Bericht teilweise zugrunde liegende Vorhaben CORAL wird mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS24077 A-D gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

14 https://coral-project.github.io/index.html; *Besucht am 26. März 2025.*

Interessenkonflikte

Es liegen keine Interessenkonflikte der Autoren vor.

Literaturverzeichnis

Du, Keli. 2023. „Understanding the impact of three derived text formats on authorship classification with Delta“. In *DHd Open Humanities Open Culture. 9. Tagung des Verbandes.* Zenodo. https://doi.org/10.5281/ZENODO.7715299.

Du, Keli und Christof Schöch. 2025. „Shifting Sentiments? What happens to BERT-based Sentiment Classification when derived text formats are used for fine-tuning“. In *DH-Conference 2024.* To appear.

Iacino, Gianna, Pawel Kamocki, Keli Du, Christof Schöch, Andreas Witt, Philippe Genêt und José Calvo Tello. 2024. „Legal status of Derived Text Formats“. *Recht und Zugang* 3 / 2024.

Iacino, Gianna, Pawel Kamocki und Peter Leinen. 2023. *Assessment of the Impact of the DSM-Directive on Text+.* Technischer Bericht. Zenodo. https://doi.org/10.5281/ZENODO.12759960.

Kugler, Kai, Simon Münker, Johannes Höhmann und Achim Rettinger. 2023. „InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline“. *Journal of Computational Literary Studies* 2 (1). https://doi.org/10.48694/JCLS.3572.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann und Jörg Röpke. 2020. „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen“. *Zeitschrift für digitale Geisteswissenschaften* 5. https://doi.org/10.17175/2020_006.

Shaping the Data Steward Community in Germany

Jens Dierkes [1,*], Daniela Hausen [2], Ute Trautwein-Bruns [3]

[1] University and City Library, University of Cologne
[2] Heinrich Heine University Düsseldorf
[3] RWTH Aachen University
[*] Corresponding author: jens.dierkes@uni-koeln.de

Research data management and data stewardship are no longer new concepts. In a national context, there is a lack of a commonly accepted framework for describing data stewards. Despite this, more and more RDM professionals are describing themselves as data stewards and a community of practice (CoP) seems to be emerging. Here we describe the first steps of this new community, which is creating a vision, establishing a communication structure and finding like-minded RDM professionals to join. A CoP will act as a social learning space to advance research data stewardship by sharing good practices and initiate discussions on job profiles, education and training of data stewards, as well as effective integration of data stewards into research teams and career paths for this group of professionals who have become an important part of today's research enterprise.

Keywords: Data Steward, Community of Practice, Community Building, Research Data Management

1 Introduction

Improving data quality through data stewardship is gaining increasing scientific and practical relevance, as evidenced by the growing interest in workshops such as "Data Stewardship goes Germany" (DSgG, see Hausen et al. 2024). The last two workshops in 2023 and 2024 revealed that there is currently a lack of adequate support for data stewards in Germany. Although it has been recognised that data stewards can play a strategic

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23937 (CC BY-SA 4.0).

and highly supportive role in research, this role is often not actively implemented. Meanwhile, international projects – for example, in the Netherlands (Jetten et al. 2021) and the national project DataStew (Seidlmayer et al. 2023) – have addressed the roles and profiles of data stewards, and it is now necessary to translate these theoretical insights into practice.

To systematically capture these challenges and develop solutions, the last two DSgG workshops focused on this theme. At DSgG 2023, it became apparent that, unlike in other countries, a dedicated community for data stewardship does not yet exist in Germany, although the need is clearly present (Krause et al. 2024). This situation was then systematically assessed at DSgG 2024 using the Community Canvas (Pfortmüller, Luchsinger, and Mombartz 2017), with small groups working on individual aspects such as identity, structure, and experiences. It became clear that data stewardship is much more than simply sharing experiences; it is essential to clearly define the role of data stewards, to promote their integration into research teams, and to establish professional career paths as well as institutional strategies. By building a nationally networked community, a virtual space is created where both technical and content-related challenges can be jointly discussed and innovative solutions developed (Hausen, Della Chiesa, et al. 2025; Hausen, Windeck, et al. 2025).

This paper will elaborate on these outcomes through both a lightning talk (Dierkes and Hausen 2025) and an interactive poster (Hausen, Dierkes, and Trautwein-Bruns 2025), thereby advancing the strategic importance of data stewardship and community within the research ecosystem.

2 Perspectives on the Emerging Data Steward Community in Germany

The interactive poster "Shape the Data Steward Community! Your Ideas Count ..." was designed to gather diverse perspectives from key stakeholders – researchers, infrastructure members, NFDI[1] members, and Data Stewards themselves – on essential questions related to the development of the emerging Data Steward Community in Germany. By providing an interactive space for feedback and exchange, the poster encouraged participants to share their own experiences and expectations regarding the community. The aim was to explore external viewpoints on the community while specifically engaging researchers and infrastructure staff in the discussion.

1 NFDI is the German National Research Data Infrastructure, https://nfdi.de; *Visited May 21st 2025.*

What are data steward tasks and skills?

Data Stewards describe themselves as "an important link between the technical infrastructure and the scientists", emphasizing the need for "field-specific knowledge (e.g. methods)", "technical expertise", and a "hands-on mentality". Infrastructure members highlight their role as an "important contact point for external partners" and stress the importance of "technical programming skills". NFDI members focus on interpersonal skills, stating that Data Stewards should be able to "listen carefully and identify problems".

The responses highlight that Data Stewards require a combination of technical expertise, domain-specific knowledge, and strong communication skills. They act as mediators between researchers and infrastructure (Rothfritz 2019).

What characterises successfully integrated data stewardship in daily research?

According to Data Stewards, successful integration means that "scientists are happy to contact the Data Steward", that there are "networking opportunities, community building", and that Data Stewards serve as a "low-barrier point of contact". Infrastructure members emphasize visibility, stating that a well-integrated Data Steward is a "person who is seen and approached (especially beneficial in smaller institutions)". Researchers define success in practical terms: A data steward "benefits / improves my workflows".

Successful integration of data stewardship into everyday research is therefore characterised by high visibility of data stewards and the practical support for the research process.

How is data stewardship knowledge managed?

Data Stewards question whether knowledge management is even necessary: "Is it managed? Does it have to?". Infrastructure members see institutional anchoring of the data steward community as a key challenge: "Where should it be institutionally anchored – DHd, NFDI? A question of long-term sustainability!". Researchers focus on reusability: "to reuse resources".

There is no clear consensus on knowledge management within *data stewardship*. While some question its necessity, others emphasize the importance of institutional anchoring and sustainability. An established CoP could play a key role by enabling regular exchange and documenting best practices, which would facilitate the efficient reuse and sharing of resources.

What do you expect from the data steward community and what are you able / willing to contribute?

For the data stewards, the community provides a space for learning from each other: "input: my experience; output: new skills through exchange". Infrastructure members seek practical guidance, referring to "low-hanging fruits: how do I reach researchers, overview" and "first steps, how to get started". The NFDI members expressed interest in "connecting with the *RDM-Helpdesk* network (working group in the NFDI Section Education & Training; Engel et al. 2024), with plans to establish a base service for Helpdesk networking (Support4RDM)". "Support4RDM plans to provide, among other things, a ticketing system and an *RDM-Helpdesk* catalogues as well as a *federated helpdesk network*".

The Data Steward Community is expected to serve as a hub for knowledge exchange and skill development. Stakeholders see it as a place to gain practical guidance and connect with existing support networks, highlighting the need for structured collaboration and accessible resources.

How will the world be different by having this community?

Data Stewards see a clear impact: "support for onboarding new Data Stewards", "strengthening science, innovation, societal good", and "making science more rewarding (& fun?)". Infrastructure members reflect on sustainability, asking whether the community will make research data management "somewhat more sustainable after the projects have run out?".

The emerging data steward community could contribute to the long-term professionalization and institutionalization of *data stewardship*. By providing support for newcomers and strengthening research as a whole, it would enhance the sustainability and impact of *data stewardship* practices.

3 Discussion and next Steps

As described above, there is a growing interest in networking and joining forces to think together, learn and shape data stewardship as a CoP (Figure 1). This was clearly articulated in the community building exercise at DSgG 2024 using the Community Canvas (Oset García et al. 2023; Hausen, Windeck, et al. 2025; Pfortmüller, Luchsinger, and Mombartz 2017). As a result, three working groups were formed to

- flesh out a vision for the CoP, including strategic elements such as education and career paths, *as well as mapping national and international initiatives*,
- collect challenges faced by data stewards to inform further community activities, and
- establish communication channels and a kind of yellow pages for effective community communication.

The discussion and results of the interactive poster provide further support for this development. They also show the importance of considering different perspectives and, therefore, different expectations. In particular, the boundary-spanning characteristics of data stewards are seen by all stakeholders.

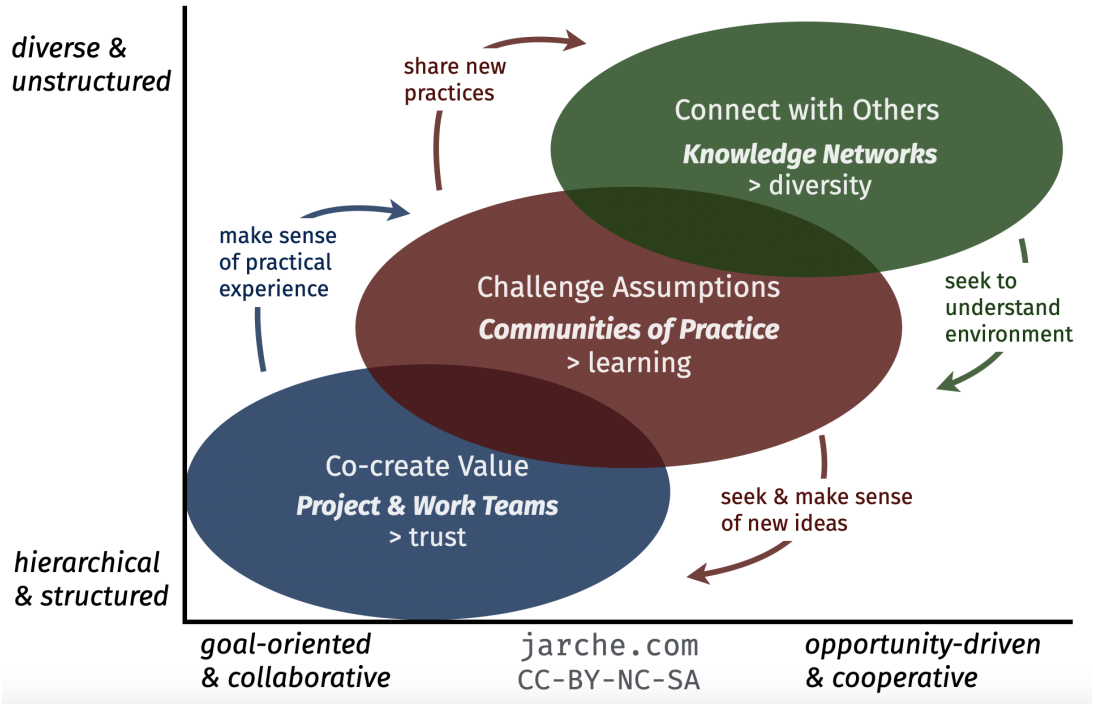


Figure 1: Illustration of the relationship between networks, communities of practice and project teams in terms of collaboration and structure/diversity. Image credits: Harold Jarche[2].

Many of the issues raised in section 2 are already part of the established working groups. *A very interesting question arises regarding the actual need to manage knowledge.* For example, the articulated need for onboarding and collaboration across RDM helpdesks requires a managed explicit body of knowledge. However, there are suggestions that a better understanding of the data management situation is possible if explicit, tacit, individual and group knowledge are treated as four distinct and coequal forms of knowledge, and if knowledge and knowing (as action) are seen as mutually enabling (Cook and Brown 1999). Discussions in the working groups are already recognising these characteristics, and as a CoP all four dimensions can be addressed.

2 https://jarche.com/2020/08/working-smarter-with-pkm-transcripts/; *Visited on March 29, 2025.*

The three working groups are carefully considering topics that could form the basis of a CoP, with the aim of demonstrating added value and enabling effective communication. Only then will it be possible to involve more members and grow as a CoP. At the same time, many parallel developments are taking place in the area of data stewardship. Several of NFDI's disciplinary consortia, in particular, employ data stewards, thereby strengthening the role of data stewardship. In this context, the connections with the RDM-Helpdesk (Support4RDM) must be considered to avoid duplication and foster synergies. The NFDI Section Education and Training, as well as the RDMTraining4NFDI base service, also address the skills required by data stewards. It is important to emphasise that the emerging CoP can act as a unifying element. It is therefore important to maintain the current momentum and to demonstrate the potential added value to the wider community. It is currently being considered to link the community to an existing, established network, such as the DINI/nestor Research Data Working Group[3], with an underlying organisational infrastructure for communication and the organisation of workshops, for example.

Acknowledgements

We would like to thank all members of the emerging data steward community for their enthusiasm and input. In particular, we would like to thank Stefano Della Chiesa, Stefan Kirsch and Hans März for organising and facilitating the current working groups.

Authorship Contributions

All co-authors contributed equally to the manuscript.

Conflict of Interest

The authors state no conflicts of interest.

3 https://www.forschungsdaten.org/index.php/AG_Forschungsdaten; *Visited on March 30, 2025.*

Bibliography

Cook, Scott D. Noam, and John Seely Brown. 1999. "Bridging Epistemologies: The Generative Dance Between Organizational Knowledge and Organizational Knowing". *Organization Science* 10 (4): 381–400. https://doi.org/10.3280/so2008-002003.

Dierkes, Jens, and Daniela Adele Hausen. 2025. *Improving Data Quality through Data Stewardship: An emerging Community of Data Stewards in Germany.* Zenodo. https://doi.org/10.5281/ZENODO.15032606.

Engel, Judith, Patrick Helling, Robert Herrenbrück, Marina Lemaire, Hela Mehrtens, Marcus Schmidt, Martha Stellmacher, Lukas Weimer, Cord Wiljes, and Wolf Zinke. 2024. *Working Group Charter. RDM Helpdesk Network.* Zenodo. https://doi.org/10.5281/ZENODO.14035822.

Hausen, Daniela Adele, Stefano Della Chiesa, Jens Dierkes, Dzulia Terzijska, Ute Trautwein-Bruns, and Jürgen Windeck. 2025. *Data Stewardship in Germany: First Steps Towards a Stronger Community – Workshop Results 2024.* Zenodo. https://doi.org/10.5281/zenodo.15114377.

Hausen, Daniela Adele, Jens Dierkes, and Ute Trautwein-Bruns. 2025. *Community Building with the Community Canvas: Initial Results from the Data Steward Community.* Zenodo. https://doi.org/10.5281/ZENODO.15064085.

Hausen, Daniela Adele, Andrea Jaek, Lina-Louise Kaulbach, and Theresa Lennartz. 2024. *Data Stewardship goes Germany 2024.* Zenodo. https://doi.org/10.5281/ZENODO.14191227.

Hausen, Daniela Adele, Jürgen Windeck, Sabine Schönau, and Jens Dierkes. 2025. *Results of the Community Canvas Sessions –Data Stewardship goes Germany 2024.* Zenodo. https://doi.org/10.5281/ZENODO.15112753.

Jetten, Mijke, Marjan Grootveld, Annemie Mordant, Mascha Jansen, Margreet Bloemers, Margriet Miedema, and Celia W.G. Van Gelder. 2021. *Professionalising data stewardship in the Netherlands. Competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship.* Zenodo. https://doi.org/10.5281/ZENODO.4623713.

Krause, Marlies, Arne Rümmler, Katharina Zinke, Janna Little, and Kay-Michael Würzner. 2024. *Data Stewardship goes Germany.* Zenodo. https://doi.org/10.5281/ZENODO.10514268.

Oset García, Paula, Veerle Van den Eynden, Ziad Choueiki, and Tessa Acar. 2023. *Building a community of data stewards in Flanders: lessons learned from the FRDN Knowledge Hub's experience with the Community Canvas framework.* Zenodo. https://doi.org/10.5281/ZENODO.10039571.

Pfortmüller, Fabian, Nico Luchsinger, and Sascha Mombartz. 2017. *The Community Canvas Guidebook. The Guide to build meaningful communities.* Visited on May 21, 2025. https://community-canvas.org/.

Rothfritz, Laura. 2019. "Data Stewardship als Boundary-Work". *Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft* 443. https://doi.org/10.18452/20636.

Seidlmayer, Eva, Fabian Hoffmann, Jens Dierkes, Birte Lindstädt, Ralf Depping, and Konrad Ulrich Förstner. 2023. "Forschung unterstützen – Empfehlungen für Data Stewardship an akademischen Forschungsinstitutionen". In *Ergebnisse des Projektes DataStew.* PUBLISSO. https://doi.org/10.4126/FRL01-006441397.

Perplexity-Inspired Metasearch-Based Alternatives to FAIR GPT: Open-source AI Consultants for Research Data Management

Thomas Schmidt [1,*], Renat Shigapov [1], Jan Kamlah [1], Irene Schumm [1]

[1] University Library, University of Mannheim, Germany
[*] Corresponding author: thomas.schmidt@uni-mannheim.de

Chatbots and virtual assistants are becoming increasingly popular for user questions and support. With FAIR GPT, the Mannheim University Library released a virtual assistant for research data management (RDM) in 2024, designed to help researchers and institutions in making their data FAIR (Findable, Accessible, Interoperable, Reusable). FAIR GPT provides various RDM services, e.g. metadata enhancement, repository selection and FAIR assessment. However, FAIR GPT has numerous disadvantages: As a 'Custom GPT' of OpenAI, it is proprietary software that only outputs sources for the generated answers if it uses its internal web search tool (which cannot be controlled by the user) and therefore lacks transparency. Reliance on external cloud-based services leads to privacy concerns when dealing with sensitive (meta)data and the chatbot is still prone to hallucinations, thus reducing its trustworthiness.

These issues led us to explore alternative open-source solutions. We searched for open-source alternatives to *Perplexity.ai*, a system known for its ability to provide citations for the information it retrieves through web searches. We identified three candidates available on GitHub: *Perplexica*, *sensei*, and *farfalle*. These tools use local instances of the metasearch engine SearXNG to perform internet searches, using the results as input for Large Language Models (LLMs). We modified these tools to focus specifically on RDM tasks, releasing the new versions on GitHub openly under the names *FAIRplexica*, *FAIR-sensei* and *FAIR-farfalle*.

Keywords: Research Data Management, LLMs, Chatbot, RDM Assistants, FAIR data

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23938 (CC BY-SA 4.0).

1 AI consultants for research data management

Chatbots such as ChatGPT provide intuitive access to information through graphical user interfaces and conversations in natural language. Users can ask individual questions and get answers tailored to their needs. Due to their intuitive usability, such virtual assistants can be implemented particularly well for support tasks: instead of tediously navigating through extensive documentation or FAQ collections, users receive targeted answers to individual questions, even if the Large Language Models (LLMs) on which these systems are based on have disadvantages, such as hallucinations or a lack of transparency regarding the sources on which the response is built, if they are not connected to tools that are able to perform web searches (Fecher et al. 2023).

Regarding RDM, such virtual assistants could support "level one" advisory services as defined in the RISE framework (Rans and Whyte 2017) or complement the service areas "communication and training" as defined in RISE-DE (Hartmann, Jacob, and Weiß 2019). They could answer common questions from researchers in a targeted manner: What are the FAIR principles and why should I publish my research data FAIR? What are best practices for documenting my dataset and which repository is best suited for my research? Virtual assistants could be used to answer such entry-level RDM questions more easily, quickly and flexibly, provided that the response is reliable, robust and transparent.

2 FAIR GPT

A first step towards a virtual AI assistant for RDM is FAIR GPT (Shigapov and Schumm 2024), which has been released as an OpenAI 'Custom GPT', "that combines instructions, extra knowledge, and any combination of skills" (OpenAI 2023). To reduce hallucinations and improve accuracy for certain tasks, FAIR GPT uses external APIs (Institut Français de Bioinformatique 2025; MaastrichtU-IDS 2025; TIB – Leibniz Information Centre for Science and Technology 2025; re3data.org 2013) and uploaded RDM resources (Horizon 2020 guidelines and UB Mannheim 2024a). Its functionalities include metadata enhancement, dataset organization, repository selection, FAIRness assessment, license recommendations, and generating documentation such as data management plans, README files, and codebooks.

However, FAIR GPT has limitations. It does not always provide sources for its answers (depending if it is using an internal web search tool or not), which reduces transparency and trust in its outputs. As part of OpenAI's Custom GPTs, FAIR GPT is not open source, which limits customization, and it lacks an API for integration into existing RDM workflows. Reliance on external cloud-based services leads to privacy concerns when dealing with sensitive (meta)data.

3 Open-source alternatives to FAIR GPT

To mitigate these limitations open-source alternatives to FAIR GPT were sought, with Perplexity.ai being taken as an inspiration. Like a search engine, Perplexity.ai generates answers to user questions by consulting internet sources. The sources used are referenced in the LLM-generated answers by means of footnotes. Users are thus able to scrutinize the LLM's answers for truthfulness, which emphasizes the robustness and reliability of the information and at the same time improves the disadvantages in terms of transparency that apply to FAIR GPT.

Three open-source alternatives to FAIR GPT that follow the Perplexity.ai model were found on GitHub: Perplexica (ItzCrazyKns 2025), sensei (jjleng 2024) and farfalle (rashadphz 2024). Although slight differences in the tech stack of each tool exist, the architectures are comparable:

1. A client-side chat interface provides an input for user questions.
2. Based on the user questions web searches with the open-source metasearch engine SearXNG (2025) are performed and relevant sources retrieved.
3. These sources are scraped, parsed and post-processed using LLMs (API services like OpenAI, Groq and Anthropic and local LLMs via HuggingFace or Ollama are implemented).
4. The generated answer is streamed to the client-side chat interface; Footnotes are included for sentences/text chunks that are based on the content of a specific web source; The footnotes link directly to the corresponding source.

A key point in identifying suitable FAIR GPT alternatives was the open-source nature of all its components. Perplexica as well as sensei and farfalle are published under open-source licenses and are based on the open-source metasearch engine SearXNG. In addition to the use of proprietary API providers such as OpenAI, Anthropic or Google, all tools also implement the use of local open-source LLMs, which offer major advantages in terms of data protection issues that arise when using cloud-based API providers (cf. Table 1).

After identifying and evaluating all three tools, the forks FAIRplexica (UB Mannheim 2025b), FAIR-sensei (UB Mannheim 2024b) and FAIR-farfalle (UB Mannheim 2025a) were created. For FAIR-sensei and FAIR-farfalle the changes to the code base mainly focus on updating the web search by introducing a configurable global context variable that is appended to the search query, which is then processed by SearXNG. Those changes as well as smaller fixes and UI changes are documented in the commit history of the respective repositories (cf. UB Mannheim 2024b, 2025a).

Table 1: Components and additional criteria for the assessment of FAIR GPT alternatives.

	Perplexica	**sensei**	**farfalle**
Components			
LLM API Providers	OpenAI, Anthropic, Google, Groq	OpenAI, Anthropic	OpenAI, Groq
Local LLMs	Ollama, custom OpenAI API endpoints	HuggingFace	Ollama, litellm
Web search	SearXNG	SearXNG	SearXNG
Additional criteria			
License	MIT	Apache 2.0	Apache 2.0
Last Commit	March 2025	October 2024	September 2024

4 FAIRplexica

FAIRplexica was selected for further development based on two criteria: Firstly, the original Perplexica repository was actively maintained, which led to a continuous development activity with numerous bug fixes and updates to the code base until March 2025. In contrast, the other two tools identified had not had any development updates since September (farfalle) or October 2024 (sensei). Secondly, Perplexica's MIT license offered the greatest possible openness in the further development of the software.

The adjustments made by FAIRplexica relate to these areas (cf. to the commit history of UB Mannheim 2025b for further details):

1. Restricting the web search with SearXNG to RDM-specific sources by appending a customisable global context variable to each query, ensuring domain-relevant results.
2. Implementation of an admin dashboard to make settings for LLMs and APIs in a password-protected admin area. The changes introduce a credential-based authentication with NextAuth (an open-source authentication solution for Next.js applications), a new "admin" role as well as several adaptations to the UI. These changes allow the instance to be used directly as an easy-to-use pubic chatbot without the need for a new wrapper.
3. UI customizations to the CI/CD of the research data center.

5 Conclusions

The search for open-source alternatives to FAIR GPT resulted in three tools freely available on GitHub that could be adapted for the purposes of an AI assistant for RDM. Perplexica proved to be the most promising tool and was therefore developed further under the name of FAIRplexica.

FAIRplexica eliminates numerous disadvantages of FAIR GPT: it enables the use of an AI assistant whose components are completely open-source, uses and cites specific RDM sources from the Internet to counteract hallucinations and enable the verifiability of sources and (when using local LLMs, e.g. in a university network) also prevents data protection problems.

Despite these promising results, some limitations remain. Running FAIRplexica on a server with local LLMs requires a suitable, often costly GPU. Despite these technical limitations a next step should focus on evaluating FAIRplexica to assess the quality of its generated responses. Identifying and testing suitable evaluation procedures (as well as relevant data sets) are therefore necessary tasks. Frameworks such as DeepEval (Confident AI 2025) may provide a foundation for the evaluation, but other evaluation strategies will be considered as well. We anticipate that a thorough evaluation will generate added insights and value into LLM driven assistants for research data management.

Acknowledgements

We wish to thank the anonymous peer reviewers for their valuable feedback. The project on which this report is based was partially funded by the Federal Ministry of Education and Research (BMBF) under the funding code 13IHS264B TransforMA.

Authorship Contributions

- Thomas Schmidt: Writing – original draft
- Renat Shigapov: Writing – review & editing
- Jan Kamlah: Writing – review & editing
- Irene Schumm: Writing – review & editing

Conflict of Interest

The authors are not aware of any potential conflicts of interest.

Bibliography

Confident AI. 2025. *DeepEval – The LLM Evaluation Framework.* GitHub Repository. Visited on May 26, 2025. https://github.com/confident-ai/deepeval.

Fecher, Benedikt, Marcel Hebing, Melissa Laufer, Jörg Pohle, and Fabian Sofsky. 2023. "Friend or Foe? Exploring the Implications of Large Language Models on the Science System". *arXiv,* https://doi.org/10.48550/ARXIV.2306.09928.

Hartmann, Niklas K., Boris Jacob, and Nadin Weiß. 2019. *RISE-DE – Referenzmodell für Strategieprozesse im institutionellen Forschungsdatenmanagement.* Zenodo. https://doi.org/10.5281/zenodo.3585556.

Institut Français de Bioinformatique. 2025. *FAIR-Checker: Assessing FAIR Principles of Web Resources.* Visited on March 23, 2025. https://fair-checker.france-bioinformatique.fr/.

ItzCrazyKns. 2025. *Perplexica: An AI-Powered Search Engine.* GitHub Repository. Visited on March 23, 2025. https://github.com/ItzCrazyKns/Perplexica.

jjleng. 2024. *Sensei: An Open-Source AI-Powered Answer Engine.* GitHub Repository. Visited on March 23, 2025. https://github.com/jjleng/sensei.

MaastrichtU-IDS. 2025. *FAIR Enough API: Evaluating FAIRness of Digital Resources.* Visited on March 23, 2025. https://api.fair-enough.semanticscience.org/.

OpenAI. 2023. *Introducing GPTs.* Visited on March 23, 2025. https://openai.com/index/introducing-gpts/.

Rans, Jonathan, and Angus Whyte. 2017. *Using RISE, the Research Infrastructure Self-Evaluation Framework.* Digital Curation Centre. Visited on March 23, 2025. http://www.dcc.ac.uk/sites/default/files/documents/publications/UsingRISE_v1_1.pdf.

rashadphz. 2024. *Farfalle: AI-Powered Search Engine.* GitHub Repository. Visited on March 23, 2025. https://github.com/rashadphz/farfalle.

re3data.org. 2013. *Registry of Research Data Repositories.* https://doi.org/10.17616/R3D.

SearXNG. 2025. *SearXNG: A Free Internet Metasearch Engine.* GitHub Repository. Visited on March 23, 2025. https://github.com/searxng/searxng..

Shigapov, Renat, and Irene Schumm. 2024. “FAIR GPT: A virtual consultant for research data management in ChatGPT”. *arXiv,* https://doi.org/10.48550/ARXIV.2410.07108.

TIB – Leibniz Information Centre for Science and Technology. 2025. *TIB Terminology Service: Access to Multidisciplinary Ontologies.* Visited on March 23, 2025. https://terminology.tib.eu/ts.

UB Mannheim. 2024a. *Awesome RDM: A Curated List of Research Data Management Resources.* GitHub Repository. Visited on March 23, 2025. https://github.com/UB-Mannheim/awesome-RDM.

———. 2024b. *FAIR-sensei: An Open-Source Perplexity Analogue for Research Data Management.* GitHub Repository. Visited on March 23, 2025. https://github.com/UB-Mannheim/FAIR-sensei.

———. 2025a. *FAIR-farfalle: An Open-Source AI-Powered Search Engine for Research Data Management.* GitHub Repository. Visited on March 23, 2025. https://github.com/UB-Mannheim/FAIR-farfalle.

———. 2025b. *FAIRplexica: An Open-Source AI Assistant for Research Data Management.* GitHub Repository. Visited on March 23, 2025. https://github.com/UB-Mannheim/FAIRplexica.

Licensing Research Software Made Easy: Introducing the License Checker Tool

Dorothea Iglezakis [1,2,*], Iqra Imran [1,2]

[1] Research Data Competence Center, University of Stuttgart, Stuttgart, Germany;
[2] University Library, University of Stuttgart, Stuttgart, Germany;
[*] Corresponding author: fokus@izus.uni-stuttgart.de

Developing research software involves decisions about software licenses. These decisions can get complicated, especially if the code to be licensed builds on already licensed components. The challenge is to select a license that both aligns with the project's goals and remains compatible with predefined licenses.

The *License Checker* helps to find an appropriate license by providing an overview of common open-source software licenses, by analysing existing code for license information and by providing information about the compatibility of these licenses. It builds on a license ontology and offers a webbased interface. Target groups are research software engineers that have a basic understanding about software licenses in general but no deeper knowledge about license compatibility. It can be used to check the compatibility of an arbitrary set of licenses, but does not handle licenses for other research artefacts than software. It offers functionality to extract licenses from existing code and analyses dependency files, but so far only for Python and Javascript. The License Checker is also not handling complex licensing situations and does not distinguish between linking against a dependency and integrating a dependency into the code for the compatibility check.

Keywords: Research Software, Open-Source Licenses, Software Licenses, License Ontology

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23939 (CC BY-SA 4.0).

1 Introduction

A license serves as a clear way to state what is allowed with a piece of software by defining permissions, conditions and restrictions associated with code usage. However, it is also a legal document, and few research software engineers feel confident interpreting legal texts (e.g., Almeida et al. 2017). Moreover, developers frequently integrate external libraries or build on code with predefined licenses to enhance the functionality of their code. Then, the challenge is to find a license that aligns with the project's goals and is compatible with the predefined licenses at the same time.

2 The License Checker

The *License Checker*[1] is a webbased tool, that helps researchers to choose an appropriate license for their research software.

2.1 License Search

The tool offers an overview over the most established open-source licenses to compare and choose the licenses that fit to the intended usage of the software. An overview page (see Figure 1) displays the license type, conditions, permissions, and limitations of each software license and allows to identify compatible licenses to any chosen set of licenses.

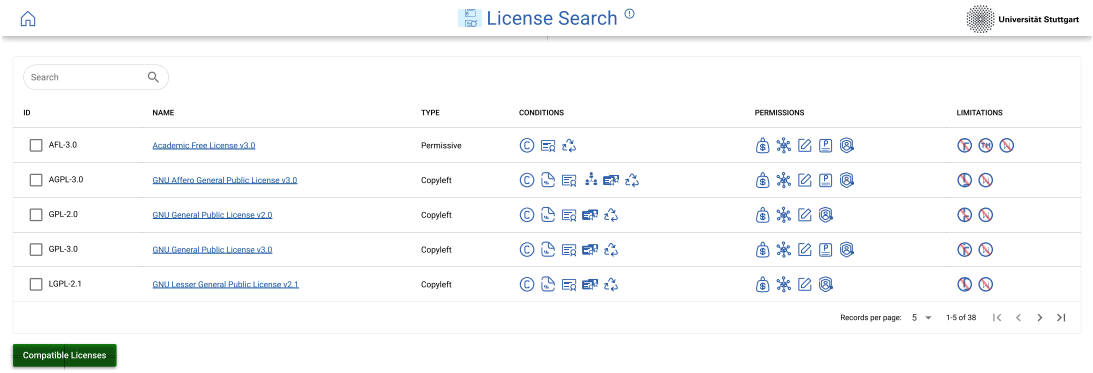


Figure 1: License Overview.

A detail page provides for each license additionally compatible licenses and the license text (see Figure 2).

1 https://checklicense.uni-stuttgart.de; *Visited on March 28, 2025.*

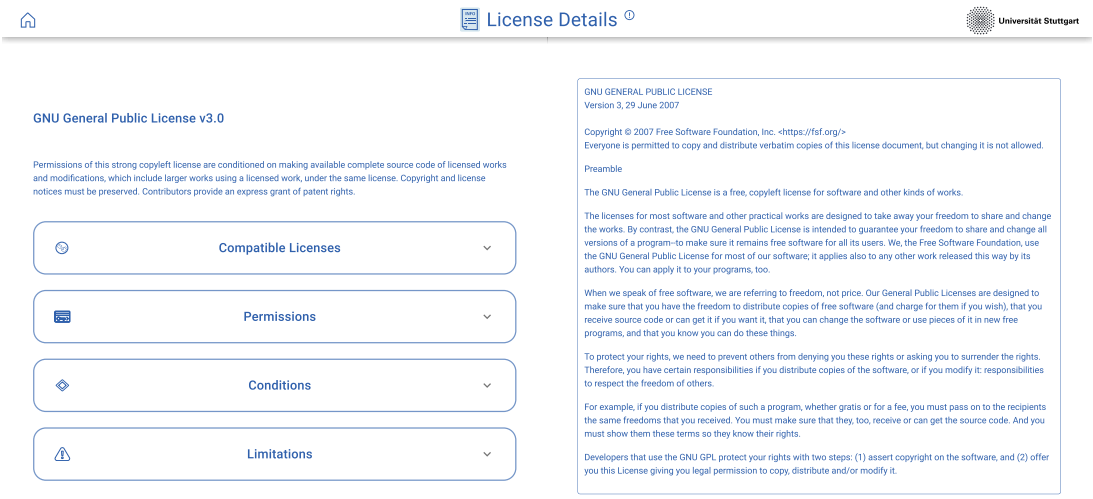


Figure 2: License Details.

2.2 License Recommendation

The *License Checker* also assists users in finding compatible licenses, allowing them to seamlessly incorporate external pieces of software codes without risking licensing conflicts. Existing code can be uploaded to the tool (see Figure 3) to be scanned for available license files with the help of fossology, an open source license compliance software system and toolkit[2].

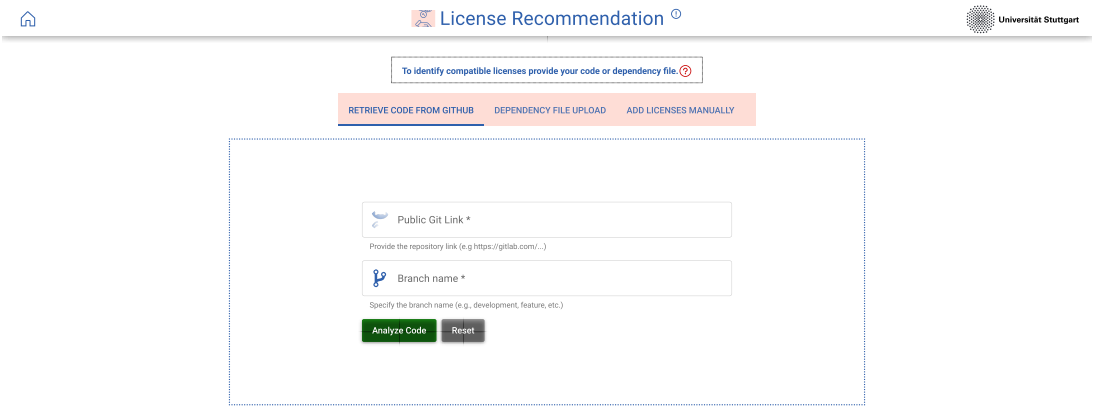


Figure 3: Upload Code.

By uploading a dependency file, the licenses of additional dependencies can be added to the list of licenses that should be taken into account (see Figure 4).

Based on a license ontology, the *License Checker* then takes this list of used licenses as input and provides a list of licenses as output that contains only licenses compatible with each of the input licenses (see Figure 5).

