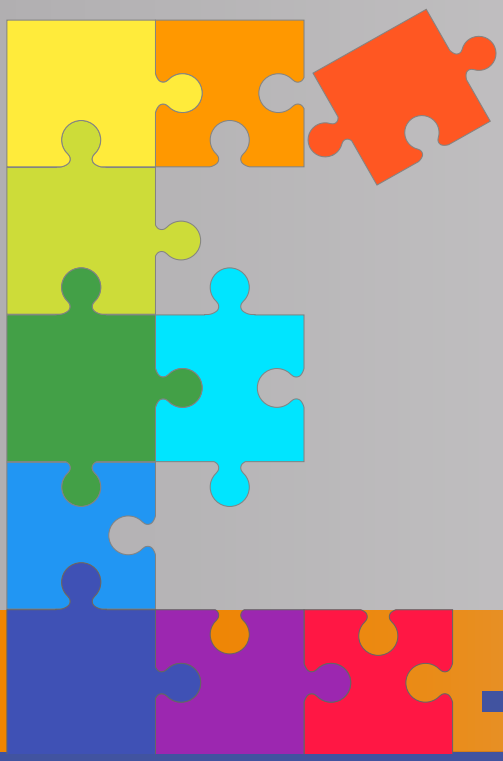
Vincent Heuveline
Fabian Gebhart
Nina Mohammadianbisheh
(Hrsg.)



-Science-Tage 2019

Data to Knowledge

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

E-Science-Tage 2019

E-Science-Tage 2019

Data to Knowledge

Herausgegeben von

Vincent Heuveline, Fabian Gebhart und

Nina Mohammadianbisheh

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

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

Publiziert bei heiBOOKS,
Universitätsbibliothek Heidelberg 2020.

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-598-2
doi: https://doi.org/10.11588/heibooks.598

ISBN 978-3-948083-15-1 (Softcover)
ISBN 978-3-948083-14-4 (PDF)

Inhaltsverzeichnis

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

Theresia Bauer

E-Science-Tage 2019: „Data to Knowledge"

Der systematische Zugang zu digitalen Datenbeständen wird für neue wissenschaftliche Erkenntnisse sowie für Innovationen und Technologietransfer immer wichtiger. Die Qualität unserer Modelle und Prognosen und die Wirksamkeit unserer Maßnahmen und Therapien, sei es zur Begrenzung der globalen Erwärmung oder zur Bekämpfung einer Krankheit, ist abhängig von der Qualität der Daten, mit denen wir unsere Modelle und Prognosen füttern. Deswegen müssen wir die Möglichkeiten zur Nutzung und Nachnutzung von Forschungsdaten über alle Fachdisziplinen hinweg verbessern.

Das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg hat bereits 2014 gemeinsam mit Vertretern aus Hochschulen und Forschungseinrichtungen des Landes in einer bundesweiten Vorreiterfunktion mit dem Fachkonzept „E-Science" zur Weiterentwicklung der wissenschaftlichen Infrastruktur in Baden-Württemberg Handlungsoptionen für den Zugang zu und die Nutzung von digitalisierten Datenbeständen entwickelt und deren Umsetzung mit mehreren Fördermaßnahmen vorangetrieben.

Mit der Förderung des Aufbaus von vier Science Data Centers in Baden-Württemberg gehen wir 2019 einen Schritt weiter: Forschungsdaten sollen für die gesamte Wissenschaft und darüber hinaus nach den FAIR-Prinzipien nachnutzbar gemacht werden. Zusätzlich werden die Science Data Centers bedarfsgerechte Workflows, Analyse-Werkzeuge und Dienste-Portfolios entwickeln und zur Verfügung stellen. Als Kompetenzzentren für die digitale datengetriebene Forschung werden sie weiterhin standortübergreifend bei der Entwicklung von Standards, Schnittstellen und Rahmenbedingungen der Archivierung und Nutzung von Forschungsdaten zusammenwirken.

Für den Wissenschaftsstandort Baden-Württemberg und für Deutschland sind wissenschaftsgetriebene Datenzentren von herausragender Bedeutung – nicht nur, weil sie Voraussetzung für Spitzenforschung und Exzellenz sind. Mit der Entwicklung und dem Betrieb von Datenzentren gestalten Wissenschaftlerinnen und Wissenschaftler gemeinsam mit den Rechenzentren und den wissenschaftlichen Bibliotheken die Rahmenbedingungen für die Verwahrung von und für den Zugang zu Forschungsdaten sowie für die Verwendung und Analyse der Daten.

Mit dem Motto „Data to Knowledge" haben die Veranstalter der E-Science-Tage 2019 genau diese Frage des Brückenschlags von den Konzepten, der Technik, den Infrastrukturen und den Analysewerkzeugen hin zur Forschung und zum Erkenntnisgewinn aufge-

Theresia Bauer: Grußwort der Ministerin für Wissenschaft, Forschung und Kunst Baden-Württemberg.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 11–12. DOI: https://doi.org/10.11588/heibooks.598.c8410

griffen. Ich freue mich sehr über das lebhafte Interesse an diesem Thema und über die Bereitschaft aller Beteiligten, die Ergebnisse der Tagung in einem Tagungsband zu veröffentlichen.

Theresia Bauer MdL
Ministerin für Wissenschaft, Forschung
und Kunst des Landes Baden-Württemberg

Grußwort des Rektors der Universität Heidelberg

Prof. Dr. Dr. h.c. Bernhard Eitel
zur Eröffnung der E-Science-Tage am 28.03.2019

Sehr geehrte Frau Ministerin, liebe Frau Bauer,
sehr geehrter Herr Staatssekretär Meister,
lieber Herr Heuveline,
liebe Kolleginnen und Kollegen,
liebe Kommilitoninnen und Kommilitonen,

herzlich begrüße ich Sie alle hier in der Aula der Neuen Universität zu den E-Science-Tagen 2019. Data to Knowledge, so Ihr Thema, zeigt, worum es geht: Daten spielen in unserer Welt eine zunehmend wichtigere Rolle, Chancen wie Risiken wachsen. Umso mehr ist es von Bedeutung, dass sich die Wissenschaft mit dem Management wissenschaftlicher Daten und dem Wissen, das sich aus der Datenflut generieren lässt, beschäftigt. Die Digitalisierung aller Lebensbereiche schreitet voran, immer mehr Energie wird die Menschheit für diese Technologien bereitstellen (müssen), immer relevanter wird die Lösung sich andeutender potentiell wachsender Konflikte zwischen Digitalisierung und Umweltschutz. Daher ist es richtig, dass das Thema nicht nur von Mathematikern, Informatikern und Computer Scientists angegangen wird, sondern längst Eingang in alle Wissenschafts- und Gesellschaftsbereiche bis hin zur Umweltökonomie, Philosophie und Theologie gefunden hat. Daher sind die E-Science-Tage an einer Forschungsuniversität wie Heidelberg gut aufgehoben: Ein breites Fächerspektrum verspricht perspektivenreiche Einsichten und das besondere Flair der Stadt bereitet dafür das richtige Environment. Sie tagen also zur richtigen Zeit mit dem richtigen Thema am richtigen Ort - an der ältesten Universität in Deutschland, stets dynamisch mit einem höchst relevanten Thema.

Ich wünsche Ihnen allen, die Sie von nah und fern nach Heidelberg gekommen sind, einen fruchtbaren Austausch und viele Anregungen für Forschung, Lehre und Anwendung. Und wenn Sie wieder zurück fahren an Ihre Heimatinstitutionen, so hoffe ich, dass Sie ein wenig Heidelberg mitnehmen werden, ein wenig von diesem besonderen Lebendigen Geist, der in dieser Stadt und dieser Universität als Lebensform wirkt – ganz so, wie es die Widmung am Eingangsportal zu diesem Gebäude seit 1929 verkündet: "Dem Lebendigen Geist". Lassen Sie sich von ihm inspirieren!

Prof. Dr. Dr. h.c. Bernhard Eitel
Rektor der Universität Heidelberg

Bernhard Eitel: Grußwort des Rektors der Universität Heidelberg.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 13. DOI: https://doi.org/10.11588/heibooks.598.c8411

Vorwort der Herausgeber

Vincent Heuveline, Fabian Gebhart und Nina Mohammadianbisheh

Das gegenwärtige Informationszeitalter hat den Zugriff auf Wissen demokratisiert. War der Erkenntnisgewinn aus niedergeschriebenen Informationen früher ein Privileg für wenige, ist heute das Gegenteil der Fall: Gigantische Mengen an wissenschaftlichen Daten sind global zugänglich und in Sekundenschnelle abrufbar. Neueste Forschungsergebnisse können inzwischen nebenbei auf dem Smartphone gelesen werden, im Café oder im Zug auf dem Weg zur Arbeit. Weltumspannende digitale Netzwerke bilden das kollektive Gedächtnis einer großen globalen Forschungsgemeinschaft. Die digitale Informationsfülle ist allgegenwärtig und für die Arbeit von Wissenschaftlerinnen und Wissenschaftlern inzwischen unverzichtbar. Sie können Forschungsdaten digital erfassen, analysieren, archivieren und wiederverwerten. Digitale Daten vereinfachen die wissenschaftliche Recherche und verringern den dafür notwendigen Zeit- und Ressourcenaufwand immens. Durch diese Errungenschaften hat die Wissenschaft eine starke Effizienzsteigerung erlebt, ohne dabei an Qualität einzubüßen.

Der Überfluss an Daten und Informationen bietet also viele Chancen, birgt aber auch große Herausforderungen: Eine hochwertige Strukturierung und Archivierung der Daten ist zwingend erforderlich, um im Stande zu bleiben, den Überblick zu wahren und das kollektiv verfügbare Wissen sinnvoll zu nutzen. Neben modernen Technologien und ausgeklügelten Algorithmen werden universitätsübergreifende Strukturen benötigt, welche die Weiternutzung der Daten ermöglichen. Ohne diese Art von Strukturbildung, wären die Wissenschaften einer scheinbar unbezwingbaren Flut an Daten ausgesetzt.

Schon als die E-Science-Tage im Jahr 2017 erstmalig veranstaltet wurden, war es unser Anliegen, genau diese Chancen und Herausforderungen in den Blick zu nehmen und einen sinnbildlichen Leuchtturm in der „Datenflut“ zu schaffen. Standen damals vor allem die Forschungsdateninfrastruktur und die nachhaltige Aufbewahrung von Daten im Vordergrund, ging es bei den E-Science-Tagen 2019 mit dem Motto „Data to Knowledge“ um eine ganz fundamentale Frage: Wie kann aus Daten Wissen gewonnen werden? Der vorliegende Tagungsband demonstriert eindrücklich, wie komplex, vielseitig und spannend sich dieser Weg „Data to Knowledge“ gestaltet und wie wichtig eine enge und partnerschaftliche Zusammenarbeit zwischen Technik und Forschung dabei ist. Die Beiträge des Bandes spiegeln damit den abwechslungsreichen und erkenntnisbringenden Austausch wider, den wir im Rahmen der E-Science-Tage 2019 erleben durften – einen Austausch, den wir auch in Zukunft intensiv fortsetzen möchten und müssen.

Gerne nutzen wir diese Gelegenheit, um uns noch einmal außerordentlich bei allen Förderern, Mitwirkenden und Besuchern der E-Science-Tagen 2019 zu bedanken. Ebenso gilt unser Dank natürlich den Autorinnen und Autoren, die zur Entstehung dieses Tagungs-

Vincent Heuveline, Fabian Gebhart und Nina Mohammadianbisheh: Vorwort der Herausgeber.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 14–15. DOI: https://doi.org/10.11588/heibooks.598.c8412

bandes beigetragen haben. Wir hoffen, dass der Band den Weg durch die „Datenflut“ erleichtert und viele Impulse und Anregungen für die weitere Arbeit im Bereich der Forschungsdaten und des Forschungsdatenmanagements geben kann.

Vincent Heuveline
Fabian Gebhart
Nina Mohammadianbisheh

Teil I.

Vorträge

Konzeption und Institutionalisierung des FDM – aus der Erfahrung eines Forschungsprojekts in den digitalen Geisteswissenschaften

Martin Spenger[1] und Stephan Lücke[2]
[1]Universitätsbibliothek der Ludwig-Maximilians-Universität München, Deutschland;
[2]IT-Gruppe Geisteswissenschaften, Ludwig-Maximilians-Universität München

Dieser Beitrag zeigt am Beispiel des Forschungsdatenmanagements an der LMU, wie Wissenschafts- und Infrastrukturpartner erfolgreich zusammenarbeiten können. Im Rahmen des Modellprojekts „eHumanities – interdisziplinär“ wird die Zusammenarbeit anhand des Pilotpojekts VerbaAlpina verdeutlicht. Der erste Teil des Aufsatzes beschreibt VerbaAlpina und die Besonderheiten aus der Perspektive der digitalen Geisteswissenschaften. Der zweite Teil bringt als Infrastrukturpartner die Universitätsbibliothek der LMU ins Spiel. Abschließend folgt ein kurzer Überblick über das Projekt „eHumanities – interdisziplinär“.

1. Teil: Das Projekt VerbaAlpina

Das Projekt VerbaAlpina[1](VA) ist ein von der DFG gefördertes Langfristvorhaben[2] mit einer Perspektive bis 2025 und befindet sich derzeit in der zweiten Förderphase, die noch bis zum Herbst 2020 andauert. Es handelt sich um ein interdisziplinäres Projekt im Umfeld der Digital Humanities (DH). Der Schwerpunkt des Interesses liegt auf den Sprachwissenschaften, daneben erfolgt jedoch auch eine intensive und streckenweise exemplarische Auseinandersetzung mit den Herausforderungen der Informationstechnologie. Das von VA zusammengetragene und analysierte Datenmaterial kann darüber hinaus auch für andere Disziplinen wie etwa die Ethnographie, die Archäologie oder auch die Geschichtswissenschaften von Interesse sein. VA betreibt unter anderem eine intensive Methodenreflexion, die hauptsächlich in der Rubrik „Methodologie“ auf der Projektwebseite dokumentiert ist. Zu einigen der im Folgenden thematisierten Punkte finden sich dort ausführlichere Darlegungen, auf die hier generell hingewiesen sei.

Initiatoren und Träger des Projekts sind Thomas Krefeld vom Romanistischen Seminar der LMU sowie Stephan Lücke von der IT-Gruppe Geisteswissenschaften der LMU. In der laufenden zweiten Förderphase verfügt VA über je zwei Doktorandenstellen im Bereich der Sprachwissenschaft (einer zuständig für die Romanistik/Slawistik, der andere für die Germanistik) und in der Informatik (für Datenbank- und Frontendentwicklung). Für die

[1] http://www.verba-alpina.gwi.uni-muenchen.de/
[2] http://gepris.dfg.de/gepris/projekt/253900505

Martin Spenger und Stephan Lücke: Konzeption und Institutionalisierung des FDM – aus der Erfahrung eines Forschungsprojekts in den digitalen Geisteswissenschaften. In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge, Heidelberg: heiBOOKS, 2020, 19–31. DOI: https://doi.org/10.11588/heibooks.598.c8414

umfangreiche Koordinationsarbeit mit den zahlreichen Projektpartnern steht eine weitere Doktorandenstelle zur Verfügung, deren Inhaberin ihre Doktorarbeit im thematischen sprachwissenschaftlichen Umfeld des Projekts vorbereitet. Unterstützt wird das Team durch eine Reihe von Hilfskräften, die vor allem für die strukturierte Datenerfassung eingesetzt werden.

Vorrangiges Ziel von VA ist die systematische Erfassung und Analyse der im Alpenraum verbreiteten morpholexikalischen Typen, die zur Bezeichnung ausgewählter „Objekte“ (Konzepte[3] , Begriffe) Verwendung finden oder auch fanden. Vereinfacht gesagt, steht dabei die Frage im Zentrum, welche Konzepte an welchen Orten mit welchen Wörtern bezeichnet werden. So wird z.B. das Konzept BUTTER (VA verwendet zur Bezeichnung der Konzepte stets Versalien, um damit den Unterschied zu den Bezeichnungen/Wörtern klar zu machen) in unterschiedlichen Regionen des Alpenraums mit unterschiedlichen morpholexikalischen Typen bezeichnet: In Bayern und Österreich herrscht der Typ *Butter* vor, im Alemannischen ist *Anke* weit verbreitet, in Italien nennt man die BUTTER *burro*. VA ist fokussiert auf die jeweiligen morpholexikalischen Typen, das heißt, phonetische Variationen werden zwar vielfach dokumentiert, jedoch nicht systematisch erfasst oder gar analysiert. Eine vollständige Erfassung des morpholexikalischen Materials ist unmöglich. Daher beschränkt VA seine Dokumentation und Untersuchung auf typisch alpine Konzeptdomänen wie etwa die Almwirtschaft (Schwerpunkt der ersten Projektphase von 2014 bis 2017) oder auch Natur und Umwelt (laufende Projektphase) sowie Tourismus (geplant für die Projektphase ab 2020).

Der geographische Rahmen des Untersuchungsgebiets ist aus pragmatischen Gründen auf die Ausdehnung der sogenannten Alpenkonvention[4] festgelegt worden. Als Datengrundlage dienen in erster Linie traditionell in Buchform publizierte sogenannte Sprachatlanten[5], die konzeptorientiert („onomasiologisch“) in Kartenform die Verbreitung von morpholexikalischen Typen für die Bezeichnung vorgegebener Konzepte präsentieren. Die systematische Erfassung dieser Daten ist kaum automatisierbar und bedeutet einen erheblichen Aufwand an Handarbeit. Die Schwierigkeit liegt dabei weniger im (nicht eingesetzten) OCR-Verfahren, sondern vielmehr an der kartographischen Zuordnung bestimmter Bezeichnungen zu den einzelnen auf den Karten verzeichneten Punkten.

Erschwert wird diese automatische Erfassung überdies durch die Verwendung von Symbolen auf der Karte, die die Orthographie des entsprechenden Typs also unterdrücken. Die Daten aus den Sprachatlanten werden ergänzt um Daten aus Wörterbüchern, die, anders als die Sprachatlanten, von den morpholexikalischen Typen ausgehen und die jeweiligen durch sie bezeichneten Konzepte dokumentieren. Es werden jedoch nur solche Wörterbücher berücksichtigt, die auch Aufschluss über die geographische Verbreitung der jeweiligen Bedeutungen geben. VA verfügt über eine große Anzahl nationaler und internationaler Partner[6], die vielfach über eigene Sprachdatensammlungen verfügen, die nach Möglichkeit und Eignung ebenfalls in den Datenbestand von VA übernommen werden. In der Summe ergibt sich ein kartierbares Netz der im Alpenraum verbreiteten Bezeichnungen der

[3] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=182&letter=K#37
[4] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=182&letter=A#103
[5] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=182&letter=Q#48
[6] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=185&db=181

ausgewählten Konzepte, die auf einer interaktiven Online-Karte[7] präsentiert werden. Ein wesentlicher Mehrwert des Einsatzes der zeitgemäßen Online-Medien ist dabei der problemfreie und schnelle Wechsel zwischen der onomasiologischen und der semasiologischen Perspektive. VA stellt sämtliche von ihm erzeugten Inhalte – wozu auch alle Softwareentwicklungen gehören – nach Möglichkeit unter einer offenen Lizenz[8] zur Verfügung (CC BY-SA). Ausnahmen bestehen nur für Daten, die anderweitig unter restriktiven Lizenzen geführt werden. VA versteht sich als konsequent digitales[9] Online-Projekt und betrachtet die traditionellen Publikations- und Kommunikationsgepflogenheiten im Wissenschaftsbetrieb als überholt. Das Projekt verzichtet vollständig auf den Einsatz herkömmlicher Drucktechnologie und hält auch den Einsatz von PDF-Dokumenten als eine wenig geeignete Publikationsform, die gegenüber der Webtechnologie entscheidende Einschränkungen besitzt. Die zentralen Projektdaten werden im relationalen Datenformat in einer MySQL-Datenbank verwaltet. Als Frontend dient eine generische WordPress-Installation, für die von den Projekt-Informatikern eine Reihe von spezifischen, auf die Projektbedürfnisse zugeschnittenen, Plugins entwickelt wurden, die über GitHub[10] der Allgemeinheit unter der CC BY-SA-Lizenz zur Verfügung gestellt werden. Das Projektportal ist multifunktional. Es dient gleichermaßen als Arbeitsinstrument für die Projektmitarbeiter wie auch als Publikations- und Dokumentationsplattform und schließlich zur wissenschaftlichen Kommunikation, wobei allerdings letztere Funktionalität noch nicht konsequent ausgebaut ist. Die Idee ist, dass sich Wissenschaftler und Laien auf dem Portal registrieren und das System für den wissenschaftlichen Austausch und auch als Instrument zur Verwaltung eigener Daten verwenden.

VA unterscheidet nicht zwischen „Forschungsdaten“ einerseits und auswertenden Daten andererseits, eine Vorstellung, von der offenkundig auch die aktuelle Diskussion des Forschungsdatenmanagements geprägt ist. Sämtliche Projektdaten sind aufeinander bezogen und somit untrennbar miteinander verbunden. Aus diesem Grund verfolgt VA das Ziel, die Gesamtheit der Daten, also die gesammelten Sprachdaten sowie alle darauf bezogenen analytischen und erläuternden Texte, kartographischen Repräsentationen und sonstige Derivate als Ganzes dauerhaft und in zuverlässig zitierbarer Weise zu erhalten. Angesichts ständig ausgebauter Speicherungskapazitäten und dem im Vergleich mit mancher naturwissenschaftlichen Disziplin geringem Datenvolumen kann die Datenmenge nicht als Argument dafür gelten, auf die dauerhafte Bewahrung des kompletten Datenbestands verzichten zu müssen.

Der aus den stets nur auf Teilbereiche des Alpenraums beschränkten Sprachatlanten und Wörterbüchern zusammengetragene gleichsam „historische“ Datenbestand weist notwendigerweise eine ganze Reihe von Inkonsistenzen auf. So sind regionale Unterschiede hinsichtlich Belegdichte und Dokumentation der ausgewählten Konzepte festzustellen. Um diese Inkonsistenzen auszugleichen, wurde von VA ein Crowdsourcing-Tool entwickelt, das Nutzern im Internet erlaubt, regionale Bezeichnungen für die projektrelevanten Konzepte beizutragen. Aktuell sind auf diesem Weg exakt 11546 (Stand 30.4.19) morpholexikalische

[7] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=133&db=xxx
[8] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=182&letter=L#41
[9] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=182&letter=D#15
[10] https://github.com/VerbaAlpina/

Einzelbelege in den Datenbestand von VA gelangt. Neben dem geographischen und konzeptbezogenen Ausgleich gelangt über das Crowdsourcing auch eine (zusätzliche, da auch die konventionellen Datenquellen unterschiedliche Zeiträume dokumentieren) diachrone Perspektive in den Datenbestand, die einen Einblick in den Sprachwandel und dessen Dynamik erlaubt.

VA steht im engen Verbund mit im Wesentlichen zwei an der LMU beheimateten Institutionen: Der IT-Gruppe Geisteswissenschaften[11](ITG) sowie der Universitätsbibliothek[12](UB). Die ITG besteht seit knapp 20 Jahren und ist ursprünglich aus einer Stelle für rechnergestützte Forschung an der Fakultät für Kulturwissenschaften der LMU hervorgegangen. Sie ist zuständig für und ist getragen von sämtlichen geisteswissenschaftlichen Fakultäten, wobei ihre zentralen Aufgaben in der Planung und Betreuung der IT-Infrastruktur, der Unterstützung bei und der Durchführung von digitaler Forschung und Lehre sowie im Management der von den Wissenschaftlern erzeugten Forschungsdaten liegen. Die bei VA beschäftigten Informatiker sind direkt an der ITG angesiedelt und haben dort die Möglichkeit zum fachlichen Austausch mit anderen Beschäftigten, die in einer ganzen Reihe von weiteren DH-Projekten mit vergleichbaren Aufgaben beschäftigt sind. Dieses strukturelle Konzept gewährleistet ein hohes Maß an Synergieeffekten, von denen auch die anderen an der ITG betriebenen Projekte profitieren können. VA ist bestrebt, die für das Projekt entwickelte Software, soweit möglich und sinnvoll, generisch und modular zu konzipieren, so dass sie mit möglichst geringem Aufwand auch in anderen Kontexten weiter- bzw. wiederverwendet werden kann.

Vor allem im Hinblick auf die dauerhafte Bewahrung der Projektdaten spielt wiederum die UB die entscheidende Rolle für VA. Generell ist VA der Meinung, dass vor allem die Bibliotheken die natürlichen Ansprechpartner für alle Fragen der Datenbewahrung sind. Dies ist begründet zum einen durch die jahrhundertelange Tradition der entsprechenden Zuständigkeit, zum anderen durch die feste institutionelle Verankerung, die eine langfristige Bestandsgarantie verspricht, wie sie kaum eine andere Einrichtung in vergleichbarem Umfang besitzt. Hinzu kommt ein hohes Maß an Kompetenz, das an der UB sowohl im Hinblick auf die bibliothekarischen wie auch die informatischen Erfordernisse vorhanden ist. Drittmittelgeförderte Repositoriums-Projekte mit begrenzter Existenzperspektive erscheinen problematisch, werden von VA jedoch zusätzlich genutzt. So wird aktuell ein Datenexport an das CLARIN-D Centre Leipzig[13] vorbereitet, mit dem VA eine Kooperationsvereinbarung abgeschlossen hat.

Über den Kontakt zur UB und ITG ist VA auch in das vom Bayerischen Staatsministerium für Wissenschaft und Kunst geförderte Projekt eHumanities – interdisziplinär[14] eingebunden. VA nimmt dort die Rolle eines Pilotprojekts ein, dessen Daten exemplarisch mit Metadaten angereichert und schließlich in das institutionelle Repositorium der UB (Open Data LMU[15]) gelangen, wo sie in versionierter Form dauerhaft gesichert und auch zitierbar sind. Die Entwicklung bzw. Anpassung der Metadatenmodelle an die spe-

[11] https://www.itg.uni-muenchen.de/index.html
[12] https://www.ub.uni-muenchen.de/index.html
[13] https://clarin.informatik.uni-leipzig.de/repo/
[14] https://www.fdm-bayern.org/
[15] https://data.ub.uni-muenchen.de/

zifischen Projekterfordernisse erfolgt in enger Zusammenarbeit von VA-, ITG– und UB-Mitarbeitern. Dieser Prozess, der sich als ausgesprochen effizient erweist, ist in dieser Form nur durch die enge lokale Nachbarschaft der beteiligten Institutionen möglich. Vor dem Hintergrund dieser Erfahrung steht VA allen Konzepten, die im Hinblick auf das Forschungsdatenmanagement Lösungen mit spezialisierten zentralen Institutionen favorisieren, die dann unter Umständen sehr weit vom Ort der Projekttätigkeit liegen, skeptisch gegenüber. Demgegenüber betrachtet VA die skizzierte enge Verzahnung der beteiligten Akteure als modellhaft. Zumindest theoretisch sollte eine Übertragung dieses Konzepts auf andere Universitätsstandorte angesichts der doch weitgehend flächendeckenden Verbreitung von Universitätsbibliotheken möglich sein.

VA steht seit Längerem auch in Kontakt mit dem vom BMBF geförderten Projekt-GeRDI, das als Datenaggregator betrachtet werden kann, dessen Ziel es ist, Forschungsdaten der verschiedensten Disziplinen über einen zentralen Zugang unter Einsatz von Metadaten zugänglich zu machen. Durch die oben beschriebene Kooperation im Rahmen des Projekts eHumanities – interdisziplinär gelangen die VA-Daten über die UB auch in den Datenbestand von GeRDI.[16]

Im Zusammenhang mit dem Forschungsdatenmanagement wird für die Aufbereitung von Forschungsdaten seit einiger Zeit auch die Erfüllung der FAIR-Prinzipien propagiert bzw. bisweilen auch gefordert. Das Projekt VA hält die in diesem Akronym versammelten Postulate für durch und durch berechtigt und ist bestrebt, diesen Kriterien möglichst weitgehend zu entsprechen. Die Auffindbarkeit (findable) und Zugänglichkeit (accessible) der VA-Daten ist durch die in Zusammenarbeit mit der UB erfolgte Anreicherung um Metadaten mit deren anschließender Einbindung in Kataloge (z.B. OPAC) und Aggregatoren-Dienste (GeRDI) sowie durch das angewendete offene Lizenzmodell hinreichend gewährleistet. Die Forderung der Interoperabilität (interoparable) und bis zu einem gewissen Grad auch der Nachnutzbarkeit (reusable) erscheint nur möglich, wenn das vom Projekt gesammelte strukturierte Datenmaterial in möglichst feiner Granulierung vorliegt und auf die einzelnen Datensätze über eine URL eindeutig referenziert werden kann. Aus diesem Grund wird der Kerndatenbestand von VA nach Einzelbelegen, morpholexikalischen Typen, Konzepten und Gemeinden gruppiert, die jeweiligen Gruppen mit persistenten Identifikatoren versehen und in dieser Form, zusammen mit allen übrigen Projektdaten, versionsweise an die UB übertragen. Nach der Anreicherung um Metadaten und der Ablage im institutionellen Repositorium ist es sodann möglich, jede einzelne Instanz innerhalb der genannten Gruppierungen über eine DOI anzusprechen. Damit sind de facto projektspezifische Normdaten erzeugt, darüber hinaus ist eine der wesentlichen Forderungen erfüllt, die den Datenbestand von VA als „Linked Open Data“ qualifizieren (die Existenz einer persistenten URL). Derzeit fehlen jedoch noch die ebenfalls erforderlichen RDF-Metadaten im XML-Format, deren Erzeugung jedoch geplant ist. Die UB erwägt außerdem, zusätzlich zu den DOIs eigene persistente URLs zu erzeugen, deren Vergabe und Betreuung in der alleinigen Verantwortung der UB liegen. VA begrüßt diese Perspektive, zumal die VA-Ressourcen zusätzlich zu den DOIs über ein weiteres, davon unabhängiges System persistenter Adressen erreichbar sein werden.

[16] `https://www.gerdi-project.eu/`

Die größte Herausforderung des Forschungsdatenmanagements besteht in der langfristigen Konservierung „lebender Systeme“, wie das Projektportal von VA eines ist. VA betrachtet dieses von ihm entwickelte Projektportal als eine zeitgemäße Publikationsform, die in ihrer primären Funktion – der Veröffentlichung – mit der traditionellen Buchpublikation vergleichbar ist, aber natürlich darüber hinausgehende Möglichkeiten bietet. Der Wunsch wäre, dieses Webportal möglichst ad infinitum online verfügbar zu halten, vergleichbar mit der Bewahrung eines Buches in einer Bibliothek. Leider stehen diesem Ideal technische Schwierigkeiten entgegen, die struktureller Natur und daher bislang nicht lösbar sind. Das Problem besteht hauptsächlich in der ständigen Weiterentwicklung der Software-, konkret: Serverumgebung, innerhalb derer ein solches Projektportal läuft. Die projekt- bzw. portalspezifische Software muss in größeren Abständen immer wieder an die veränderte Umgebung angepasst werden, bedarf also mehr oder minder permanenter Pflege. VA ist zwar insofern strategisch gut aufgestellt, als das Projektportal von der ITG betreut wird, die über eine unbefristete Bestandsperspektive verfügt und im Rahmen ihrer personellen Möglichkeiten die Betreuung des VA-Webportals auch über das Projektende von VA hinaus übernehmen wird, jedoch kann nicht ausgeschlossen werden, dass in mittelbis langfristiger Perspektive derart großer Aufwand für den Fortbetrieb des Portals geleistet werden müsste, der die Kapazitäten der ITG übersteigt. Versuchsweise wurde eine ältere Version des VA- Webportals in einem sog. Docker-Image auf einem Server der UB[17] abgelegt, jedoch erscheint auch dies nicht als absolut zuverlässige Dauerlösung. Als derzeit einzig vernünftiges Konzept zur dauerhaften Bewahrung auch des Webportals erscheint nur die von VA betriebene möglichst ausführliche Dokumentation der Funktionalität der Webseite zusammen mit der Archivierung des entwickelten Softwarecodes auf GitHub sowie im institutionellen Repositorium der UB (Letzteres wird derzeit noch projektintern diskutiert). Späteren Generationen sollte es dann zumindest theoretisch möglich sein, das Gesamtsystem einschließlich all seiner Funktionen mit der dann verfügbaren Technik „nachzubauen“.

2. Teil: Die Perspektive der Bibliothek

Wie aufgezeigt spielen Bibliotheken eine entscheidende Rolle im Umgang mit Forschungsdaten. Dabei nimmt die bereits an den Einrichtungen vorhandene Expertise in der Erschließung und Zugänglichmachung von Informationen eine zentrale Rolle im Prozess des Forschungsdatenmanagements ein.

Neben der Erschließung der Forschungsdaten und der Verknüpfung mit Normdaten kann die generische und fachspezifische Anreicherung mit Metadaten als ein Kompetenzbereich der Bibliotheken betrachtet werden. Oft besteht zusätzlich eine entsprechende Infrastruktur an den Einrichtungen, die sich mit der Vergabe von persistenten Identifikatoren (PID) befasst. Bibliotheken können in der Regel auf eine langjährige Erfahrung mit der Vergabe und dem Einsatz von Digital Object Identifiern (DOI) und Uniform Resource Names (URN) zurückgreifen.

Diese Aufgabenbereiche bilden die Grundlage, um Daten zugänglich und auffindbar zu machen. Neben der Beratung zum Forschungsdatenmanagement finden sich Bibliotheken

[17] `https://verba-alpina-archiv.ub.uni-muenchen.de/`

zudem immer häufiger in der Position des „Data Publishers“, also des Datenveröffentlichers. Mit der Veröffentlichung von Forschungsdaten – beispielsweise auf institutionellen Repositorien – entstehen zusätzliche Aufgabenfelder für die Bibliotheken. In der Regel verfügen die Publikationsplattformen bereits über eine Infrastruktur, die es ermöglicht, die Metadaten über Schnittstellen an weitere Suchmaschinen oder Discovery-Systeme zu liefern. Bibliotheken kennen dabei auch die Recherche- und Nutzungs-Bedürfnisse der Forschenden und sorgen dafür, dass auch die Forschungsdaten über geeignete Plattformen für eine breite Nutzergruppe zugänglich sind. Während mit der Vergabe von PIDs bereits eine dauerhafte Zitierbarkeit gegeben ist, müssen sich die „Data Publishers“ auch mit der Frage auseinandersetzen, wie Forschungsdaten langfristig verfügbar bleiben können. An vielen Bibliotheken bestehen bereits entsprechende Workflows für digitale Publikationen, die sich teilweise auch auf Forschungsdaten übertragen lassen. Gemäß den Regeln der guten wissenschaftlichen Praxis sollen Daten mindestens zehn Jahre aufbewahrt werden[18]. Dies erscheint jedoch, verglichen mit „traditionellen“ Beständen von Bibliotheken, sehr kurz. Während beispielsweise im Bereich „Altes Buch“ Medien bewahrt werden, die mehrere hundert Jahre alt sein können, ist es unklar, ob heute erstellte Forschungsdaten in zehn Jahren noch lesbar sind. Es wird daher an verschiedenen Lösungen gearbeitet, von Technologien wie Bitstream Preservation bis hin zur Langzeitarchivierung, damit auch Informationen aus digitalen Daten langfristig verfügbar sind.

Fallbeispiel Universitätsbibliothek der Ludwig-Maximilians-Universität München:

Die Themen Open Access und Forschungsdaten sind an der Universitätsbibliothek der Ludwig-Maximilians-Universität München (UB der LMU) Teil des Alltagsgeschäfts. Neben elektronischen Publikationsplattformen für Zeitschriften (Open Journals LMU), Hochschulschriften (Elektronische Hochschulschriften) und weiteren wissenschaftlichen Publikationen (Open Access LMU), werden auch hybride Publikationsformen (z. B. Open Publishing LMU) angeboten.[19] Die Veröffentlichung von Forschungsdaten ist seit 2010 über das institutionelle Repositorium Open Data LMU möglich.

Primär richtet sich das Repositorium an Wissenschaftler/innen aller Fakultäten der LMU sowie kooperierender Institutionen. Die Ausrichtung ist interdisziplinär, und es wurden bereits Forschungsdaten aus über 15 verschiedenen Fachgebieten veröffentlicht. Nutzer/innen können nach erfolgreicher Registrierung ihre Daten eigenständig auf den Server hochladen. Eine einheitliche Forschungsdaten-Policy wurde bisher nicht eingeführt, es wird aber empfohlen, Daten im Sinne der Budapester Open Access Initiative[20] und der Berliner Erklärung[21] über offenen Zugang zu wissenschaftlichem Wissen der Allgemeinheit zur Verfügung zu stellen.

Das Repositoriums Open Data LMU läuft unter der Open-Source-Software EPrints.[22] An der UB der LMU ist die Version 3.3.15 im Einsatz. EPrints wird in einer Linux-

[18] https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf
[19] https://www.ub.uni-muenchen.de/schreiben/open-access-publizieren/index.html
[20] https://www.budapestopenaccessinitiative.org/
[21] https://openaccess.mpg.de/Berliner-Erklaerung
[22] https://www.eprints.org/

Umgebung aufgesetzt und benötigt Perl sowie eine MySQL-Datenbank. Eine OAI-Schnittstelle erlaubt den Export von Metadaten in vielen Standard-Formaten, darunter DataCite, Dublin Core oder RDF.

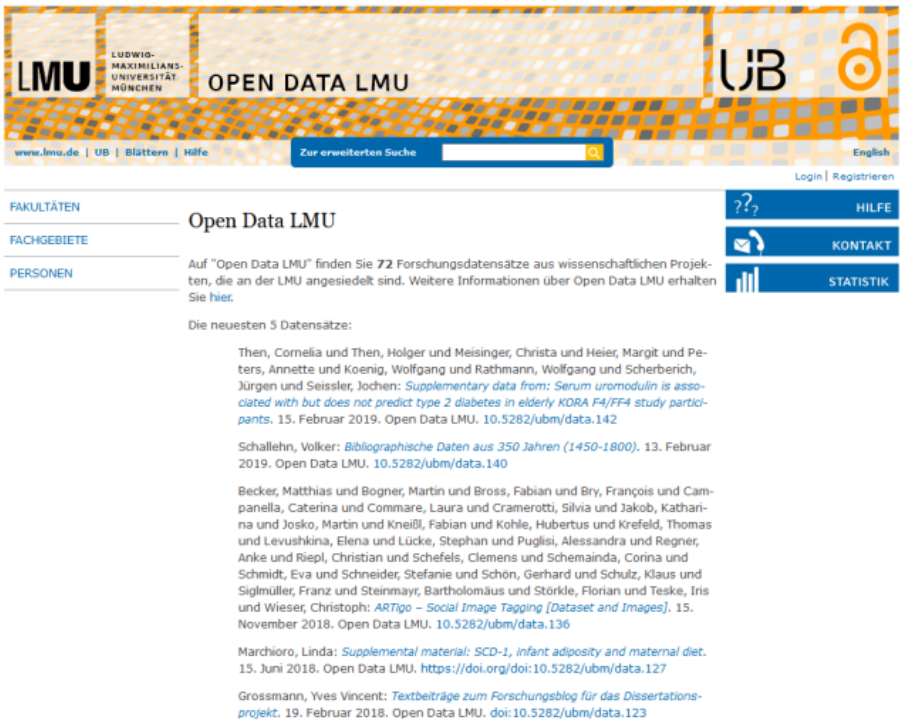


Abbildung 1.: Das institutionelle Repositorium Open Data LMU

Durch die stetig wachsenden Anforderungen an das Forschungsdatenmanagement wurde an der UB der LMU alternative Repositorien-Software evaluiert. Eine Alternative sollte alle Funktionen, die EPrints bietet, beinhalten sowie einen komfortableren und vielseitigeren Umgang mit Forschungsdaten ermöglichen. Zentrale Anforderungen sind beispielsweise eine hohe Skalierbarkeit sowie die Anbindung an neue Technologien wie Linked Open Data und Semantic Web.

Die Wahl fiel schließlich auf Fedora Repositories.[23] Fedora ist ebenfalls ein Open Source-Produkt und wird von DuraSpace entwickelt. Hinter der Organisation steht eine große und aktive Community, die Lösungen für unterschiedliche Bedarfe anbietet. Das populärste Produkt von DuraSpace ist das Repositorium DSpace, welches mittlerweile zu den am häufigsten eingesetzten Datenrepositorien an deutschen Forschungseinrichtungen zählt. Während sich DSpace verhältnismäßig einfach installieren und einrichten lässt, fungiert Fedora als Middleware und bedarf tiefer greifender Entwicklungs- und Programmierarbeiten. Wie genau sich Fedora als Middleware einsetzen lässt, zeigt folgende Skizze, die in der Entwicklungsphase an der UB der LMU erstellt wurde. Als Pilotprojekt diente dabei das in Teil 1 des Textes behandelte Projekt VerbaAlpina (VA). Die Projektwebseite

[23] `https://duraspace.org/fedora/`

hat seit 2019 eine Schnittstelle (API), die Forschungsdaten in unterschiedlichen Formaten bereitstellt, wahlweise als csv, xml oder json.[24]

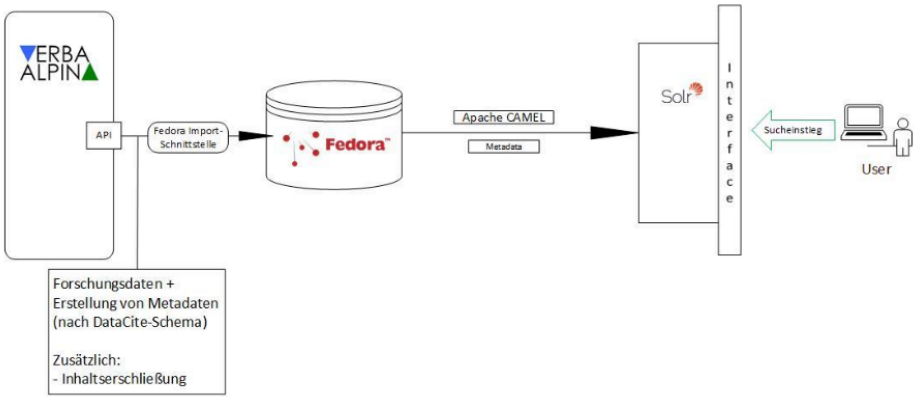


Abbildung 2.: vereinfachte Skizze der Infrastruktur

Über die Schnittstelle gesammelte Daten werden anschließend um Metadaten angereichert. Dabei arbeiten Metadaten-Experten der UB der LMU mit Vertretern aus den Fachdisziplinen zusammen. Im Beispiel von VA entstand in Zusammenarbeit mit den Projektmitarbeiter/innen ein detailliertes Datenmodell, das anschließend in ein geeignetes Metadatenschema übertragen wurde.

Das Metadaten-Management ist die Kernaufgabe im frühen Stadium des Forschungsdatenmanagements und entscheidet darüber, wo und von wem die Daten aufgefunden werden können. Die UB der LMU verwendet dabei den Metadaten-Standard von DataCite[25]. Dieser Standard kann auch zur Registrierung von DOIs genutzt werden. Zusammen mit der ITG und dem Leibniz-Rechenzentrum (LRZ) der Bayerischen Akademie der Wissenschaften wurden in einer Arbeitsgruppe Best-Practice-Empfehlungen für die Verwendung des Metadaten-Schemas erarbeitet.[26] Dies ist insofern sinnvoll, da den Forschenden eine Empfehlung gegeben werden kann, wie das Metadaten-Schema in ihrem Fachbereich am besten zu verwenden ist. Neben der Vergabe von Metadaten spielt bei der Erstellung von DOIs das Thema Granularität eine große Rolle. Je nach Disziplin kann die Granularität stark variieren. Da die Möglichkeit besteht, PIDs für Forschungsdaten zu vergeben, stellt sich unweigerlich die Frage, auf welcher Ebene dies geschehen soll. Im DOI-Handbook werden folgende Möglichkeiten genannt:

> A DOI name can be assigned to any object, regardless of the extent to which that object might be a component part of some larger entity. DOI names can be assigned at any desired degree of precision and granularity that a registrant deems to be appropriate. For example, for granularity in textual materials, separate DOI names can be assigned to a novel as an abstract work, a specific edition of that novel, a specific chapter within that edition of the novel, a

[24] `https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=8844`
[25] `https://schema.datacite.org/meta/kernel-4.3/`
[26] `http://doi.org/10.5281/zenodo.3559800`

> single paragraph, a specific image, or a quotation, as well as to each specific manifestation in which any of those entities are published or otherwise made available.[27]

Im Falle von VA haben die Einzelbelege in der Forschung einen besonderen Status. Deshalb ist im Projektkontext die Überlegung, PIDs auf Einzeldatensatzebene zu vergeben. Bei VA handelt es sich derzeit (April 2019) um ca. 81.000 Datensätze. Damit die Unterscheidung und Suchbarkeit der Einzelbelege sinnvoll gestaltet werden kann, fließen in die Metadaten auch Sacherschließung und geographische Informationen mit ein. Die in Teil 1 erwähnten Verknüpfung der Einzelbelegte mit den Entitäten „Gemeinden“, „morpholexikalische Typen“ und „Konzept“ werden ebenfalls in den Metadaten berücksichtigt.

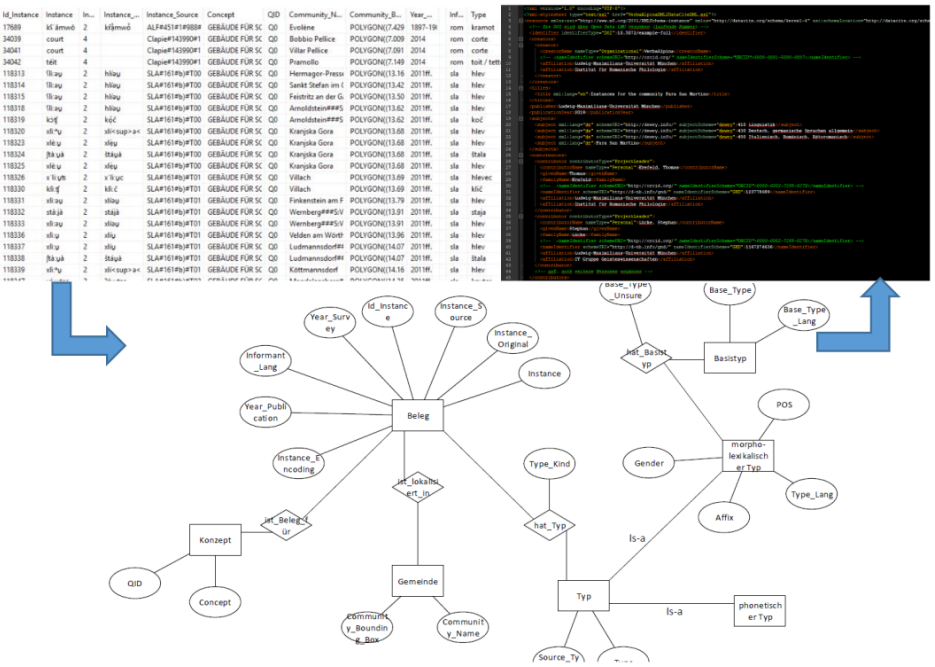


Abbildung 3.: Transformation der Daten: Von den Roh-Daten aus einer CSV-Datei, über das Datenmodell hin zum DataCite-XML(Grafiken erstellt von Sonja Kümmet, UB der LMU)

Die Transformation erfolgt dabei über mehrere Schritte. Anhand eines detaillierten Datenmodells werden die Rohdaten der Schnittstelle bearbeitet und um entsprechende Metadaten angereichert. Dies erfolgt in Abstimmung mit den Forschenden. Anschließend wird eine DataCite-XML-Datei erstellt, die alle relevanten Metadaten beinhaltet. Diese XML-Datei wird von der UB der LMU auch für die Erstellung von DOIs verwendet. Fedora bietet an dieser Stelle verschiedene Möglichkeiten für einen Ingest der Daten. Neben XML besteht auch die Möglichkeit, RDF-Dateien für das Anlegen der Datensätze zu verwenden. Im Testbetrieb wird an der Universitätsbibliothek der Import beider Varianten erprobt.

