Forschungsdatenmanagement für ein interdisziplinäres Verbundprojekt

Matthias Grönewald [1], Rainer Niekamp[2], Oliver Gutfleisch [1] und Jörg Schröder [2]

[1]Funktionale Materialien, Materialwissenschaft, Technische Universität Darmstadt
[2]Institut für Mechanik, Universität Duisburg-Essen

Ein verantwortungsbewusster und transparenter Umgang mit Forschungsdaten ist für die Qualität und das Ansehen der wissenschaftlichen Forschung von wesentlicher Bedeutung. Digitale Technologien bestimmen zunehmend die wissenschaftliche Arbeit mit Forschungsdaten. Dies beeinflusst Forschungsthemen, Fragen und Methoden einer Disziplin und das Selbstverständnis von Disziplinen. Der unter dem Begriff „digitaler Wandel“ zusammengefasste Transformationsprozess [1] hat eine große Dynamik und verändert auch die Denkweise im Bereich der Materialwissenschaften und die Zusammenarbeit mit benachbarten Disziplinen durch gemeinsame oder kombinierte Forschungsdaten. Vor diesem Hintergrund gibt es im Projekt Z-INF des DFG geförderten SFB/TRR 270 HoMMage zwei wichtige wissenschaftliche Visionen: nachhaltige und wiederverwendbare Forschungsdaten, die für die materialwissenschaftliche Gemeinschaft verfügbar sind, und die Eignung der Daten für maschinelles Lernen für digitale Zwillinge. Die Schaffung einer gemeinsamen Infrastruktur ist der erste Schritt.

1 Einleitung

Nach dem Verständnis natürlicher Phänomene durch Anwendung der Paradigmen der experimentellen Wissenschaft (seit Jahrtausenden), theoretischen Wissenschaft (seit Jahrhunderten) und computergestützten simulierenden Wissenschaft (seit Jahrzehnten) haben wir seit den letzten Jahren die Möglichkeit der rein datengetriebenen Wissenschaft durch die Verwendung von Algorithmen aus dem Bereich des Maschinellen Lernens [2]. Dieser neueste Ansatz wird in dem multidisziplinären Gemeinschaftsprojekt, SFB/TRR 270 HoMMage verfolgt, um neue Magnetmaterialien mit hervorragenden Eigenschaften für eine effiziente Energiekonversion zu finden. Das zugehörige INF-Projekt bietet Infrastruktur und Unterstützung zum Sammeln und Speichern, der durch physikalische oder in-silico-Experimente erzeugten Daten und deren Wiederverwendung im Sinne des FAIR-Prinzips (eng.: *findable, accessible, interoperable, reusable*) [3]. Eine Basis dieser Infrastruktur ist das elektronische Laborbuch (ELB), die Schnittstelle für die Wissenschaftler, um die experimentellen Ergebnisse in strukturierter Form mit den notwendigen Metadaten abzulegen. Eine weitere Basis ist die dezentrale Speicherlösung für das gesamte gemeinsame Projekt.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13734

In dieser Veröffentlichung möchten wir den Beginn dieses interinstitutionellen Forschungsdatenmanagements (FDM) vorstellen, mit Schwerpunkt auf den Eigenschaften des elektronischen Laborbuchs, der Integration des FDM in die Infrastruktur der Universitäten und betrachteten Nutzungsszenarien.

2 Ziel und Rahmenbedingungen

Das Verbundforschungsprojekt SFB/TRR 270 *Hysteresis design of magnetic materials for efficient energy conversion* - kurz *HoMMage* - widmet sich der Erforschung neuer magnetischer Materialien. In modernen Technologien zur Energieumwandlung sind sowohl Permanentmagnete mit maximierter Hysterese, als auch Weichmagnete mit minimierter Hysterese wichtige Komponenten. Beide Magnettypen haben vielfältige Anwendungsfelder, Permanentmagnete mit hoher gespeicherter Energiedichte unter anderem im Bereich der Windgeneratoren oder der Elektromobilität, Weichmagnete z.B. im Bereich magnetische Kühlung unter Anwendung des magnetokalorischen Effekts. Allen Anwendungen ist gemein, dass sie von, für den jeweiligen Einsatz gezielt verbesserten, neuen Magnetmaterialien profitieren. Deshalb suchen innerhalb des SFB/TRR 270 Forschende aus unterschiedlichen Disziplinen wie der Materialwissenschaft, der Physik, der Chemie oder der Fertigungstechnik, verteilt über mehrere Arbeitsgruppen an fünf Standorten, nach neuen innovativen Materialien und Verarbeitungswegen. Die Forschungsansätze finden dabei auf unterschiedlichsten Skalen, von Manipulation auf atomarer Ebene bis hin zu Verformungstechniken an großen Werkstücken, statt. Auf der Basis dieser Verbindung von theoretischen und experimentellen Gruppen, die ihre Ergebnisse und ihr Wissen kontinuierlich austauschen und verknüpfen, sollen so auch Wege entwickelt werden mittels computergestützter Methoden neue vielversprechende Materialzusammensetzungen vorherzusagen. Ein zentrales Forschungsdatenmanagement ist dabei ein entscheidendes Element. Das INF-Projekt zielt als zentrales Serviceprojekt deshalb innerhalb des SFB/TRR 270 darauf ab, die Verwaltung von Forschungsdaten auf nachhaltige Weise gemäß den FAIR-Grundsätzen [4] für alle Teilnehmer:innen bereitzustellen. Zusammengefasst sind die wichtigsten Herausforderungen: Komplexität, Heterogenität und Größe der Daten.