2 https://www.fossology.org/; *Visited on March 29, 2025.*

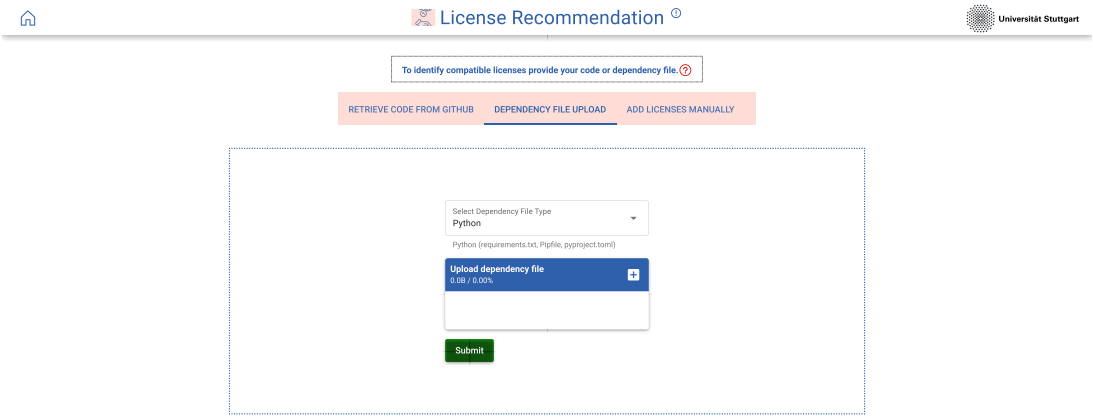


Figure 4: Upload Dependency File.

List of Compatible Licenses with AFL-3.0 ARTISTIC-2.0

Search

ID	NAME	TYPE
AFL-3.0	Academic Free License v3.0	Permissive
APACHE-2.0	Apache License 2.0	Permissive
ARTISTIC-2.0	Artistic License 2.0	Permissive
BSD-2-CLAUSE	BSD 2-Clause "Simplified" License	Permissive
BSD-3-CLAUSE	BSD 3-Clause "New" or "Revised" License	Permissive

Figure 5: Compatibility Check Result.

2.3 The License Ontology

The license ontology models common open source licenses and license types together with their compatibility. Figure 6 visualizes the main classes and properties. The ontology builds on the SPDX license list[3] and the appendix of choosealicense.com[4].

Each software license has a type and a set of permissions, conditions and limitations. Compatibility between licenses is defined between license types (e.g, Permissive isForwardCompatibleWith Copyleft) for the general rules and between licenses for specific exceptions (e.g., EUPL-1.1 isExplicitlyCompatibleWith BSL-1.0).

3 https://github.com/spdx/license-list-data; *Visited on March 28, 2025.*
4 https://choosealicense.com/appendix/; *Visited on March 28, 2025.*

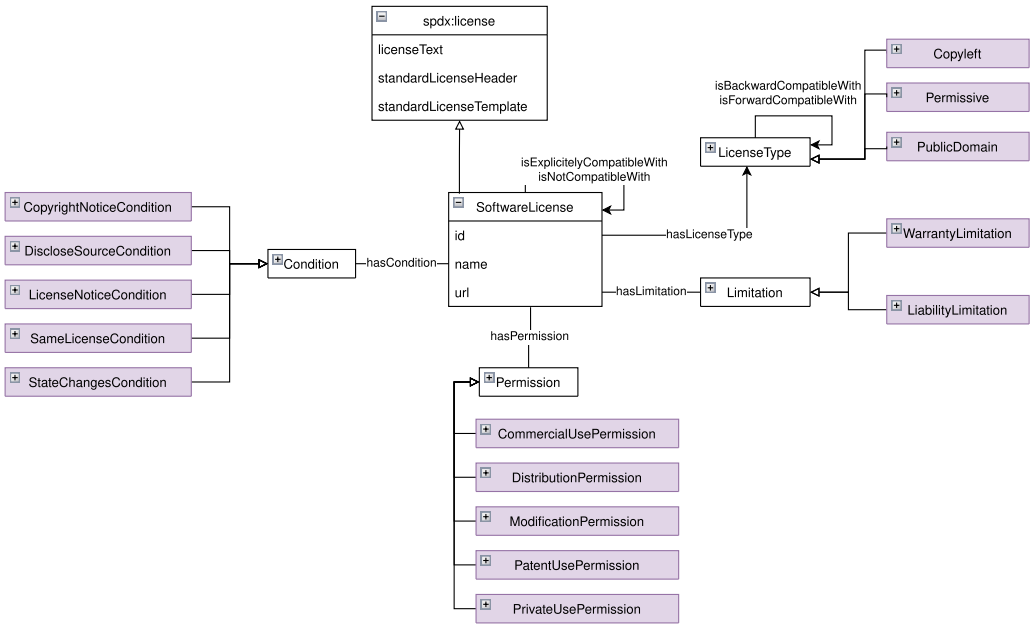


Figure 6: License Ontology.

2.4 Architecture

The license checker consists of several components, depicted in Figure 7. All components are available as docker containers[5] and can be deployed using docker compose. The *License Checker* frontend is a single page application using the frameworks VueJS[6] and Quasar[7]. The license engine uses the fossology toolkit to extract licensing information from existing code and provides the license texts. The *License Checker* backend, a REST API written in Python, uses the license ontology to provide information about software licenses and their compatibility.

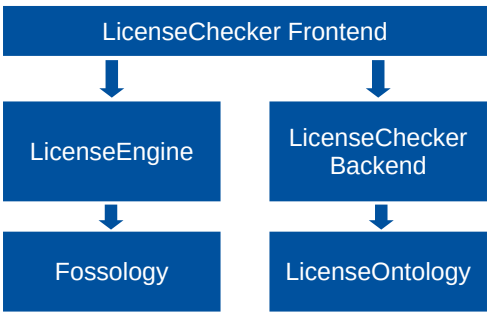


Figure 7: Components of the License Checker.

5 https://github.com/izus-fokus/LicenseChecker; *Visited on March 28, 2025.*
6 https://vuejs.org/; *Visited on March 31, 2025.*
7 https://quasar.dev/; *Visited on March 31, 2025.*

3 Outlook

The license checker is actually in a usable protoype stage and available for testing. It can not replace legal advice and does also not provide legal certainty, but offers a possibility to get aware of possible license conflicts. Further development is planned within the national research data infrastructure NFDI in a newly founded working group "Licensing & Agreement Rule Automation (LARA)" of the NFDI section ELSA for ethical, legal and social aspects[8] by checking the legal soundness of the license ontology and integrating the tool into a framework to support researchers to license not only research software, but also research data.

Acknowledgements

The development of the *License Checker* Tool was funded by the German Research Foundation (DFG) within the ReSUS project with the grant number 425911815.

In addition to the authors, other people were involved in the development of the license checker components. Evgenia Yangulova and Saskia Holland conducted most of the research on the license ontology. Markus Hirsch worked on the license ontology and the licence checker backend, while Michael Zimmermann developed the license engine. Faisal Khan implemented main parts of the license checker backend. Usman Ghani Amin worked on the backend and on the frontend of the license checker. Iqra Imran did most of the work in designing and implementing the license checker frontend.

Authorship Contributions

Dorothea Iglezakis drafted and wrote the first version. Iqra Imran commmented and improved.

Conflict of Interest

The authors declare no conflicts of interest

8 https://www.nfdi.de/section-elsa/; *Visited on March 28, 2025.*

Bibliography

Almeida, Daniel A., Gail C. Murphy, Greg Wilson, and Mike Hoye. 2017. "Do Software Developers Understand Open Source Licenses?" In *2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC),* 1–11. https://doi.org/10.1109/ICPC.2017.7.

Die GesundFDM Spring School: Praxisnaher Erwerb von Datenkompetenzen für Studierende

Katharina Koch [1,*], Stephanie Werner [2], Robert Werth [3], Nadine Probol [2], Eike Quilling [1]

[1] Hochschule Bochum, Bochum, Deutschland
[2] Hochschule Darmstadt, Darmstadt, Deutschland
[3] Frankfurt University of Applied Sciences, Frankfurt am Main, Deutschland
[*] Kontakt: katharina.koch@hs-bochum.de

Mit der fortschreitenden Digitalisierung steigen die Anforderungen an das Forschungsdatenmanagement (FDM), und Datenkompetenz avanciert zu einer Schlüsselqualifikation innerhalb und außerhalb der Hochschulen. Für Studierende fehlen bislang systematische Lehrangebote in diesem Bereich. Das BMBF-geförderte Projekt „Angewandtes Forschungsdatenmanagement für die gesundheits- und pflegebezogenen Wissenschaften“ (GesundFDM) begegnet diesem Mangel mit einer Spring School, die praxisnahe, fachspezifische und wiederverwendbare Lehrmaterialien bereitstellt – speziell für Studierende ohne Vorkenntnisse. Ziel ist es, eine frühzeitige und reflektierte Auseinandersetzung mit FDM-Themen zu fördern und nachhaltige Datenpraktiken zu etablieren. Grundlage der Lerninhalte bildet die „Lernzielmatrix zum Thema Forschungsdatenmanagement“ (Petersen u. a. 2023), ergänzt durch fachspezifische Beispiele. Ein zentrales didaktisches Element bilden fiktive Personae, die Studierende gesundheits- und pflegebezogener Studiengänge mit unterschiedlichen Bedarfen und Herausforderungen repräsentieren und der Veranschaulichung abstrakter Inhalte dienen.

Keywords: Datenkompetenz, Data Literacy, Forschungsdatenmanagement, Studierende, Gesundheits- und pflegebezogene Wissenschaften

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23940 (CC BY-SA 4.0).

1 Einleitung

Angesichts der fortschreitenden Digitalisierung und steigender Datenmengen wachsen die Anforderungen an das Forschungsdatenmanagement (FDM). Vor diesem Hintergrund entwickelt sich Datenkompetenz zu einer Schlüsselqualifikation für Forschende, Personen in den Unterstützungsstrukturen, aber auch für Studierende. Definiert als die Fähigkeit, Daten reflektiert zu sammeln, zu managen, zu bewerten und anzuwenden (Ridsdale u. a. 2015), bildet Datenkompetenz die Grundlage für fundiertes FDM und damit für effiziente Forschung. In den gesundheits- und pflegebezogenen Wissenschaften, wie z. B. der Pflegewissenschaft oder der Physiotherapiewissenschaft, ist diese jedoch vergleichsweise gering ausgeprägt (Klocke u. a. 2023). Neben der Beratung und Begleitung von Forschenden durch geschultes FDM-Personal ist es essenziell, Studierende – und damit potenzielle Nachwuchsforschende – an FDM-Themen heranzuführen und so frühzeitig ihre Datenkompetenz zu stärken. Obwohl Studierende im Rahmen von Praxisprojekten und Abschlussarbeiten Forschungsdaten erheben, aufbereiten und analysieren, ist die Vermittlung entsprechender Kompetenzen noch nicht explizit in die Lehrpläne der Studiengänge integriert. Vielmehr werden diese Themen in sehr unterschiedlichem Maße in der Lehre behandelt und kaum systematisch vermittelt. Einzelne Aspekte werden u. a. in Modulen zum wissenschaftlichen Arbeiten angesprochen. Bestehende außercurriculare Weiterbildungsangebote zu FDM richten sich selten an Studierende (Mertzen u. a. 2023) und sind bislang generisch ausgerichtet (z. B. der Zertifikatskurs „Forschungsdatenmanagement für Studierende“ der Landesinitiative für Forschungsdatenmanagement in Brandenburg).

Hier setzt das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Projekt „Angewandtes Forschungsdatenmanagement für die gesundheits- und pflegebezogenen Wissenschaften“ (GesundFDM, 2022–2025) an und hat eine Spring School entwickelt, die darauf abzielt, die Datenkompetenz von Studierenden der gesundheits- und pflegebezogenen Wissenschaften frühzeitig zu fördern. Die GesundFDM Spring School ist als außercurriculares Angebot konzipiert und stellt nachnutzbare, fachspezifische sowie praxisorientierte Lehr- und Lernmaterialien bereit. Diese können sowohl in digitalen Lernumgebungen und im Eigenstudium als auch in Präsenzveranstaltungen eingesetzt werden. Kennzeichnend ist die fachspezifische Ausrichtung, die darauf abzielt, FDM-Inhalte für die Zielgruppe der Studierenden anwendungsbezogen zu vermitteln. Dies ermöglicht Besonderheiten der einzelnen Disziplinen zu thematisieren, z. B. den Umgang mit sensiblen personenbezogenen Gesundheitsdaten im Forschungsprozess. Das vorgestellte Lehrkonzept fördert die frühzeitige und anwendungsbezogene Heranführung von Nachwuchsforschenden an Themen des FDM und befähigt sie zu einem nachhaltigen und reflektierten Umgang mit eigenen oder nachgenutzten Daten, u. a. im Rahmen von Abschlussarbeiten. Der frühzeitige Kontakt mit FDM-Themen trägt dazu bei, dass diese zu einem selbstverständlichen Bestandteil der eigenen wissenschaftlichen Praxis werden. Dieses Wissen ist auf späteren Karrierestufen – innerhalb wie außerhalb der Wissenschaft – von großem Wert.

2 Das Konzept der GesundFDM Spring School

Die GesundFDM Spring School richtet sich an Studierende auf Bachelor- und Masterniveau ohne Vorkenntnisse im FDM. Erwartet werden jedoch grundlegende Kenntnisse im wissenschaftlichen Arbeiten und in den jeweils üblichen Forschungsmethoden der Fachdisziplinen.

Inhaltsentwicklung mit System: Lernziele und Praxisbezug

Ausgangspunkt der inhaltlichen Konzeption der GesundFDM Spring School war die „Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards (Version 2)“ (Petersen u. a. 2023). Diese Lernzielmatrix bündelt für das FDM relevante generische Vermittlungsinhalte und Lernziele für die Qualifikationsstufen Bachelor, Master, PhD und Data Steward. Für die Konzeption der Spring School lag der Fokus auf den Qualifikationsstufen der Bachelor- und Masterstudierenden. Für diese Zielgruppen wurden spezifische Lernziele und relevante Themenschwerpunkte identifiziert. In weiteren Schritten erfolgte eine fachspezifische Ausarbeitung dieser Inhalte. Dabei standen unter anderem Aspekte wie Datenschutz und Forschungsethik im Fokus. Die Inhalte wurden durch anwendungsorientierte Beispiele ergänzt, – darunter eigens mithilfe Künstlicher Intelligenz generierte, fiktive qualitative und quantitative Datensätze, die exemplarische Szenarien aus der Forschungspraxis der gesundheits- und pflegebezogenen Wissenschaften darstellen.

Die inhaltliche Struktur der GesundFDM Spring School

Die GesundFDM Spring School gliedert sich in fünf thematische Module, die jeweils einem Veranstaltungstag entsprechen. Konzipiert wurde sie als Blockveranstaltung in Präsenz, die pro Tag einen zeitlichen Rahmen von vier Stunden (inkl. 45 Minuten Mittagspause) umfasste. Bereits in den Modultiteln spiegelt sich das Ziel der GesundFDM Spring School wider, FDM-Inhalte in einer zielgruppengerechten Sprache zu vermitteln und das häufig sehr technische Vokabular zu entmystifizieren. Die fünf Module decken den gesamten Forschungsdatenlebenszyklus ab:

- Modul 1: Wozu das Ganze? – Grundlagen des Forschungsdatenmanagements
- Modul 2: Wer hat Recht? – Rechtliche und ethische Aspekte im Umgang mit Forschungsdaten in gesundheitsbezogenen Disziplinen
- Modul 3: Das Chaos beherrschen – Ordnung und Struktur für Forschungsdaten
- Modul 4: Kontext schafft Qualität – Sicherung der Nachvollziehbarkeit von Forschungsdaten durch Datendokumentation und Metadaten

- Modul 5: Jedes Ende kann ein Anfang sein – Daten publizieren, finden und nachnutzen

Modul 1 vermittelt grundlegende Prinzipien und Ziele des FDM und zeigt dessen Einbettung in den Forschungsprozess. Behandelt werden zentrale Konzepte, wie Open Science, die FAIR-Prinzipien, die Leitlinien guter wissenschaftlicher Praxis sowie ethische Aspekte im Umgang mit gesundheitsbezogenen Forschungsdaten und deren praktische Anwendung. Zudem stehen Forschungsdatenpolicies sowie Beratungsangebote und Unterstützungsstrukturen zur Umsetzung des FDM im Fokus. Die Teilnehmenden erhalten eine Einführung in die Erstellung von Datenmanagementplänen mit dem Ziel, sie anhand praxisorientierter Leitfragen schrittweise zur eigenständigen Entwicklung eines eigenen Plans zu befähigen.

Modul 2 widmet sich rechtlichen und ethischen Aspekten im Umgang mit Forschungsdaten. Ein besonderer Fokus liegt auf dem Thema Datenschutz. Dies umfasst sowohl die korrekte Erstellung informierter Einwilligungen als auch den Schutz erhobener personenbezogener Daten durch Strategien der Anonymisierung und Pseudonymisierung. Übergeordnetes Ziel ist es, ein Verständnis für den datenschutzkonformen Umgang mit eigenen und nachgenutzten Daten zu vermitteln und die Teilnehmenden dazu zu befähigen, dies in ihren eigenen Projekten umzusetzen.

Modul 3 zielt darauf ab, Wissen darüber zu vermitteln, wie Forschungsdaten effizient verwaltet und nachvollziehbar abgelegt werden können. In diesem Rahmen werden die Themen Benennungskonventionen, Ordnerstrukturen und Versionskontrolle sowie die adäquate Verwendung von Datentypen und -formaten behandelt. Weiterhin finden Aspekte der Datenspeicherung und Datensicherheit Berücksichtigung. Die Teilnehmenden lernen Eignung und Risiken einzelner Maßnahmen kennen und erstellen für ein Beispielprojekt eine Backup-Strategie.

Modul 4 rückt Aspekte der Datendokumentation in den Fokus. Vermittelt werden grundlegende Kenntnisse zu gängigen Dokumentationsformen, um die Teilnehmenden zur eigenständigen Anwendung in der Forschungspraxis zu befähigen. Anhand von Metadaten, Metadatenstandards und Terminologien erlernen sie Strategien zur Beschreibung von Daten, um diese auffindbar und nachvollziehbar zur Nachnutzung bereitzustellen.

Modul 5 widmet sich den Themen Publikation und Nachnutzung von Forschungsdaten. Dies beinhaltet die Nutzung von Repositorien, sowohl generischer als auch fachspezifischer Art. Die Teilnehmenden lernen Strategien zur Recherche und Auswahl geeigneter Repositorien kennen und üben, gezielt nach nachnutzbaren Datensätzen entlang ihrer Forschungsinteressen zu recherchieren. Ergänzend erhalten sie einen Überblick über die rechtlichen Rahmenbedingungen, die eine ordnungsgemäße und rechtskonforme Nachnutzung von Daten gewährleisten.

Die Spring School zeichnet sich durch einen steten Wechsel von Theorie und Praxis aus, indem die Teilnehmenden das neu erlangte Wissen zu einem Thema direkt auf fachbezogene Fallbeispiele anwenden, bevor sich der nächste theoretische Input anschließt. Die praktischen Einheiten sind als Einzel- oder Kleingruppenarbeiten konzipiert. Sie ermöglichen eine gezielte Überprüfung des Lernfortschritts und bieten Raum zur unmittelbaren Klärung aufkommender Fragen.

Mit fiktiven Charakteren zu echtem Verständnis: Der Einsatz von Personae im FDM-Kontext

Ein zentrales Alleinstellungsmerkmal und zugleich eine didaktische Besonderheit der GesundFDM Spring School ist der Einsatz von Personae zur Wissensvermittlung. Dieses aus dem Marketingbereich stammende Instrument dient der Darstellung von Verhaltens- und Denkweisen, Zielen und Motivationen von Nutzendengruppen anhand fiktiver, deskriptiver Modelle. Ziel ist es, Nutzende auf persönlicher und emotionaler Ebene zu verstehen und anzusprechen (Urbaum und Werner 2024). Im Kontext der GesundFDM Spring School dienen die Personae dazu, spezifische Bedarfe und Herausforderungen der einzelnen Fachgebiete in den gesundheits- und pflegebezogenen Wissenschaften in Bezug auf FDM zu verdeutlichen. Konzipiert als Studierende gesundheits- und pflegebezogener Studiengänge, bieten diese fiktiven Charaktere anhand ihrer Bedarfe, Probleme und Fragen einen anwendungs- und lebensweltbezogenen Zugang, um die Anforderungen, Herausforderungen und Vorteile des FDM kennenzulernen. Die Personae dienen somit der Veranschaulichung abstrakter Inhalte und ermöglichen es den Teilnehmenden, sich mit diesen zu identifizieren und so den Transfer auf persönliche Hürden im Umgang mit Forschungsdaten anzuregen. Hieraus ergeben sich konkrete Anhaltspunkte, in welcher Weise die Teilnehmenden die erworbenen Kompetenzen in ihrem (Forschungs-)Alltag nutzen können. Insgesamt wurden fünf Personae entwickelt, die die Disziplinen Gesundheitswissenschaften (mit Schwerpunkt auf Ökonomie und Management), Angewandte Therapiewissenschaften, Pflegewissenschaften, Hebammenwissenschaft und Gesundheitsbezogene Arbeit repräsentieren. Ihre Beschreibung beinhaltet Angaben zur Ausbildung und zum Studium, zu beruflichen und wissenschaftlichen Zielen sowie zu individuellen Charakteristika und Interessen. In der didaktischen Praxis orientiert sich die Auswahl der Personae an den fachlichen Hintergründen der Teilnehmenden und kann in homogenen Gruppen begrenzt werden. Zu Beginn der Spring School werden die Personae vorgestellt und begleiten anschließend alle Module als durchgängiges didaktisches Element. Sie kommen insbesondere im Rahmen praxisorientierter Anwendungsfälle zum Einsatz, etwa als Akteur*innen im Format der Scary Tales, das potenzielle Konsequenzen fehlerhaften Datenmanagements exemplarisch veranschaulicht. Zudem werden am Beispiel der Forschungsinteressen der Personae spezifische qualitative und quantitative Forschungsdesigns eingeführt, zu denen die Teilnehmenden im Verlauf der Spring School Datenmanagementpläne erstellen.

3 Eindrücke aus der Pilotierung

Die GesundFDM Spring School wurde im März 2025 als Präsenzveranstaltung auf dem Gesundheitscampus der Hochschule Bochum pilotiert. Ziel der Pilotierung war es, das didaktische Material sowie die Inhalte in der Praxis zu erproben und aus der Perspektive der Studierenden gezielte Hinweise für die Weiterentwicklung zu gewinnen. Unter anderem galt es zu evaluieren, inwiefern die Inhalte die tatsächlichen Bedarfe der Studierenden dieser Fachdisziplinen widerspiegeln. Das Evaluationskonzept umfasste eine Kombination aus (1) einer Erwartungsabfrage zu Beginn der Veranstaltung, (2) Tagesevaluationen sowie (3) einer übergreifenden Abschlussevaluation des Gesamtkonzepts. Unter den elf Teilnehmenden befanden sich überwiegend fortgeschrittene Masterstudierende sowie zwei angehende Promovierende. Viele brachten aufgrund ihres Vorwissens konkrete Fragen aus ihrem Forschungsalltag mit. Fachlich stammten die Teilnehmenden überwiegend aus den Gesundheitswissenschaften; vertreten waren zudem Pflegewissenschaft, Hebammenwissenschaft und Gesundheitsmanagement. Das fachspezifische und anwendungsorientierte Konzept der GesundFDM Spring School hat sich insgesamt bewährt. Die Teilnehmenden zeigten großes Interesse an den Praxiselementen. Diese erforderten eine intensive Betreuung durch die Dozierenden. Positiv bewertet wurden auch die Erstellung von Datenmanagementplänen sowie die Fallbeispiele und Scary Tales, die das Verständnis förderten. Die Nutzung der Personae wurde gelobt, allerdings zeigte sich der Bedarf an einer ausführlichen Einführung und regelmäßigen Wiederholungen zur Sicherung des Wiedererkennungswerts. Alle Inhalte wurden als relevant eingeschätzt, insbesondere die Themen Datenschutz, Ethik und Recht, was deren hohe Bedeutung in den gesundheits- und pflegebezogenen Wissenschaften unterstreicht.

4 Ausblick

Die Materialien der GesundFDM Spring School werden im Sommer 2025 als Open Educational Resources (OER) in der GesundFDM-Zenodo-Community[1] veröffentlicht. Dies umfasst das Konzept der GesundFDM Spring School, das Konzept des GesundFDM Trainer Tandems, das Konzept zur Verwendung der Personae, das Modulhandbuch, eine Sammlung selbst erstellter Grafiken zu FDM-Themen sowie alle verwendeten Lehr- und Lernmaterialien inklusive Drehbüchern und Foliensätzen. Durch die Veröffentlichung dieser Materialien erhalten Lehrende in den Disziplinen der gesundheits- und pflegebezogenen Wissenschaften einsatzbereite, kommentierte Materialien, die sie mit geringem Eigenaufwand, auch auszugsweise, in ihre Lehre integrieren können. Darüber hinaus existieren vielfältige Visionen zur Weiterentwicklung dieses vielversprechenden Konzepts: Es bestehen Überlegungen, die Inhalte zu digitalisieren und als Lehrvideos zur Nachnutzung für Studierende und Lehrende verfügbar zu machen. Zudem können die Lehrmaterialien

1 https://zenodo.org/communities/gesund-fdm/records?q=&l=list&p=1&s=10&sort=newest; *Besucht am 27. Mai 2025.*

weiter ausdifferenziert und stärker auf studentische Zielgruppen mit unterschiedlichem Vorwissen zugeschnitten werden. Beispielsweise ist ein Angebot speziell für Masterstudierende denkbar, die kurz vor ihrer Abschlussarbeit stehen. Darüber hinaus können die bestehenden Materialien in Kleinsteinheiten für kurze Impulse in der Lehre aufbereitet werden, um FDM bestmöglich an alle Studierenden heranzutragen. Zusammenfassend ist festzuhalten, dass im Rahmen der GesundFDM Spring School ausgeprägtes Interesse und ein erheblicher Bedarf an FDM-Wissen insbesondere bei fortgeschrittenen Studierenden deutlich wurden, welche bislang nicht in ausreichendem Maße durch Angebote im Rahmen der curricularen wie auch der extracurricularen Lehre aufgegriffen werden. Vor diesem Hintergrund leistet die GesundFDM Spring School mit ihren nachnutzbaren Materialien einen wichtigen Beitrag zur Förderung der Datenkompetenz von Studierenden in den gesundheits- und pflegebezogenen Wissenschaften.

Danksagungen

Die GesundFDM Spring School wurde im Rahmen des BMBF-geförderten Projekts „Angewandtes Forschungsdatenmanagement für die gesundheits- und pflegebezogenen Wissenschaften“ (GesundFDM) durchgeführt.

Beiträge

Die Autor*innen haben gleichermaßen zu allen Sektionen des Papers beigetragen.

Interessenkonflikte

Es sind uns keine Interessenkonflikte bekannt.

Literaturverzeichnis

Klocke, Andreas, Robert Werth, Arnela Balic und Meike Backes. 2023. *Entwicklung und Verbreitung von Forschungsdatenmanagement an Fachhochschulen und Hochschulen für Angewandte Wissenschaften (EVER_FDM): Schlussbericht zum Projekt EVER-_FDM.* Forschungszentrum Demografischer Wandel (FZDW), Frankfurt University of Applied Sciences, Frankfurt am Main. https://doi.org/10.2314/KXP:1877275336.

Mertzen, Daniela, Heike Neuroth, Carsten Schneemann und Kathrin Woywod. 2023. „Zertifikatskurs „Forschungsdatenmanagement für Studierende“ der Landesinitiative für Forschungsdatenmanagement in Brandenburg“, https://doi.org/10.18420/INF2023_08.

Petersen, Britta, Claudia Engelhardt, Tanja Hörner, Juliane Jacob, Tatiana Kvetnaya, Andreas Mühlichen, Hermann Schranzhofer u. a. 2023. *Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards.* Zenodo. https://doi.org/10.5281/ZENODO.8010617.

Ridsdale, Chantel, James Rothwell, Mike Smit, Michael Bliemel, Dean Irvine, Daniel Kelley, Stan Matwin, Brad Wuetherick und Hossam Ali-Hassan. 2015. *Strategies and Best Practices for Data Literacy Education Knowledge Synthesis Report.* https://doi.org/10.13140/RG.2.1.1922.5044.

Urbaum, Dorothee und Stephanie Werner. 2024. *Zielgruppenanalyse für Open Educational Resources (OER) – Eine Handreichung zur Erstellung von Personae.* Zenodo. https://doi.org/10.5281/ZENODO.13774050.

Modulare standortübergreifende Schulungsangebote: Das FDM-Curriculum der Universitätsallianz Ruhr

Julia Stapels [1], Tuba Güden-Silber [2], Jessica Stegemann [3]

[1] Forschungsdatenservice, TU Dortmund, Deutschland;
[2] Research Data Services, Ruhr-Universität Bochum, Deutschland;
[3] Research Data Services, Universität Duisburg-Essen, Deutschland

Im Zuge des Wandels der Wissenschaften nimmt das Forschungsdatenmanagement an Bedeutung zu und damit der Bedarf der Forschenden nach Schulungsangeboten. Um diesen Anforderungen bedarfsorientiert und ressourceneffizient zu begegnen, sind innovative Konzepte gefragt. Ein Beispiel dafür sind modulare standortübergreifende Schulungsangebote wie das FDM-Curriculum der Universitätsallianz (UA) Ruhr (Ruhr Universität Bochum, TU Dortmund, Universität Duisburg-Essen). Das FDM-Curriculum der UA Ruhr besteht aus einem Angebot an Grundlagenschulungen und ergänzenden Vertiefungsmodulen, die in der Regel online und teilweise in Präsenz, an den verschiedenen Standorten angeboten werden. Die Vertiefungsmodule ermöglichen es den Forschenden, sich intensiver mit spezifischen Themenbereichen auseinanderzusetzen, in Hands-on-Workshops den Umgang mit Tools auszuprobieren und praxisnahe Tipps für ihren Forschungsalltag zu erhalten. Bei Teilnahme an einer Grundlagenschulung und zwei Vertiefungsveranstaltungen wird ein FDM-Badge ausgestellt. Mit diesem Nachweis wird ein zusätzlicher Anreiz zur Teilnahme an den Schulungen geschaffen. Das Curriculum steht in dieser Form allen Forschenden der UA Ruhr seit dem Wintersemester 2024/25 offen. Der standortübergreifende Ansatz ermöglicht es, eine breite, sowohl fachbereichsspezifische als auch interdisziplinäre Auswahl für Forschende mit ca. 20 Veranstaltungen pro Semester zu gestalten und dabei die Ressourcen der beteiligten Hochschulen schonend sowie die vorhandenen Expertisen zielgerichtet einzusetzen. Die Erfahrungen bei der Konzeption und Umsetzung des Curriculums sollen in diesem Beitrag weitergegeben werden.

Keywords: FDM, RDM, curriculum, workshops, education

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23941 (CC BY-SA 4.0).

1 Das Schulungsangebot FDM-Curriculum der UA-Ruhr

Forschende an Hochschulen und Forschungseinrichtungen sehen sich zunehmend mit Herausforderungen im Bereich des Forschungsdatenmanagements (FDM) und den damit verbundenen Forderungen von beispielsweise Fördermittelgebern konfrontiert. Eine der zentralen Aufgaben dieser Institutionen ist es, ihre Forschenden bei der Bewältigung dieser Herausforderungen zu unterstützen. Ziel ist es, die Wissenschaftler:innen in die Lage zu versetzen, verantwortungsvoll und nachhaltig mit ihren Forschungsdaten umzugehen, fachliche Standards einzuhalten und ihre Daten sowie Erkenntnisse für die Wissenschaft und Gesellschaft bereitzustellen. Um ihre Forschenden mit ihren unterschiedlichen Bedürfnissen flächendeckend unterstützen zu können, haben viele Einrichtungen FDM-Servicestellen eingerichtet, die unter anderem Beratungen und Schulungen anbieten. Allerdings steht jede Einrichtung vor der Herausforderung begrenzter Personalressourcen, die die vielfältigen Bedarfe der Forschenden abdecken sollen.

Die Universitäten der Universitätsallianz Ruhr (UA Ruhr) haben die Notwendigkeit erkannt, Synergien zu nutzen und Ressourcen zu bündeln. Unter anderem haben sie ihr bestehendes Schulungsangebot im Bereich FDM zusammengeführt. Dabei ging es nicht nur um die gegenseitige Öffnung von Workshopangeboten. Um ein strukturiertes, ineinandergreifendes und der gestiegenen Nachfrage von Forschenden entsprechendes Angebot machen zu können, wurde ein neues gemeinsames Schulungskonzept entwickelt – das FDM-Curriculum der UA Ruhr. Dabei galt es, sowohl den Anforderungen von Sonderforschungsbereichen / Transregios und Graduiertenprogrammen als auch von Forschenden ohne Verankerung in strukturierten Programmen gerecht zu werden. Neben Grundlagen des FDM sollten auch fachspezifische Themen und Funktionen von Tools vermittelt werden. Daher braucht es ein Angebot, das zeitliche Flexibilität bietet und die Aneignung von Überblickswissen in gleichem Maße ermöglicht, wie eine individuelle Spezialisierung.

Vor diesem Hintergrund wurde an der Universität Duisburg-Essen (UDE) durch die Research Data Services (RDS) Anfang 2022 damit begonnen, das FDM-Curriculum als strukturiertes Schulungsprogramm zu konzipieren. Dabei konnte zwar auf nachnutzbare Schulungsmaterialien zurückgegriffen werden (siehe z. B. Forschungsdatenmanagement-Initiative HU Berlin 2025; Krähwinkel u. a. 2022), dennoch mussten sowohl auf inhaltlicher als auch didaktischer Ebene zielgruppen- und standortspezifische Anpassungen getätigt werden. Im Wintersemester 2022/2023 wurde das FDM-Curriculum erstmals an der UDE angeboten (siehe Leimer u. a. 2023; Leimer u. a. 2025). Seit dem Sommersemester 2024 wird das Angebot durch Schulungen der Technischen Universität Dortmund (TU Dortmund) ergänzt. Der Einstieg der Ruhr-Universität Bochum (RUB) im Wintersemester 2024/2025 war schließlich der Startschuss für das standortübergreifende FDM-Curriculum der UA Ruhr, das damit allen Forschenden der drei Standorte zur Verfügung steht.

2 Struktur des FDM-Curriculums

Als wiederkehrendes Schulungsangebot wird das FDM-Curriculum in jedem Semester angeboten. Mit dieser Regelmäßigkeit wird eine Planbarkeit sowohl für die durchführenden Trainer:innen, als auch für die interessierten Forschenden ermöglicht. Ein großes Maß an Flexibilität wird dabei nicht nur auf zeitlicher Ebene geschaffen, die Untergliederung in ein Basis- und ein Vertiefungsmodul schafft auch Raum für inhaltliche Ausgestaltungen (siehe Abbildung 1).