[27] `https://www.doi.org/doi_handbook/2_Numbering.html#2.3.2`

Der Fluss der Forschungs- und Metadaten ist daher sehr komplex. Ein daraus entwickeltes Aufgabenmodell wirft zudem einige Fragen auf, die im Rahmen des Projekts geklärt werden müssen:

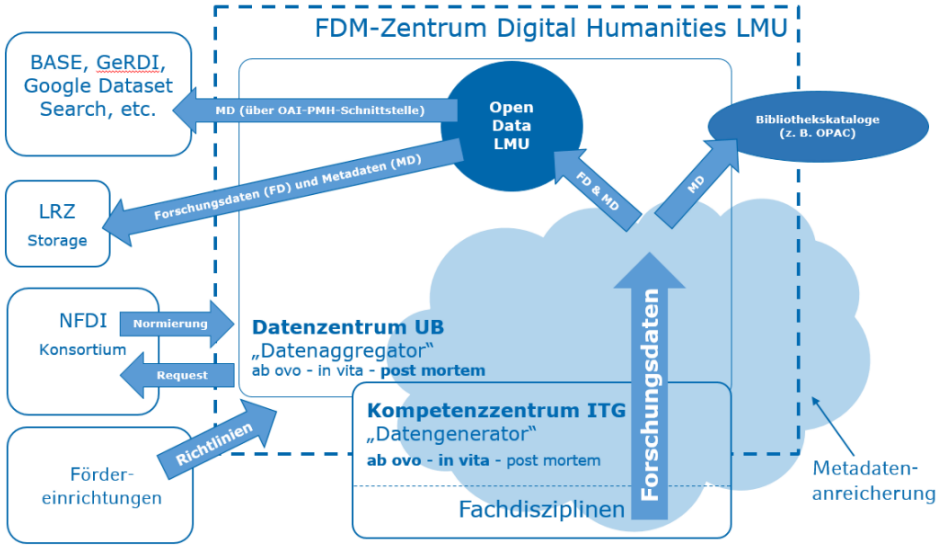


Abbildung 4.: Erweitertes Aufgabenmodell zum FDM in den digitalen Geisteswissenschaften an der LMU

Das erweiterte Aufgabenmodell zeigt, wie die Aufgaben der ITG und der UB der LMU zusammenhängen. Die Anreicherung mit Metadaten erfolgt an der UB der LMU in enger Zusammenarbeit mit der ITG und den Forschenden aus den Fachdisziplinen. Sobald die Forschungsdaten um entsprechende Metadaten angereichert sind, werden Forschungs- und Metadatensätze in das Repositorium Open Data LMU übertragen. An diesem Punkt werden die Forschungsdaten zum ersten Mal über das institutionelle Repositorium durchsuch- und auffindbar. Um die Daten einer breiteren Nutzergruppe zur Verfügung zu stellen, werden über entsprechende Schnittstellen (z. B. OAI-PMH) Metadaten an weitere Dienste geliefert. Daten von Open Data LMU werden momentan von BASE, Google Dataset Search oder GeRDI indexiert.

Zur Verfügbarkeit der Forschungsdaten bietet Open Data LMU eine sogenannte Bitstream Preservation. Die Daten werden gemäß den Regeln der guten wissenschaftlichen Praxis mindestens zehn Jahre aufbewahrt. Mit dem Umstieg auf eine neue Repositorien-Software soll gleichzeitig auch die Strategie der Datenarchivierung überarbeitet werden. Zukünftig soll ein Schwerpunkt nicht nur auf die Archivierung von Forschungs-, sondern auch der zugehörigen Metadaten gelegt werden. Eine sich an den NESTOR-Richtlinien[28] orientierende Langzeitarchivierung könnte dabei eingerichtet werden. Im Falle des Pilotprojekts VA bietet auch die Projektwebseite mit ihrer multimedialen Kartenansicht Informationen, die durch eine reine Ablage der Forschungsdaten verloren gehen könnten. Die UB der LMU arbeitet hier erneut im Tandem mit der ITG und dem VA-Projekt,

[28] `https://www.langzeitarchivierung.de/`

um die Bedürfnisse der Wissenschaft zu erkennen und umzusetzen. In der Vergangenheit gab es des Öfteren Kooperationen mit dem LRZ, die sich mit der Archivierung von Daten beschäftigt haben. Auch für zukünftige Projekte soll in diesem Bereich kooperiert werden.

Für die langfristige Planung wird ebenfalls erwogen, Forschungsdaten über bestehende Recherche-Systeme an Bibliotheken auffindbar zu machen. Dies könnte über Bibliothekskataloge erfolgen. Dabei ist angedacht, analog zu den oben genannten Discovery-Diensten nur die Metadaten weiterzugeben. Sobald die Infrastruktur um das Fedora-Repositorium zuverlässig im Produktivbetrieb läuft, wird diese Möglichkeit evaluiert. Die Zusammenarbeit zwischen UB und Wissenschaftler/innen der LMU hat gezeigt, dass es von großem Vorteil sein kann, auf bestehende Strukturen aufzubauen. Sowohl die Universitätsbibliothek, als auch die geisteswissenschaftlichen Institute sind ein fester Bestandteil der Universität und können auch längerfristige Vorhaben besser umsetzen, als beispielsweise Projekte mit begrenzter Laufzeit. Die Zusammenarbeit hat zudem auf beiden Seiten die Kompetenz der Forschungsdaten-Ansprechpartner/innen vergrößert. Dabei spielt auch die Beratung auf verschiedenen Stufen des Forschungsprozesses eine zentrale Rolle, die sowohl von Infrastruktur-, als auch von Wissenschaftspartnern übernommen werden kann. Wenn die Universitätsbibliothek schon frühzeitig in die Planung und Durchführung von Forschungsdatenvorhaben involviert wird, erzeugt dies Synergien, die ein erfolgreiches Vorhaben leichter realisierbar machen. ITG und UB der LMU bilden bereits de facto ein institutionelles Kompetenz-und Datenzentrum für Forschungsprojekte in den Digitalen Geisteswissenschaften an der LMU. Die Zusammenarbeit kann perspektivisch neue Kooperationen und Finanzierungsmöglichkeiten entstehen lassen.

Das Modellprojekt „eHumanities – interdisziplinär“:

Die in Teil 1 und Teil 2 genannten Infrastruktur- und Wissenschaftspartner arbeiten nicht nur innerhalb der LMU zusammen, sondern sind auch Teil des Projekts: „eHumanities – interdisziplinär“. Dort arbeiten ITG und die UB der LMU unter der Federführung der UB der Friedrich-Alexander-Universität Erlangen-Nürnberg (UB der FAU) an Fragestellungen zum Management von Forschungsdaten in den digitalen Geistes- und Sozialwissenschaften. Das Projekt wird vom Bayerischen Staatsministerium für Wissenschaft und Kunst gefördert und hat eine Laufzeit von drei Jahren (März 2018 – März 2021).

Im Projekt beschäftigen sich die Mitarbeiter/innen mit der Konzeption und Evaluierung neuer Hilfsmittel und der Erarbeitung von Best-Practice-Empfehlungen. Die Verbindung von digitaler Bibliotheksexpertise mit informatischen und fachmethodischen Schnittstellenkompetenzen steht dabei im Vordergrund.

Projektergebnisse werden über die Projektwebseite[29] bekanntgegeben. Der Zwischenbericht über das erste Projektjahr[30] wurde bereits veröffentlicht und zukünftige Berichte und Ergebnisse werden ebenfalls mit der Community geteilt. Ziele sind dabei auch, Erfahrungsberichte zu Fedora zu veröffentlichen und Programmierarbeiten über die Plattform GitHub zur Verfügung zu stellen.

[29] `https://www.fdm-bayern.org`
[30] `https://zenodo.org/record/2645935`

Durch die Veröffentlichung von Quellcode in Kombination mit den Berichten soll eine Transferierbarkeit der Projektergebnisse auf weitere Vorhaben und/oder Disziplinen möglich sein.

Bereits nach einem Drittel der Projektlaufzeit hat sich gezeigt, dass die Zusammenarbeit von Infrastruktur- und Wissenschaftspartnern neue Sichtweisen auf altbekannte Probleme ermöglicht, wie z. B. die Fragen nach der Langzeitarchivierung, Granularität und Auffindbarkeit. Durch eine Kooperation wird zudem vermieden, dass Bibliotheken und Wissenschaftler/innen Insellösungen für ihre Projekte aufbauen. Die Zusammenarbeit sorgt dafür, dass sich an der LMU aus der Erfahrung mit Pilotprojekten der digitalen Geisteswissenschaften feste Workflows entwickeln und möglicherweise schließlich etablieren lassen, von denen alle Teilnehmenden profitieren und mit deren Hilfe wiederkehrende Themen schneller bearbeitet werden können.

Die Bereitstellung von Datenmodellen und das Erarbeiten von Best-Practice-Empfehlungen für Metadatenschemata können zudem in Datenmanagementpläne übernommen werden. Dabei handelt es sich um ein weiteres Arbeitspaket des Modellprojekts, mit dem sich die UB der FAU beschäftigt. Dort wurde im Frühjahr 2019 eine RDMO-Instanz eingeführt, die anschließend auch auf Server der UB der LMU übertragen werden soll. Um die Wissenschaftler/innen für das Thema „Forschungsdaten“ zu sensibilisieren, wird an der UB der FAU in einem weiteren Arbeitspaket ein digitales Lern- und Informationsangebot erstellt. Die Schulungs- und Videomaterialien werden anschließend als Open Educational Resources auch anderen Interessenten zur Verfügung gestellt und in den Digitalen Campus Bayern integriert. Dort können sie beispielsweise zum Bestandteil von Curricula im Bereich der Digital Humanities werden. Mit leichten Veränderungen sind diese Materialien auch auf andere Institutionen und verwandte Disziplinen transferierbar.

Adapting Established Software Engineering Techniques and Technologies into an Assistance System for Neuroscientists

Thorsten Arendt and Alexander C. Schütz
Department of Psychology, Philipps-Universität Marburg, Germany

In the information infrastructure project NOWA (NeurOscientific Workflow Assistance) of the collaborative research center CRC/TRR 135 *Cardinal mechanisms of perception: Prediction, Valuation, Categorization* we develop and combine tools for the sharing of research data. In this context, NOWA aims to create an organizational and technological framework that supports workflows throughout the entire research data lifecycle. This article addresses the conceptual and technical part of NOWA. We first discuss the specific requirements of the scientists and then compare them with proven solutions from the computer science and software engineering domain, respectively. These so-called best practices are then adapted to the specific needs within the CRC. Finally, we conclude the article by an overview of the planned technical and technological implementation of the NOWA assistance system.

1. NeurOscientific Workflow Assistance (NOWA)

The sensory organs are the "window to the world" since they enable the sensation of signals from the environment. But how does the brain process this information? How does perception work? The DFG funded collaborative research center CRC/TRR 135 *Cardinal Mechanisms of Perception: Prediction, Evaluation, Categorization* deals particularly with these questions. More precisely, twenty interdisciplinary working groups at the Justus-Liebig-Universität Gießen (JLU) and the Philipps-Universität Marburg (UMR) use a combination of behavioral experiments, physiological measurements and computational modeling to gain a comprehensive understanding of prediction, evaluation and categorization. The goal is to delineate these cardinal mechanisms behaviorally, to identify their underlying neural substrates and to explain their functions with computational models.

Already during the first funding phase of the CRC, the researchers increasingly pursued the implementation of the Open Science principles. In addition to the publication of research results in peer-reviewed articles, the researchers published relevant research data (Open Access) using the publicly accessible repository Zenodo[1] (Open Data). Doing this, they increased traceability and reproducibility of their results and so the overall transparency of their research. However, the high effort when selecting and packaging the research data and creating the meta data revealed the need for a more fundamental

[1] `https://zenodo.org/`

Thorsten Arendt und Alexander C. Schütz: Adapting Established Software Engineering Techniques and Technologies into an Assistance System for Neuroscientists. In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge, Heidelberg: heiBOOKS, 2020, 32–46. DOI: https://doi.org/10.11588/heibooks.598.c8415

approach. For the second funding phase, this led to the information infrastructure project NOWA (NeurOscientific Workflow Assistance) [2], which is in the focus of this article.

In NOWA we develop new and combine existing tools for the sharing of research data. In order to enhance reproducibility of all steps in the lifecycle of a scientific study – from planning experiments over collecting and analysing data to publishing them– this information infrastructure project aims at creating an organizational and technological framework, supporting workflows along the entire data lifecycle. The long-term goal is to share research data within both, the single working groups and the entire CRC, right from the beginning of the study and to make them publicly accessible at the "press of a button" at the time the results are published. NOWA is implemented in close cooperation with the researchers of the CRC and the local infrastructure facilities including the joint project HeFDI (Hessische Forschungsdateninfrastrukturen).[2] The University Computer Center of the UMR thereby provides the technical infrastructure.

This article particularly addresses the conceptual and technical part of NOWA. We first discuss the specific requirements of the scientists –such as collaborative working, management of large files, versioning of data and code, quality assurance at data and code level, meta data management, publishing research as well as automating these steps as effectively as possible – and then compare them with established solutions from the computer science and software engineering domain, respectively (e.g., version control using git and GitLab, continuous integration/testing, style guides, and packaging). These so-called best practices are then adapted to the specific needs within the CRC. Finally, in the conclusion of the article we provide an overview of the planned technical and technological implementation of the NOWA assistance system.

2. What Neuroscientists need

A neuroscientific study in the CRC involves typically a sequence of consecutive steps from designing experiments to collecting data and analysing data (see Figure 1). Since each study usually consists of several sequential experiments that are building-up on each other, this cycle is traversed several times. During each of those steps, different types of research data are produced: Meta data include information about the design of the experiment, the settings of technical equipment and software, personal data about the tested participants or information about the type of data transformation and analysis. Software consists of scripts for experiments and for data analysis as well as for computational modelling. Primary data refer to the immediate outputs of the experiments, while processed data refers to analysed and aggregated data and statistical results.

2.1. Sharing and Publishing Research Data

One objective of the CRC is to facilitate the collaboration of researchers within and beyond the CRC. Sharing data with researchers in and outside of the CRC poses additional requirements to research data management, which are specified below.

[2] https://www.uni-marburg.de/de/forschung/kontakt/forschungsdatenmanagement/projekte/hefdi-hessische-forschungsdateninfrastrukturen

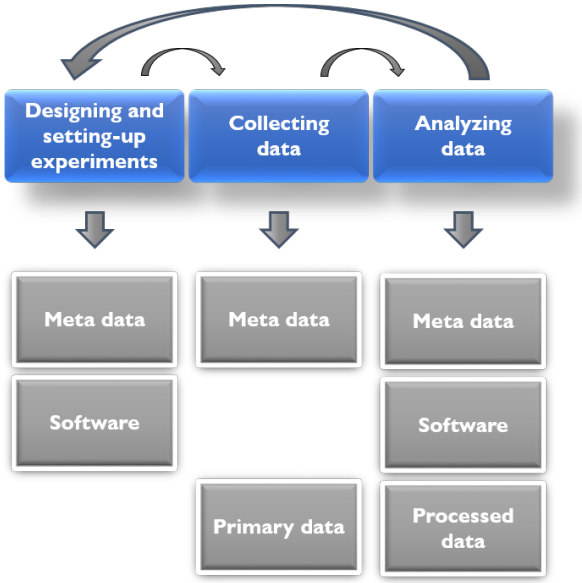


Figure 1.: The research process and where research data occur

Private and collaborative work Sharing data with other researchers requires comprehensive meta data regarding all aspects of a study, including the experimental design, the collection of primary data and the analysis of processed data, such that all involved researchers can comprehend the study material. The sharing of research data can be accomplished more effectively if those requirements are met from the beginning of the study. Working collaboratively on a study also requires a user management that allows to specify access privileges for different researchers and that keeps track of which researcher is working on the material. Of course, collaborative work also requires an easy exchange of research data and (semi-)automatic comparison of files and detection of modifications.

Publication Ultimately, all research data of the CRC should be made publicly available within the constraints of ethical regulations. To publish the data of a single study, this requires that all research data of the study are bundled and that each entity is tagged according to whether it can be made public or not. A direct export filter to existing research data repositories could facilitate the publication of data, such that only a single button-press is required.

2.2. Managing Research Data

An efficient management of research data is an important tool to guarantee the reproducibility of research. In the following, we highlight the most important requirements of research data management in the CRC.

Meta data, primary data, and code As mentioned before, different types of research data are produced during the life-cycle of a study. Several objectives need to be met by efficient research data management: For instance, different types of research data

need to conform with each other and redundancies in the data should be eliminated. In addition, the work flow should be optimized such that the additional work load imposed by research data management is minimized. The efficient reuse of materials, for instance analysis scripts, in other projects should be facilitated.

Data versioning A study in the CRC progresses usually through several stages, from an initial test phase, the presentation of first results at a conference to the publication of the final results in a scientific journal and the final data in a data repository. Usually, all research data evolve during this life cycle of a study. To make the research progress transparent and reproducible, it is important to keep track of all changes and to be able to return to any previous stage of the study and the associated materials. This goal cannot be achieved without a continuous versioning of all data that belong to the study. Ideally, versioning of all study-related material automatically generates a digital lab book that allows a complete overview of the work progress in the study.

Quality of data A further objective of NOWA is to maintain a high quality of research data to ensure the reproducibility of the research and results of the CRC. Several issues can arise during a study that will ultimately compromise the quality of research data: Meta data, for instance, might be incomplete because the necessary documentation is not carried out immediately. Primary data might be incomplete or might contain invalid data due to technical or organizational problems during data collection. Software might contain unknown limitations or bugs due to insufficient testing. An active research data management should include automated checks for data quality and highlight potential issues early on when they can be corrected easily.

2.3. Overview

To summarize (see Figure 2), researchers in the CRC need an efficient research data management spanning the whole data life-cycle from the design of the experiments over data collection and data analysis to the publication of data. The research data management should include all types of research data, such as meta data, primary and processed data and software. Primary goals of the research data management are to ensure a high quality of the data, to track different versions of data and to facilitate the collaboration between different researchers and working groups.

3. What Computer Science and Software Engineering provides

In this section, we discuss some of the challenges that have been identified in the domains of computer science and software engineering over the last 40 years, as well as solutions developed for them. The objective is to discover similarities to the needs of the researchers presented in the previous section in order to adapt the best practices from the domains of computer science and software engineering to the domain of neuroscience.

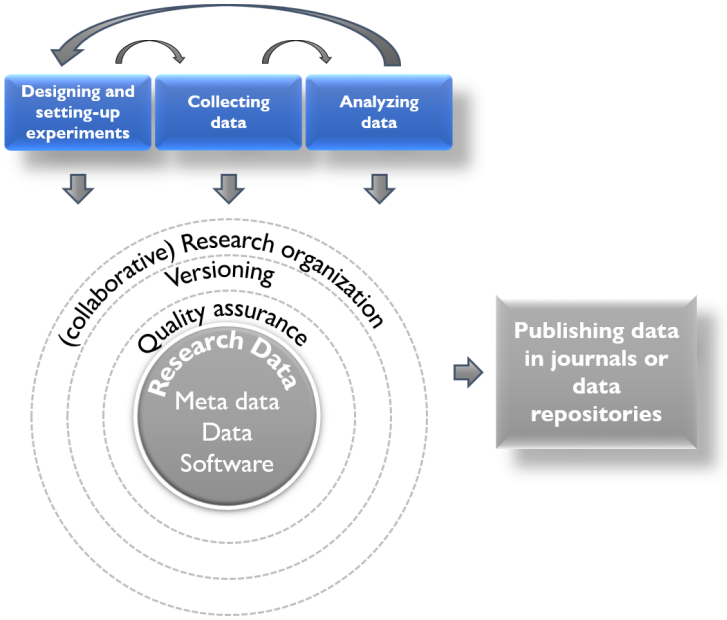


Figure 2.: Overview on what CRC/TRR 135 neuroscientists need

3.1. Collaboration and Versioning

Especially in recent decades, software systems have become ever larger and more complex. Consequently, the software engineers were confronted with a number of challenges in the area of so-called configuration management. For example, because large and complex software systems consist of a large number of collaborating components, the dependencies between these components must be managed. As these components evolve over time, each with its own extent and speed, the versions of the components and the entire system must be stored properly. Furthermore, old versions of files and the entire project should be easily and effectively recoverable, for example if the changes that are made lately prove to be insufficient. Additionally, for maintainability of the system, it would be very helpful to be able to easily understand how the components evolved between versions. Finally, it would be beneficial to have some sort of implicit backup system to counter the risks of a single point of failure when the code is managed on a single system. However, the core of these specific challenges is more general. Large and complex software systems are being developed by a large number of project members, many of whom are spread over different locations around the world. Therefore, the core of the challenges is to enable collaborative and distributed work within the software development project.

In order to meet the challenges in the field of configuration management, the computer science community developed so-called Version Control Systems (VCS) [14]. Such a system tracks changes to project (a set of files) and for each change, it records the date and time, the person who made the change, and the differences between the file contents before and after the change. As a result, the system can re-create a consistent snapshot of the project at any point in time. Two different approaches to version control systems have evolved over time. In the centralized approach, the VCS consists of a server that stores the only master copy of the entire project. Examples for centralized VCSs are Concurrent

Versions System (CVS)[3] and Subversion (SVN)[4]. However, since the centralized approach depends on a server, working on the project is only possible if there is a network connection. Furthermore, this approach carries the risk of a single point of failure. In contrast, the distributed approach generally does not have a server. Each user has a full copy of the whole project with the full history on the computer. Popular distributed VCSs are Mercurial[5] and git [6]. The distributed approach represents the state of the art.

In terms of supporting collaborative and distributed work within a software development project, a number of useful solutions have been established: groupware, computer-supported collaborative work, CASE tools, collaboration platforms, etc. [13, 9]. All of these solutions share the core features co-ordination, collaboration, and community building –with communication as cross-cutting function. Furthermore, these systems can be classified into their spatial and temporal dimensions. The spatial (or geographic) distribution in the software development process distinguishes spatially close co-located and distributed organizational units. In the temporal distribution, the focus is on whether the cooperation, in particular the communication, takes place with a time delay. Here, information and software artifacts can be either IT-buffered (asynchronous) or concurrently (synchronously) exchanged respectively edited. A combination of the two aforementioned concepts represent web-based collaboration platforms having a VCS as core technology. Here, state-of-the-art platforms are GitHub[7], GitLab[8] and Bitbucket[9].

3.2. Data, Meta Data and Software

Digital data has been stored in files of various formats for decades. However, this syntactic separation is often in contrast to the semantic contexts within the data. In order to summarize data semantically, applications such as ZIP and RAR were developed early on. They package a set of heterogeneous files into an archive and simultaneously compress their contents. This is similar with software systems. Even small applications consist of a large number of components that provide the required functionality only in interaction and depend on each other to a different extent. Appropriately developed software packagers [6, 8] carry out the combination of these components –in the corresponding versions– into an installable and executable application.

In most cases, data has a content-related structure. Uniformly structured and logically related data volumes are managed in databases. The essential task of a database is to store large volumes of data efficiently, without contradictions and permanently and to provide the necessary subsets in different, needs-based forms of presentation for users and application programs. At the next level, a data repository is an infrastructure of databases that collect, manage and store varying data sets. In fact, a data repository can be seen

[3] `https://savannah.nongnu.org/projects/cvs`
[4] `https://subversion.apache.org/`
[5] `https://www.mercurial-scm.org/`
[6] `https://git-scm.com/`
[7] `https://github.com/`
[8] `https://gitlab.com/`
[9] `https://bitbucket.org`

as a general term that comprises several concrete ways to collect and store data[10]. Such a repository can be implemented in one of the following ways. Data warehouses are large data repositories that aggregate data from multiple sources or segments of a business, without the data being necessarily related [7]. Data lakes are large data repositories that store unstructured data that is classified and tagged with meta data. An example can be found at [12]. Finally, data cubes are lists of data with three or more dimensions stored as a table [11]. The common structure of data can be specified using so-called meta data. Meta data are often described as information about data, i.e., the information required to understand data, including data set contents, context, quality, structure, and accessibility. The meta data explains where the data originated, how it was captured, and what it represents. In most instances, meta data are descriptions of things, whether physical objects (books, items in a warehouse) or information objects (spreadsheets, web pages, publications, etc.). Meta data repositories store data about data and databases. An example of a meta data repository for technical information in digital medical images can be found at [15].

Today, information and knowledge in the form of data and programs are widely published on the World Wide Web. Often, this information is in a packetized and archived format, as described above. However, even at this higher level, there are relationships between data (sets) stored at different locations in the internet. In order to make these relationships more explicit, this information can be coupled through the concepts of Linked Data [5] and the Resource Description Framework (RDF)[11], respectively.

3.3 Software Quality

Software of inferior quality is hardly accepted by the users. In extreme cases, this can mean that the software is rarely or not used at all. As a consequence, achieving a high quality is one of the major challenges in software development. But what does software quality exactly mean? To explicitly describe the term software quality, several so-called quality models for software products gradually evolved over the last 50 years. All of these quality models have in common that they describe certain quality aspects that own a software. These quality aspects are structured using a tree-based hierarchy to form the model. For example, the ISO 9126 standard[12] specifies six independent, high-level quality characteristics that are further sub-divided into 21 sub quality criteria. Following this standard, a software should meet the intended needs, perform well under stated conditions, and provide appropriate performance relative to given resources. Moreover, it should be understandable and easy to use, be modifiable with minimal effort, and transferable to another environment. Altogether, the need to develop high quality software resulted in establishing a software quality assurance process that ensures that the developed software meets and complies with defined or standard quality specifications.

The evaluation of software quality can be effectively automated using the state-of-the-art continuous testing & integration approach [18]. This approach works as follows. When

[10] `https://www.cbronline.com/opinion/what-are-data-repositories`
[11] `http://www.w3.org/TR/rdf-concepts/`
[12] `https://www.iso.org/standard/22749.html`

developer check in their changes to the code repository the VCS triggers a continuous integration (CI) process on a dedicated server. This server builds the software, runs several specified (static and dynamic) tests and analyses, returns the results to the developer, and potentially releases the new version of the software. In modern version control platforms like GitLab the CI server is integrated and and represents a central component for the assurance of the software's quality.

Already in the early years of software engineering, the developers noticed that they re-implemented many similar and even identical algorithms repeatedly. This led to the notion of software reuse representing the process of creating software systems from existing software rather than building software systems from scratch. Especially the upcoming paradigm of object-oriented software development forwarded and empowered the software reuse approach. For enabling software reuse, appropriate code is outsourced into so-called libraries, which can then be integrated into the actual software system to be implemented. Such software libraries can be provided either internally within the software company only or externally, e.g., as open source. A survey on software reuse libraries can be found in [19], for example.

Coding does not enforce problems as long as the code is syntactically correct. However, software code also has other quality characteristics than just syntactic correctness. Here, the use of standards and guidelines –e.g., for the Java programming language[13]– with a consistent style helps to develop high-quality code. A consistent style improves the readability, and therefore, maintainability of code. Furthermore, it facilitates sharing of code among different programmers, especially teams of programmers working on the same project. Finally, it saves development time, once the guidelines are learned, by allowing programmers to focus on the semantics of the code, rather than spend time trying to determine what particular format is appropriate for a given situation. Compliance with these guidelines can be verified by appropriate tools, e.g., during the CI process. In 1998, Kent Beck and Martin Fowler developed the concept of code smells [4] which has been further adapted to other software artifacts such as software models [1]. Code smells represent suspicious parts that are potential candidates for improvements, i.e., they are not synonyms for problems but are worthy of an inspection. In recent years, also approaches to use machine learning techniques to detect code smells have been developed (see [10] for an overview). Finally, the concept of quality assurance has been also adapted to assessing the quality of meta data. A comprehensive overview in this field can be found at [20].

3.4 Adaptation

In this section, we compare the needs of the scientists within the CRC with those of the computer science and software engineering domain. In doing this, we reuse solutions and best practices developed to the challenges of software engineering and adapt them to the domain of neuroscience. These adaptations will then provide a roadmap for the current and future work in building up the NOWA system.

As can be easily seen, the requirements of researchers and those of computer scientists are similar in many ways. In the remainder of the section, we discuss the common

[13] `https://google.github.io/styleguide/javaguide.html`

requirements in the three categories of data management, collaborative work, and quality assurance of artifacts. In the category data management, the correspondence is quite obvious. Both domains deal with data, meta data, and software code. The same holds for the category collaborative work. Members of the team want to share their artifacts with other people. Furthermore, researchers might ask the following questions similar to those in configuration management:

- Which version of our research data is part of a particular publication?
- How did our analysis script evolve during the data analysis, and why?
- Who made the last change in the current version of our research paper, and which one?

Finally, there is also a large overlap in the category quality assurance. The main goal of researchers is that their research is correct and trustworthy. This implies that research results are reproducible through the information provided about the methods used and the associated research data. As a consequence, the research data must be of high quality. For example, the data must match the corresponding meta data, or the previously mentioned analysis scripts must be implemented correctly. Figure 3 shows which best practices from the software engineering domain can be adapted and applied to which challenges in the CRC. Table 1 summarizes, in a before and after comparison, how the application of these best practices will impact the current research workflows in the CRC – again along the above categories and the common goal of simultaneously publishing research results and research data.

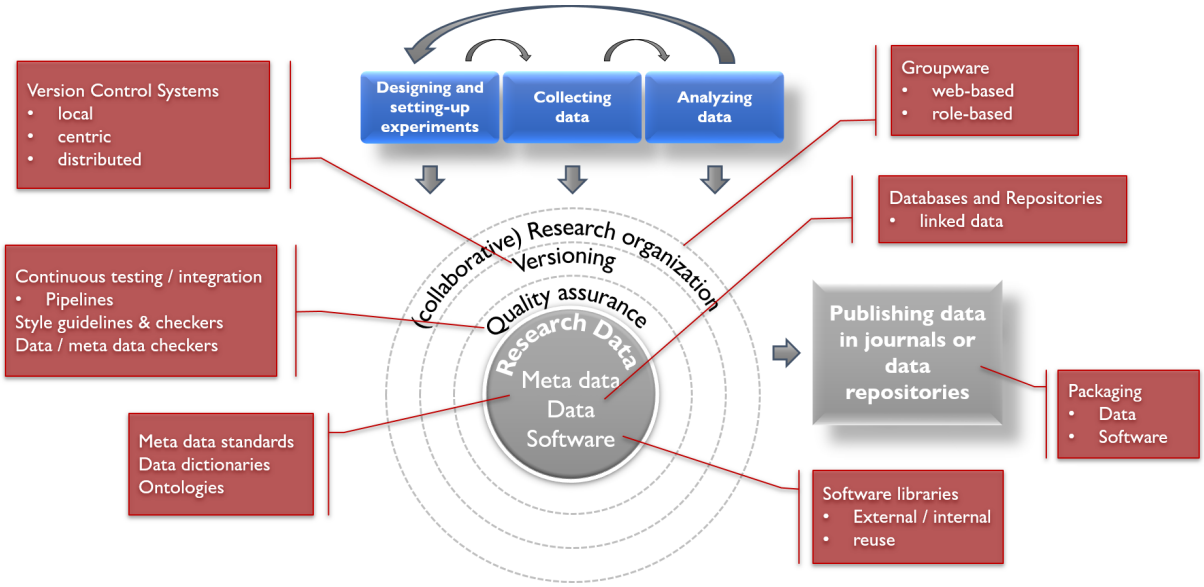


Figure 3.: Adaptation of best practices from the domains of computer science and software engineering to the needs of the researchers within the CRC

With regard to research data management, concrete data can be stored in databases in special repositories with a given structure. The data can then be linked with each other

if necessary. This reduces the current high number of less-structured files and introduces more unified, clearer and more basic file structures. Meta data can be centrally managed and stored with the help of standards, data dictionaries or ontologies[14]. As a result, the previously rather implicit or multi-managed meta data become explicit and through the standardization clearer and more comprehensible. With regard to the research software to be created, more software libraries can be used. These include external standard libraries, but also libraries built internally in the working groups or in the entire CRC. Thus, the more error-prone and time-consuming copy-paste programming can be reduced and the software can be created faster and more comprehensible.

With regard to collaboration, we can adapt a combination of the concepts groupware and version control systems. More specifically, we will use a local instance of the version control collaboration platform GitLab as the basis for the NOWA assistance system. By using GitLab, documents can be shared in a central yet distributed environment. The complexity that has resulted from the previous use of different methods for file exchange (e.g., via e-mail or file hosting services) is thereby greatly reduced. Furthermore, data and documents are managed more clearly, since the different file names used so far for the versioning are omitted.

With regard to the quality assurance of research artifacts, various methods and techniques can be adopted from software development, in particular, of course, for the creation of the necessary research software. The so far only fundamentally existing quality considerations such as syntactic correctness checks are supplemented by more sophisticated methods such as the use of style guides. However, the effort to check the artifacts for various new quality assurance techniques must be kept to a minimum –after all, the actual research is still the focus of the researchers. This can be done in particular by using the continuous integration technology, which is already integrated in the version control collaboration platform GitLab.

Finally, with regard to the joint publication of research results and data, the methods of packaging used in software engineering, both for data and for software, can be adapted. The previous complexity in the challenge of compiling the relevant research data can be significantly reduced by using the tagging feature provided by the VCS GitLab. Combined with the use of a structured quality assurance process started early in the research workflow, researchers within the CRC can ensure the overall reproducibility of their results.

[14] These best practices were not covered in this article.

	Before	After
Research Data	• High number of less-structured files • Implicitly given meta data • Copy-Paste programming	• Unified file structures • Explicitly given, standardized meta data • Software reuse
Collaboration	• Document exchange via Email or file hosting services • File versioning by different names with different nomenclatures	• VCS server with central and distributed repositories • Versioned files with permanent names
Quality Assurance	• Basic quality (syntactical checks only, redundancy, etc.) • Nearly no automation	• High quality (conformity, comprehensibility, reusability, etc.) • CI pipeline within VCS
Publication	• Complex (time-consuming, error-prone, etc.) • Low utility (data reuse difficult)	• Effortless due to using tagged data versions • High utility (data reuse easy)

Table 1.: Before/after comparison when adapting best practices from software engineering

4. Current and Future Work

As already mentioned in the previous section, we will set up a local instance of the GitLab web application at the UMR University Computer Center that is based on the distributed version control system git. This will represent the core of the NOWA workflow assistance system. By using the possibilities given by GitLab, a continuous and distinctive versioning of the active research data in the CRC projects will be supported in the best possible way. In particular, the Git LFS (Large File Storage)[15] extension optimally manages large files such as fMRI data. Additionally, using GitLab will facilitate improved collaborative work within the interdisciplinary working groups of the CRC. For this purpose, we will prepare and integrate specific security and authorization concepts. Parallel to the set-up of the assistance system, we analyze the research workflows within the CRC in close collaboration with the researchers involved. These workflows will be consecutively optimized by the possibilities provided by GitLab. Together with good practices for the general use of GitLab we will pass on the improved workflows to the researchers in regular training courses.

Since GitLab also provides a continuous integration and testing subsystem, our workflow assistance system will provide measures to ensure the quality of research data on top of this. On the one hand, we integrate existing quality assurance techniques for the software code that is currently used within the CRC research processes. On the other hand, we plan to develop further –possibly project-specific– quality assurance techniques together with the researchers which will then also be integrated into the assistance system. Moreover, NOWA will provide a collection of reusable libraries for research software such as data analysis scripts. The basis for this can be, for example, established libraries for the languages Python or MATLAB as provided in [17, 21] or [16, 22], respectively. These are then successively extended by own implementations. NOWA will initially make these libraries available to the researchers of the CRC and, ultimately, in a public repository of the GitLab instance.

In addition to the opportunities provided by GitLab, the workflows of the researchers will be supported by additional services offered in the infrastructure facilities of the participating universities. On the one hand, the collaborative work can be supplemented by the Sync&Share solution HessenBox, which is currently being tested. On the other hand, active research data management can be optimally supported by the use of HeRDMO, the HeFDI-administered prototypical installation of the research data management solution RDMO (Research Data Management Organizer)[16] also funded by the DFG. This tool captures all relevant planning information in data management plans and manages all data management tasks throughout the entire research data lifecycle. In passive research data management, the publication of research results and associated research data can then finally be achieved via the existing publication servers in conjunction with the data repository DSpace[17] currently under test at the UMR computer center.

The mere publication of research results and data of any kind within the World Wide

[15] `https://git-lfs.github.com/`
[16] `https://rdm1organiser.github.io/`
[17] `https://duraspace.org/dspace/`

Web together with their linking in a standardized format often does not guarantee the reproducibility of the underlying research. Current approaches and efforts to overcome these challenges include, for example, the concept of so-called research objects [3][18] as well as the publication of all relevant information in a reproducible article of the eLife Science Magazine[19]. In the future, we will also pursue and integrate these approaches in our NOWA assistance system.

Acknowledgements

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 222641018 – SFB/TRR 135 TP INF.

Bibliography

[1] Thorsten Arendt. *Quality Assurance of Software Models - A Structured Quality Assurance Process Supported by a Flexible Tool Environment in the Eclipse Modeling Project*. Doctoral thesis, Philipps-University Marburg, pp. 1–385 (2014). DOI: 10.17192/z2014.0357

[2] Thorsten Arendt, Ortrun Brand, Christian Krippes, Andreas Gabriel, Matteo Valsecchi, Clemens Helf, Karl R. Gegenfurtner, and Alexander C. Schütz. *NeurOscientific Workflow Assistance (NOWA)*, Poster presented at DINI Jahrestagung 2018, Bielefeld, Germany (2018). DOI: 10.17192/es2019.0002

[3] Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, and Carole Goble. *Why linked data is not enough for scientists*, In: Future Generation Computer Systems, vol. 29, pp. 599 – 611 (2013). DOI: 10.1016/j.future.2011.08.004

[4] Kent Beck and Martin Fowler. *Bad Smells in Code*. In: Refactoring: Improving the Design of Existing Code, 2nd Edition, Addison-Wesley Professional (2018)

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. *Linked Data: The Story so Far*, In: Semantic Services, Interoperability and Web Applications: Emerging Concepts, ed. Amit Sheth, pp. 205–227 (2011). DOI:10.4018/978-1-60960-593-3.ch008

[6] Alexandre Decan, Tom Mens, and Philippe Grosjean. *An empirical comparison of dependency network evolution in seven software packaging ecosystems*, In: Empirical Software Engineering, vol. 24, pp. 381–416 (2019). DOI: 10.1007/s10664-017-9589-y

[7] Barry Devlin and Lynne Doran Cote. *Data Warehouse: From Architecture to Implementation*, Addison-Wesley Longman Publishing Co., Inc. (1996).

[18] `http://www.researchobject.org/`
[19] `https://elifesciences.org/`

[8] Shouki A. Ebad and Moataz Ahmed. *Software Packaging Approaches – A Comparison Framework*, In: Software Architecture, Springer Berlin Heidelberg, LNCS, vol. 6903, pp. 438–446 (2011)

[9] Clarence A. Ellis, Simon J. Gibbs, and Gail Rein. *Groupware: Some Issues and Experiences*. In: Communications of the ACM, vol. 34, pp. 39–58 (1991). DOI: 10.1145/99977.99987

[10] Francesca Arcelli Fontana, Mika V. Mäntylä, Marco Zanoni, and Alessandro Marino. *Comparing and experimenting machine learning techniques for code smell detection*. In: Empirical Software Engineering, vol. 21, pp. 1143–1191 (2016). DOI: 10.1007/s10664-015-9378-4

[11] Steven Geffner, Divakant Agrawal, and Amr El Abbadi. *The Dynamic Data Cube*, In: Advances in Database Technology – EDBT 2000, Springer Berlin Heidelberg, LNCS, vol. 1777, pp. 237–253 (2000)

[12] Rihan Hai, Sandra Geisler, and Christoph Quix. *Constance: An Intelligent Data Lake System*, In: Proceedings of the 2016 International Conference on Management of Data (SIGMOD), ACM, pp. 2097–2100, (2016). DOI: 10.1145/2882903.2899389

[13] Tobias Hildenbrand, Franz Rothlauf, and Armin Heinzl. *Ansätze zur kollaborativen Softwareerstellung*. Arbeitspapier, Universitätsbibliothek Mannheim (2006)

[14] Konrad Hinsen, Konstantin Läufer, and George K. Thiruvathukal. *Essential Tools: Version Control Systems*, In: Computing in Science and Engineering, vol. 11, pp. 84–91(2009). DOI:10.1109/MCSE.2009.194

[15] Hans-Erik Källman, Erik Halsius, Magnus Olsson, and Mats Stenström. *DICOM Metadata repository for technical information in digital medical images*. In: Acta Oncologica, vol. 48, pp. 285–288, Taylor & Francis (2009). DOI: 10.1080/02841860802258786

[16] Mario Kleiner, David H. Brainard, and Denis Pelli. *What's new in Psychtoolbox-3*. In: Perception, 36 (2007). DOI: 10.1068/v070821.

[17] Florian Krause and Oliver Lindemann. *Expyriment: A Python library for cognitive and neuroscientific experiments*. In: Behavior Research Methods, vol. 46, pp. 416–428 (2014). DOI: 10.3758/s13428-013-0390-6

[18] Mathias Meyer. *Continuous Integration and Its Tools*. In: IEEE Software, vol. 31, pp. 14–16 (2014). DOI: 10.1109/MS.2014.58

[19] A. Mili, R. Mili, and R.T. Mittermeir. *A survey of software reuse libraries*. In: Annals of Software Engineering, vol. 5, pp. 349–414 (1998). DOI: 10.1023/A:1018964121953

[20] Jung-Ran Park and Yuji Tosaka. *Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms*. In: Cataloging & Classification Quarterly, Routledge, vol. 48, pp. 696–715 (2010). DOI: 10.1080/01639374.2010.508711

[21] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. *PsychoPy2: experiments in behavior made easy.* In: Behavior Research Methods, vol. 51, pp. 195–203 (2019). DOI: 10.3758/s13428-018-01193-y

[22] Sabine Verboven and Mia Hubert. *LIBRA: a MATLAB library for robust analysis*. In: Chemometrics and Intelligent Laboratory Systems, vol. 75, pp. 127–136 (2005). DOI:10.1016/j.chemolab.2004.06.003

Implementierung der FAIR-Prinzipien im Forschungsdatenmanagement: Eine Terminologie-basierte Strategie für die inhaltliche Beschreibung numerischer Faktendatensätze

Giacomo Lanza[1], Joachim Erich Meierl[1], Ulrich Schwardmann[2] und Thomas Wiedenhöfer[1]

[1]Physikalisch-Technische Bundesanstalt (PTB) ;
[2]Gesellschaft für die Wissenschaftliche Datenverarbeitung Göttingen (GWDG)

Abstract - Deutsch

In der Open Science-Ökonomie stellen numerische Faktendaten eine disziplinübergreifende Herausforderung für die praktische Umsetzung der vier FAIR-Prinzipien dar. Die zu erwartende unüberschaubar große Anzahl von Datensätzen und die heute schon in den verschiedenen Disziplinen gängige große Vielfalt an verwendeten Datenformaten und -strukturen sind wesentliche Ursache des Problems. Diese Heterogenität erschwert die Festlegung eines einheitlichen Standards zur Metadaten-Beschreibung und Archivierung von Forschungsdaten unterschiedlichen Ursprungs, und folglich deren Nachnutzung: z.B. sieht das *DataCite*-Metadatenschema keine Felder für eine detaillierte Beschreibung numerischer Faktendaten jenseits der Angabe unkontrollierter Schlagwörter vor. Vor diesem Hintergrund ist bereits das F-Prinzip (Auffindbarkeit) nur eingeschränkt umsetzbar: zielgerichtetes, feingranulares Suchen und präzises Finden auf Datenrepositorien-übergreifender Ebene ist auf dem derzeitigen Stand nicht möglich.

Wir setzen zur Lösung des Problems bei den typischen Eigenschaften numerischer Faktendaten an. Diese sind gekennzeichnet durch Messgrößen, Maßeinheiten, numerische Werteangaben, Rollen (z.B. Messgröße, Messvariable, Messparameter) und, bei quantitativ bewerteter Zuverlässigkeit der Faktendaten, auch die Messunsicherheitsangabe und das Messunsicherheitsmodell. Als Module eines *Metrology Terminology Directory* (MTD) werden in abgegrenzten Namensräumen kontrollierte Vokabulare für Messgrößen, Maßeinheiten, Messverfahren und verschiedene Charakteristiken von Messobjekten mehrsprachig entwickelt und jeweils über spezifische *Persistent Identifiers* in einer sogenannten Data Type Registry sprachübergreifend adressierbar gemacht. In von uns neu entwickelten Faktendaten-spezifisch strukturierten Metadatenmodulen dienen diese Vokabulare zur Beschreibung des Faktendatensatzes. Auf diese Weise werden die wesentlichen Eigenschaften numerischer Faktendatensätze für komplexes Suchen und Finden mit geeigneten Retrieval-Werkzeugen zugänglich gemacht.

Giacomo Lanza et al.: Implementierung der FAIR-Prinzipien im Forschungsdatenmanagement: Eine Terminologie-basierte Strategie für die inhaltliche Beschreibung numerischer Faktendatensätze.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge, Heidelberg: heiBOOKS, 2020, 47–57. DOI: https://doi.org/10.11588/heibooks.598.c8416

Durch Implementierung der Metadatenschema-Module seitens der Hersteller von digitalen Messgeräten, digitalen Sensoren, Messdatenverarbeitungs-Software, könnte zukünftig erreicht werden, dass die Metadatenbeschreibung schon bei der erstmaligen analog/digital-Wandlung von Messdaten beginnen und dann über die weitere Verarbeitungskette bis zur Archivierung und Publikation angereichert werden kann. Das würde die Dokumentationsarbeit erleichtern, einen gewissen Qualitätsstandard einführen und somit zu einer effizienten Umsetzung der FAIR-Prinzipien beitragen können.

Über eine Pilotrealisierung der MTD und ausgewählter Metadaten-Module werden wir im Vortrag berichten.

Abstract - Englisch

Within the Open Science economy, a major challenge for the practical realisation of the FAIR principles is represented by numerical factual data. This is a result of the overwhelming big quantity of data sets, as well as of the big variety of data structures currently used in the different disciplines. This heterogeneity makes it difficult to compare data structures of different origins, as well as the establishment of unique standards for their description through metadata: for instance, DataCite metadata scheme doesn't include any fields for the detailed description of factual data, other than the inclusion of free-text – thus not controlled – keywords. Under these conditions there are limited possibilities only to implement even the first of the FAIR principles, Findability: a guided advanced search of factual data over a plurality of repositories is currently not feasible.

We are approaching a solution to this problem via the typical characteristics of numerical factual data, which are commonly described by quantity names, units, numerical value ranges, roles (measured quantity, variable or parameter) and some estimate of the data reliability, such as the uncertainty budget and the uncertainty model. Within a newly defined "Metrology Terminology Directory"(MTD), we are developing a collection of multilingual controlled vocabularies for physical quantities, measuring units, experimental techniques and selected information about research objects; the correct designation of each item takes place in a language-independent manner via a persistent identifier. These vocabularies are then applied within an ad hoc defined metadata module for numerical factual data for a thorough description of a dataset. That way, the relevant features of numeric factual data are made accessible for complex searches and finding with suitable retrieval tools.

If the proposed metadata module is adopted and implemented by the producers of digital measuring instruments, digital sensors and software for data analysis, it will be possible in the future to implement a standardised metadata description already at the point of the first analog-to-digital conversion, and then propagate and enrich it somehow automatically along all subsequent data transformation steps. This would facilitate documentation work considerably, as well as would contribute to the realisation of all elements of the FAIR principles.

Within the talk a pilot realisation of the MTD will be presented, along with selected metadata modules.

1. Hintergrund und Problemstellung

Der freie Austausch von Informationen und Wissen ist eine der tragenden Säulen der Wissenschaft. Mit zunehmender Digitalisierung in der Wissenschaft und mit der Verbreitung der Open-Science-Philosophie werden in wachsendem Ausmaß auch Forschungsdaten der Öffentlichkeit verfügbar gemacht. Diese Daten fallen in einer Vielzahl offener und proprietärer Formate an. Für deren Beschreibung werden Metadaten verwendet, deren Umfang i. d. R. von den Möglichkeiten der Datenportale oder von Fachgemeinschaften bestimmt werden. Die Qualität der Metadatenerfassung ist dabei abhängig von der fachlichen Expertise und dem Vollständigkeitsanspruch der mit dem Datenkuratieren beauftragten Personen.

Eine Nachnutzung dieser Daten setzt die Fähigkeit voraus, sie mittels geeigneter Suchmaschinen selektiv recherchieren und filtern zu können. Eine feingliedrige, „erweiterte“ Suche bedarf einer gemeinsamen „Sprache“ zwischen den Suchmaschinen und den Datenrepositorien. Das wurde zum Teil erreicht mit der Festlegung standardisierter Metadatenschemata (DataCite [1]) und Protokolle für das *Metadata Harvesting* (OAI-PMH); darüber hinaus wenden einige Fachdisziplinen zusätzliche Metadatenmodule an, die eine feingranulare Beschreibung ermöglichen. In gewissen Fällen wird auch eine Suche im Datenbestand selbst angeboten.