Eine der ersten Aufgaben war die Definition von Datenerfassungs- und Verwaltungsplänen (engl. data management plans, DMPs). Mit ihrer Hilfe wird die Auswertung geeigneter Metadatenformate für experimentelle und simulierte Daten möglich. Ein elektronisches Laborbuch auf Basis der FLOS-Software *eLabFTW* [5] wurde dem SFB/TRR 270 zur Verfügung gestellt. Dieses wird ständig aktualisiert und erweitert. Eine zusätzliche Benutzeroberfläche zum Hochladen, Suchen und Herunterladen unter anderem von Softwarebibliotheken für Datenkonvertierung und -reduktion wird das Toolset vervollständigen, das für die Analyse der gesammelten Forschungsdaten und Metadaten mithilfe von Algorithmen für maschinelles Lernen erforderlich ist. Diese Analyse wird die Entdeckung neuer Materialien zu unterstützen. Eine weitere wesentliche Aufgabe für INF ist die Schulung durch regelmäßige Workshops und Beratung aller Mitglieder des SFB/TRR 270 im Bereich FDM sowie deren Sensibilisierung.

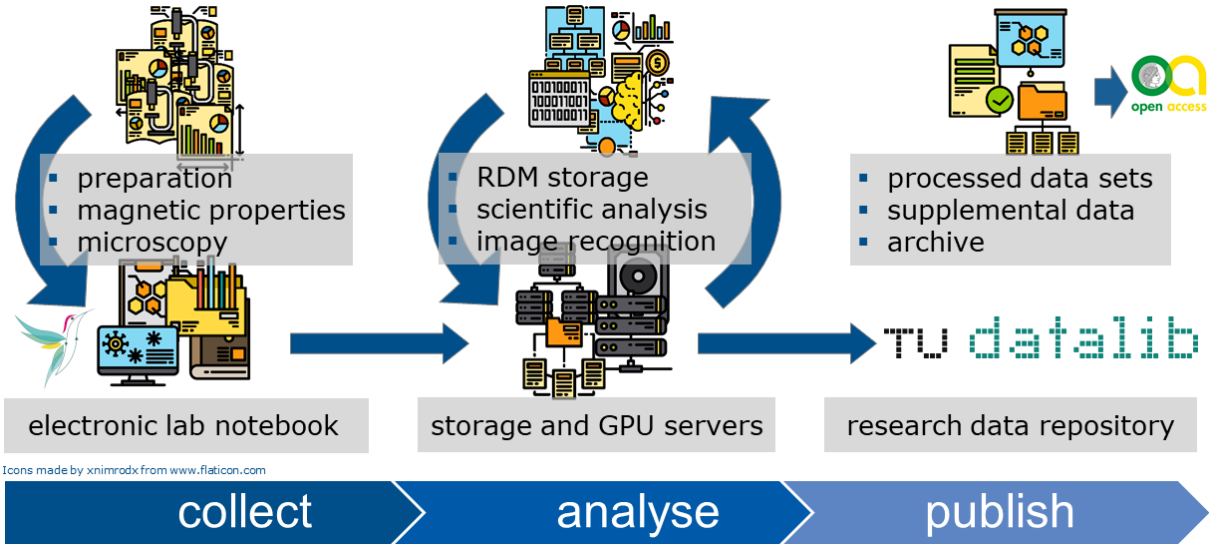

Abbildung 1: Die drei Hauptelementen der SFB/TRR 270 FDM-Infrastruktur: gemeinsam genutztes ELB, dezentraler „micro-cluster" aus Datenservern und Systemen optimiert für Maschinelles Lernen, institutionelle Forschungsdatenrepositorien zu Langzeitarchivierung.