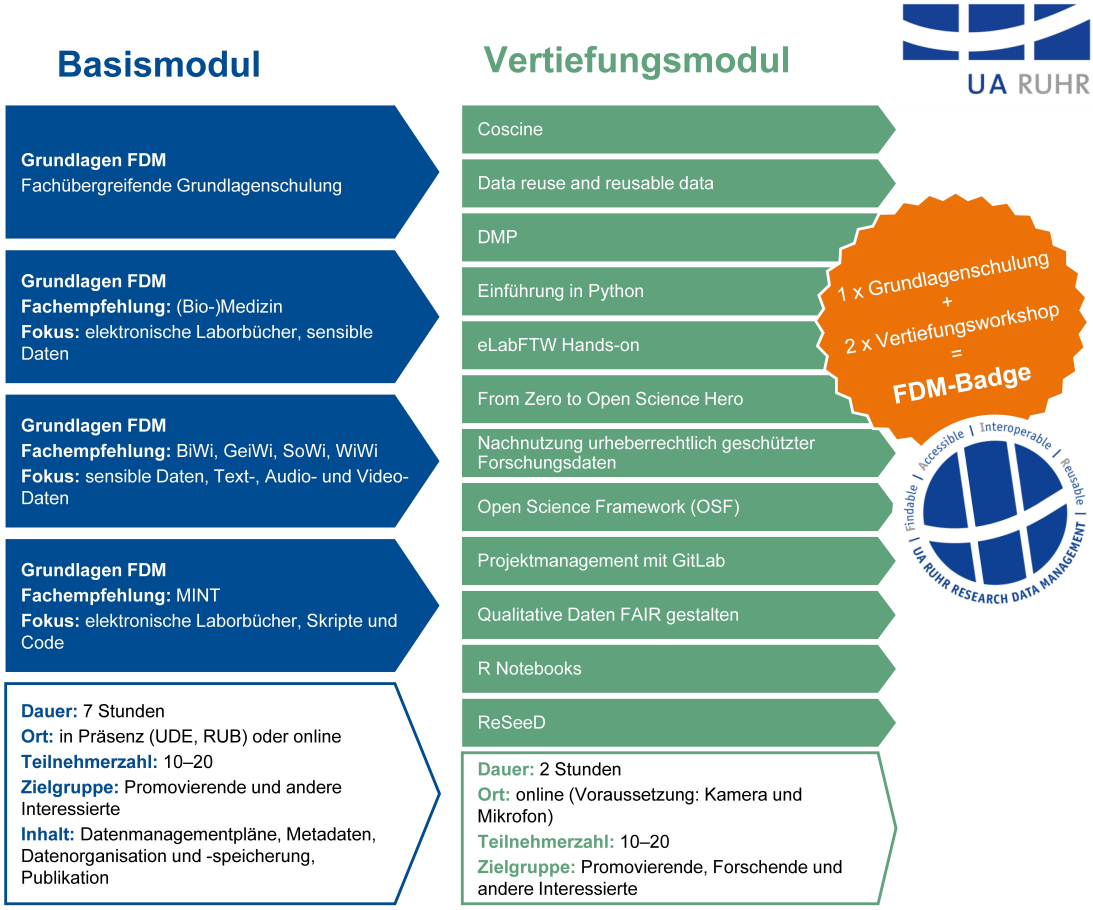


Abbildung 1: Modulübersicht FDM-Curriculum der UA Ruhr. Abbildung beruht auf Stapels, Güden-Silber und Stegemann (2025): FDM-Curriculum der UA Ruhr und Badge.

Im Basismodul haben die Forschenden die Möglichkeit, in einer 7-stündigen Schulung die Grundlagen des FDM kennenzulernen und einen Überblick über die relevanten Themenfelder zu erhalten. Das Grundkonzept der UDE sah zunächst drei verschiedene Grundlagenschulungen in Präsenz mit einer Fokussetzung auf bestimmte Datentypen – z. B.

sensible Daten mit der Fachempfehlung für bildungs-, geistes-, sozial- und wirtschaftswissenschaftliche Fächer – vor. Mit der Ergänzung durch Dortmund und Bochum konnte das Angebot auf inzwischen sieben Grundlagenschulungen erweitert werden, die nun teilweise auch online und ohne fachspezifischen Fokus angeboten werden. Im Vertiefungsmodul werden in 2-stündigen Workshops (i.d.R. online) verschiedene FDM-Themen vertiefend behandelt und Anwendungen trainiert. Das Angebot reicht von Hands-On-Workshops zu spezifischen FDM-Tools wie GitLab, eLabFTW, institutionellen Forschungsdatenrepositorien u. a. über Themen wie FAIR, Open Science und Datenmanagementpläne bis hin zu rechtlichen Fragestellungen wie der Nachnutzung urheberrechtlich geschützter Forschungsdaten. Durch die Kooperation der drei UA Ruhr-Standorte konnte das ursprünglich aus sieben Vertiefungsworkshops bestehende Angebot auf ein breiteres, attraktiveres Portfolio von zurzeit 12 Schulungen erweitert werden, aus dem die Forschenden die für sie aktuell relevantesten Themen auswählen können. Bei Vorlage der Teilnahmebescheinigungen für eine Grundlagenschulung und zwei Vertiefungsworkshops wird das „FDM-Badge“ ausgestellt. Das Badge dient als einfacher Nachweis grundständigen Wissens über FDM-Maßnahmen und -Tools. In der standortübergreifenden Weiterbildung gibt es natürlich aussagekräftigere Zertifikate über den Erwerb von FDM-Wissen (siehe Slowig u. a. 2022), diese dienen aber vornehmlich der tiefergehenden Ausbildung von zukünftigen Data Stewards, erfordern einen höheren Zeitaufwand und setzen Prüfungsleistungen voraus. Mit dem „FDM-Badge“ sollte ein niederschwelliger Anreiz zur Teilnahme an mehreren Veranstaltungen geschaffen werden, der zeitliche Kapazitäten von in Dissertationsvorhaben und Lehre eingespannten Promovierenden und arrivierten Forschenden berücksichtigt und praktische Anwendungen beim FDM stärker fokussiert als theoretische Wissensvermittlung. Das Badge beruht dabei auf einem Gamification-Ansatz als didaktischem Mittel (siehe Tolks und Sailer 2021), welcher die freiwillige Beschäftigung mit dem Thema FDM stimulieren soll.

3 Potentiale und Herausforderungen

Die standortübergreifende Ausrichtung des Curriculums bietet sowohl für die Teilnehmenden als auch für die Standorte zahlreiche Vorteile. Die Teilnehmenden profitieren von zusätzlichen Terminen für die Grundlagenschulungen, da nun auch ein verpasster Termin einfach noch im selben oder im folgenden Semester nachgeholt werden kann. Die bedarfsgerechte Ergänzung von Vertiefungsveranstaltungen hat zur Steigerung der Attraktivität des Programms beigetragen. Schließlich dient das Badge als „Auszeichnung“ über das erworbene FDM-Wissen, das den Teilnehmenden in ihrer zukünftigen Laufbahn zugutekommen kann.

Auch die drei Standorte profitieren von der standortübergreifenden Ausrichtung. Zum einen konnte das Schulungsprogramm zum FDM ressourcenschonend erweitert werden. Somit ist eine Spezialisierung der Standorte auf bestimmte Themen möglich, und nicht jeder Standort muss Schulungen in jedem Spezialbereich zur Verfügung stellen. Hierbei wur-

den die Vertiefungsschulungen auf Basis der Bedarfe der Forschenden (Kletke u. a. 2024), der Kompetenzen der Lehrenden und des Angebots an Tools ausgewählt und iterativ auf Basis des Feedbacks in den Schulungen angepasst. Zum anderen steigt der Bekanntheitsgrad des Angebots durch eine dreifache Verfügbarkeit der Informationskanäle an. Weiterhin bieten die gemeinsame Teilnehmendenverwaltung und Statistik einen einfachen Überblick sowie eine Erleichterung der Verwaltung im Gegensatz zum einzelnen Angebot. Schließlich kommt sowohl den Teilnehmenden als auch den Standorten zugute: Durch die Kooperation, die Verständigung auf gemeinsame Standards und den offenen Austausch von Schulungsmaterialien in einer kooperativ gepflegten überregionalen Hochschulcloud findet eine standortübergreifende Qualitätssicherung und ständige Verbesserung der Lehre sowie der Arbeitsabläufe statt. Durch die Arbeitsteilung innerhalb der Kooperation werden zudem Kapazitäten frei, um die Erfahrungen mit dem Curriculum in die FDM-Community einzubringen, sei es durch Vorstellung bei einschlägigen Tagungen oder auch durch die Mitarbeit in verschiedenen regionalen Arbeitsgruppen (z. B. der Landesinitiative für Forschungsdatenmanagement – fdm.nrw) und überregionalen Initiativen (z. B. in der Sektion EduTrain der Nationalen Forschungsdateninfrastruktur – NFDI). Somit wird sichergestellt, dass das Curriculum eine FDM-Weiterbildung auf qualitativ hohem Niveau bietet und sich ständig den sich weiterentwickelnden Bedarfen anpassen kann.

Trotz aller Vorteile bringt eine standortübergreifende Abstimmung natürlich auch organisatorische Herausforderungen. Initial wurden gemeinsame Arbeitsabläufe entwickelt, die Dokumentation eingerichtet, sowie die Erstellung von Teilnahmebescheinigungen und die Ausstellung der Badges abgestimmt. Informationen, die die verschiedenen Standorte betreffen, müssen fortlaufend aktualisiert werden und etwaige Änderungen zeitnah kommuniziert werden. Hierfür hat sich ein monatliches Treffen mit Curriculums-Verantwortlichen aller drei Standorte bewährt. Zusammengefasst überwiegen die Vorteile die organisatorischen Herausforderungen bei weitem und die Standorte konnten von der gemeinsamen Optimierung der Arbeitsabläufe profitieren.

4 Rezeption und Ausblick

Insgesamt sind schwankende, aber tendenziell steigende Teilnahmezahlen am Curriculum zu beobachten. Während im Sommersemester 2023 114 Teilnahmen verzeichnet werden konnten, waren es im Wintersemester 2023/24 78 Teilnahmen. Im Sommersemester 2024 waren es 98 und zum offiziellen Start des Regelbetriebes im Wintersemester 2024/25 205 Teilnahmen. In diesem Semester gab es von der UDE 108 Teilnahmen, von der RUB 56 Teilnahmen und von der TU Dortmund 41 Teilnahmen (siehe Abbildungen 2 und 3). Bisher (Stand: 17.03.2025) wurden 31 Badges ausgestellt, bei 495 Teilnahmen insgesamt.

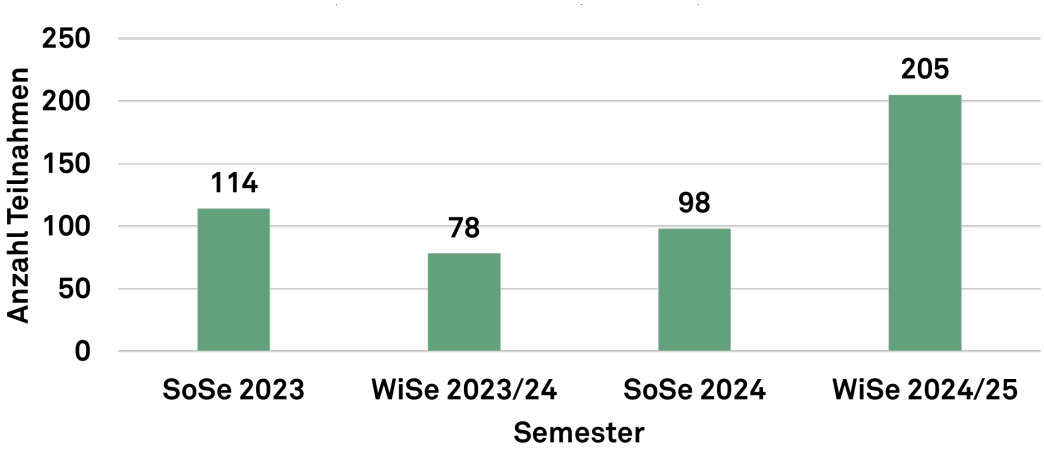


Abbildung 2: Teilnahmen gesamt (Stand: 17. März 2025, N = 495).

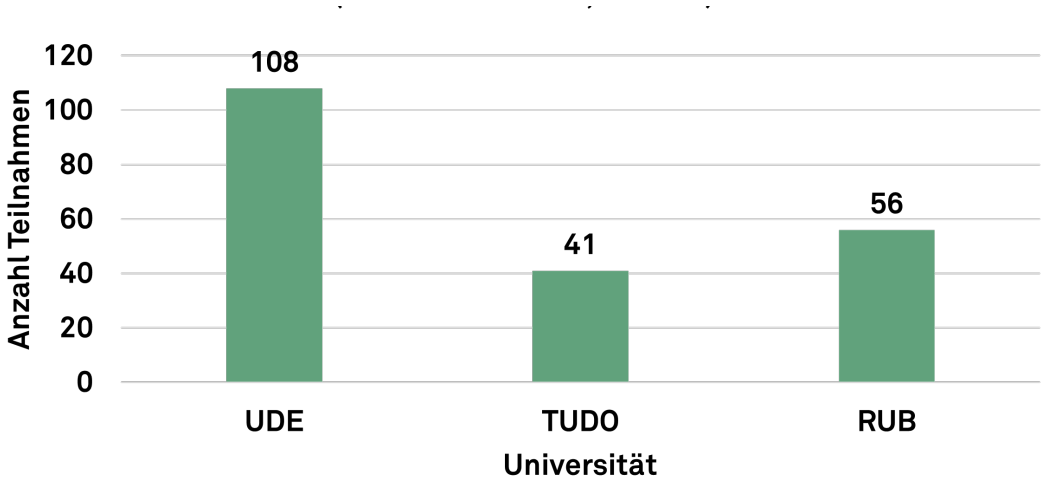


Abbildung 3: Teilnahmen im WiSe2024/25 bislang (Stand: 17. März 2025, N = 205).

Eine erste Evaluation der Schulungsangebote, die an der TU Dortmund durchgeführt wurden, zeigte eine positive Rezeption. Die Ausweitung der Evaluation wird zurzeit in der standortübergreifenden Arbeitsgruppe vorbereitet und soll ab dem Wintersemester 2025/2026 nicht nur Hinweise auf Verbesserungsbedarf in Bezug auf die Schulungsperformance liefern, sondern auch Informationsbedarfe der Teilnehmenden nach weiteren FDM-Themen und -Tools abfragen. Die Ergebnisse dienen auch dazu, die laufenden Aktualisierungen und den Ausbau des Angebots bedarfsgerecht zu gestalten. In der Diskussion ist zudem, einige der Schulungsmaterialien auch standortübergreifend als Open Educational Resources (OER) zu veröffentlichen. Auch hier ist aber ein Abwägen notwendig: Der Arbeitsaufwand bei der Erstellung von gut nachnutzbaren OER, die frei von standortspezifischen Bezügen sind und über eine bloße Veröffentlichung von Präsentationsfolien hinausgeht, muss einkalkuliert werden. Hinzu kommt, dass zwar speziell einige Hands-on-

Übungen aus den Vertiefungsmodulen neues, attraktives Material für eine Nachnutzung durch die FDM-Community bieten, sie drohen aber auch schnell zu veralten, sobald z. B. Änderungen in der Benutzungsoberfläche von Tools (z. B. des Open Science Framework oder von Elektronischen Laborbüchern) vorgenommen werden. Hier gilt es also, nicht nur für die Erstellung von OER sinnvolle Konzepte zu entwickeln, sondern auch die laufende Pflege standortübergreifend zu organisieren.

5 Fazit

Der standortübergreifende Ansatz des Curriculums ermöglicht ein breit aufgestelltes Schulungsangebot, sowohl fachbereichsspezifisch, als auch interdisziplinär. Die Teilnahme kann den Forschenden niederschwellig ermöglicht werden. Das Curriculum wird gut angenommen und die steigende Nachfrage nach Badges zeigt eine wachsende FDM-Awareness bei den Forschenden. Auf institutioneller Ebene stärkt der zielgerichtete Einsatz von Ressourcen und Expertise die Idee der UA Ruhr.

Danksagungen

Wir danken der AG Schulung und den Forschungsdatenservices der RUB, UDE und TU Dortmund für die Mitarbeit und Weiterentwicklung des Curriculums.

Beiträge

Alle Autorinnen haben am ersten Entwurf mitgearbeitet und den Text überarbeitet.

Interessenkonflikte

Keine.

Literaturverzeichnis

Forschungsdatenmanagement-Initiative HU Berlin. 2025. *Medienkollektion der Forschungsdatenmanagement-Initiative an der Humboldt-Universität zu Berlin.* Besucht am 10. Juni 2025. https://rs.cms.hu-berlin.de/dataman/pages/home.php.

Kletke, Olaf, Inga Larres, Kathrin Höhner, Wibke Kleina, Julia Stapels, Bernd Zey und Niels Kasties. 2024. „Bestands- und Bedarfserhebung im Forschungsdatenmanagement". *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB* 11 (2): 1–18. https://doi.org/10.5282/O-BIB/6014.

Krähwinkel, Esther, Patrick Langner, Robert Lipp und Andre Manuel Pietsch. 2022. *HeFDI Data OER: Forschungsdatenmanagement – eine Online-Einführung.* Zenodo. https://doi.org/10.5281/ZENODO.6373596.

Leimer, Sophia, Marlen Externbrink, Sonja Hendriks, Anna Köhler, Lisa Korte, Jessica Stegemann, Sarah Stock, Henning Timm und Stephanie Rehwald. 2025. *Forschungsdatenmanagement-Curriculum der Research Data Services an der Universitätsbibliothek Duisburg-Essen (Preprint).* https://doi.org/10.17185/DUEPUBLICO/83316.

Leimer, Sophia, Sonja Hendriks, Lisa Korte, Jessica Stegemann, Sarah Ann Stock, Henning Timm und Stephanie Rehwald. 2023. „Research Data Management Curriculum of the Research Data Services at the University Library Duisburg-Essen". In *Proceedings of the Conference on Research Data Infrastructure,* Bd. 1. TIB Open Publishing. https://doi.org/10.52825/cordi.v1i.209.

Slowig, Benjamin, Mirjam Blümm, Konrad U. Förstner, Marvin Lanczek, Birte Lindstädt, Rabea Müller, Ulrike Nickenig, Stephanie Rehwald und Lioba Schreyer. 2022. „Zertifikatskurs Forschungsdatenmanagement in NRW". *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB* 9 (3/2022). https://doi.org/10.5282/o-bib/5833.

Stapels, Julia, Tuba Güden-Silber und Jessica Stegemann. 2025. *Modulare standortübergreifende Schulungsangebote: Das FDM-Curriculum der Universitätsallianz Ruhr.* https://doi.org/10.11588/HEIDOK.00036253.

Tolks, Daniel und Michael Sailer. 2021. „Gamification als didaktisches Mittel in der Hochschulbildung". In *Digitalisierung in Studium und Lehre gemeinsam gestalten,* 515–532. Springer Fachmedien Wiesbaden. ISBN: 9783658328498. https://doi.org/10.1007/978-3-658-32849-8_29.

Open Science in den Humanities: Mehr Spielräume und Nachvollziehbarkeit durch Galaxy

Daniela Schneider [1], Jan Leendertse [2]

[1] Galaxy Team Freiburg, Albert-Ludwigs Universität Freiburg;
[2] Central Data Facility, Rechenzentrum, Albert-Ludwigs Universität Freiburg

In den Geisteswissenschaften kann der Einstieg in digitale Analysen mit etlichen Hürden verbunden sein. Fehlende Programmierkenntnisse, hohe Kosten für Spezialsoftware oder fehlende Administratorenrechte können eine Herausforderung sein. Dazu soll gute wissenschaftliche Praxis umgesetzt und gutes Forschungsdatenmanagement betrieben werden. Je mehr dieser Anforderungen zusammenkommen, desto herausfordernder wird die Umsetzung eines neuen digitalen Projektes.

Dieser Artikel untersucht, wie die Open Source Analyseplattform Galaxy an genau diesen Herausforderungen ansetzen kann. Am Praxisbeispiel eines Textvergleichs aus einer geschichtswissenschaftlichen Dissertation zeigt der Artikel, wie die in den Lebenswissenschaften entwickelte Plattform auch in den Geisteswissenschaften sinnvoll einsetzbar ist. Die Beispielanalyse, die ursprünglich manuell durchgeführt wurde und sehr zeitaufwändig war, wurde automatisiert, was die Effizienz und die Aussagekraft der Ergebnisse erheblich steigerte. Der in Galaxy erstellte Workflow ermöglicht eine transparente und reproduzierbare Analyse, wodurch die Prinzipien der guten wissenschaftlichen Praxis gefördert werden. Zudem wird gezeigt, dass auch in den Humanities eine Offenheit für Open Science möglich ist, selbst wenn die zugrunde liegenden Daten aus proprietären Quellen stammen. Dieses Beispiel verdeutlicht nicht nur die Nützlichkeit von Galaxy als Werkzeug zur Textanalyse, sondern durch das Baukastenprinzip der Plattform auch das Potenzial für die breite Anwendung in verschiedenen Fachbereichen. Durch die Implementierung von Galaxy wird es Forschenden ermöglicht, sich stärker auf ihre Kernkompetenzen zu konzentrieren und gleichzeitig die Sichtbarkeit und Nachvollziehbarkeit ihrer Methoden zu verbessern, wodurch die Herausforderungen der digitalen Analyse in den Geisteswissenschaften adressiert werden.

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23955 (CC BY-SA 4.0).

Keywords: Humanities, Open Science, FAIR

1 Open-Science in den Humanities?

Open Science ist als Prinzip breit zustimmungsfähig. In den Geisteswissenschaften sehen viele Forschende für sich jedoch keinen schnellen oder praktikablen Weg, dem zu folgen. Open Science wird mit Digitalisierung assoziiert, die in den Humanities vielerorts Abwehr auslöst. Darüber hinaus soll in den Projekten gute wissenschaftliche Praxis befolgt und gutes Forschungsdatenmanagement betrieben werden. Diese Anforderungen an gute wissenschaftliche Qualität und Nachvollziehbarkeit können für Forschende jedoch auch eine Hürde darstellen. Kommen zu diesen noch weitere Barrieren wie hohe Kosten für Software oder fehlende Administratorenrechte, um Programme auf dem eigenen Computer zu installieren, oder müssen erst Programmiersprachen für die Analyse erlernt werden, wird der Einstieg in digitale Analysemethoden schnell zur Herausforderung.

2 Textanalyse mit Galaxy

Am Beispiel einer Dissertation aus den Geschichtswissenschaften wird gezeigt, wie die Open Source Plattform Galaxy (The Galaxy Community u. a. 2024) als Self-Service-Tool den Einstieg in digitale Textanalyse erleichtert (Schneider und Leendertse 2025). Obwohl Galaxy in den Lebenswissenschaften entwickelt wurde, wird hier deutlich, dass die Plattform großes Potential für die Geisteswissenschaften birgt. Die Beispielanalyse wurde in der Dissertation mit bekannten geschichtswissenschaftlichen Methoden durchgeführt und war nicht genuin als digitaler Prozess geplant. Methodisch wurde die Quellenlage durch den Vergleich von Zeitungsartikeln hergestellt, die in unterschiedlichen politischen Räumen entstanden sind. Sie stimmen zum großen Teil überein, weisen aber sichtbare Unterschiede in Form von Ersetzungszeichen auf, die auf Zensureingriffe zurückgehen. Der Textvergleich wurde ursprünglich manuell durchgeführt, war sehr zeitaufwändig und repetitiv. Hier setzt Galaxy[1] als Open Source Plattform an.

Galaxy bietet eine grafisch orientierte Oberfläche, die zum Experimentieren einlädt. Auf Wunsch kann das Ausprobieren aufgezeichnet, im Nachgang bereinigt und als Prozedur gespeichert werden. Solche Workflows werden so zum Ausgangspunkt weiterer Verfeinerung, sind teilbar, transparent und im besten Fall reproduzierbar. Die FAIR-Prinzipien bleiben kein schwer erfüllbarer Anspruch, sondern werden beiläufig gelebtes Prinzip. Für viele Fachbereiche gibt es bereits zahlreiche Trainingsmodule und Best-Practice-Workflows in Galaxy, die ohne Barrieren als Open Educational Resources (OER) zugänglich sind (Hiltemann u. a. 2023).

1 https://usegalaxy.eu/; *Besucht am 4. Juni 2025.*

Das Beispiel zum Textvergleich chinesischer historischer Zeitungen macht dies deutlich. Statt zäher manueller Gegenüberstellung können in Galaxy zensierte und unzensierte Zeitungsartikel mit chinesischen Zeichen schnell und übersichtlich gegenübergestellt werden. Der Vergleich ist visuell aufschlussreich und kann mit Standardtools wie 'diff'[2] ausgewertet werden. Dadurch können auch große Textmengen schnell analysiert werden. Technisch überraschend einfach, aber mit hoher Aussagekraft hinsichtlich der Forschungsfrage, welche Inhalte in den Zeitungen zensiert wurden. Bei der Auswertung ist weiterhin spezifische Erfahrung aus den jeweiligen Fachgebieten gefragt, besonders bei der kritischen Einordnung von Zwischenergebnissen und daraus folgend der weiteren Bearbeitung und tiefergehenden Analyse des Materials. Die Nutzung von Galaxy zeigt, wie Teile der Analyse auch in den Geisteswissenschaften teilautomatisiert werden können, so dass größere Textmengen verarbeitet werden können. So werden Ergebnisse nachvollziehbarer. Gleichzeitig bleibt für die eigentliche Auswertung mehr Zeit.

Der Workflow, die einzelnen Analyseschritte und die Daten jedes Teilschritts können dabei gemäß der guten wissenschaftlichen Praxis geteilt und wiederholt werden. Obwohl die Ausgangsdaten im Beispiel wegen ihres proprietären Ursprungs nicht zugänglich sind, kann so der Prozess geteilt werden. Das zeigt gleichzeitig, dass Open Science selbst dort möglich ist, wo Daten aus guten Gründen nicht offen sind.

3 Methodisches Vorgehen

Für einen Textvergleich auf Zeichen- und Zeilenebene eignet sich das Tool 'diff'[3], das aus der Unix-Welt stammt, aber auch für Windows portiert ist. Der Gebrauch dieses Kommandozeilentools ist nicht Teil des historischen Studiums und erfordert autodidaktische Einarbeitung. Zusätzlich eine Umgebung aufzubauen, die das Vorgehen reproduzierbar macht, erfordert weitere Kenntnisse, die nicht zum Curriculum von Historiker:innen gehören.

Der Beispielworkflow, der diesem Artikel zu Grunde liegt (Schneider 2025), zeigt, wie diese Textanalyse schrittweise im Galaxy erarbeitet werden kann. Diese Analyseplattform protokolliert das Vorgehen, das als Workflow gespeichert und verfeinert werden kann.

Das grundlegende Forschungsinteresse ist, welche Zeichen in einem chinesischen Zeitungsartikel aus Hong Kong in den 1930er Jahren zensiert wurden. Dazu werden zwei Zeitungsartikel miteinander verglichen. Da die der Arbeit zugrunde liegenden Artikel aus einer proprietären Datenbank stammen, wurden für ein Tutorial, das im Nachgang zur eigentlichen Forschung angeschlossen wurde, stattdessen mit Sprachmodellen Dummy-Texte erstellt. Mit diesen Texten kann die Analyse schrittweise nachgearbeitet werden, um das

2 Das Programm ermöglicht eine automatisierte Gegenüberstellung von Texten und zeigt deren Unterschiede.

3 https://pubs.opengroup.org/onlinepubs/9799919799/utilities/diff.html, *Besucht am 27. März 2025.*

Verfahren auf ähnliche Forschungsfragen anzuwenden. Mit diesen Texten und dem Tutorial bleiben die originalen Quellen zwar weiterhin unter Verschluss, doch die Methoden werden offengelegt und können nachgenutzt werden.

Einer der Artikel enthält offensichtliche Zensurzeichen in Form von ×, welches einzelne chinesische Zeichen ersetzte. Durch intensive Recherche konnte im Original eine frühere Version desselben Artikels unzensiert recherchiert werden, der in Festlandchina veröffentlicht wurde.

Statt die beiden Texte mühevoll manuell miteinander zu vergleichen, um herauszufinden, welche Zeichen mit × ersetzt wurden, war dies mit der Plattform Galaxy automatisiert möglich. Der Workflow lässt sich dabei in verschiedene Schritte unterteilen, die mit jeweils unterschiedlichen Programmen (in Galaxy Tools), die sich zu einem Workflow miteinander verbinden lassen. Zunächst werden die Texte hochgeladen und für den Vergleich vorbereitet. Der Übersichtlichkeit halber, werden sie als ein Zeichen pro Zeile dargestellt und Leerzeichen gelöscht. ‚Diff' stellt beide nun identisch vorbereiteten Texte visuell gegenüber und markiert Veränderungen farblich. Auslassungen werden rot und die Einfügungen grün markiert.

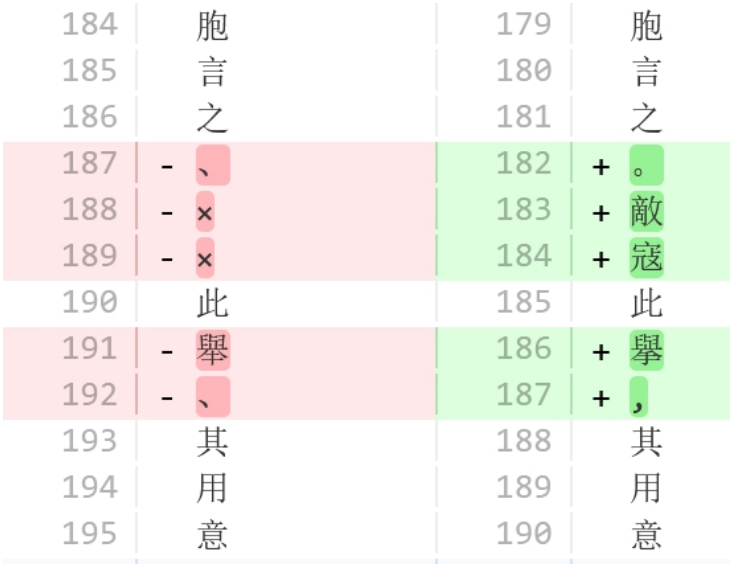


Abbildung 1: Screenshot des Textvergleichs mit ‚Diff'-Tool in Galaxy.

Hier werden auch die Zeilen farblich hervorgehoben, in denen die Zensurzeichen vorkommen. Nach ihnen wird im nächsten Schritt gefiltert, sodass das Dokument nur noch × und ihre unzensierten Äquivalente enthält. Für zwei verschiedene Arten der Visualisierung werden die extrahierten Zeichen nun einmal quantifiziert und in einer Tabelle sortiert und einmal als Wordcloud grafisch ausgegeben. Der Vergleich der Artikel zeigt, dass Zeichen wie 敵 (Feind) und 寇 (Eindringling) häufig zensiert wurden. So beispielsweise auch in den Zeilen 188 und 189 der obigen Abbildung. Dies verweist auf die Zensurpolitik der britischen Regierung, die bestrebt war, anti-japanische Inhalte zu vermeiden, um mögliche

Spannungen mit Japan zu verhindern. Japaner in den in Hong Kong veröffentlichten Zeitungsartikeln als Feinde oder Eindringlinge zu bezeichnen, offenbarte eine pro-chinesische Haltung der Zeitungen. Diese zensierte die britische Hong Kong Regierung, die im Konflikt neutral bleiben wollte, um nicht selbst in den Konflikt verwickelt zu werden. Die Gegenüberstellung der Zeitungstexte im Beispieltutorial unterstreicht diese Erkenntnisse.

4 Erschließbarkeit der Ergebnisse durch Galaxy

Analysen werden parametrisierbar und schaffen eine neue Vergleichbarkeit von Ergebnissen. Daten und Inhalte von Plattformen, wie beispielsweise Zenodo, können in Galaxy importiert und in einem Workflow algorithmisch analysiert werden. Die Ergebnisse sind wiederum in einem frei wählbaren Repositorium (Zenodo, FreiData, etc.) als FAIRes Digitales Objekt (FDO), beispielsweise RO-Crate, speicherbar. FDO sind auf Export und maschinenlesbare Weiterverarbeitung ausgelegt.

Galaxy bietet Open Educational Resources (OER) – ebenso Arbeitsfeld von Open Science – im Galaxy Training Network (GTN), das bereits über 400 Tutorials enthält, die über offenene Lizenzen bereitgestellt werden. Neben den praktischen Tutorials bietet es Beispieldatensätze, E-Learning-Material, Videos, Aufzeichnungen und komplette Schulungsworkshops, die Forscher:innen weltweit zum Selbststudium zur Verfügung stehen. Das Schulungsnetzwerk wird von einer globalen Community unterstützt, die neuen Nutzer:innen den Einstieg erleichtert und in kollaborative Arbeitsformen einführt. Durch diesen Ansatz haben verschiedene Wissenschaftsfelder Galaxy für sich entdeckt. Erste Analyseprojekte aus den Humanities finden sich bereits auf der Plattform.[4] So versucht Galaxy Forschende dabei zu unterstützen, sich auf das Kerngeschäft ihrer Arbeit zu fokussieren: die Forschung und Untersuchung/Analyse von Material und Daten. Reproduzierbarkeit und die Grundlage für gutes Forschungsdatenmanagement sind automatisches „Nebenprodukt“ der Analyse und verringern so die Hürden, diese Erwartungen an gute, FAIRe Forschung tatsächlich umzusetzen. Technisch gibt es sehr viele bereits integrierte Programme, die kostenlos, ohne Installation und ohne Programmierkenntnisse im Browser abrufbar sind. Dies ist auch in den Geisteswissenschaften von großem Interesse für die Forschenden.

4 Liste von Digital Humanities Tutorials im Galaxy Training Network: https://training.galaxyproject.org/training-material/topics/digital-humanities/; *Besucht am 9. Juli 2025 durch die Herausgeber*innen.*

5 Fazit

Zusammenfassend zeigt das Fallbeispiel, wie Galaxy als Open Source Plattform digitale Analysen in den Geisteswissenschaften unterstützen kann. Da Galaxy ohne Installation im Internet-Browser zugänglich ist und Analysetools mit wenig Einarbeitungsaufwand zugänglich macht, verringert sie die eingangs erwähnte Hürde für digitale Textanalysen. Durch das Baukastenprinzip ist Galaxy zudem sehr flexibel einsetzbar, und kann leicht für das eigene Forschungsvorhaben angepasst werden. Die einzelnen Schritten innerhalb eines protokollierten Workflows werden automatisch mit den Zwischenergebnissen dokumentiert. Die Ergebnisse lassen sich auch kleinschrittig teilen. Auch der Workflow als Ablauf der Analyse ist teilbar und reproduzierbar. Damit wird über die Analyse auch das methodische Vorgehen für andere erschließbar. Sie können als Ro-Crate Format exportiert und in weitere Repositorien leicht übernommen wernden. Das Beispiel eines algorithmisch unterstützten Textvergleichs mit ungewöhnlichem Zeichensatz zeigt, wie FAIRes Forschungsdatenmanagemen in den Humanities praktischen Nutzen hat.

Danksagungen

Wir bedanken uns für die finanzielle Unterstützung durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Nationalen Forschungsdateninfrastruktur – NFDI 46/1 – 501864659164 – NFDI4BioImage. Darüber hinaus wurden wir von DataPLANT 442077441 durch die Nationale Forschungsdateninitiative (NFDI 7/1), de.NBI-RBC und de.KCD unterstützt.

Beiträge

Daniela Schneider und Jan Leendertse haben beide gleichwertig als Co-Autoren den obigen Text verfasst.

Interessenkonflikte

Keine.

Literaturverzeichnis

Hiltemann, Saskia, Helena Rasche, Simon Gladman, Hans-Rudolf Hotz, Delphine Lariviè-re, Daniel Blankenberg, Pratik D. Jagtap u. a. 2023. „Galaxy Training: A powerful framework for teaching!“ Herausgegeben von Francis Ouellette. *PLOS Computational Biology* 19 (1): e1010752. https://doi.org/10.1371/journal.pcbi.1010752.

Schneider, Daniela. 2025. *Text-Mining Differences in Chinese Newspaper Articles.* Tutorial. Besucht am 21. Mai 2025. https://gxy.io/GTN:T00530.

Schneider, Daniela und Jan Leendertse. 2025. *Open Science in the Humanities: Galaxy as a Flexible and yet Reproducible Platform.* Poster. Heidelberg. Besucht am 21. Mai 2025. https://doi.org/10.11588/HEIDOK.00036288.

The Galaxy Community, Linelle Ann L Abueg, Enis Afgan, Olivier Allart, Ahmed H Awan, Wendi A Bacon, Dannon Baker u. a. 2024. „The Galaxy Platform for Accessible, Reproducible, and Collaborative Data Analyses: 2024 Update“. *Nucleic Acids Research* 52 (W1): W83–W94. https://doi.org/10.1093/nar/gkae410.

Bridging the Troubled Waters of Peer Review with ing.grid: Towards Openness, Fairness and F.A.I.R.-ness

Agnes Kleinhans [1], Michaela Leštáková [1], Kevin Logan [1], Peter F. Pelz [1]

[1] Chair of Fluid Systems, Technical University of Darmstadt

Peer review is a method for ensuring the soundness and trustworthiness of scientific results. It has become an essential part of the publishing process in science. However, the practice of peer review has been receiving more and more criticism, highlighting the challenges it faces. We present our own experience with pushing peer review forward in the independent, scholarly-led journal ing.grid founded in 2021. ing.grid is a diamond OA journal for FAIR data management in engineering sciences that accepts manuscripts as well as software and data submissions. To improve the review process, we have implemented a post-publication Open Peer Review Process that allows for fast publication of the first version of the submission as a preprint while also aiming for fairer, high quality reviews. Through the implementation of Open Peer Review, ing.grid has established a publishing process that is more transparent and accessible, aligning with the principles of Open Science and FAIR data management.

Keywords: Open Peer Review, Open Science, FAIR Data, Scholarly Publishing, Engineering Science

1 Introduction

Open peer review (OPR) is an evolving model of peer evaluation that emphasises transparency by revealing the identities of both authors and reviewers, or through the (anonymised) publication of reviews. This approach contrasts with traditional peer review methods, such as double-anonymised reviews and single-anonymised reviews, which maintain (reviewer) anonymity.