Angesichts der wachsenden Datenmengen und der anfallenden multidisziplinären Fragestellungen wird der Bedarf immer offensichtlicher an einer interoperablen Struktur für die dokumentarische Beschreibung der Daten, die das zuverlässige Finden, Filtern und Vergleichen von Forschungsdaten unterschiedlichen Ursprungs ermöglicht. Die weitgehend KI-unterstützten Methoden der *Big Data*-Analyse und des *Machine Learning* erfordern, dass die Daten nicht nur für Menschen zugänglich, sondern auch maschinenlesbar und -verständlich sind. Dies erfordert eine Auswahl standardisierter, langlebig lesbarer Datenformate, sowie eine eindeutige Kodierung der Metadatenbeschreibungen mit Standardisierung sowohl der Attributfelder, als auch deren zulässiger Inhalte (Attributwerte). Die für Forschungsdatenmanagement eingeführten vier FAIR-Prinzipien [2], die Daten sollen auffindbar (*Findable*), zugänglich (*Accessible*), interoperabel (*Interoperable*) und wiederverwendbar (*Reusable*) sein, geben die Ziele vor, machen aber keine Angaben für Lösungen dieser nicht trivialen Aufgabenstellung.

Sehr stark automatisierungs- und messtechnisch geprägte Industrieunternehmen sind sich dieser Herausforderungen seit längerem bewusst. Firmeneigene Schemata und Protokolle für die digitale Übertragung von Informationen sind bereits entwickelt [3, 4, 5, 6].

2. Lösungsansatz

Forschungsdaten ohne eine standardisierte, feingranulare, den Suchmaschinen zugängliche Metadaten-Beschreibung sind schwierig zu finden bzw. mit zunehmender Komplexität schwer oder nicht immer eindeutig interpretierbar.

Um die Interoperabilität von Forschungsdaten zu gewährleisten, ist die Einführung einer gemeinsamen Grundlage für die Metadaten-Beschreibung unentbehrlich. Diese entsteht aus (1) einer definierten Palette von Attributen (Metadatenfelder) sowie (2) je einer kontrollierten Liste zulässiger Werte, die eine eindeutige, reproduzierbare und zweifelsfreie

Erfassung ermöglicht. Das Ganze sollte international und womöglich fachübergreifend abgestimmt werden. Ein solcher Standard besteht aus den folgenden Bausteinen:

- Ein Metadatenschema, also eine hierarchische Anordnung kontrollierter Metadatenfelder mit festgelegten Benennungen und vorgegebenen Regeln zur inhaltlichen Belegung (Variabeltyp, Freitext / kontrollierter Text). Erstrebenswert ist hier ein modularer Aufbau, bei dem neben einem gemeinsamen Kernschema (z.B. eine Erweiterung des aktuell verwendeten DataCite-Metadatenschema) unterschiedliche fachspezifische oder methodenspezifische Metadaten-Module optional verwendet werden können.

- Mehrere kontrollierte Vokabulare, die für die Inhalte ausgewählter Metadatenfelder eine Liste zulässiger Werte (Terme) zur Verfügung stellen. Je nach Komplexität können diese Vokabulare als Thesauri oder Ontologien realisiert werden. Um die Mehrdeutigkeiten der natürlichen Sprache zu überwinden, ist es ratsam, die einzelnen Begriffe mit langlebigen Kennzeichen (Persistent Identifiers) zu versehen. Hierdurch wird die Verfügbarkeit der Vokabulare in mehreren (menschlichen) Sprachen unter Beibehaltung der logischen Struktur möglich.

  Eine derartige Vorgehensweise überwindet die Begrenztheit generischer Metadatenschemata (wie z.B. das von DataCite) und ermöglicht gleichzeitig die eindeutige Interpretation (Disambiguierung) der Benennungen von Feldern und Inhalten, die in unterschiedlichen Fachbereichen unterschiedliche Bedeutung tragen könnten.

3. Beispiel: Größen und Einheiten

In den quantitativen Wissenschaften (z. B. Chemie, Physik, Materialwissenschaften und andere technische Wissenschaftsgebiete) spielen numerische Faktendaten eine zentrale Rolle. Die grundlegenden Informationen für deren eindeutige Identifizierung sollten Angaben beinhalten über

- Versuchsobjekt: Probentyp, Proben-ID, chemische Identität, Hersteller, Auftraggeber;

- Datengenerierung: experimentelle Methode bzw. Simulationsprozedur, Identifikation des Messplatzes, Zeitpunkt, Mitarbeiter. . . ;

- Faktendaten-Merkmale: Größe, Maßeinheit, numerischer Wertebereich, numerische Auflösung, Messunsicherheit, Rolle (Messgröße, Variable oder Parameter).

Im vorgestellten Vorgehen wurde zunächst die Darstellung der Datenwerte behandelt. In der Folge wurde ein atomares Metadatenschema für numerische Faktendaten und Vokabulare für physikalische Größen und Einheiten entwickelt.

4. Ergebnisse: Metadatenschema

Im Rahmen des innerhalb Horizon 2020 europäisch geförderten Projektes *SmartCom* wurde für industrielle Anwendung ein atomares Metadatenschema (**D-SI** [7]), basierend auf internationalen Richtlinien der Messtechnik, für die Beschreibung numerischer Faktendaten definiert. Das Schema sieht Felder für die Angabe eines Messwerts mit Maßeinheit und Messunsicherheit vor. Die Unsicherheit kann entweder als absolute erweiterte Unsicherheit samt Erweiterungsfaktor, oder als Intervall samt Intervallgrenzen dargestellt werden; in beiden Fällen wird die Überdeckungswahrscheinlichkeit angegeben. Optionale Angaben dazu sind der Größenname, der Zeitstempel sowie die angenommene Wahrscheinlichkeitsdichteverteilung des Messunsicherheitswertes. Die Qualitätskontrolle legt den größten Wert auf die Angabe der Einheiten, wobei SI-Basiseinheiten ohne Präfixe stark empfohlen werden.

Tabelle 1.: Vorgeschlagene Metadatenfelder für die Beschreibung numerischer Faktendaten. Für jede Größe werden ein PID, die Rolle, die Einheit, die numerischen Extremwerte und die Unsicherheit angegeben. Die **fett** markierte Felder stellen eine Erweiterung im Vergleich zum heutigen **D-SI**-Schema dar.

Feld	**Beschreibung**	**Beispiel**	***Obligation***
`block`	Einzeldatei, oder zusammenhängender Tabellenbereich		M
`block/npoints`	Anzahl Messwerte	941	M
`block/quantity`	In einer Datei enthaltene Größe		M
`block/quantity/`**`role`**	Funktion der aufgeführten Größe: "measurand", "parameter" oder "variable"	`https://unserDatenraum.boh/role:variable`	M
`block/quantity/`**`qty_id`**	Identifikator der Größe	`https://unserDatenraum.boh/qty:wavelength`	M
`block/quantity/description`	Ausführliche Beschreibung	Wellenlänge des Lasers zur Bestrahlung der Probe	O
`block/quantity/list/label`	Kurzbezeichnung / Symbol	X	O
`block/quantity/list/real[label="min"]/value`	Minimalwert	230	M
`block/quantity/list/real[label="max"]/value`	Maximalwert	700	M
`block/quantity/list/`**`unit_id`**	Identifikator der Einheit	`https://unserDatenraum.boh/unit:nm`	M
`block/quantity/list/uncertainty`	Erweiterte Messunsicherheit (wenn role="measurand")	1	MA
`block/quantity/list/coverageFactor`	Erweiterungsfaktor (default=2)	2	MA

Für die Beschreibung von Forschungsdaten für Open Science-Anwendungen wird dieses Schema erweitert. Für jede im Datensatz auftauchende Größe werden sowohl der Wertebereich (Minimum und Maximum) als auch die Unsicherheit angegeben. Die Benennung

der Größe ist verpflichtend und soll möglichst aus einem normierten Vokabular stammen. Zusätzlich wird die Rolle der Größe (Messgröße, Messparameter oder Messvariable) sowie eine wörtliche Erläuterung angegeben.

Das vorgeschlagene Schema bildet die hierarchische Anordnung der relevanten Felder (Tabelle 1) ab. Die korrekte Eingabe der Metadatenwerte wird durch Anwendungsregeln unterstützt. Struktur und Regeln sind bewusst syntaxneutral angelegt und ermöglichen deshalb die äquivalente Auszeichnung bzw. Export der Metadaten in XML, JSON, YAML oder einem anderen beliebigen Auszeichnungsformat bzw. Notation.

5. Ergebnisse: kontrollierte Vokabulare

Die entwickelten Vokabulare werden in einem *Metrology Terminology Directory* gesammelt und sollen in maschinenlesbarem Format (JSON, RDF oder OWL) für die Wissenschaft, die Wirtschaft und die Öffentlichkeit zur Verfügung gestellt werden. Eine Testinstanz befindet sich auf dem von der GWDG betriebenen ePIC-Server [8].

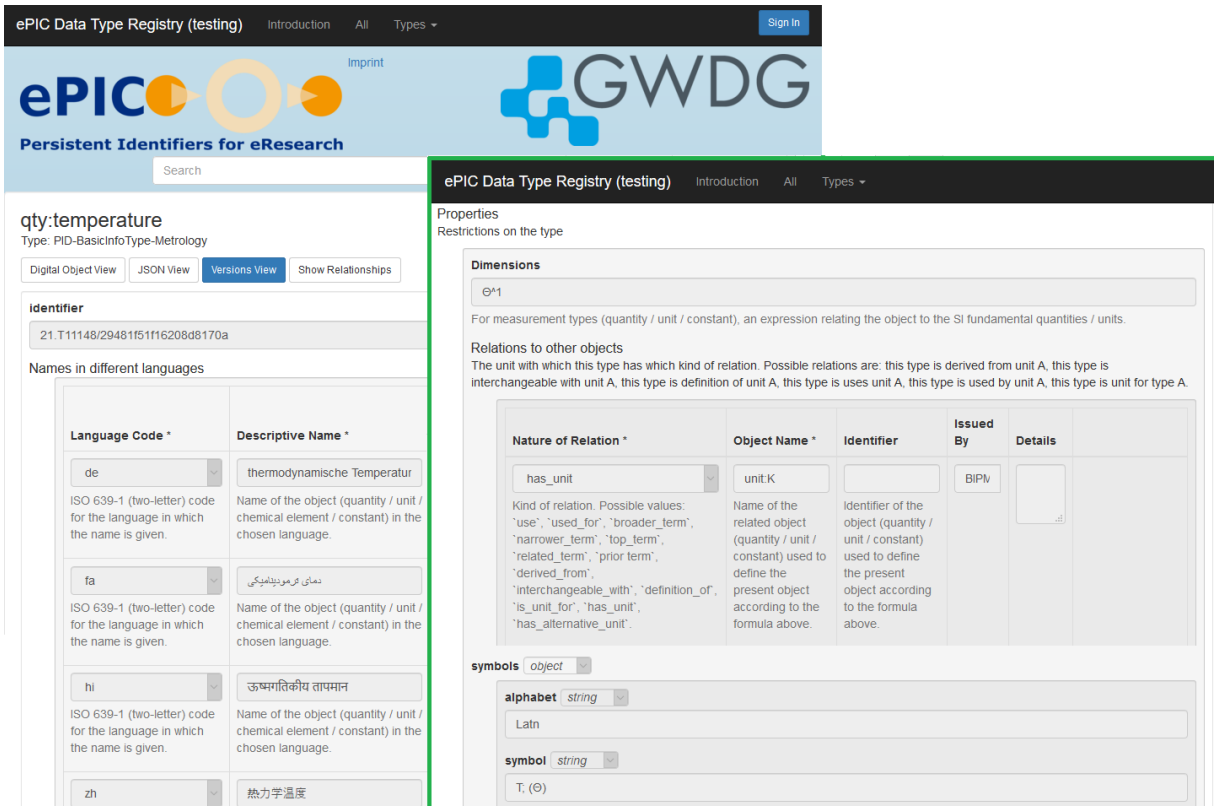


Abbildung 1.: Ein Beispiel aus dem Vokabular für physikalische Größen (`qty`=*quantity*). Für jeden Eintrag (in diesem Fall, die Größe "Thermodynamische Temperatur") werden ein Identifikator und standardisierte mehrsprachige Bezeichnungen zugewiesen.

Das Vokabular für physikalische Größen besteht derzeit aus ca. 400 Einträgen, von denen ein Beispiel in Abbildung 1 wiedegegeben wird. Für jeden Eintrag werden die folgenden Attribute angeboten:

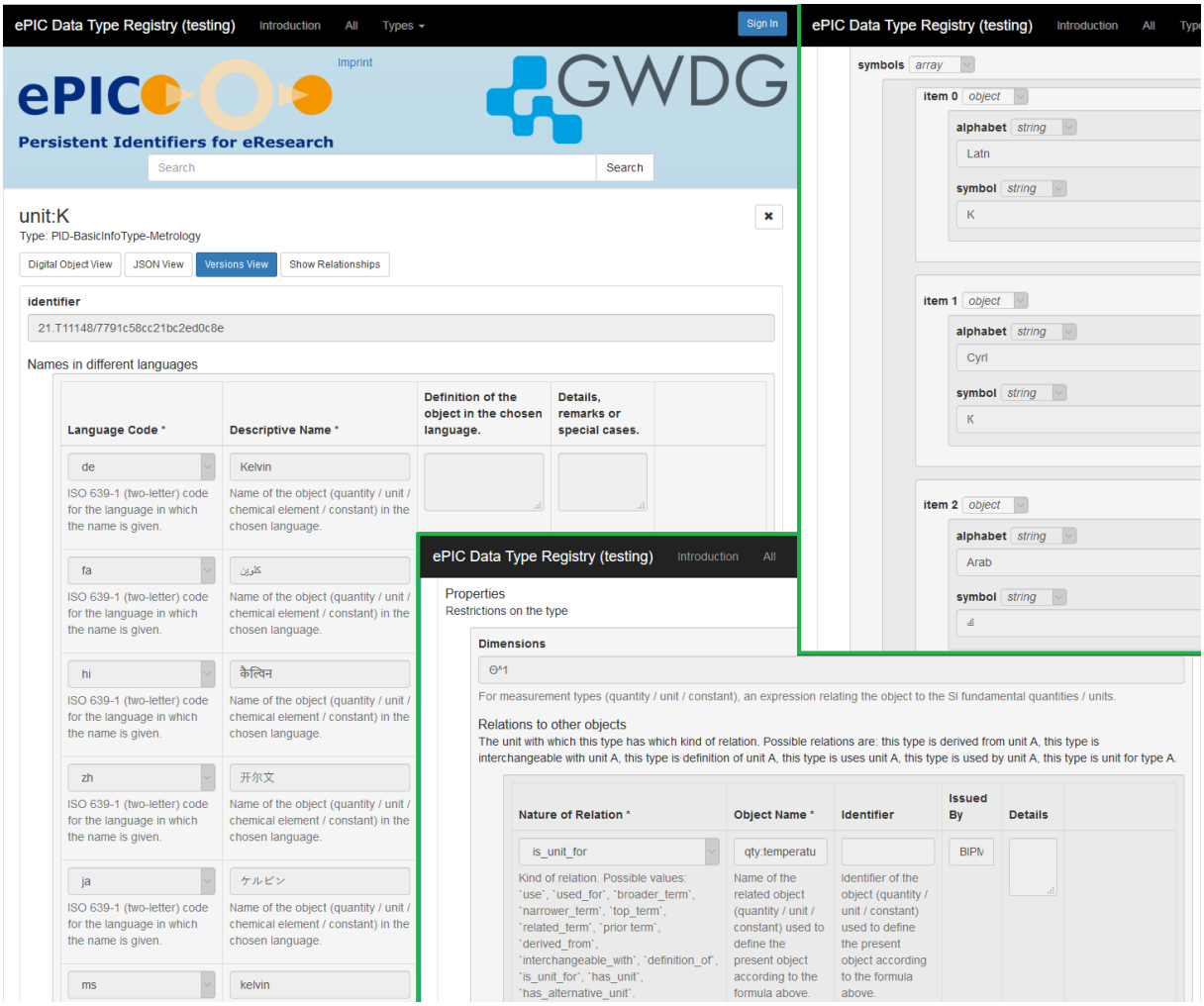


Abbildung 2.: Ein Beispiel aus dem Vokabular für Einheiten (`unit`). Zum Eintrag (in diesem Fall die Einheit "Kelvin") gehören ein Identifikator und die Benennungen in mehreren Sprachen. Das Einheitensymbol in verschiedenen Schriftarten (lateinisch, kyrillisch, arabisch, LaTeX-Kodierung) wird ebenso angezeigt.

- Benennung, derzeit in 18 Sprachen.
- Symbol und Definitionsformel.
- Dimensionen nach BIPM Broschüre [9]; Bezug zur vorgegebenen SI-Einheit.
- Kurzer Definitionstext und Anmerkungen.
- Notation aus einer mehrstufigen Klassifikation.

Das Vokabular für Naturkonstanten, in Anlehnung an die Liste von CODATA [10], enthält derzeit 84 Einträge. Für jede Naturkonstante werden die schon bei Größen gelisteten Attribute angegeben, und zusätzlich:

- Der numerische Wert der Naturkonstante mit absoluter und relativer Standardunsicherheit und die Gültigkeitszeitspanne dieses Wertes (basiert auf den CODATA-Ausgaben von 1969, 1973, 1986, 1998, 2002, 2006, 2010, 2014 und 2018).
- Die Korrelationskoeffizienten zu den anderen Naturkonstanten, mit Gültigkeitszeitspanne.

Das Vokabular für Einheiten listet derzeit ca. 100 Einträge auf; ein Beispiel wird in Abbildung 2 gezeigt. Dazu gehören, konform zur BIPM-Broschüre, die 22 kohärenten Einheiten und die derzeit noch akzeptierten Nicht-SI-Einheiten, sowie deren Kombinationen nach Summen und Produkten (z. B. J/K, N·m). Für die Einheiten werden folgende Attribute vergeben:

- Benennung, derzeit in 18 Sprachen.
- Symbol in lateinischem, kyrillischem und arabischem Alphabet, sowie LaTeX-Quellcode.
- Definitionsformel mit Umrechnungsfaktor, sofern notwendig.
- Dimensionen nach BIPM-Broschüre und Bezug zu den damit verbundenen Größen und Konstanten.
- Kurzer Definitionstext und Anmerkungen.

Geplant ist ein viertes Vokabular, das auf Basis des VIM (Internationales Wörterbuch der Metrologie) [11] und des GUM (Guide to the expression of uncertainty in measurement) [12] die grundlegenden metrologischen Begriffe mehrsprachig und maschinenlesbar darstellen wird. Unter anderen werden die Grundbegriffe in Bezug auf Messverfahren, Messunsicherheit sowie die Rolle einer Größe aufgeführt.

6. Beispielanwendung

Das hier eingeführte Schema samt Vokabularen stellt eine mögliche Grundlage dar für die interoperable Erfassung von Forschungsdaten in einem Datenportal oder Repositorium, das eine feingranulare Suche ermöglicht. Dazu soll eine spezialisierte Suchmaschine eingerichtet werden, die nach Größennamen und Wertebereichen indexieren und filtern kann.

Ein möglicher Anwendungsfall: Ein Materialwissenschaftler könnte nach aktuellen Erkenntnissen über die Temperaturabhängigkeit (*Variable*) verschiedener Materialeigenschaften (*Messgrößen*) einer Substanz unter gewissen Umweltbedingungen (*Parameter*) suchen. Eine solche Recherche sollte alle Datensätze finden und wiedergeben, in denen die gewünschte Substanz in der Probenbeschreibung, die Materialeigenschaft als Größe vorkommt und die Temperaturwerte zwischen vorgegebenen Grenzwerten (z. B. zwischen 200 K und 1000 K) liegen.

Eine solche Suchplattform, der *MessdatenMetaViewer* (Abbildung 3 (a)), wurde in unserer Arbeitsgruppe entwickelt. Die Anwendung basiert auf einer Datenbank, in der die

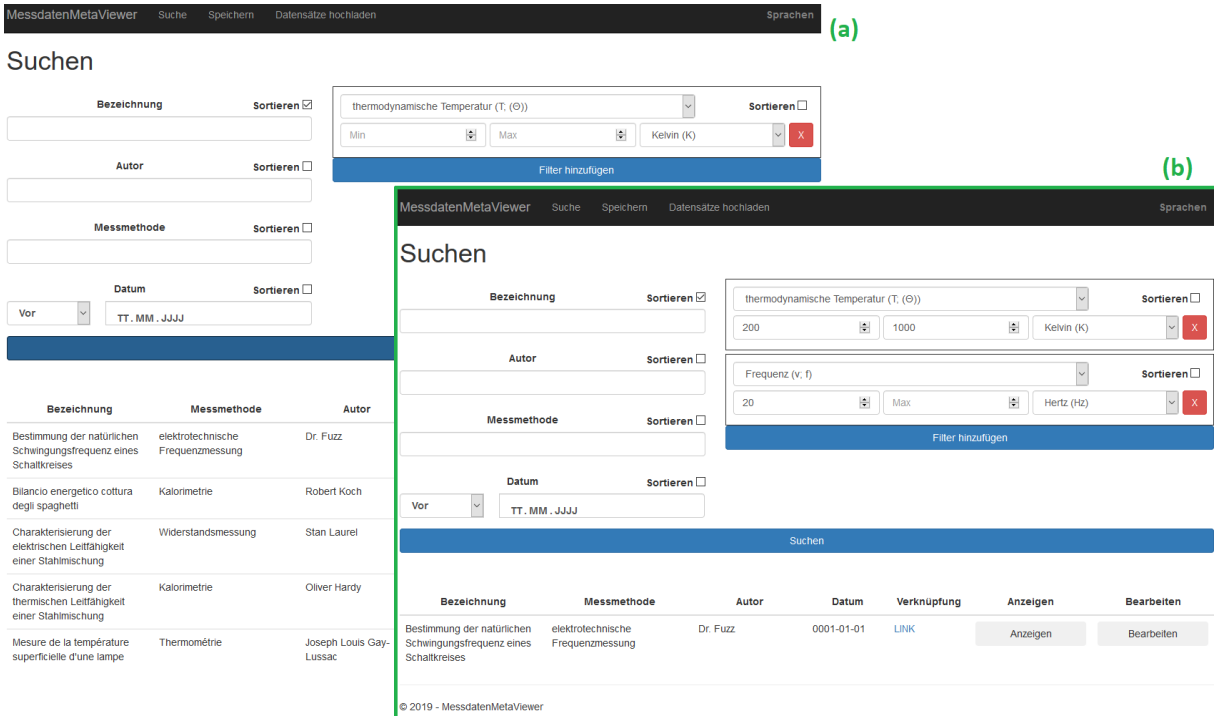


Abbildung 3.: Die Suchplattform *MessdatenMetaViewer*, die eine erweiterte Suche auf Basis der von uns definierten Felder und der festgelegten Vokabulare ausführt und nach Größen und Wertebereichen filtern kann. (a): Ein Schnappschuss aus einer kleinen Forschungsdatensammlung, mit der Suchmaske. (b): Die gleiche Datensammlung nach einer Suche nach den Größen *Thermodynamische Temperatur* und *Frequenz*.

Datensätze nach Autor, Datum, Messmethode und zugehörigen Größen erfasst sind; für jede Größe werden zusätzlich die dazugehörige Einheit und der Wertebereich angegeben. Die Suchfunktion filtert sowohl nach Größen, als auch nach Wertebereichen und macht es deswegen möglich, die passenden Datensätze aus der Menge auszuwählen. Ein Beispiel wird in Abbildung 3 (b) gezeigt, in der nur die Datensätze mit Frequenz- und Temperaturangabe in einem gewissen Bereich wiedergegeben werden.

7. Schlussfolgerung

Das vorgeschlagene Metadatenmodul ermöglich eine feingranulare Erfassung von Datensätzen durch wenige bedeutsamen Metadatenfelder und lässt sich leicht als modulare Erweiterung in andere, auch fachspezifische Metadatenschemata integrieren. Die Speicherung der Metadatenwerte als Persistent Identifiers garantiert die eindeutige maschinelle Lesbarkeit der Inhalte und ermöglicht eine schnelle Übersetzung in beliebige menschliche Sprachen für diverse Anwendungen. Da jeder erfassten Größe eine eindeutige Rolle (Mess-Größe, -Parameter, -Variable) und das Intervall deren numerischen Werte zugewiesen wird, ermöglicht das die zielgerichtete Suche nach numerischen Faktendaten.

Die Verwendung persistenter Identifikatoren für die kontrollierten Vokabulare für die

Benennung der Metadatenfelder und -inhalte aus einem Thesaurus ermöglicht die Kommunikation mit Suchmaschinen, die über verschiedenen Datenbanken recherchieren. Somit ermöglicht der hier vorgestellte terminologische Ansatz einen vielversprechenden Weg, numerische Faktendaten nachhaltig und interoperabel zu beschreiben, zu speichern und gezielt wieder zu finden.

8. Ausblick

Schon heute werden bei einer softwaregesteuerten Messung wichtige Parameter über den Messablauf von den Messgeräten selbst in den Kopfzeilen der erzeugten Rohdaten oder in zusätzlichen begleitenden Dateien protokolliert. Auch erlaubt bestimmte Software für die wissenschaftliche Datenauswertung die Speicherung von Informationen über enthaltene Variablen und deren numerischen Werte. Die Lesbarkeit und Nachnutzbarkeit dieser Metadaten hängt vom Format und von der verwendeten Syntax ab. Wären sie in einem standardisierten Format abgespeichert, so könnten sie leicht von anderer Software gelesen, interpretiert und genutzt werden. Somit würde die Metadatenbeschreibung eines Messobjektes entlang des Datenverarbeitungsprozesses stetig angereichert werden und am Ende würde die gesamte Dokumentation über den Messprozess und seine Ergebnisse vorliegen. Metadatenerfassung durch einen Menschen würde so weitgehend vermieden und beschränkte sich daher auf diejenigen Felder, die nicht im Verlauf des Prozesses automatisch beschrieben werden können. Weitere wesentliche Vorteile dieser Vorgehensweise sind, dass fehleranfällige nachträgliche „händische“ Erfassung von Metadaten minimiert wird und die Kooperationsbereitschaft für Forschungsdatenmanagement mit Zielrichtung Open Science beim für die Datengewinnung zuständigen Personal wächst.

Um die Interoperabilität, eine breite Akzeptanz und die Praxistauglichkeit sicherzustellen, wird die Einbeziehung von Messgeräteherstellern und Herstellern wissenschaftlicher Software (ELN - Elektronische Laborbücher, LIMS - Laborinformationsmanagementsysteme) bei der Festlegung des Formats angestrebt.

Danksagung

Wir bedanken uns bei unseren PTB-Kollegen Robin Becker für die Entwicklung der Suchplattform *MessdatenMetaViewer* und Sascha Eichstädt für nützliche Diskussionen.

Literaturverzeichnis

[1] The DataCite Schema. https://schema.datacite.org

[2] FORCE 11: the FAIR data principles.https://www.go-fair.org/fair-principles/

[3] ASAM, the ASAM Open Data Services format.
https://www.asam.net/standards/detail/ods/wiki/

[4] Woopsa Protocol Specifications. http://www.woopsa.org/specifications/

[5] eCl@ss - Standard für Stammdaten und Semantik für die Digitalisierung. https://www.eclasscontent.com/index.php

[6] Das Referenzarchitekturmodell Industrie 4.0 (RAMI 4.0). https://www.plattform-i40.de/I40/Redaktion/DE/Downloads/Publikation/struktur-der-verwaltungsschale.pdf

[7] EMPIR 17IND02 (SmartCom). D-SI: XML implementation of the Digital SI (D-SI) meta data model for the exchange of metrological data - version 1.0.1-beta. https://www.ptb.de/si/smartcom/d-si/v1_0_1/SI_Format.xsd

[8] Projekt ePIC: Persistent identifiers for research. http://dtr-test.pidconsortium.eu/

[9] SI, Le Système International d'Unités—The International System of Units, 9th edn. BIPM 2019, ISBN 978-92-822-2272-0. https://www.bipm.org/fr/publications/si-brochure/

[10] P. J. Mohr, D. B. Newell and B. N. Taylor. CODATA recommended values of the fundamental physical constants: 2014. https://ws680.nist.gov/publication/get_pdf.cfm?pub_id=920686

[11] Bureau International des Poids et des Mesures, JCGM 200:2012. Vocabulaire international de métrologie - Concepts fondamentaux et généraux et termes associés (VIM). https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf

[12] Bureau International des Poids et des Mesures, JCGM 100:2008. Évaluation des données de mesure - Guide pour l'expression de l'incertitude de mesure (GUM). https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_F.pdf

Kollaborative Forschungsunterstützung: Ein Integriertes Probenmanagement

Marius Politze[1], Annett Schwarz[1], Sebastian Kirchmeyer[2], Florian Claus[1] und Matthias S. Müller[1]

[1]IT Center, RWTH Aachen University ;
[2]Lehrstuhl für Physikalische Chemie II und Institut für Physikalische Chemie, RWTH Aachen

Bei der zielgerichteten Unterstützung des Forschungsdatenmanagements stehen zentrale Infrastruktureinrichtungen wie Bibliothek und Rechenzentrum vor enormen Herausforderungen: Es gilt, die Vielfalt der wissenschaftlichen Disziplinen abzubilden. Zugleich müssen Unterstützungsangebote gut skalieren und auf Basis zentraler Services realisierbar sein. Im Rahmen des SFB 985 „Funktionelle Mikrogele und Mikrogelsysteme“ wurde ein integriertes Probenmanagement entwickelt, das auf Basis der hochschulweit genutzten Kollaborationsplattform wissenschaftliche Zusammenarbeit mit Experimentaldaten unterstützt. Die Plattform wird im SFB bereits für klassisches Dokumentenmanagement genutzt und ist den Forschenden daher für Informationsaustausch und der gemeinsamen Arbeit an Dateien bekannt. Innerhalb eines Arbeitsbereichs haben Forschende nun die Möglichkeit, Proben zu beschreiben und begleitende Dokumente sowie Messdaten strukturiert zu hinterlegen und direkt online zu bearbeiten. Für die Integration dieser virtuellen Arbeitsbereiche mit den Proben im Labor ermöglicht die Anwendung das Erstellen spezieller Etikettenvordrucke. Diese QR-Codes können in den Laboren und an den Arbeitsplätzen mit dafür vorgesehenen Lesegeräten eingelesen werden und erlauben das direkte Navigieren zum Arbeitsbereich der jeweiligen Probe. Ziel des Probenmanagements ist es, so den gesamten Forschungsdatenlebenszyklus von der Erstellung, Verarbeitung und Analyse bis hin zur langfristigen Speicherung und Archivierung dieser Daten zu unterstützen.

1. Einleitung

Forschungsdaten und die dazugehörigen Infrastrukturen stehen aktuell im Fokus. Diese bezieht sich vor allem auf nationale, europäische und internationale Strukturen wie die geplante nationale Forschungsdaten-Infrastruktur (NFDI)[1] oder die European Open Science Cloud (EOSC)[2] und entsprechende Projekte wie der EOSCPilot oder EOSC Hub -[3].

Jenseits von (inter-)nationalen Strukturen sind an einer Universität die zentralen Infrastruktureinrichtungen wie Bibliothek und Rechenzentrum in der Pflicht. Zugleich stehen sie aufgrund der großen Vielfalt von Disziplinen vor enormen Herausforderungen. So dominieren an der RWTH Aachen University zwar zumindest zahlenmäßig Natur- und Ingenieurwissenschaften. Aber wie sich aus den bisherigen Erfahrungen im Projekt

Marius Politze et al.: Kollaborative Forschungsunterstützung: Ein Integriertes Probenmanagement.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 58–67. DOI: https://doi.org/10.11588/heibooks.598.c8417

„Forschungsdatenmanagement an der RWTH“ [4] gezeigt hat, divergieren die Bedürfnisse selbst innerhalb z.B. der Ingenieurwissenschaften stark. Wie kann nun auf die vielen unterschiedlichen fachspezifischen Bedürfnisse reagiert werden?

Trotz vielfältiger Angebote stellen Publizieren, öffentliches Nachweisen, aber auch langfristiges Speichern von Forschungsdaten ein für Forschende häufig ungelöstes Problem dar [5]. Andererseits zeigt sich, dass spezialisierte Lösungen, die gut auf Bedarfe der Forschenden abgestimmt sind, hohe Akzeptanz genießen, wie die im Rahmen von TR CRC 32- [6] und LIFE [7] entstandenen Repositorien zeigen. Ein Datenmanagementsystem muss sich entsprechend an den konkreten Bedarfen der Forschenden orientieren. Betrachtet man das Datenmanagement jedoch als rein institutionelle Aufgabe sind die Mehrwerte für Forschende oft nur indirekt und somit weniger ersichtlich. Die Integration in die Abläufe der Forschenden stellt die wesentliche Herausforderung für die Akzeptanz eines Datenmanagementsystems dar.

Im Rahmen des SFB 985 „Funktionelle Mikrogele und Mikrogelsysteme“ beteiligen sich Forschergruppen der RWTH Aachen, des Leibnitz Instituts für Interaktive Materialien, des Forschungszentrums Jülich sowie des Fraunhofer-Instituts für Lasertechnik. Zur Unterstützung der interdisziplinären Zusammenarbeit wurde ein integriertes Probenmanagement entwickelt, das auf Basis einer Kollaborationsplattform wissenschaftliche Zusammenarbeit mit Experimentaldaten unterstützt. Innerhalb des SFB dient die Kollaborationsplattform u.a. zum internen Informationsaustausch, zur Erstellung von Textpublikationen oder Versuchsskizzen in und zwischen Projektgruppen. Der Informationsaustausch über institutionelle Grenzen innerhalb der RWTH, aber auch zwischen den beteiligten Forschungseinrichtungen, stellt eine Herausforderung für die beteiligten Forschenden und Infrastrukturdienstleister dar, sodass sich diese zentrale Anlaufstelle für den SFB als sinnvoll herausgestellt hat.

Neben der gemeinsamen Arbeit an Dokumenten dient die Kollaborationsplattform auch zum Austausch und zur Verwaltung von Forschungsdaten, die im Rahmen von Experimenten anfallen. Stark vereinfacht ist der Ablauf im SFB wie folgt: In den beteiligten Instituten werden Proben von Mikrogelen synthetisiert und deren Eigenschaften analysiert. Die für die Synthese und Analyse notwendigen Versuchsaufbauten und Verfahren werden in individuellen Laborbüchern dokumentiert. Für eine effiziente Zusammenarbeit müssen diese Informationen allen Beteiligten zur Verfügung stehen. Das in die Kollaborationsplattform integrierte Probenmanagement soll den einfachen Austausch dieser Informationen innerhalb des SFB ermöglichen. Ziel des Probenmanagements ist es, den gesamten Forschungsdatenlebenszyklus (siehe Abbildung 1) von der Erstellung, Verarbeitung und Analyse bis hin zur langfristigen Speicherung und Archivierung dieser Daten zu unterstützen. Die den Forschenden angebotenen Workflows orientieren sich dabei an den FAIR-Prinzipien (findable, accessible, interoperable, re-usable).

1.1. Unterstützter Forschungsworkflow

In der Kollaborationsplattform werden die bei der Synthese einer Probe anfallenden Metadaten dokumentiert. Dadurch wird für jede Probe ein eigener Bereich erstellt, der es ermöglicht, Dokumente und andere Dateien zu speichern, auszutauschen und gemeinsam

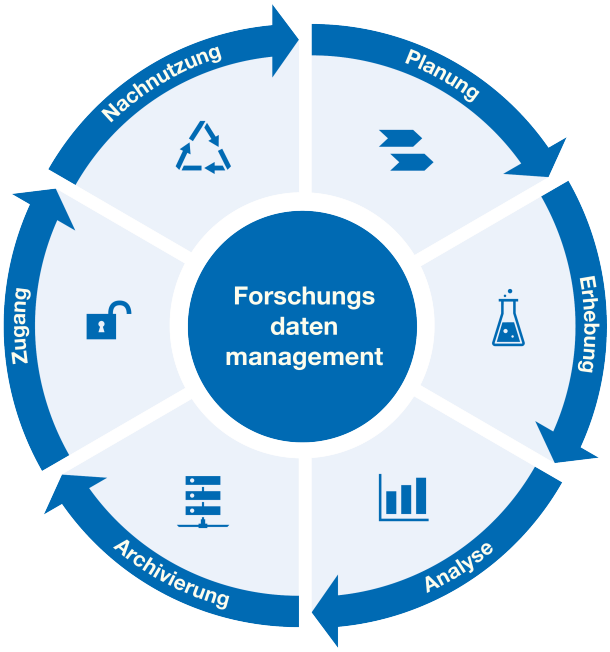


Abbildung 1.: Forschungsdaten Lebens-Zyklus der RWTH Aachen University.

zu bearbeiten. Die abgelegten Dateien stehen anderen Forschenden im SFB direkt zur Verfügung und können so wiederverwendet werden. Um die physische Probe mit dem virtuellen Bereich zu verbinden, erstellt das Probenmanagement einen Etikettenvordruck, der zusätzlich zu den Metadaten mit einem QR-Code über eine URL auf den Bereich der Probe verweist. Durch Auslesen der QR-Codes an PCs mit Handscannern oder Kamera-Apps auf Smartphones oder Tablets wird der Bereich geöffnet. Neben den Proben können durch Aufkleber auch Seiten im Laborjournal mit den virtuellen Inhalten auf der Kollaborationsplattform verknüpft werden. Das Probenmanagement verbindet so die virtuellen und physischen Arbeitsumgebungen der Forschenden.

Zusammen mit den Metadaten erhält jede Probe durch diesen Prozess einen eindeutigen Bezeichner (PID), der über institutionelle Grenzen hinweg für die Beschreibung der Probe verwendet werden kann. Eingesetzte Proben können so jederzeit projektübergreifend identifiziert und verfolgt werden. Durch die direkte Verknüpfung der Proben und dazugehöriger Dateien, wie Messergebnisse oder Versuchsprotokolle lassen sich doppelte Untersuchungen vermeiden und Resultate sind direkt für Kollaborationspartner verfügbar.

1.2. SharePoint als Kollaborationsplattform zur Forschungsunterstützung

SharePoint wird als Kollaborations- und Dokumentenmanagement-Lösung bereits in vielen Kontexten an der RWTH Aachen eingesetzt. Bei verschiedenen Arbeitsgruppen in SFBs, Forschungsprojekten oder Instituten finden dabei im Wesentlichen Features zur Zusammenarbeit an Dokumenten oder für die Versionskontrolle Anwendung. Daneben bietet die Plattform Möglichkeiten zur Strukturierung von Informationen in sogenannten Listen.

Diese Listen enthalten, ähnlich wie Datenbanktabellen, Elemente, deren Eigenschaften über Spalten verschiedener Datentypen definiert werden. Jede Arbeitsgruppe erhält dafür eine sogenannte „Site“, die unabhängig von allen anderen an die eigenen Bedürfnisse angepasst werden kann. Dazu werden in einem Workshop zunächst die Anforderungen der Arbeitsgruppe, z.B. eines SFB, aufgenommen, analysiert und mit Bordmitteln bestmöglich abgebildet. An der RWTH wird diese initiale Konfiguration durch Mitarbeiter des IT Centers begleitet und kann daraufhin von Mitarbeitern der Arbeitsgruppen weitgehend selbstständig angepasst werden. Insgesamt unterstützt das IT Center mit dem Angebot aktuell gut 100 Arbeitsgruppen mit SharePoint „Sites“.

Neben der klassischen Verwendung als Dokumentenmanagementsystem wird SharePoint bereits im Kontext FDM bei der Implementierung verschiedener, disziplinspezifischer Use Cases an der RWTH eingesetzt. So wird die Plattform zum einen als Repositorium mit disziplinspezifischen Metadaten eingesetzt [8], und zum anderen zur Unterstützung individueller und forschungsnaher Lehre [9]. Für die Entwicklung individueller Angebote kann auf eine Vielzahl grundlegender Funktionalitäten wie zum Beispiel Nutzer- und Rechtemanagement, strukturierte Datenablage, Metadaten oder Volltextsuchen zurückgegriffen werden. Die Plattform bietet eine Vielzahl von Schnittstellen, mit denen die Funktionalitäten erweitert und an Anforderungen angepasst werden können. Durch diese flexiblen Erweiterungs- und Konfigurationsmöglichkeiten stellt SharePoint somit eine gut geeignete Basis für die Unterstützung individueller Forschungsabläufe dar.

2. Implementierung

Damit die Unterstützung von individuellen und disziplinspezifischen Forschungsprozessen als zentrale Dienstleistung skaliert muss der Softwareentwicklungsprozess entsprechend verschlankt werden. Kurze Feedback- und Weiterentwicklungs-Zyklen sind eine wichtige Rahmenbedingung. SharePoint dient dabei als „Rapid Development“-Plattform: Dabei werden die vorhandenen Strukturelemente für die Strukturierung und Bearbeitung von Inhalten wiederverwendet, um schnell spezifische Abläufe zu implementieren.

2.1. Publikationsimport aus der Hochschulbibliographie

Wissenschaftliche Publikationen haben nach wie vor den höchsten Stellenwert in der Bewertung der wissenschaftlichen Leistung. Sie sind somit ein wichtiger Bestandteil der Zusammenarbeit im wissenschaftlichen Umfeld.

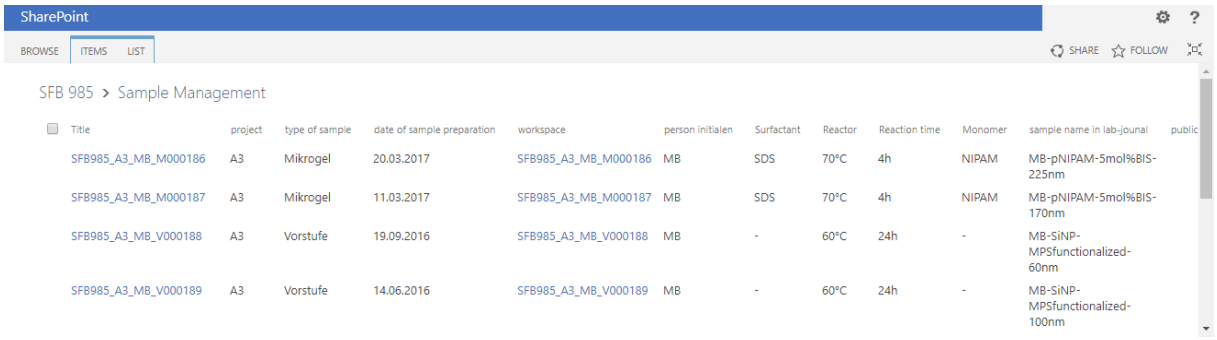
Zentrale Anlaufstelle für alle im Kontext der RWTH entstandenen Veröffentlichungen ist der Nachweis- und Publikationsserver RWTHPublications. Alle Organe der RWTH sind dazu angehalten, jeden wissenschaftlichen Output dort zu verzeichnen und mit entsprechenden Metadaten zu versehen. Neben typischen bibliographischen Metadaten erfasst RWTHPublications auch die Zuordnung zu Lehrstühlen, Instituten und Projekten an der RWTH. Alle für den SFB985 relevanten Publikationen sind somit bereits strukturiert aufgenommen.

Damit Publikationen auch innerhalb der Kollaborationsplattform sichtbar sind, werden diese automatisch synchronisiert. Über die Projektzuordnung werden nur die Publikatio-

nen übernommen, die mit dem SFB assoziiert sind. Neben der reinen Darstellung werden die Publikationen so als Elemente in der Plattform verfügbar und können, über bekannte und erprobte Implementierungen, mit anderen Inhalten in der Kollaborationsplattform verknüpft werden.

2.2. Proben und Probenmetadaten

Zur Verwaltung der Proben in der Kollaborationsplattform wird zunächst eine angepasste Liste verwendet. Diese liefert die grundlegende Datenstruktur für eine flexible Sammlung aller relevanten fachspezifischen Metadaten, die während der Synthese anfallen. Zudem ist eine teilautomatisierte Erfassung organisatorischer Metadaten wie Autoren oder Erstellungsdatum möglich, sowie die Zuordnung des eindeutigen Probenbezeichners. Abbildung 2 zeigt einen Ausschnitt der Listendarstellung im Webbrowser.



Title	project	type of sample	date of sample preparation	workspace	person initialen	Surfactant	Reactor	Reaction time	Monomer	sample name in lab-journal	public
SFB985_A3_MB_M000186	A3	Mikrogel	20.03.2017	SFB985_A3_MB_M000186	MB	SDS	70°C	4h	NIPAM	MB-pNIPAM-5mol%BIS-225nm	
SFB985_A3_MB_M000187	A3	Mikrogel	11.03.2017	SFB985_A3_MB_M000187	MB	SDS	70°C	4h	NIPAM	MB-pNIPAM-5mol%BIS-170nm	
SFB985_A3_MB_V000188	A3	Vorstufe	19.09.2016	SFB985_A3_MB_V000188	MB	-	60°C	24h	-	MB-SiNP-MPSfunctionalized-60nm	
SFB985_A3_MB_V000189	A3	Vorstufe	14.06.2016	SFB985_A3_MB_V000189	MB	-	60°C	24h	-	MB-SiNP-MPSfunctionalized-100nm	

Abbildung 2.: Screenshot des Probenmanagement mit Metadaten zur Synthese von Proben.

Das zugrundeliegende Datenmodell erlaubt eine einfache Anpassung der erfassten Metadaten durch den Forschenden selbst, sodass sich im Projektverlauf weitere Metadaten erfassen lassen. Verschiedene Feldtypen und Werteüberprüfungen in einer Excel-artigen Syntax erlauben es, die Einhaltung von Standards oder Nomenklaturen zu überprüfen und durchzusetzen.

Einen Teil der fachspezifischen Metadaten stellt die hierarchische Verknüpfung von Proben dar. Forscher können so angeben, ob eine Probe als Basis für eine andere Probe verwendet wurde und diese Beziehung direkt in den Metadaten abbilden. In der Listendarstellung im Webbrowser werden die Proben dann über einen Link miteinander verknüpft. Forschende erhalten so einen besseren Überblick über den Kontext einer Probe und können zwischen verwendeten und verwendenden Proben navigieren. Um Proben nicht nur im Labor-Kontext, sondern auch im wissenschaftlichen Kontext zu erschließen, können Forschende zudem Proben mit Textpublikationen verknüpfen, die aus der Hochschulbibliographie importiert wurden. Durch diese Verknüpfung werden auch gleichzeitig einige Metadaten der Publikation, wie zum Beispiel der Titel oder die DOI in die Metadaten der Probe mit aufgenommen.

Durch das Anlegen der Metadaten einer Probe wird automatisch ein Arbeitsbereich für die Ablage von Dokumenten und Daten erzeugt. Zur einfacheren Navigation wird dieser direkt in den Probenmetadaten verlinkt und erlaubt so ein einfaches Wechseln zwischen

den verschiedenen Ansichten. Für die Automatisierung des Ablaufs verwendet das Probenmanagement sogenannte „Event Receiver“. Mit dieser von SharePoint breitgestellten Programmierschnittstelle können eigene Programmteile definiert werden, die bei jeder Erstellung oder Änderung an einem Listenelement aufgerufen werden und ermöglichen so eine Automatisierung verschiedener Abläufe.

2.3. Arbeitsbereiche

Mit der Erstellung der Arbeitsbereiche werden Probenmetadaten automatisch aus der Probenliste auf die entsprechenden Arbeitsbereiche übertragen und synchronisiert. Änderungen der Probenmetadaten erfolgen immer in der Probenliste und werden auf den Arbeitsbereich übertragen. Für die technische Umsetzung der Arbeitsbereiche in SharePoint wird eine Dokumentenbibliothek mit Dokumentenmappen verwendet. Im Gegensatz zu den, ebenfalls in SharePoint vorhandenen, Ordnern können Dokumentenmappen mit eigenen Metadaten beschrieben werden und diese automatisch an Dokumente in den Dokumentenmappen vererben. Abbildung 3 zeigt einen Ausschnitt der Darstellung im Webbrowser.