3 Umsetzung

Das INF-Projekt betreibt dabei nur die Teile der Infrastruktur selbst, die nicht bereits durch zentrale Dienste der beteiligten Institutionen abbildbar sind. Die Vernetzung und Integration der unterschiedlichen Angebote in ein zentrales überinstitutionelles und interdisziplinäres FDM stellen dabei ihre eigenen Herausforderungen. Dabei werden prinzipiell alle Phasen eines typischen Datenlebenszyklus [6] angesprochen, wobei sich die Betreuung auf einige zentrale Bausteine fokussiert. Generell lässt sich die im Rahmen dieses Projekts unterstützte Infrastruktur auf drei Kernelemente zusammenfassen: ein gemeinsam genutztes elektronisches Laborbuch, ein kleiner Verbund aus Servern die neben Datenspeicher auch spezielle Rechenkapazität für Maschinelles Lernen bieten, sowie die Nutzung von Forschungsdatenrepositorien, unter anderem in Form der institutionell betriebenen *TU-datalib* [7].

Das elektronische Laborbuch

Die zentrale Schnittstelle zwischen den Wissenschaftler:innen im SFB/TRR 270 und dem Datenmanagement ist das verwendete elektronische Laborbuch. Auf dem Markt sind dutzende kommerzielle und wenige Open-Source Lösungen zu finden. Die wichtigsten Auswahlkriterien für dieses interdisziplinäre Verbundprojekt waren:

- Einfache intuitive Handhabung
- Verlinkbarkeit von Experimenten untereinander und mit Datenbankobjekten

- Freie Definierbarkeit von Vorlagen
- Die Möglichkeit beliebige Datenformate hochzuladen
- Support bei technischen Fragen und Erweiterungen der Funktionalität
- Vorzugsweise Open-Source

Eine Vorauswahl ergab mehrere Kandidaten aus denen *eLabFTW* nach einer Testphase, an der sowohl Forschende als auch Mitarbeiter:innen des Infrastrukturbetriebs beteiligt waren, *elabFTW* als beste Lösung gewählt wurde. Die Grundeinheit der Daten sind die Experimente und Proben mit den Metadaten, den experimentellen Daten aus Versuchen und Simulationen, erläuternden Texten und Bildern sowie den Referenzen auf andere Objekte innerhalb der Datenstruktur des ELB. Diese Objekte können sowohl andere Experimente als auch Messapparaturen oder Methodiken sein, die in der Datenbank des ELB abgelegt wurden. Die Gesamtstruktur dieser Daten wird nicht in Form einer Datenhierarchie dargestellt, sondern ergibt sich durch die Vernetzung der Objekte vergleichbar eines *knowledge graphs*. Mit Hilfe von Suchparametern und über eine API können daraus Forschungsdaten metadatengestützt abgerufen werden.

Integration des SFB/TRR 270 FDM mit der lokalen Infrastruktur

Die überinstitutionelle Nutzung der gemeinsamen Infrastruktur und die Vernetzung mit den bestehenden Diensten ist allgemein nicht trivial. Neben vielen Herausforderungen wurden aber auch bereits einige Lösungen gefunden.

Eine dieser erfolgreich angewandten Lösungen baut auf dem Werkzeug RDMO, das als Dienst an mehreren Standorten des Verbundforschungsprojekts zur Verfügung steht. Bereits zu Beginn des Projekts wurden damit DMPs erstellt und im weiteren Verlauf aktualisiert. Auf Basis derer können gemeinsame Standards (Datenformate, Software, etc.) identifiziert werden und ein allgemeiner Überblick über die Anforderungen an das FDM gewonnen werden. Unter anderem wurden so früh der große Speicherbedarf für kollaborative „lebende Forschungsdaten", also solche die ausgetauscht werden und in nachgelagerten Schritten ausgewertet, prozessiert oder umgewandelt werden, erfasst.

Ebenfalls ein Erfolg ist das durch eine der teilnehmenden Universitäten zur Verfügung gestellte zentrale ELB mit Erweiterung durch einen Forschungsdatenspeicher. Auch durch kontinuierliche Anpassung kann es dem Bedarf der heterogenen Gruppe der Forschenden gerecht werden. Die Anpassungen finden immer in Abstimmung mit zentralen Infrastrukturbereichen statt, um möglichst interoperables FDM zu gewährleisten. Aktuell bestehen Bemühungen auch Kapazitäten des HHLR-Zentrums der TU Darmstadt in Form eines Datenprojekts, mit dem große Mengen an Forschungsdaten gespeichert werden können, die direkt für Computerprojekte der theoretischen Gruppen verfügbar sind, in das FDM des Forschungsverbundes zu integrieren.