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23942 (CC BY-SA 4.0).

ing.grid is an independent, scholarly-led, diamond open-access journal for FAIR (Findable, Accessible, Interoperable, Reusable) data management in engineering science. Established in 2021, ing.grid promotes transparency and openness by following the principles of Open Access, Open Review, Open Data, and the FAIR principles. By integrating OPR into its publishing process, ing.grid reflects its commitment to these ideals.

This paper investigates the significance of OPR in the changing landscape of academic publishing, particularly emphasising its application in ing.grid. It reviews the benefits and challenges of OPR on research transparency and integrity. Through this exploration, the paper aims to enrich ongoing conversations about the future of peer review and the shift towards more open, inclusive, and transparent publishing practices.

2 Why open peer review?

Since 2001, there has been a constant emergence of Open Peer Review (OPR) in the publishing industry. The number of open report journals has increased from 38 journals in 2001 to at least 617 journals by the end of 2019 (Wolfram et al. 2020). Remarkably, a mere five publishers are responsible for 81 % of all recognised OPR journals (MDPI, SDI, BioMed Central, Frontiers Media S.A., Kowsar, see Wolfram et al. 2020). This highlights the significant influence publishers have had so far in advancing open report initiatives. OPR is especially practiced in Medical and Health science disciplines, in contrast to Humanities and Technology, with optional open reports being more prevalent in those fields (Wolfram et al. 2020).

The benefits of adopting OPR include improving transparency in the scientific publishing process. By publicly sharing reviews, OPR enhances the visibility of scientific discourse. It also promotes accountability, resulting in more valuable feedback. Furthermore, reviewers gain enhanced acknowledgment for their input, contrasting with conventional peer review systems, where their identities are kept confidential.

This transparency can lead to more fair assessments, as reviewers are held accountable for their comments (Holst, Eggleton, and Harris 2022) and discourage non-constructive or even harmful feedback. By making reviewers accountable for their comments, OPR can promote a more respectful and professional peer review process. In some cases, OPR has been shown to result in longer and more detailed reviews, as reviewers may feel a greater sense of responsibility when their identities are known (Wolfram, Wang, and Park 2019).

In traditional peer review models, reviewers often remain anonymous, which ultimately leads to not receiving recognition for their contributions. OPR addresses this limitation by publicly acknowledging reviewers' efforts, thereby promoting greater transparency and credit within the scholarly community (Henriquez 2023).

While OPR offers a variety of benefits, it also introduces particular challenges that need to be considered. Reviewers can be hesitant to provide negative feedback to authors they have a personal or professional relationship with (Vidal and Leaver 2015). In addition, research has indicated that reviewers are more inclined to offer favourable evaluations to authors from their own country; this tendency has been statistically significant in certain instances, suggesting a possible bias related to national affiliation (Thelwall et al. 2019). Addressing these biases requires careful consideration and the implementation of strategies to ensure fairness and neutrality in the peer review process.

In spite of these obstacles, open peer review has the potential to enhance the quality and transparency of scientific publishing. The ongoing transition toward OPR signifies a larger trend in academia aimed at enhancing openness, accountability, and accessibility in scholarly publishing.

3 FAIRness and openness in ing.grid's open peer review process

ing.grid aims at advancing transparency and openness in the scientific publishing process by following an OPR policy and the FAIR (findable, accessible, interoperable, reusable) principles as guidelines. Submissions are required to follow the FAIR principles, as these principles offer a framework to improve the transparency, accessibility, and reproducibility of scientific research while establishing a baseline standard for submissions to ing.grid.

For all categories of submissions – software, data, and manuscripts – ing.grid mandates that materials be *findable* by both humans and machines. This is primarily achieved through the assignment of a Digital Object Identifier (DOI) to each accepted and published submission, providing a unique and persistent global reference. In addition, software and data submitted as supplementary material to ing.grid's submissions must be deposited in publicly accessible repositories that support persistent identification. Software repositories must be directly linked in the software descriptor and include a version ID. Likewise, datasets must be stored in repositories that assign DOIs, which must be cited in the metadata of the data descriptor. For manuscripts, any supplementary software or datasets must be made discoverable through clearly referenced DOIs or permanent links. While formal descriptors for supplementary materials are not mandatory, authors are required to clearly reference all used tools or data.

For *accessibility*, all resources must be retrievable via standardised protocols. Software repositories must facilitate anonymous browsing, cloning, and viewing, guaranteeing that access is not restricted by any registration or barriers. Datasets need to be openly available and with an option for reviewers to access it anonymously. In manuscripts, authors are

responsible for ensuring that any supplementary software or data is publicly accessible and properly cited with URLs or DOIs.

Interoperability requires that software and data can effectively communicate with other tools, systems, and workflows. For software, this includes the implementation of standardised APIs, metadata, and the incorporation of dependency and environment management systems like requirements.txt, package.json, CMakeList.txt, or Docker. For datasets, metadata should be articulated in commonly accepted formats, and the data itself must be stored in non-proprietary, community-standard formats. Manuscript submissions, especially those that feature additional software or datasets, are encouraged to follow the same principles, facilitating the integration of the material into broader research. If possible, (meta)data should use standardised vocabularies.

To achieve *reusability*, licensing, documentation, and transparency are essential. Software must be released under an Open Source Initiative-recognised license and include comprehensive README files, installation guides, usage instructions, and examples. The documentation should outline contribution rules and testing procedures. Software submissions need to refer to executable code and describe how users can use and integrate the code in other projects. Data should also be open-licensed, typically CC-BY 4.0, and come with thorough documentation on creation, validation, and uncertainty measures. Manuscripts must cite additional software or datasets, clarifying their definitions, licenses, and providing basic usage examples or tutorials.

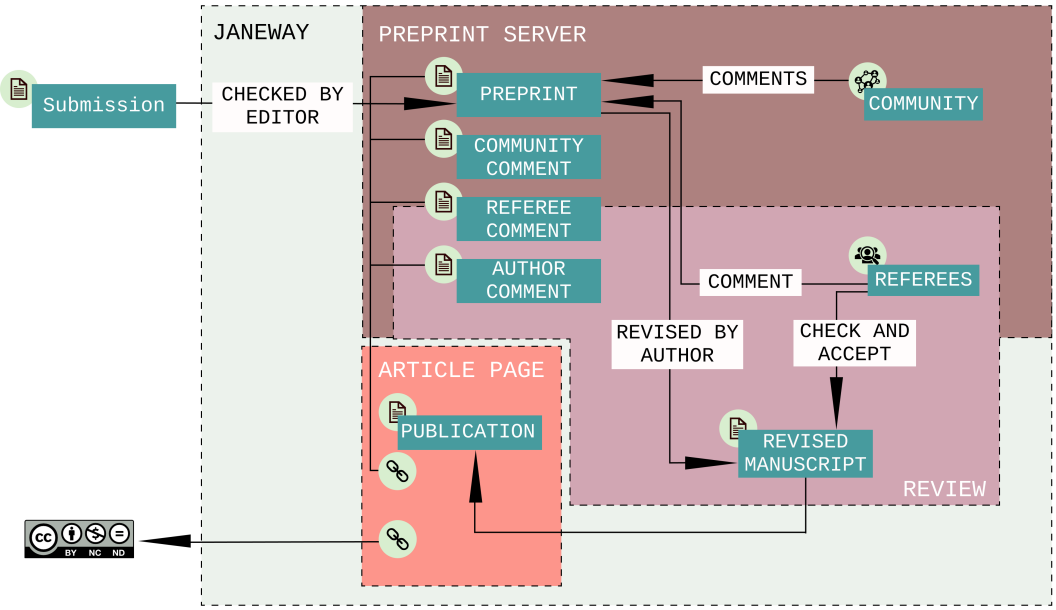


Figure 1: Open Peer Review Process at ing.grid.

The publishing and review process for ing.grid (Figure 1) involves several stages: First, authors submit their work to ing.grid as a preprint – the submission will be immediately

released on ing.grid's preprint server. During the review process, the authors' identities are known to the reviewers. Conversely, ing.grid strongly encourages reviewers to publish their identities together with their feedback, yet allows them the choice of remaining anonymous, temporarily or permanently.

However, the decision-making process is carried out with complete transparency – the reviews and the final editorial decisions and justifications are made available to both authors and the research community as comments below the submission on the preprint server. Authors must answer reviewer feedback in detailed comments, usually within four weeks. The comments are also published as part of the review discussion beneath the preprint. This open process ensures that all parties in the review process understand the reasoning behind acceptance, revision, or rejection, and thereby strengthening accountability and trust within the publication and research system. In addition, ing.grid also invites the broader community to actively engage with the discussion by providing their feedback and insights directly below the submission on the preprint server. This approach guarantees that knowledge is not restricted to a limited cohort of reviewers but rather aims upon drawing a heterogeneous spectrum of perspectives and increasing the visibility of the scholarly work.

Once a submission is accepted, the final version of the submission is published on ing.grid's journal and permanently linked to the preprint version, including the entire review discussion. In order to honor the scientific work behind reviews, reviewers can publish their reviews optionally linked with their ORCID, and receive formal review certificates, granting them academic acknowledgment for their work.

4 Insights and challenges in ing.grid's open peer review process

Our experience with ing.grid's OPR process has yielded significant insights regarding its advantages and limitations. A key characteristic is the transparency of review reports, which has got positive feedback from all participating parties. The authors appreciate the openness of the review process due to the public availability of review comments on the preprint server, while reviewers and the broader scientific community acknowledge the accessibility and transparency of the entire review process.

Although reviewers have the choice to remain anonymous during the review process, our results indicate that the identities of reviewers are often disclosed. With 52 % percent of reviewers choosing to reveal their identities, a considerable portion of reviewers operates with transparent identities.

While ing.grid promotes and allows open engagement from the broader scientific community, this feature has hardly been used so far. Even though the comment section beneath the articles on the preprint server is openly accessible to the community (provided an account is registered), contributions from researchers or the public beyond the reviewers, authors and editors have not been submitted so far. This implies that while the opportunity to contribute is intended to be inclusive and open, additional initiatives may be necessary to actively encourage involvement.

A significant challenge we have identified is the comparatively lengthy duration of the OPR process. Nevertheless, it remains uncertain whether this is inherently linked to the transparency of the method. As previously stated, we provide reviewers with the choice to stay anonymous, and around half opt for this. Typically, the duration of ing.grids' review process can reach up to 41 weeks, from the initial submission to the final publication in the ing.grid journal.

5 Future directions and conclusion

ing.grid's OPR process has illustrated significant advantages in transparency and openness towards authors, reviewers and the scientific community. To further enhance its efficiency and encourage broader adoption, there remain areas for further improvement.

An objective for the future is to simplify the connection between the preprint server and the journal platform. A more seamless workflow would reduce the administrative efforts between both websites and reduce the potential for confusion of the users. The most pressing area for improvement, is the duration of the review process. The current average review duration takes is 41 weeks – nonetheless, if researchers seeking a timely publication of their submission, the preprint article is published immediately. Lastly, broadening the acceptance criteria to incorporate preprints from other platforms would create the opportunity for researchers who have previously shared their work elsewhere.

Looking ahead, ing.grid aims to enhance the reach, accessibility and speed of scientific publishing, and thereby further empowering researchers, reviewers and the broader scientific community.

Funding & Acknowledgements

The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and

support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) – project number 442146713.

The authors would further like to thank the University and State Library Darmstadt and the scholarly publishing platform infrastructure Janeway.

Authorship Contributions

- Agnes Kleinhans – Conceptualisation, Project Administration, Writing (original draft, review, editing)
- Michaela Leštáková – Conceptualisation, Project Administration, Writing (original draft, review, editing)
- Kevin Logan – Conceptualisation, Project Administration, Writing (original draft, review, editing)
- Peter F. Pelz – Conceptualisation, Funding acquisition, Project Administration, Supervision

Conflict of Interest

The authors declare that they have no conflict of interests.

Bibliography

Henriquez, Tania. 2023. "Open peer review, pros and cons from the perspective of an early career researcher". *mBio* 14 (5). https://doi.org/10.1128/mbio.01948-23.

Holst, Faye, Kim Eggleton, and Simon Harris. 2022. "Transparency Versus Anonymity: Which Is Better to Eliminate Bias in Peer Review?" *The UKSG Journal* 35 (0): 16. https://doi.org/10.1629/uksg.584.

Thelwall, Mike, Verena Weigert, Liz Allen, Zena Nyakoojo, and Eleanor-Rose Papas. 2019. "Does the Use of Open, Non-Anonymous Peer Review in Scholarly Publishing Introduce Bias? Evidence from the F1000 Post-Publication Open Peer Review Publishing Model." *arXiv: Digital Libraries* 13 (1): 45–47. https://arxiv.org/abs/1911.03379..

Vidal, Jordi Blanes i, and Clare Leaver. 2015. "Bias in Open Peer-Review: Evidence from the English Superior Courts". *The Journal of Law, Economics, and Organization* 31 (3): 431–471. https://doi.org/10.1093/jleo/ewv003.

Wolfram, Dietmar, Peiling Wang, Adam Hembree, and Hyoungjoo Park. 2020. "Open peer review: promoting transparency in open science". *Scientometrics* 125 (2): 1033–1051. https://doi.org/10.1007/s11192-020-03488-4.

Wolfram, Dietmar, Peiling Wang, and Hyoungjoo Park. 2019. "Open peer review: the current landscape and emerging models". In *Proceedings of the The 17th International Conference on Scientometrics & Informetrics,* edited by Giuseppe Catalano, Martina Gregori Cinzia Daraio, Henk F. Moed, and Giancarlo Ruoccoh, 387–398. ISSI Society.

The CREATIVE Project – Customizing a Generic Repository for Domain Scientists and Creating a Data Steward Network

Sibylle K. Hassler [1,2], Peter Braesicke [3], Mirko Mälicke [2], Jörg Meyer [4], Carlos Zuleta Salmon [1]

[1] Institute of Meteorology and Climate Research – Atmospheric Trace Gases and Remote Sensing (IMKASF), Karlsruhe Institute of Technology (KIT);
[2] Institute for Water and Environment – Hydrology, Karlsruhe Institute of Technology (KIT);
[3] German Weather Service (DWD);
[4] Scientific Computing Center, Karlsruhe Institute of Technology (KIT)

The project CREATIVE aims at facilitating and supporting the publication of data from the scientific communities represented within the Climate and Environment Centre (CEC) at KIT in the generic institutional repository RADAR4KIT. We developed a customised template for domain-specific metadata and an online tool for metadata entries. Additionally, we facilitate establishing a network of data stewards at the CEC as a sustainable support structure for the data publication process.

Keywords: metadata, template, repository, data steward, data publication

1 Introduction

Data publication at the Karlsruhe Institute of Technology (KIT) is supported by the institutional, generic repository RADAR4KIT. While this process is well-established and supported by the research data management (RDM) team at the KIT library and Scientific Computing Center (SCC), the repository cannot fully support any domain-specific metadata description of the data by default. However, such domain-specific metadata options are desirable to better describe environmental data and support its findability in aggregating services that can use such descriptors. With the CREATIVE project we

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23943 (CC BY-SA 4.0).

aim to find a solution to this gap so that data from the KIT researchers working in the earth and environment domain can be published in a well-described and FAIR (findable, accessible, interoperable and reusable) way. In the following sections we will outline the different entities that play a role in the CREATIVE project and portray our project goals, approach, status and outlook.

2 Entities and interactions within the CREATIVE project

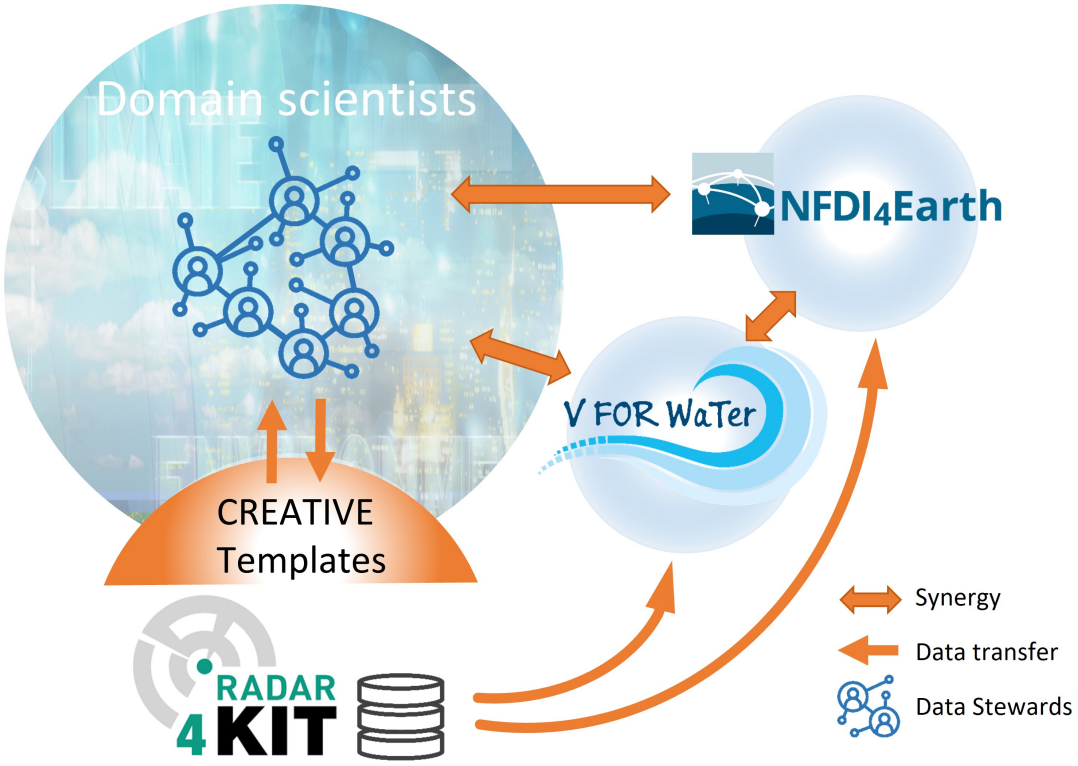


Figure 1: Overview of the CREATIVE project entities and interactions.

2.1 Domain scientists at the Climate and Environment Centre of KIT

The KIT Climate and Environment Centre (CEC)[1] is an organisational structure within KIT which encompasses institutes that conduct basic and applied research and generate knowledge on climate and environmental change. This includes concepts, observations and

1 https://www.klima-umwelt.kit.edu/english/index.php; *Visited on March 28, 2025.*

models as well as nature-based and technological solutions for safeguarding the natural basis of life and shaping the living environment. The CEC offers a broad range of expertise in the natural, technical, and social sciences with about 700 scientists from 30 institutes.

2.2 Institutional repository RADAR4KIT

The repository RADAR4KIT[2] is the generic institutional repository of KIT. It is based on the RADAR[3] service which is offered by FIZ Karlsruhe[4], with data storage realised on the IT infrastructure of KIT, located at the SCC. The repository allows publication and archiving of data packages from scientific studies and projects, including a coherent description with metadata. After a curation step the data receives a DOI and is openly accessible via the RADAR4KIT portal. The metadata description of RADAR4KIT is generic in order to cater to all research disciplines at KIT. Before uploading a dataset, the metadata needs to be entered by hand or can be uploaded as an XML file. To ensure the quality and consistency of these metadata records the XML files are validated against XML Schema Definition (XSD) files that define the structure and rules of the templates. This validation process checks for structural integrity, data type consistency, and compliance with controlled vocabularies.

2.3 Virtual research environment V-FOR-WaTer

V-FOR-WaTer[5] is a virtual research environment (VRE) that has been developed in close collaboration between hydrologists (KIT Institute for Water and Environment) and computer scientists (KIT SCC). It caters to the needs of hydrologists and environmental scientists by providing quick and easy access to hydrological data and tools. The main purpose of V-FOR-WaTer is to support quick pre-processing of data from diverse data sources and thus accelerating scientific workflows. Shared tools and a storable workspace support reproducible data analysis.

2.4 NFDI4Earth

The NFDI4Earth[6] (National Research Data Infrastructure for Earth System Science, ESS) is a consortium of around 70 institutions within Germany, including universities, research

2 https://radar.kit.edu/radar/en/home; *Visited on March 28, 2025.*
3 https://www.fiz-karlsruhe.de/en/produkte-und-dienstleistungen/radar; *Visited on March 28, 2025.*
4 https://www.fiz-karlsruhe.de/en; *Visited on March 28, 2025.*
5 https://www.vforwater.de/english/; *Visited on March 28, 2025.*
6 https://nfdi4earth.geo.tu-dresden.de/; *Visited on March 28, 2025.*

organisations, infrastructure providers, governmental institutions as well as scientific associations and networks. It aims to provide FAIR, coherent, and open access to all relevant ESS data, to innovative research data management and data science methods. NFDI4Earth was constructed in a community-driven process and offers a range of services. One of them is the OneStop4All[7], the web-based key entry point for all NFDI4Earth contents and resources. It was developed to fit the needs of a wide range of users and rests on a backbone called the KnowledgeHub where the metadata of ESS relevant resources are collected and referenced.

3 CREATIVE goals and methods

3.1 Goals of the CREATIVE project

The project CREATIVE aims at facilitating and supporting data publications in the generic repositorium RADAR4KIT for the scientific communities within the CEC at KIT. This is achieved by creating custom templates and entry forms for domain-specific metadata that satisfy the requirements of the CEC researchers. At the same time, suitable interfaces ensure that the domain-specific (meta-)data in RADAR4KIT are accessible from the national research data infrastructure NFDI4Earth and the virtual research environment V-FOR-WaTer. With further activities, e.g. organising specific RDM workshops, a network of data stewards in the individual institutes of CEC is established, strengthening sustainable and advanced RDM at KIT.

3.2 Methods of the CREATIVE project

In a series of interviews and workshops with CEC researchers we collected and defined the metadata requirements for a comprehensive description of the relevant datasets, we co-developed the templates, collected feedback and adapted the templates accordingly. In close collaboration with the RADAR4KIT and RADAR team we established how the templates can be integrated and verified in the repository and how a future harvesting of the data would work. The template was developed on the existing metadata schema of V-FOR-WaTer (based on ISO19115 and DataCite) to ensure the processing possibility within the VRE. Further, communicating with the data model team of NFDI4Earth will also enable findability of the data in the OneStop4All as soon as the RADAR4KIT repository is harvested for data originating from the CEC researchers. As a direct integration of the template as an entry form in RADAR4KIT was not (yet) possible, we implemented an independent web form to enter the metadata and export it in different formats, e.g. XML which can be used while uploading the metadata in the RADRA4KIT repository.

7 https://onestop4all.nfdi4earth.de/; *Visited on March 28, 2025.*

A procreative exchange among CEC researchers and data stewards at the workshops supported the cultural change towards more FAIR data publications and open data. While dedicated data stewards are lacking at many CEC institutes, the exchange formats brought together researchers interested in data publications and data managers from the Helmholtz Research Field Earth and Environment DataHub. The interaction between these different parties can be seen as a starting point for a network among CEC data stewards and creates a climate of mutual help and support.

4 Transferability

While the CREATIVE metadata template was targeted to support CEC researchers in publishing their data on the local repository RADAR4KIT, the approach itself is easily transferable to other disciplines and other repositories and is meant to serve as a blueprint of the possibility to customise generic repositories. This kind of approach has also been used successfully in bioinformatics (Brandt et al. 2023).

The code behind the metadata editor web form is available open source[8] and can be reused to cater to different constellations of metadata requirements and repositories.

5 Status and outlook

The CREATIVE project resulted in the collaborative design of domain-specific metadata templates, their incorporation into RADAR4KIT, and the development of a standalone web form application. This app enables researchers to upload datasets easily and generate XML files that comply with the required metadata structure, which are validated using the developed XSD validation schema for CREATIVE within RADAR4KIT. Once validated in RADAR4KIT, they are made discoverable with proper metadata standards, including DOI assignment. The metadata from the template was based on V-FOR-WaTer, ensuring compatibility, however, the direct connection of the two systems still needs to be established. The metadata is also compatible to the NFDI4Earth schema so it can be harvested automatically as soon as the connection to RADAR4KIT is established.

Further development steps are in progress and include, for example, the integration of existing ontologies, and the automatic completion of metadata fields from well-structured and widely used file formats at CEC such as netCDF. While the full implementation of interfaces to NFDI4Earth and V-FOR-WaTer is still ongoing, the current state is already a clear improvement for sharing and publishing datasets generated by CEC researchers.

8 https://github.com/KIT-HYD/creative_frontend; *Visited on March 28, 2025.*

The CEC data steward network was initialised and generates interest. Further steps in this regard will be visiting the CEC institutes and presenting the tool in hands-on sessions so that it can be thoroughly tested and promoted. In case the metadata requirements change or further institutes that were not involved in previous workshops join in, the template can also be updated. Directly supporting the researchers in using the tools will also strengthen the institutional effort strengthening data publications at KIT.

Acknowledgements

We acknowledge the support of the RADAR4KIT and RADAR teams in the development of the template, the support by the NFDI4Earth in the project and the fruitful exchange and discussions with the IMK data management team (supported by the Helmholtz Research Field Earth and Environment DataHub) and the other V-FOR-WaTer team members during the project.

Authorship Contributions

SH developed the project concept, was mainly responsible for the project proposal, initiated contacts between the project staff, CEC researchers and data managers, fostered connections to the NFDI4Earth and V-FOR-WaTer and supported the technical developments. She presented the poster at the E-Science-Tage and mainly wrote the article. PB facilitated RDM at IMK as co-chair of the Helmholtz RF E&E DataHub and as co-spokesperson of the NFDI4Earth. MM developed and implemented the web form, implemented the link between the web form and V-FOR-WaTer. JM contributed to the project concept, served as contact for V-FOR-WaTer. CZ co-developed the subject-specific metadata templates for RADAR4KIT, fostered collaboration between researchers and data stewards across the CEC institutes. PB, MM, JM and CZ all contributed to the writing of the article.

Conflict of Interest

There are no conflicts of interest.

Bibliography

Brandt, Olaf, Holger Gauza, Jan Kaltenbach, Maximilian E. Müller, Gabriel Schneider, and Claus Zinn. 2023. “Ein Werkzeug zur XSD-basierten Metadatenannotation”. In *E-Science-Tage 2023: Empower Your Research – Preserve Your Data,* edited by Vincent Heuveline, Nina Bisheh, and Philipp Kling, 276–281. heiBOOKS. https://doi.org/10.11588/heibooks.1288.c18088.

FIZ-OAI: Eine skalierbare Open-Source-Implementierung zur Aggregation und Dissemination von Metadaten mit OAI-PMH

Stefan Hofmann [1*], Kerstin Soltau [1], Michael Hoppe [1], Felix Bach [1]

[1] FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur;
[*] Kontakt: stefan.hofmann@fiz-karlsruhe.de

Das Protocol for Metadata Harvesting OAI-PMH stammt aus dem Jahr 2001, ist aber immer noch weit verbreitet. Mit FIZ-OAI hat FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur[1] einen skalierenden, robusten und einfach zu wartenden OAI-Provider als Open-Source-Software entwickelt, um große Mengen an Metadaten aus unterschiedlichen Forschungsdateninfrastrukturen zur freien Nachnutzung verfügbar zu machen. Aggregatoren wie CrossRef[2], die DNB[3], Europeana[4], BASE[5] oder B2Find[6], aber auch viele weitere Partner nutzen diesen Service, um Metadaten abzurufen (zu „harvesten“). In diesem Beitrag stellen wir die Architektur und wesentlichen Funktionen von FIZ-OAI vor und skizzieren aktuelle Anwendungsfälle von Forschungs- und Publikationsumgebungen in verschiedenen wissenschaftlichen Disziplinen.

Keywords: OAI-PMH, Metadaten, Metadata Harvesting, Software, Open-Source

1 https://www.fiz-karlsruhe.de; *Besucht am 16. Mai 2025.*
2 https://www.crossref.org/; *Besucht am 16. Mai 2025.*
3 https://www.dnb.de; *Besucht am 16. Mai 2025.*
4 https://www.europeana.eu/; *Besucht am 16. Mai 2025.*
5 https://www.base-search.net/; *Besucht am 16. Mai 2025.*
6 https://b2find.eudat.eu/; *Besucht am 16. Mai 2025.*

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23944 (CC BY-SA 4.0).

1 Hintergrund und Motivation

Die rasche Zunahme der Menge an digitalen Forschungsdaten und Publikationen (Bornmann und Mutz 2015) stellt die wissenschaftliche Informationslandschaft vor große Herausforderungen. Für eine effektive Nachnutzung, Archivierung und den interdisziplinären Austausch sind präzise, standardisierte Metadaten unverzichtbar, da sie Inhalte, Entstehungsumstände und wichtige Verfügbarkeitsinformationen (z. B. Lizenz, Speicherort) der Datensätze beschreiben. Nur so lassen sich Datensätze in wissenschaftlichen Repositorien systematisch finden, vergleichen und weiterverarbeiten (Baca 2016).

Neben der Speicherung großer Datenmengen rückt die Frage in den Fokus, wie die Daten nachhaltig bereitgestellt und automatisiert weiterverarbeitet werden können. Dabei hat sich seit mehr als zwei Jahrzehnten das Protocol for Metadata Harvesting (Lagoze und Van de Sompel 2001) der Open Archives Initiative (Lagoze u. a. 2002) bewährt. Dieses ermöglicht den Aufbau verteilter Dateninfrastrukturen, indem Metadaten verschiedener Dienste über eine standardisierte Schnittstelle zentral abgerufen werden können.

Obwohl das OAI-PMH-Protokoll bereits im Jahr 2001 eingeführt wurde und mittlerweile zahlreiche technische Alternativen, wie Linked Open Data[7] oder ResourceSync[8] verfügbar sind, findet es weiterhin breite Anwendung[9] in fast 7.000 verfügbaren OAI-Providern. Diese fortwährende Nutzung lässt sich primär auf seine umfassende Community-Unterstützung, die HTTP-basierte Architektur und die präzise definierte Spezifikation[10] zurückführen.

Stand 2019 betrieb FIZ Karlsruhe eine Vielzahl von Anwendungen (DDB, Thieme Connect, RADAR, zbMATH), deren Metadaten zur freien Nachnutzung verfügbar gemacht werden sollten. Es wurde ein einheitlicher OAI-Provider gesucht, welcher in allen Projekten eingesetzt und in die bestehenden Infrastrukturen integriert werden sollte. Aus den Projekten heraus wurden die folgenden Anforderungen an den Provider definiert:

- Einfache Installation und Betrieb über Docker-Container
- Langfristig wartbare Software auf Java-Basis, die aktuelle Sicherheitsstandards einhält
- Skalierbarkeit für Millionen von Metadatensätzen
- Vollständige OAI-PMH-Implementierung
- Open-Source zur Förderung einer aktiven Community

Die Open Archives Initiative listet auf ihrer Homepage über 30 frei verfügbare OAI-Provider[11] als potentielle Kandidaten. Die Wahl fiel auf das erweiterbare OAICat-Frame-

7 https://de.wikipedia.org/wiki/Linked_Open_Data; *Besucht am 16. Mai 2025.*
8 https://www.openarchives.org/rs/toc; *Besucht am 16. Mai 2025.*
9 https://www.openarchives.org/Register/BrowseSites; *Besucht am 16. Mai 2025.*
10 https://www.openarchives.org/OAI/openarchivesprotocol.html; *Besucht am 16. Mai 2025.*
11 https://www.openarchives.org/pmh/tools/; *Besucht am 16. Mai 2025.*

work[12]. Es basiert auf Java, steht unter einer Open-Source-Lizenz und unterstützt eine flexible Datenhaltung, welche auch die Anbindung von clusterfähiger Datenbanken beinhaltet. Somit waren alle genannten Anforderungen erfüllt. Auf Basis von OAICat wurde bei FIZ Karlsruhe die Software FIZ-OAI entwickelt. Dabei handelt es sich um einen leistungsstarken und modular aufgebauten OAI-Provider, der in der Lage ist, Millionen von Metadatensätzen zu verwalten und bereitzustellen. Die Software steht unter einer Open-Source-Lizenz (Apache 2.0)[13] frei zur Verfügung, so dass wissenschaftliche Einrichtungen einen eigenen OAI-Provider selbstständig und flexibel aufsetzen können.

2 Architektur

Aufbauend auf den geschilderten Zielsetzungen erläutern wir im Folgenden, wie sich diese in der konkreten Systemarchitektur widerspiegeln. Der nächste Abschnitt gibt detaillierte Einblicke in die technischen Komponenten und deren Zusammenspiel. Abbildung 1 veranschaulicht die Mehrschichten-Architektur von FIZ-OAI, bestehend aus einem Webserver (FIZ-OAI-Provider), einem Anwendungsserver (FIZ-OAI-Backend), einer Datenbank (Cassandra) und einem Suchindex (ElasticSearch / Apache Solr). Hinzu kommt noch eine datengebende (Loader) Komponente, welche das OAI-Backend mit Metadaten befüllt.

Der *FIZ-OAI-Provider*[14] verarbeitet OAI-PMH-Anfragen und beruht auf dem *OAICat-Framework* von OCLC. Beim „Harvesten“ generiert der OAI-Provider für jedes angefragte OAI-PMH-Verb die passende Antwort durch Suchanfragen an das Backend. Die genauen Anfrage-/Antwortformate und Fehlerbehandlungen sind in der OAI-PMH Spezifikation[15] zu finden.

Das *FIZ-OAI-Backend*[16] bietet eine REST-API zur Speicherung und Suche von Metadaten, die in einem Apache Cassandra[17] Datenbank-Cluster abgelegt werden. Cassandra ist eine verteilte NoSQL-Datenbank und ermöglicht linear skalierbare Lese- und Schreibleistungen für große Datenmengen. Um Suchanfragen von OAI-PMH-Verben wie ListIdentifiers/ListRecords effizient bearbeiten zu können, kommt wahlweise *ElasticSearch*[18] oder *Apache Solr*[19] zum Einsatz. Beide Suchplattformen können in einem Cluster betrieben werden und sind extrem leistungsfähig in der Verwaltung von Volltext- und Metadatenindizes.

12 https://www.oclc.org/research/areas/data-science/oaicat.html; *Besucht am 16. Mai 2025.*
13 https://www.apache.org/licenses/LICENSE-2.0; *Besucht am 16. Mai 2025.*
14 https://github.com/ER-FIZKarlsruhe/fiz-oai-provider; *Besucht am 16. Mai 2025.*
15 https://www.openarchives.org/OAI/openarchivesprotocol.html#ProtocolMessages; *Besucht am 16. Mai 2025.*
16 https://github.com/ER-FIZKarlsruhe/fiz-oai-backend; *Besucht am 16. Mai 2025.*
17 https://cassandra.apache.org; *Besucht am 16. Mai 2025.*
18 https://www.elastic.co/de/elasticsearch; *Besucht am 16. Mai 2025.*
19 https://solr.apache.org/; *Besucht am 16. Mai 2025.*

Eine datengebende Anwendung (*LOADER*), spricht über die REST-API mit dem Backend. Über HTTP werden die Metadaten in einem XML-Eingangsformat übermittelt. Alle Daten werden in Cassandra persistiert und die für OAI-PMH relevanten Felder (Identifier, Datum, Format, Set) für die Suche indexiert. Weiterhin kann das Backend über vom LOADER bereitgestellte XSLT-Transformationen (Crosswalks) das XML-Eingangsformat in weitere XML-Formate transformieren, um den Bedürfnissen unterschiedlicher Communities gerecht zu werden. Durch das Gruppieren von Datensätzen in Sets[20] können spezifische Metadaten-Kollektionen separat abgerufen werden.

Ein zentrales Ziel bestand darin, den gesamten Software-Stack via Docker-Container so aufzubereiten, dass er einfach installiert und gewartet werden kann, wodurch Betreiber rasch produktive Instanzen in lokalen Rechenzentren oder Cloud-Umgebungen einrichten können. Hierfür steht ein unterstützendes Projekt[21] auf GitHub bereit. Das Projekt besteht aus einer Docker-Compose Datei, in welcher obige Komponenten zu einer funktionalen Einheit zusammengefügt werden. Mit Hilfe eines Install-Script, kann in wenigen Minuten eine FIZ-OAI Instanz aufgesetzt werden.