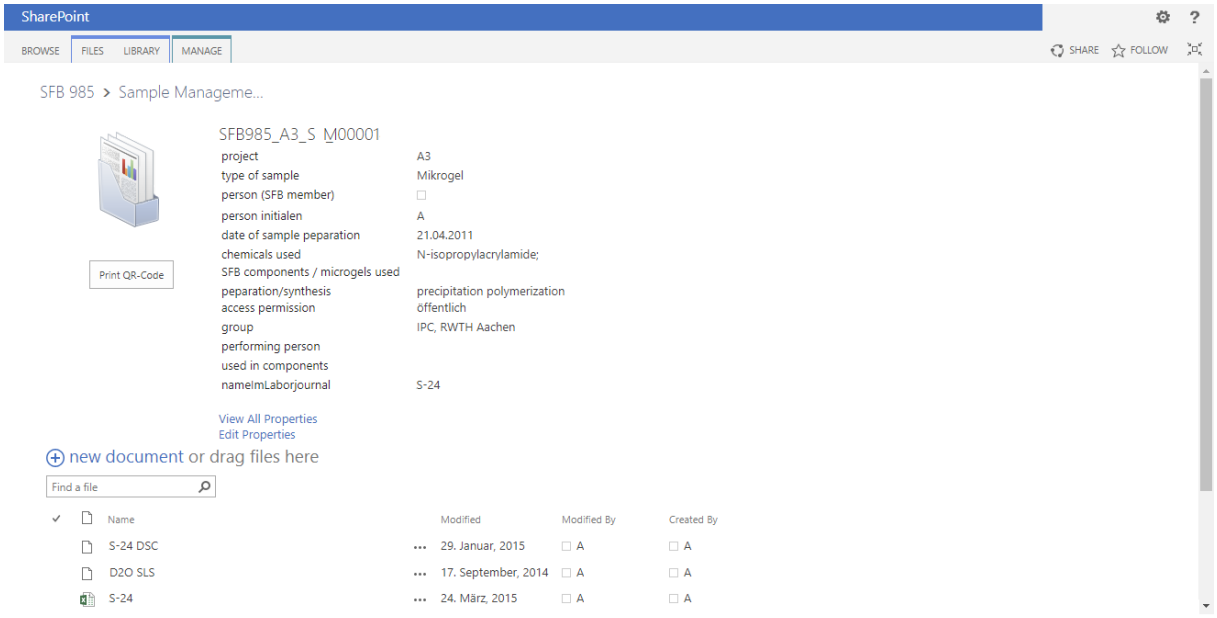


Abbildung 3.: Screenshot eines Arbeitsbereichs zur Verwaltung von Probendaten.

Abhängig von den Metadaten werden dann Zugriffsberechtigungen auf die Inhalte der Dokumentenmappen vergeben. So kann der Zugriff zunächst auf Projektbeteiligte eingeschränkt werden. Eine Freigabe der Daten auch für andere Projekte ist jederzeit zusätzlich möglich.

Innerhalb des Arbeitsbereichs haben Forschende die Möglichkeit, Dokumente oder andere Daten strukturiert zu hinterlegen. Die Dateien sind somit direkt mit der Probe assoziiert und können jederzeit auf ihren Ursprung zurückverfolgt werden. Zudem liefert die Plattform eine Versionierung für beliebige Dateitypen. Veränderungen an Daten

können so nachverfolgt und ggf. rückgängig gemacht werden. Zu den abgelegten Dateien werden automatisch technische Metadaten, wie Erstellungsdatum oder Autor erfasst. Weitere Metadaten werden für Dateien nicht manuell erfasst, sondern leiten sich über die Dokumentenmappe von den Probenmetadaten ab.

Da SharePoint ursprünglich als dokumentenbasierte Kollaborationsplattform konzipiert ist zeigen sich bei der Interaktion mit Dateien die besonderen Stärken. Insbesondere Office Dokumente (PDF, DOC, XLS, PPT) lassen sich direkt in der Plattform ansehen und betriebssystemunabhängig im Browser bearbeiten. Für die abgelegten Forschungsdaten ist dies insbesondere für Messreihen, die im Excel-Format aufgezeichnet wurden, relevant. Diese lassen sich so ohne „Medienbrüche“ durch Hoch- und Runterladen von Dateien in Abläufe im Labor- oder Forschungsalltag integrieren. Zudem ist eine Synchronisation von geöffneten Dateien auf einem Arbeitsplatzrechner, sowie gleichzeitiges Bearbeiten von Dateien mit mehreren Nutzern möglich.

Zur Übertragung größerer Datenmengen oder zur Automatisierung bietet die Plattform die Möglichkeit, die Arbeitsbereiche über WebDav einzubinden und direkt auf einem Arbeitsplatzrechner mit den dort gespeicherten Dokumenten zu interagieren. Zudem existiert mit „OneDrive for Business“ ein Synchronisationsclient, der automatische Synchronisation und Offlinekopien auf einem Arbeitsplatzrechner ermöglicht.

2.4. QR-Code Erzeugung

Für die Integration der virtuellen Arbeitsbereiche in der Kollaborationsplattform mit den Proben im Labor ermöglicht die Anwendung das Erstellen spezieller Etikettenvordrucke. Dazu wurden die SharePoint Dokumentenmappen um eine entsprechende Interaktionsmöglichkeit erweitert.

Diese Etiketten lassen sich auf Probenbehälter aufkleben und zeigen neben menschenlesbaren Metadaten, insbesondere Namen und Identifikationsnummer der Probe, auch einen QR-Code. Die Codes werden dabei über ein zweistufiges Verfahren erzeugt. Zunächst wird für die URL des Arbeitsbereichs eine Kurz-URL generiert. Diese wird dann für die Codierung des QR-Codes verwendet. Dieser Zwischenschritt ist notwendig, da die von SharePoint für die Arbeitsbereiche vergebenen URLs sehr lang sind. In diesem Fall wären QR-Codes mit höherer Auflösung zur Kodierung der Inhalte notwendig. Aufgrund der verfügbaren Fläche auf den Proben ist die Größe der QR-Codes jedoch begrenzt, so dass ein Format mit möglichst wenigen Datenpunkten vorzuziehen ist.

Die so erzeugten QR-Codes lassen sich in den Laboren und Arbeitsplätzen mit dafür vorgesehenen Lesegeräten auf einem Computer einlesen und erlauben den Forschenden effizient zum Arbeitsbereich der Proben zu navigieren. Die Probe und die zugehörigen Daten werden eindeutig verknüpft und sind von den Mitarbeitenden leicht auffindbar.

Neben der Beschriftung der Proben können die Etiketten auch für die Verknüpfung der handschriftlichen Laborbücher mit den im Arbeitsbereich abgelegten Daten genutzt werden. Über diese Verknüpfung kann direkt aus den handschriftlichen Aufzeichnungen auf die virtuelle Arbeitsumgebung verwiesen werden. Die Einbindung in den bereits bei den Forschenden etablierten Forschungsablauf ist somit gewährleistet.

3. Ausblick und Weiterentwicklung

Das Probenmanagement stellt auf Basis einer Kollaborationsplattform eine disziplinspezifische Arbeitsumgebung für den SFB985 dar. Durch die Verwendung hochschulweit genutzter und zentral bereitgestellter Infrastrukturen kann der Zugang zu den Forschungsdaten nachhaltig sichergestellt werden. Bei der Entwicklung der Anwendung lag der Fokus daher zunächst auf der Unterstützung der lokalen und individuellen Abläufe der Forschenden. Im Zuge der Weiterentwicklung soll das Probenmanagement nun an weitere zentrale Workflows und Infrastrukturen für das Forschungsdatenmanagement angebunden werden.

Obwohl Proben innerhalb der Kollaborationsplattform über die Identifikationsnummer eindeutig zuzuordnen sind, ist es im Sinne des Datenmanagements wünschenswert, dass der zu einer Probe entstandene Datensatz auch global eindeutig identifizierbar ist. Im Kontext von Forschungsdaten haben sich inzwischen verschiedene Systeme etabliert. Insbesondere für nicht veröffentlichte Datensätze scheint sich das EPIC System, das wie DOI auf dem Handle Netzwerk basiert, zu eignen [10].

Ein Grundkonzept der Kollaborationsplattform ist die Zusammenarbeit und somit die Veränderbarkeit der abgelegten Dateien innerhalb der Arbeitsbereiche. Sind Daten jedoch Grundlage für wissenschaftliche Erkenntnisse, müssen nachträgliche Modifikationen nachvollziehbar und der Stand der Daten zum Zeitpunkt der Veröffentlichung wiederherstellbar sein. Mit simpleArchive existiert an der RWTH eine Lösung zum Archivieren von Dateien, die über eine API in andere Prozesse eingebunden werden kann [11]. Daten können so in einem Versionsstand festgehalten werden, der sich über einen Identifier mit einer Textveröffentlichung verknüpfen lässt oder als Grundlage für die Publikation der Forschungsdaten dient.

Über die in vorherigen Projekten erprobte Schnittstelle zur Anbindung von domänenspezifischen Vokabularen in SharePoint [8] kann ein strukturierter Zugriff auf die im Probenmanagement gespeicherten Metadaten realisiert werden. Über eine Integration mit dem ebenfalls an der RWTH entwickelten Metadaten Manager [12] könnte so die Sichtbarkeit und Auffindbarkeit der Forschungsdaten verbessert werden.

Über solche Verknüpfungen mit disziplinübergreifenden Forschungsdatenworkflows lassen sich die von den Forschenden erhobenen Daten mittel- und langfristig im Sinne des institutionellen Forschungsdatenmanagements verwalten und können nachhaltig aufgefunden und nachgenutzt werden.

Literaturverzeichnis

[1] Rat für Informationsinfrastrukturen (RfII), „Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland.“. 2016.

[2] Realising the European open science cloud: First report and recommendations of the Commission high level expert group on the European open science cloud. Publications Office of the European Union, 2016.

[3] Prompting an EOSC in practice: Final report and recommendations of the Commission 2nd High Level Expert Group [2017-2018] on the European Open Science Cloud (EOSC). Publications Office of the European Union, 2018.

[4] Hausen, Daniela A., Ulrike Eich, Bela Brenger, Florian Claus, Benedikt Magrean, Matthias S. Müller und Elke Müller et al., „Introducing Coordinated Research Data Management at RWTH Aachen University. A Brief Project Report.“.

[5] Dreyer, Malte und Andreas Vollmer, „An Integral Approach to Support Research Data Management at the Humboldt-Universität zu Berlin.“ in Proceedings of the 22nd EUNIS Congress, hrsg. von Yiannis Salmatzidis, 319–326. Thessaloniki, Greece, 2016.

[6] Curdt, Constanze, Dirk Hoffmeister, Christian Jekel, Krischan Udelhoven, Guido Waldhoff und Georg Bareth, „Implementation of a centralized data management system for the CRC Transregio 32 'Patterns in Soil-Vegetation-Atmosphere-Systems'.“ in Proceedings of the 2nd Data Management Workshop, hrsg. von Constanze Curdt und Christian Wilmes, 27–33, Kölner Geographische Arbeiten 96. Cologne, Germany, 2016.

[7] Kirsten, Toralf, Alexander Kiel, Jonas Wagner, Mathias Rühle und Markus Löffler, „Selecting, Packaging, and Granting Access for Sharing Study Data.“ in INFORMATIK 2017: Digitale Kulturen: Beitrage der 47. Jahrestagung der Gesellschaft für Informatik e.V. (GI), hrsg. von Maximilian Eibl und Martin Gaedke, 1381–1392, GI Edition Lecture Notes in Informatics Proceedings (LNI). Bonn, Germany: Köllen, 2017.

[8] Politze, Marius und Bernd Decker, „Ontology Based Semantic Data Management for Pandisciplinary Research Projects.“ in Proceedings of the 2nd Data Management Workshop, hrsg. von Constanze Curdt und Christian Wilmes, Kölner Geographische Arbeiten 96. Cologne, Germany, 2016.

[9] Politze, Marius und Simon Consoir, „A general architecture for content driven mobile applications: building an interactive tour guide for historical sites.“ in International Conference on Education and New Learning Technologies, hrsg. von Luis Gómez Chova, Agustín López Martínez und Ignacio Candel Torres, 133–139, EDULEARN proceedings. IATED Academy, 2016.

[10] Kálmán, Tibor, Daniel Kurzawe und Ulrich Schwardmann, „European Persistent Identifier Consortium - PIDs für die Wissenschaft.“ in Langzeitarchivierung von Forschungsdaten: Standards und disziplinspezifische Lösungen, hrsg. von Reinhard Altenhöner und Claudia Oellers, 151–164. Berlin, Germany: Scivero Verl., 2012.

[11] Politze, Marius und Florian Krämer, „simpleArchive – Making an Archive Accessible to the User.“ in Proceedings of the 23rd EUNIS Congress, hrsg. von Raimund Vogl, 121–123. Münster, Germany, 2017.

[12] Politze, Marius und Florian Krämer, „Towards a distributed research data management system.“ in Proceedings of the 22nd EUNIS Congress, hrsg. von Yiannis Salmatzidis, 184–186. Thessaloniki, Greece, 2016.

Lessons learned from Virtualized Research Environments in today's scientific compute infrastructures

Dirk von Suchodoletz[1], Jonathan Bauer[1], Oleg Zharkov[1], Susanne Mocken[1] and Björn Grüning[2]

[1]Department of eScience, University of Freiburg, Germany;
[2] Department of Bioinformatics, University of Freiburg, Germany;

The Virtual Open Science Collaboration Environment project (ViCE) aimed to promote Virtualized Research Environments (VRE) to be transparently used on various research infrastructures available in Baden-Württemberg. VREs provide researchers with more freedom and flexibility using infrastructures for research and teaching ranging from high performance computing (HPC) and cloud resources to lecture PC pools. The project managed to shape new future operational models of HPC clusters and scientific clouds and to separate contradictory demands regarding software environments. The project reached varying results ranging from a rather broad uptake in the domain of the simpler virtual teaching and working environments for desktop operation compared to the more complex scientific workflows characterized by further external dependencies. Requirements like special filesystem access, a fast message passing interface or the use of special purpose hardware like graphics processing units limit the flexibility of the VRE approach to certain degrees. VREs formalize the abstraction of (complex) scientific workflows from the underlying hardware to make them more versatile, exchangeable and both archivable and reusable in the long run. Abstraction helps to complement the research data management of results and primary data sets in the future. The broader application of VREs directly relates to the business and operation models of the large scale research infrastructures in Baden-Württemberg like bwHPC and bwCloud. The gained technical flexibility is not necessarily matched to well-established financing and compensation models for the infrastructure providers.

1. Motivation

The exponential growth of computational power in the past decades has greatly contributed to scientific advances in all fields. One of the key success strategies in science is to recognize recurring patterns and exploit them via templates. First, find out which part of a problem is static or invariant – this becomes the template. Then iterate over the variable part of the problem to search for the solution. Research projects should enjoy a quick start without tedious workflows to procure and set up the necessary IT infrastructure. Especially compute resources need to scale up and down to follow the demands of

Dirk von Suchodoletz et al.: Lessons learned from Virtualized Research Environments in today's scientific compute infrastructures.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 68–81. DOI: https://doi.org/10.11588/heibooks.598.c8418

the individual project progress. At the same time, students and research assistants need to be integrated efficiently into research workflows. Virtual Machines (VM) can help by allowing prepared software environments to be copied, avoiding setting up the complete hardware, operating system, and application stack including configuration (Fig. 1). Additionally, individual researchers and workgroups should gain more flexibility to set up their own derived versions of research environments and workflows [1, 2, 43].

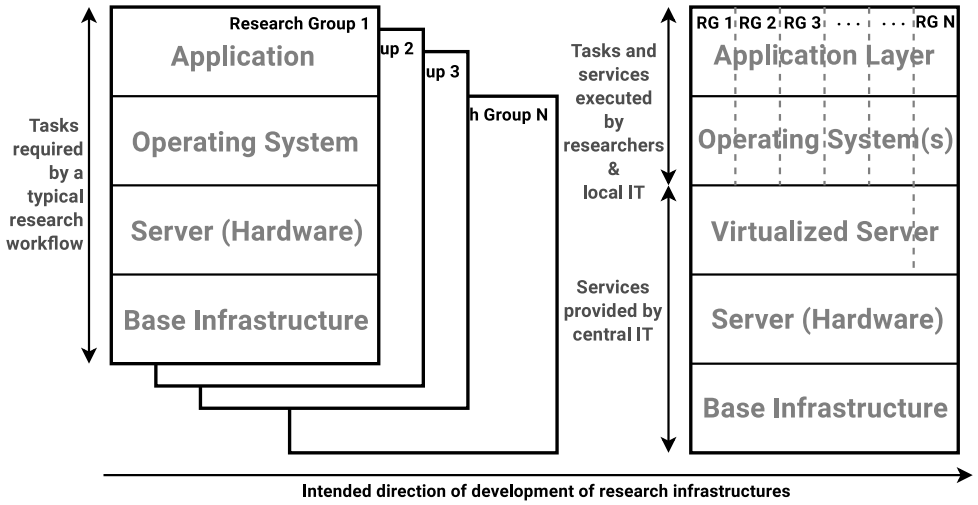


Figure 1.: Virtualization of research infrastructure helps both to provide instantaniously available resources and answer to flexible demands of each user.

The development of hardware virtualization for the x86 platform in the last two decades [4] and the cloud revolution [5, 6] also led to a paradigm shift for university computer centers. The way IT resources are provided and which services should accompany them is changing. University computer centers find themselves in the position of being pushed out of the driver's seat regarding technology development. They are now pushed by the fast technological pace set by the IT giants and big data companies.

The ubiquitous use of digitalized workflows and the Fourth Paradigm in science demand an ever-increasing amount and variety of IT-based research infrastructures. To avoid handing over sizeable proportions of infrastructure-providing activities to the commercial domain – for reasons ranging from privacy and security to expertise considerations – computer centers have to find new ways to offer a significant range of infrastructures in an efficient way [1, 14]. It should provide comparable offerings regarding features and pricing[1] as well as to avoid overextending existing personnel resources when scaling up. Demands for hardware often come up on short notice and for project periods well below the cost-amortization period of five to six years that is typical for digital equipment. Having decentralized and often duplicated personnel to select, procure and operate all the various research infrastructure components is too expensive to sustain in the long run.

[1] The term "pricing" is used in a wider sense here, as it is necessary to consider different models in basic free services, cost recovery or extension of infrastructure by bringing in project money.

Further challenges of university computer centers and faculty IT units are rooted in the very diversity of scientific communities and their broad set of demands with respect to software, tools or scientific workflows. This creates varied and often contradictory demands with respect to software environments. From the operator's point of view, it is a matter of balancing the needs of the various user groups with regard to future operating models, which can be much better represented by virtualization of resources.

University computer centers no longer offer full support as in the early days of IT and are increasingly less proficient in the specific scientific tools used in the various disciplines. Researchers from different disciplines find that the services offered by the computer center do not really fit their needs. They increasingly look at offers from the commercial sector and find services there that may also not be a perfect match either, but are cheaper or free of charge and immediately available.[2] Virtual Research Environments (VREs) can be a means and a starting point to reverse this trend. The standard services offered, such as storage or server hosting, can thus be prepared according to the target group and provide effective relief for the individual disciplines. Depending on the expected task or upcoming workflow of the individual working group or discipline, VREs are a scalable technology that relies on existing basic infrastructures of the respective data center or the responsible collaborative services. To achieve this, VREs must be technically state of the art, i.e. they must be able to handle the entire range from containerization to virtualization. Memory should be available in different forms, ranging from a fast scratch space to highly available or redundant setups, which can be integrated directly inside VREs or mediated via the host system. VREs, however, require acceptance by the respective disciplines in order to develop their full effectiveness. Due to the largely free design of the contents of a VRE, complex coordination processes are almost completely eliminated. In addition, the starting time from the project idea or approval of a research project to the first calculation and processing steps is shortened. Together with their local IT administrators, the researchers can approach their questions in a much more focused way and concentrate on content aspects.

The project ViCE – Virtual Open Science Collaboration Environment – aimed to clarify organizational questions concerning the development of sustainable business and control models for the cooperation of different expert communities with data centers on the basis of VREs. ViCE accompanied the increasing cooperation of the operator locations of the large research infrastructures like bwCloud and bwHPC and discussed possible operating and business models in this context. The project also offered the occasion to evaluate new operating models and containerization solutions in various combinations and setups [1, 43]. The project has led to joint endeavors such as the cooperation with the *de.NBI* (German Bioinformatics Network) and the grant approval for the Science Data Center *BioDATEN* in the field of bioinformatics which started in mid 2019.

[2] Many cloud service providers offer a free basic package like many Sync and Share services. Amazon has special free offers to researchers for AWS.

2. Implementation example: Bioinformatics

Bioinformatics and life sciences are fast evolving and complex fields. In contrast to physics, for example, life science research environments need to adapt nearly weekly to new specific requirements. New techniques and methods are published daily and the set of tools that need to interoperate is in the thousands. Under these circumstances, it is very challenging to maintain VREs and at the same time offer the latest methods in an accessible and reproducible way.

To address these needs the *conda* package manager was utilized, which addresses in particular the needs of a scientific community. The approach is architecture independent, programming language independent, user-space enabled, and capable of isolated virtual environments. Specifically for the bioinformatics domain the *Bioconda* project was founded, which has created more than 6600 packages over the last three years [8]. In addition, a technique that converts conda packages to containers (currently *Docker*, *rkt* and *Singularity* are supported) was developed [9]. All of those packages can be combined in complex VREs and are used by projects like *Cyverse*, *Snakemake*, *Nextflow* and *Galaxy*.

Another challenge in life sciences is that the user groups are very heterogeneous. Only a minority of users that collect data are able to setup a VM, use containers, a terminal or write a program. To address this, the European Galaxy server[3] was launched. Galaxy is a graphical web-based gateway to more than 2000 different tools, ranging from genomics and proteomics, to statistics and machine learning. The European Galaxy Server has currently 1000 active users and more than 100,000 jobs every month, which create 50 TB of data. It is supported by the BMBF funded de.NBI project, the *ELIXIR ESFRI* and the *European Open Science Cloud*.

3. Lessons learned

Facilitating virtualization revolutionized IT operations. Resource virtualization both helps to separate the requirements of different scientific user groups as well as separating hardware and operating system administration from researchers' workflows. As many resources in research infrastructures are underutilized for certain time periods, tapping into cloud strategies can help to significantly save on investment with respect to hardware resources. A welcomed by-product are the savings on rackspace and energy. VREs are a way to tap into these developments. ViCE analyzed use cases from different disciplines ranging from humanities to natural sciences to evaluate the necessary steps towards virtualization or containerization. Software and infrastructural dependencies become apparent if deployed in a VRE and provide insights into the challenges of long term access to scientific workflows and associated data, particularly with regard to system access, user management and provisioning of storage resources in the long run. The often tight ties to such network and parallel file systems need to be loosened or, even better, replaced by another technology such as object stores to become truly independent of location. Such modern day storage solutions offer token-based access management and simplify global and long-term access for researchers.

[3] Project homepage: `https://usegalaxy.eu` (visited on 20.08.2019).

Various degrees of success were achieved compared to the initial project goals. While the broad one-fits-it-all VRE is an illusion, there is a rather broad common base for general purpose hardware and service provisioning. A wide range of different use cases were adapted to and brought onto the underlying bwHPC, bwCloud and bwLehrpool infrastructures.[4] Good results with virtualized desktop environments were achieved on the bwLehrpool platform, as the dependencies on, for example, user authentication or locally mounted network shares were rather low. More complex VREs featuring core scientific workflows required additional considerations and adaptations [43]. As research data management gained momentum, a better understanding of the correlation of data and scientific workflows needed to be developed. Reproducible scientific results require reproducible environments. Either VREs in the form of VMs or containers need to be kept functioning over longer periods of time or VREs need to be defined in a declarative and reproducible manner with tools like Ansible, Packer and Kickstart while using Jenkins for continuous integration [2].

An ongoing challenge, in a wider sense, is the handling of sensitive data on shared resources like HPC and cloud infrastructures. The requirements of the implemented data protection ruling are to be honored. ViCE discussed data management issues stemming from the handling of sensitive data. It quickly became clear, however, that comprehensive organizational processes were required to master this task, for example by certifying the underlying infrastructure. For many infrastructure providers, it will be necessary to accommodate research projects with sensitive data. Thus, the formalization of infrastructure operations in adherence to the European General Data Protection Regulation becomes inevitable. As a result of these findings and a growing number of requests, a certification of the de.NBI cloud infrastructure is envisioned within the coming two years.

3.1. Integration of special purpose hardware

Special purpose hardware like GP-GPUs (general purpose graphics processing units), Infiniband or Omni-Path (both low latency, high bandwidth compute node interconnects) infrastructures are not easily virtualized as limitations regarding hardware and software support still exist. Therefore, such resources are not easily available from inside VREs and cannot be trivially shared among VREs running on a single host system, although there are a couple of ways to dedicate such resources to single VM instances [10]. Nevertheless, a fully virtualized VRE is less dependent on the existence of hardware components and thus easier to share and move across different host systems.

To allow the sharing of GPU resources within a tier 3 HPC cluster like NEMO,[5] a Docker or Singularity container was created which allows direct access to the necessary hardware and to the parallel file system at the same time. PCI passthrough is one of the options to allow VMs to access hardware in the host system, albeit exclusively.

[4] These large scale research and teaching infrastructures are provided via various state-sponsored or co-financed past and ongoing projects. Background information is e.g. available from [11]. Additional information on the use cases is found at `https://www.forschungsdaten.info` (visited on 20.08.2019) within the ViCE project pages.

[5] The bwForClusters NEMO is hosted in Freiburg and part of a state wide federated HPC research infrastructure.

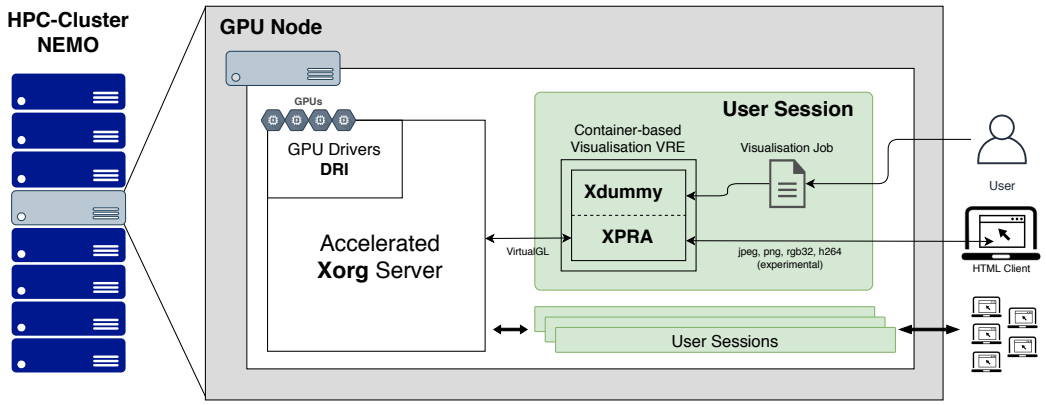


Figure 2.: During the last phase of ViCE a special purpose VRE for remote visualization of research results near to the location of the (large scale) data got implemented and put to production in NEMO.

Nevertheless, it can help to share a well-equipped GPU node among different users and their respective software environments. Up to now, the experiments with Docker and Nvidia GPUs demonstrated a couple of kernel and driver challenges as software versions need to be tightly matched in the host and Docker containers, reintroducing dependencies meant to be overcome by virtualization in the first place.

A use case for a VRE accessing and using a GPU for both rendering of data and creation of a remote interactive viewing stream was implemented for a microsystems technology working group for remote visualization of large data sets (Fig. 2). It would have been impractical to copy files of up to several Terabytes just for viewing portiongs of the data and then discarding the file. The setup is much more flexible than operating on the separate user desktops as it can easily accessed from different machines. At the moment each viewing instance requires it's own GPU as shared access would require a special license.

3.2. Resource sharing and scaling

Resource abstraction by virtualization or containerization facilitates the sharing of resources and is thus economically and organizationally attractive. The flexibility of the underlying infrastructures is either inherently available, in e.g. the bwCloud and bwLehrpool services, or was added, e.g. to NEMO. This allows the fast integration of new user groups into the existing research infrastructures, which previously operated their own compute and storage systems with considerable effort. Scientists can participate in larger infrastructures by investing a comparably small amount of funds. As such, larger infrastructures usually have fluctuating loads, it is possible to first accommodate new users and evaluate futher investments at a later point in time. Such researchers can start their work much faster than if they had to define, tender, install and administer the complete software and hardware stacks themselves. Such consolidations help research institutions through more efficient use of resources and dissemination of modern concepts.

In the commercial world, the cloud is a "pay-as-you-go" business model which is not

applicable to the university domain. A comparably clear approach could be implemented if just money for hardware is brought in, as was done with the de.NBI cloud, where a proportional share is dedicated to the shareholder. The models of bwUniCluster and bwForClusters could also be seen as a suitable basis for the discussion of operational and business models for VREs. The additional funds required for larger requirements are collected in the rectorates and at the universities of applied sciences according to the known pattern [12]. Even an operating costs levy would be conceivable. The refinancing of the hardware could also be cushioned by a 143c co-financing. The necessary personnel will be paid in the existing infrastructure locations from a bwHPC-S5 that may be extended. If, comparable to bwHPC, consumption billing is largely dispensed with and a certain fair-share factor is applied instead, the administrative effort can remain comparatively moderate. The question of control still needs to be clarified, which may not have been optimally handled by the previous state user committee due to the sometimes significantly different user structures. The cooperation necessary for a comprehensive use of VREs will generally not be carried out without control, whereby the life cycle of the services involved introduces a further level to be considered. So far, decision-making on continuation, change or discontinuation has hardly been carried out offensively. This also applies to the provision of platforms for VREs.

3.3. External dependencies

One of the resources a research or teaching project might need in significant quantities is storage. Data and the software for the various scientific workflows often do not live together on the same machine but are brought together by some data infrastructure. Often e.g. home directories, source and destination shares or software module collections are mounted from a central resource and secured by defining IP ranges to which an export is allowed. If used in a VRE, especially on top of different resources at different sites or if meant to be shared among different colleagues in a distributed group, this option is no longer suitable. A similar problem arises from latencies if the shared resource is not available from within the hosting site but a couple of network hops away.

If a scientific research environment runs as a virtualized instance right from the start, the local hardware dependency is loosened and the subsequent relocation of the environment to a new virtualization platform is significantly facilitated. For example, a scientist could develop simulation software on the local desktop in a container or a VM, simplifying debugging by interactively testing algorithms and processes with direct feedback. The successfully tested simulation can then be scaled up by running on a cloud or in a cluster. The resources required for this can be provided by data centers in the form of a compute cloud. Virtualization right from the start ensures that the VM image can be copied directly into the cloud environment and that one or more virtual instances appropriate to the problem can be started. At the same time, the image can be made available to cooperation partners for direct use or adaptation to one's own problem as well as to third parties for verification of the workflow or can be used as part of a course. This makes it easier for young scientists to start productive research, as they no longer have to spend their time installing operating systems and software packages without any guarantee of success.

3.4. Network storage

The options for the delivery of input data and the storage of output is dependent on the amount of data processed in each step and the bandwidth between storage and computation resources provided. The VRE use case developed together with the CMS group of particle physicists at KIT [43] read the input from a locally provided CVMS proxy[6] and wrote the data back over the network to the dedicated storage system at the KIT. Thus, only a small amount of local disposable scratch data was required in the VRE.

A different approach evaluated was the deployment of *SDS@hd* [13] to VREs. This state-wide service offers storage for (federated) research projects in Baden-Württemberg.[7] The service specifies that the storage is intended for data in active use, not for long-term storage or backups. This means that it could be beneficial in use cases with the requirement of permanent storage – cloned or parallel projects requiring access to shared data or a space to save their results.

Conveniently, SDS@hd also offers an existing test project that can be quickly and easily connected to in order to determine if the service is suitable for a specific project or in the given infrastructure. For a productive use of this service, an entitlement must be granted: first an entitlement by the institute, then a request for a specific amount of storage with justification must be submitted; after receiving provisional approval, a contract must be signed and submitted before the allocation can be approved. This process has to be completed once for every storage project, but once it is done the project owner can easily invite other users to the project.

Once approved, the storage project could be accessed by various methods – SSHFS access was easy and instantly available using a password of one's own choosing; NFSv4 access required quite a bit of human interaction (providing personal data and information regarding the machine that would be used to make the connection) in order to generate a keytab for access; SMB is also an option, but was not tested in the course of the ViCE Project.

Higher latencies with jitter usually hurt the performance of traditional file systems. Having heard complaints of slow data transfers with SSHFS as a potential negative outweighing the ease of connection, several tests were run to compare performance. While initial results confirmed the assumption that NFS would be faster, further tests were run using different ciphers, resulting in comparable results using both connection types. Tests showed that from bwCloud to the SDS@hd storage project, NFS was able to handle writes faster than SSHFS, and the inverse was true for reads. Thus, a good understanding of the usage patterns for each project and some preparation at setup time can pay off in the longer term for a project with intensive reads or writes.

4. Cooperation

The diffusion of IT in almost all scientific disciplines and the increasing digitalization of formerly non-technical workflows in research has increased considerably, which is reflected

[6] Before the use of the proxy, the amount of data copied over the network was significant.
[7] Subsidised by the university for researchers in Heidelberg, at a fee for external users.

in both qualitatively and quantitatively increased demands on local IT support and the respective computer centers. However, their structures, both in terms of size and orientation of personnel and financial resources, do not necessarily grow with the wishes of users and their needs. To a certain extent, this means that it will be more difficult to add new services to the catalog if the portfolio of services has grown. At this point, cooperation projects, in which new services can be jointly provided and offered in a network without all participants having to assign their own personnel, offer a possible way out.

The direct and continuous contact between the computer centers and the researchers is a necessary prerequisite for a better planning of the basic offers of computer centers and for responding to the requirements of the researchers in the best possible way. Such communication structures can be ensured by project-accompanying or topic-related governance structures. State-sponsored projects such as ViCE point the way here to moderating the introduction process of novel services for individual scientists or research groups. They show how research can overcome IT-related limitations in time and space and find new forms of division of labor and cooperation. Necessary basic infrastructures of the computer centers are prepared in such a way that they can easily be integrated by different disciplines and without delay.

4.1. Organizational challenges

After two and a half years of ViCE project duration, the organizational hurdles proved to be greater than the technical ones. Although there is one (or more) clear business model(s) for the implementation of cooperation with mutual service provision and settlement,[8] there is a lack of coordinated activities in this direction.[9] Very well-equipped – in terms of hardware, software and personnel – projects such as bwHPC work thanks to generous funding. In the case of rather simple structures such as bwLehrpool, where two partners provide the services for all others and one institution is responsible for billing, it took quite a long time for the desired legal framework to be created.

The current situation is characterized by the fact that more or less all IT projects initiated by the ALWR-BW and funded by the Ministry of Science, Research and the Arts, Baden-Württemberg independently try to find answers to the questions of sustainability and cost allocation (Fig. 3). When this is seriously attempted, it ties up considerable resources of project personnel, who often have little expertise in this field. The ViCE project is no exception. The big step to set up a company, registered society, or association of any kind for the handling of project and additional tasks (related to the classic data-center business) is discussed and dismissed regularly.[10]

[8] Paal et. al. [14] present the general options; examples of ongoing cooperative projects with financial compensation in one form or another are outlined in [11].

[9] See the discussion in the report `https://www.forschungsdaten.info/typo3temp/secure_downloads/67417/0/c5e104aa380decfca16882404c15bf6ab4eeeaaa/BetriebsmodelleForschInfra.pdf` (visited on 20.08.2019).

[10] For a more in-depth elaboration, refer to the report `https://www.forschungsdaten.info/typo3temp/secure_downloads/67417/0/c5e104aa380decfca16882404c15bf6ab4eeeaaa/GeschaeftsmodelleForschInfra.pdf` (visited on 20.08.2019).

However, from the point of view of individual projects, implementation is not pursued further as it is deemed clearly too costly from a limited resource project's perspective. The necessity of considering how future achievements are to be described, provided and invoiced, becomes in the authors' view ever more urgent. While the necessity for the development of corresponding country concepts in the area of service allocation initially existed with some key projects, the country concepts in other areas (such as HPC) developed their own dynamics, justifying an expansion of considerations in these areas. The aim of such considerations should therefore be to create a concept for cross-national governance structures that organize both the handling of additional funds and the burden of sharing between the institutions, since developments at the state level clearly point in this direction. This also applies to all other state cooperation projects – since pure direct exchange of services cannot reflect the complexity – as well as to ongoing projects and concepts. However, if this aspect is not tackled with the same vigour and effort, it is to be feared that the good or very good position of the country in the medium and long term will be endangered. The locations are becoming increasingly hesitant about the question of whether certain projects should be implemented cooperatively.

From the point of view of the operators, it also became clear during the various workshops and training courses that the topic of *Secure or data-protected compute infrastructures* is becoming increasingly important. Special challenges arise when dealing with personal data in both HPC and cloud environments.[11] This also applies to the secure storage of such data records. In the meantime, requests have been received from a number of fields. The next logical step is to start a certification process for the involved infrastructures and services.[12]

4.2. Cross-institutional compensation

With regard to the classic university computer center operation, there is no very close link between payment and service provision. The existing flat-rate model of data centers clearly reaches its limits here and, in extreme cases, leads to misplanning and misallocation of resources (services continue to be operated because employees want them and not so much because there is a significant demand). Therefore there must be mechanisms in the fast-changing technological framework in which data centers operate to determine how this change can be carried out. In the discussion with project partners and participants on the provider side, it became clear that the players' expectations of legal security and long-term predictability in cooperation are increasing. These developments lead to a situation in which cooperation in the medium term must move towards a common set of values with long-term common goals and ideas. This is an evolution of the short-term common interest groups that gathered to acquire project funding in the first place. Central to this development is the mutual trust of the partners and the commitment, overarching the necessary level of personal individual contacts.

[11] This was not an initial criterion at the start of the project, but there were increasing inquiries from researchers of various domains and corresponding requirements.

[12] More science funders start to require a certain certification to handle sensitive data.

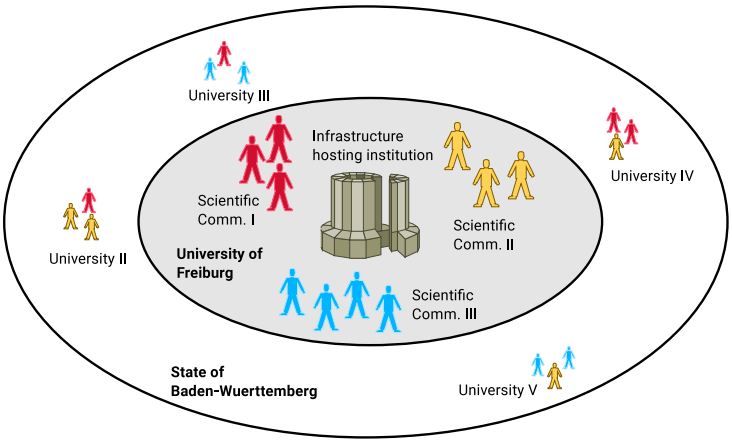


Figure 3.: While it is possible to negotiate compensation and distribution of costs among different stakeholders within a single university it is getting much more tedious even within the context of a single state.

A couple of questions came up during the project runtime illustrating the cross-institutional financial considerations and challenges: Do special funds professors receive when negotiating their new position – which typically come directly from the budget of the institution hiring that person and are usually invested locally at the same site – go directly to another location to expand the infrastructure that person wants to use? Who finances the basic equipment costs (ranging from facility and energy to personnel) to complement the money raised for individual projects? What could be the distribution key for cost compensation: efficiency gains through centralization vs. locally incurred resource costs? What about the formation of new (large scale) research infrastructures which are of mutual interest for more than one local scientific community?

A substantial need for clarification exists with respect to applicable operation and business models (Fig. 3). If long-term "large solutions" are the goal, such as the establishment of an association, non-profit limited, registered society or similar [14], definition of long-term goals and (often) appropriate political support are required. For shorter and smaller projects, achievement exchanges between partners might be sufficient. When considering these different formal forms of organization for cooperation, it becomes clear that cooperation and committee structures do not differ much. They are determined by the cooperation partners and their agreed goals. The "policy" in one form or another, be it national policy or the promotion of certain developments through project lines, was perceived by the actors as an essential factor.

5. Conclusion

ViCE was mainly active in the areas of virtual research environments and research data management. Thanks to ViCE, it has been possible to bundle the essential methods for selected specialist communities in a VRE and to connect them with other infrastructures in order to achieve largely seamless access to the data and storage of (interim) results. VREs make it easier for young scientists to gain access to existing large-scale infrastruc-

tures without having to submit their own time-consuming funding applications (Fig. 4). VREs can contribute to the improvement of teaching: in this regard, an increased use of virtualized teaching and working environments could be achieved. VREs provide more freedom and flexibility for scientists when using the provided infrastructures for research and teaching such as bwHPC, bwCloud or the bwLehrpool lecture PC pools. The wider introduction of VREs into the scientific workplace leads to a redistribution of tasks: researchers focus on the application side whereas the computer centers provide scalable research infrastructures. In addition to the level of scientific applications, a number of organizational issues, including broad technical access to federal infrastructures, were and are still to be clarified. The discussion on operating and business models could be advanced, as well as considerations regarding financing. However, it became apparent that the political level also has to be involved in order to pursue sustainable approaches.

The interdisciplinary character of the project's approach became particularly clear in the use cases of physics, bioinformatics and with the creation of a common corpus access by the english studies/computer linguistics. E-science environments of this kind require a new assessment of existing infrastructures, as was demonstrated by the integration of remote data sources and sinks. From the scientists' point of view, they must be easily and reliably available, which is only possible to a limited extent with conventional storage systems. Further efforts are therefore necessary, which will be tackled within the framework of bwHPC-S5, bwSFS and within the Science Data Center project BioDATEN for bioinformatics which started in July 2019.[13]

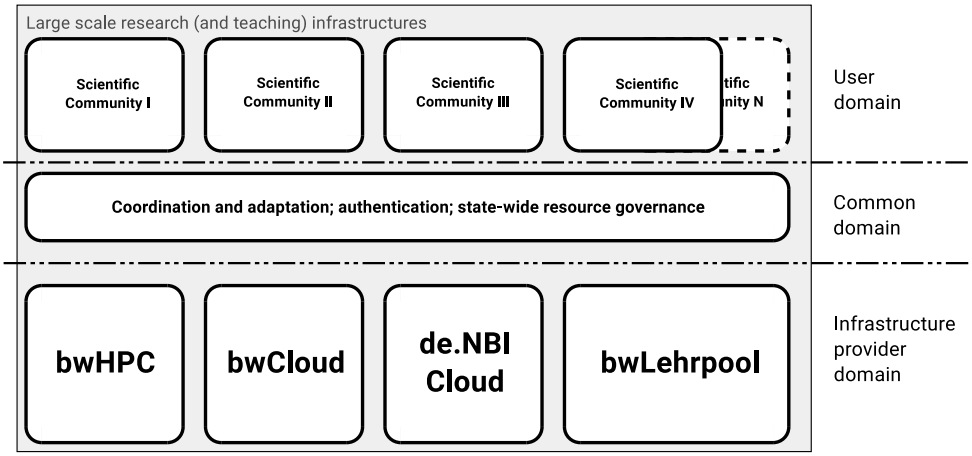


Figure 4.: Virtual research environments allow an easy mapping of multiple scientific communities to multiple (federated) large scale research infrastructures.

Standardized services and virtualized infrastructures are easy to use across sites and scale when supporting more communities (Fig. 4). To achieve this goal, cooperation is key: proper (legal, financial) frameworks for inter-institutional exchange are required. Above the base layer of services the various scientific communities expect more diverse software stacks on the middle layer, which provides less common ground for (widely) shared VREs than expected at the project's beginning. The project thus changed its approach towards

[13] See `http://www.biodaten.info` (visited on 20.08.2019).

the registry from a project specific implementation [2, 15] to a more general approach reusing existing established solutions [13]. To mitigate the involved changes in the course of the project, comprehensive information was provided for the participating scientific communities and later beyond, with the aim of identifying new ways of using existing large-scale research infrastructures and eliminating access obstacles. This included information and guidance on research data management. By using containerization and packaging, ViCE discussed the basics for long-term access to research environments characterized by data and processes.

Acknowledgement

The work outlined in this publication was done during the ViCE project funded by the Ministry of Science, Research and the Arts, Baden-Württemberg, Germany. The support is gratefully acknowledged.

Bibliography

[1] Konrad Meier, Björn Grüning, Clemens Blank, Michael Janczyk, and Dirk von Suchodoletz. Virtualisierte wissenschaftliche Forschungsumgebungen und die zukünftige Rolle der Rechenzentren. In *10. DFN-Forum Kommunikationstechnologien, 30.-31. Mai 2017, Berlin, Gesellschaft für Informatik eV (GI)*, pages 145–154, 2017.

[2] Jonathan Bauer, Dirk von Suchodoletz, Jeannette Vollmer, and Helena Rasche. Game of templates. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 245–262. TLP, Tübingen, 2019.

[3] Felix Bührer, Frank Fischer, Georg Fleig, Anton Gamel, Manuel Giffels, Thomas Hauth, Michael Janczyk, Konrad Meier, Günter Quast, Benoît Roland, Ulrike Schnoor, Markus Schumacher, Dirk von Suchodoletz, and Bernd Wiebelt. Dynamic Virtualized Deployment of Particle Physics Environments on a High Performance Computing Cluster. *Computing and Software for Big Science*, 2018.

[4] Amir Ali Semnanian, Jeffrey Pham, Burkhard Englert, and Xiaolong Wu. Virtualization technology and its impact on computer hardware architecture. In *2011 Eighth International Conference on Information Technology: New Generations*, pages 719–724. IEEE, 2011.

[5] Pradip K. Sarkar and Leslie W. Young. Sailing the cloud – a case study of perceptions and changing roles in an australian university. In *ECIS*, page 14, 2011.

[6] Marios D Dikaiakos, Dimitrios Katsaros, Pankaj Mehra, George Pallis, and Athena Vakali. Cloud computing – distributed internet computing for it and scientific research. *IEEE Internet computing*, 13(5):10–13, 2009.

[7] Dirk von Suchodoletz, Janne Chr. Schulz, and Jan Leendertse. Abstraktion erlaubt neue Aufgabenverteilung – Virtualisierung, Clouds und die zukünftige Rolle wissenschaftlicher Rechenzentren. *Wissenschaftsmanagement*, 4:31–35, 2017.

[8] Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, and Johannes Köster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7):475–476, jul 2018.

[9] Björn Grüning, John Chilton, Johannes Köster, Ryan Dale, Nicola Soranzo, Marius van den Beek, Jeremy Goecks, Rolf Backofen, Anton Nekrutenko, and James Taylor. Practical computational reproducibility in the life sciences. *Cell Systems*, 6(6):631–635, jun 2018.

[10] Vishakha Gupta, Ada Gavrilovska, Karsten Schwan, Harshvardhan Kharche, Niraj Tolia, Vanish Talwar, and Parthasarathy Ranganathan. GViM: GPU-accelerated virtual machines. In *Proceedings of the 3rd ACM Workshop on System-level Virtualization for High Performance Computing*, pages 17–24. ACM, 2009.

[11] Dirk von Suchodoletz, Janne Chr. Schulz, Jan Leendertse, Hartmut Hotzel, and Martin Wimmer, editors. *Kooperation von Rechenzentren Governance und Steuerung – Organisation, Rechtsgrundlagen, Politik*. de Gruyter, 2016.

[12] Dirk von Suchodoletz, Stefan Wesner, and Gerhard Schneider. Überlegungen zu laufenden Cluster-Erweiterungen in bwHPC. In Dirk von Suchodoletz, Janne Chr. Schulz, Jan Leendertse, Hartmut Hotzel, and Martin Wimmer, editors, *Kooperation von Rechenzentren: Governance und Steuerung – Organisation, Rechtsgrundlagen, Politik*, pages 331–342. De Gruyter, 2016.

[13] Martin Baumann, Vincent Heuveline, Oliver Mattes, Sabine Richling, and Sven Siebler. SDS@hd–Scientific Data Storage. In Jens Krüger and Thomas Walter, editors, *Proceedings of the 4th bwHPC Symposium October 4th, 2017, Alte Aula Eberhard Karls Universität Tübingen*, pages 32–36, 2017.