Ambivalenter sind die Ergebnisse der Bestrebungen einer flexiblen Nutzerauthentifizierung und einer kollaborativen Plattform einschließlich gemeinsam zugänglicher Sync&-

Share-Lösungen zu bewerten. Für eine gemeinsame Authentifizierung bestehen allgemein Optionen [8], diese aber praktisch in die einzelnen Dienste zu integrieren, stellt viele komplexe Herausforderungen im Detail. Hier sind auch die überregionalen Möglichkeiten noch ausbaufähig. Die Initiativen im Rahmen der NFDI [9] bieten derzeit die Chance auf Verbesserung. Momentan ist der Bedarf an Kollaborationsplattformen noch nicht gedeckt.

Zuletzt bestehen auch spezifische Anforderungen, für die es momentan keinerlei sinnvolle Integration in bestehende Infrastruktur gibt. Im Rahmen des SFB/TRR 270 ist dabei die große Datenmenge, maßgeblich aufgrund einer Vielzahl eingesetzter bildgebender Messverfahren, eine der zentralen Ursachen. In Verbindung mit dem Forschungsziel diese auch für Maschinelles Lernen zur Verfügung zu stellen, entsteht so ein zu spezialisierter Bedarf als das zentrale Einrichtungen aktuell Lösungen bieten können. In der Regel verfügen Angebote entweder nicht über die notwendigen Hardwarekapazitäten oder bieten nicht die Softwareumgebung, die nötig ist um eine Integration mit den anderen Bausteinen des FDM zu ermöglichen. Ein im Rahmen des INF-Projekts betriebener Kleinstverbund („micro cluster") aus spezialisierten Komponenten, kann hier den Anforderungen gerecht werden. Die Kombination aus Datenspeichern und auf Maschinelles Lernen hin optimierte Server stellt dabei die notwendige Speicher- und Rechenkapazität für das gesamte Verbundprojekt. Auch besteht so durch den direkten Zugriff auf die Datenstrukturen die Möglichkeit mit den lebenden Forschungsdaten Untersuchungen und Analysen mit dem Zweck von Struktur-Eigenschafts-Vorhersagen zu erproben, ohne aufwendig neue Datenstrukturen außerhalb der eigentlichen FDM Infrastruktur anlegen zu müssen. Gleichzeitig wird die Lücke zwischen der Rechenleistung eines Arbeitsplatzrechners und einem Rechenprojekt am HPC-Zentrum geschlossen, die Forschenden niederschwelligen Zugang für Vorversuche bietet.

Interaktion der Nutzer und die betrachteten Nutzerszenarien

Eine entscheidende Größe für die Akzeptanz des Systems bei vielen Forschenden ist der niederschwellige Einstieg zur Erfassung von Forschungsdaten und Forschungsmetadaten. Daneben kann sich erfolgreiches Forschungsdatenmanagement aber auch durch einen wissenschaftlichen Mehrwert auszeichnen. Die Verwendung gemeinsamer Vorlagen für Datenstrukturen begünstigt beide Aspekte. Bestehende Vorlagen können einfach nach genutzt und bei Bedarf angepasst werden, umso den Aufwand für die einzelnen Wissenschaftler:innen zu minimieren. Durch diesen iterativen Prozess werden die den Forschenden innerhalb des Verbundprojekts zur Verfügung stehenden Vorlagen kontinuierlich optimiert. Gleichzeitig wird eine Standardisierung der Dokumentation über einzelne Gruppen hinaus gefördert und Metadatenbeschreibungen angeglichen.

Die im Fokus des INF-Projekts liegenden Nutzungsszenarien greifen dabei auf die modellhafte Beschreibung der archetypischen „Datenlieferant:innen" und „Datenanalyst:innen" zurück. Die „Datenlieferant:innen" sind experimentelle Gruppen, die Forschungsdaten, wie Eigenschaftsmessungen, Bilddaten und zugehörige Metadaten, über das ELB zur Verfügung stellen. Die „Datenanalyst:innen" beschreiben theoretisch arbeitende Gruppen, die