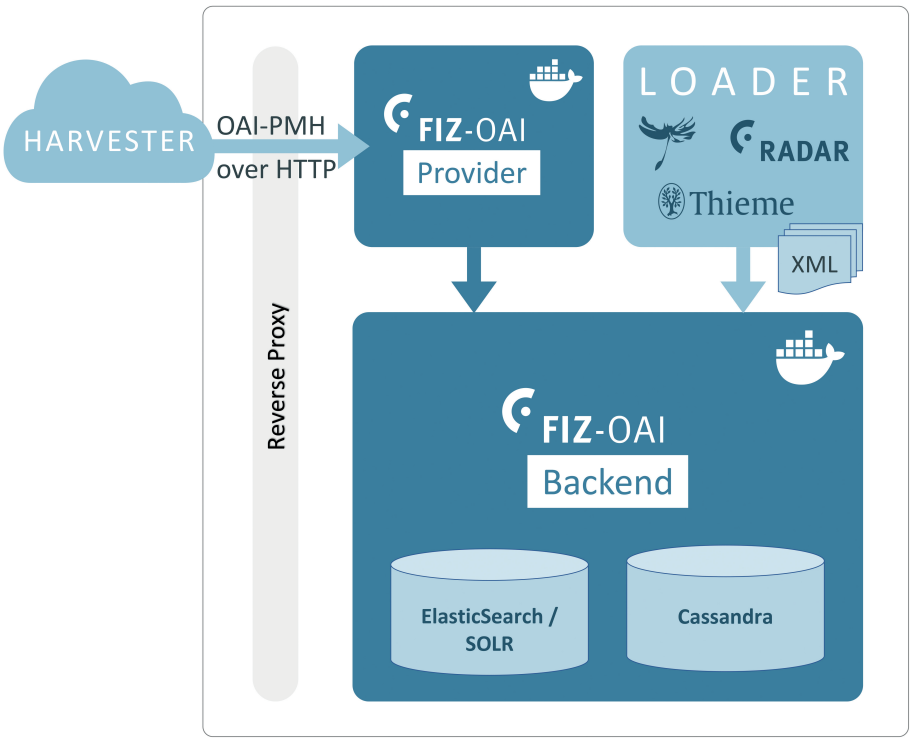


Abbildung 1: FIZ-OAI Architektur.

20 https://www.openarchives.org/OAI/openarchivesprotocol.html#Set; *Besucht am 16. Mai 2025.*
21 https://github.com/ER-FIZKarlsruhe/fiz-oai-docker; *Besucht am 16. Mai 2025.*

3 Anwendungsfälle

Auf dem OAI-Provider von *Thieme Connect*[22] stehen aktuell über eine Million Metadatensätze in den Formaten NLM, JATS, MARC zur Verfügung, welche bevorzugt für Zeitschriftenartikel und Bücher verwendet werden. Dies ermöglicht die Bereitstellung von medizinischen Fachartikeln für Aggregatoren wie der Deutschen Nationalbibliothek.

Der OAI-Provider der *Deutschen Digitalen Bibliothek (DDB)*[23], verdeutlicht insbesondere die gute Skalierbarkeit des Systems. Mit über 50 Millionen Metadatensätzen, die in 3 Formaten (DDB, EDM, OAI-DC) abgerufen werden können, sowie 4.232 verfügbaren Sets, handelt es sich um die größte aktuell laufende FIZ-OAI Installation. Diese Formate sind spezifisch für Objekte wie Bilder oder Texte aus Kultur- und Wissenssammlungen.

Auch für Forschungsdatenrepositorien wie *RADAR*[24] und die geplanten Dienste in der Nationalen Forschungsdateninfrastruktur (NFDI) bietet FIZ-OAI einen stabilen, wartungsarmen Zugang zum Austausch von Metadaten. Beispielsweise wird gegenwärtig eine Instanz für *NFDI4Chem* aufgebaut, um chemiebezogene Forschungsdaten zentral auffindbar zu machen. Eine *zbMATH* Instanz mit Metadaten aus der Domäne Mathematik soll zeitnah über 4 Millionen Metadatensätze ausliefern.

Tabelle 1: Übersicht der Metadatenfelder.

Instanz	**Dokumente**	**Formate**	**Harvester**
DDB	∼ 50.000.000	DDB, EDM	Europeana
Thieme	∼ 1.000.000	NLM, JATS, MARC	TIB, DNB, ZB MED
RADAR	∼ 2.000	RADAR, DATACITE	B2FIND, BASE, NFDI4Chem
zbMATH Open	∼ 4.000.000	DATACITE	MaRDI, DataCite

Danksagungen

Wir danken allen Partnerinstanzen und „Harvestern“, die uns stets konstruktiv bei der Optimierung und Weiterentwicklung der Software zur Seite stehen. Wir laden ausdrücklich zur aktiven Mitwirkung und Weiterentwicklung von FIZ-OAI ein.

22 https://www.thieme-connect.de/oai/OAIHandler?verb=Identify; *Besucht am 16. Mai 2025.*
23 https://oai.deutsche-digitale-bibliothek.de/?verb=Identify; *Besucht am 16. Mai 2025.*
24 https://www.radar-service.eu/oai/OAIHandler?verb=Identify; *Besucht am 16. Mai 2025.*

Beiträge

Die Grundstruktur des vorliegenden Textes basiert auf einem initialen Entwurf, der mithilfe von ChatGPT generiert wurde.

Interessenkonflikte

Die Autor*innen erklären, dass keine Interessenkonflikte bestehen.

Literaturverzeichnis

Baca, Murtha. 2016. *Metadata and the Web.* 3. Aufl. Herausgegeben von Murtha Baca. Los Angeles: Getty Publications. Besucht am 27. Oktober 2023. https://www.getty.edu/publications/intrometadata/.

Bornmann, Lutz und Rüdiger Mutz. 2015. „Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references". *Journal of the Association for Information Science and Technology* 66 (11): 2215–2222. https://doi.org/10.1002/asi.23329.

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson und Simeon Warner. 2002. *Open Archives Initiative Protocol for Metadata Harvesting Version 2.0.* Besucht am 21. Mai 2025. https://www.openarchives.org/OAI/openarchivesprotocol.html.

Lagoze, Carl und Herbert Van de Sompel. 2001. „The open archives initiative: building a low-barrier interoperability framework". In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries,* 54–62. JCDL01. ACM. https://doi.org/10.1145/379437.379449.

Uniting the DIFferences With the DALIA Interchange Format (DIF) – a Minimal Metadata Specification for the DALIA OER Platform

Jonathan D. Geiger [1], Petra C. Steiner [2], Abdelmoneim Amer Desouki [2], Henrika M. Hüppe [3], Jan-Michael Haugwitz [4]

[1] Academy of Sciences and Literature Mainz;
[2] University and State Library Darmstadt, Technical University of Darmstadt;
[3] Institute of Inorganic Chemistry, RWTH Aachen University;
[4] IT Center, RWTH Aachen University

The Data Literacy Alliance (DALIA) project aims to make open educational resources (OER) on data literacy and research data management skills accessible and interoperable. A knowledge graph has been developed to store the information about the materials, using the DALIA Interchange Format (DIF) to ensure transparency and interoperability. This paper focuses on the application of the DIF as a metadata specification for teaching and learning materials of the National Research Data Infrastructure (NFDI). It describes the format and attributes of the DIF and seeks to foster ongoing collaboration with the professional community by inviting all interested parties to provide feedback and make contact.

Keywords: Open Educational Resources, Metadata Specification, Data Literacy, Research Data Management, Search Platform

1 Introduction

Teaching data literacy is a key issue for society and encompasses a broad field: it begins with creating awareness and extends to practical applications, such as programming skills.

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23945 (CC BY-SA 4.0).

A wide range of data literacy materials and tools is already available, with some being commercial and others available as open educational resources (OER). An essential step in the development of data literacy is therefore the indexing, linking, visualizing, and making accessible of existing materials. The scientific infrastructure project DALIA takes up these tasks within the context of the National Research Data Infrastructure (NFDI), federal RDM initiatives, data competence centers, and beyond.

Since most scientific disciplines and fields in Germany deal with the issue of adequate training of data literacy for their students and researchers, it is reasonable to exploit synergies, share experiences, technically link existing materials, and identify gaps in both general and subject-specific offerings. The Section Training & Education (EduTrain)[1] of the NFDI is committed to overseeing this task; the DALIA project is responsible for the technical implementation, anchoring, and establishment within the NFDI.

A knowledge graph based on the MoDALIA Ontology is used to store information about the educational materials. It is based on semantic web technologies RDF, SPARQL, and OWL, allowing it to be linked to the Linked Open Data (LOD) cloud. To facilitate the ingestion process and serve as a means of communication with the community, the DALIA Interchange Format (DIF) was developed. It is compatible with the ontology.

This short paper provides an overview of the background of the DALIA project and focuses on the DIF as a metadata specification for teaching and learning materials. It also aims to foster ongoing communication with the professional community by inviting all interested parties to provide feedback and make contact.

Section 2 offers an overview of the DALIA project and platform. Section 3 describes the essentials of the DIF in greater detail. Section 4 will discuss future developments and enhancements planned for the near future.

2 DALIA: Project, Platform, and Knowledge Graph

The DALIA project started at the end of 2022. It combines a broad cross-disciplinary competence spectrum of teaching, research and infrastructure, supported by the Technical University of Darmstadt (Chair of Fluid System Technology, and University and State Library Darmstadt), the RWTH Aachen University (Chair of Bioinorganic Chemistry and IT Center), the Academy of Sciences and Literature Mainz (Digital Academy), the University Medical Center Göttingen (Department of Medical Informatics), and the German National Library of Science and Technology Hannover (TIB). Collaborating together, the project partners and the NFDI e.V. and its sections EduTrain, Ethical, Legal & Social As-

1 https://www.nfdi.de/section-edutrain/; *Visited on March 24, 2025.*

pects[2] , and (Meta)data, Terminologies, Provenance[3] aim to consistently provide teaching and learning materials. The DALIA project promotes the learning and teaching of data literacy by providing a curated collection of OER. This empowers everyone to collect, analyze, critically evaluate, and use data in a FAIR way. Hence, the DALIA platform aims to be a single point of entry for all different areas of science, career levels, and levels of research data management expertise – from those beginning with RDM to data stewards – providing access to a collection of OER focused on data literacy (Figure 1).

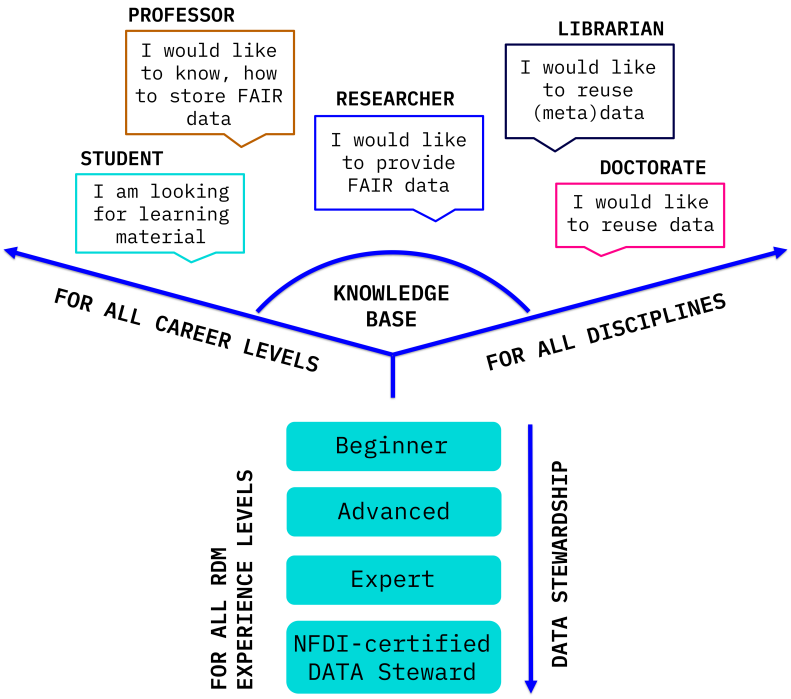


Figure 1: Visualization of the DALIA concept: Teaching and training materials for all career levels, research data management experience levels, and disciplines in one knowledge base.

To link and integrate the diverse needs of all career and competence levels, and academic disciplines, DALIA is developing a knowledge graph that establishes interoperability between existing resources and ensures findability, quality assurance, and adherence to the FAIR Principles (Garcia et al. 2020).

To ensure the interoperability of teaching and learning materials for the DALIA knowledge graph, a schematization of their metadata is required. This is done by the DIF, which provides a framework to make the metadata of teaching and learning materials transparent, comparable, and smoothly integrable into the DALIA platform. It describes and explains the data fields for the online publication of educational resources. Version 1.3 of the DIF was published on Zenodo in June 2024 (Geiger et al. 2024).

2 https://www.nfdi.de/section-elsa/; *Visited on March 24, 2025.*
3 https://www.nfdi.de/section-metadata/; *Visited on March 24, 2025.*

3 The DALIA Interchange Format (DIF)

The DIF was developed in close collaboration with the scientific community. The development process included several feedback rounds, leading to a clear structure, and transparent and user-oriented definitions. This process also allowed consideration of existing data standards as well as the requirements of the professional community. The selection of the description elements is based on the Dublin Core Application Profile (DCAP, see Coyle 2022), a set of elements in tabular form that can be used for the application and validation of metadata. Like the DCTAP, the DIF can be converted to RDF format. The description elements are as follows:

- The attribute name, which is a descriptive label
- The priority level *mandatory*, *recommended*, or *optional*
- A description of the attribute, including a definition and if required instructive information
- The values to be used, consisting of the constraint type and constraint
- Examples, for elucidating the definition
- The cardinality, for indicating if the attribute is repeatable or not
- The data type, which provides information on the format of the attribute
- The propertyID: a standardized identifier
- Optionally a comment, for example, more instructions on exceptions

The PropertyIDs that specify the attributes are particularly important: Common standards such as the DCMI Metadata Terms (DCMI Usage Board 2020) and Schema.org are used here. As an application profile, the DIF comprises elements compiled from other namespaces. However, when standardized definitions of the required attributes are unavailable, custom IDs are defined in the underlying MoDALIA ontology (mo)[4] to meet formal requirements such as range and domain. The priority levels were carefully selected. In practice, often some information about the resources is missing. However, this should not impede the process of ingestion to the platform. To address this, only four DIF attributes are mandatory (*Authors*, *License*, *Link*, *Title*). Eleven are strongly recommended (*Community*, *Description*, *Discipline*, *FileFormat*, *Keywords*, *Language*, *LearningResourceType*, *MediaType*, *ProficiencyLevel*, *PublicationDate*, *TargetGroup*), and three are optional (*RelatedWork*, *Size*, *Version*). Table 1 provides an example of the metadata for Biernacka et al. (2024).

4 https://purl.org/ontology/modalia; *Visited on March 24, 2025.*

Table 1: Example for the metadata of a resource according to the DIF. *Authors*, *Description*, and *LearningResourceType* have been abbreviated for conciseness.

Attribute Name	**Example**
Authors	Biernacka, Katarzyna :https://orcid.org/0000-0003-4864-8036 * Castillo Upiachihua, Claudia Beatriz :https://orcid.org/0009-0009-2186-3212 * ...
License	CC-BY-4.0
Link	https://doi.org/10.5281/zenodo.13927614
Title	Train-the-Trainer Concept on Research Data Management
Community	DINI/nestor UAG Schulungen/Fortbildungen (SR)
Description	This version of the 'Train-the-Trainer Concept on Research Data Management' is the English translation of the fifth updated version of the German 'Train-the-Trainer-Konzept zum Forschungsdatenmanagement' ...
Discipline	https://w3id.org/kim/hochschulfaechersystematik/n0
FileFormat	.pdf * .zip * .txt * .docx * .rtb * .pptx * .xlsx
Keywords	research data * research data management * RDM * train-the-trainer * OER * training * Forschungsdaten * Forschungsdatenmanagement * Didactics * ...
Language	en
Learning Resource Type	Book * https://w3id.org/kim/hcrt/course * https://w3id.org/kim/hcrt/lesson_plan * https://w3id.org/kim/hcrt/worksheet * ...
MediaType	text * presentation
ProficiencyLevel	novice * advanced beginner * competent
PublicationDate	2024-10-31
Target Group	data steward * teacher (higher education)
Related Work	isNewerVersionOf: https://doi.org/10.5281/zenodo.10122153 * isNewerVersionOf: https://doi.org/10.5281/zenodo.4071471
Size	195.2
Version	5.1

Further restrictions apply to the cardinalities. For instance, each OER should possess exactly one license and one title. The latter is for the sake of simplicity, as the DIF is a minimal standard. The attributes are explained in the description, and the technical constraints are made explicit. Controlled vocabularies are specified by picklists. Examples round off the DIF. Geiger et al. (2025) provide an example of the same OER showing the cardinalities. The DIF is partly compatible with the General Metadata Profile for Educational Resources (Pohl et al. 2024) and the controlled vocabulary used therein. It also considers the LRMI (Barker and Sutton 2022), and the RDA ETHRD-IG minimum metadata set recommendations (Hoebelheinrich et al. 2022). Additionally, a crosswalk to Schema.org attributes on learning resources is provided. To facilitate the contribu-

tion of metadata for OER to the DALIA platform, we developed a .csv template. This template was published alongside the DIF version 1.3 on Zenodo (Geiger et al. 2024). It includes a column for each attribute, allowing contributors to submit metadata for one or multiple OERs in a single document. This document is interoperable with the backend implementation of the DALIA platform.

4 Further Developments

Detailed feedback and inquiries from the community have shown the necessity and importance of such a metadata specification. Particularly during the preparation of metadata collections following the DIF structure, numerous questions arose, underscoring the potential for further improvements.

Currently, we are conducting a detailed analysis of the feedback and questions received. This process has led to the preparation of a revised version of the DIF (version 1.4). On the one hand, there was further feedback from research data management projects with their respective needs and perspectives. While these are largely addressed by the DIF, they also introduce new requirements in certain areas. On the other hand, the practical application of the DIF by the projects has revealed ambiguities, which can be resolved through more precise definitions, adapted examples, and, in particular, more detailed and extended picklists, for example in the area of scientific disciplines or domain-specific keywords. Compatibility with the DIF version 1.3 is important here. Similarly, future versions will also be backward compatible. This iterative development approach supports us to provide solutions that fit the community best and achieve widespread acceptance. In the near future, metadata for OER will no longer need to be submitted via .csv file in a manual process. Instead, DALIA will provide a curation form to registered contributors. This will enhance the useability through features like autocomplete and API-based data import from platforms such as Zenodo. Consequently, metadata can be added in accordance with the DIF and the MoDALIA Ontology with less effort and reduced error potential, thereby improving the quality of the metadata. As registered users will be able to edit metadata through a reviewed process, DALIA will support the RDM and data literacy communities with functionalities for enhancing both the quantity and quality of metadata, advancing the goal of FAIR OER.

Acknowledgements

This project with the funding code 16DWWQP07A was funded by the Federal Ministry of Education and Research (BMBF) and the funding measure from the EU's Capacity Building and Resilience Facility.

Authorship Contributions

All authors were writing, reviewing, and editing.

Conflict of Interest

There are no conflicts of interest.

Bibliography

Barker, Phil, and Stuart Sutton, editors. 2022. "LRMI Terms (RDF) 2022-06-14". Visited on March 27, 2025. https://www.dublincore.org/specifications/lrmi/lrmi_terms/2022-06-14/.

Biernacka, Katarzyna, Claudia Beatriz Castillo Upiachihua, Ron Dockhorn, Claudia Engelhardt, Laura Grunwald-Eckhardt, Kerstin Helbig, Juliane Jacob, et al. 2024. "Train-the-Trainer Concept on Research Data Management". https://doi.org/10.5281/zenodo.13927614.

Coyle, Karen, editor. 2022. "Elements for DC Tabular Application Profiles". With contributions from Tom Baker, Phil Barker, John Huck, and Nishad Thalhath. Visited on March 27, 2025. https://www.dublincore.org/specifications/dctap/elements/.

DCMI Usage Board. 2020. "DCMI Metadata Terms". Visited on March 27, 2025. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

Garcia, Leyla, Bérénice Batut, Melissa L. Burke, Mateusz Kuzak, Fotis Psomopoulos, Ricardo Arcila, Teresa K. Attwood, et al. 2020. "Ten simple rules for making training materials FAIR". Edited by Scott Markel. *PLOS Computational Biology* 16 (5): e1007854. https://doi.org/10.1371/journal.pcbi.1007854.

Geiger, Jonathan D., Petra Steiner, Abdelmoneim Amer Desouki, Henrika Maria Hüppe, Frank Lange, Sonja Herres-Pawlis, Alessandra Kuntz, Ulrich Sax, and Jochen Ortmeyer. 2025. *Uniting the DIFferences with The DALIA Interchange Format (DIF) – a Minimal Metadata Specification for the DALIA OER Platform.* Zenodo. Poster at E-Science-Tage, Heidelberg, Germany. https://doi.org/10.5281/zenodo.15020137.

Geiger, Jonathan D., Petra Steiner, Abdelmoneim Amer Desouki, and Frank Lange. 2024. *DALIA Interchange Format (Version 1.3).* Zenodo. https://doi.org/10.5281/ZENODO.11521029.

Hoebelheinrich, Nancy J., Katarzyna Biernacka, Michelle Brazas, Leyla Jael Castro, Nicola Fiore, Margareta Hellström, Emma Lazzeri, et al. 2022. *Recommendations for a minimal metadata set to aid harmonised discovery of learning resources.* Zenodo. https://doi.org/10.15497/RDA00073.

Pohl, Adrian, Axel Klinger, Boris Hartmann, Carl Schuurbiers, Manuel Kummerländer, Manuel Oellers, Mikey Stengel, et al. 2024. "Allgemeines Metadatenprofil für Bildungsressourcen (AMB)". Entwurf vom 21. November 2024. Visited on March 27, 2025. https://dini-ag-kim.github.io/amb/draft/.

FORTH-BW: Erarbeitung und Bereitstellung eines bedarfsorientierten FDM-Workflows für HAWs

Karoline Herrmann [1,*], Ina Elisabeth Klingele [2], Lioba Petter [1], Carsten Herbes [1]

[1] Hochschule für Wirtschaft und Umwelt Nürtingen-Geislingen
[2] Hochschule für öffentliche Verwaltung Kehl
[*] Kontakt: karoline.herrmann@hfwu.de

In den letzten Jahren ist die Bedeutung des Forschungsdatenmanagements (FDM) auch an Hochschulen für Angewandte Wissenschaften (HAW) in Baden-Württemberg immer größer geworden. Im Verbundprojekt FORTH-BW haben sich die Verbundpartner zum Ziel gesetzt, FDM an HAW in Baden-Württemberg als Bestandteil des Forschungsprozesses zu etablieren und dafür die notwendigen Unterstützungsangebote zur Verfügung zu stellen. Der *FORTH-BW Forschungsdatenmanagement-Workflow* führt schrittweise durch die Antragsphase, die Forschungsphase und die Abschlussphase eines Forschungsprojekts. Der Bereich „Rechtliches“ komplettiert diese drei Phasen als Querschnittsthema und umfasst die rechtlichen Fragestellungen, die im Laufe des FDM-Prozesses geklärt werden müssen. Ergänzt wird der Workflow durch Handreichungen, die spezifische Aspekte des FDM vertiefen, sodass Forschende schnell und bedarfsgerecht Informationen finden. Der Workflow ist niedrigschwellig und selbsterklärend durch phasenspezifische Farben übersichtlich aufgebaut. Die Handreichungen sind durch einheitliche Kodierung einzelnen Prozessschritten zugeordnet.

Keywords: Workflow, FDM-Tool, Forschungsdatenmanagement, Hochschulen für Angewandte Wissenschaften

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23947 (CC BY-SA 4.0).

1 Ausgangslage

In den letzten Jahren ist die Bedeutung der Forschung auch an HAW in Baden-Württemberg immer größer geworden. Die Höhe der eingeworbenen Drittmittel steigt, was auch für die Zukunft als Ziel in den Struktur- und Entwicklungsplänen der Hochschulen festgelegt ist. Forschungsergebnisse werden in Journals, auch peer-reviewed, publiziert. Weil diese Journals die Autor*innen immer mehr zu Open Data drängen (vgl. Prosser u. a. 2022), ist es zum Teil heute schon und wird zukünftig vermehrt wichtig sein, die Daten in fachspezifischen Repositorien oder Datenzentren bereitzustellen und neben dem eigentlichen Artikel ebenfalls zu publizieren. War es in der Vergangenheit üblich, Forschungsdaten in unterschiedlichen Datenformaten, die nicht unbedingt für die Langzeitarchivierung geeignet sind, abzuspeichern, beliebig zu archivieren und dabei keinem einheitlichen Vorgehen zu folgen, so ist dies heute kein gangbarer Weg mehr für Forschende an HAWs und widerspricht der guten wissenschaftlichen Praxis (Deutsche Forschungsgemeinschaft e.V. 2019). Nicht nur verlangen Verlage als Grundlage für die Veröffentlichung qualitativ hochwertige Daten, auch Fördergeber (vgl. Montague-Hellen und Montague-Hellen 2023) – etwa die Europäische Union, in Deutschland bspw. die Deutsche Forschungsgemeinschaft (DFG) oder die Bundesanstalt für Landwirtschaft und Ernährung (BLE) – fordern immer öfter Datenmanagementpläne schon bei der Beantragung eines Vorhabens.

Damit Forschende an HAW die vielen unterschiedlichen Anforderungen von Fördergebern und Verlagen sowie die Vorgaben der guten wissenschaftlichen Praxis und der gesetzlichen Regelungen erfüllen können, benötigen sie zweierlei: Beratungsressourcen, die im gesamten FDM-Prozess – von der Antragsphase eines Forschungsprojekts bis hin zur Langzeitarchivierung – unterstützen, und strukturierte Anleitungen für die Organisation der Daten, ihre Aufbereitung und Dokumentation sowie Archivierung – unter Beachtung und Beantwortung der rechtlichen Fragestellungen. An HAW existieren diese Strukturen und Anleitungen bislang höchstens in Ansätzen. Die Personalressourcen für die Unterstützung beim FDM sind minimal: So sind die Servicestellen der HAW – Forschungsreferate (IAF, Institutszentren für Angewandte Forschung), IT-Abteilungen und Hochschulbibliotheken – mit den vorhandenen Ressourcen nicht in der Lage, diese Unterstützung zu bieten (vgl. Heister u. a. 2025, S. 7).

Im Verbundprojekt „Entwicklung und Implementierung eines bedarfsgerechten Forschungsdatenmanagements an HAW in Baden-Württemberg“ (FORTH-BW) haben sich die Projektpartner Hochschule für Wirtschaft und Umwelt Nürtingen-Geislingen (HfWU, Prof. Dr. Herbes, Herrmann, Petter), Hochschule für öffentliche Verwaltung Kehl (Prof. Dr. Klingele), Hochschule Reutlingen, HTWG Hochschule Konstanz, Hochschule für Technik Stuttgart und Hochschule der Medien Stuttgart unter der Leitung der HfWU zum Ziel gesetzt, FDM an HAW in Baden-Württemberg als Bestandteil des Forschungsprozesses zu etablieren und dafür die notwendigen Unterstützungsangebote zur Verfügung zu stellen.

2 Forschungsstand

Zahlreiche Studien in verschiedenen Disziplinen haben sich mit den Faktoren beschäftigt, die das Verhalten von Forschenden bezogen auf Forschungsdaten beeinflussen. Die Studien zu Einflussfaktoren auf FDM-Aktivitäten beziehen sich dabei vor allem auf den Aspekt des Teilens von Forschungsdaten, aber auch die Nachnutzung von Daten, die durch andere Forscher zur Verfügung gestellt werden.

Die Faktoren, die das Teilen von Forschungsdaten beeinflussen, lassen sich im Wesentlichen in drei Hauptkategorien einteilen: die erste Kategorie umfasst die institutionellen Faktoren, einschließlich der Richtlinien der Förderorganisationen und Zeitschriften sowie der Verträge mit Sponsoren. Ein weiterer Faktor sind die vorhandenen informationstechnischen Ressourcen, einschließlich der Repositorien (auch) für Forschungsdaten und der zugehörigen Metadaten. Die dritte Kategorie umfasst individuelle Faktoren wie Persönlichkeitsmerkmale, den empfundenen Nutzen, wie eine erhöhte Sichtbarkeit und eine erhöhte Zitationsrate, den Aufwand und das empfundene Risiko (vgl. Fecher, Friesike und Hebing 2015; Kim und Stanton 2015; Zuiderwijk und Spiers 2019).

Neben den Faktoren, die das Data Sharing positiv beeinflussen, wurden auch die Hinderungsgründe und Bedenken untersucht. Gründe, die Forschende davon abhalten, ihre Daten zu teilen sind die Notwendigkeit, zuerst den eigenen Artikel, dem die Daten zugrunde liegen, zu veröffentlichen. Denn viele Wissenschaftler*innen fürchten durch das Bereitstellen der Daten von Konkurrent*innen bei der Veröffentlichung von Forschungsergebnissen und Ideen überholt zu werden (vgl. Pook-Kolb 2021, S. 191). Weitere Hinderungsgründe sind Zeitmangel, fehlende Berechtigungen und die fehlende Finanzierung (vgl. Tenopir u. a. 2020, S. 21). Eine verbreitete Angst unter Forschenden ist auch, dass bei der Wiederverwendung der Daten durch andere, Fehler oder Diskrepanzen in der eigenen Forschung aufgedeckt werden können (vgl. Gorgolewski, Margulies und Milham 2013, S. 1). Ein weiterer Hinderungsgrund beim Teilen von Daten ist, dass Forschende bei der Bereitstellung der Daten nicht ohne Weiteres den gesamten Kontext, in dem die Daten entstanden sind, mitteilen können, was zu Fehlinterpretationen führen kann. Viele Forschende sehen die Daten außerdem noch immer als ihr persönliches „Eigentum“ an – obwohl es juristisch kein Eigentum an Daten, sondern Nutzungsrechte gibt (vgl. Hübner 2024, S. 1) – auch wenn diese Denkweise als egoistisch kritisiert wird (vgl. Kurata, Matsubayashi und Mine 2017, S. 3). Es existiert eine Diskrepanz zwischen der positiven Wahrnehmung des Data Sharing und dem tatsächlichen Verhalten der Forschenden (vgl. Tenopir u. a. 2020, S. 23).

Die Frage nach der Anwendung und der Akzeptanz von Datenmanagementplänen wurde bisher nur wenig untersucht. Blümm u. a. geben in ihrer Studie aus 2023 eine Bestandsaufnahme sowie Empfehlungen zu Datenmanagementplänen an Hochschulen. Die Analyse von FDM-Policies und Experteninterviews ergab, dass es an Bewusstsein und Sichtbarkeit für Datenmanagementpläne, aber auch entsprechenden Unterstützungsangeboten man-

gelt (vgl. Blümm u. a. 2023, S. 117). Laut Schmiederer und Kuberek (2022, S. 7) ist die Akzeptanz für die Nutzung von Datenmanagementplänen seit dem Programm Horizon 2020 gestiegen, jedoch ist die Verpflichtung von Seiten der Fördermittelgeber bei der tatsächlichen Nutzung von DMPs entscheidend und hier differieren die Vorgaben zwischen den Forschungsförderern erheblich. Die Literaturrecherche ergab außerdem, dass die Unterstützung beim Erstellen von Datenmanagementplänen gering ist und die Forschenden unzufrieden mit den vorhandenen Unterstützungsangeboten sind (vgl. Tenopir u. a. 2020, S. 13).

3 Forschungsdatenmanagement-Workflow

Eine Erhebung mittels eines mixed-methods-Ansatzes (32 qualitative Interviews und Umfrage mit 279 Teilnehmenden unter verschiedenen Stakeholdern: Forschende, IT- und Bibliotheksmitarbeitende, Justiziare) des FDM an den beteiligten HAW im Verbundprojekt hat gezeigt: Die Wichtigkeit, Forschungsdaten zu managen, ist bekannt und als soziale Norm verinnerlicht. Das Verhalten der Forschenden zeigt, dass diese Norm noch nicht umgesetzt wird, denn die intrinsische Motivation, FDM zu betreiben, ist gering, der Aufwand wird als hoch angesehen. Darüber hinaus werden rechtliche Aspekte als Hinderungsgründe für die Veröffentlichung von Forschungsdaten betont. Auch fehlen an HAW das nötige Wissen, das Handwerkszeug und die Strukturen. Diese projektinternen Erkenntnisse decken sich mit Erfahrungsberichten von Mitarbeitenden in Serviceeinreichtungen an anderen HAW in Deutschland (vgl. Heister u. a. 2025) und mit den Befunden der oben genannten Literatur (bspw. Kim und Stanton 2015; Tenopir u. a. 2020).

Auf diese Bedarfe hat das Projektteam mit der Bereitstellung des *FORTH-BW Forschungsdatenmanagement-Workflows* reagiert. Der Workflow wurde seit Oktober 2022 basierend auf der Bedarfsermittlung und unter Berücksichtigung von Aspekten des Open-Science, rechtlicher Grundlagen sowie existierender IT-Tools erarbeitet. Dafür stand die Expertise des interdisziplinären FORTH-BW-Projektteams zur Verfügung. Der Workflow ist niedrigschwellig und selbsterklärend durch phasenspezifische Farben übersichtlich aufgebaut und kann so intuitiv und selbständig erfasst werden. Die Handreichungen und Tools sind durch einheitliche Kodierung ebenso wie verschiedene Ansprechstellen einzelnen Prozessschritten zugeordnet. Abbildung 1 zeigt den schematischen Aufbau des Workflows.

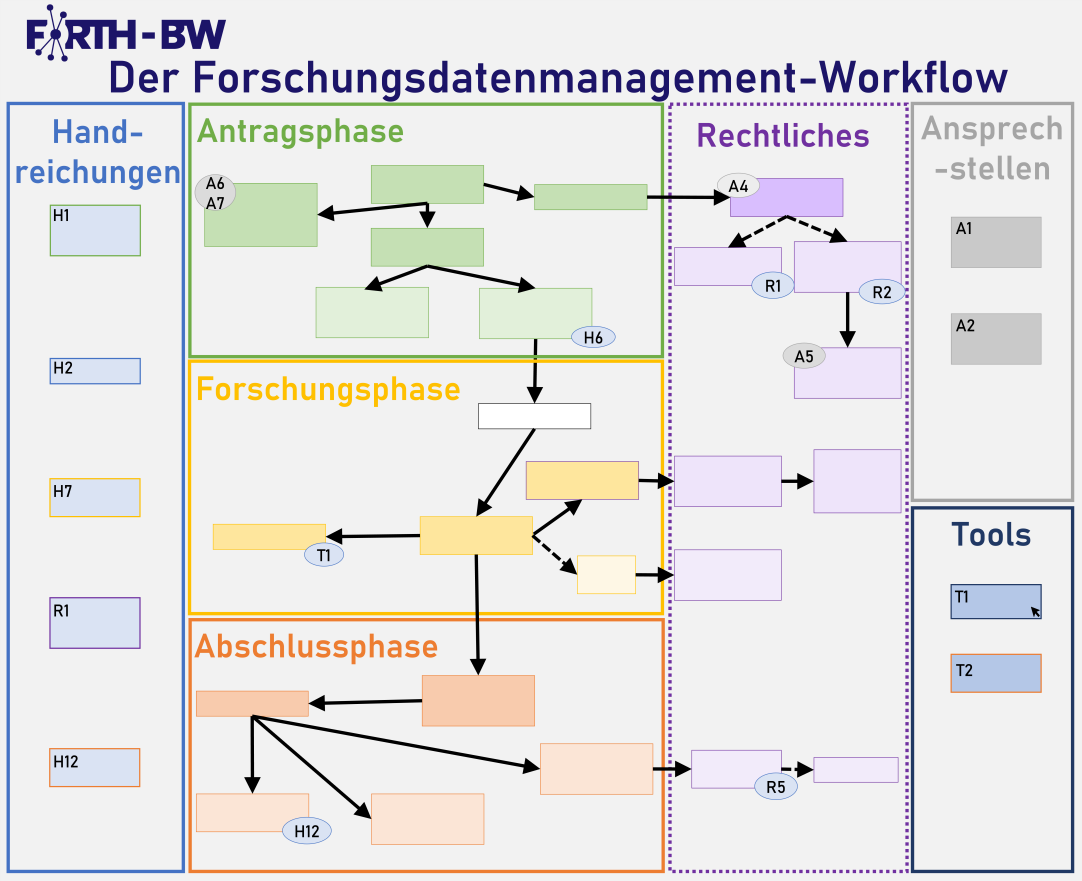


Abbildung 1: Schematische Darstellung des FORTH-BW FDM-Workflows.