[14] Dirk von Suchodoletz, Janne Chr. Schulz, Jan Leendertse, Hartmut Hotzel and Martin Wimmer. Vorbetrachtungen. In Dirk von Suchodoletz, Janne Chr. Schulz, Jan Leendertse, Hartmut Hotzel, and Martin Wimmer, editors, *Kooperation von Rechenzentren: Governance und Steuerung – Organisation, Rechtsgrundlagen, Politik*, pages 315–329. De Gruyter, 2016.

[15] Christopher B. Hauser and Jörg Domaschka. ViCE Registry: An Image Registry for Virtual Collaborative Environments. In *2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 82–89, 2017.

[16] Sarah Berenji Ardestani, Carl Johan Hakansson, Erwin Laure, Ilja Livenson, Pavel Stranák, Emanuel Dima, Dennis Blommesteij n, and Mark van de Sanden. B2share: An open escience data sharing platform. In *2015 IEEE 11th International Conference on e-Science (e-Science)*, pages 448–453. IEEE, 2015.

Welche Unterstützungsangebote benötigen Disziplinen für die systematische Verankerung von E-Science in der universitären Forschung und Lehre? Perspektiven und Anforderungen am Beispiel der Germanistischen Linguistik

Michael Beißwenger[1], Hubert Klüpfel[2], Ania López[3] und Stephanie Rehwald[3]
[1]Institut für Germanistik, Universität Duisburg-Essen;
[2]Referent des CIO Universität Duisburg-Essen;
[3]Universitätsbibliothek, Universität Duisburg-Essen

In diesem Beitrag diskutieren wir Anforderungen und Perspektiven für die Unterstützung von E-Science an Universitäten. Am Beispiel der Germanistischen Linguistik skizzieren wir die Bedeutung der Arbeit mit digitalen Daten und -infrastrukturen in Forschung und Lehre und leiten daraus Anforderungen ab, die sich bei der Integration entsprechender Ressourcen in den Forschungsprozess, in die Lehre und in die Ausbildung des wissenschaftlichen Nachwuchses ergeben. Wir geben einen Überblick über außeruniversitäre Initiativen und Netzwerke, die für die Disziplin relevante Infrastrukturen und Services bereitstellen oder solche entwickeln. Diese Angebote können die in dieser und anderen Disziplinen bestehenden Anforderungen bislang aber nur zu einem Teil abdecken. Für die Universitäten stellt sich die Frage, inwieweit sie – ergänzend zu vorhandenen Initiativen und Netzwerken – die Dissemination und Weiterentwicklung von E-Science als Institution aktiv mitgestalten möchten und welche Unterstützungsmaßnahmen sie dazu im Rahmen ihrer Aufgaben bereitstellen können. Am Beispiel der Universität Duisburg-Essen zeigen wir, wie Unterstützungsmaßnahmen, insbesondere auch mit Blick auf die gegenwärtig in Aufbau befindlichen Forschungsdateninfrastrukturen auf nationaler (NFDI) und auf Länderebene, konzipiert und weiter ausgebaut werden können.

1. Einleitung

In diesem Beitrag diskutieren wir Perspektiven und Strukturen für die Unterstützung von E-Science an der Universität Duisburg-Essen (UDE). Am Beispiel der Germanistischen Linguistik und mit Bezug zu konkreten Projekte beschreiben wir die Anforderungen, die sich bei der Einbindung digitaler Forschungsressourcen und -methoden in Forschungsprozesse und in die Lehre stellen. Wir geben einen groben Überblick, welche außeruniversitären Initiativen und Unterstützungsangebote existieren, auf die Forschende und Lehrende für die Arbeit mit digitalen Ressourcen zurückgreifen können. Wir skizzieren, welche

Michael Beißwenger et al.: Welche Unterstützungsangebote benötigen Disziplinen für die systematische Verankerung von E-Science in der universitären Forschung und Lehre? Perspektiven und Anforderungen am Beispiel der Germanistischen Linguistik. In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 –Data to Knowledge, Heidelberg: heiBOOKS, 2020, 82–94. DOI: https://doi.org/10.11588/heibooks.598.c8419

Unterstützungsangebote an der Universität Duisburg-Essen – wie vermutlich auch an anderen Hochschulen – bereits vorhanden sind und wie diese die Anforderungen aus dem Fach bedienen können. Davon ausgehend diskutieren wir, welche Rolle die Universität, ergänzend und komplementär zu existierenden außeruniversitären Angeboten, in Bezug auf die Weiterentwicklung von E-Science in der universitären Forschung und Lehre einnehmen könnte. Speziell im Zusammenhang mit der im Aufbau befindlichen Nationalen Forschungsdateninfrastruktur (NFDI) und dazu korrespondierenden Initiativen auf Landesebene werden die Universitäten nicht umhinkommen, ihr Selbstverständnis und ihre Strategie in Bezug auf die Förderung und Verbreiterung von E-Science in den Disziplinen zu bestimmen.

2. Bedeutung von digitalen Forschungsressourcen für die Linguistik

Die Germanistische Linguistik greifen wir in unserem Beitrag als ein Paradebeispiel für eine Disziplin heraus, in der digitale Forschungsressourcen und Infrastrukturen die Möglichkeiten und Methoden empirischer Forschung in den letzten zwei Jahrzehnten substanziell verändert haben. Die Arbeit mit digitalen Ressourcen und Werkzeugen ermöglicht es, wissenschaftliche Hypothesen mit computergestützten Methoden an dokumentierten Stichproben zum Sprachgebrauch zu erkunden sowie qualitativ und quantitativ zu überprüfen. Digitale Infrastrukturen, über die diese Ressourcen und Werkzeuge genutzt werden können und die unter Beteiligung von Forscher*innen im Fach weiter ausgebaut werden, bilden daher „eine wichtige und in der Zukunft noch wichtigere Säule für das Fach“ (Hinrichs 2018: 47).

Als Datengrundlage für linguistische Forschungen spielen u.a. sog. *Korpora*eine wichtige Rolle. Darunter versteht man Sammlungen authentischer Sprachdaten (mündlich, schriftlich, multimodal), die für Zwecke der linguistischen Analyse aufbereitet sind. Typische Aufbereitungsschritte sind z. B. die Anreicherung der Daten um *Metadaten*, die für die Analyse benötigt werden (bei Texten z. B. Angaben zum Autor, zur Textsorte, zum Erscheinungsort und -jahr, zu möglichen Vor- und Nebenversionen; bei Gesprächen und Daten aus internetbasierter Kommunikation z. B. Angaben zu den Interaktionsbeteiligten und zu ihren sozialen Beziehungen, zu ihrem Alter und Bildungsgrad, zu verwendeten Sprachen, zum Zweck und zur Vorgeschichte der Interaktion usw.), sowie das Hinzufügen von linguistischen Annotationen (z. B. Wortart, Flexionsform, syntaktische Struktur) zu den enthaltenen Wörtern, syntaktischen Aufbaueinheiten und Sätzen. Mit Annotationen wird es möglich, bei der Korpus-Recherche und -analyse über die reine Volltextsuche hinaus nach sprachlichen Mustern und Strukturen zu suchen (Lemnitzer & Zinsmeister 2015; Lüdeling/Kytö 2008/2009).

Eine besondere Rolle spielen sog. *Referenzkorpora*. Das sind Korpora, die nicht nur in der Eigenforschung des Korpus-Erstellers eine Rolle spielen, sondern die der Scientific Community als Grundlage für beliebige sprachbezogene Forschungen bereitgestellt werden. Forschung, die auf Grundlage solcher Korpora durchgeführt wird, ist nachprüfbar, weil die Daten offen zugänglich sind. Eine Vernetzung der Forschungsergebnisse mit

den verwendeten Ressourcen wird möglich. Dadurch, dass unterschiedliche Forschungsfragen auf denselben Referenzkorpora durchgeführt werden, wird es möglich, Ergebnisse zu unterschiedlichen Forschungsfragen über den Bezug auf dieselben Daten zueinander in Beziehung zu setzen und sie zu vergleichen. Dadurch können neue Forschungsfragen entstehen.

Große, aufbereitete Referenzkorpora und Korpus-Sammlungen, wie sie z. B. vom Leibniz-Institut für Deutsche Sprache (IDS), Mannheim, mit dem Deutschen Referenzkorpus DeReKo (Lüngen 2017) und der Datenbank Gesprochenes Deutsch (Schmidt 2017) oder von der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) über die Korpusplattform des Projekts www.dwds.de (Geyken et al. 2017) angeboten werden und die über Online-Schnittstellen abgefragt werden können, bieten Wissenschaftler*innen für viele Domänen des Sprachgebrauchs bereits eine gute Ressourcenlage. Für andere, noch unzureichend abgedeckte Domänen wird die Ressourcenlandschaft in diversen Projekten weiter ausgebaut. Eine solche Domäne ist aktuell die Sprachverwendung in Social-Media-Anwendungen bzw. in der internetbasierten Kommunikation (Beißwenger et al. 2017, Beißwenger 2018). Neben Korpora spielen im Bereich der Linguistik aber auch noch andere digitale Sprachressourcen (DSR) eine Rolle – zum Beispiel digitale Wörterbücher und Nachschlagewerke zur Sprache (z.B. Grammatiken), Sammlungen historischer Texte, lexikalisch-semantische Netze oder Werkzeuge für die automatische Sprachanalyse (Tokenisierer, Part-of-Speech-Tagger, Lemmatisierer, syntaktische Parser u.a.). Als sehr produktiv für den Aufbau und die Aufbereitung von Sprachkorpora und anderen digitalen Sprachressourcen hat sich die Zusammenarbeit mit dem Bereich Sprachtechnologie/Computerlinguistik erwiesen, der Verfahren für die automatische linguistische Analyse und für die Annotation von Sprachdaten entwickelt. Nationale und internationale Initiativen unterstützen den Ausbau und die Bereitstellung von Sprachressourceninfrastrukturen im Bereich der Linguistik und der Geisteswissenschaften.

3. Welche Anforderungen stellen sich in der Disziplin?

In diesem Abschnitt skizzieren wir Anforderungen an die Arbeit mit digitalen Sprachressourcen (DSR) in der universitären Forschung und Lehre im Fach Germanistik.1 Dazu unterscheiden wir zunächst sechs verschiedene Perspektiven, unter denen DSR für Wissenschaftler*innen im Fach im Zusammenhang mit ihren Aufgaben in Forschung und Lehre relevant werden. Welche dieser Perspektiven für den oder die einzelne*n Wissenschaftler*in tatsächlich praxisrelevant sind, hängt ab vom Stellenprofil und von individuellen thematischen Schwerpunkten. Begreift man die Arbeit mit DSR und mit darauf bezogenen Methoden als einen wichtigen Baustein einer modernen, empirisch ausgerichteten Linguistik, so darf angenommen werden, dass für einzelne Wissenschaftler*innen jeweils mehrere der Perspektiven in der Forschungs- und Lehrpraxis eine Rolle spielen. Für an der Universität tätige Linguist*innen ist der Umgang mit digitalen Sprachressourcen (DSR) unter unterschiedlichen Perspektiven relevant:

1. Als Forschende nutzen sie DSR als Werkzeuge im eigenen Forschungsprozess.

2. Als Lehrende vermitteln sie Möglichkeiten der Arbeit mit DSR an Studierende, um diese mit modernen empirischen Methoden des linguistischen Arbeitens vertraut zu machen.

3. Als Hochschullehrer*innen bilden sie wissenschaftlichen Nachwuchs aus, für dessen Karriere ein reflektierter Umgang mit DSR unerlässlich ist.

4. Als empirisch Forschende leiten sie Forschungsprojekte, in deren Rahmen Sprachdaten erhoben werden, um diese linguistisch zu analysieren. Forschungsförderer wie z. B. die DFG erwarten, dass in Projekten erhobene Forschungsdaten bzw. Sprachkorpora nachnutzbar vorgehalten werden, und geben dafür datentechnische und rechtliche Standards vor (vgl. DFG 2015a, 2015b).

5. Ggf. sind sie auch an Forschungs- und Entwicklungsprojekten beteiligt, deren genuines Ziel im Aufbau von DSR besteht, die eine Lücke in der Korpuslandschaft schließen und der Scientific Community als Forschungsressourcen bereitgestellt werden sollen.

6. Als Akteur*innen im Bereich ‚Digital Humanities' und/oder als Mitglied von Fachgesellschaften beteiligen sie sich ggf. an der Weiterentwicklung von Standards für die Arbeit mit DSR und am Ausbau entsprechender Infrastrukturen.

Diese Perspektiven dürften sich *mutatis* mutandis auch auf die Arbeit mit digitalen Ressourcen in anderen geisteswissenschaftlichen Disziplinen übertragen lassen. Je nach Perspektive ergeben sich dabei unterschiedliche Anforderungen an Unterstützungsangebote, die benötigt werden, wenn man die Weiterentwicklung der damit beschriebenen Aktivitäten in Forschung und Lehre als ein strategisches Entwicklungsziel des Faches und der das Fach einbettenden Institutionen (Universität, Fachgesellschaften, Forschungsdateninfrastrukturen) begreift. Die folgenden Anforderungen lassen sich – als erweiterbare Liste – formulieren; der Fokus liegt auf der Arbeit mit Korpora.

1. Nutzung von DSR im Forschungsprozess: Benötigt werden Schnittstellen und Werkzeuge, anhand derer existierende DSR flexibel für die Nutzung in unterschiedlichen linguistischen Forschungskontexten angepasst werden können. Aus DSR für Forschungszwecke extrahierte Daten sollten, auch nach weiterer Aufbereitung, jederzeit auf die entsprechenden Teile der Ausgangsressource rückbeziehbar sein. Projektspezifisch vorgenommene Erweiterungen von Ausschnitten bestehender Ressourcen (z. B. zusätzliche linguistische Annotationen) sollten so zur Verfügung gestellt werden können, dass sie für die Scientific Community als Erweiterungen der Ausgangsressource nutzbar sind.

2. Arbeit mit DSR in der universitären Lehre: DSR sollten mit Nutzerschnittstellen ausgestattet sein, die es auch Semi-Experten der Domäne nach entsprechender Einarbeitung erlauben, in vollem Umfang von den Abfrage- und Auswertungsmöglichkeiten Gebrauch zu machen. Um die Potenziale von DSR für die empirische Forschung praxisnah zu vermitteln, ist die Bereitstellung didaktisch aufbereiteter Praxisbeispiele für Forschungsprojekte auf unterschiedlichen Niveaustufen wünschenswert.

3. Ausbildung des wissenschaftlichen Nachwuchses: Wünschenswert sind regelmäßige Weiterbildungsangebote, z. B. auf Ebene von Fächergruppen oder Fakultäten, zu existierenden DSR und zu Methoden der Ressourcen-gestützten linguistischen Forschung, in denen gängige Werkzeuge, Standards und Verfahren für die Modellierung und Annotation von Textdaten sowie für die Abfrage und computergestützte Auswertung von Korpora vorgestellt werden. Solche Weiterbildungen sind nur als fachnah konzipierte Angebote sinnvoll denkbar. Expertise zu Standards, Werkzeugen und Infrastrukturen im Bereich Digital Humanities sowie personelle Ressourcen für deren Vermittlung ist für eine nachhaltige Verankerung solcher Angebote unabdingbar.

4. Selbst (z. B. im Rahmen von geförderten Projekten) erhobene Forschungsdaten nachnutzbar aufbereiten und im Rahmen etablierter Forschungsinfrastrukturen bereitstellen: Hier besteht Bedarf an Beratung und Unterstützung, um in Projekten erhobene Daten in Repräsentationsformate zu konvertieren, die für die Langzeitarchivierung in Repositorien und ggf. für die Bereitstellung als DSR in existierenden Forschungsinfrastrukturen als Standards etabliert sind. Erfahrungsgemäß ist hierzu Beratung bereits in der Konzeptionsphase der Projekte erforderlich, damit idealerweise schon vor der Erhebung der Projektdaten Modellierungsentscheidungen getroffen werden können, die eine nachhaltige Repräsentation erlauben. Auch rechtliche und forschungsethische Fragen sind entsprechend frühzeitig im Erhebungsprozess zu behandeln, wenn nach Abschluss des Projekts eine Archivierung und/oder Weitergabe an Dritte vorgesehen (oder vom Projektträger erwünscht) ist. Um die Integration von Projektdaten in existierende übergreifende Infrastrukturen (z. B. CLARIN) zu organisieren, sollten Projektleiter mit den entsprechenden Zentren in Kontakt gebracht und sollten technische Fragen der Datenübergabe sowie der vor- oder nachgängigen Datenaufbereitung durch eine Supportstelle mit Expertise im Bereich Digital Humanities unterstützt werden.

5. Aufbau und Erweiterung des DSR-Ressourcenlandschaft: Der Aufbau neuer DSR und die Beteiligung am Ausbau von Forschungsdateninfrastrukturen sowie die damit zusammenhängenden konzeptionellen und technischen Aufgaben sollten als Forschungsleistungen – beispielsweise im Rahmen von Ziel- und Leistungsvereinbarungen – gewürdigt werden. Technische Unterstützungsmaßnahmen – z. B. die Bereitstellung projektspezifisch eingerichteter Server – und konzeptionelle Services – z. B. Unterstützung bei der Entwicklung geeigneter Datenhaltungsstrukturen – ist in der Entwicklungsphase neuer DSR erforderlich, da in den Geisteswissenschaften typischerweise kein Personal für solche Aufgaben vorhanden ist. Nach Abschluss der Entwicklungsarbeiten, der Datenerhebung und Aufbereitung sollten die resultierenden DSR dann in übergreifende Forschungsdateninfrastrukturen überführt werden (siehe 4.).

6. Beteiligung an der Weiterentwicklung von Standards: Die Weiterentwicklung von Standards für die Arbeit mit DSR und am Ausbau entsprechender Infrastrukturen sollte als Forschungsleistung – beispielsweise im Rahmen von Ziel- und Leistungsvereinbarungen – gewürdigt werden.

4. Existierende Infrastrukturen und Netzwerke außerhalb der Universitäten

Die Dissemination der Nutzung von DSR und von Werkzeugen für die computergestützte Analyse von DSR für die empirische linguistische Forschung und für eine zeitgemäße wissenschaftliche Ausbildung von Studierenden und von wissenschaftlichem Nachwuchs wird u.a. in Initiativen wie CLARIN-EU[1]bzw. CLARIN-D[2] und DARIAH[3] vorangetrieben. Auch verschiedene Fachgesellschaften widmen sich dem Thema mit eigenen Sektionen – so beispielsweise die Sektion „Computerlinguistik“ in der deutschen Gesellschaft für Sprachwissenschaft (DGfS)[4], der Forschungsfokus „Digitale Infrastrukturen für die Angewandte Linguistik“ (GAL-DIAL) in der Gesellschaft für Angewandte Linguistik (GAL e.V.)[5] sowie verschiedene Arbeitskreise in der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL e.V.)[6].

Der Förderung einer Vernetzung zwischen Sprachressourcenanbietern und linguistischen Nutzer*innen verpflichtet sind diverse Aktivitäten (internationale Workshops, Master Classes, Summer Schools) von CLARIN-ERIC[7] zum Thema „User Involvement“. Mit der Entwicklung eines Curriculums für die „Digitalen Geisteswissenschaften“ beschäftigt sich u. a. die AG „Referenzcurriculum Digital Humanities“ in der Fachgesellschaft Digital Humanities im deutschsprachigen Raum (DHd)[8].

Das deutsche CLARIN-Verbundprojekt CLARIN-D ist organisiert in Zentren als Pfeilern einer aufzubauenden, im europäischen Kontext eingebundenen, nationalen Sprachressourcen- und Forschungsinfrastruktur. Die Zentren stellen Forschenden DSR und darauf bezogene Services bereit und kuratieren DSR in Hinblick auf die Langzeitarchivierung und Bereitstellung für die Scientific Community. Mit der Förderung der „Nationalen Forschungsdateninfrastruktur“ (NFDI) durch Bund und Länder wird eine Bündelung und Koordination bestehender und neuer Aktivitäten im Forschungsdatenmanagement angestrebt. Wie in Abbildung 1 gezeigt, liegt die NFDI quer zu den vorhandenen Strukturen und folgt dem Leitgedanken, Konsortien zu bilden, die verschiedene Säulen des Wissenschaftssystems miteinander vernetzen. Wie bei dem Aufbau der European Open Science

[1] CLARIN: Common Language Resources and Technology Infrastructure. https://clarin.eu,besucht:26.04.2019.

[2] CLARIN-D: Deutsches Konsortium im Rahmen von CLARIN-EU:https://www.clarin-d.net,besucht:05.04.2019.

[3] DARIAH: Digital Research Infrastructure for the Arts and Humanities.https://dariah.eu,besucht:26.04.2019.

[4] Deutsche Gesellschaft für Sprachwissenschaft / Sektion Computerlinguistik, 2019. https://dgfs.de/de/cl/, be-sucht: 26.04.2019.

[5] Gesellschaft für Angewandte Linguistik / Forschungsfokus Digitale Infrastrukturen für die Angewandte Linguis-tik (GAL-DIAL):https://gal-ev.de/gal-forschungsfokus-digitale-infrastrukturenfuer-die-angewandte-linguistik-gal-dial/, besucht: 26.04.2019.

[6] Arbeitskreise in der Gesellschaft für Sprachtechnologie und Computerlinguistik: https://gscl.org/sigs, besucht: 04.05.2019.

[7] CLARIN-ERIC:https://www.clarin.eu/, besucht: 04.05.2019.

[8] DHd-AG Referenzcurriculum Digital Humanities: https://dig-hum.de/ag-referenzcurriculum-digital-humanities, besucht: 26.04.2019.

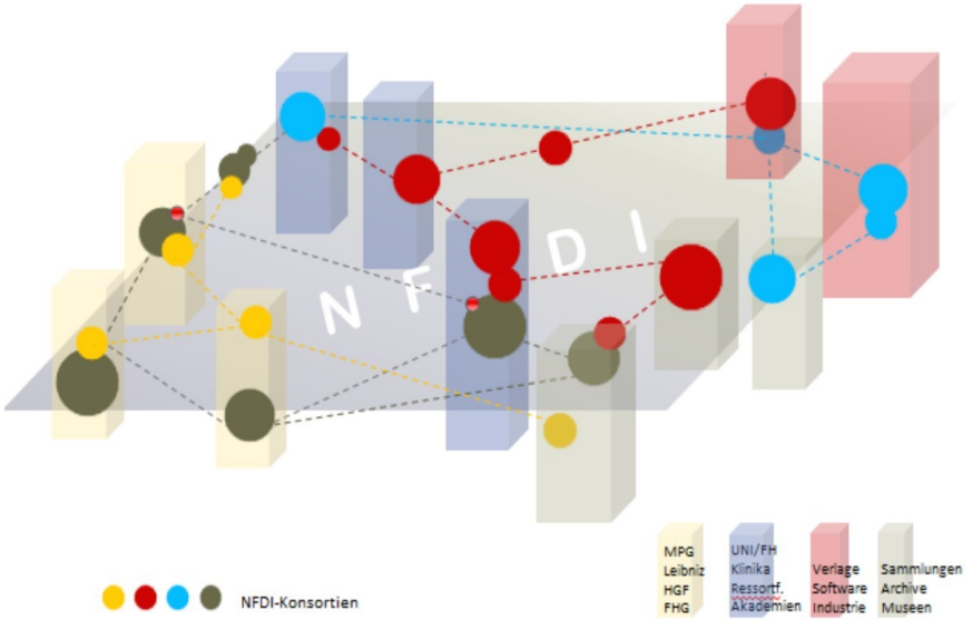


Abbildung 1.: Die Nationale Forschungsdateninfrastruktur als Querschnittsstruktur. Die NFDI besteht aus Konsortien, die sich aus Akteuren und Institutionen außeruniversitärer und Hochschulforschung, Industrie und weiteren Bereichen bilden (Quelle: Wambsganß, 2019, Folie 21).

Cloud[9] spielen die fachlichen Communities für den Aufbau der Nationalen Forschungsdateninfrastruktur eine zentrale Rolle: Die Vernetzung erfolgt thematisch und wird daher stark von den in den Fachcommunities diskutierten und entwickelten Standards und Curricula beeinflusst.

Abseits des wissenschaftspolitischen Prozesses rund um die NFDI haben einige Bundesländer landesweite Aktivitäten zu Forschungsdatenmanagement (FDM) zu verzeichnen (Grasse/López/Winter 2018). Im Land Nordrhein-Westfalen werden die entsprechenden Aktivitäten in der Landesinitiative NFDI[10] gebündelt. Diese nimmt eine Scharnierfunktion zwischen Landes- und Bundesaktivitäten wahr. Die wesentlichen Aufgaben der Landesinitiative sind die Vernetzung relevanter Stakeholder an Hochschulen und außeruniversitären Forschungseinrichtungen in NRW, sowie die Begleitung und Initiierung von hochschulübergreifenden Aktivitäten im Kontext von FDM, sodass mittel- bis langfristig NRW-weite Empfehlungen und Lösungen für FDM etabliert und darüber hinaus in einer nationalen Gesamtstrategie erfolgreich eingebracht und platziert werden können (Curdt, 2018). Als Ansprechpartnerin für Hochschulleitungen und Infrastrukturpartner im Bereich FDM, betreibt die Landesinitiative Trendscouting und sorgt für nationale sowie internationale Vernetzung. Aktuell befindet sich die Phase II für den Zeitraum 2019 bis 2021 in Planung und wird einen speziellen Fokus auf die Weiterbildung im Bereich E-Science und Forschungsdaten legen.

[9] `https://www.egi.eu/about/newsletters/what-is-the-european-open-science-cloud/`
[10] Landesinitiative NFDI der Digitalen Hochschule NRW. `https://fdm-nrw.de/`, besucht: 26.04.2019.

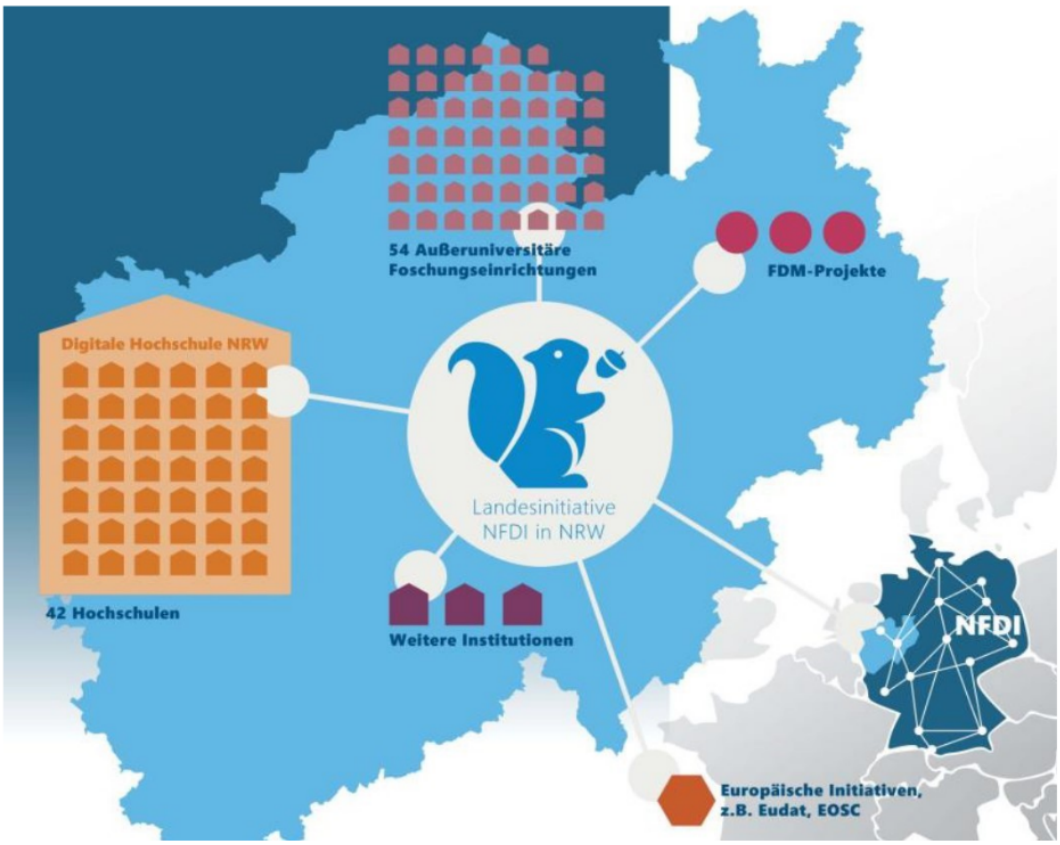


Abbildung 2.: Die Landesinitiative Nationale Forschungsdateninfrastruktur der Digitalen Hochschule NRW bietet Unterstützungsangebote für die Hochschulen und Forschungseinrichtungen in NRW im Kontext von Forschungsdatenmanagement und Fragen zur NFDI.

5. Überlegungen zur Rolle der Universitäten bei der Förderung und Weiterentwicklung von E-Science in den Disziplinen

Für die Universitäten stellt sich die Frage, ob sie die Bereitstellung entsprechender Angebote vor allem als eine Aufgabe von Fachgesellschaften und der Scientific Community bzw. einer nationalen Forschungsdateninfrastruktur ansehen oder ob sie die Dissemination und Weiterentwicklung von E-Science als Institution aktiv mitgestalten möchten. Hier besteht ein enger Zusammenhang mit der Schaffung von Rahmenbedingungen für wissenschaftliche Innovationen.

Bei den oben genannten sechs Perspektiven wurde am Beispiel der DSR schon deutlich, dass der Zugang zu bereits bestehenden Angeboten nicht ohne weiteres gegeben ist. Dies ist weitgehend übertragbar auf Forschungsdatenmanagement in anderen Fachdisziplinen. Die einzelne Wissenschaftlerin steht, gerade am Anfang ihrer Karriere, vor der Aufgabe, sich mit ihren spezifischen Forschungsinteressen und Karrierezielen in den in Abschnitt

2 skizzierten Perspektiven zu verorten und die Rolle von E-Science-Angeboten wir DSR im Rahmen der eigenen Aufgaben in Forschung und Lehre zu bestimmen. Die Universität Duisburg-Essen – als Beispiel – hält dazu bereits eine Reihe von Unterstützungsangeboten in den zentralen Einrichtungen Bibliothek (UB), Rechenzentrum („Zentrum für Informations- und Mediendienste“ (ZIM)) und Forschungsförderung („Science Support Center“ (SSC)) bereit.

Eine zentrale Rolle sieht die Universität dabei derzeit in Unterstützungsangeboten im Bereich Forschungsdatenmanagement (FDM) und baut hierzu die Servicestelle „Research Data Services“ (RDS) auf (siehe Abbildung 3),Das Konzept betont die zentrale Bedeutung von wechselseitigem Erfahrungsaustausch und Wissensaufbau zwischen den Wissenschaftler*innen der UDE und den Infrastruktureinrichtungen, um gezielt Unterstützungsangebote weiterzuentwickeln und mit bestehenden Initiativen der Fachcommunity zu verknüpfen. Im Rahmen eines kollegialen Kompetenznetzwerks soll der Austausch zwischen Wissenschaftler*innen untereinander sowie mit den Research Data Services dauerhaft etabliert werden.

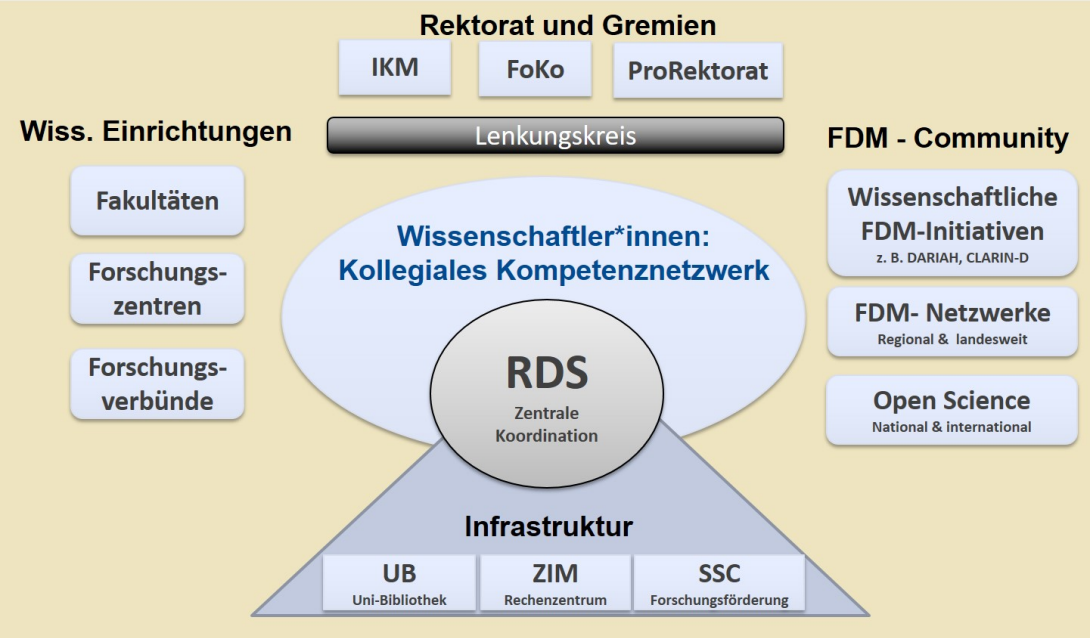


Abbildung 3.: Konzept eines Unterstützungsangebots „Research Data Services“ (RDS) an der Universität Duisburg-Essen (in Planung).

An dieser Stelle sei darauf hingewiesen, dass der Begriff E-Science in verschiedenen Fächern mit unterschiedlicher Schwerpunktsetzung verwendet wird und mit unterschiedlichen Anforderungen verknüpft ist. So ist z.B. das High Performance Computing (HPC) ein Bereich von E-Science, der in der Germanistischen Linguistik bislang noch keine prominente Rolle spielt, der in Bereichen insbesondere in den Natur- und Ingenieurwissenschaften hingegen von zentraler Bedeutung ist. Dort wo der Messaufbau und das analysierte Ergebnis für den Forschungsprozess meist wichtiger als die Rohdaten selbst sind, werden diese i.d.R. auch nicht aufbewahrt.

Im Gegensatz dazu sind in den Geisteswissenschaften oft die Rohdaten singulär oder originär in dem Sinne, dass sie durch Wiederholung nicht oder nur mit hohem Aufwand reproduziert werden könnten und dass einmal erhobene und mit sprach- und korpustechnologischen Verfahren aufbereitete Daten einen großen Wert darstellen, weil sie die Grundlage für anspruchsvolle empirische Forschung – im Falle von Referenzkorpora sogar für ganz unterschiedliche Forschungsfragen in einer Vielzahl von Projekten – bilden können. In Bezug auf die Germanistische Linguistik lassen sich aus den oben skizzierten Anforderungsbereichen Unterstützungsnahmen ableiten, die die Universität im Rahmen der RDS bieten bzw. anbieten könnte.

1. eine Lotsenfunktion übernehmen, die Forschende (insbesondere auch den wissenschaftlichen Nachwuchs) an für ihre Forschungsvorhaben geeignete Forschungsdateninfrastrukturen heranführt und sie mit entsprechenden Ansprechpartner*innen, Netzwerken und Zentren in Kontakt bringt;

2. die Verankerung von DSR in der Lehre unterstützen (unter anderem die Dokumentation von „Good Practices“, die für die weitere Dissemination genutzt werden können);

3. die Einbindung entsprechender Trainings- und Weiterbildungsmaßnahmen in universitäre Strukturen zur Förderung des wissenschaftlichen Nachwuchses wie Graduiertenkollegs, Graduate Centers initiieren und fördern;

4. Beratung in Bezug auf Forschungsinfrastrukturen und digitale Ressourcen – möglichst nah zu den Anforderungen im jeweiligen Fach bzw. Fächercluster – bei der Vorbereitung von Drittmittelanträgen anbieten;

5. technische und konzeptionelle Unterstützung für Forscher*innen bieten, die sich als Teil ihrer Forschungsarbeit am Auf- und Ausbau von digitalen Forschungsressourcen und -infrastrukturen beteiligen; und

6. das Engagement von Wissenschaftler*innen in Netzwerken und Fachgesellschaften unterstützen, die sich außeruniversitär mit der Weiterentwicklung von E-Science (Infrastruktur, Standards, Good Practices, Curricula) beschäftigen.

Dabei werden selbstverständlich vorhandene zentrale Infrastruktureinrichtungen, insbesondere die Bibliothek, das Rechenzentrum (ZIM)und die Forschungsförderung (SSC) eine wichtige Rolle spielen und durchgängig einzubinden sein. Gleichzeitig gilt: die Aufgaben beim Forschungsdatenmanagement gehen über das traditionelle Aufgabengebiet der Bibliotheken und Rechenzentren hinaus. Das gilt ganz besonders für die *Digital Humanities*, bei denen das *Computing* wie oben erläutert nicht im Mittelpunkt steht (Hinrichs, 2018). Wie bereits eingangs postuliert, sehen wir die Universitäten vor der Aufgabe, ihre Strategie zur E-Science in den Disziplinen weiterzuentwickeln – das Beispiel der Germanistische Linguistik zeigt, wie sehr die Umsetzung der eng verwobenen Anforderungen an Wissenschaftler*innen und Infrastruktureinrichtungen auf ein etabliertes Netzwerk mit gemeinsamen Kompetenzaufbau angewiesen ist.

6. Zusammenfassung und Ausblick

Die Möglichkeiten der Arbeit mit digitalen Daten, Ressourcen und Infrastrukturen verändern einerseits Forschungsprozesse und führen andererseits zu neuen Anforderungen in Bezug auf die Planung und Abwicklung von Forschungsprojekten, auf den Umgang mit und die Pflege von Forschungsdaten sowie auf die Vermittlung der für eine zeitgemäße, digital gestützte Forschung benötigten Kompetenzen an Studierende und Nachwuchswissenschaftler*innen. Am Beispiel der Germanistischen Linguistik haben wir den Stand der Arbeit mit digitalen Ressourcen und Infrastrukturen dargestellt und daraus eine Reihe von Anforderungen abgeleitet, die sich Forschenden und Lehrenden stellen, wenn sie entsprechende Ressourcen und die darauf bezogenen Methoden nachhaltig in Forschung und Lehre einsetzen möchten. Die Germanistische Linguistik diente uns dabei als eine Beispiel-Disziplin; die skizzierten Anforderungen lassen sich mit den erforderlichen Anpassungen sicherlich auf andere geisteswissenschaftliche Fächer übertragen. In anderen Disziplinen, etwa in den MINT-Fächern und im Bereich Medizin, mögen sich die Anforderungen in Bezug auf E-Science anders darstellen; auch hier gibt es aber Bedarfe, die für eine nachhaltige Verankerung der Arbeit mit digitalen Ressourcen und Infrastrukturen stellen.

Die Universitäten sind gefordert, ihr Selbstverständnis und ihre Rolle in Bezug auf die Förderung und Weiterentwicklung von E-Science in den Disziplinen zu entwickeln – im Sinne einer Bereitstellung von Rahmenbedingungen für wissenschaftliche Innovationen. Am Beispiel der Universität Duisburg-Essen haben wir skizziert, welche Unterstützungsangebote bereits existieren und wie diese weiterentwickelt werden können. Sehr sinnvoll erscheint es uns, den Prozess der Weiterentwicklung entsprechender Unterstützungsangebote unter Einbezug von Akteur*innen aus den Disziplinen bzw. Fakultäten zu gestalten („user Involvement“), um zwischen dem institutionell Realisierbaren und dem aus Sicht von Forschung Lehre Wünschenswerten einen guten und passgenauen Kompromiss zu finden und dabei geeignete Schnittstellen zu außeruniversitär existierenden Infrastrukturen, Services und Initiativen zu definieren. Die Förderung von E-Science und der nachhaltige Auf- und Ausbau von Forschungsdateninfrastrukturen ist eine grundsätzlich überuniversitär – auf nationaler und internationaler Ebene – zu leistende Aufgabe der Forschungsförderung (vgl. z. B. Hinrichs 2018). Die Anbindung universitärer Forschung und Lehre an diese Infrastrukturen und die Verankerung ihrer Nutzung in den Disziplinen kann aber sinnvoll nur auf Ebene der Universitäten geleistet werden. Dazu sind einerseits technische und konzeptionelle Unterstützungsangebote erforderlich, die möglichst nahe an den Bedarfen einzelner Disziplinen und Disziplinen-Cluster ansetzen, andererseits Anreizsysteme, um die Dissemination der Potenziale digitaler Forschungsinfrastrukturen und deren Nutzung in den Disziplinen nachhaltig zu verankern.

Literaturverzeichnis

[1] Beißwenger, Michael (2018): Internetbasierte Kommunikation und Korpuslinguistik: Repräsentation basaler Interaktionsformate in TEI. In: Henning Lobin, Roman Schneider and Andreas Witt (Hrsg.): Digitale Infrastrukturen für die germanistische Forschung. Berlin/New York: de Gruyter 2018 (Germanistische Sprachwissenschaft um 2020, Bd. 6), 307-349. Open-Access-Publikation: https://doi.org/10.1515/9783110538663-015

[2] Beißwenger, Michael; Chanier, Thierry; Erjavec, Tomaž; Fišer, Darja; Herold, Axel; Lubešic, Nikola; Lüngen, Harald; Poudat, Céline; Stemle, Egon; Storrer, Angelika; Wigham, Ciara (2017): Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries. In: Lars Borin (Ed.): Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure (Linköping University Electronic Conference Proceedings 136), 1-18.

[3] Curdt, Constanze, Grasse, Marleen, Hess, Volker, Kasties, Nils, López, Ania, Magrean, Benedikt, Winter, Nina. (2018). ZUR ROLLE DER HOCHSCHULEN - Positionspapier der Landesinitiative NFDI und Expertengruppe FDM der Digitalen Hochschule NRW zum Aufbau einer Nationalen Forschungsdateninfrastruktur. Zenodo. http://doi.org/10.5281/zenodo.1217527 DFG (2015a): Handreichung: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora. www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

[4] DFG (2015b): Handreichung: Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora.https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf

[5] Grasse, M., López, A. and Winter, N., (2018). Landesinitiative NFDI – a Central Point of Contact for RDM for Higher Education Institutions in the German State of North Rhine-Westphalia. Data Science Journal, 17, p.25.DOI:http://doi.org/10.5334/dsj-2018-025

[6] Geyken, Alexander, Adrien Barbaresi, Jörg Didakowski, Bryan Jurish, Frank Wiegand &Lothar Lemnitzer (2017): Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS). In: Zeitschrift für germanistische Linguistik 45 (2), 327–344.

[7] Hinrichs, Erhard (2018): Digitale Forschungsinfrastrukturen für die Sprachwissenschaft. In: Lobin, H. (Hrsg.): Digitale Infrastrukturen für die germanistische Forschung, S. 81–133. De Gruyter, Berlin.

[8] Lemnitzer, Lothar & Heike Zinsmeister (2015): Korpuslinguistik. Eine Einführung. 3., überarbeitet und revidierte Auflage. Tübingen: Narr (Narr Studienbücher).

[9] Lüdeling, Anke; Kytö, Merja (2008/2009): Corpus Linguistics. 2 Bde. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 29).

[10] Lüngen, Harald (2017): DeReKo – Das Deutsche Referenzkorpus. Schriftkorpora der deutschen Gegenwartssprache am Institut für Deutsche Sprache in Mannheim. In: Zeitschrift für germanistische Linguistik 45 (1), 161–170.

[11] Schmidt, Thomas (2017): DGD – die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim. In: Zeitschrift für germanistische Linguistik 45 (3), 451-463.

[12] Wambsganß, Joachim (2019): Rat für Informationsinfrastrukturen (RfII) - Einblicke in aktuelle Themen. und Arbeiten. E-Science-Tage Heidelberg, Keynote.

Transforming data silos into knowledge: Early Chinese Periodicals Online (ECPO)

Matthias Arnold[1] and Lena Hessel[2]

[1]Heidelberg Research Architecture - Heidelberg Centre for Transcultural Studies, Universität Heidelberg;

[2]Institute of Chinese Studies, Universität Heidelberg

Abstract

This paper introduces the project "Early Chinese Periodicals Online (ECPO)" [1]. ECPO joins several important digital collections of the early Chinese press and puts them into a single overarching framework. In a first phase, several databases on early women's periodicals and entertainment publishing were created: "Chinese Women's Magazines in the Late Qing and Early Republican Period" (*WoMag*), "Chinese Entertainment Newspapers" (*Xiaobao*), and databases hosted at the Academia Sinica in Taiwan. These systems approach the material in two ways: in the *intensive approach* we record all articles, images, advertisements, and related agents and assign them to a complete set of scanned pages, while in the *extensive approach* we record the main characteristic features of publications.

ECPO has begun to join these various materials in a second, ongoing phase of the project. Today, ECPO provides open access to 267 publications comprising over 280.000 pages of print. A key aspect is to make entire issues available, front-to-back, including illustrations, advertisements, and even blank pages. For 138 publications we also provide descriptions of individual items in Chinese with Pinyin transcription. These records also contain genre and column information, basic content analysis, as well as names and roles of agents associated with an item.

Our new cross-database agent service allows us to manage the approximately 47.000 names recorded in *WoMag* and ECPO: we a) merge identical names across databases, b) identify agents and assign names to them, and c) link agent records to authority data (GND, VIAF, Wikidata, Baidu, DBpedia). Besides creating a curated list of agents occurring in the publications, we also aim to add missing persons to authority files like the GND.

One crucial aspect of ECPO is full text capability. Unfortunately, OCR software cannot be used out-of-the-box, for a number of reasons: document analysis fails to recognize complex newspaper layout, character recognition fails when it faces emphasis marks next to characters, and recognized passages have to be grouped in the right semantic order.

Das hier beschriebene Paper ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00027325` veröffentlicht.

The paper will discuss approaches to further exploring and analyzing the knowledge hidden in these publications, together with efforts to open the collection's data for reuse. We demonstrate workflows in the Agents service and cross-database record curation. We also present results from a crowdsourced approach to newspaper segmentation to generate segments that can easier be OCRed. In addition, we introduce first ideas to create a module for encoding text in TEI and relate it to the database.

1. Introduction

Research data can occur in various forms and formats. In the humanities, the systematic collection of data driven by a set of distinct research questions very often is a challenge on its own: The research data collection becomes a major part of the research data output. This is also the case for the project we introduce in this paper:[1]"Early Chinese Periodicals Online (ECPO)"[1].

The project aims at a systematic examination of the Chinese periodical press during the first four decades of the 20th century. Chinese periodicals of this era remain understudied, even though they dominated the contemporary print market and provide access to what Raymond Williams has called "actual culture".[2] They present researchers with a number of challenges: They are physically dispersed and often poorly preserved, voluminous, multi-generic, and intellectually demanding. We approached these challenges in two ways. Firstly, we formed a multidisciplinary research team. Through a number of international interdisciplinary projects the research team engaged in over the past decade[3], we have developed a new methodology for approaching materials from the Chinese popular press.[4] Secondly, we created a set of unique digital resources developed by the Heidelberg Research Architecture (HRA) and located at the Heidelberg Centre for Transcultural Studies (HCTS). These resources formed the basis for conceptualizing and implementing a database prototype, ECPO -Early Chinese Periodicals Online.

ECPO is distinguished from other existing databases of Chinese periodicals in that it not only provides image scans, but also preserves materials often excluded in reprint, microfilm or digital editions, such as advertising inserts and illustrations. Our workflow is even designed to establish a definitive page sequence for citations and references. In addition, it incorporates a sophisticated body of metadata in both English and Chinese,

[1] This paper was made possible with funding from the Heidelberg Centre for Transcultural Studies and The Center of East Asian Studies at Heidelberg University. The authors wish to thank Prof. Barbara Mittler (Heidelberg University) and Prof. Joan Judge (York University) for their continuous support and invaluable input.

[2] Williams (1961) p.70.

[3] "A New Approach to the Popular Press in China: Gender and Cultural Production, 1904-1937" funded by the Canadian SSHRC and the German Humboldt-Foundation (TransCoop-Program); "The Stuff Stars are made of: International Politics, Mass Media and the Rise of dan Actors in the Republican Era (1910s–1930s)" funded by the Cluster of Excellence "Asia and Europe", the Krupp Foundation and the Institute of Chinese Studies, University of Heidelberg; "Building Early Chinese Periodicals Online (ECPO)—Expanding and Refining a Database Prototype for Historical Media Studies" funded by the Chiang Ching-kuo Foundation, Republic of China.

[4] See, for example, Hockx, et al (2018), and Sung, et al (2014).

including keywords and biographical information on agents—editors, authors and other individuals, groups, institutions, or corporations mentioned in articles and represented in illustrations and advertisements. We strongly believe in the importance of this manual editing of individual records for each item—article, image, advertisement, because this data analysis still offers information beyond what one might retrieve through a full text search, especially since much of the corpus is not yet available in full text for text mining. In addition, the bilingual analytical data makes these important sources available to non-readers of Chinese.