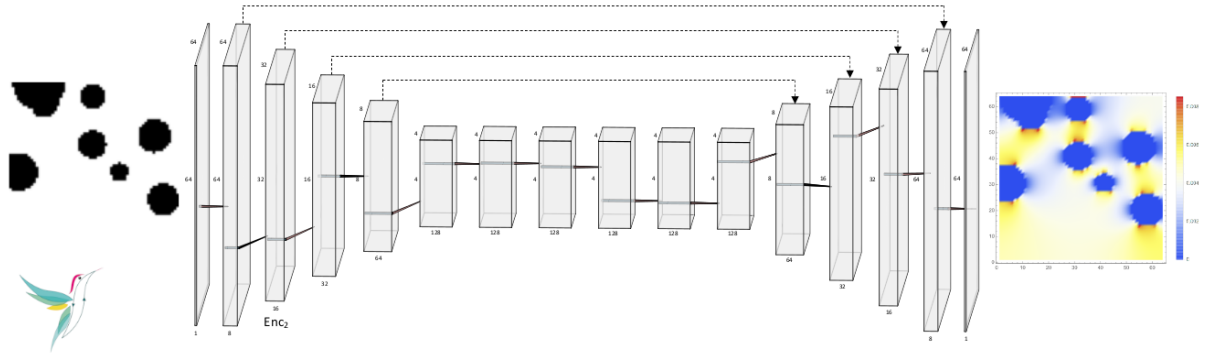

Abbildung 2: Im Prozesse des Maschinellen Lernens werden die von den erhobenen Forschungsdaten von den „Datenanalyst:innen" strukturiert und auf Deskriptoren reduziert und damit dann mehrstufig ein künstliches neuronales Netzwerk trainiert, um damit Strukturen mit interessanten Eigenschaften zu vorherzusagen.

diese Daten abfragen, ggf. mit Hilfe der Metadaten strukturieren, und dann mittels Maschinellen Lernens auswerten (vgl. Abb. 2). Aus diesem kooperativen Arbeitsablauf sollen Kandidaten von Materialkombinationen und Herstellungsbedingungen für neue Stoffe mit hervorstechenden Eigenschaften ermittelt werden.

4 Zusammenfassung

An das FDM innerhalb eines interdisziplinären Verbundprojektes werden viele Herausforderungen gestellt. Neben technischen Anforderungen und Inkompatibilitäten bei der Integration bestehender Infrastruktur erzeugt die Kollaboration zwischen den fachlich verschiedenen forschenden Gruppen auch neuen zum Teil spezialisierten Bedarf. Das INF-Projekt innerhalb des SFB/TRR 720 HoMMage hat neben dem Ziel diesem allem mit einer Kombination selbstverwalteter und zentraler Infrastruktur zu begegnen auch die Aufgabe den Forschenden innerhalb des Verbunds Wissen und Unterstützung zu einem besseren FDM zu bieten. So kann mit Hilfe strukturierter Forschungsdaten und modernen Computerverfahren ein erfolgreicher Beitrag zur Entwicklung neuer vielversprechender magnetischer Materialien geliefert werden.

Danksagungen

Mit besonderem Dank für die Unterstützung durch Sascha Sczyrba (UDE), Stefan Beyer (UDE), Andreas Hönl (TUDa), Stephan Diefenbach (TUDa), sowie dem gesamten HRZ-Team der TU Darmstadt und dem des ZIM an der Universität Duisburg-Essen.

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 405553726 – TRR 270, Teilprojekt Z-INF.

ORCID IDs

- Matthias Grönewald https://orcid.org/0000-0002-3480-9102
- Oliver Gutfleisch https://orcid.org/0000-0001-8021-3839
- Jörg Schröder https://orcid.org/0000-0001-7960-9553

Literaturverzeichnis

[1] DGM e.V., DGM: Digitaler Wandel in der Wissenschaft: Herausforderungen und Chancen für das Fachgebiet Materialwissenschaften und Werkzeugtechnik, Anmerkungen der Fachkollegien Materialwissenschaft und Werkstofftechnik der Deutschen Forschungsgemeinschaft, (Stand 2018).

[2] A. Takbiri, H. Kazemi, N. Nasrabadi. A data-driven surrogate to image-based flow simulations in porous media. Computers & Fluids, 2020.

[3] C. Draxl, M. Scheffler, NOMAD: The FAIR concept for big data-driven materials science, MRS Bulletin, 43, 676-682, 2018.

[4] M. D. Wilkinson, et al., The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 2016.

[5] `https://www.elabftw.net/`, (Stand 23.04.2021).

[6] Rat für Informationsinfrastrukturen, Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen, S. 160, 2016.

[7] `https://tudatalib.ulb.tu-darmstadt.de/`, (Stand 23.04.2021).

[8] `https://www.aai.dfn.de/`, (Stand 23.04.2021).

[9] `https://www.nfdi.de/`, (Stand 23.04.2021).