Die Darstellung des Datenlebenszyklus ist im FDM geläufig. In einzelnen Beratungssituationen an den beteiligten HAW ist jedoch deutlich geworden, dass die überschriftartigen Bezeichnungen der einzelnen Zyklusphasen zu unspezifisch sind, um eine direkte Anwendung durch die Forschenden zu ermöglichen. Daher wurden diese Phasen im FORTH-BW-Projekt in eine Prozessdarstellung überführt: den FDM-Workflow. Dieser führt schrittweise durch die Antragsphase, die Forschungsphase und die Abschlussphase eines Forschungsprojekts. Der Bereich „Rechtliches“ komplettiert diese drei Phasen als Querschnittsthema und beantwortet die rechtlichen Fragestellungen, die im Laufe des FDM geklärt werden müssen. Ergänzt wird der Workflow auf der einen Seite durch Ansprechstellen – wie die Forschungsreferate, Datenschutzbeauftragte oder die Hochschulbibliothek – die in einzelnen Prozessschritten punktuell zur Verfügung stehen und für einzelne Beratungsleistungen angefragt werden können. Auf der anderen Seite finden sich empfehlenswerte Tools wie der Research Data Management Organiser[1] für die Erstellung von Datenmanagementplänen, und Handreichungen, die spezifische Aspekte des FDM so konzise wie möglich vertiefen, sodass Forschende schnell und bedarfsgerecht Informationen finden. Somit ist keine kontinuierliche individuelle Beratung in allen Forschungspha-

1 https://rdmorganiser.github.io/; *Besucht am 9. Mai 2025.*

sen notwendig, was die Servicestellen an den HAW entlastet. Folgende Handreichungen flankieren den Workflow inhaltlich: *Formulierungshilfen für die Antragstellung*, *Repositoriensuche*, *Repositorienvergleich*, *Repositorienwahl*, *Fachspezifische Repositorien, Datenspeicherung, Gute Datenorganisation, Sicherheitsmanagement, Dateiformate, Datendokumentation, RDMO-Kurzanleitung (deutsch und englisch), Datenbewertung und -löschung, Vergleich generischer Datenjournale.*

Hervorzuheben sind Handreichungen zu rechtlichen Aspekten, welche die große Bedarfslücke auf diesem Gebiet schließen: *Erstellung einer anonymen Umfrage, Datenschutz bei personenbezogenen Daten, Gemeinsame Verantwortlichkeit, Datenmanagement mit Auftragsverarbeitern, Nutzungsrechte, Lizenzen für Forschungsdaten.* Diese Handreichungen werden um Vorlagen bzw. Verweise auf bestehende Muster ergänzt: *Information personenbezogene Daten, Vertrag zur Verarbeitung personenbezogener Daten, Vertrag zur gemeinsamen Verantwortlichkeit, Beitrag zum Verzeichnis der Verarbeitungstätigkeiten, Datennutzungsvereinbarung, Satzung Ethikkommission.* Insbesondere die Handreichungen zu rechtlichen Aspekten sind in Beratungsgesprächen sehr nachgefragt.

4 Nachnutzung

An der HfWU wird der Workflow in der individuellen FDM-Beratung bereits erfolgreich eingesetzt. In der aktuellen Projektlaufzeit bis zum Abschluss im September 2025 wird der Workflow sowohl an den Verbundhochschulen als auch im FORTH-BW-Netzwerk in verschiedenen Projektphasen und Forschungsdisziplinen erprobt, sodass einzelne Prozessschritte ggf. nachgebessert oder geschärft werden können. Die Handreichungen werden sukzessive auf den projekteigenen Wikiseiten[2] veröffentlicht. Hier steht auch der Workflow zum Download bereit.

Das FORTH-BW – Projektteam ist im deutschlandweiten Netzwerk „FDM@HAW“ aktiv, in dem sich Akteure von HAW – bspw. der Projekte FDM-ndsHAW, FDM-HAWK, FDM@HAW.rlp, FitforFDM der BMBF-Förderrichtlinie zur Förderung von Projekten zum Thema Nachnutzung und Management von Forschungsdaten an Fachhochschulen – zu FDM-relevanten Themen austauschen und an FDM-Entwicklungen partizipieren. Im Zuge des weiteren Wissenstransfers werden der Workflow sowie die Handreichungen zum Projektende unter CC BY-NC – Lizenz auf dem Repositorium Zenodo bereitgestellt. Mögliche Weiterentwicklungen werden im Zug der Verstetigung an den einzelnen Verbundhochschulen stattfinden.

2 https://forth-bw.hfwu.de; *Besucht am 19. März 2025.*

Danksagungen

Das Projekt „FORTH-BW: Entwicklung und Implementierung eines bedarfsgerechten Forschungsdatenmanagements an HAW in Baden-Württemberg“ wird unter FKZ 16FDFH-120A (HfWU) und 16FDFH120C (Hochschule Kehl) vom Bundesministerium für Forschung und Entwicklung mit Mitteln der Aufbau- und Resilienzfazilität der Europäischen Union, NextGenerationEU gefördert.

Beiträge

- Karoline Herrmann: Konzept, Verschriftlichung – Originalentwurf, Projektadministration.
- Ina Klingele: Konzept, Verschriftlichung – Überprüfen und Redigieren.
- Lioba Petter: Verschriftlichung – Originalentwurf.
- Carsten Herbes: Verschriftlichung – Überprüfen und Redigieren, Supervision, Akquisition der Fördermittel.

Interessenkonflikte

Es gibt keinen Interessenkonflikt.

Literaturverzeichnis

Blümm, Mirjam, Christine Burkart, Maria Chlastak, Katharina Fritsch, Heike Neuroth, Andreas Schieberle, Stefan Schmunk und Stephanie Werner. 2023. „Datenmanagementpläne an Fachhochschulen /Hochschulen für Angewandte Wissenschaften Eine Bestandsaufnahme und Empfehlungen“. In *Schriften zur Informationswissenschaft,* Bd. 77. Hochschulverband Informationswissenschaft (HI) e.V. https://doi.org/10.5281/ZENODO.8275739.

Deutsche Forschungsgemeinschaft e.V. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis.* Zenodo. https://doi.org/10.5281/zenodo.6472827.

Fecher, Benedikt, Sascha Friesike und Marcel Hebing. 2015. „What Drives Academic Data Sharing?“ *PLoS ONE* 10 (2): e0118053. https://doi.org/10.1371/journal.pone.0118053.

Gorgolewski, Krzysztof J., Daniel S. Margulies und Michael P. Milham. 2013. „Making Data Sharing Count: A Publication-Based Solution“. *Frontiers in Neuroscience* 7. https://doi.org/10.3389/fnins.2013.00009.

Heister, Eske, Karoline Herrmann, Diana Schmidt und Esther Schneidenbach. 2025. *Forschungsdatenmanagement an HAWs/FHs: Strategien, Herausforderungen und Lösungsansätze.* Zenodo. https://doi.org/10.5281/ZENODO.15062819.

Hübner, Andreas. 2024. *Wem "gehören"Forschungsdaten?* Zenodo. https://doi.org/10.5281/ZENODO.11077412.

Kim, Youngseek und Jeffrey M. Stanton. 2015. „Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis“. *Journal of the Association for Information Science and Technology* 67 (4): 776–799. https://doi.org/10.1002/asi.23424.

Kurata, Keiko, Mamiko Matsubayashi und Shinji Mine. 2017. „Identifying the Complex Position of Research Data and Data Sharing Among Researchers in Natural Science“. *Sage Open* 7 (3). https://doi.org/10.1177/2158244017717301.

Montague-Hellen, Beth und Kate Montague-Hellen. 2023. „Publishers, funders and institutions: who is supporting UKRI-funded researchers to share data?“ *Insights* 36. https://doi.org/10.1629/uksg.602.

Pook-Kolb, Michaela. 2021. *Teilen oder nicht teilen: Die Logik des Schützens von Forschungsdaten.* Springer Fachmedien Wiesbaden. ISBN: 9783658353001. https://doi.org/10.1007/978-3-658-35300-1.

Prosser, Annayah M. B., Richard J. T. Hamshaw, Johanna Meyer, Ralph Bagnall, Leda Blackwood, Monique Huysamen, Abbie Jordan, Konstantina Vasileiou und Zoe Walter. 2022. „When open data closes the door: A critical examination of the past, present and the potential future for open data guidelines in journals“. *British Journal of Social Psychology* 62 (4): 1635–1653. https://doi.org/10.1111/bjso.12576.

Schmiederer, Simon und Monika Kuberek. 2022. „Forschungsdaten-Policy für Forschungsprojekte im Spannungsfeld zwischen Kooperationsvertrag und Datenmanagementplan“. *Bausteine Forschungsdatenmanagement,* Nr. 2, https://doi.org/10.17192/BFDM.2022.2.8446.

Tenopir, Carol, Natalie M. Rice, Suzie Allard, Lynn Baird, Josh Borycz, Lisa Christian, Bruce Grant, Robert Olendorf und Robert J. Sandusky. 2020. „Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide“. Herausgegeben von Sergi Lozano. *PLoS ONE* 15 (3): e0229003. https://doi.org/10.1371/journal.pone.0229003.

Zuiderwijk, Anneke und Helen Spiers. 2019. „Sharing and re-using open data: A case study of motivations in astrophysics“. *International Journal of Information Management* 49:228–241. https://doi.org/10.1016/j.ijinfomgt.2019.05.024.

LLM4DDC: Adopting Large Language Models for Research Data Classification Using Dewey Decimal Classification

Gautam Kishore Shahi [1], Renat Shigapov [2], Oliver Hummel [1]

[1] Technical University of Applied Sciences, Mannheim, Germany;
[2] University of Mannheim, Mannheim, Germany

Classifying research data in institutional repositories is time-consuming and challenging. While the Dewey Decimal Classification (DDC) system is widely used in subject classification for texts, its application to research data metadata has been limited so far. This study explores the possible use of large language models (LLMs) and small language models (SLMs) for the automatic classification of research data in the context of DDC. This study uses sample data from an existing dataset compiled from different institutions mainly in Germany. We use a prompt engineering approach for LLMs, and fine tuning for SLMs, where we use RoBERTa as a baseline. Our results show that LLMs with prompt engineering currently are not able to classify metadata of research data into DDC classes as good as SLMs with fine tuning. To foster adoption, we openly release our models, code, and datasets for integration into research data infrastructures at GitHub.[1]

Keywords: Dewey Decimal Classification, Research Data, Large Language Model, Automated Classification

1 Introduction

As the volume of research data continues to grow, accurately classifying this data in institutional repositories remains a significant challenge, as it requires significant human

1 https://github.com/TransforMA-WP3/LLM4DDC; *Visited by the editors on June 30, 2025.*

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23948 (CC BY-SA 4.0).

effort to assign the appropriate subject classification. While the Dewey Decimal Classification (DDC) system is widely used for automatic subject indexing (Golub 2021) in the context of libraries, its application to automating the creation of metadata for research data is at best a fledgling discipline (Weber et al. 2020). We address this gap by exploring the use of large language models (LLMs) and small language models (SLMs) in automating the detection of research areas for a DDC classification of research data (Hu et al. 2024). This has practical implications for numerous data repositories, including the German National Research Data Infrastructure (NFDI), where accurate metadata is crucial for effective research data management and retrieval.

Different approaches have been tested for the automatic classification of research data. Recent research (Ho et al. 2024) used traditional Machine Learning (ML) models for identifying DDC classifications, having DDC classes in the range of 400 to 499. Our current focus is on identifying DDC classes at the domain and subject level out of the three-level DDC classification system (ref. Section 2.2) to better understand the feasibility of such an approach. For the LLMs we were aiming on optimizing model parameters such as temperature and leveraging prompt-engineering strategies with zero-shot and few-shot prompts (Shahi and Hummel 2025). For SLMs, we used fine-tuned versions of pre-trained models, such as RoBERTa from Hugging Face[2]. In both settings, we used title and description of the research data to identify the best matching DDC classes and evaluated the results with precision, recall and F1-score.

The remainder of the paper is organized as follows; Section 2 describes the background, Section 3 describes an experiment, Section 4 describes the results and their analysis, before finally, Section 5 provides the conclusion and discusses future work.

2 Background

2.1 Contemporary Language Models

Large Language Models (LLMs) have evolved significantly over the last three years based on the advancement of deep learning which was mainly caused by improved computing performance through GPUs as well as the availability of large-scale datasets for model training (Zhou et al. 2024). Early language models, such as word embeddings (Word2Vec Church 2017), laid the foundation for contextualized and semantic text representations. The introduction of transformer-based architectures, such as Google's bidirectional transformers for language understanding (BERT; Devlin et al. 2019), revolutionized natural language processing (NLP) by enabling models to generate coherent and contextually rich text. However, BERT is still considered being a Small Language Model (SLM)[3].

2 https://huggingface.co/; *Visited by the editors on June 30, 2025.*

3 https://www.ibm.com/think/topics/small-language-models; *Visited by the editors on June 30, 2025.*

In scholarly communication, LLMs have been increasingly utilized to enhance various aspects of scientific publishing. Researchers have also been leveraging LLMs for subject tagging (Shahi and Hummel 2025) or extracting metadata from papers (Watanabe, Ito, and Matsubara 2025). Moreover, SLMs have already been used for assigning DDC classes to scientific papers (Ho et al. 2024). Hence, it seems worthwhile to explore both SLMs and LLMs for DDC classification in this study.

2.2 DDC

The Dewey Decimal Classification (DDC) system is a method for organizing books and other materials in libraries. Created by Melvil Dewey in 1876 (Dewey 1899), it assigns a unique number to each subject, making it easy to group publications and find related materials. We used the 23rd edition of DDC released in 2011 (Dewey et al. 2011). The classification system is divided into three levels; the first level consists of ten domains, then each domain has ten subjects, each comprising ten more specific topics; hence the "decimal" in the name. Each domain, subject, and topic is assigned a number (Dewey 1899). For example, 500 is assigned for domain *science*, while 530 stands for *Physics* and 532 is *Fluid Mechanics* as subcategory of science and physics. Due to its simplicity, in combination with a comparatively good topic coverage, the DDC is widely used in libraries worldwide (Wang 2009).

3 Approach & Experiments

The approach we used to classify research data into their DDC classes is illustrated in Figure 1. The pipeline consists of dataset collection, data cleaning, building classification models based on contemporary LLMs and SLMs, and finally evaluating the results obtained. Each step is explained in more detail below. The replication package for this work is available at (Shahi, Shigapov, and Hummel 2025).

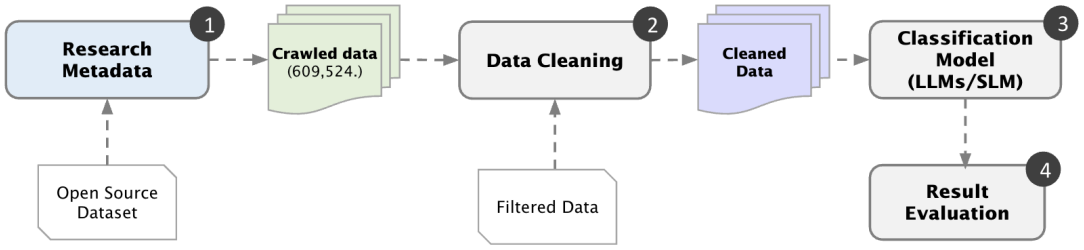


Figure 1: Overview of evaluation approach.

3.1 Dataset

We used a list of research data collected from Weber et al. (2020) as ground truth. This dataset is compiled from 29 different sources such as various German university libraries, Arxiv and some other data sources. In total, Weber et al. have collected around 16 million records, however, the final dataset they have shared consists of only roughly 610,000 records. Out of this, we filtered around 222,000 records actually having valid DDC classes.

Moreover, the dataset was highly imbalanced, so we randomly sampled 1,000 entries from those 14 subjects actually containing more than 1,000 entries, obtaining the total of 14,000 records used in this study. The subjects spanned 6 different domains (Science, Technology, Computer Science Information & General Works, Language, Philosophy & Psychology, Social Sciences).

Preprocessing & Cleaning

Prior to selecting the actual records for our study, we cleaned the dataset by removing research data titles or texts that were not in English or had an obviously wrong DDC number assigned, such as *488348*. Moreover, we filtered out records that included only a broad domain without specifying a subject. For example, DDC number 551 corresponds to the domain *Science*, but no specific subject is provided, which would have caused issues in our experiments.

3.2 Classification Model

We used the cleaned data for the evaluation of our classification models. A concatenation of text and title was used to yield DDC number and class name from the LLMs and SLMs. Both classification models were implemented for the domain and subject levels of DDC (cf. Section 3.1).

Evaluated Models

We used the following freely available open-source LLMs based on the Ollama framework for our experiments.[4] Ollama is a tool designed to run LLMs containing a large variety of LLMs. It allows downloading LLMs with various parameter sizes. We chose to use Llama,

4 https://ollama.com/library; *Visited by the editors on June 30, 2025.*

Mistral, and Gemma as described in the following: *Llama 3.1* is a state-of-the-art open-source language model developed by Meta for natural language processing tasks. In our experiments, we used the 70-billion-parameter version. *Gemma* is a family of lightweight language models created by Google DeepMind for efficient NLP. We utilized Gemma 2.7, which has 27 billion parameters. *MistralLite* is a compact and efficient language model developed by Mistral AI, optimized for fast inference and strong performance across a range of NLP tasks. It is the smallest model in our setup, with only 7 billion parameters.

To invoke the above LLMs, we used zero and few-shot prompts as has proven successful in our previous work (Shahi and Hummel 2025). In the zero-shot (prompt 1), we ask the LLM to classify research data for DDC without providing any additional explanation. In the few-shot prompt 2, we provided the hierarchical structure of DDC domain and subjects and asked the LLM for an appropriate classification. In both prompts, we asked for DDC numbers and classes, which we then used for evaluation (cf. Section 4).

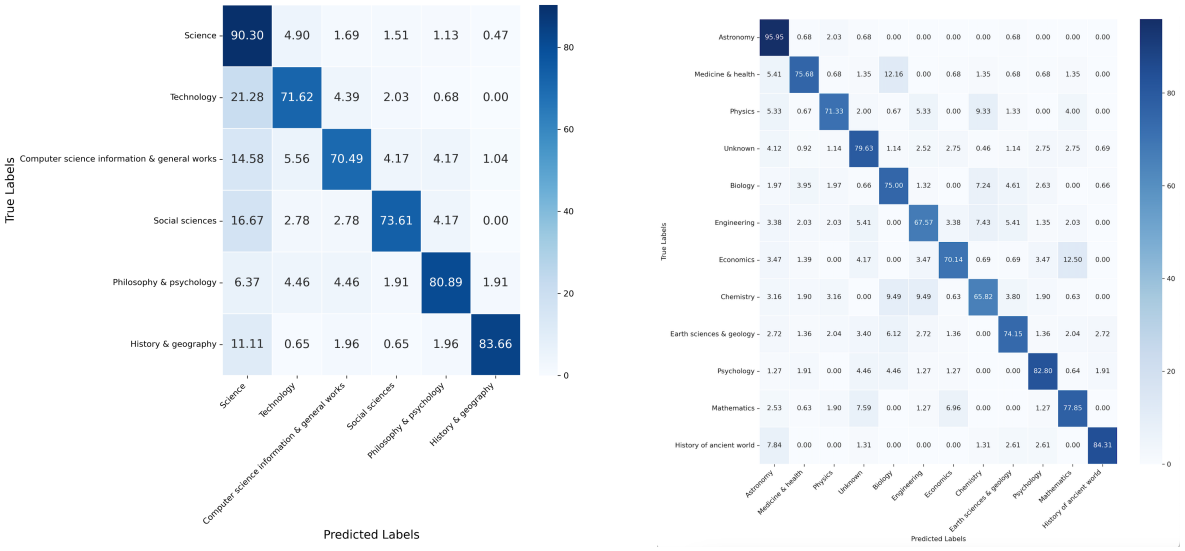
For the baseline comparison with an SLM, we used the pre-trained baseline models *BERT* (Devlin et al. 2019), *RoBERTa* (Liu et al. 2019), and *DistilBERT* (Sanh et al. 2019) model from Huggingface and fine-tuned them with with a part of the dataset described above. For these classification models, text was used as feature and domain and subject names were provided as target variable.

4 Evaluation & Result Analysis

We evaluated the classification performance with precision, recall, and F1 scores, calculated separately for both classification levels: domain and subject. At the domain level, we considered six different domains, while the subject level test data included 14 distinct classes overall.

For the LLMs, the F1 score ranged from 0.15 to 0.43, and they struggled to predict both DDC numbers and subjects accurately. In contrast, the fine-tuned SLMs performed much better, as we found that RoBERTa performed best from all three models (BERT, RoBERTa, DistilBERT), achieving F1 scores between 0.83 and 0.96 (in the subject Astronomy 520). Figure 2 shows the confusion matrix for the best domain and subject classification yielded with RoBERTa.

A manual analysis of the LLM's outputs revealed that they often provided only partial DDC classifications. For instance, Llama (LLM) predicted Philosophy instead of the correct class Philosophy & Psychology, suggesting a potential leverage for relatively simple future improvements. Overall, classification at the subject level was more accurate than at the domain level. The best results were obtained for the Science domain and its subjects, followed by Technology.



(a) Top domain. (b) Subjects.

Figure 2: Confusion matrix showing results for different classification approach.

Overall, the presented approach still faces numerous challenges and leaves room for improvement. One is that there are different versions of DDC available so that the development of a classification system remains challenging, especially for LLMs without specific training. The relatively poor quality of the reference data that merely partially covers the DDC, currently also hinders the improvement of this approach. Although, different bibliographic sources and libraries provide data aligned with DDC classes, they are often skewed to specific domains or subjects, such as in the above dataset, where more than 75 % of the records belong to the Science domain.

5 Conclusion & Future Work

In the presented study, we used large and small language models to classify research metadata based on the DDC. The results show that finetuned SLMs still easily outperform generic LLMs for the classification of research data. We also discussed the main challenges that hinder LLM performance such as different variants of the DDC classification system. However, based on the obtained results, the SLM-based approach has a good potential to be used as an automatic classification model for assigning DDC classes to research data in the near future.

For future work, potential extensions include incorporating more training and test data from more diverse sources, which should also lead to a coverage of additional domains and subjects from the DDC. This could improve the overall performance and ensure a more

balanced distribution of subjects, enhancing generalizability and quality especially with pre-trained SLM-based models.

Data and code availability statement

The replication package, including the code and dataset used in this paper, is openly available at GitHub[5] and is archived[6] (Shahi, Shigapov, and Hummel 2025). If you reuse the dataset, additional attribution to Weber (2019) is needed. The replication package is licensed under the MIT license (for code) and Creative Commons Attribution 4.0 International license (for everything else).

Acknowledgements

This work has been carried out under the TransforMA project, which has received funding from the federal-state initiative "Innovative Hochschule" of the Federal Ministry of Education and Research (BMBF) in Germany.

Authorship Contributions

- Gautam Kishore Shahi: Writing – original draft, experiments;
- Renat Shigapov: Writing – review & editing;
- Oliver Hummel: Writing – review & editing.

Conflict of Interest

Beyond the funding declared under Acknowledgements, all authors have no conflict of interest to declare.

5 https://github.com/TransforMA-WP3/LLM4DDC; *Visited by the editors on June 30, 2025.*

6 https://www.doi.org/10.7801/479 or https://madata.bib.uni-mannheim.de/479; *Visited by the editors on June 30, 2025.*

Bibliography

Church, Kenneth Ward. 2017. “Word2Vec”. *Natural Language Engineering* 23 (1): 155–162. https://doi.org/10.1017/s1351324916000334.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers),* 4171–4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423.

Dewey, Melvil. 1899. *Decimal Classification and Relativ Index for Libraries: Clippings, Notes, Etc.* Library bureau.

Dewey, Melvil, Joan S. Mitchell, Julianne Beall, Rebecca Green, Giles Martin, and Michael Panzer. 2011. *Dewey decimal classification and relative index.* 23rd edition. Dublin, Ohio: OCLC Online Computer Library Center.

Golub, Koraljka. 2021. “Automated Subject Indexing: An Overview”. *Cataloging and Classification Quarterly* 59 (8): 702–719. https://doi.org/10.1080/01639374.2021.2012311.

Ho, Clara Wan Ching, Tobias Weber, Thorsten Fritze, and Thomas Risse. 2024. “Towards Multilingual LLM-Based Approaches for Automatic Dewey Decimal Classification”. In *International Conference on Theory and Practice of Digital Libraries,* 23–33. Springer, Springer Nature Switzerland. ISBN: 9783031724404. https://doi.org/10.1007/978-3-031-72440-4_3.

Hu, Linmei, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. “Llm vs small model? large language model based text augmentation enhanced personality detection model”. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 38:18234–18242. 16. https://doi.org/10.1609/aaai.v38i16.29782.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* https://doi.org/10.48550/ARXIV.1907.11692.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* arXiv. https://doi.org/10.48550/ARXIV.1910.01108.

Shahi, Gautam Kishore, and Oliver Hummel. 2025. “On the Effectiveness of Large Language Models in Automating Categorization of Scientific Texts”. In *Proceedings of the 27th International Conference on Enterprise Information Systems,* 544–554. SCITEPRESS – Science / Technology Publications. https://doi.org/10.5220/0013299100003929.

Shahi, Gautam Kishore, Renat Shigapov, and Oliver Hummel. 2025. *Replication package for "LLM4DDC: Adopting Large Language Models for Research Data Classification Using Dewey Decimal Classification"*. Dataset. Visited on June 30, 2025. https://doi.org/10.7801/479. https://madata.bib.uni-mannheim.de/479/.

Wang, Jun. 2009. "An extensive study on automated Dewey Decimal Classification". *Journal of the American Society for Information Science and Technology* 60 (11): 2269–2286. https://doi.org/10.1002/asi.21147.

Watanabe, Yu, Koichiro Ito, and Shigeki Matsubara. 2025. "Capabilities and Challenges of LLMs in Metadata Extraction from Scholarly Papers". In *International Conference on Asian Digital Libraries,* 280–287. Springer. https://doi.org/10.1007/978-981-96-0865-2_23.

Weber, Tobias. 2019. *s-sized Training and Evaluation Data for Publication "Using Supervised Learning to Classify Metadata of Research Data by Field of Study"*. Zenodo. https://doi.org/10.5281/ZENODO.3490396.

Weber, Tobias, Dieter Kranzlmüller, Michael Fromm, and Nelson Tavares de Sousa. 2020. "Using supervised learning to classify metadata of research data by field of study". *Quantitative Science Studies* 1 (2): 525–550. https://doi.org/10.1162/qss_a_00049.

Zhou, Shutian, Zizhe Zhou, Chenxi Wang, Yuzhe Liang, Liangyu Wang, Jiahe Zhang, Jinming Zhang, and Chunli Lv. 2024. "A User-Centered Framework for Data Privacy Protection Using Large Language Models and Attention Mechanisms". *Applied Sciences* 14 (15): 6824. https://doi.org/10.3390/app14156824.

New Developments in RADAR: Safeguarding and Publishing Research Data for Long-term Usability

Kerstin Soltau [1,*], Stefan Hofmann [1], Sandra Göller [1], Felix Bach [1]

[1] FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur
[*] Corresponding author: info@radar-service.eu

RADAR[1], launched in 2017 as a cloud service, is a generic research data repository designed to securely archive, publish, and optionally review research data. Developed in a cooperation project funded by the DFG, it is now operated by FIZ Karlsruhe – Leibniz Institute for Information Infrastructure[2]. RADAR stores data according to the BagIt standard in a geo-redundant manner, maintaining three copies across two academic data centers, located at KIT[3] and TU Dresden[4] using tape archives as storage medium. With a minimum retention period of 25 years for published data and flexible archival options compliant with the DFG Code of Conduct, RADAR serves a variety of scientific disciplines. Agile software development has driven the development of the system, resulting in numerous improvements in terms of metadata annotation and FAIRness of metadata (e.g. OAI provider, FAIR signposting, Schema.org integration, a knowledge graph, SPARQL endpoints etc.) as well as functional enhancements, especially in terms of uploading research data (e.g. WebDAV protocol support and Git integration). In addition, community-tailored publication services developed in the framework of the NFDI[5] as well as alternative operating variants (RADAR Local and Hybrid) expand the range of services offered by RADAR and meet the needs of additional target groups. This paper provides an overview of RADAR's technical architecture, service enhancements, and its significant role in advancing research data management.

1 https://www.radar-service.eu/; *Visited on May 16, 2025*
2 https://www.fiz-karlsruhe.de; *Visited on May 16, 2025.*
3 https://www.kit.edu/; *Visited on May 16, 2025.*
4 https://tu-dresden.de/; *Visited on May 16, 2025.*
5 https://www.nfdi.de/; *Visited on May 16, 2025.*

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23949 (CC BY-SA 4.0).

Keywords: RADAR, Research Data, Research Data Management, Repository, FAIR data

1 Introduction

The exponential growth of research data over recent decades has necessitated robust, secure, and long-term data management solutions. Research institutions worldwide are increasingly challenged by the need to archive and share datasets while ensuring compliance with policies related to data retention, security, and accessibility. In this context, RADAR was conceived as a cross-disciplinary research data repository to address these challenges. Initially funded by the German Research Foundation (DFG)[6] during a 3-year project phase and launched as a cloud service for academic institutions in 2017, RADAR represents a significant step towards establishing an integrated infrastructure that supports the findability, accessibility, interoperability, and reusability of research data.

The primary objective of RADAR is to provide a secure, scalable, and user-friendly platform for archiving and publishing research data independent of scientific discipline. By leveraging cloud technologies and reliable archival techniques, RADAR offers a comprehensive solution tailored to the needs of research institutions, scientific communities, individual researchers, regional networks, and funding organizations. This paper presents an overview of the system architecture and data integrity, metadata management and the implementation of the FAIR principles, recent functional enhancements and user-focused features, and community-specific services and RADAR operating variants.

2 System Architecture and Data Integrity

RADAR was designed as a cloud-based service to address the challenges inherent in managing long-tail research data. The infrastructure is built on a modular architecture that ensures both scalability and resilience. At the heart of this system is a secure archival process that employs geo-redundancy – each dataset is stored in three copies across two different academic data centers at SCC[7] and ZIH[8]. This approach guarantees data integrity and availability even in the event of hardware failures and also mitigates risks associated with natural disasters or localized cyberattacks. The primary storage medium for RADAR are tape archives. Though often slower in terms of data retrieval compared to disk-based systems, it offers significant cost benefits and stability for archival purposes in the long term. The system is engineered such that the retention period for

6 https://www.dfg.de; *Visited on May 16, 2025.*
7 https://www.scc.kit.edu; *Visited on May 16, 2025.*
8 https://tu-dresden.de/zih; *Visited on May 16, 2025.*

published data is set to a minimum of 25 years, while archived data can be stored for flexible periods (e.g. 10 years in accordance with the DFG Code of Conduct[9]).

Data integrity is a key aspect of RADAR's design. Datasets are provided with a checksum (MD5 hash value) before being transferred to a tape archive. This checksum is validated after each copying process. The data centers also validate the checksums internally after migration to new tapes.

3 Metadata Management and FAIR Principles

A dataset annotated with meaningful and interoperable metadata increases its visibility, discoverability, reusability and reproducibility. Thus effective data management goes hand in hand with comprehensive metadata annotation. RADAR has developed its own generic metadata schema which is listed as FAIRsharing Standard (FAIRsharing Team 2023). It is based on the DataCite Metadata Kernel and comprises 10 mandatory and 13 optional elements. Over time, however, it became obvious that this generic approach is not sufficient for all use cases and that sometimes researchers need to describe their data according to their discipline- or project-specific needs. To meet these needs, in 2021 RADAR introduced a flexible subject-specific annotation option for custom defined XSD schemas that are either RADAR-compliant or generate the mandatory RADAR metadata by an additionally provided XSLT transformation.

The FAIR principles are integral to modern research data management. So far, RADAR has already implemented several approaches to maximise the FAIRness of datasets:

- **Operation of a dedicated OAI provider**, via which the metadata of all data publications are available for harvesting by third parties in various formats.
- **Integration of standard data in RADAR metadata**, e.g. for individuals (ORCID), organisations (ROR), funding organisations (Crossref Funder Registry), and keywords (GND, the Integrated Authority File).
- **Schema.org Integration**: Adoption of this widely recognized metadata standard, facilitating broader machine-readability, discoverability, and interoperability by internet search engines, like Google Dataset Search.
- **FAIR Signposting**: Enrichment of the source code of landing pages with typed links, ensuring machine readability and actionability in a uniform way by making data resources easy to find and reference.
- **RADAR Knowledge Graph and SPARQL Endpoint**: The development of a RADAR knowledge graph enables the semantic linking of data elements, while the SPARQL endpoint[10] provides researchers with powerful querying capabilities.

9 https://www.dfg.de/en/basics-topics/basics-and-principles-of-funding/good-scientific-practice/code-of-conduct; *Visited on May 16, 2025.*

10 https://www.radar-service.eu/sparql; *Visited on May 16, 2025.*

In 2023, RADAR participated in a support measure of the EU project FAIR-IMPACT (Soltau et al. 2024). During this FAIRness assessment process, the *F-UJI score*[11] as indicator for the FAIRness of data publications was significantly increased (up to 91 %). Starting in spring 2025, RADAR integrates the *TS4NFDI terminology service*[12] developed as part of the NFDI, which allows to include a large number of different discipline specific terminologies and terminology collections in several metadata fields (keywords, subject area etc.).

4 Functional Enhancements and User-Focused Features

To accommodate the various workflows of researchers and to allow data transfer even for larger data volumes, RADAR continuously evolves its data ingestion methods. Since early 2024, an alternative upload option via the *WebDAV protocol* has been available. This method offers a familiar interface for users who prefer a flexible, network drive-like process that simplifies the upload process also for larger data volumes. End of 2024, RADAR introduced *Git integration* allowing researchers to seamlessly import content from both public and private GitHub repositories into RADAR. Integration with GitLab will be released in the first half of 2025.

Further, RADAR will soon offer more flexible access to research data and expand its download options. Currently, datasets stored on tape are only available for download as packed .tar files. Around mid 2025, users will be able to *download individual files or folders* from these datasets via selective streaming. In future, this enhancement will also enable data previews and facilitate data transfer prior to versioning. At the same time, access to data by external reviewers will be simplified by additionally offering the dataset content as downloadable .zip file.

The RADAR team follows an *agile software development* workflow, continuously releasing improvements based on user feedback, discussions and prioritization with the RADAR User Advisory Council, as well as emerging needs and developments in the research data ecosystem, including the NFDI. *Future improvements* include the versioning of datasets, enhanced search and terminology driven functionalities, optional integration of artificial intelligence for automated metadata annotation, and integration with other research data management tools and authentication infrastructures like IAM4NFDI[13].

11 https://www.f-uji.net/; *Visited on May 16, 2025.*
12 https://base4nfdi.de/projects/ts4nfdi; *Visited on May 16, 2025.*
13 https://www.nfdi-aai.de/; *Visited on May 16, 2025.*

5 Community-Specific Service Offerings and Alternative Operating Variants

Over time, FIZ Karlsruhe has not only continuously enhanced RADAR functionality, but also expanded the service to new target groups and introduced alternative operating variants for use by academic institutions.

FIZ Karlsruhe is actively involved in several NFDI consortia. Analyses in NFDI4Chem, NFDI4Culture, and NFDI4Memory demonstrated the need for a generic repository based on RADAR to complement the portfolio of existing specialized repositories. So far, FIZ Karlsruhe has developed three community-specific data publication services, which are free of charge and directly address researchers independent from their institutional affiliation. The RADAR team will continuously enhance these services, for example by integrating discipline-specific terminologies and metadata standards developed by the respective consortia:

- **RADAR4Culture**[14]: for communities in the field of cultural heritage.
- **RADAR4Chem**[15]: for the chemistry community.
- **RADAR4Memory**[16]: for the historically working humanities.

Further, in response to the needs of academic institutions and their diverse IT environments, FIZ Karlsruhe launched alternative operating variants in 2021. Unlike the cloud service, these allow to integrate own IT resources within the RADAR system:

- **RADAR Local**: Enables institutions to host a local RADAR instance on their own IT and under their own domain operated by FIZ Karlsruhe, providing complete control over data management.
- **RADAR Hybrid**: Allows institutions to store research data in an alternative archive storage (either locally or externally), while continuing to use RADAR´s cloud-based software and infrastructure.

As of spring 2025, seven RADAR Local instances are either operated by FIZ Karlsruhe or currently in the implementation phase. The instance at the University of Potsdam is being expanded into a cross-university research data portal for the federal state of Brandenburg, as part of the project FDM-BB[17]. This centrally provided, state-wide data repository based on RADAR could serve as a model for other federal states.

14 https://radar4culture.radar-service.eu/; *Visited on May 16, 2025.*
15 https://radar4chem.radar-service.eu/; *Visited on May 16, 2025.*
16 https://radar4memory.radar-service.eu/; *Visited on May 16, 2025.*
17 https://fdm-bb.de/; *Visited on May 16, 2025.*

6 Conclusion

All future developments remain firmly aligned with our core principles: long-term data security and integrity, demand-driven and sustainable software development, and the FAIR principles. To do so, we are in close consultation with the institutions using RADAR and with the NFDI consortia in which we participate, and are committed to the ongoing integration of RADAR into regional, national and the European research data ecosystems.

Acknowledgements

We thank all institutions using RADAR, especially the User Advisory Council, our partners, and all colleagues within and beyond the NFDI community for their ideas and constructive feedback.

Authorship Contributions

The basic structure of the text was created with the help of ChatGPT.

Conflict of Interest

The authors have no conflicts of interest to declare.