ECPO combines several important digital collections of the early Chinese press into a single overarching framework. To date, ECPO has focused on a body of rich but heretofore undervalued materials—women's and entertainment magazines. It is open to further additions: Currently, we are adding a selection of literary, art and women's magazines, e.g. Banyue半月 (The Half Moon Journal), *Tianyi* 天義(Tien yee), as well as western-language publications produced in China, e.g. The Canton Press.

The core of ECPO consists of several databases on early women's periodicals and entertainment publishing: "Chinese Women's Magazines in the Late Qing and Early Republican Period" (*WoMag*), "Chinese Entertainment Newspapers"(*Xiaobao*), and various databases hosted through the Academia Sinica in Taiwan.

WoMag [2] focuses on four influential women's magazines published between 1904 and 1937. It records all articles, images, advertisements, and related agents and assigns them to a complete set of scanned pages. This database is the model for what we have called the *intensive approach* within our database structure. *Xiaobao* [3] provides basic publication data and characteristics of the contents of some 22 entertainment newspapers (小報, *xiaobao)* from the late Qing and Republican periods. This database is the model for what we have called the *extensive approach* in our database structure.

The Academia Sinica, and in particular its Institute of Modern History, have in recent years digitized large parts of their collections of periodicals and built a database for the *Funü zazhi* 婦女雜誌 (The Lady's Journal) [4]. All these resources follow the model for what we call the *extensive approach*.

In a second, ongoing phase of the project, ECPO has begun to integrate these materials.[5] It is our aim to make the various individual digital collections accessible through a single search interface. We continue to acquire new publications to broaden the project's material base, and thus, to enrich and expand its potential for data analysis. We strive to open up our system to share data with the community, enable data re-use, and are adapting to principles of FAIR use.

As it currently stands, ECPO provides the research community with open access to more than 280 publications, mostly from the Early Republican period, comprising over 280.000 printed pages. A key and unique aspect of the project is to make entire issues available, front-to-back, including illustrations, advertisements, and even blank pages.

[5] During this phase, the project received a grant from the Chiang Ching-kuo Foundation 2012-2015 for a collaboration between Heidelberg University and the Academia Sinica. Since 2015, the Institute of Chinese Studies and the Heidelberg Centre for Transcultural Studies (HCTS) at Heidelberg University have continued to support the project. All technical development is coordinated through the Heidelberg ResearchArchitecture (HRA).

For approximately half of the publications, we provide descriptions and bibliographic metadata of individual items (articles, images, advertisements) in Chinese with Pinyin transcription.[6] These annotated records also contain genre and column information, basic content analysis through bilingual keywords, as well as the names and roles of agents associated with an item, including "mentioned in article" or "depicted in image". Overall, the project followed the five guidelines for the digital archiving of periodicals as formulated by Latham and Scholes.[7]

To further increase the impact of ECPO and in order to sustain the information, ECPO has begun to develop dynamic data services to provide data for re-use as open data. We implemented a MODS XML API [5] to provide bibliographic information of all annotated items in the database. In addition, we installed an IIIF image service for all page scans, and are developing an API to output each publication's detailed publishing information. Data sets will be published in heiDATA,[8] the Heidelberg research data repository.

2. The Agents Service

Before *WoMag* was linked to ECPO, both databases recorded personal names separately. At the same time, names were recorded without distinguishing between names and actual agents. We subsequently have begun to build a cross-database agent service, to create a single, central resource that hosts all agent-related data and can be referenced from other databases. Within this service, we combine the agent data from both *WoMag* and ECPO, keep the database open for the addition of new sets of data, and expand and refine the records of personal names into entries on agents, i.e. the persons behind a name, with biographical information, notes, and links to external authorities.

Our records of personal names were created manually by research assistants as part of the workflow to record individual articles, images, and advertisements. As with all manual input, some oversights and errors occurred in this process. Names were accidentally recorded more than once, name variants were not recognized as belonging to the same person, both of which led to the creation of more than one entry for a single individual. On the other hand, identical names were sometimes not recognized as belonging to different persons, resulting in only one entry for several persons. And sometimes, names were simply overlooked. In the process of joining ECPO and *WoMag* name records, we created a lot of duplicates, the same was the case when we ingested new data from other projects. Since all these issues have to be solved manually, we have implemented a number of tailored functions into the agent service to simplify these tasks.

The agent service allows us to: a) distinguish between "names" and "agents", b) assign names to agents, c) merge identical names (that refer to identical agents) across databases, and d) link agent records to authority data (GND, VIAF, Wikidata). At the moment

[6] Currently, (March 2019) 40.936 issues in total. Number of annotated items: 46.931 articles, 20.532 images, 18.639 advertisements.

[7] Latham and Scholes (2006), p. 524.

[8] heiDATA is an institutional repository for research data of Heidelberg University, see `https://heidata.uni-heidelberg.de/`. The ECPO data set will be available at `https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/Z3J0DV`.

(March 2019) we have 1220 Agents with references to VIAF, 985 with references to GND, 848 with references to Wikidata.[9] Recently, we started to also include Baidu Baike and DBpedia, and both authorities are referred to by 9 agents.

Before we can implement distinctions between "names" and "agents", i.e. separate names into different agent entries, merge them into one, or eliminate duplicates, we have to carry out research. In order to understand which person is meant by a particular name or visual representation in an item, we have to understand the item and its context; we have to find persons who share that name or that visual representation; and we have to know something of their biographical background. Only then can we identify an item as referring to a specific person and assign a respective agent entry to that item. Through this research process, we have often come across heretofore little known persons, events and phenomena.

One such phenomenon, which illustrates the effort we put into editing each agent entry, is the world tour of the Islington Corinthians Football Club.[10] Between 1937 and 1938, this London-based amateur football club sent a group of its players on a tour around the world, to engage in friendly matches in Europe, South Asia, East Asia and Northern America. In April 1938, their itinerary led them to China where they made headlines in a local entertainment newspaper.[11] In this instance, our task started with those newspaper articles and the question, which London-based football club in the 1930s was represented through the Chinese name *yi shi lin dun* 衣士林頓. Having found out that it likely refers to the Islington Corinthians F.C., we began to research which club members were sent on tour in order to create respective agent entries and match them to the Chinese version of their names; then, we looked into those club members' biographies in order to identify the people and organizations named in the Chinese newspaper articles. Finally, there were the names of the quasi-obscure foreign players and referees of the Shanghai Football Association against whom the Islington Corinthians played a match on April 3, 1938. Luckily in this case, we could rely on several outside digital resources,[12] such as a short Wikipedia entry for the Islington Corinthians F.C., contemporary newspaper articles, present-day web articles, and web sites of hobby historians and collectors of sports memorabilia. Our aim is to connect database users to these kinds of resources. Therefore, each agent entry is supposed to include assignments to ECPO items, short biographical information and references to the sources we used.

[9] VIAF: The Virtual International Authority File combines multiple national and international name authority files into a single OCLC-hosted name authority service; `https://viaf.org/`. GND: Gemeinsame Normdatei, or Integrated Authority File, is hosted by the German National Library; `https://www.dnb.de/EN/Standardisierung/GND/gnd_node.html` Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation; `https://www.wikidata.org`. Baidu Baike 百度百科 is a Chinese-language encyclopedia by the Chinese search engine Baidu; http://baike.baidu.com/. DBpedia aims to allow users to semantically query structured information from Wikipedia resources; `http://dbpedia.org/`

[10] "Islington Corinthians F.C.", ECPO Agents, last accessed May 5, 2019, `https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/agent-information.php?agentid=9814.`

[11] For example, "Ba qian qiu mi da wei sao xing: Yishilindun zu qiu dui zuo jing quan jun fu mei" 八千球迷大爲掃興：依士林頓足球隊昨竟全軍覆沒, Jing bao 晶報, April 4, 1938, p. 2, available at ECPO, `https://uni-heidelberg.de/ecpo/publications.php?magid=1&isid=3841&ispage=2&itemid=5325&itype=2.`

[12] See above, "Islington Corinthians F.C.".

To provide this information in a sustainable way, we also offer permanent links to all of our digital sources as archived in the project OpenDACHS [6].

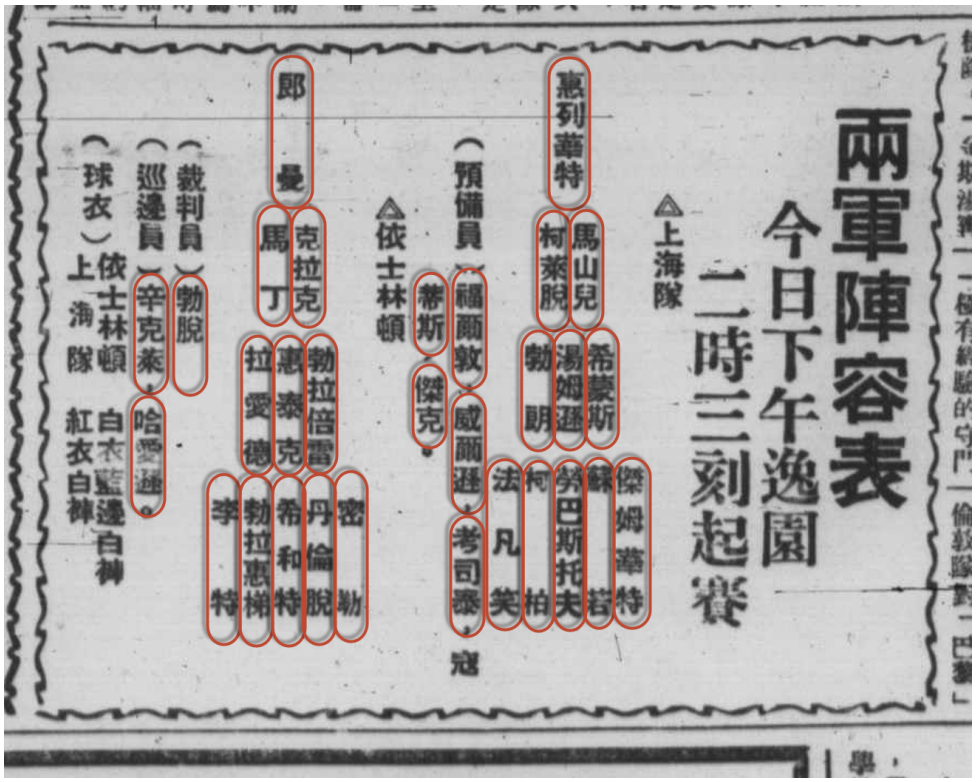


Figure 1.: The players' line-up for the April 3 match between the Islington Corinthians and an all-stars team of the Shanghai Football Association. All names are only printed in their Chinese versions. Cf. *Jing bao* 晶報(The Crystal), April 3, 1938, special issue, page 1, available via ECPO at `https://uni-heidelberg.de/ecpo/publications.php?magid=1&isid=4873` (emphasis markers by the authors).

Besides creating a curated list of agents occurring in ECPO items and connecting ECPO users to external resources, we also aim to open up and share our data. We plan to add missing persons to Authority files, using the German National Authority file (GND). The missing persons we plan to add range from people of minor, local importance, like the players of the Shanghai Football Association mentioned above; to people who played more prominent roles in world history, for example Sir Herbert Phillips,[13] Consul-General of the United Kingdom in Shanghai during the late 1930s, or Arminio de Mello Franco,[14] Brazilian ambassador to China in 1927-1928, neither of whom currently has a VIAF or GND record, or a Wikidata page. It is certainly neither meaningful nor feasible to add every individual from ECPO resources to the international authority.

[13] "Phillips, Herbert, Sir", ECPO Agents, last accessed on May 5, 2019, `https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/agent-information.php?agentid=9928`.

[14] "Franco, Arminio de Mello", ECPO Agents, last accessed on May 5, 2019, `https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/agent-information.php?agentid=33930`.

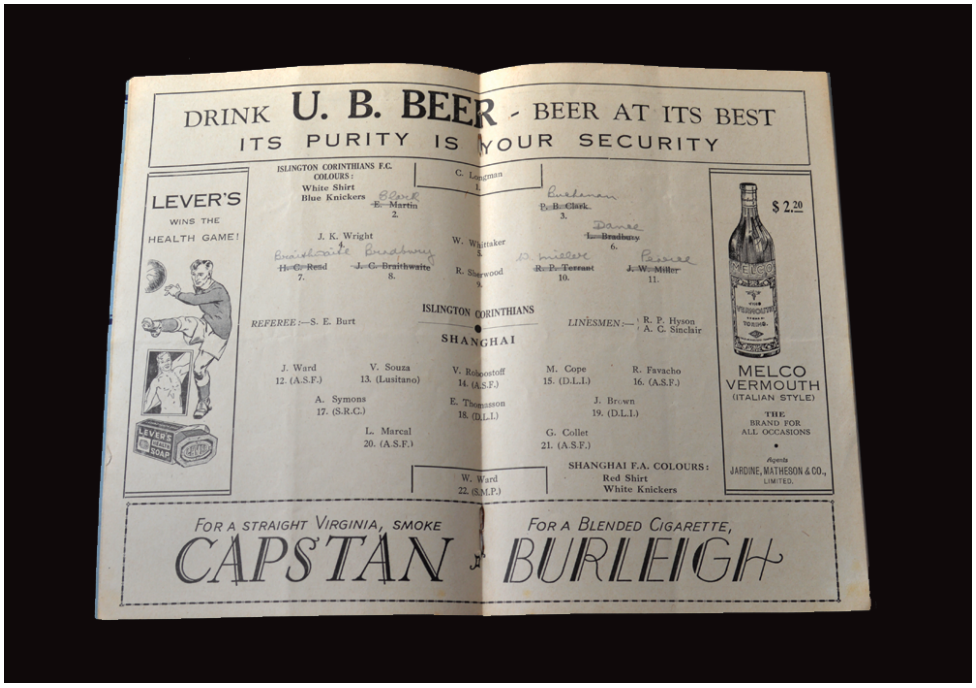
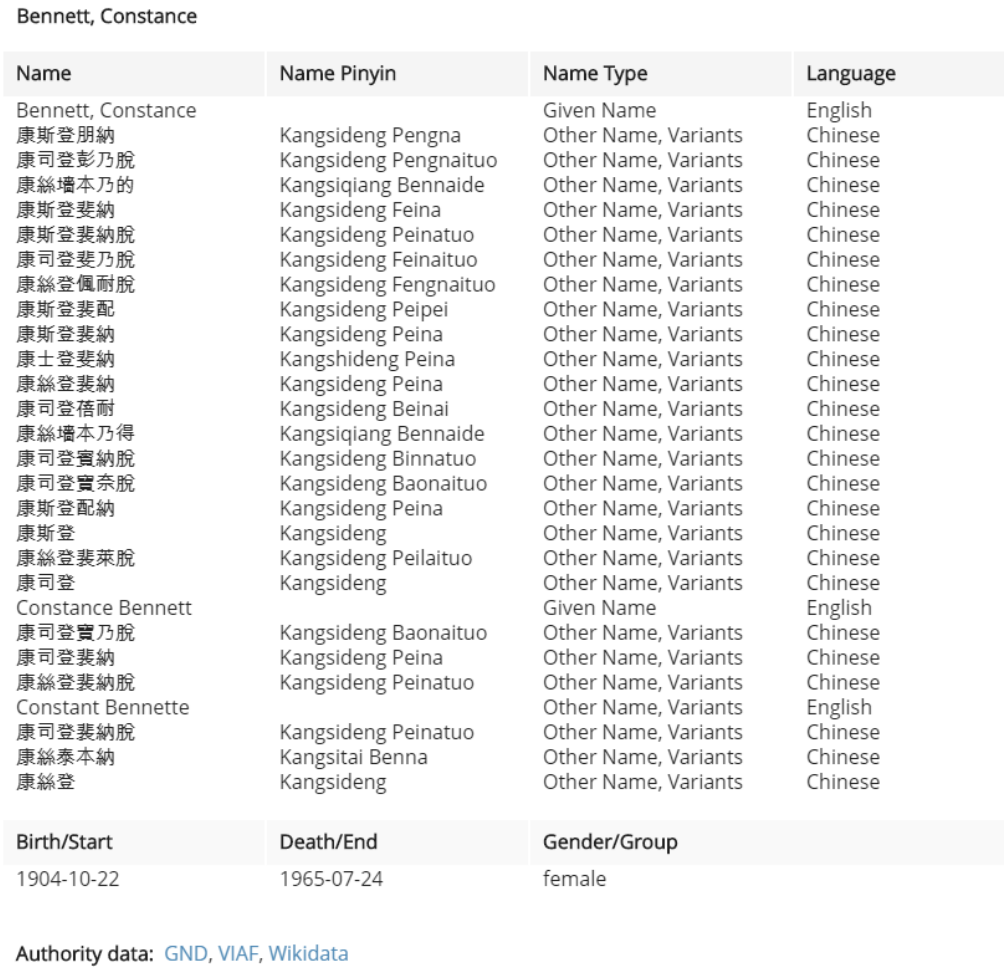


Figure 2.: The players' line-up for the April 3 match between the Islington Corinthians and an all-stars team of the Shanghai Football Association, as it was printed in the Shanghai Football Association's booklet for the game; all the participants' names are printed in Latin script. Photos of the booklet are hosted on the website 10footballs.com. They are part of an impressive private collection of football memorabilia, with a focus on the Islington Corinthians' world tour ("Islington Corinthians - Shanghai [team line up]", in *1938 - Islington Corinthians vs Shanghai: Sunday, 3rd April at 2.45 p.m. Canidrome*, edited by The Shanghai Football Association (Shanghai: The Printing Press, 1938), center pages, accessed February 7, 2019, `http://10footballs.com/wp-content/uploads/2018/10/063.png`).

But we will begin with more prominent figures and those agents that occur more frequently. Besides missing persons, we are preparing to add missing names to Authority files, especially Chinese variants of foreign names. In the Republican period, standardized Chinese renderings of foreign names did not yet exist. As a result, for some agents we have registered over twenty Chinese name variants. Most, if not all of these variants, are usually missing from Authority files such as GND and VIAF. Therefore, we are developing an API to provide our agents' data in machine readable format.

For example, the actress Constance Bennett[15] is mentioned by twenty-five different Chinese names in ECPO, and one variation of her English name.

[15] "Bennett, Constance", ECPO Agents, last accessed May 5, 2019, `https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/agent-information.php?agentid=34635`.



Bennett, Constance

Name	Name Pinyin	Name Type	Language
Bennett, Constance		Given Name	English
康斯登朋納	Kangsideng Pengna	Other Name, Variants	Chinese
康司登彭乃脫	Kangsideng Pengnaituo	Other Name, Variants	Chinese
康絲墻本乃的	Kangsiqiang Bennaide	Other Name, Variants	Chinese
康斯登斐納	Kangsideng Feina	Other Name, Variants	Chinese
康斯登裴納脫	Kangsideng Peinatuo	Other Name, Variants	Chinese
康司登斐乃脫	Kangsideng Feinaituo	Other Name, Variants	Chinese
康絲登佩耐脫	Kangsideng Fengnaituo	Other Name, Variants	Chinese
康斯登裴配	Kangsideng Peipei	Other Name, Variants	Chinese
康斯登裴納	Kangsideng Peina	Other Name, Variants	Chinese
康士登斐納	Kangshideng Peina	Other Name, Variants	Chinese
康絲登裴納	Kangsideng Peina	Other Name, Variants	Chinese
康司登蓓耐	Kangsideng Beinai	Other Name, Variants	Chinese
康絲墻本乃得	Kangsiqiang Bennaide	Other Name, Variants	Chinese
康司登賓納脫	Kangsideng Binnatuo	Other Name, Variants	Chinese
康司登寶奈脫	Kangsideng Baonaituo	Other Name, Variants	Chinese
康斯登配納	Kangsideng Peina	Other Name, Variants	Chinese
康斯登	Kangsideng	Other Name, Variants	Chinese
康絲登裴萊脫	Kangsideng Peilaituo	Other Name, Variants	Chinese
康司登	Kangsideng	Other Name, Variants	Chinese
Constance Bennett		Given Name	English
康司登寶乃脫	Kangsideng Baonaituo	Other Name, Variants	Chinese
康司登裴納	Kangsideng Peina	Other Name, Variants	Chinese
康絲登裴納脫	Kangsideng Peinatuo	Other Name, Variants	Chinese
Constant Bennette		Other Name, Variants	English
康司登裴納脫	Kangsideng Peinatuo	Other Name, Variants	Chinese
康絲泰本納	Kangsitai Benna	Other Name, Variants	Chinese
康絲登	Kangsideng	Other Name, Variants	Chinese

Birth/Start	Death/End	Gender/Group
1904-10-22	1965-07-24	female

Authority data: GND, VIAF, Wikidata

Figure 3.: Cut-out of the agent entry for Constance Bennett ("Bennett, Constance", ECPO Agents, last accessed May 5, 2019, `https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/agent-information.php?agentid=34635`.)

In GND, she appears only under her English name.[16] In VIAF, she is listed with twenty-two name variants in multiple languages.[17] However, only one of those variants is Chinese. Constance Bennett's example highlights a less obvious contribution the agent service offers to researching foreigners in China. In order to find information in contemporary Chinese sources, the Chinese version of a foreigner's name needs to be known. Constance Bennett's Authority file would provide a wide range of possible Chinese name variants of any other person sharing the name "Bennett" (which is quite a common name), and thereby become a starting point for researching Republican sources. The agents service could act as a resource for Chinese name variants of common English or other foreign names.

[16] "Bennett, Constance", GND, last accessed May 5, 2019, `http://d-nb.info/gnd/129660760`.
[17] VIAF ID: 34718115 (Person), last accessed May 5, 2019, `http://viaf.org/viaf/34718115`.

3. Towards Full Text

ECPO is now also beginning to focus on full text capabilities, with the aim of producing machine readable texts that can then be used for further analysis, like text mining. While some records occasionally feature full text passages in the metadata, for example some advertisements[18], this task is too big to solve manually – we need automated workflows.

However, after a number of first experiments it soon became clear that we cannot use OCR software out-of-the-box, for a number of reasons:

a) In many cases we are working with secondary material in sometimes sub-optimal quality. Images can be blurry, or show noise, stains, scratches, etc.

b) Document analysis fails to recognize the complex and very densely set page layout. This is especially true for newspapers.

c) Character recognition fails when it faces special characters, like emphasis marks next to characters. Typically, titles (e.g. of articles) or texts within illustrations are handwritten or feature special calligraphic styles, which also cannot be recognized.

While the production of quality full text can be a challenge even with western language material[19]—Chinese characters are very complex, and significantly different from Latin-based script systems[20]—there are different ways to approach these problems. One way is double-keying, but this approach is extremely labor- and cost-intensive.[21] Only very few such endeavors have been undertaken even in China.[22] Although OCR has been significantly improved in recent years, it still largely fails with Chinese texts from before the 1970s.[23]

[18] ECPO already contains more than 5.000 advertisement records with full-text data, for example, the advertisement for a publication by Bao Tianxiao 上海春秋第二集出版(Shanghai Chunqiu di er ji chu ban), 晶報 *Jing bao*, volume 1, issue 706, Monday, 1925-01-12, page 1: `https://uni-heidelberg.de/ecpo/publications.php?magid=1&isid=286&ispage=2&itemid=4754&itype=4`.

[19] The German OCR-D initiative `http://ocr-d.de` was started to coordinate the developments OCR for printed historical texts with a focus on 16th to 19th ct. German language material. It continues the efforts of the EUC project IMPACT `http://www.impact-project.eu/` and can be seen as parallel approach to the current Horizon 2010 project READ `https://eadh.org/projects/read`, which focuses on handwritten text recognition of archival records.

[20] With all language specific variants, the 26 Latin letters form a group of about 500, while there are over 50.000 Chinese characters alone, without taking the many variants into account. In the current version 12 of the Unicode standard a total number of 96190 Ideographic code points, or 87.887 CJK Unified ideographs are defined. [7]

[21] A typical quote for a periodical of the early 20th century is 3 USD per 1000 characters. Adopting these figures to the *Jingbao* newspaper (21 years, about 10.000 double-pages) would result in an estimated quarter million USD.

[22] For example the new edition of late 19th ct. Shanghai Daily Shenbao produced by double-keying specialist Greenapple Changsha. The biggest non-commercial program is the full-text digitization of the complete Buddhist canon by the Dharma Drum Institute in Taiwan, which took more than 12 years to complete and was largely carried out by volunteers from the worldwide Buddhist community.

[23] Commercial double-keying agencies have separate (higher) pricing for material even dating to "before 1990."

Figure 4.: Detail of a digitized microfilm page, showing stains and scratches, as well as different fonts, illustrations, and emphasis characters. *Jing bao* 晶報 (The Crystal), March 6, 1919, page 2 (upper part), available via ECPO at `https://uni-heidelberg.de/ecpo/publications.php?magid=1&isid=649&ispage=2`.

Our tests with OCR software like Abbyy Cloud OCR SDK [8] showed that processing full pages does not work.[24] Even when pages were manually segmented and the segments then processed, the OCR failed if emphasis characters were present. This only changed for segments that were manually optimized: emphasis characters removed and images optimized. There the recognition rate went up to over 60%. Although an error rate of up to 40% is still far from a good recognition result, it still indicates the potential of this approach. Tests by a possible Chinese partner[25] show different options to develop automatic workflows for further processing these segments, detecting and removing emphasis characters, running image optimizations, detecting lines and characters, and recognizing characters.

[24] Only less than 10% of the characters were recognized.

[25] Computational Knowledge Lab, Department of Engineering Science and Ocean Engineering, Taiwan National University, `http://www.cklab.org/`.

Recently, other platforms have started promising projects with processing non-Latin character scripts.[26]

Within ECPO we therefore focused on segmentation. We began a pilot project with Pallas Ludens [9], a local start-up specializing in crowdsourcing solutions. The pilot was performed by a non-Chinese-speaking crowd, who manually identified information blocks on newspaper pages and qualified them with a label. Identification was very good, but as the crowd was unable to read Chinese, they were not able to identify semantic groups, e.g. decide, which segments belong to one article and which to another. The grouping of individual boxes into meaningful semantic units was then done in a second run by a reader of Chinese. This, too, turned out to be quite successful. Unfortunately for us, Pallas Ludens was then bought out by a larger company and had to stop all external co-operation—including the one with us.

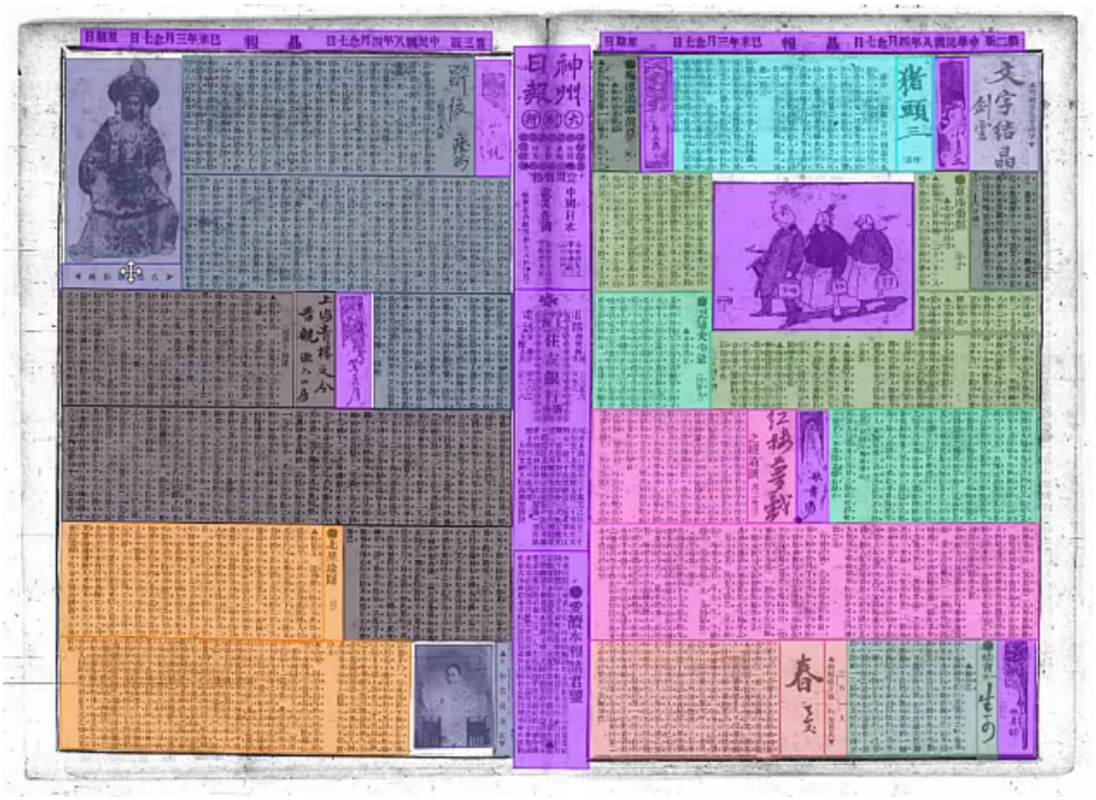


Figure 5.: A page where bounding boxes are grouped into semantic units, e.g. titles and text of articles, image and caption. From *Jing bao* 晶報(The Crystal), April 27, 1919, page 2, available via ECPO at `https://uni-heidelberg.de/ecpo/publications.php?magid=1&isid=9&ispage=2`.

Nevertheless, the outcome of the pilot taught us: crowdsourcing our material is not only possible, but can produce very good results, especially, if participants can read Chinese,

[26] For example, the *Naval Kishore Press – digital* project initiated by Heidelberg's South Asia Institute Library and Heidelberg University Library processed selected Hindi and Sanskrit titles in Devanagari script with Transkribus (`http://transkribus.eu/`). After training with ground truth from 200 pages an error rate of 5.59% was reached. See Merkel-Hilf (2018).

are supervised, and if user interfaces with excellent usability are provided.[27] We have just launched a new tool for annotating and semantic grouping that stores annotations with coordinates and labels in web annotation format in an XML database. The annotated segments can later be served in the database using the coordinates with our IIIF image service, and also be sent to a future OCR workflow. We are currently looking for partners to develop and implement these workflows to automate page segmentation and are in contact with computer vision labs and partners of the OCR-D initiative.

With the advent of basic segmentation and OCR workflows, we will soon have various full text segments. This calls for new solutions, as for example the text segments need to be linked with their respective metadata records, and the search has to be expanded to include unstructured texual data. The same is true when it comes to encoding full issues of magazines with structured markup, e.g. using TEI XML. The reason are a number of Asia-specific features, or perhaps even features of non-Western publications, for which good practice conventions still need to materialize. Within ECPO, we have started creating TEI records for the magazine *Tian yi* 天義(Tien yee).[28]

One major issue is the handling of emphasis characters in the text.[29] In *Tian yi*, for example, up to 6 different emphasis characters occur. Sometimes their use for emphasis is mixed with punctuation marks, in some texts almost every character is emphasized. Although there is a way to encode these characters using the list of "kenten" characters,[30] we eventually decided to not include these marks at all. The main reasons for this are: a) the workload for encoding and distinguishing emphasis characters and punctuation marks is huge, b) the added value of encoding these characters is limited, c) the focus currently rests on the creation of a machine readable text, and this additional level of information can be added at a later time. However, one type of emphasis is encoded by us, that is the font size. Texts can be set in double-size characters - indicating emphasized passages, in "normal" size, and in half-size characters - indicating comments or inserts.

[27] Outcome: *Jingbao* 1919-22 completely segmented with adjusted bounding boxes, for the first April issue (1919) all boxes are clustered in semantic groups.

[28] This amends an endeavor where we are adding research analysis of a project from the late 1990s into ECPO: As part of research focus "Frauenbewegungen - kultureller und sozialer Wandel" funded by the Hessian Ministry of Higher Education, Research and the Arts, Prof. Monika Übelhör (University of Marburg) received a grant for her project "Die Zeitschrift Tianyi (1907-1910) als Plattform für eine grundsätzliche Neubestimmung der Stellung der Frau in der chinesischen Gesellschaft"(The magazine Tianyi (1907-1910) as a platform for a fundemental redefinition of the position of women in Chinese society) in 1996. Project member Gabriele v. Sivers-Sattler, M.A. studied this publication and created a classified and annotated inventory. Cf. Sivers (2001) and Sivers-Sattler (2001). We are currently adopting their research outcome to the database ECPO database structure for data ingest.

[29] The authors are grateful for the valuable suggestions they got in various discussions with the following specialists: Duncan Paterson, Christian Wittern, Marcus Bingenheimer, Wolfgang Meier. Problematic cases and possible solutions are collected in an online document, cf. Arnold (2018).

[30] The list of *kenten* (jap.圏点) or "emphasis dots" comprise about 10 glyphs with their Unicode code points typically used for emphasis. Cf. the the W3C recommendations in Etemad and Ishii (2013), and the overview in the Japanese Wikipedia `https://ja.wikipedia.org/wiki/` 圏点 (page last changed 2019-03-30). Since CSS3 is still only a recommendation, browsers may render elements differently; some may not even be supported. For more information see Andrew (2017).

This is a typical feature in Asian texts, and we use CSS attributes with TEI elements `<emph> @type` and `<hi> @rend` to encode them.

Other editorial decisions we have to make are related to spaces. These can be indentations of full paragraphs, or single spaces. Indentation of longer passages is a common feature in magazines to indicate, for example different levels of argumentation, or longer commentaries. Individual spaces can also be used to emphasize the following character(s), but in some cases there may be technical reasons. While that is still a task for future research, there are multiple ways to encode these spaces. One is to use the `<space>` element with `@unit` and `@extent`, e.g.`<space unit="char" extent="1"/>`. Another suggestion is to use a U+3000 "Ideographic Space". At the moment we use the first method, but we are open for further discussion.

4. Summary

In this paper we discussed a number of approaches to further exploring and analyzing the contents of collected publications, together with efforts to open the collection's data for re-use. We demonstrated workflows in the Agents service, which assists in curating agent records across databases and forms the basis for enhancing authority records. We also presented results from a crowd-sourced approach to newspaper segmentation to generate segments that can easier be OCRed. In addition, we introduced our efforts to develop a proper markup for encoding full Chinese periodicals in TEI XML using *Tian yi* as example.

ECPO started as a typical information-silo: a sophisticated data structure, but no "outside" connections. While the collection of these materials itself is a huge contribution in providing the community with research data, we are continuously enhancing the metadata. Content analysis through keywords for each publication is amended by information about the publishing history, which is as comprehensive as possible. With the separation of meta-/data from the end-user interface we are able to start providing data sets in machine readable formats. We have also started to adopt FAIR principles: we are implementing DOI records for each publication, and connect our authority data to international authority files. We are publishing our resources and metadata Open Access, including metadata on publication, issue and individual item levels. We provide access to data sets through API's (e.g. bibliographic data in MODS format) and are preparing the publication of IIIF manifests. And we are publishing data sets on the Heidelberg research data platform heiDATA.

We still are expanding our data corpus: a project funded by the HCTS adds more than 100.000 pages from Foreign Press published in China. In co-operation with Erlangen University we are expanding our agent service with features like relations, and location services. We are also able to use ECPO to store and make available output from former research projects. In addition, ECPO will slowly grow into a data platform for other material from the CATS library. This will not only allow users to access the data, but also to further enhance and share it with the research community.

With its rich material base, ECPO is growing in different directions. A small project cannot do this alone. We are actively involved in initiatives like the DH-d Working Group

Newspapers and Journals, the non-Latn scripts interest group, the TEI East Asia SIG, and in close contact with the FID Asien (Cross Asia), as well as with members of OCR-d and READ/Transkribus. Only in collaboration with these groups and individuals can we successfully work with our own material and further develop ECPO.

Literature

Andrew (2017): Andrew, Rachel. „Christmas Gifts for Your Future Self: Testing the Web Platform". *24 Ways* (blog), 10. Dezember 2017. https://24ways.org/2017/testing-the-web-platform/.

Arnold (2018): Arnold, Matthias. "Tianyi bao - Questions related to TEI". 2018. https://docs.google.com/document/d/1xsE4kavEe-LdL7JKwDpaXtGZQbXLsRLtzSJ8sM4MTbw/edit?usp=sharing

Etemad and Ishii (2013): Elika J. Etemad and Koji Ishii (eds). "CSS Text Decoration Module Level 3", W3C Candidate Recommendation 1 August 2013, https://www.w3.org/TR/2013/CR-css-text-decor-3-20130801/#text-emphasis-style-property

Hockx, et al (2018): Hockx, Michel, Joan Judge, and Barbara Mittler, eds. *Women and the Periodical Press in China's Long Twentieth Century: A Space of Their Own?* Cambridge: Cambridge University Press, 2018.

Merkel-Hilf (2018): Merkel-Hilf, Nicole. „Naval Kishore Press – Digital: From Hidden Treasure to Open Access". *International Institute for Asian Studies - The Newsletter*, Autumn 2018. https://iias.asia/the-newsletter/article/naval-kishore-press-digital-hidden-treasure-open-access.

Latham and Scholes (2006): Sean Latham and Robert Scholes, "The Rise of Periodical Studies," *PMLA: Publications of the Modern Language Association*, 121 (2006), 517-531.

Sivers-Sattler (2001): Gabriele von Sivers-Sattler. "He Zhens Forderungen zur Namensgebung von Frauen im vorrevolutionären China: Untersuchungen zur anarchistischen Zeitschrift *Tian Yi (Naturgemäße Rechtlichkeit)* (1907-1908)". In Gimpel, Denise & Hanz, Melanie (editors). *Cheng - All in Sincerity: Festschrift in Honour of Monika Übelhör*, 275-284. [Hamburger Sinologische Schriften 2]. Hamburg: Hamburger Sinologische Gesellschaft e.V., 2001.

Sivers (2001): Gabriele von Sivers-Sattler. "Die mythische Figur Nügua in der anarchistischen Zeitschrift *Naturgemäße Rechtlichkeit (TianYi)*,1907-1908". In Übelhör, Monika (ed) *Zwischen Tradition und Revolution: Lebensentwürfe und Lebensvollzüge chinesischer Frauen an der Schwelle zur Moderne* [Beiträge zu einem Symposium des Fachgebietes Sinologie der Philipps-Universität Marburg vom 26. bis 28. November 1999], 105-130. [Schriften der Universitätsbibliothek Marburg, 107]. Marburg: Verlag der Universitätsbibliothek, 2001.

Sung, et al (2014): Sung, Doris, Liying Sun and Matthias Arnold. “The Birth of a Database of Historical Periodicals: Chinese Women’s Magazines in the Late Qing and Early Republican Period.” In *Tulsa Studies in Women’s Literature* 33, no. 2 (2014): pp. 227-37. http://muse.jhu.edu/article/564237.

Williams (1961): Williams, Raymond. *The Long Revolution*. Harmondsworth: Penguin Books, 1961.

Bibliography

[1] http://uni-heidelberg.de/ecpo

[2] http://uni-heidelberg.de/womag

[3] http://xiaobao.uni-hd.de/

[4] http://mhdb.mh.sinica.edu.tw/fnzz/

[5] http://kjc-sv034.kjc.uni-heidelberg.de/ecpo/api/mods

[6] http://www.asia-europe.uni-heidelberg.de/index.php?id=4425

[7] http://www.unicode.org/Public/UCD/latest/ucd/PropList.txt

[8] ABBYY Cloud OCR SDK http://ocrsdk.com/.

[9] Pallas Ludens http://pallas-ludens.com.

Defining the future scientific data flow for multi-disciplinary research data

Felix Bartusch[1], Kolja Glogowski[2], Ulrich Hahn[1], Michael Janczyk[2], Steve Kaminski[1], Jens Krüger[1], Volker Lutz[1], Gerhard Schneider[2], Mark Seifert[2], Dirk von Suchodoletz[2], Thomas Walter[1] and Bernd Wiebelt[2]

[1] Zentrum für Datenverarbeitung, University of Tübingen ;
[2] Rechenzentrum, University of Freiburg

Digital data and computerized workflows are at the core of almost every domain in science. Data is not only the base for scientific publication but can become equally important by itself. The discovery of new insights from huge amount of (unstructured) data for completely unrelated fields already have made big data a valuable asset for scientific findings. The value of the ever-increasing amounts of data for subsequent use and the requirements of funding agencies generate the need for formalized Research Data Management (RDM). Modern digital workflows involve more than one system to generate, compute or visualize ever-larger data sets. Thus, the operators of the large scale federated research infrastructures at the involved HPC computing centers in Baden-Württemberg face the challenge of providing suitable storage services. Such a Storage-for-Science (SFS) represents an essential building block for the anticipated state-wide data federation. In addition to the integration of the various pre-existing infrastructures, the long-term identification of data sets, their owners, and the definition of necessary metadata becomes a challenge. The implementation and provisioning of a RDM system needs to be organized together with the scientific communities and has to fit well into the growing Research Data Repositories landscape.

1. Introduction

Modern scientific work has become significantly digital, meaning it uses various devices, programs and tools to gather and process data in a multitude of ways. The involved digital research workflows are getting more complex and the tide of data processed or created through them is ever rising. Data intensive computing (DIC) involving big data or methods like deep learning provide a new perspective on existing data. The unstructured approaches to store and manage data of the past are not viable in the modern world of multi-disciplinary and multi-institutional research as well as over geographically distributed locations. A research data management system (RDMS) needs to answer how to properly store and present data on the long run, from short living projects and in an environment of high fluctuation of researchers as well as bridging from the existing landscape of network file systems into a world of flexible scientific workflows. The

Corresponding author:jens.krueger@uni-tuebingen.de

Felix Bartusch et al.: Defining the future scientific data flow for multi-disciplinary research data.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 110–127. DOI: https://doi.org/10.11588/heibooks.598.c8421

long-term identification of data sets, their owners and the definition of necessary metadata presents a challenge as well as the integration of large-scale object storage concepts. Future RDMS should consider the complete data life cycle, spanning from data acquisition over the various stages of computation, visualization to long-term archiving and publication. Additional components on top of the RDMS are desired to offer added-value services like special purpose repositories, semantic search, indexing or versioning. Individual scientific communities should be enabled to provide tailored services for their specific data management tasks as well without the need to run their own RDM enabled storage systems.

To help individual researchers to adhere to the FAIR principles [1] modern data management should extend beyond the traditional data-handling performed by now. Researchers often do not standardize metadata, making interoperability and sharing difficult. Data curation, the selection of data sets of relevance, and the removal of irrelevant data are often not formalized steps in the workflow. A RDMS can help to solve these shortcomings in today's workflows by providing tools and services to support researchers in their data management tasks and automation of various workflows. To address the researchers' needs the HPC cluster sites of the BinAC[1] in Tübingen and the NEMO[2] in Freiburg complement their compute infrastructures by a holistic approach to a RDMS. The Storage-for-Science (SFS) is designed to run in a cooperated, federated way spanning both locations. The system will become an integral part of the Baden-Württemberg data federation [2, 3] offering high performance data paths to other research infrastructures like HPC clusters and cloud systems. The establishment of the Science Data Center BioDATEN[3] will provide further input from one of the core scientific communities onto the system design and intended additional services.

The following paper is structured as follows: It gives an overview of the current state of data management nationally and internationally. The requirements stemming from the data life cycle, today's and future workflows of HPC and DIC user communities will be discussed. Further, it explores options to extend and optimize existing scientific workflows. From this discussion it tries to provide an overview on the concept of the Baden-Württemberg data federation and a coherent framework for the design of a research data management aware large scale data facility. The SFS is a RDMS supported by the DFG and state funds to provide joint storage and research data management functionality for various research groups in Tübingen, Freiburg, Ulm and Stuttgart. The texts extends upon the inputs provided in a paper published by the North Rhine-Westphalia RDMS group [4], an article presented at the DFN forum in 2017 [5] and a discussion on researchers' needs in the HPC domain presented at the eScience days 2019 in Heidelberg [6].

[1] `https://www.binac.uni-tuebingen.de`
[2] `https://www.hpc.uni-freiburg.de/nemo`
[3] `http://www.biodaten.info`

2. Related Work

The Research Data Alliance (RDA), has been brought into existence to tackle the challenges of modern research data management from an infrastructure perspective [7]. From the perspective of research funding institutions there is an increasing movement to impose a good scholarly practice by requiring scholars and scientists to plan and execute sound data management.[4]

Initiatives in various countries to provide infrastructures, support and services exist for quite a while: The UK as well as the Netherlands were among the first movers in Europe to provide a range of discipline-specific data centres or to offer support for data archiving for scientists from various disciplines [8]. The Digital Curation Centre[5] is one of the most relevant national providers of expertise. In the Netherlands the Data Archiving and Network Services (DANS)[6] focuses on the data archiving of the social sciences and humanities. It started the development of cost models from the very beginning to provide a sustainable service in the long run [9]. To address the needs of data globalization and high performance access Onedata in Poland follows an approach relying on a global registry for mediating metadata synchronization and file transfer [10]. Overseas in the US, the National Science Foundation supports several projects to build sustainable infrastructure by trying to support all sides: scientists, software developers, librarians, archivists, and information scientists as well users to deeply engage in a joint process. Two prominent examples are the Science Gateway Institute USA[7] which supports domain developers to create tailored portals for their communities [11] and the US Research Software Sustainability Institute[8] helping to preserve related software. A similar institution was founded in the UK, the Software Sustainability Institute[9] which is also strongly engaged in training and education [12]. Federated infrastructures like EUDAT can provide guidance on which services are to be provided and how the several challenges are tackled [13, 14]. In Germany the developments gather momentum, one significant step forward is the development of the National Research Data Management Infrastructure [15]. Several regional, often federated initiatives work on a RDMS, like the consortium of several universities in North Rhine Westphalia lead by the university of Aachen [4]. A flexible and scalable storage system was created using a Quobyte[10] solution for the de.NBI Cloud in Tübingen, providing a broad range of functionalities. It features both a full integration with OpenStack and additionally a multipurpose S3 object storage as well [16]. The ViCE project [17] worked on different use cases for virtualization or containerization. It evaluated the various efforts to access external data stores in relation to the actual location of the virtual research environment [1]. A discussion on the design of a RDMS can be found in [5] as well as a RDMS in relation to other large scale infrastructures in [6].

[4] See e.g. German Research Foundation [19], the Federal Ministry of Education and Research [20] or the German university rectors" conference [21, 22].
[5] For more information see, `http://www.dcc.ac.uk`
[6] For more information see, `http://www.dans.knaw.nl`
[7] For more information see, `https://sciencegateways.org`
[8] For more information see, `http://urssi.us`
[9] For more information see, `https://www.software.ac.uk`
[10] `https://www.quobyte.com/case-studies/uni-tuebingen`

The aforementioned programs, initiatives and implementations can only give a cross-section overview about the vast number of worldwide activities dealing with modern research data management in its broadest sense.

3. Data Life Cycle

Extensive data analysis has become an irreversible trend in modern science. The increasing scale of scientific data is invalidating classic methods that had previously been considered to be good enough. For example, storing large amounts of data in a single filesystem with the Data Management Plan (DMP) consisting in the proper naming of files and directories is probably one of the least scalable methods. As a popular approach, it just requires a naming convention to start with and disciplined scientists adhering to it. It might even work reasonably well for a limited time within a small group of scientists and a moderate amount of data. However, further developing this method into a proper DMP, spanning the complete life cycle of the data and extending it to larger groups of scientists, is a tedious if not impossible task.