Bibliography

FAIRsharing Team. 2023. *RADAR Metadata Schema.* FAIRsharing record. https://doi.org/10.25504/FAIRsharing.e26f92.

Soltau, Kerstin, Stefan Hofmann, Agnes Jasinska, and Clara Linés. 2024. *FAIRness assessment as a catalyst of chemistry research repository transformation in Germany.* Zenodo. https://doi.org/10.5281/ZENODO.13135009.

Was machen eigentlich Data Champions?

Anne Gärtner [1,*], Kristina Grüttemeier [2], Markus Rump [1], Johanna Vompras [1]

[1] Kompetenzzentrum Forschungsdaten, Universität Bielefeld, Deutschland;
[2] Servicezentrum Medical Data Science, Universität Bielefeld, Deutschland;
[*] Kontakt: anne.gaertner@uni-bielefeld.de

Data Champions können eine wichtige Rolle bei der Verbesserung der Sichtbarkeit und Effektivität des Forschungsdatenmanagements (FDM) an Hochschulen einnehmen. Das Konzept, das inital an den Universitäten TU Delft (Clare 2019) und der University of Cambridge (Savage und Cadwallader 2019) entwickelt wurde, hat sich dort als ein nachhaltiges Modell bewährt. Auch in Deutschland gewinnt die Adaption dieses Ansatzes zunehmend an Bedeutung und wird an verschiedenen Hochschulen, z. B. der TU Dortmund[1] oder der RWTH Aachen[2], in bestehende FDM-Services integriert.

Dieser Beitrag stellt die Implementierung eines Data-Champions-Netzwerks an der Universität Bielefeld vor, das seit fünf Jahren erfolgreich besteht. Ziel des Beitrags ist es, sowohl die Entstehung des Netzwerks vorzustellen als auch dessen Struktur, Funktionsweise und seine Vorteile für das hochschulweite FDM. Zudem wird die konkrete Ausgestaltung des Netzwerks präsentiert. Zuletzt thematisieren wir die Herausforderungen, die wir im Verlauf identifiziert haben, diskutieren deren Implikationen und zeigen Potenziale für eine zukünftige Weiterentwicklung auf.

Keywords: Data Champions, Forschungsdatenmanagement, FDM-Strategie, Netzwerkbildung

1 https://fdm.tu-dortmund.de/fdm-an-der-tu-dortmund/data-champions/; *Besucht am 30. Mai 2025.*
2 https://blog.rwth-aachen.de/forschungsdaten/2024/04/11/data-champions-rwth-carolin-victoria-schneider/; *Besucht am 30. Mai 2025.*

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23950 (CC BY-SA 4.0).

1 Einleitung

Data Champions bezeichnen engagierte Forschende, die sich für gutes FDM einsetzen und als Multiplikator*innen in ihren Fachbereichen wirken können. Anders als z. B. Data Stewards, die häufig institutionell verankert und operativ in FDM-Prozesse eingebunden sind, übernehmen Data Champions diese Rolle meist ehrenamtlich und ohne formale Zuständigkeit. Durch die enge inhaltliche Überlappung sind Data Stewards häufig auch Data Champions. Eine Tätigkeit als Data Steward ist jedoch keine Voraussetzung für ein Engagement als Data Champion.

Die konkrete Ausgestaltung der Rolle eines Data Champion variiert stark zwischen Hochschulen, wie im Austausch deutschsprachiger Angebote im Rahmen des 71. Jour Fixe FDM der Landesinitiative für Forschungsdatenmanagement fdm.nrw deutlich wurde[4]. An manchen Standorten werden einzelne Forschende für herausragendes Datenmanagement in ihren Projekten ausgezeichnet, an anderen Standorten bilden Data Champions ein aktives Netzwerk, das regelmäßig miteinander und mit FDM-Servicestellen im Austausch steht.

2 Genese und Konzept des Data-Champions-Netzwerks an der Universität Bielefeld

Die Entstehung des Data-Champions-Netzwerks an der Universität Bielefeld ist eng mit der Entwicklung des Forschungsdatenmanagements an der Hochschule verbunden. Ein entscheidender Meilenstein im Jahr 2018 war die Etablierung des Kompetenzzentrums Forschungsdaten[5], einer von der Universitätsbibliothek (UB) und dem Hochschulrechenzentrum (BITS) getragenen Serviceeinheit. Diese bietet ein generisches Dienstleistungsportfolio für zahlreiche Aspekte des Forschungsdatenmanagements, wie Datenmanagementpläne, Publikation und Archivierung von Forschungsdaten, Beratung zu rechtlichen und methodischen Fragen sowie Schulungsangebote für Forschende. Ein rein zentraler Ansatz stößt jedoch schnell an personelle Kapazitätsgrenzen, wenn fachspezifische Feinheiten einzelner Disziplinen Berücksichtigung finden müssen. Das Dienstleistungsportfolio bleibt in diesem Fall oberflächlich und verhindert eine nachhaltige Verankerung von FDM-Praktiken in den Fachbereichen und bei den Forschenden. Ein Data-Champions-Netzwerk, das zentral organisiert wird und Mitglieder aus unterschiedlichen Disziplinen rekrutiert, erhöht nicht nur Fachspezifität und Sichtbarkeit des FDM-Portfolios, sondern bietet auch eine einzigartige Möglichkeit des niedrigschwelligen Austauschs und der effizienten Weiterentwicklung des FDM-Bereichs.

4 https://www.fdm.nrw/archiv-jour-fixe-fdm; *Besucht am 30. Mai 2025.*

5 Kompetenzzentrum Forschungsdaten: https://data.uni-bielefeld.de; *Besucht am 30. Mai 2025.*

Angesichts dieser Ausgangslage suchten wir nach einer Lösung, um FDM-Themen gezielt in die Fachbereiche zu integrieren. Bei der Konzeptionierung und Gründung des Bielefelder Data-Champions-Netzwerks, lieferten unter anderem die Erfahrungen der TU Delft und der University of Cambridge wertvolle Orientierung. Die dort entwickelten Data-Champions-Modelle zeigten, dass dezentrale Netzwerke eine effektive Möglichkeit bieten, Forschende mit FDM-Affinität miteinander zu vernetzen. In Bielefeld entschieden wir uns bewusst dazu, auf das ehrenamtliche Engagement der Forschenden zu setzen: Unsere Data Champions sind in der Regel Personen, die bereits Berührungspunkte mit disziplinspezifischem Forschungsdatenmanagement haben – oft unabhängig von den zentralen Servicestellen und im Rahmen eigener Projekte oder Fakultätsinitiativen. Darüber hinaus steht die Rolle des Data Champions allen Forschenden offen, die ein Interesse am Forschungsdatenmanagement haben und sich in diesem Bereich engagieren möchten.

Durch die Vernetzung dieser Personen können zentrale FDM-Servicestellen nicht nur bestehende Aktivitäten und Infrastrukturen besser überblicken, sondern bekommen zusätzlich fachspezifischen Input durch die punktuell agierenden Expert*innen. Dieser Input kann dann als Anforderung an zukünftige FDM-Services gezielt in die strategische Entwicklung des FDM-Portfolios einfließen.

Die hier vorgestellten Überlegungen wurden 2019 in Form eines Konzeptpapiers für das Data-Champions-Netzwerk der Hochschulleitung und relevanten Gremien vorgestellt. Der offizielle Start des Netzwerks folgte 2020 mit einer (digitalen) Informationsveranstaltung. Für eine von Beginn an hohe hochschulweite Sichtbarkeit wurde die Veranstaltung und das Netzwerk über den hochschulweiten Emailverteiler des Prorektors für Forschung beworben und von Beginn an mit einem eigenen Webauftritt[6] begleitet. Ausgehend von der Informationsveranstaltung, lässt sich die Gründung des Netzwerks in drei Schritten zusammenfassen:

Schritt 1: Informationsveranstaltung für Forschende und Vorstellung der ersten Data Champions. Diese wurden während der Konzeptionsphase direkt vom Kompetenzzentrum identifiziert und angesprochen.

Schritt 2: Persönliche Gespräche mit Forschenden, die Interesse an der Mitwirkung als Data Champion hatten.

Schritt 3: Start des Data-Champions-Netzwerks: Kick-Off Meeting aller Data Champions und Vorstellung aller Data Champions auf der Website des Kompetenzzentrums Forschungsdaten.

6 Webauftritt des Data-Champions-Netzwerks der Uni Bielefeld: https://www.uni-bielefeld.de/ub/digital/forschungsdaten/data-champions/; *Besucht am 30. Mai 2025.*

Seit 2023 wird das Data-Champions-Netzwerk gemeinsam mit dem Servicezentrum Medical Data Science[7] betrieben.

3 Struktur und Funktionsweise des Netzwerks

Das Data-Champions-Netzwerk an der Universität Bielefeld besteht aktuell aus einer wachsenden Gruppe engagierter Forschender aus verschiedenen Fachbereichen. Derzeit nehmen etwa 10 bis 15 Data Champions regelmäßig an den Netzwerktreffen teil. Auf der Webseite des Netzwerks sind sie mit ihrer Expertise und ihrem Fachbereich sichtbar gelistet, sodass andere Forschende gezielt Kontakt aufnehmen können.

Data Champions treffen sich zwei- bis dreimal jährlich zu interdisziplinären Austauschrunden. Diese Treffen finden digital statt. Die FDM-Servicestellen bereiten einen initialen Input vor, beispielsweise zu neuen FDM-Tools, Infrastrukturentwicklungen oder relevanten Publikationen. Der anschließende inhaltliche Austausch geht von den Data Champions selbst aus. Sie bringen Themen aus ihren jeweiligen Bereichen und Forschungsprojekten in die Treffen ein. Anliegen, die eine vertiefte Diskussion erfordern, können im Vorfeld angemeldet oder nach einer ersten Besprechung im Netzwerk identifiziert werden. In diesem Fall können die zentralen Servicestellen die Vorbereitung für das nächste Treffen übernehmen.

Durch die Teilnahme erhalten Data Champions eine erhöhte Sichtbarkeit innerhalb der wissenschaftlichen Gemeinschaft. Sie profitieren vom interdisziplinären Austausch, erhalten Zugang zu aktuellen Informationen im Bereich Forschungsdatenmanagement und können sich über neue Entwicklungen informieren. Dies ermöglicht ihnen, ihre eigenen FDM-Strategien gezielt weiterzuentwickeln. Die erworbenen Kompetenzen sind nicht nur für ihre eigene Forschung von Nutzen, sondern auch ein Vorteil bei der Einwerbung von Drittmitteln, für die immer häufiger Datenmanagementanforderungen bestehehen.

Aber auch für die Hochschule leisten die Data Champions einen wertvollen Beitrag in mehreren zentralen Bereichen. Sie fungieren als Multiplikatoren für grundlegende Prinzipien des FDM innerhalb ihrer Disziplinen und tragen dazu bei, das Bewusstsein für FDM-Themen zu schärfen. Gleichzeitig bringen sie spezifische Bedarfe aus ihren Fachbereichen in das Netzwerk ein, wodurch sich nicht nur potenzielle Synergien mit Forschenden ergeben, die an ähnlichen Fragestellungen arbeiten, sondern auch gezielte Rückmeldungen an die zentralen FDM-Servicestellen ermöglicht werden. Darüber hinaus engagieren sie sich aktiv in Netzwerkveranstaltungen und Schulungsangeboten, etwa beim NRW-weiten

7 Servicezentrum Medical Data Science: https://www.uni-bielefeld.de/fakultaeten/medizin/forschung/infrastruktur/medicaldatascience/; *Besucht am 30. Mai 2025.*

Tag der Forschungsdaten[8] oder der Bielefelder Woche der Forschungskompetenzen[9]. Ein weiterer wichtiger Aspekt ihrer Arbeit ist die Entwicklung und Verbreitung fachspezifischer Ressourcen und Handreichungen, die als praxisnahe Unterstützung für Forschende dienen.

Bestehende Data-Champions-Kontakte können in diversen Bereichen als unterstützende Glieder gesehen werden. Wir wollen im folgenden einige Beispiele für die bestehende Zusammenarbeit vorstellen:

Data Champions werden über eine Mailingliste zu FDM-bezogenen Veranstaltungen, Ausschreibungen und Förderlinien informiert und können diese Informationen gezielt in der eigenen Fakultät streuen. Die zentralen FDM-Servicesstellen erarbeiten mit den Data Champions gemeinsam Best-Practice-Empfehlungen, die in den Fachdisziplinen detaillierter spezifiziert und den Forschenden sowie Studierenden zur Verfügung gestellt werden können. Ein Beispiel dafür ist eine aktuell in Bearbeitung befindliche Handreichung zur Wahl geeigneter Speicherorte für Forschungsdaten. Für diese Handreichung stellen die zentralen FDM-Servicestellen eine fachgenerische Vorlage zu Speicherorten entlang des Data Life Cycles bereit. Anschließend ergänzen interessierte Data Champions fachspezifische Informationen.

Das Ergebnis enthält dann praxisnahe Empfehlungen, die den Anforderungen unterschiedlicher Fachdisziplinen entsprechen. Gleichzeitig erhalten die zentralen FDM-Servicestellen wertvolle Einblicke in bestehende fachliche Best Practices sowie in lokale Infrastrukturen zur Datenverwaltung und Speicherung. Diese Einblicke ermöglichen es, in Beratungen gezielt auf bestehende Expert*innen oder verfügbare Tools in den Fakultäten zu verweisen, was wiederum den Informationsfluss und das Networking innerhalb der Hochschule stärkt.

4 Koordination des Netzwerks

Ein funktionierendes Data-Champions-Netzwerk benötigt nicht nur engagierte Forschende, sondern auch eine koordinierende Instanz. Die beiden FDM-Servicestellen der Universität Bielefeld spielen dabei eine zentrale Rolle:

- Sie organisieren die digitalen Netzwerktreffen und stellen Kommunikationsplattformen (Emailverteiler) bereit. Bei den Treffen geben sie aktuelle, für alle Fachdisziplinen relevante Informationen zu FDM-Entwicklungen (lokale oder auch landesweite) gebündelt an die Data Champions weiter.

8 NRW-weiter Tag der Forschungsdaten: https://fdm-nrw.coscine.de/#/4.TdF; *Besucht am 30. Mai 2025.*

9 Bielefelder Woche der Forschungskompetenzen: https://www.uni-bielefeld.de/ub/digital/forschungsdaten/veranstaltungen/datenwoche/; *Besucht am 30. Mai 2025.*

- Sie erfassen fachdisziplinspezifische Bedarfe durch den gegenseitigen Dialog, beraten bei individuellen Fragen und stehen auch im Nachgang der Treffen als Anlaufstelle für Data Champions zur Verfügung. Das aus dem Data-Champions-Netzwerk neu erworbene fachspezifische Wissen kann somit in die Beratungs- und Unterstützungsangebote miteinfließen.

5 Herausforderungen und Lösungsansätze

Die beschriebenen Aufgaben erfordern sowohl von zentralen Servicestellen als auch von den Data Champions zeitliche Ressourcen. Ohne eine langfristige und feste Stelle, die die koordinierenden Aufgaben des Netzwerks übernimmt und sicherstellt, dass die Data Champions effektiv entsprechend ihrer Kapazitäten und Kompetenzen eingebunden werden können, sehen wir folgende Herausforderungen und Lösungsansätze:

1. **Ressourcen:** Data Champions engagieren sich freiwillig und haben dafür nur begrenzte Kapazitäten. Eine mögliche Lösung ist die institutionelle Verankerung der Rolle, um langfristige Kontinuität zu gewährleisten. Insbesondere bei neu ausgeschriebenen Stellen für Data Stewards, die sich im Netzwerk engagieren sollen, könnte ein fester Stundenumfang für das Engagement als Data Champion vorgesehen werden.
2. **Fluktuation:** Da Data Champions oft Forschende sind, die in befristeten Projekten arbeiten, gibt es eine hohe Fluktuation. Hier helfen strukturierte Onboarding-Prozesse und kontinuierliche Rekrutierungsmaßnahmen. Gleichzeitig bleibt das Netzwerk damit ein niedrigschwelliges Angebot, mit dem die sich stetig ändernden Bedarfe und Anforderungen aufgegriffen werden können.
3. **Zusatzbelastung für Data Champions:** Solange das Engagement als Data Champion nicht als fester Bestandteil von Stellen eingeplant wird, ist jede Art der Beteiligung eine Zusatzbelastung für die teilnehmenden Forschenden. Um dem entgegen zu wirken, setzt das Data-Champions-Netzwerk auf flexible Beteiligungsmöglichkeiten und niedrigschwellige Kommunikationswege.

6 Fazit und Zukunftsperspektiven

Das Data-Champions-Netzwerk an der Universität Bielefeld hat sich bislang als ein wertvolles Instrument zur Vernetzung von FDM-Akteuren erwiesen und leistet einen guten Beitrag zur Sichtbarkeit und Weiterentwicklung des Forschungsdatenmanagements. Vor allem für Hochschulen, an denen noch kein Netzwerk aus in Fachdisziplinen oder drittmittelgeförderten Großprojekten beschäftigten Data Stewards aufgebaut wurde, eignet sich dieser Ansatz hevorragend. In einem Data-Champions-Netzwerk können darüber hinaus

auch kleinere Bereiche vertreten sein, d. h. es kann auch den „Long-Tail“ der datenbezogenen Aktivitäten der Hochschule sichtbar machen. Seine Gründung kann sehr niedrigschwellig erfolgen, erfordert keine größeren Investitionen oder Aufwände und bietet eine gute Basis, um die FDM-Aktivitäten an der Hochschule aufzudecken und persönlich mit den Akteur*innen in Kontakt zu treten. Das Data-Champions-Netzwerk schafft einen Rahmen für disziplinübergreifende Vernetzung und ermöglicht den direkten Austausch zwischen Gleichgesinnten und FDM-Interessierten. Auch mit lediglich ehrenamtlichem Engagement gelingt es den Data Champions, feine Nuancen und ein erhöhtes Bewusstsein für FDM-Themen in ihre Fachdisziplinen zu tragen und dort ein Umdenken anzuregen.

Aus den in Bielefeld gewonnenen „Lessons Learned“ erachten wir es als äußerst sinnvoll, das Data-Champions-Konzept weiterzuentwickeln – beispielsweise durch eine formale Anerkennung der Rolle oder eine stärkere Einbindung in hochschulweite strategische Entscheidungsprozesse im Kontext des FDM. Wir würden es begrüßen, wenn durch die Vernetzung der Akteur*innen gemeinsame Anliegen zur Infrastrukturentwicklung kooperativ vorangetrieben werden könnten. Bisher mangelt es jedoch an Ressourcen für die Etablierung von Pilotprojekten, gezielten Initiativen oder kleineren Arbeitsgruppen. Zudem erweist es sich als schwierig, Data Champions als festen Bestandteil der Governance-Struktur zu verankern, da ihre Mitwirkung unverbindlich und rein ehrenamtlich erfolgt.

Im Zuge strategischer Bestrebungen an der Universität Bielefeld wird aktuell u. a. die Schaffung neuer Ressourcen im Bereich des FDM in Aussicht gestellt. Damit geht perspektivisch auch die Möglichkeit einer institutionellen Verankerung der Data Champions einher.

Beiträge

- Anne Gärtner: Konzept, Projektorganisation, Verschriftlichung – Originalentwurf
- Kristina Grüttemeier: Konzept, Verschriftlichung – Überprüfung und Bearbeitung
- Markus Rump: Konzept
- Johanna Vompras: Konzept, Verschriftlichung – Originalentwurf

Interessenkonflikte

Die Autor*innen erklären, dass kein Interessenkonflikt besteht.

Literaturverzeichnis

Clare, Connie. 2019. *How to build a community of Data Champions: Six Steps to Success.* Zenodo. https://doi.org/10.5281/zenodo.3383814.

Savage, James L. und Lauren Cadwallader. 2019. „Establishing, Developing, and Sustaining a Community of Data Champions“. *Data Science Journal, 18(1),* https://doi.org/10.5334/dsj-2019-023.

Automated Metadata Extraction Compliant with Machine-actionable Software Management Plans

Dhwani Solanki [1], Suhasini Venkatesh [1], Dietrich Rebholz-Schuhmann [1], Leyla Jael Castro [1,*]

[1] ZB MED Information Centre for Life Sciences, Cologne, Germany
[*] Corresponding author: ljgarcia@zbmed.de

Research software and the FAIR for Research Software (FAIR4RS) principles are gaining more attention from research communities in different domains due to their role in the reproducibility of science. The Software Management Plans (SMPs) are a nice complement to the FAIR4RS principles and research software good practices. A machine-actionable layer providing semantically structured metadata describing the research software would make it easier for machines to process the data and would enable, for instance, the creation of Knowledge Graphs for research software metadata and related artifacts, e.g., data processed by the software. To this end, we have created the machine-actionable SMPs (maSMPs) metadata schema based on schema.org, and compatible with Bioschemas and Codemeta. To make it easier for researchers, we are also working on a tool to automatically extract such metadata from GitHub repositories. Here we introduce our approach towards maSMPs and present our preliminary work on automatic metadata extraction from GitHub API.

Keywords: Software Management Plans, machine-actionability, metadata

1 Introduction

Research software and its corresponding FAIR for Research Software (FAIR4RS) principles (Barker et al. 2022) are gaining more attention from research communities in different

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23951 (CC BY-SA 4.0).

domains as they contribute to the reproducibility of science. Two major obstacles linked to research software are: (i) lack of documentation with enough details to support its reuse by other researchers, and (ii) lack of structured metadata to support machine-actionability. Some of the commonly missing elements are key to FAIR4RS, for both humans and machines, e.g., authorship details, license, and unique identification. Some of these elements align with good practices that should be considered during the life cycle of research software.

To make it easier for researchers to adopt such good practices and improve the FAIR-ness of their software, the ELIXIR Software Best Practices group proposed a Software Management Plan (SMP, see Alves et al. 2021) questionnaire, i.e., a text-based document. SMPs are similar to Data Management Plans but for software rather than pure data, e.g., datasets. SMPs focus on research software good practices complementing, but not replacing, project management approaches. SMPs are meant to raise awareness of good practices while also providing means to collect information aligned to such practices. They are not a software development methodology or a full software project management approach. In fact, they can be used jointly with other approaches for software development but also on their own, which is an advantage for researchers who are not necessarily software developers, but still have to deal with code to process and data as needed for their research-oriented analyses.

While text-based SMPs serve humans well, a semantic layer was needed to enable machine-actionability, turning SMPs into machine-actionable SMPs (maSMPs). In addition to schemas supporting the machine-actionability, there is also a need to automate the metadata extraction process, making it easier for researchers to provide good basic metadata that is already present in, for instance, information about the source code repository. Here we introduce our approach towards maSMPs (Castro et al. 2024a) and present our preliminary work on a tool to extract metadata using the GitHub API (Venkatesh et al. 2025). In the future, we will expand the tool to also integrate metadata from GitLab source code repositories, and corresponding releases (for GitHub and GitLab) deposited in Zenodo.

2 Machine-actionable Software Management Plans

We have developed a maSMP metadata schema including types and properties (Castro et al. 2024a) mapping the information collected in SMPs together with usage profiles (Castro et al. 2024b), i.e., recommendations and guidelines of use. Our metadata schema builds on top of schema.org (Guha, Brickley, and Macbeth 2016), Codemeta (Jones et al. 2016), and Bioschemas (Gray, Goble, and Jimenez 2017). The first version released mid 2022 was created on top of the ELIXIR SMP (Alves et al. 2021). New versions aim to better align with other SMPs, and are commonly developed with the help of the community during hackathons. Currently, our maSMP has been aligned with the SMP

(Martinez-Ortiz et al. 2022) proposed by the eScience Centre in the Netherlands, and the SMP (Grossmann and Franke 2023) by the Max Planck Digital Library.

In Figure 1, we present the main elements of our maSMP metadata schema. As observed in the figure, the schema considers the connection between DMPs and SMPs. We do not consider elements related to the research project or budget as those have been already discussed in the maDMPs (Miksa, Walk, and Neish 2019) proposed by the Research Data Alliance DMP Common Standards Working Group. The main elements in our schema are: SoftwareManagementPlan, SoftwareSourceCode, and SoftwareApplication (corresponding to software releases).

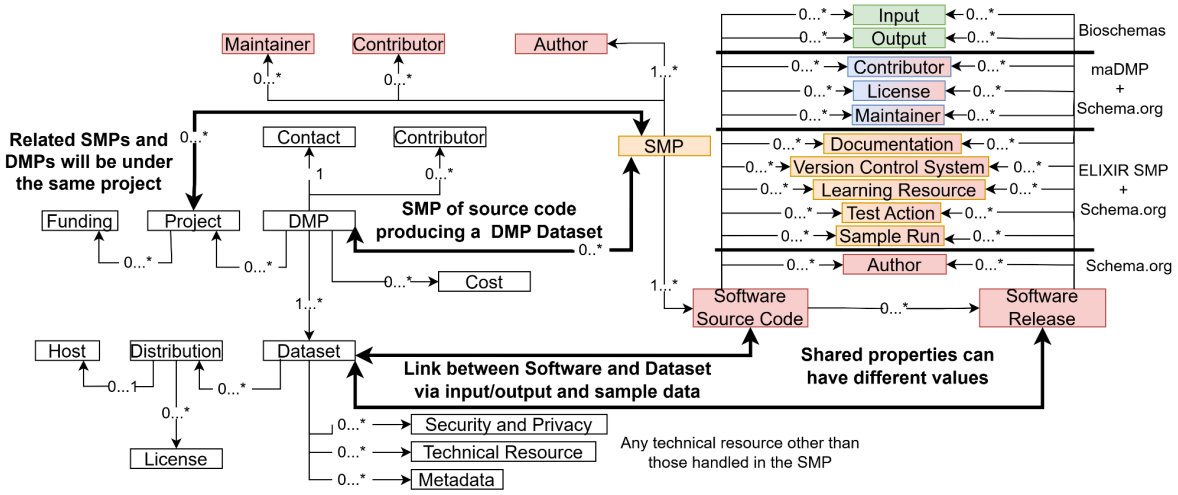


Figure 1: An extract of the maSMP metadata schema and its alignment to maDMPs.

3 Metadata extraction from GitHub

Although our maSMP metadata schema is already a step forward towards FAIR4RS and reproducibility, providing tools to automatically extract such metadata would (i) reduce the overhead for researchers and (ii) make it easier for such metadata to become part of research software repositories, e.g., hosted in GitHub, improving FAIRness of software directly at its source. There are already tools addressing the automatic extraction of metadata from GitHub repositories. For instance, the Software Metadata Extraction Framework (SOMEF, see Mao, Garijo, and Fakhraei 2019) creates a metadata file compatible with CodeMeta from information extracted via GitHub API together with machine-learning predictions, while SOMESY (Soylu et al. 2024) relies on its own file (somesy.toml) to create also a Codemeta file.

We are developing a REST API and interface (Venkatesh et al. 2025) to align metadata extraction to our maSMP metadata schema, see Figure 2. Different from SOMEF, we keep

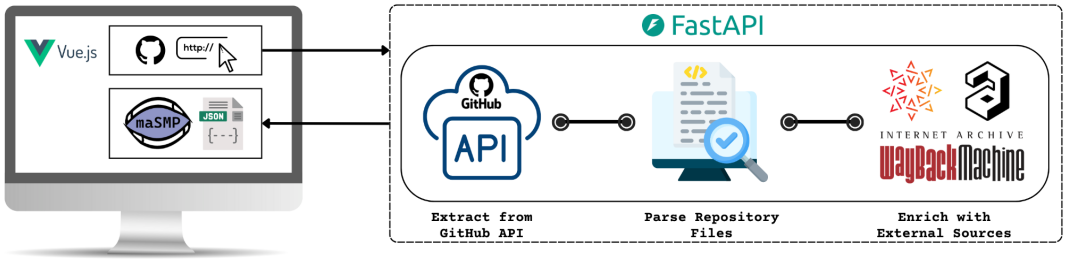


Figure 2: Our maSMP metadata extraction tool.

a clear separation between metadata obtained by direct means, e.g., GitHub API, and that one obtained as predictions using Machine Learning models. Different from SOMESY, we rely on GitHub API rather than on a toml file. At the moment, our extraction approach focuses on directly extracted metadata and distinguishes between the source code hosted in GitHub and the corresponding releases archived in external platforms. This work will help improve documentation, reproducibility, and the overall quality of research software. By automating metadata extraction and aligning it to maSMP and FAIR4RS, we aim to foster better practices in research software management across scientific domains.

4 Future work

One of the main challenges of the maSMP metadata schema is keeping it up-to-date with respect to changes in schema.org, Codemeta and SMP approaches. We will continue working with the community, e.g., via hackathons, to get feedback and improve the schema. As for the metadata extraction, the main challenge relates to the scarce data that can be directly retrieved from the GitHub API. We plan to use Large Language Models to fill in the blanks. The metadata extractor tool should also provide feedback and recommendations on good practices. Furthermore, we will work on automatic synchronization, GitLab coverage, and integration to the Research Data Management Organizer (RDMO, see Klar et al. 2025). Thanks to the integration in RDMO, together with automated metadata extraction tools, we aim to facilitate a broader adoption of maSMPs. The integration to RDMO will also enable capturing some metadata elements that otherwise would be difficult to achieve, e.g., links to tutorials about the software.

Acknowledgments

This work is part of the NFDI4DataScience project funded by the German Research Foundation (DFG), project number 460234259.

Authorship Contributions

DS, SV, DRS, and LJC contributed to the writing of the original draft, review, and editing of this manuscript. DRS contributed to the methodology. LJC contributed to the methodology and conceptualization.

Conflict of Interest

None declared.

Bibliography

Alves, Renato, Dimitrios Bampalikis, Leyla Jael Castro, José María Fernández, Jennifer Harrow, Mateusz Kuzak, Eva Martin, Fotis E. Psomopoulos, and Allegra Via. 2021. „ELIXIR Software Management Plan for Life Sciences“, https://doi.org/10.37044/osf.io/k8znb.

Barker, Michelle, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, et al. 2022. „Introducing the FAIR Principles for research software“. *Scientific Data* 9 (1). https://doi.org/10.1038/s41597-022-01710-x.

Castro, Leyla Jael, Olga Giraldo, Lukas Geist, Nelson Quiñones, Dhwani Solanki, and Dietrich Rebholz-Schuhmann. 2024a. *machine-actionable Software Management Plan Ontology (maSMP Ontology).* Zenodo. https://doi.org/10.5281/ZENODO.10582073.

———. 2024b. *Usage guidance (aka profiles) for the machine-actionable Software Management Plan Ontology.* Zenodo. https://doi.org/10.5281/ZENODO.10582121.

Gray, Alasdair J. G., Carole Goble, and Rafael C. Jimenez. 2017. *From Potato Salad to Protein Annotation.* ISWC Posters and Demo Session. Visited on May 28, 2025. http://ceur-ws.org/Vol-1963/paper579.pdf.

Grossmann, Yves Vincent, and Michael Franke. 2023. „Sustainable Research Software Through Software Management Plans“, https://doi.org/10.17617/2.3525108.

Guha, Ramanathan. V., Dan Brickley, and Steve Macbeth. 2016. „Schema.org: evolution of structured data on the web“. *Communications of the ACM* 59 (2): 44–51. https://doi.org/10.1145/2844544.

Jones, Matthew Bentley, Carl Boettiger, Abby Cabunoc Mayes, Arfon Smith, Peter Slaughter, Kyle Niemeyer, Yolanda Gil, et al. 2016. *CodeMeta: an exchange schema for software metadata. KNB Data Repository.* https://doi.org/10.5063/SCHEMA/CODEMETA-1.0.

Klar, Jochen, Olaf Michaelis, David Wallace, Max Schröder, Heinz-Alexander Fütterer, Claudia Malzer, Giacomo Lanza, David Martínez Muñoz, Dario Pilori, and Harry Enke. 2025. *Research Data Management Organiser (RDMO).* Zenodo. https://doi.org/10.5281/ZENODO.596581.

Mao, Allen, Daniel Garijo, and Shobeir Fakhraei. 2019. „SoMEF: A Framework for Capturing Scientific Software Metadata from its Documentation“. In *2019 IEEE International Conference on Big Data (Big Data),* 3032–3037. IEEE. https://doi.org/10.1109/bigdata47090.2019.9006447.

Martinez-Ortiz, Carlos, Paula Martinez Lavanchy, Laurents Sesink, Brett G. Olivier, James Meakin, Maaike de Jong, and Maria Cruz. 2022. *Practical guide to Software Management Plans.* Zenodo. https://doi.org/10.5281/ZENODO.7248877.

Miksa, Tomasz, Paul Walk, and Peter Neish. 2019. *RDA DMP Common Standard for Machine-actionable Data Management Plans.* https://doi.org/10.15497/RDA00039.

Soylu, Mustafa, Anton Pirogov, Volker Hofmann, and Stefan Sandfeld. 2024. *somesy.* Zenodo. https://doi.org/10.5281/ZENODO.13120456.

Venkatesh, Suhasini, Nelson Quiñones, Dhwani Solanki, Dietrich Rebholz-Schuhmann, and Leyla Jael Castro. 2025. *v0.1.0-beta pre-release maSMPs metadata extraction.* Zenodo. https://doi.org/10.5281/ZENODO.14918500.

elAPI: A Powerful, Extensible API Client for eLabFTW

Mahadi Xion [1], Alexander Haller [1], Philipp Kling [1], Martin Baumann [1], Nick Kepper [1]

[1] University Computing Centre, Heidelberg University, Heidelberg, Germany

The adoption of electronic lab notebooks (ELNs) has gained considerable traction, driven by the need for efficient data management, reproducibility, and automation, as well as the broader shift toward digitizing workflows in research environments. At Heidelberg University, the Research Data Unit[1] – a joint initiative of the University Library and the University Computing Centre – has introduced the ELN eLabFTW[2] as a general solution for the entire university[3]. To improve the management of our instance, we initially set out to develop a simple API script, which gradually evolved into elAPI – a full-fledged API client for eLabFTW. This paper outlines the motivation for developing elAPI, its architecture, use cases, and the impact it has had so far.

Keywords: Electronic Lab Notebook (ELN), eLabFTW, RDM-Tool, Automation, API

1 Introduction

Application Programming Interfaces (APIs) are essential for connecting software systems and automating data exchanges, improving efficiency across various domains. elAPI is a specialized API client developed to interact seamlessly with eLabFTW (CARPi, Minges, and Piel 2017), a widely used electronic lab notebook (ELN). ELNs play an important role in modern scientific research by allowing researchers to securely document experiments,

1 https://www.data.uni-heidelberg.de; *Visited on May 24, 2025.*
2 https://www.elabftw.net; *Visited on May 24, 2025.*
3 https://www.urz.uni-heidelberg.de/en/service-catalogue/software-and-applications/elabftw; *Visited on May 24, 2025.*

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23952 (CC BY-SA 4.0).

manage data, and ensure regulatory compliance. However, the complexity of laboratory workflows often necessitates advanced data management and automation capabilities. elAPI addresses these needs by simplifying interactions with eLabFTW, enabling users to automate key functions and efficiently manage experimental data.

2 Context and Motivation

In 2023, we initially set out to create a simple automated script to collect user statistics and generate invoices based on user data. At the time, only the official `elabapi-python`[4] library for eLabFTW was available, along with a few limited, single-purpose solutions from third parties or older projects that were based on the now-retired API v1[5].

The official Python library was, at the time, a relatively new, auto-generated client for the new API v2. It exhibited some inconsistencies with the actual API endpoints and, ultimately, we decided to use raw HTTP API calls via the popular HTTPX[6] library in our script. This decision, combined with a desire for abstraction, and better support for our specific automation needs, led to the development of a completely new API client for eLabFTW.

Once we realized that we had the opportunity to elevate our code into a complete API client with relatively little additional effort, we decided to take that chance and create an attractive solution for both researchers and administrators. Our goal was to develop a client that is easy to configure and use, yet flexible, with safe and sensible defaults.

2.1 Feature Overview and Comparison

The eLabFTW project first released their official API V2 client, `elabapi-python` in December, 2022[7], and elAPI was released almost a year later. elAPI, so far, is the only mature API client alternative to `elabapi-python`. While `elabapi-python` works well as a reference client, it has some caveats as automation scripts become increasingly advanced, creating a clear demand for a scalable, extensible, and robust framework. elAPI fills this gap while simplifying code development for both researchers and software developers. Although elAPI offers many additional features to streamline automation; here we will focus on the *structural* differences between elAPI and `elabapi-python`.