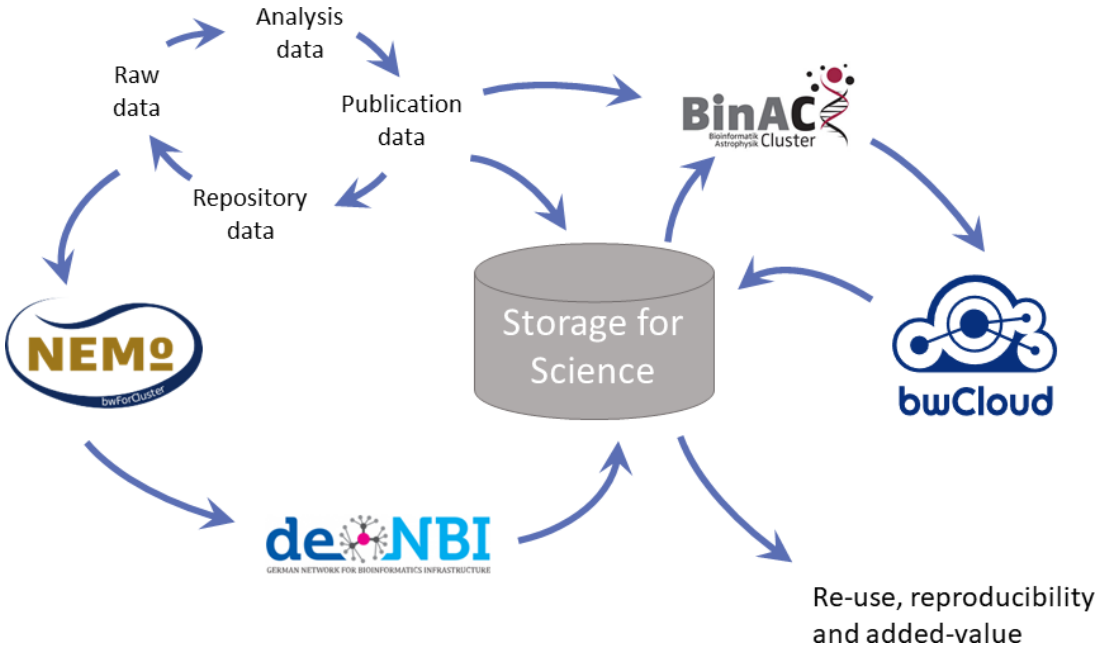


Figure 1.: Depiction of the data life cycle involving the bwForClusters NEMO and BinAC, the de.NBI Cloud in Freiburg and Tübingen as well as the bwCloud, relying the on the Storage-for-Science.

A modern DMP has to consider the complete life cycle of the data (see Figure 1) residing on multiple geographically distributed resources [23]. Typical examples for the provenance of raw scientific data include experiments and scientific simulations. In both cases, if large amounts of raw data are generated, a sufficiently large and performant storage resource is required (LSDF – large scale data facility). This storage resource needs to be either directly attached to the scientific device generating the data (e.g. microscope,

compute cluster) or it needs to be connected via a high speed network and protocol. Alternatively, the scientific device could be fitted with a fast local storage that caches the data on acquisition and forwards it over time to the large scale data facility. After the raw data has been collected it has to be annotated and enriched with an initial set of qualifying metadata. This includes both, metadata for scientific processing and metadata for governance purposes. To enable scientific reproducibility, data is stored in an immutable way after acquisition and major processing and refinement steps. Changes are recorded in a way they can be tracked and possibly rolled back and reapplied. This is the core requirement for offering the data as a public repository to a wider scientific audience. Furthermore, this enables the publishing of research data respecting all four FAIR principles (findable, accessible, interoperable, re-usable [1]).

4. Annotation

Throughout a modern research data life cycle (see Figure 1), an appropriate metadata management is crucial for the immediate and long-term preservation, discovery, publication and reuse of scientific data. From the very beginning of a research project, a DMP should provide information about discipline specific metadata and related vocabulary needed for the enrichment of data objects. Especially the continuation of a provenance information chain on data objects throughout the life cycle is a key aspect of metadata management.

Research projects are usually in the need for support from infrastructure providers regarding metadata management. They need to offer an ecosystem (see Figure 2) including storage, search systems, pid services, interfaces and presentation layers for metadata management and have to take care of keeping metadata standardized (e.g. METS/PREMIS [29, 30]) and interoperable between systems (via protocols like OAI-PMH [31]). Software platforms together with rich metadata must provide an abstraction layer to enable researchers to effortlessly track, move and collect their distributed data objects in complex environments. In distributed systems, a search for data objects from rich metadata information is considered to be far superior to manual browsing and filename search.

The enormous amount of digital objects created in modern HPC environments makes an entirely manual generation and management of metadata impractical. In fact, a phenomenon known as metadata bottleneck [11] must be circumvented via the usage of automated processes to support researchers and system administrators. One step further, automated parsing of basic information from e.g. running HPC jobs into metadata records, such as job-id, user-id, number of allocated cpu and so forth, could help to complete metadata information. Other processes could check for changes on data files and automatically generate and parse provenance information into related metadata records. Based on the specifics of the research project, several types of metadata information like file ownership, access rights and license information for data reuse could be conveniently configured as predefined values for all data objects within the project scope. With correspondingly higher effort, even parts of discipline specific metadata could be automatically collected via script based scanning through job output and log files. Based on automatically created metadata records available in a searchable data store, researchers must furthermore be

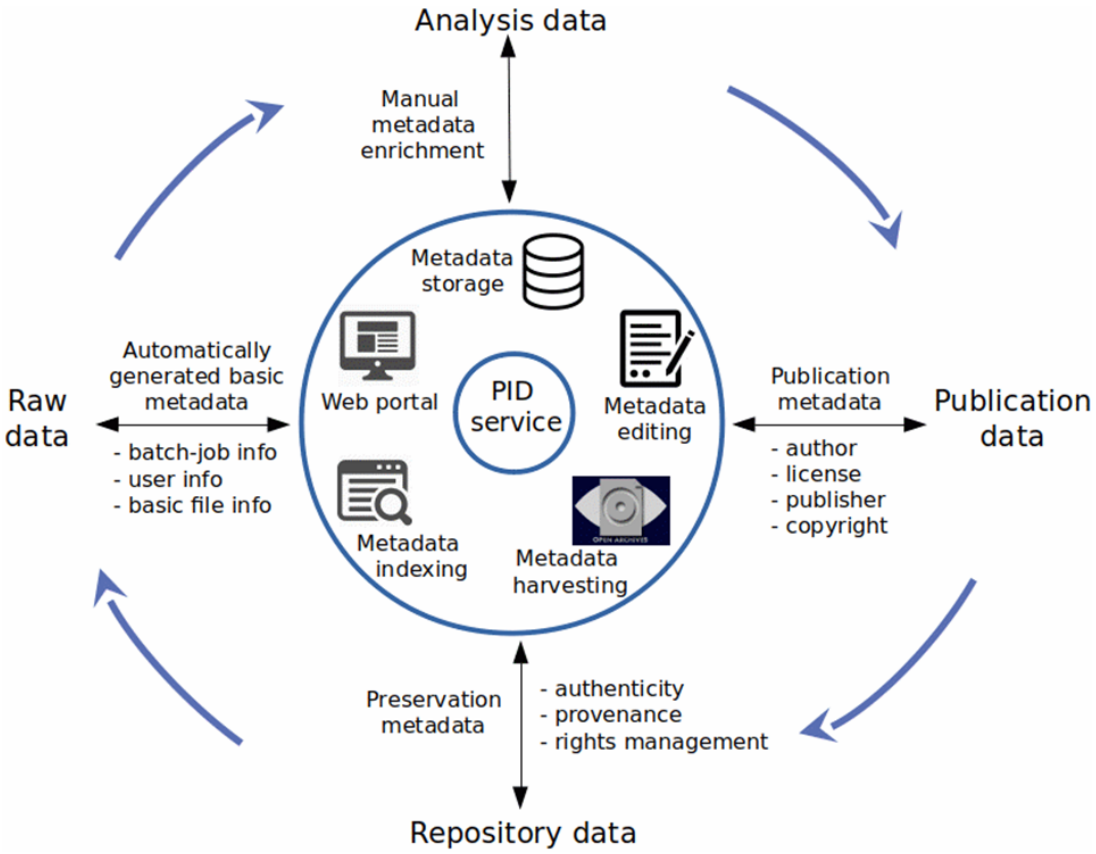


Figure 2.: Illustration of a metadata management chain. Icons where taken from [24, 25, 26, 27, 28].

enabled to manually add descriptive information via appropriate tools and standardized workflows provided by the technical infrastructure.

Metadata information can also support data staging processes on federated storage systems such as the SFS based on file status information available from the metadata. Data records with the status „cold“ could be semi-automatically moved to affordable and less performant storage media, usually until the end of a research project, whereas data annotated as „hot“ must be kept on expensive and performant storage for subsequent data analysis.

5. Data Staging

Research data are often produced on storage systems at different locations from where they are further processed or stored for future use. One example for this is a simulation that writes its results to the parallel file system of an HPC cluster. These results are then often analyzed using compute resources that have no direct access to the original parallel file system (e.g. cloud-based virtual machines (VMs), dedicated high-memory servers or visualization workstations, or external HPC clusters). Other examples are scientific instruments (e.g. telescopes, microscopes or sequencers) that create large amounts of

observational data on local cache storage systems. After acquisition the data need to be moved to other storage systems, presumably at a different geographic location, where they can be stored safely, calibrated, visualized and analyzed.

In very simple cases, direct copies can be a viable solution for transferring data between different locations, but handling data manually gets quickly complicated when more complex scenarios and environments are involved. Especially when data need to be replicated at multiple locations on storage systems belonging to different organizations, or when data sets are shared between collaborators, the utilization of a dedicated RDMS becomes inevitable.

Starting from data stored in mutable files on a file system that can be directly manipulated by the user, a first step towards a managed system is to define data sets or collections of files, and ingest these data sets into such a system, creating well-defined objects that can be associated with persistent identifiers. This way data sets can be distributed asynchronously to any number of locations, without worrying about changes to local files. The transition from mutable files to well-defined, managed objects also marks a point where data sets can be augmented with additional governance metadata that define ownership, access permissions and data retention in a way that is independent of location-specific administrative domains [5], and allows collaborations to manage their data using a uniform, location-independent namespace.

With regard to the anticipated data federation in Baden-Württemberg, a suitable data management system needs to be vendor-independent, and must support heterogeneous storage solutions as well as a variety of access protocols. In order to gain high acceptance from scientific communities, it also needs to be able to support different workflows and scale well for large data sets and high numbers of files. It should ideally be an actively maintained open-source software that is based on standard software components and provides modern application programming interfaces (like REST APIs) for its subsystems. Two popular examples of existing open-source data management systems are iRODS[11] and Rucio[12] [33].

6. Reproducible Research Environments

The term research environment encompasses the software stack, explicit description of workflows, custom scripts, and settings a researcher used for processing the data sets. As these components of the research environment are stored on computational resources, they should be also considered as data. As a major consequence of this, the whole data life cycle, as well as the FAIR principles, are applicable to the research environment. Leveraging the data life cycle, research environments should be versioned and archived in order to enable reproducible computations.

The explicit notation of an analytical workflow in a workflow language is good scientific practice and the publication of these workflows according to FAIR principles is essential for transparency and reusability [34]. Analogous to open data, open research environments

[11] `https://irods.org`
[12] `https://rucio.cern.ch`

increase reproducibility and – not unselfish for an author – also increases the number of citations [35]. It is obvious, that a simple listing of software used by a pipeline is not sufficient for other researchers to reproduce a complex computational analysis [36, 37]. In recent years software containerization techniques like Docker[13] and Singularity [38] were developed, allowing researchers to package a specific software stack, including an operating system and additional data like custom scripts, in a single entity called container image. This technique enables versioning and archiving of research environments and pipelines, as the environment is bundled in one image [39, 40]. It also enables researchers to share this single entity via scientific data repositories, public container hubs, or institutional repositories residing on systems such as the federated SFS.

In contrast to a static research environment installed on bare metal, the container is portable and can be deployed on arbitrary computational resources. When performing computations on big datasets, it could be faster to move the containerized research environment to the data, instead of moving the datasets to a computational resource. One advantage of a state-wide data federation would be, that researchers can decide to move their data to the containerized research environment, or vice-versa.

A disadvantage of software containers is their black-box character, as it is not obvious which software versions and scripts are bundled. As stated in Section 3, organizing the containerized environments using a simple naming scheme is not easy to implement and would not scale very well. Using a standardized metadata schema[14] to describe containers and their contents would break this black-box character to some extent and make searchers in an environment registry possible. If the containerized research environment was used for processing datasets, recording this container in the metadata of resulting datasets increases the result's provenance. Ideally, the researcher is able to create a seamless provenance chain starting from the instrument, that produced the raw data, over several processing steps using software containers to the final results.

The results of two projects, both funded by the state of Baden-Württemberg, could be used to provide researchers tools for handling containerized research environments in the proposed concept. The ViCE project created a prototype of an image registry, which can be used to manage research environments scientists tailored, based on their needs. Such an image registry can then be used to exchange images amongst platform borders, e.g. different HPC clusters [11]. Although container images allow a reproducible deployment of complex software stacks, in terms of long-time archiving and executability the container runtime is an additional point of failure. The project CiTAR (Citing and Archiving Research)[15] provides a platform for archiving virtual machine and container images. They are normalized to one standardized format (e.g. OCI image format), for which the runtime (e.g. runc) is available in the CiTAR service. A citable handle is assigned to the archived environment. A data federation, that promotes and supports the tools and workflows introduced in this section, would allow researchers to create and manage reproducible research environments.

[13] `https://www.docker.com/`
[14] `https://github.com/opencontainers/image-spec/blob/master/annotations.md`
[15] `http://citar.eaas.uni-freiburg.de/`

6.1. Virtualization

For the application of the bwForCluster „NEMO“ [41] one of the key requirements was to provide the complex computing environments for the experimental particle physics collaborations like the „Compact Muon Solenoid“ (CMS) or „A Toroidal LHC ApparatuS“ (ATLAS) at CERN. Inside this environment they had to provide services, which enable researchers to access scientific software and data. After discussion with the scientific groups the solution was to provide virtualization based on OpenStack [17]. As preparation for the bwForCluster NEMO a test cluster was created to develop this solution. The CMS groups at the KIT already had developed a resource broker „Responsive On-demand Cloud Enabled Deployment“ (ROCED) which communicates between different batch systems. This had to be enhanced with a connector to Moab/Torque for NEMO and SLURM for the ATLAS groups in Freiburg [42]. When NEMO started August 2016 the solution based on ROCED was already established and fist jobs already were computed [43].

For the upcoming replacement of the bwForCluster „BinAC“ [44] a similar strategy is anticipated, to be able to separate the environment for sensible data from the broader pool of compute and storage resources. Since virtualization encapsulates the environment and creators of this „Virtualized Research Environment“ have root access inside the virtual machine, the OpenStack environment has to be separated from the cluster. Parallel and home file systems like BeeGFS, Lustre, NFS, etc. are usually not secured for performance reasons and usually this is not necessary, since traditional clusters are black boxes and can only be accessed through special login nodes [6]. Having a closed environment without external mounts data for processing and scientific software has to be streamed or copied from external services. New approaches for caching data and software (like CVMFS, XrootD, Squid, etc.) have to be implemented and tested [45].

7. Baden-Württemberg Data Federation

The state of Baden-Württemberg has recognized the relevance and importance of RDM for sustainable and future-oriented organization of research data, its availability and publication. This is addressed in the jointly developed eScienve concept[16] and through two statewide funding programs (Research Data Management and Virtualized Research Environments).

The plan of action for development and deployment of federated research data management infrastructures follows the recommendations of the German Information Infrastructure Council.[17] Existing infrastructures (LSDF in Karlsruhe and Heidelberg) are becoming enriched with research data management capabilities. Emerging infrastructures (SFS in Freiburg, Tübingen, Stuttgart and Ulm) are already procured with research data management as a core functionality. Hence, researchers are not just simply provided with larger storage systems. Additionally, they will be given the tools they need to devise and implement DMPs suited for their field of research. To prevent isolated non-sustainable

[16] For more information see, `https://idw-online.de/de/attachmentdata37340.pdf`

[17] For more information see, `http://www.rfii.de/download/rfii-recommendations-2016-performance-through-diversity`

provisional solutions, these efforts are coordinated in a data federation. Establishment of a federation by connecting existing individual data management systems can be studied from several examples: Initiatives in the Helmholtz Association (Supercomputing and Big Data program)[18], activities at the national level (BMBF, DFG: LIS programs)[19], activities on the European scale (EUDAT)[20], and possibly EOSC[21] in the future) and finally attempts on the international level (Research Data Alliance)[22].

Access to storage systems requires a variety of protocols to meet the diverse needs of scientists. These include remote file system protocols such as NFS and CIFS/SMB as well as object store protocols such as Amazon S3. Additionally, performant network connections have to be offered for the coupling of data and HPC systems.

In many areas of application, sensible data is subject to complex and restrictive usage rules, which have to be taken into account. These restrictions are typically due to storage of personal data (e.g. medicine, social sciences, psychology and mobility research) or storage of secret data (e.g. economics and engineering). A data federation can establish, support and enforce policies to govern the adherence to these rules.

The basis for the Baden-Württemberg data federation are the existing and emerging infrastructures for data management and data storage in the state (see Figure 3). These include the parallel file systems of the HPC systems, LSDFs, specialized data analysis systems, repositories and archiving systems such as bwDataArchiv and bwDataDiss. In addition to these statewide storage systems and repositories, gateways need to be established to other national and international systems operated by their respective communities.

8. Storage-for-Science

The concept of SFS is a distributed federated storage system that spans over four different university locations in Baden-Württemberg. Besides offering a large storage capacity for research the focus is on additional RDM functionality to enable researchers to properly annotate their data from the start to the end of their project data life cycle. Figure 4 shows the main building blocks for the SFS. Caching systems provide access to existing data and allow data import from all kinds of measuring instruments. The main sites in Freiburg and Tübingen provide central mass storage for active data as well as a geo-replicated long-term storage infrastructure for „cold“ and archived data. Configurable data movers will move the data between the already mentioned hierarchy levels of SFS. The data movers and their APIs will enable the control of the data flows within the SFS system, to the HPC systems such as NEMO and BinAC and to other storage systems, thus spanning part of the data federation.

In addition to generic RDM methods for capturing and handling metadata, the development of additional RDM tools in collaboration with the research groups on the SFS is

[18] For more information see, `https://www.helmholtz.de/en/research/key_technologies/supercomputing_big_data`
[19] For more information see, `https://www.dfg.de/foerderung/programme/infrastruktur/lis`
[20] For more information see, `https://eudat.eu`
[21] For more information see, `https://www.eosc-portal.eu`
[22] For more information see, `https://rd-alliance.org`

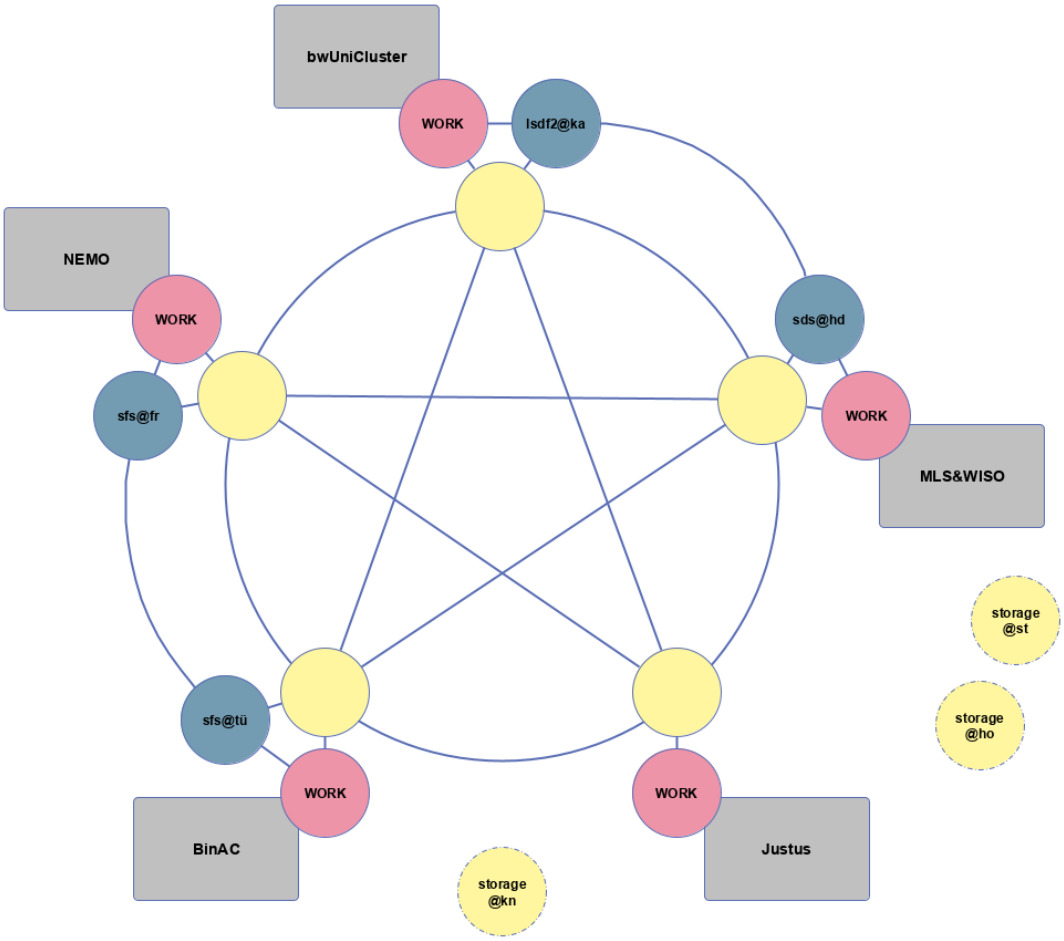


Figure 3.: Depiction of the anticipated state-wide data federation, involving Storage-for-Science (blue circles on the left).

key for the integration of community based metadata standards. Manual and automatic generation of metadata will be supported, therefore the data movers will always move metadata alongside data. Consequently, data and metadata consistency is given on every level of the SFS hierarchy thus generating an integrated research data management. Beside the benefits of getting integrated data management for project data life cycles and research data life cycles further improvements for research by utilizing the data mover capabilities to implement workflows that resemble the data analysis and project workflows in research are expected. The integration of existing workflow engines like Galaxy[23], UNICORE[24] and NextFlow[25] is envisioned.

The technical building blocks for the SFS system will be file system based NAS and object storage systems. While the file system based technology will be used for the cache level and the largest part of the central storage level, the object storage systems will

[23] For more information see, `https://galaxyproject.org`
[24] For more information see, `https://www.unicore.eu`
[25] For more information see, `https://www.nextflow.io`

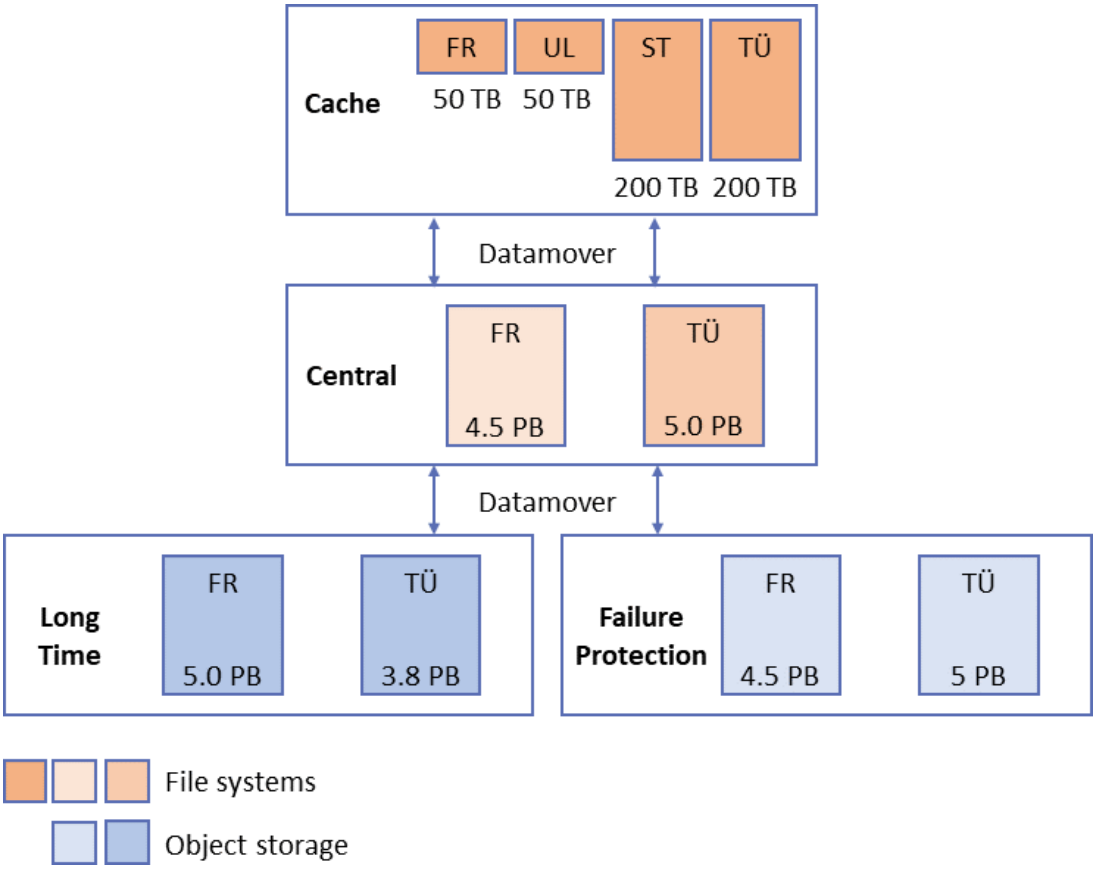


Figure 4.: Planned hierarchy levels of the distributed federated Storage-For-Science system including the anticipated size of the individual components.

be used on the long-term storage and data protection levels. The integration of object storage into SFS will increase the flexibility of data protection and data archiving, as provisioning of storage resources with dedicated levels of data protection and wide area geo-replication are already implemented. Software components such as the data movers will provide transparent access to data for researchers on all levels and additional RDM functionality. The integration of existing RDM solutions as well as interfaces for external and community repositories will provide further benefits for research. The SFS system user management needs to be only loosely coupled to the individual local IDM systems of the involved sites. It will adapt concepts from established AAI federations to ensure the usability across all participating universities, institutions and support widely cooperating scientists.

Similar to the „Central Application Site for bwHPC“[26] a self-service interface is envisioned. It would serve multiple purposes, like allowing project managers to declare their storage needs and provide information on their project, while storage administrators will also get informed about their users' needs. The descriptive and administrative metadata getting collected helps both to plan for storage volumes required and provide information

[26] For more information see, `https://www.bwhpc.de/en/zas_info_bwforcluster.php`

to university research information systems or grant providing agencies. Typically such information would include project title, containing additionally organizational names and persons involved, contact information, research grant and a short project summary to provide a context for the data, e.g. a description from a relevant ontology, an abstract of a related or planned paper, or excerpts of a relevant proposal. Further information should include the type (e.g. „hot“ or „cold“ data; file system, object storage, repository, long-term archive), redundancy expectations and the expected capacity and duration for storing the data. Especially the latter information would help for long-term planning and development of the system as well to plan for refinancing.

9. Conclusion

In the process of grant application and procurement for the federated SFS essential core characteristics and abstract features were identified together with the served scientific communities. The system will offer a good compromise of price per terabyte, performance and capacity. It features both traditional file systems as well as object storage to address the various needs of the scientists. It provides various levels of defined services including geo-redundancy. For trusted, reproducible scientific workflows the well-annotated, archived data will be immutable. It allows the automation of workflows by providing appropriate interfaces like REST APIs allowing asynchronous operation. The SFS system implements an identity mapping for users adapting concepts from established AAI federations. It thus abstracts from site-specific identity management to support the linking of long-term object identifiers to data owners and their individual requirements like embargo or retention periods. The SFS will handle the core project descriptive, administrative and technical metadata leaving the freedom to use specific scientific metadata to the respective communities.

The joint procurement and later operation of the system deepens the cooperation between the involved computing centers and communities. The ongoing process generates insights to be used for general application of research data management in the involved universities. The expertise is shared within the context of the data federation in Baden-Württemberg and the bwHPC-S5 project [3].

Acknowledgements

The research infrastructures described in this publication are part of the bwHPC project sponsored by the Ministry of Science, Research and the Arts, Baden-Württemberg (MWK) and the German Science Foundation (INST 37/935-1 FUGG and INST 39/963-1 FUGG). The same applies to the work presented from the CiTAR and ViCE projects which got supported by the MWK as part of it’s eScience initiative. Part of the work presented here was also supported through the BMBF funded project de.NBI (031 A 534A). The support is gratefully acknowledged.

Bibliography

[1] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Sant os, Philip E. Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 2016. doi:10.1038/sdata.2016.18.

[2] Gerhard Schneider, Vincent Heuveline, Karl-Wilhelm Horstmann, Bernhard Neumair, Petra Hätscher, Josef Kolbitsch, Simone Rehm, Michael Resch, Thomas Walter, Stefan Wesner, and Peter Castellaz. Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das High Performance Computing (HPC), Data Intensive Computing (DIC) und Large Scale Scientific Data Management (LS^2DM). In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 3–16. TLP, Tübingen, 2019. doi:10.15496/publikation-29040.

[3] Robert Barthel and Jürgen Salk. bwHPC-S5: Scientific Simulation and Storage Support Services – Unterstützung von Wissenschaft und Forschung beim leistungsstarken und datenintensiven Rechnen sowie großskaligem Forschungsdatenmanagement. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 3–16. TLP, Tübingen, 2019. doi:10.15496/publikation-29039.

[4] Thomas Eifert, Ulrich Schilling, Hans-Jörg Bauer, Florian Krämer, and Ania Lopez. Infrastructure for Research Data Management as a Cross-University Project. In *International Conference on Human Interface and the Management of Information*, pages 493–502. Springer, 2017.

[5] Dennis Wehrle, Bernd Wiebelt, and Dirk von Suchodoletz. Design eines FDM-fähigen Speichersystems. In *10. DFN-Forum Kommunikationstechnologien, 30.-31. Mai 2017, Berlin*, pages 115–124. Gesellschaft für Informatik eV (GI), 2017. `https://dl.gi.de/bitstream/handle/20.500.12116/470/paper10.pdf`.

[6] Dirk von Suchodoletz, Ulrich Hahn, Bernd Wiebelt, Kolja Glogowski, and Mark Seifert. Storage infrastructures to support advanced scientific workflows. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 263–279. TLP, Tübingen, 2019. doi:10.15496/publikation-29058.

[7] Andrew Treloar. The Research Data Alliance: globally co-ordinated action against barriers to data publishing and sharing. *Learned Publishing*, 27(5):S9–S13, 2014.

[8] Peter Doorn and Heiko Tjalsma. Introduction: archiving research data. *Archival science*, 7(1):1–20, 2007.

[9] Anna S. Palaiologk, Anastasios A. Economides, Heiko D. Tjalsma, and Laurents B. Sesink. An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS. *International journal on digital libraries*, 12(4):195–214, 2012.

[10] Łukasz Dutka, Michal Wrzeszcz, Tomasz Lichoń, Rafał Słota, Konrad Zemek, Krzysztof Trzepla, Łukasz Opioła, Renata Słota and Jacek Kitowski. Onedata – A Step Forward towards Globalization of Data Access for Computing Infrastructures. *International Conference On Computational Science, ICCS 2015*, 51:2843–2847, 2015. doi:10.1016/j.procs.2015.05.445.

[11] Sandra Gesing, Nancy Wilkins-Diehr, Maytal Dahan, Katherine Lawrence, Michael Zentner, Marlon Pierce, Linda Hayden and Suresh Marru. Science Gateways: The Long Road to the Birth of an Institute. *Hawaii International Conference on System Sciences (HICSS)*, 2017. doi:10.24251/HICSS.2017.755.

[12] Stephen Crouch, Neil Chue Hong, Simon Hettrick, Mike Jackson, Aleksandra Pawlik, Shoaib Sufi, Les Carr, David De Roure, Carole Goble and Mark Parsons. The Software Sustainability Institute: Changing Research Software Attitudes and Practices. *Computing in Science and Engineering*, 15(6):74–80, 2013. doi:10.1109/MCSE.2013.133.

[13] Sarah Berenji Ardestani, Carl Johan Hakansson, Erwin Laure, Ilja Livenson, Pavel Stranák, Emanuel Dima, Dennis Blommesteijn, and Mark van de Sanden. B2share: An open escience data sharing platform. In *2015 IEEE 11th International Conference on e-Science (e-Science)*, pages 448–453. IEEE, 2015.

[14] Damien Lecarpentier, Peter Wittenburg, Willem Elbers, Alberto Michelini, Riam Kanso, Peter Coveney, and Rob Baxter. EUDAT: a new cross-disciplinary data infrastructure for science. *International Journal of Digital Curation*, 8(1):279–287, 2013. doi:10.2218/ijdc.v8i1.260.

[15] Rat für Informationsinfrastrukturen. Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland. `http://www.rfii.de/download/rfii-empfehlungen-2016`, 2016.

[16] Benjamin Gläßle, Maximilian Hanussek, Felix Bartusch, Volker Lutz, Ulrich Hahn, Werner Dilling, Thomas Walter, and Jens Krüger. de.NBI Cloud Storage Tübingen. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 201–215. TLP, Tübingen, 2019. doi:10.15496/publikation-29054.

[17] Konrad Meier, Björn Grüning, Clemens Blank, Michael Janczyk, and Dirk von Suchodoletz. Virtualisierte wissenschaftliche Forschungsumgebungen und die zukünftige Rolle der Rechenzentren. In *10. DFN-Forum Kommunikationstechnologien, 30.-31. Mai 2017, Berlin, Gesellschaft für Informatik eV (GI)*, pages 145–154, 2017. `https://dl.gi.de/bitstream/handle/20.500.12116/473/paper13.pdf`.

[18] Jonathan Bauer, Dirk von Suchodoletz, Jeannette Vollmer, and Helena Rasche. Game of Templates. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 245–262. TLP, Tübingen, 2019. doi:10.15496/publikation-29057.

[19] Deutsche Forschungsgemeinschaft (DFG). Leitlinien zum Umgang mit Forschungsdaten. https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf, September 2015.

[20] Bundesministerium für Bildung und Forschung (BMBF). Leitlinien für das neue EU-Rahmenprogramm für Forschung und Innovation. https://www.bmbf.de/files/Bundesregierung_FP9_Leitlinienpapier_September_2017.pdf, September 2017.

[21] Hochschulrektorenkonferenz (HRK). Management von Forschungsdaten – eine zentrale strategische Herausforderung für Hochschulleitungen – Empfehlung der 16. HRK-Mitgliederversammlung. https://www.hrk.de/positionen/beschluss/detail/management-von-forschungsdaten-eine-zentrale-strategische-herausforderung-fuer-hochschulleitungen, May 2014.

[22] Hochschulrektorenkonferenz (HRK). Wie Hochschulleitungen die Entwicklung des Forschungsdatenmanagements steuern können. Orientierungspfade, Handlungsoptionen, Szenarien – Empfehlung der 19. HRK-Mitgliederversammlung. https://www.hrk.de/positionen/beschluss/detail/wie-hochschulleitungen-die-entwicklung-des-forschungsdatenmanagements-steuern-koennen-orientierungsp, November 2015.

[23] Alex Ball. *Review of data management lifecycle models*. University of Bath, IDMRC, 2012. https://purehost.bath.ac.uk/ws/portalfiles/portal/206543/redm1rep120110ab10.pdf.

[24] http://chittagongit.com/icon/webpage-icon-10.html.

[25] http://chittagongit.com/icon/online-form-icon-11.html.

[26] http://chittagongit.com/icon/website-search-icon-17.html.

[27] Gregor Cresnar. https://www.flaticon.com/free-icon/database_159252#term=database&page=1&position=2. Icon made by [Gregor Cresnar] from www.flaticon.com.

[28] https://www.openarchives.org/images/OA200.gif.

[29] Linda Cantara. METS: The metadata encoding and transmission standard. *Cataloging & classification quarterly*, 40(3-4):237–253, 2005.

[30] PREMIS Working Group and others. PREservation Metadata: Implementation Strategies, 2004.

[31] Herbert Van de Sompel, Michael L. Nelson, Carl Lagoze, and Simeon Warner. Resource harvesting within the OAI-PMH framework. *D-lib magazine*, 10(12), 2004.

[32] Christopher B. Hauser and Jörg Domaschka. ViCE Registry : An Image Registry for Virtual Collaborative Environments. In *2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 82–89, 2017. doi:10.1109/CloudCom.2017.11.

[33] Martin Barisits, Thomas Beermann, Frank Berghaus, Brian Bockelman, Joaquin Bogado, David Cameron, Dimitrios Christidis, Diego Ciangottini, Gancho Dimitrov, Markus Elsing, Vincent Garonne, Alessandro di Girolamo, Luc Goossens, Wen Guan, Jaroslav Guenther, Tomas Javurek, Dietmar Kuhn, Mario Lassnig, Fernando Lopez, Nicolo Magini, Angelos Molfetas, Armin Nairz, Farid Ould-Saada, Stefan Prenner, Cedric Serfon, Graeme Stewart, Eric Vaand ering, Petya Vasileva, Ralph Vigne, and Tobias Wegner. Rucio – Scientific Data Management. *arXiv e-prints*, page arXiv:1902.09857, February 2019. `https://arxiv.org/abs/1902.09857`.

[34] FAIR principles for data stewardship. *Nature Genetics*, 48(4):343–343, 2016. doi:10.1038/ng.3544.

[35] Heather A. Piwowar, Roger S. Day, and Douglas B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, 2(3), 2007. doi:10.1371/journal.pone.0000308.

[36] Daniel Garijo, Sarah Kinnings, Li Xie, Lei Xie, Yinliang Zhang, Philip E. Bourne, and Yolanda Gil. Quantifying reproducibility in computational biology: The case of the tuberculosis drugome. *PLoS ONE*, 8(11):1–11, 2013. doi:10.1371/journal.pone.0080278.

[37] Jun Zhao, Jose Manuel Gomez-Perez, Khalid Belhajjame, Graham Klyne, Esteban Garcia-Cuesta, Aleix Garrido, Kristina Hettne, Marco Roos, David de Roure, and Carole Goble. Why Workflows Break – Understanding and Combating Decay in Taverna Workflows. In *2012 IEEE 8th International Conference on E-Science*, pages 1–9, 2012. doi:10.1109/eScience.2012.6404482.

[38] Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):1–20, 2017.

[39] Carl Boettiger. An introduction to Docker for reproducible research. *SIGOPS Operating Systems Review*, 49(1):71–79, 2015. doi:10.1145/2723872.2723882.

[40] Felix Bartusch, Maximilian Hanussek, and Jens Krüger. Containerization of Galaxy Workflows increases Reproducibility. In *Proceedings of the bwHPC Symposium*, pages 16–19, 2017. doi:10.15496/publikation-25200.

[41] Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt. bwForCluster NEMO. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors,

Proceedings of the 5th bwHPC Symposium, pages 29–50. TLP, Tübingen, 2019. doi:10.15496/publikation-29041.

[42] Felix Bührer, Anton J. Gamel, Benoît Roland, Benjamin Rottler, Markus Schumacher, and Ulrike Schnoor. Integration of NEMO into an existing particle physics environment through virtualization. In *Proceedings of the 5th bwHPC Symposium*, pages 187–200. TLP, Tübingen, 2019. doi:10.15496/publikation-29053.

[43] Felix Bührer, Frank Fischer, Georg Fleig, Anton Gamel, Manuel Giffels, Thomas Hauth, Michael Janczyk, Konrad Meier, Günter Quast, Benoît Roland, Ulrike Schnoor, Markus Schumacher, Dirk von Suchodoletz, and Bernd Wiebelt. Dynamic Virtualized Deployment of Particle Physics Environments on a High Performance Computing Cluster. *Computing and Software for Big Science*, 2018.

[44] Jens Krüger, Volker Lutz, Felix Bartusch, Werner Dilling, Anna Gorska, Christoph Schäfer, and Thomas Walter. Bioinformatics and Astrophysics Cluster (BinAC). In Sabine Richling, Martin Baumann, and Vincent Heuveline editors, *Proceedings of the 3rd bwHPC Symposium*, pages 91–95. Heidelberg: heiBOOKS, 2017. doi:https://doi.org/10.11588/heibooks.308.418.

[45] Christoph Heidecker, Matthias J. Schnepf, Florian von Cube, Manuel Giffels, and Günter Quast. Dynamic Resource Extension for Data Intensive Computing with Specialized Software Environments on HPC systems. In *Proceedings of the 5th bwHPC Symposium*, pages 161–172. TLP, Tübingen, 2019. doi:10.15496/publikation-29051.

Management of Research Data in Computational Fluid Dynamics and Thermodynamics

Björn Selent[1], Hamzeh Kraus[2], Niels Hansen[2], Björn Schembera[3], Anett Seeland[4,5] and Dorothea Iglezakis[5]

[1]Institute of Aerodynamics and Gasdynamics, University of Stuttgart, Germany;
[2]Institute of Thermodynamics and Thermal Process Engineering, University of Stuttgart;
[3]High-Performance Computing Center Stuttgart;
[4]Communication and Information Center, University of Stuttgart;
[5]University Library, University of Stuttgart

The performance increase of the available resources in the HPC area offers the possibility to investigate fundamental questions of fluid mechanics with numerical tools in high temporal and spatial resolution. In particular, turbulent flows, which account for the largest part of flow processes in nature and technology and do not exhibit a closed analytical solution, can be investigated with increasing precision. Nowadays, one trillion data points are stored and processed per simulation. It is not clear from the outset which data is relevant for understanding the physical processes. Therefore, hundreds to thousands of simulations are carried out in the course of a research project in order to investigate the influence of individual parameters. Similarly molecular simulations underwent a huge increase in algorithmic and technical performance making it now simple to generate large amounts of data. The pure amount of data can be reduced by suitable data compression algorithms. However, it remains an important and challenging task to manage these simulations in a structured way to ensure reproducibility, retrievability and clarity. Equally important is the subsequent question of how the methodology, process and results of the research project can be secured in the long term and shared if necessary. So far, publication of the data is not anchored in the professional culture and is not easy to achieve due to the technical circumstances. Archiving data for more than 3-4 years seems neither possible nor sensible. Based on this initial situation, members of the IAG and ITT, in cooperation with the infrastructure facilities (UB, TIK, HLRS) of the University of Stuttgart, develop and test a working process in which, immediately after data generation, metadata is automatically extracted from log and input files, supplemented by rarely changing information on authors and projects, and stored in the DaRUS data repository. The repository, based on the open-source software *Dataverse*, is used for the local administration of the data. The easy searchability of the descriptive data helps to reuse the existing data and to document the research process. Last but not least, the data already described can be published or meaningfully archived without much additional effort. For the future, we are also striving for clear criteria for the selection, quality control and retention period of the data and mechanisms for the automated linking of datasets.

Björn Selent et al.: Management of Research Data in Computational Fluid Dynamics and Thermodynamics.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 128–139. DOI: https://doi.org/10.11588/heibooks.598.c8422

1. Initial Situation and Requirements

1.1. Fluid Mechanics

Fluid mechanics is a branch of mechanics that considers the motion of liquids and gases and the forces associated with them. Besides analytical and experimental methods, numerical fluid mechanics (CFD) is a means of capturing and analyzing fluid mechanical processes. CFD facilitates the formulation and calculation of the underlying equations of conservation for mass, momentum and energy even if there is no closed solution available. In comparison to experiments, complex geometries and environmental conditions can also be varied more easily and economically. Numerical methods have been used in

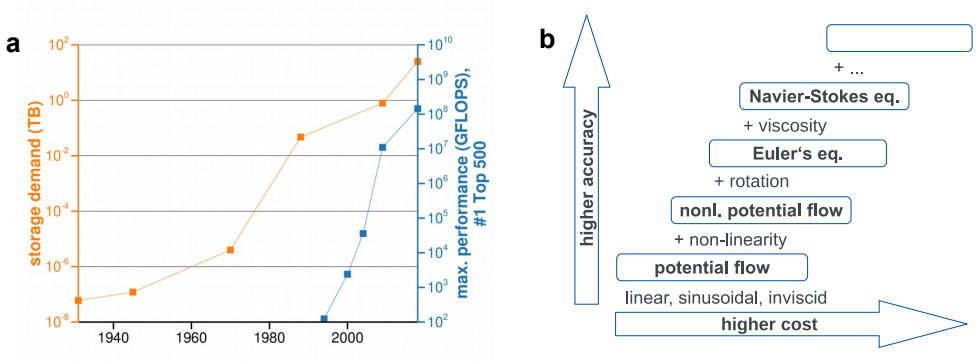


Figure 1.: (a) Storage demand and performance increase of fastest supercomputers. (b) Complexity over cost of equations governing fluid flows.

fluid mechanics for about 90 years (e.g. [3]), albeit they saw their breakthrough with the arrival of digital computers. Since 1990 the fastest supercomputer has seen a performance boost of six orders of magnitude as can be seen in Figure 1(a). This massive increase in computational performance has led to the ability to use numerical methods based on more evolved equations which model the physical processes in much more detail (cf. Fig.1(b)). The gain in additional accuracy though, led to an increase in both the complexity of the numerical set-up and the associated computational effort. This is exemplified in Figure 1(a) by plotting the storage demand of recent CFD simulations. The amount of data produced by recent computations lies in the range of several trillion data points ($\mathcal{O}(10)$ TB). The extended parameter space furthermore demands more simulations to merely test the computational set-up. And finally when the production set-up has been established the number of parameters to be examined is also a lot larger than in previous decades. This makes a proper management of the computational campaign and the associated data an ever increasing and urgent demand. Data management in this context is not limited to the eventually published scientific article but to all data necessary to produce these published results along the complete research cycle.

The remainder of the present text describes the development of tools necessary to support researchers in the handling of research data and processes throughout the full lifespan of their respective project and beyond.

In order to develop a suite of tools for research data management (RDM) five fundamental stages of a research cycle stage have been defined. These stages albeit deduced from a particular project in the field of CFD may, due to their general character, more or less be found in any research area. Each of the stages has specific demands on the RDM which are partially induced by mutual dependencies as shown in Figure 2.

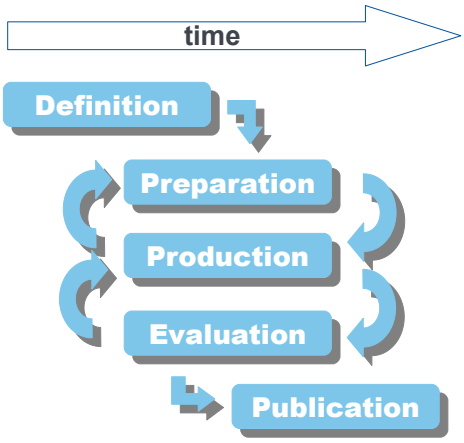


Figure 2.: Stages of research cycle.

The *Definition* stage consists of determining the project's goals, matching them with the state-of-the-art knowledge in its scientific field, collecting information of existing tools and methods to accomplish these goals and making a timetable of the project's progress. These tasks are usually data-poor, i.e. the recordings and collections consist of a manually manageable amount of documents. The data is typically generated on a personal workstation and transferred to any type of repository system. The RDM has to take care of the data being findable and saved from loss, at least for the project's lifespan.

The *Preparation* stage is devoted to preparing the numerical set-up. This includes so-called mesh studies to identify the numerical parameters associated with the computation and a thorough testing of any additional numerical methods involved with the models. Assuming a 3-dimensional computational domain, a set of four different boundary conditions and three additional parameters, a total number of $N = 3^3 \cdot 4^3 \cdot 3 = 5184$ simulations can easily occur. The simulations are usually run on some kind of High Performance Computing (HPC) system. At this stage it is vitally important that the RDM processes are automated and transparent for the researcher. The important metrics of the simulations have to be captured and transferred into a searchable format on a long lasting repository. The main focus of RDM thus lies in the preservation and documentation of the processes leading to a functional environment for succeeding numerical simulations and in making this information retrievable at later times. Albeit simulations at later stages are done for settings determined during the preparation, what might occur is that some limited additional testing is necessary. The RDM tools should therefore allow for dynamic changes at any time.

During the *Production* stage the number of simulations is usually one order smaller than during preparation. But now the results, i.e. actual output data, are valuable as well, at least in parts. The processes during the production stage involve further tools to extract and summarize the results from large datasets. This includes programs to compute derived data and procedures to visualize the results in graphs and pictures. Thus any suitable RDM should allow extending the simulations' metrics with post-processing steps and the generated graphics and data-tables. The RDM has to ensure a strong link between the set-up and the result metrics in order to enable any repetition at later times.

The *Evaluation* stage is strongly interwoven with the *Production* stage, the main difference being that the results are not only observed but also checked for plausibility and that results from several runs are compared with each other and with existing literature. This almost automatically leads to a further requirement for the RDM at this point, namely compound results depicted in a graph should be assigned to the correct simulation's metrics. Finally it is worth mentioning that the integrity of the data has to be ensured as well in order to preserve good scientific practice. *Preparation*, *Production* and *Evaluation* are data-rich stages which lead to the strong demand that any RDM measures should alleviate the work flow and must not add any additional load. Otherwise the acceptance of any RDM is strongly undermined.