4 https://github.com/elabftw/elabapi-python; *Visited on May 24, 2025.*

5 See (a) https://gitlab.com/iam-cms/workflows/extra-nodes/elabapy-cli,(b) https://metacpan.org/pod/ELab::Client and (c) https://github.com/elabftw/elabftw/blob/6079d11f4c9c5695de8c7484f659dd4f26d25529/README.md#deprecated-projects-using-retired-api-v1; *Visited on May 24, 2025.*

6 https://www.python-httpx.org; *Visited on May 24, 2025.*

7 https://github.com/elabftw/elabapi-python/releases/tag/0.1.3; *Visited on May 28, 2025*

1. Simplicity: For small and simple automation tasks, writing a whole script can be cumbersome and counterproductive. elAPI is designed to be both a CLI tool and a library to enable high versatility. E.g., for a task like getting a list of experiments that have been assigned a certain status (e.g., "success") can be done with elAPI by running the following single line command: `elapi get experiments --query '{"status": "<desired status ID>"}'`. The same task with `elabapi-python` would require at least a dozen lines of code, which include some boilerplate code for server authentication.

2. Scalability: `elabapi-python` uses the Python built-in `threading` module to handle asynchronous connections[8]. Multi-threading is fraught with perils (Smith 2017). elAPI handles asynchronous connections with `asyncio` and `async/await` (co-routines) via HTTPX, which provides better documentation and control over the connections.

3. Configuration management: Any script written with `elabapi-python` requires a number of boilerplate code to manage authentication with the eLab server. elAPI offers a local and global configuration management system along with strong validation methods where boilerplate server authentication information can be punted off. This makes the script written with elAPI clutter-free and allows the developer to focus on the core logic.

4. Extensibility: elAPI was designed to enable unification of various eLab automation solutions, hence a robust third-party plugin system has been built into elAPI. Reusing an `elabapi-python` script that was developed by a different institute involves a completely different setup and environment to be able to run the script, depending on the script's requirements. elAPI's third-party plugin system promotes a shared ecosystem of eLab automation solutions where various institutes can easily share, distribute, and manage each other's solutions. Existing production-ready plugins support billing, data obfuscation, and user/team management.

3 Architecture and Design

We draw our inspiration from the microkernel architecture (Richards and Ford 2020; Thomas 2025) for the overall architecture principle and layered architecture for architecture partitioning. There was no strong rationale for choosing a microkernel architecture over alternatives such as a microservice architecture, other than the desire to start with something simple that still allows for extensibility. We intend elAPI to serve as the foundation for all eLabFTW automation solutions. In other words, eLabFTW users should

8 HTTP APIs can be found in `api_client.py`. Source: https://pypi.org/project/elabapi-python/5.2.0/#files; *Visited on July 4, 2025.*

be able to package their automation scripts as plugins, thereby extending elAPI's functionality. We refer to this architecture as the Simple Microkernel Architecture (SMA). In the very beginning, SMA was designed for elAPI only. Later we have improved and generalized it enough to be useful for any software design.

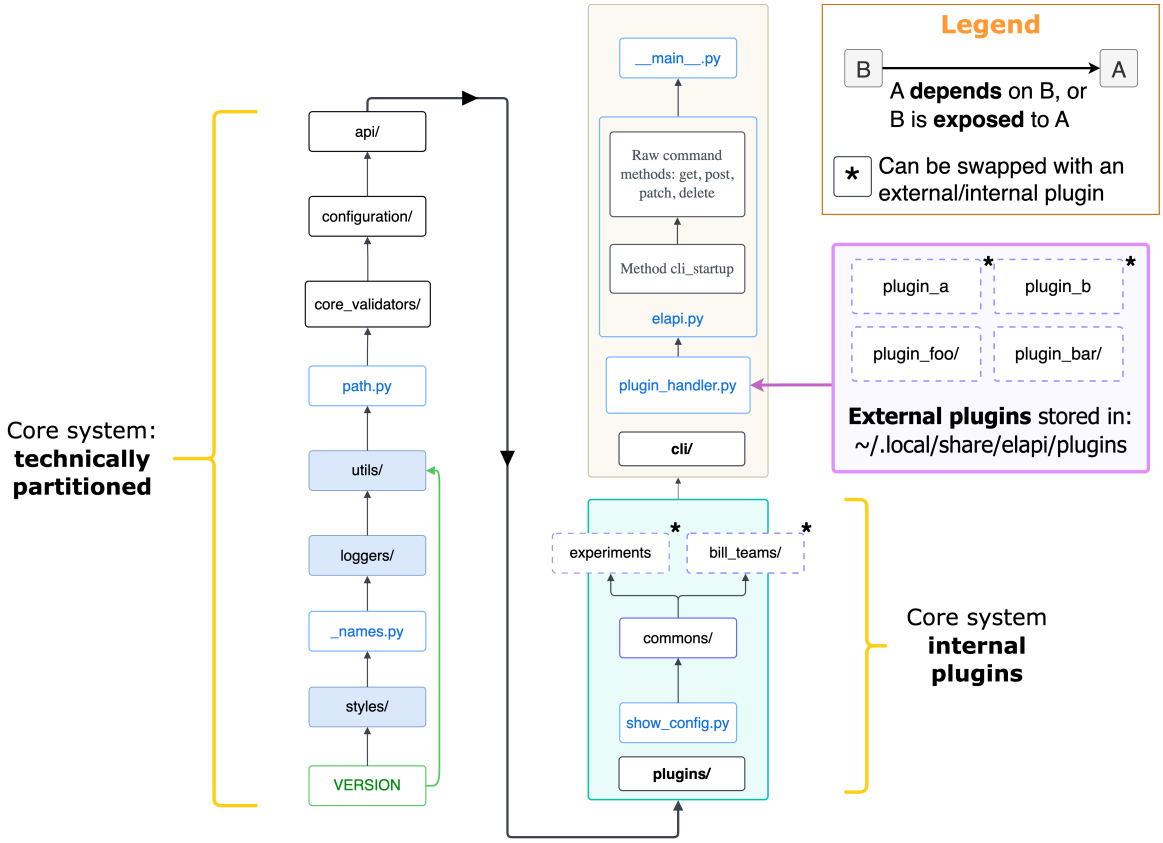


Figure 1: elAPI SMA architecture overview.

3.1 SMA Principles

SMA follows most of the principles found in microkernel architecture guidelines with a few additions.

1. A core system must provide basic functionalities: The core system principle is adopted straight from the textbook microkernel architecture principles. In SMA, the core system must provide the basic features. Here, features that would be considered as basic would depend on the software domain. At a minimum, the core system must enable the plugin system. The core system in elAPI offers interface style guides, logging functionality, helper functions, OS path handlers, basic validation abstractions, configuration manager, extended HTTPX API interfaces to communicate with eLabFTW endpoints, and plugin

handlers that enable the entire plugin system. elAPI's core system can be extracted out and used as the foundation for the SMA implementation for any standalone program.

2. Interfaces can be flexible across internal plugins: Typically, a microkernel architecture separates plugins from the core system entirely. SMA modifies this approach by introducing internal plugins, which reside inside the core codebase, and external plugins, which can exist anywhere. Both types of plugins can share the same plugin *interface*. This is also the default with elAPI's SMA implementation. Depending on flexibility preference, however, having a unique interface or distinct interfaces for internal plugins is also allowed where a unique interface is enforced upon external plugins.

3. Internal plugins can be part of the core system: In SMA, internal plugins can be part of the core system such that they are meant to be used by external plugins in the same way core system functionalities are used by all plugins in a typical microkernel architecture. In elAPI, public APIs offered by the internal plugin *"experiments"* are meant to be leveraged by external plugins to make updates or advanced modifications to eLabFTW experiments, among other features. Note, if external plugins are disabled for a SMA implementation, then internal plugins should not be part of the core system.

4. Hybrid partitioning is allowed: SMA recommends technical partitioning for the basic functionality layers. If domain partitioning is desired, then it is recommended to apply domain partitioning to internal and external plugins. The combined partitioning helps enforce a design based on strict separation of concerns.

5. A strict unidirectional dependency must be maintained between technically partitioned layers: SMA imposes a dependency invariance between technically partitioned layers. In Figure 1, we see a top layer always "depends on" a bottom layer, or in other words, a bottom layer's APIs are always exposed to a top layer. This is denoted with an upward arrow. A layer that sits at the most bottom is an independent layer, as it doesn't depend on any other layer. The third-party libraries (typically whose code do not reside in the codebase) do no take part in the invariance, and are considered floating layers. In that sense, an independent layer is also a floating layer. In elAPI, the styles layer is an independent layer. styles layer designate user interface definitions and methods, e.g., the color scheme used by elAPI CLI. Styles being the independent layer indicates that all dependent layers must adhere to the same UI guidelines.

6. Plugins need not be independent: This principle is in line with the third principle. Microkernel architecture imposes that plugins must be independent, and for good

reasons, i.e., to avoid "Big Ball of Mud" (Foote and Yoder 1999). SMA still lets the developer decide the rules internal and external plugins should follow. For example, external plugins can depend on internal plugins (as they are part of the core system). An external plugin vendor is free to decide that their plugin would depend on another external plugin.

4 Use Cases at Heidelberg University

1. Internal billing: The need for automation initially arose from the requirement of a billing system for our eLabFTW instance. This necessitated the collection of monthly user statistics, cost calculations, and the generation of invoices. These objectives have now been successfully achieved through the implementation of the internal elAPI plugin 'bill-teams'[9] (see Figure 2).

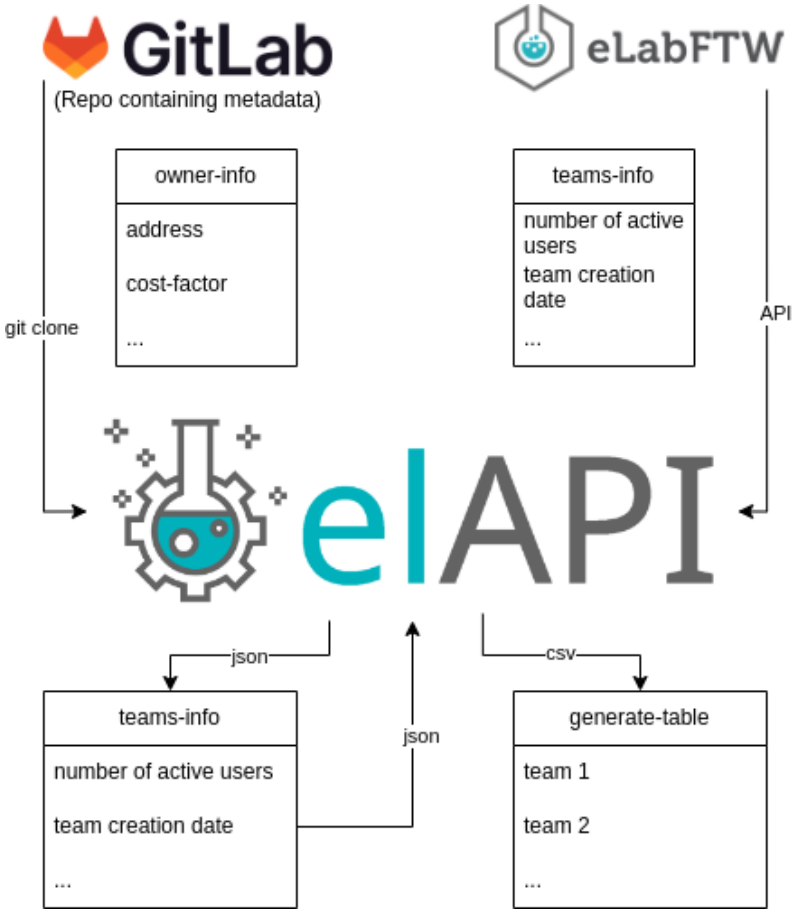


Figure 2: elAPI bill-teams workflow.

9 https://github.com/uhd-urz/elAPI/tree/ad1caad24c1cceaf9970c96093303a589118d351/src/elapi/plugins/bill_teams; *Visited on May 24, 2025.*

2. Stress testing: We utilized elAPI to generate thousands of users and hundreds of teams. Subsequently, we retrieved all relevant metadata and measured the responses to ensure that our instance can scale effectively for the foreseeable future.

3. Mass expiration of users in a team: We utilized a section of the elAPI code to create a custom script that efficiently expires all users for a given team.[10]

4. Rapid creation of plugins: An institute made an urgent request involving a fallback plan for the mass renaming of all experiments and database items across multiple teams. Using elAPI, we were able to quickly develop a custom plugin (called "Obfuscate")[11] in less than two days. This plugin saved the original title in extra fields, renamed the title, and allowed for later reversion if needed. Without a robust framework like elAPI, this task would have required significantly more time.

5. Use in teaching: In a course on research data management, students were tasked with planning and creating templates and experiments in eLabFTW. We used elAPI to demonstrate how we could later retrieve all the tags used by the team and identify frequently used terms.

5 Impact in the Community

elAPI is featured in the official eLabFTW GitHub repository[12] under the "Related Projects" section and was presented at the eLabFTW Community Meeting III[13]. We have received positive feedback from users; however, since we do not collect usage statistics, we are uncertain about the broader adoption of elAPI outside of Heidelberg University. Some usage statistics can be inferred from PyPI download counts (e.g., via pepy.tech[14]). However, these figures are often inflated by automated systems such as CI/CD pipelines and mirrors, and therefore do not accurately reflect the actual number of human users. Conversely, download statistics fail to account for other types of usage, such as users cloning the repository directly from GitHub or those who still continue to use older versions of the package.

10 https://github.com/uhd-urz/elAPI/blob/ad1caad24c1cceaf9970c96093303a589118d351/examples/python/manage_teams/expire_teams.py; *Visited on May 24, 2025.*
11 https://github.com/uhd-urz/elAPI/tree/ad1caad24c1cceaf9970c96093303a589118d351/examples/python/create_plugins/obfuscate; *Visited on May 24, 2025.*
12 https://github.com/elabftw/elabftw; *Visited on May 24, 2025.*
13 https://github.com/elabftw/elabftw/discussions/5152; *Visited on May 24, 2025.*
14 https://pepy.tech/projects/elAPI; *Visited on May 24, 2025.*

6 Limitations and Improvements

Some limitations arise from the general limitations of the eLabFTW API. eLabFTW currently does not provide unrestricted API-Level access to all actions and data. This makes it impractical to provide solutions like a global sync to repositories or archival systems that could be triggered by a sysadmin. While elAPI significantly enhances the user-friendliness of the API, it remains a CLI application and a Python library. This can pose a substantial barrier to entry for many users, especially those unfamiliar with command-line tools or programming. To improve accessibility, introducing a GUI/Web interface or enabling custom code hooks within eLabFTW would be valuable enhancements.

7 Conclusion and Outlook

elAPI has successfully addressed all our immediate needs far better than we initially anticipated. We now plan to gradually introduce new features as needed, primarily through the development of additional plugins. The first commands for the experiment-related plugins serve as an example of this approach.

For internal use cases, we have identified several potential areas for further development, including team management, automated metadata checks, and import/export functionality with heiARCHIVE[15]/heiDATA[16]. These avenues will be explored in future iterations of elAPI. We are advocating for the addition of new features in eLabFTW that would address or alleviate the limitations mentioned previously. To foster reuse and collaboration, we offer to feature community-provided plugins in our GitHub repository[17], serving both as practical tools and as examples for other developers. Additionally, we welcome contributions of general-purpose plugins, which could then be included in future releases of elAPI. We are exploring ways to link or feature elAPI and its plugins, along with examples and other related materials such as templates, in a community repository for ELN/eLabFTW resources, such as the ELN material collection[18]. A long-term goal may be to generalize the architecture of elAPI into a broader framework, with elAPI itself becoming a consumer of that framework.

15 https://heiarchive.uni-heidelberg.de; *Visited on May 24, 2025.*
16 https://heidata.uni-heidelberg.de; *Visited on May 24, 2025.*
17 https://github.com/uhd-urz/elAPI; *Visited on May 24, 2025.*
18 https://eln-resources.gitlab.io/collection; *Visited on May 24, 2025.*

Software Availability Statement

elAPI is freely available under the AGPLv3 license at the following repository: https://github.com/uhd-urz/elAPI[19]. Citation recommendation: Xion, Mahadi, and Alexander Haller. 2025. *elAPI*. Version V2. https://doi.org/10.11588/DATA/E4XHXZ.

Acknowledgments

The authors wish to thank the organizers of the E-Science-Tage conference in Heidelberg for their dedicated efforts, and our reviewers for their valuable and constructive feedback.

Authorship Contributions

- Mahadi Xion: Conceptualization, Original Draft, Review, Editing
- Alexander Haller: Conceptualization, Original Draft, Review, Editing
- Philipp Kling: Review
- Nick Kepper: Review
- Martin Baumann: Review

Conflict of Interest

Philipp Kling is leading the organization of the conference and, thus, is involved with the reviewing of contributions. Evaluation of this contribution was done anonymously by our reviewers and the final decision for the acceptance of this paper was in line with Cora F. Krömer and Sophie G. Habinger and the programme committee of the conference.

Bibliography

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. https://doi.org/10.21105/joss.00146.

19 *Visited on July 4, 2025.*

Foote, Brian, and Joseph Yoder. 1999. “Big Ball of Mud”. In *Fourth Conference on Patterns Languages of Programs.* Visited on June 27, 2025. http://www.laputan.org/mud/.

Richards, Mark, and Neal Ford. 2020. *Fundamentals of Software Architecture: An Engineering Approach.* 432. Sebastopol, CA: O’Reilly Media. ISBN: 9781492043454.

Smith, Luke. 2017. *Python: what are the advantages of async over threads?* Visited on May 24, 2025. https://stackoverflow.com/a/48021834/7696241.

Thomas, Richard. 2025. *Microkernel Architecture.* Visited on March 31, 2025. https://csse6400.uqcloud.net/handouts/microkernel.pdf.

DataPLANT Services Design – Considerations Towards a Common NFDI Landscape

Dirk von Suchodoletz [1], Marcel Tschöpe [1], Jonathan Bauer [1], Kevin Schneider [2], Timo Mühlhaus [2]

[1] Computer Center, University of Freiburg, Germany;
[2] Computational Systems Biology, RPTU Kaiserslautern-Landau, Germany

Keywords: DataPLANT, cloud oriented service infrastructure, PLANT DataHUB, NFDI, data publication

1 Motivation

Many Research Data Management (RDM) services are not limited to a single scientific domain or facility. Since no individual research institution can fully support all aspects of RDM with the required depth and domain-specific requirements for every discipline, collaboration and shared services are both logical and necessary. This approach aligns with the broader objective of One NFDI[1] within the National Research Data Infrastructure (NFDI). The NFDI aims to create a cross-disciplinary RDM ecosystem that meets the specific needs of research groups and disciplines, while also enabling the (re)use of data across traditional boundaries as well as the deployment of AI approaches. In line with this vision, one of the core objectives and contributions of the DataPLANT consortium (Martins-Rodrigues et al. 2021) is to provide tools and services that can be shared with other consortia and deployed across flexible backend infrastructures. At the heart of these technical services is the PLANT DataHUB.

1 Refers to the still informal but increasingly widespread concept and ultimate goal of a cross-disciplinary and cross-institutional approach to joint undertakings in research data management.

Published in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, and Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23953 (CC BY-SA 4.0).

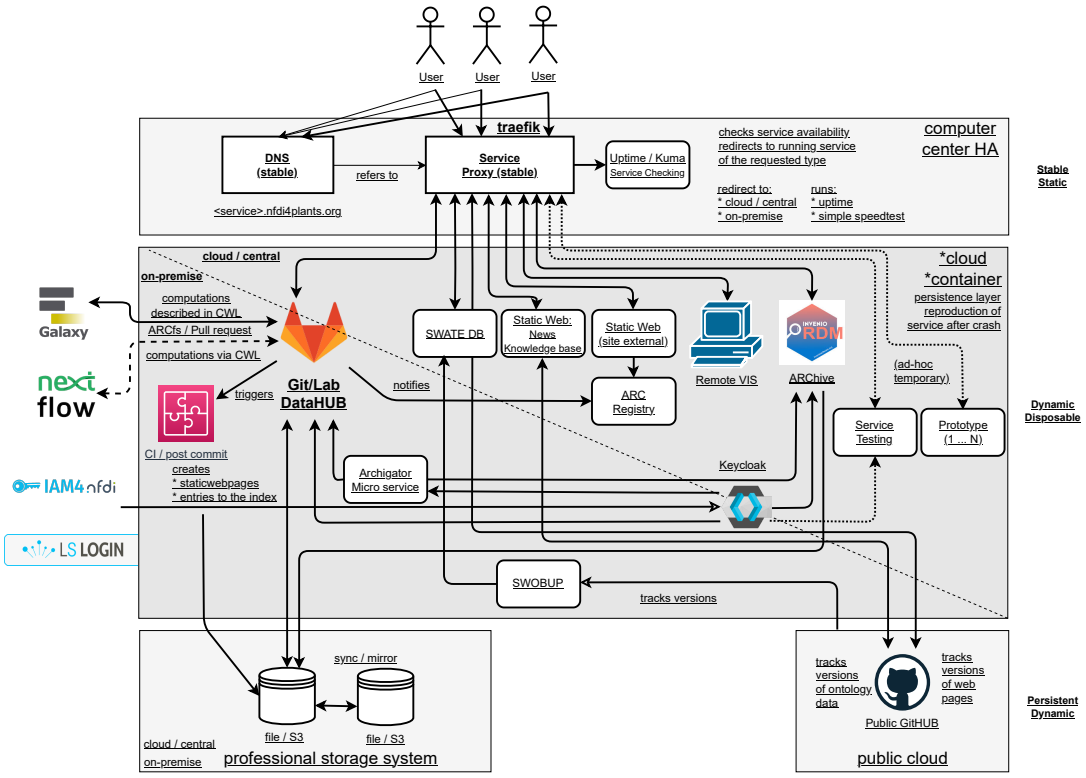


Figure 1: DataPLANT service infrastructure (Bauer, Tschöpe, and Suchodoletz 2025).

In addition to core functions, the DataHUB offers Continuous Integration (CI) pipeline templates, automated quality assurance, and other interaction services to foster DataPLANT's open participation and contribution model (see Figure 1). This model encompasses all relevant assets, including Annotated Research Context (ARC)[2], raw data, metadata templates, ontologies, code, and workflow descriptions. While originally geared towards plant scientists, both the ARC concept and the use of software development-based approaches to RDM, along with the associated services, are now being adopted by other consortia (e.g. FAIRagro, BIOIMAGE) and researchers across a wide range of disciplines. The backend services are designed as flexible, cloud-based microservices, supporting both on-premises installations and future integration with a shared NFDI infrastructure.[3] The PLANT DataHUB thus provides a range of RDM workflows to support data scientists throughout different stages of the research data lifecycle, from development to the publication of results. The bottom-up, community-driven approach requires collaboration among multiple stakeholders. In building these services, we follow design principles that offer high-level guidance and promote the creation of sustainable and maintainable applications. At DataPLANT, tool development is driven by community needs and facilitated through direct communication between researchers and developers. This process follows

2 See https://arc-rdm.org/; *Visited on May 22, 2025,* and Weil et al. (2023) for further background on the ARC concept, implementation and usage.

3 Multicloud proposal to the NFDI Section "Common Infrastructure", https://doi.org/10.5281/zenodo.6510971; *Visited on May 22, 2025.*

an incremental and iterative approach, supported by community contributions and issue tracking via the project repository.[4]

2 Service landscape & Design considerations

Over the past four years, the DataPLANT team has developed a suite of tools and microservices to extend software frameworks like GitLab for the PLANT DataHUB (Suchodoletz et al. 2023) or InvenioRDM for the "ARChive"[5] to enhance the digital service ecosystem of plant scientists (Suchodoletz, Bauer, and Tschöpe 2023). These core components, which focus on data management, versioning, sharing and publishing, are designed as portable modules that can be integrated into a broader base infrastructure. Key design priorities include beside user-friendliness, security, re-deployability, and fast recovery in case of system failures or infrastructure disruptions.

All DataPLANT core components are executed in Docker, while the configuration of these services is stored on the project's public code repository to separate development from data management and for security reasons. We aim to an "Infrastructure-as-code" principle for reproducibility and transparency. As soon as changes are made to the configuration in the repository, they are propagated to the core components. This process builds upon CI jobs, which are provided on so-called runners. These recipes for building software are used to distribute the configuration. This further supports the idea of a code repository as a core component of the science gateway – both for managing infrastructure components and for the actual DataHUB holding the scientific data as ARCs.

The process of assessing quality parameters of an ARC is further referred to as validation of the ARC against a "validation package" (VP), which is an arbitrary set of validation cases defined by e.g. a PI or research project lead that the ARC must pass to qualify as valid in regard to the VP. While a basic set of VPs is implemented, they can be significantly and individually extended for each project to accommodate different use cases and can be automatically triggered through the built-in CI infrastructure. The entire CI pipeline is organized in such a way that allows contributors to develop a VP of their own, and thus extend the ARC integration with yet other systems.

To support scientists in publishing their work, the in-house development Archigator acts as a connector and authentication microservice between the DataHUB and ARChive as well as other publication platforms for data and results. The integration of other repository

4 Everyone can contribute to developments at the project's GitHUB space, https://github.com/nfdi4plants *(Visited on May 22, 2025).* Feature requests and improvements are discussed openly on that project space or during bi-weekly data steward meetings. When necessary, final decisions are made in cross–task area speaker meetings, with support from the Scientific Advisory Board.

5 DataPLANT themed version of InvenioRDM, https://archive.nfdi4plants.org *(visited on May 22, 2025),* for data publication.

backends and workflow orchestration systems, such as Galaxy and nf-core, can be easily generalized by building on the foundational design of the Archigator tool. Archigator then serves as a data provider for these backends, utilizing their individual APIs. A virtual filesystem based on PyFilesystem called ARCfs has been created to allow for the integration of filesystem based storage in e.g. Galaxy, so that ARCs can be seamlessly integrated into such platforms (Bauer, Tschöpe, Suchodoletz, et al. 2024). This will allow scientists to connect their data in a fully integrated way throughout the data lifecycle based on GitLab. Here, we envision and strongly promote GitLab as a DataHUB to serve as a future representation of a central platform in the NFDI context (Suchodoletz et al. 2024).

In the case of Archigator and the publication of the scientists' results in i.e. ARChive, these CI processes help to assess the presence of mandatory publication-related metadata, quantify and publish the results. In addition to the core components, the in-house developments Swobup and SWATE (Mühlhaus et al. 2022), implemented as part of DataPLANT, support plant scientists in the provision and management of ontologies. These have also been developed with the principles of reusability and rapid recovery in mind. For reproducibility and provenance tracking all ontology data are published on the GitHub project space to engage the plant community. This data is retrieved by Swobup from the public repositories and integrated into the DataPLANT infrastructure as soon as changes are made to the public repositories and also when the Swobup and SWATE services are restored.[6]

All these tools and (micro-)services were programmed with on-premise operation in mind. Several communities within DataPLANT expressed the need to deploy their individual instances of the services (in particular the DataHUB). Partner institutions[7] have started to deploy their own DataHUB instances. Most of the core components have been developed as Docker containers or integrated directly into custom Docker images. We provide Docker Compose templates to configure the images for the various infrastructures available on site. The requirements for such on-premises deployments are: a virtualisation or container platform, a fast scratch space, and high-capacity storage such as NFS shares or, ideally, object storage. Although the basic requirements do not vary between institutions, the setups depend on the locally available infrastructures. The DataPLANT infrastructure team provides technical guidance to help local administrators create the required setup in terms of performance and data safety. In general, we encourage the use of a dynamic-yet-disposable model using cloud-based infrastructures (Bauer, Tschöpe, Schnürle, et al. 2024).

6 All tools, services, container descriptions are available from the project's public GitHUB referenced above.

7 At the beginning of 2025: Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Heinrich Heine Universität Düsseldorf, Eberhard-Karls Universität Tübingen, Universität zu Köln (planned).

3 Service Integration

The "glue that holds it all together" is a reverse proxy: a single entry point gateway running on a redundant virtual machine infrastructure (Bauer, Tschöpe, and Suchodoletz 2025). The proxy automatically manages Transport Layer Security certificates for new services, greatly increasing the flexibility of the infrastructure by allowing rapid changes to back-end services without having to change the static host names of user-facing services. Using the Traefik,[8] DNS names can be automatically created or changed on the fly, enabling a very dynamic structure. In addition, several monitoring services built into the proxy detect service outages, while logging services track performance data, access patterns and accounting information. The data are transferred from the proxy so that no additional disk space is used on the proxy to keep the setup simple to separate the individual components and also to take safety aspects into account. We actively participate in various working groups of the NFDI section "Common Infrastructure", primarily focusing on Multicloud, Overall Architecture, nfdi.software, Data Management Planning, Terminology Services, Identity and Access Management, and Research Software Engineering. We have also used past NFDI conferences to promote the service integration agenda. Through these efforts, we aim to encourage broader adoption by other communities and integration into the future NFDI service landscape.

The user management and accounting is realised in the DataPLANT infrastructure with Keycloak: a powerful open-source tool that acts as an AAI proxy between web-based services and either internal Keycloak accounts or external identity providers (IdP). Many external IdPs are relevant in the NFDI context, like Life Sciences AAI, ORCID and in particular the upcoming NFDI-AAI. While still in development, the NFDI-AAI aims to provide a NFDI-wide infrastructure to foster inter-consortia activities. The first steps to integrate DataPLANT Keycloak as a community AAI in the so-called NFDI-Infrastructure-Proxy were done as part of an incubator project.[9]

The further development of a conceptual Authentication, Authorization and Accounting (AAA) framework with mutual assurance (as part of IAM4nfdi) is a crucial step toward a sustainable services landscape. An effective accounting system requires coordination between the involved AAI, which provides appropriate attributes for attribution of resource reservation, and the offered services. This involves defining standardized service and infrastructure offerings that can be linked to their respective resources consumed on the provider side. For the DataHUB service relevant parameters might include storage volume and redundancy level, type of storage (e.g., object, block, or file storage), and expected performance parameters. For computing resources e.g. for the intended workflows, further parameters and limits are needed, such as CPU hours for HPC or cloud flavors with specific configurations and defined runtimes.

8 Widely used implementation of a reverse proxy, https://traefik.io; *Visited on May 22, 2025.*

9 Overview on projects within IAM4nfdi completed https://incubators.nfdi-aai.de/; *Visited on May 22, 2025.*

To enable future cost compensation, information about the infrastructure users and their institutions must be properly recorded and maintained. Additionally, it is useful to provide ongoing accounting information for monitoring purposes to both users and their funding entities. It would also be beneficial to offer forecasts on expected usage and related costs (Leendertse and Suchodoletz 2020). AAA concepts and structures need to be further developed to facilitate inter-institutional exchange of RDM services.

4 Outlook

The infrastructure stack has been in production for over four years across multiple locations. The volume of stored data is continuously increasing, CI services are being expanded, and the integration of computational workflows is being enhanced. Over the coming years, the focus will be on maintaining and improving the existing service landscape, securing broad community support, and ensuring sustainable long-term operation and development beyond the initial funding.

We will further evaluate and implement basic services provided through Base4NFDI when they become available to complement our service stack. Our infrastructure provisioning approach is modular, allowing other communities to reuse and recombine created services and tools. Individual modules can be replaced by future joint services of groups of consortia or the whole NFDI, such as for data publication. A layered model is used to separate tools and services from the underlying storage and compute infrastructure (Bauer, Tschöpe, and Suchodoletz 2025). This fits into the proposed service stack for the NFDI EOSC EU node.[10]

The service stack, either partially or completely, is made available to users beyond DataPLANT's core participants. All components are open-source and accessible from public repositories to promote collaboration and innovation. In the envisioned second funding phase of DataPLANT, a provider circle is planned to be set up to ensure continuous development and sustainability of services. Efforts will be made to integrate more closely with services from other consortia, state initiatives and international developments on research data management, enabling overarching RDM services.

Acknowledgements

We acknowledge the support for DataPLANT 442077441 through the German National Research Data Initiative (NFDI 7/1).

10 See, https://www.nfdi.de/nfdi-is-part-of-eoscs-build-up-phase; *Visited on May 22, 2025.*

Authorship Contributions

Dirk von Suchodoletz is the main author of this paper and coordinator of this project. Marcel Tschöpe and Jonathan Bauer significantly contributed to the initial design of the Archigator microservice, which interfaces with the respective APIs of the adapted GitLab and InvenioRDM packages, as well as to the implementation of the CI pipeline. They were also involved in the AAI (Authentication and Authorization Infrastructure) for DataPLANT, including the integration of Keycloak and IAM4NFDI. Kevin Schneider and Timo Mühlhaus are among the primary authors of the ARC specification, contributing to the definition and refinement of user requirements for the DataHUB, the data publication workflow, and the validation process. All authors jointly designed and developed the DataPLANT service landscape, with additional contributions from other team members over the past four years.

Conflict of Interest

There are no conflicts of interest.

Bibliography

Bauer, Jonathan, Marcel Tschöpe, Paul Chr. Schnürle, Julian Weidhase, Christoph Garth, and Timo Mühlhaus. 2024. *Cloud based flexible service infrastructure stack for the NFDI.* Structure graphics for this short paper. Visited on July 7, 2025. https://events.gwdg.de/event/658/contributions/2388/.

Bauer, Jonathan, Marcel Tschöpe, and Dirk von Suchodoletz. 2025. *DataPLANT services design – Considerations towards a common NFDI landscape.* Figure of the structure of DataPLANT service integration. https://doi.org/10.11588/heidok.00036418.

Bauer, Jonathan, Marcel Tschöpe, Dirk von Suchodoletz, Cristina Martins-Rodrigues, Julian Weidhase, Timo Mühlhaus, Christoph Garth, Gajendra Doniparthi, Holger Gauza, and Louisa Perelo. 2024. "From DataPLANT's DataHUB to DataPUB(lication)". In *Proceedings of the 15th International Workshop on Science Gateways (IWSG2023),* edited by Jens Krüger and Sandra Gesing. Publication by the 15th International Workshop on Science Gateways (IWSG2023), 13-15 June in Tübingen. https://doi.org/10.15496/publikation-100323.

Leendertse, Jan, and Dirk von Suchodoletz. 2020. “Kosten und Aufwände von Forschungsdatenmanagement”. *Bausteine Forschungsdatenmanagement,* number 1 (1): 1–7. 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, and Björn Usadel. 2021. “DataPLANT – Ein NFDI-Konsortium der Pflanzen-Grundlagenforschung”. *Bausteine Forschungsdatenmanagement,* number 2 (2): 46–56. https://doi.org/10.17192/bfdm.2021.2.8335.

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. https://doi.org/10.11588/heibooks.979.c13724.

Suchodoletz, Dirk von, Jonathan Bauer, and Marcel Tschöpe. 2023. “DataPLANT Cloud Oriented Service Infrastructure”. In *Proceedings of the Conference on Research Data Infrastructure,* volume 1. TIB Open Publishing. https://doi.org/10.52825/cordi.v1i.414.

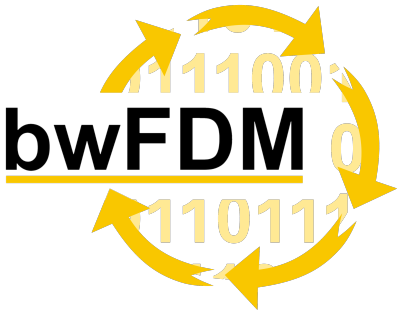
Suchodoletz, Dirk von, Dominik Brilhaus, Marcel Tschöpe, and Jonathan Bauer. 2023. *GitLab as a tool for Research Data Management.* Zenodo, 1. Conference on Research Data Infrastructure, Karlsruhe. https://doi.org/10.5281/zenodo.10021180.

Suchodoletz, Dirk von, Timo Mühlhaus, Christoph Garth, and Björn Usadel. 2024. *Software repository for the NFDI.* Structure graphics for this short paper. Visited on July 7, 2025. https://events.gwdg.de/event/658/contributions/2386/.

Weil, Heinrich Lukas, Kevin Schneider, Marcel Tschöpe, Jonathan Bauer, Oliver Maus, Kevin Frey, Dominik Brilhaus, et al. 2023. “PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research”. *The Plant Journal,* 1–15.

Veranstalter

Landesinitiative für Forschungsdatenmanagement in Baden-Württemberg



Die E-Science-Tage werden von der Landesinitiative für Forschungsdatenmanagement in Baden-Württemberg (bwFDM) 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.

Technologische und politische Entwicklungen stellen die aktuelle Forschung und das Forschungsdatenmanagement vor neue Herausforderungen. Die Bewältigung dieser erfordert ein koordiniertes Umfeld. Die E-Science-Tage 2025 boten Verantwortlichen und Interessierten die Möglichkeit, sich zu den aktuellen Entwicklungen im Bereich Forschungsdatenmanagement auszutauschen und zu vernetzen. Mit dem Thema „Research Data Management: Challenges in a Changing World“ lag der Fokus auf der Diskussion der aktuellsten Herausforderungen, die an Forschung und an das Management von Forschungsdaten gestellt werden. In diesem Rahmen wurden Lösungsansätze, neue Entwicklungen und Werkzeuge des Forschungsdatenmanagement thematisiert. Der vorliegende Tagungsband ist eine Sammlung von Vorträgen und Postern zu diesen Themen.

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386