The final stage of the cycle consists of the *Publication* of the conclusions and discoveries made. Here the typical requirements for published data are to be applied, i.e. it has to be findable, accessible, inter-operable and re-usable (FAIR) [8] for the public, verifiable and long-lasting. The repository system therefore has to allow access from the outside without endangering any non-published data on the same resources form being compromised. Nevertheless the results shown in any publication should readily be available in both graphical and tabulated form in order to allow third parties to compare them with their own results. In ideal circumstances, the whole production tool-chain should be made available to third parties, including the programs to compute the data along with all parameters and post-processing routines. To achieve this, it is advisable to resort to open and established standard formats and open source programs whenever possible. There are first attempts to share and disseminate large datasets in the community [2, 6], but there is a lack of standards to describe and exchange these datasets in a structured way.

1.2. Thermodynamics

Thermodynamics provides the scientific fundamentals for energy and material conversion processes which occur in almost all areas of modern societies. This makes thermodynamics a key discipline for pressing technological issues in our society such as the rising demand on energy supply and storage, the development of new materials and the optimization of chemical and biotechnological processes. The ITT conducts research in the fields of molecular thermodynamics, molecular simulation as well as simultaneous process and solvent design. In recent years molecular simulation in particular has matured into a powerful tool for predicting microscopic processes and material properties with a high degree of versatility. However this versatility comes at a price in that it is increasingly difficult for researchers to find their way through the maze of available computational techniques and models.

In principle, molecular models and algorithms can be separated into disjunct groups, but in practice a more or less strong coupling exists, making the connection between the molecular model and the predicted substance property sometimes ambiguous. This may also depend on algorithmic details that are not, or incompletely outlined [1, 4, 7]. As a consequence the reproducibility of literature data, only based on the textual information, is often hardly possible.

In the field of thermodynamic property measurement, the Thermodynamics Research Center (TRC) at the National Institute of Standards and Technology (NIST) established a cooperation with various journals more than 10 years ago in order to develop standards for the storage and exchange of thermophysical property data (`https://trc.nist.gov/ThermoM-L.html`). Numerous journals oblige authors to adhere to these standards. In contrast, molecular simulation standards for the exchange and communication of simulation conditions and results have yet to be developed.

Similar to the fluid mechanics workflow towards a publication, different stages can be defined for a molecular simulation study. The definition stage is comprised of determining the systems that need to be examined in order to explore a specific property, and how many variations of e.g. solvent composition, temperature or pressure are required. The system will then be prepared in the next stage, by generating an initial structure containing all atom positions, and files containing all the systems parameters. These two stages are critical and play a major role, since the actual production takes a long simulation time to create physically and statistically adequate data. Thus deep understanding and initial test cases are required before moving to the production stage.

In short, a molecular dynamics simulation numerically integrates Newton's equation of motion for all atoms. Due to physical and numerical constraints, an integration time step is in the range of femto seconds (10^{-15} s). Depending on the scientific question, simulation times up to micro seconds (10^{-6} s) might be needed, creating a computational time up to a month or longer in some cases. Once a physically converged system is achieved, dynamic and static system properties, like diffusion coefficient and density distribution, can be calculated in the evaluation stage. Depending on the results, further simulations with different parameters might be needed. In the final stage, the computed properties will be published. Subsequent to acceptance however, the produced data will be archived or discarded most of the time and also rarely made publicly available due to a data size of hundreds of gigabytes. Therefore a database for the produced data will be needed to incite more clarity into the simulation model and workflow. The approach of a data repository for input and output files of a simulation might not be a replacement for a structured simulation workflow, but it surely is an approach leading in the right direction.

2. Realization

The goal of the implementation was to support the two central requirements—to provide an overview over a large amount of data with an expanded parameter space and to prepare and link data from the different stages (preparation, production, evaluation) for a FAIR way [8] to archive and publish.

The basis for both requirements is a structured description, capturing not only all relevant search criteria to make the data findable but also a representation of the research process and the observed system.

To create and handle this metadata description together with the actual data, a toolchain that supports the researcher with automation, a safe location for the data and an intuitive interface for interaction with the data is necessary.

2.1 Description of the Data - EngMeta as a Metadata Scheme for Engineering Data

The first step of building a metadata description model was the identification of relevant categories: For search and overview, the most important criteria besides general descriptive information like author, year, and project turned out to be quite discipline-specific information about the observed system and their components (like force fields in thermodynamics, the observed and controlled variables), parameters of the used methods (like filtering in aerodynamics or the simulation metrics) and parameters of the observation/simulation itself (like the grid). To make the data understandable and reproducible information about the computing environment (for computational engineering) and used instruments (for experimental engineering), the used software and tools are also necessary.

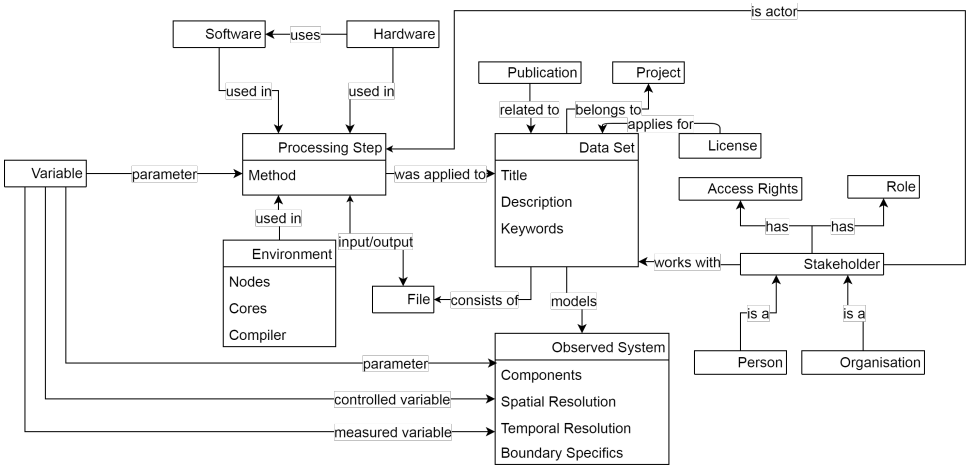


Figure 3.: Object model to describe engineering data.

The object model in Figure 3 visualizes the objects of the research process whose properties are important for the description of the data. Central are not only the characteristics of the data itself but also the steps taken to generate this data. Important for the local management of hot data was also the possibility to mark and comment on negative results. Starting from this object model we looked for existing metadata schemata covering these properties. As we did not find a scheme that addresses all criteria, we created an application profile building on existing standards like DataCite (for general descriptive information), PREMIS (for technical information), CodeMeta (for describing software), ExptML (for information about experimental instruments) and PROV (for a description of the process) and added additional description fields for the discipline-specific information. Figure 4 shows the main metadata categories and fields that later became EngMeta, a metadata model for computational engineering applications[5]. EngMeta is serialized in an XSD-scheme and publicly available.[1] For the data from the *Definition* stage, the general descriptive metadata fields of EngMeta are sufficient with the possibility to link the data to the project (context → project). For the *Preparation* stage, the capture of the simulation parameters (provenance → method/parameters, observed system → bounding

[1] `https://www.ub.uni-stuttgart.de/engmeta`, Last accessed 26/04/2019

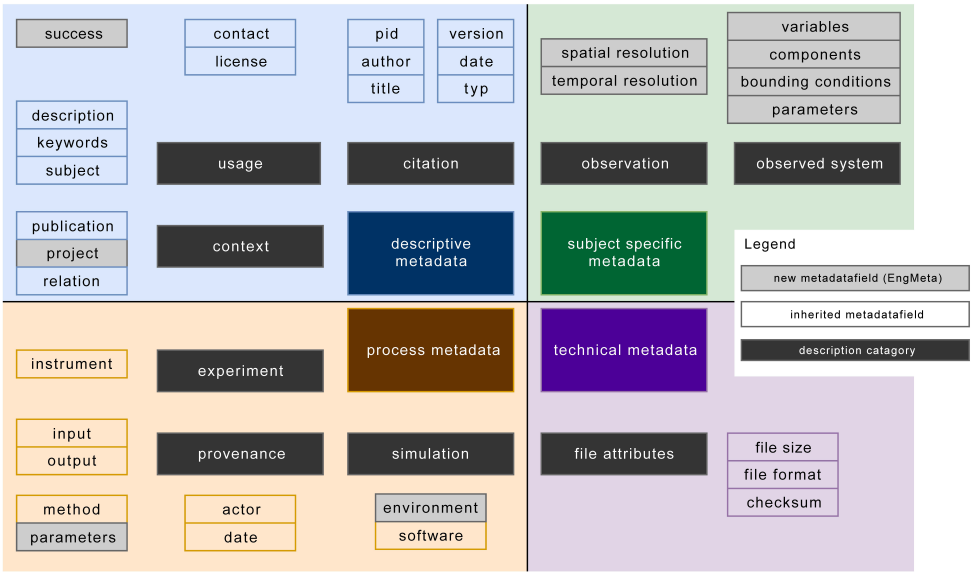


Figure 4.: Metadata categories of EngMeta.

conditions/components) and the documentation of successful and failed configurations (usage → success) is more crucial than storing the actual data. To enable reproduction and understandability, the data of the *Production* stage also has to be connected with a detailed description of the tools and environment (simulation → software/environment), methods (provenance → method/parameters) and the definition of input and output data (provenance → input/output). For the *Evaluation* stage, EngMeta provides different relation types (context → relation) to assign data from different runs and stages to each other.

2.2. Extracting Structured Metadata Using Automated Metadata Extraction

A pure metadata model leaves the annotation of metadata to the researchers. Since the effort for tagging the data in such a expanded parameter space is extraordinary high, we developed an automated metadata extraction as a second step. Lots of meta information already available through the simulation outputs can be extracted without manual intervention. This especially holds true for process metadata, such as information on the computational environment, as well as for domain-specific metadata, such as information about the controlled variables. This information is mostly available through job-, log- and output-files of the simulation run. Technical metadata is also easy to collect, since it is available through file system attributes, whereas descriptive metadata is all but impossible to collect, since it describes the data from a higher point of view. With this background in mind, we developed an automated approach as a light-weight Java tool. The metadata extraction is generic in two ways: First, the metadata information to extract from the simulation codes is configurable via a specific file. This configuration file defines the following for each metadata key: In which file can it be found? What is

the search key required to find it? What is the delimiter to separate the key from the value? We chose this approach of configuring the extraction outside the code because only in this way we could overcome one's inhibitions. This configuration file only has to be written once for a simulation code or at least only once for many runs to be accomplished. User-defined descriptive metadata can be put into extra files which can be parsed along with all the output files from the simulation runs. Second, the format of the output files, which includes all the extracted metadata values, can be changed. This was originally implemented for EngMeta, returning an XML document according to the EngMeta specification. Changing the format can only be accomplished by writing output classes in Java. However, the extractor tool additionally outputs the plain information as a (key, value) listing, so one can also convert from this plain text information to any user-defined format.

The tool itself was implemented with the Java Scanner API[2], which is feasible for log files in the size of < 20 MB (the usual log files size of simulation output is only some MegaBytes). However, we developed a parallel solution using the Apache Spark Data Analytics Framework[3], being capable of analyzing multiple 100MB of textual log-files in a very short period of time. The tool can be run directly after the actual simulation, writing a sub-directory *.metadata*. This sub-directory now contains two files: *metadata.txt* contains the plain listing whereas *engMeta.xml* contains the same information according to the EngMeta model specification as an XML document. Since the tool is written in Java, it can be run on several operating systems and architectures, ranging from workstations (including Linux and Windows machines) to clusters (tested on bwUniCluster[4]) and high-performance computing systems (tested on Cray XC40 „Hazel Hen“[5]).

2.3. Overview over the Data - Using Dataverse as a Data Repository for Metadata Management

Without a repository to index, store and manage the metadata as well as a link to the data itself, both the scheme and the extracted metadata are useless for the researchers. So, the third step was to identify a system, that is able to handle EngMeta as a metadata scheme, has a detailed role and rights management and offers an intuitive interface for data management and search. After an overview of existing software systems for (data) repositories, three systems were shortlisted for a detailed comparision: Dataverse[6], DSpace[7] and Invenio[8]. While none of the systems met all the requirements, we chose Dataverse as our data platform primarily for the following reasons: The ability to define custom metadata per collection, programmable APIs to enable automation, an intuitive web interface, an active developer community, and optimization for data instead of text publications.

[2] `https://docs.oracle.com/javase/8/docs/api/java/util/Scanner.html`, Last accessed 26/04/2019
[3] `https://spark.apache.org/`, Last accessed 26/04/2019
[4] `https://www.scc.kit.edu/dienste/bwUniCluster.php`, Last accessed 26/04/2019
[5] `https://www.hlrs.de/systems/cray-xc40-hazel-hen/`, Last accessed 26/04/2019
[6] `www.dataverse.org`, Last accessed 03/05/2019
[7] `https://duraspace.org/dspace/`, Last accessed 03/05/2019
[8] `https://invenio-software.org/`, Last accessed 03/05/2019

Dataverse is an Open Source software platform for data repositories, developed at Harvard IQSS, supported by an active international community of contributors. It is the software base of the data repository of Harvard University and another 41 installations existing around the world. Dataverse offers a user interface to upload, search, manage, share and publish datasets and different APIs for programmable access to the functionalities of the platform. A differentiated rights and role management gives the possibility to share datasets with the desired community—be it within the own working group, with international project partners or the whole research community. Each dataset is described with structured metadata and is managed within a so called dataverse, a collection of datasets with its own set of metadata categories and rights. So, this system allows to create discipline-specific data spaces inside an institutional repository. Dataverse comes out of the box with a set of metadata categories: general descriptive metadata and some discipline-specific additions for the social and geospatial sciences and astronomy. Basing on this software, DaRUS – the data repository of the University of Stuttgart – offers all its members and partners the possibility to manage, share and publish their datasets. We added the procedural and discipline-specific parts of EngMeta to allow the differentiated description of datasets from our two application areas. Normally, Dataverse is meant for the publication of datasets. With DaRUS, we are using the system mainly for the management of the research data within a research group or project, to give an overview over the produced datasets and to help keep the datasets findable and understandable for third parties.

For the data itself, object storage (Netapp StorageGRID) was introduced at our university. Unlike other storage solutions, such as file systems, data on an object storage is organized as objects with an unique identifier, some metadata information and the data itself. Object storage has its strengths in write-once-read-many scenarios, typical for research data, and can easily be scaled for huge amounts of objects. Further, it can simultaneously handle different physical storage, like discs and tapes, enabling cost-efficiency for hot and cold data. Eventually, the identifier URL does nor change, even if the object is moved to a different physical location, which makes it easy to assign DOIs to objects. To avoid data loss, the objects are replicated between two independent data centers. Dataverse supports objects storage via the S3-interface.

2.4. Automation of Data Ingest

After extracting all available metadata from the simulation files, we had to map EngMeta as a deeply nested XML schema to the flat key-value structure of Dataverse's metadata in order to automate the creation of datasets in DaRUS. For this purpose a Python API was developed. It accesses DaRUS through a provided REST-protocol. The API takes as input the generated metadata file from the extraction and other user-specific entries, like the desired database and repository name, and automatically uploads the specified files into DaRUS.

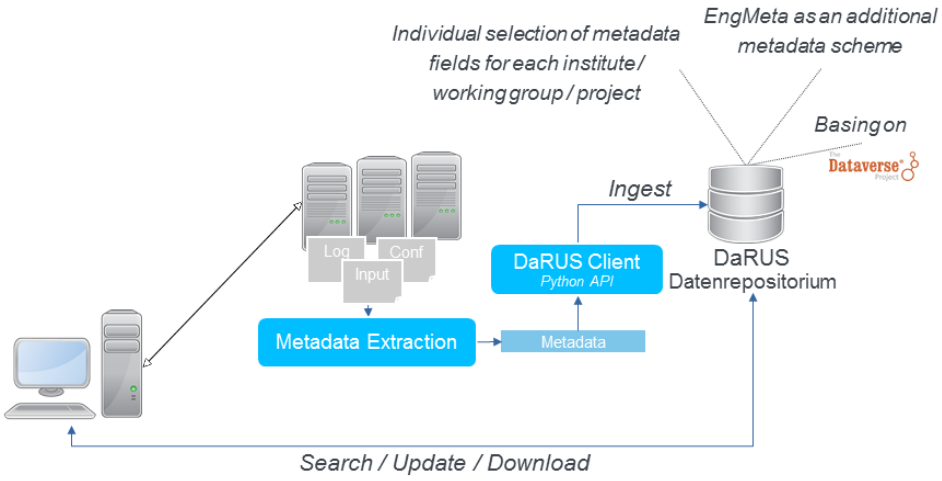


Figure 5.: Interaction of Repository, Metadata Extraction and Automation

3. Experiences

The whole process illustrated in Figure 5 is evaluated since the end of 2018 while part of the components were still in development. The major findings will be discussed briefly.

First and foremost we saw that the cooperation between researchers and infrastructural services has been very productive. Especially the metadata model could not have been developed without this combined effort. The researchers contributed deep insight to their research process and the domain-specific view while the infrastructural services contributed their knowledge on metadata standards, repositories and system integration. With the combination of the individual components we have built a fully functional toolchain for extracting, storing and managing data and processes. Even though there are still open questions we think of this as Proof of Concept for a method to assist scientists at all stages of their work in a transparent fashion without increasing their workload.

One of the major experiences was the impact of automation. We learned that automation is crucial if research data management efforts should be accepted. Without automation researchers add only the basic metadata and describe the data in a non-structured way, i.e. by specifying parameters in a general free-text description field. This approach is not sustainable in terms of FAIR data management, so automation was one of our primary goals.

This led to the development of the automated metadata extraction. One of the findings was that initially, the development of the configuration files for the automated metadata extraction process takes some additional time. One has to choose reasonable search terms and know the output files of the simulation codes well. After this initial extra effort, one can profit from the automated approach. We were able to attach the automated extraction directly to the job scripts of the simulations, so after each simulation's run, the extraction could be performed automatically.

Management of data requires a different view of the data than publishing: Dataset listings in Dataverse focus on the citation of data, but for an overview of ones own data there has to be a sortable and filterable view focused on the parameters of the data.

Dataverse offers a full text search and search facets that are good to filter on discrete metadata categories. For continuous parameters, we need different tools to filter datasets like range queries or a sortable tabular overview over the datasets.

3.1. Future Work

The amounts of data usually generated in fluid mechanics and thermodynamics are expensive to store and time-consuming to move. Dataverse has its root in the social sciences, and is therefore not designed for file ingest in the gigabyte and terabyte range. Files of this magnitude produce time-outs during the ingest and download. To enable uploads of data larger than 2 GB, the time outs for API calls are prolonged. But still, this procedure does not scale sufficiently for datasets larger than 100 GB. Therefore, we work on different ways to handle, save and manage these files. One starting point is to leave the data where it is and to connect Dataverse to several different data backends, be it through links in the metadata or through connecting Dataverse with different storage technologies in distinct locations. Another approach is the development of clear criteria to decide which part of the data shall be stored for what period of time. The data from the *Preparation* stage can be deleted quite soon, while published data has to be securely saved for a long period. In order to use the available storage space efficiently, we will need measures for the usefulness of data and automation not only for the ingest, but also for the deletion of data (records) no longer used.

Even if the data is read in automatically, the challenge remains to link data from different stages in the research process in a meaningful and automated way. Data from different simulation runs are linked, compared and integrated during subsequent research steps as described in Section I. A further goal for the future is to automate the mapping of these links into the metadata of the data records concerned and to represent and visualize the research process in a comprehensible and reproducible way.

4. Conclusions

To manage and handle large amounts of data from computational fluid dynamics and thermodynamics we used a data repository based on the Dataverse platform together with an automated description and upload of the data. We use Dataverse not only in its actual function – publishing – but in addition for managing and sharing data within a defined public. There are still steps to take in making this system a feasible management of this data, especially in the linking of datasets and the handling of large data files.

Due to their configurability, the approaches developed can be easily transferred to other fields. This applies in particular to disciplines that also use (simulation) codes to obtain their results. DaRUS is already used as an institutional repository. The content is adapted to the different subject areas through the individual configuration of the metadata categories. The automated metadata extraction can principally work on all text-based log, input or user-generated files that contain metadata information in a semi-structured form. It is currently being tested by working groups from various disciplines.

Thanks

The DIPL-ING project is funded by the Federal Ministry of Education and Research under grant number 16FDM008.

Bibliography

[1] Loeffler, Hannes H., et al. “Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages.” Journal of Chemical Theory and Computation 14.11 (2018): 5567-5582.

[2] Meneveau, Charles and Marusic, Ivan “Turbulence in the Era of Big Data: Recent Experiences with Sharing Large Datasets.” In A. Pollard, L. Castillo, L. Danaila and M. Glauser (eds.), Whither Turbulence and Big Data in the 21st Century?. Springer, (2017): 497-507

[3] Rosenhead, L. “The Formation of Vortices from a Surface of Discontinuity.” Proc. Roy. Soc. London A, 134.823 (1931): 170-192.

[4] Schappals, Michael, et al. “Round Robin Study: Molecular Simulation of Thermodynamic Properties from Models with Internal Degrees of Freedom.” Journal of Chemical Theory and Computation 13.9 (2017): 4270-4280.

[5] Schembera, Björn, and Iglezakis, Dorothea. “The Genesis of EngMeta-A Metadata Model for Research Data in Computational Engineering.” Research Conference on Metadata and Semantics Research. Springer, Cham, (2018).

[6] Sillero, Juan A. and Jiminéz, Javier “Public Dissemination of Raw Turbulence Data.” In A. Pollard, L. Castillo, L. Danaila and M. Glauser (eds.), Whither Turbulence and Big Data in the 21st Century?. Springer, (2017): 509-515.

[7] van Gunsteren, Wilfred F., et al. “Validation of Molecular Simulation: An Overview of Issues.” Angewandte Chemie International Edition 57.4 (2018): 884-902.

[8] Wilkinson, Mark D., et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Scientific data 3 (2016).

Vom Papier zur Datenanalyse. „Neue“ historische Forschungsdaten für die Wirtschaftswissenschaften

Sabine Gehrlein, Jan Kamlah, Matthias Pintsch, Irene Schumm und Stefan Weil
Universitätsbibliothek Mannheim, Deutschland

1. Ausgangslage

Die empirische Wirtschaftsforschung hat eine lange Tradition und damit auch die Erhebung und Bereitstellung der benötigten Forschungsdaten. Neben den Akteuren der amtlichen Statistik hat sich ein differenzierter und internationaler Kreis von kommerziellen Datenanbietern etabliert, die ihre Produkte hochpreisig anbieten. Zu den Schwergewichten in diesem Markt gehören etwa Bloomberg, Bureau van Dijk, Thomson Reuter's, Standard & Poor's oder Bisnode. Dabei ist die Wissenschaft nicht die Hauptzielgruppe dieser Unternehmen; im Fokus stehen eher große Finanz- und Versicherungsunternehmen und deren Marktforschungsabteilungen. Ein Blick in das Datenbank-Infosystem (DBIS) zeigt, dass aktuell nur eine kleine Zahl deutscher Hochschulen die notwendigen Mittel aufbringen kann, um ihren Forscherinnen und Forschern Zugang zu den Produkten der kommerziellen Datenanbieter zu verschaffen.

Selbst wenn dies der Fall ist, ist damit noch keine umfassende Versorgung der empirischen wirtschaftswissenschaftlichen Forschung gewährleistet. Die verfügbaren kommerziellen Forschungsdaten sind geographisch recht einseitig ausgerichtet: Insbesondere für die USA existiert ein großes und differenziertes Angebot, für Deutschland hingegen sind deutlich weniger Ressourcen verfügbar.

Zudem lassen sich viele Fragestellungen nur dann methodisch sauber bearbeiten, wenn die Forschungsdaten über einen ausreichend langen Zeitraum konsistent verfügbar sind. Diese langen Zeitreihen können die kommerziellen Datenanbieter nur selten liefern. Da die primären Kunden in der Finanz- und Versicherungsbranche vor allem an aktuellen Daten und Analysen interessiert sind, wird in die Pflege und Aufbereitung historischer Daten kaum investiert. Abbildung 1 zeigt die zeitliche und inhaltliche Abdeckung von drei zentralen kommerziellen Unternehmensdatenbanken, aus der anschaulich hervorgeht, dass historische Phänomene wie z. B. die Ölkrise der 70er Jahre auf dieser Datengrundlage nicht analysiert werden können.

In Zukunft könnte sich diese Situation sogar noch verschärfen, da viele relevante Wirtschaftsdaten, die online über Webdienste angeboten werden, verschwinden, sobald sie mit der aktuellen Version überschrieben werden. Diese Daten sind dann für die Forschung verloren.

Hier können nicht-kommerzielle Anbieter, darunter auch Bibliotheken, eine wichtige Rolle spielen (vgl. Zumstein 2016). Mit der Expertise, die sie in Fragen des Datenmana-

Sabine Gehrlein et al.: Vom Papier zur Datenanalyse. „Neue“ historische Forschungsdaten für die Wirtschaftswissenschaften.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 140–152. DOI: https://doi.org/10.11588/heibooks.598.c8423

	COMPUSTAT GLOBAL	EIKON /DATASTREAM	ORBIS / AMADEUS
Anbieter	Standard & Poor's	Thomson Reuters	Bureau van Dijk
Zeitraum	1988-heute	Ø Letzte 5 Jahre	Letzte 10 Jahre
Anzahl dt. Firmen	942	ca. 1.000	3,85 Mio.
Bilanzdaten	Ja	Ja	Ja
Governancedaten	Nein	Ja	Ja

Abbildung 1.: Übersicht über die Inhalte gängiger kommerzieller Forschungsdatenangebote (eigene Darstellung)

gements, der Erschließung und der Archivierung elektronischer Informationen aufgebaut haben, sind sie geeignete Akteure, um alte und neue Forschungsdaten zu sammeln, aufzubereiten und dauerhaft für die Wissenschaft verfügbar zu machen.

Das Angebot muss dabei an den aktuellen Bedürfnissen der wissenschaftlichen Nutzerinnen und Nutzer ausgerichtet werden. So ist es in der quantitativen Wirtschaftsforschung heute Standard, mit Auswertungs- und Analyseprogrammen wie Stata, R oder SPSS in kurzer Zeit sehr große Datenmengen zu verarbeiten und so Theorien und Hypothesen zu testen. Fragen, die früher langwierige Datensammlungen und Vorbereitungen voraussetzten, lassen sich so innerhalb weniger Stunden bearbeiten.

Voraussetzung dafür ist, dass Rohdaten in einem maschinell zu verarbeitenden Format und in hoher inhaltlicher Qualität vorliegen. Es reicht daher nicht aus, eine Datenquelle nur zu digitalisieren. Zusätzlich ist mindestens eine nutzbare Volltexterfassung über OCR nötig, besser noch die Überführung der Daten in eine strukturierte Datenbank mit vielfältigen Such- und Downloadmöglichkeiten. Zudem sollten die angebotenen Daten inhaltlich umfassend beschrieben und dokumentiert sowie einfach und kostengünstig zugänglich sein.

2. Das Projekt Aktienführer an der UB Mannheim

Die Universität Mannheim hat ein ausgeprägtes Profil in den empirisch orientierten Wirtschafts- und Sozialwissenschaften, was sich auch in national und international renommierten Fachbereichen für Betriebswirtschaftslehre und Volkswirtschaftslehre widerspiegelt. Voraussetzungen dafür sind insbesondere erfolgreiche Forschungsprojekte und daraus resultierende Publikationen, die auch auf einem erstklassigen Zugang zu Publikationen und Datenbanken beruhen. Dienstleister und Ansprechpartner in diesem Feld ist die Universitätsbibliothek (UB).

Im alltäglichen Austausch mit Forscherinnen und Forschern der Betriebswirtschaftslehre erhielt die UB wiederholt Hinweise darauf, dass es der Forschung am Zugang zu Forschungsdaten im Bereich der deutschen Unternehmens- und Kapitalmarktdaten mangelt. Konkret handelte es sich um Daten aus dem Hoppenstedt-Aktienführer, einer Publikation, die mit ihrem Vorgänger Saling's Börsenpapiere seit dem Jahr 1870 jährlich zu den deut-

schen börsennotierten Unternehmen berichtete. In jährlichen Print-Bänden und ab 2000 auf CD-ROM enthält der Hoppenstedt-Aktienführer detaillierte Daten zur Entwicklung wichtiger Unternehmen und der Volkswirtschaft Deutschlands.

Vor dem Hintergrund eines deutlich kommunizierten Bedarfs von Seiten der Forschung setzte sich die UB Mannheim das Ziel, die wertvollen Daten des Hoppenstedt-Aktienführers in ein zeitgemäßes, elektronisches Format zu überführen und der Forschung zur Verfügung zu stellen. Dieses Projekt erschien naheliegend, da die UB Mannheim aufgrund des Universitätsprofils bestens mit wirtschaftswissenschaftlichen Forschungsdaten vertraut ist und da vorhandene Kompetenzen im Feld der automatisierten Texterkennung (OCR – Optical Character Recognition) eingebracht und weiterentwickelt werden konnten. Diese Kompetenzen spielen auch im Rahmen weiterer Projekte eine große Rolle.[1]

Durch Förderung der DFG war es der UB Mannheim möglich, den Hoppenstedt-Aktienführer im Rahmen zweier Projekte von 2013 bis 2015 sowie von 2017 bis 2019 zu digitalisieren, in eine komfortable Datenbank zu überführen und der Forschung kostenfrei zur Verfügung zu stellen. Durchführung und Ergebnis, insbesondere der zweiten Projektphase, sollen weiter unten genauer diskutiert werden.

In der deutschen Unternehmens- und Kapitalmarktforschung ist der Hoppenstedt-Aktienführer (kurz: Aktienführer) eine bestens etablierte Datenquelle. Bevor der Hoppenstedt-Verlag die Publikation übernahm, erschien sie im Verlag Saling – weswegen manche den Aktienführer auch noch als „Salinger“ kennen. Derzeit liegen die Nutzungsrechte beim Unternehmen Bisnode, von dem die UB Mannheim die Rechte zur Digitalisierung und elektronischen Präsentation erworben hat.

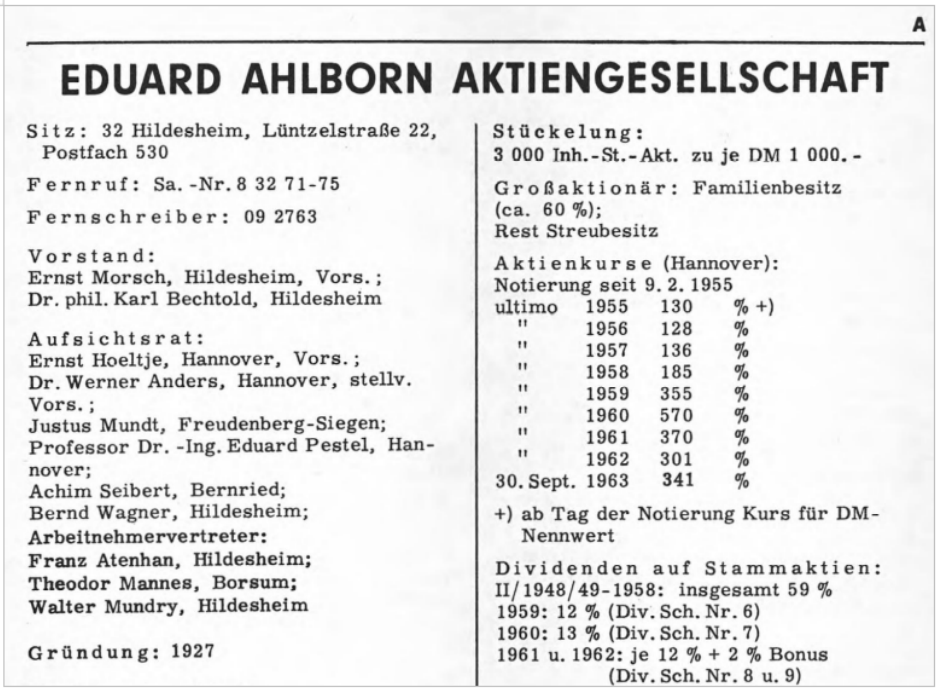
Nach dem Zweiten Weltkrieg erschien der Aktienführer erstmals wieder 1953 und dann ab 1956 in jährlich aktualisierten Print-Ausgaben, die 1999 eingestellt wurden. Ab dem Jahr 1998 und bis 2018 gab es dann eine halbjährlich aktualisierte CD-ROM, und seit 2019 bietet Bisnode die Aktienführer-Daten nur noch elektronisch als Web-Dienst an. Es ist noch zu klären, wie die Daten des Web-Dienstes für die Forschung gesichert und dauerhaft verfügbar gemacht werden können.

Die seit 1956 erschienenen Aktienführer-Bände und CD-ROMs enthalten in standardisierter Form Berichte über deutsche und ausländische Aktiengesellschaften, die an einer deutschen Börse gehandelt werden. In den Berichten finden sich unter anderem Daten zu Firmensitz und Vorstand, zu den Tätigkeitsbereichen und Beteiligungen, zur Aktionärs- und Kapitalstruktur und zu wesentlichen Positionen der Bilanz sowie Gewinn- und Verlustrechnung. Die konstant hohe Datenqualität sowie die lange Publikationshistorie des Aktienführers machen die Daten sehr wertvoll für die auf Deutschland bezogene empirische Wirtschaftsforschung. Abbildung 2 zeigt beispielhaft einen Ausschnitt aus einem Unternehmensprofil.

Durch die stringente Struktur der Unternehmensprofile in diesen Aktienführer-Bänden war es möglich, die OCR-Erkennung und Strukturerschließung weitestgehend zu automatisieren und so mit begrenztem Aufwand eine beachtliche Datenmenge in eine Datenbank zu überführen.

Die ersten Vorläufer des Hoppenstedt-Aktienführers wurden im Jahr 1870 unter dem Namen „Saling's Börsenpapiere“ publiziert. Diese erschienen regelmäßig in verschiede-

[1] Siehe auch: Ausblick – Forschungsdaten an der UB Mannheim.



A

EDUARD AHLBORN AKTIENGESELLSCHAFT

Sitz: 32 Hildesheim, Lüntzelstraße 22, Postfach 530

Fernruf: Sa.-Nr. 8 32 71-75

Fernschreiber: 09 2763

Vorstand:
Ernst Morsch, Hildesheim, Vors.;
Dr. phil. Karl Bechtold, Hildesheim

Aufsichtsrat:
Ernst Hoeltje, Hannover, Vors.;
Dr. Werner Anders, Hannover, stellv. Vors.;
Justus Mundt, Freudenberg-Siegen;
Professor Dr.-Ing. Eduard Pestel, Hannover;
Achim Seibert, Bernried;
Bernd Wagner, Hildesheim;
Arbeitnehmervertreter:
Franz Atenhan, Hildesheim;
Theodor Mannes, Borsum;
Walter Mundry, Hildesheim

Gründung: 1927

Stückelung:
3 000 Inh.-St.-Akt. zu je DM 1 000.-

Großaktionär: Familienbesitz (ca. 60 %);
Rest Streubesitz

Aktienkurse (Hannover):
Notierung seit 9. 2. 1955

ultimo	1955	130	% +)
"	1956	128	%
"	1957	136	%
"	1958	185	%
"	1959	355	%
"	1960	570	%
"	1961	370	%
"	1962	301	%
30. Sept.	1963	341	%

+) ab Tag der Notierung Kurs für DM-Nennwert

Dividenden auf Stammaktien:
II/1948/49-1958: insgesamt 59 %
1959: 12 % (Div. Sch. Nr. 6)
1960: 13 % (Div. Sch. Nr. 7)
1961 u. 1962: je 12 % + 2 % Bonus (Div. Sch. Nr. 8 u. 9)

Abbildung 2.: Ausschnitt eines Unternehmensprofils (Screenshot)

nen Teilen für die verschiedenen deutschen Börsenplätze und auch zu Auslandsbörsen. Im Unterschied zu den Nachkriegs-Aktienführern ist die Information hier aber weniger strukturiert, eher in Fließtext enthalten. Da dies für automatisierte Verfahren eine Herausforderung darstellt, wurde im Aktienführer-Projekt hier zunächst auf die Erstellung einer Datenbank verzichtet. Aber alle Bände wurden digitalisiert und mit OCR bearbeitet, so dass zumindest eine Volltext-Suche verfügbar ist, die eine schnelle Durchsicht der einzelnen Bände ermöglicht.

Die Saling's Börsenpapiere umfassen zeitlich sehr prägende Epochen der deutschen und europäischen Geschichte, was sie zu herausragenden Quellen der Wirtschafts- und Sozialgeschichte macht. Man denke hier an die Industrialisierung im Kaiserreich, die Auswirkungen des Ersten Weltkriegs und die Weltwirtschaftskrise während der Weimarer Republik.

3. Vom Papier zur Datenbank – Erschließungsverfahren für Forschungsdaten am Beispiel des Aktienführer-Datenarchivs

Auf dem Weg vom gedruckten Werk zur Datenbank müssen eine Reihe von Schritten bewältigt werden. Zuerst müssen die Rechte in Zusammenarbeit mit dem Rechteinhaber geklärt werden. Dann sind die gedruckten Werke fachgerecht und mit hoher Qualität zu scannen, so dass dann Bilddaten des Originals vorliegen. Für die Erfassung der in den Bilddaten enthaltenen Informationen gibt es nun zwei Wege – händisch oder automatisch. Bei der händischen Datenerfassung übertragen Menschen die Informationen aus den Bilddaten in ein geeignetes elektronisches Format. Bei der automatischen Datenerfas-

sung übernimmt diese Aufgabe der Computer, wobei die entsprechenden Software-Tools dazu entwickelt werden müssen. Beide Wege wurden im Aktienführer-Projekt beschritten. In der ersten Projektphase entschied man sich für eine händische Datenerfassung, in der zweiten Projektphase dann für ein automatisiertes Vorgehen. Diese Abläufe sollen nun anhand des Aktienführer-Projekts vorgestellt werden.

Rechte am Aktienführer des Hoppenstedt-Verlags ist die Firma Bisnode Inhaberin der Aktienführer-Urheberrechte. Die UB Mannheim erwarb von Bisnode das Recht, alle Bände des Aktienführers sowie seiner Vorgänger zu digitalisieren, im Volltext strukturiert zu erfassen und wissenschaftlichen Einrichtungen kostenfrei zur Verfügung zu stellen. Die Einschränkung auf eine nicht-kommerzielle Nutzung war Bisnode sehr wichtig, um eigene Geschäftsmodelle nicht zu gefährden. Damit kann das Aktienführer-Datenarchiv nur wissenschaftlichen Einrichtungen und Einzelpersonen mit nachgewiesenem wissenschaftlichem Interesse verfügbar gemacht werden. Konkret wird der Zugang über eine Nationallizenz in Verwaltung der ZBW (Deutsche Zentralbibliothek für Wirtschaftswissenschaften) organisiert.[2]

Als erster Schritt zum Aktienführer-Datenarchiv waren alle verfügbaren Bände des Aktienführers und seiner Vorgängerpublikationen zu scannen. Da nicht alle Bände an der UB Mannheim vorhanden waren, wurde die Unterstützung durch Leihgabe und Geschenke anderer Bibliotheken benötigt.[3] Die Scanarbeiten wurden intern durch das UB-Digitalisierungsteam bewältigt.

Für das Aktienführer-Datenarchiv mit den Jahren 1870 bis 2018 wurden insgesamt 173 Bände mit zusammen 237.000 Seiten gescannt. Da für die Datenbank nur die Bände von 1956 bis 1999 genutzt wurden, entfallen auf diesen Teil 44 Bände mit etwa 50.000 Seiten.

Für die Jahre 2000 bis 2018 wurden die CD-ROM Ausgaben des Aktienführers verwendet. Hier war es nötig, eine Reihe von speziellen Software-Tools zu schreiben, um die verschlüsselten Daten zugänglich zu machen und strukturiert ins Aktienführer-Datenarchiv zu übertragen.

Händische Datenerfassung in Projektphase I

In der ersten Projektphase (2013–2015) bestand das Ziel darin, die Aktienführer-Bände für die Jahre 1976 bis 1999 zu digitalisieren und in eine Datenbank zu überführen. Wie schon gesagt, entschied man sich an dieser Stelle für eine händische Datenerfassung mittels „Double Keying“, da automatisierte Methoden zu dieser Zeit noch nicht die gewünschte Zeichen- und Strukturqualität erreichen konnten.

Beim „Double Keying“ werden die betreffenden Inhalte mehrfach händisch abgeschrieben, mit Vergleich und Korrektur der Ergebnisse. Die UB Mannheim beauftragte einen indischen Dienstleister mit dieser Aufgabe. Grundlage der Zusammenarbeit war ein detaillierter Erfassungsleitfaden, der für jede Datenkategorie die Regeln festhielt, wie die entsprechenden Daten zu erfassen sind. Schließlich lieferte der Dienstleister die struktu-

[2] URL Nationallizenz des Aktienführer-Datenarchivs: `https://www.nationallizenzen.de/angebote/nlproduct.2014-03-03.9100427542?`

[3] Ein herzlicher Dank geht hier an die USB Köln, die UB Bochum, die Bibliothek der HU Berlin, die UB Heidelberg, die FH-Bibliothek Würzburg, die Lippische Landesbibliothek Detmold, die ZBW, die SUB Göttingen, die Bibliothek des Deutschen Bundestags, die UB Greifswald sowie die StB Koblenz.

rierten Daten im XML-Format an die UB Mannheim. Nach einer ersten Testphase war die gelieferte Datenqualität überwiegend gut, wobei eine ständige Qualitätskontrolle von Seiten der UB notwendig blieb. Im Ergebnis lagen die Daten der Aktienführer-Jahrgänge 1976 bis 1999 in einem strukturierten Format vor.

Ein wichtiger Zwischenschritt an dieser Stelle war die Zusammenführung der Unternehmensprofile der verschiedenen Jahrgänge trotz unterschiedlicher Schreibweisen oder Umbenennungen der Unternehmen, um einfache und konsistente Zeitreihenanalysen zu ermöglichen.

Abschließend wurde aus den Anforderungen der Forscherinnen und Forscher aus dem Bereichen Finance sowie Accounting & Taxation eine sinnvolle Datenbankstruktur und Weboberfläche entwickelt, um eine komfortable Nutzung der Forschungsdaten zu ermöglichen. Im Zentrum stand dabei die Möglichkeit, individuell zusammengestellte Datenpakete in einem Standardformat (hier CSV) zur direkten Weiterverarbeitung herunterladen zu können.

Im Ergebnis der ersten Projektphase standen die Daten des Aktienführers für die Jahre 1976 bis 1999 strukturiert zur Verfügung. Es wurde aber schnell klar, dass diese zeitliche Abdeckung für die Forschung nicht ausreichend ist. Dank weiterer DFG-Förderung konnte dem in einer zweiten Projektphase von 2017 bis 2019 abgeholfen werden.

Automatisierte Datenerfassung in Projektphase II

Auch bei der automatisierten Datenerfassung bildeten die beim Scannen erzeugten Bilddaten den Ausgangspunkt der Arbeit. Im Unterschied zum händischen Vorgehen war jetzt aber eine Reihe von Zwischenschritten zu absolvieren, um die Bilddaten in Text und schließlich in eine strukturierte Datenbank umzuwandeln. Eine Übersicht dazu gibt die folgende Abbildung 3.

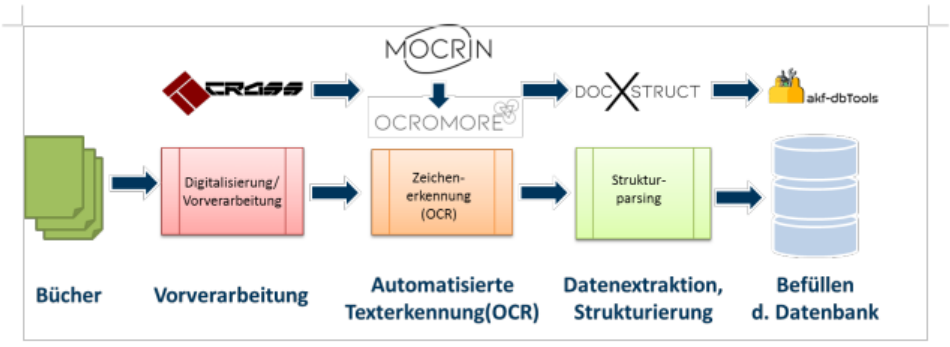


Abbildung 3.: Arbeitsschritte zur automatisierten Datenerfassung (eigene Darstellung)

Die in Abb. 3 dargestellten vier Arbeitsschritte sollen nun im Einzelnen dargestellt werden. Für alle Arbeitsschritte hat das Projektteam der UB Mannheim spezielle Software-Tools geschrieben, die frei verfügbar auf GitHub sind und als Open Source nachgenutzt werden können.[4]

[4] Die Aktienführer-Datenarchiv-Tools auf GitHub: `https://github.com/UB-Mannheim/`Aktienfuehrer-Datenarchiv-Tools

Schritt 1 – Vorverarbeitung

Mittels des Tools „crass“ wurden die gescannten Bilddaten für die folgende Texterkennung vorbereitet. Dazu wurde die 2-Spaltenstruktur der Aktienführer-Unternehmensprofile aufgelöst und in eine leichter zu verarbeitende 1-Spaltenstruktur überführt. Hinzu kommen das Zusammenfassen einzelner Textsegmente und die Korrektur von Verzerrung, die beim Scannen erfolgt sein können.

Schritt 2 – Automatisierte Texterkennung (OCR)

Im zweiten Schritt geht es darum, in den vorliegenden Bilddaten den enthaltenen Text zu erkennen. Zu diesem Zweck wurden im Aktienführer-Projekt drei verschiedene, populäre OCR-Programme parallel genutzt: Tesseract, Ocropus und ABBYY FineReader. Das heißt, jede Bilddatei wurde mehrfach mit OCR bearbeitet mit dem Ziel, die Stärken aller drei OCR-Programme zu kombinieren. Dies bedeutet aber auch, dass die entstandenen drei OCR-Ergebnisse in einem weiteren Schritt wieder zu einem gemeinsamen Ergebnis zusammengefügt werden müssen, welches möglichst die vorhandenen Erkennungsfehler minimiert.

Um diesen Prozess abzubilden, wurden zwei spezielle Tools geschrieben: Mocrin und Ocromore.

Mocrin ermöglicht, die drei OCR-Programme Tesseract, Ocropus und FineReader über ein einziges Interface zu steuern und die Ergebnisse im hOCR-Dateiformat in einer einheitlichen Ordnerstruktur abzulegen.[5] Dabei sind die Erkennungsergebnisse mit Konfidenzen versehen. Diese Konfidenzen geben an, mit welcher Sicherheit ein bestimmtes Zeichen vom OCR-Programm erkannt wurde.

Ocromore übernimmt dann die Zusammenführung zu einem optimierten Endergebnis auf Basis der von den OCR-Programmen gelieferten Konfidenzen. Ein Beispiel für diesen Prozess ist in Abbildung 4 dargestellt.

Darin wird deutlich, wie Ocromore die Ergebnisse der drei OCR-Programme (R1, R2, R3) vergleicht und zusammenführt. Am Beispiel des dritten Buchstabens aus „Eduard“ sieht man, dass R1 ein „u“ mit einer Konfidenz von 99 Prozent erkannt hat, R2 erkennt ein „o“ mit 60 Prozent Konfidenz und R3 wieder ein „u“ mit 90 Prozent Konfidenz. Zusammengefasst gibt das für „u“ einen Konfidenzwert von 189 Punkten und für „o“ nur 60 Punkten, weswegen Ocromore das „u“ an dieser Stelle auswählt. Auf diese Art und Weise gelingt es Ocromore, die Erkennungsergebnisse zu vergleichen und zusammenzuführen.

Dieses Vorgehen ermöglicht eine deutliche Erhöhung der Erkennungsgenauigkeit, wie in Abbildung 5 dargestellt.

Dabei wird für den Aktienführer eine Erkennungsgenauigkeit von 99,6 Prozent erreicht, was einer Fehlerreduktion von ca. 33 % im Vergleich zum besten der einzelnen OCR-Ergebnissen entspricht. Ebenfalls sieht man eine Erhöhung der Erkennungsgenauigkeit für den englischen Standardkorpus UNLV (University of Nevada Las Vegas Standardized Test Set).[6]

[5] Das hOCR-Format ist ein offener Standard zur Darstellung von OCR-Ergebnissen, vgl. Breuel 2007 und `http://kba.cloud/hocr-spec/1.2/`.

[6] Zu Ocromore vgl. auch Kamlah /Stegmüller 2018.

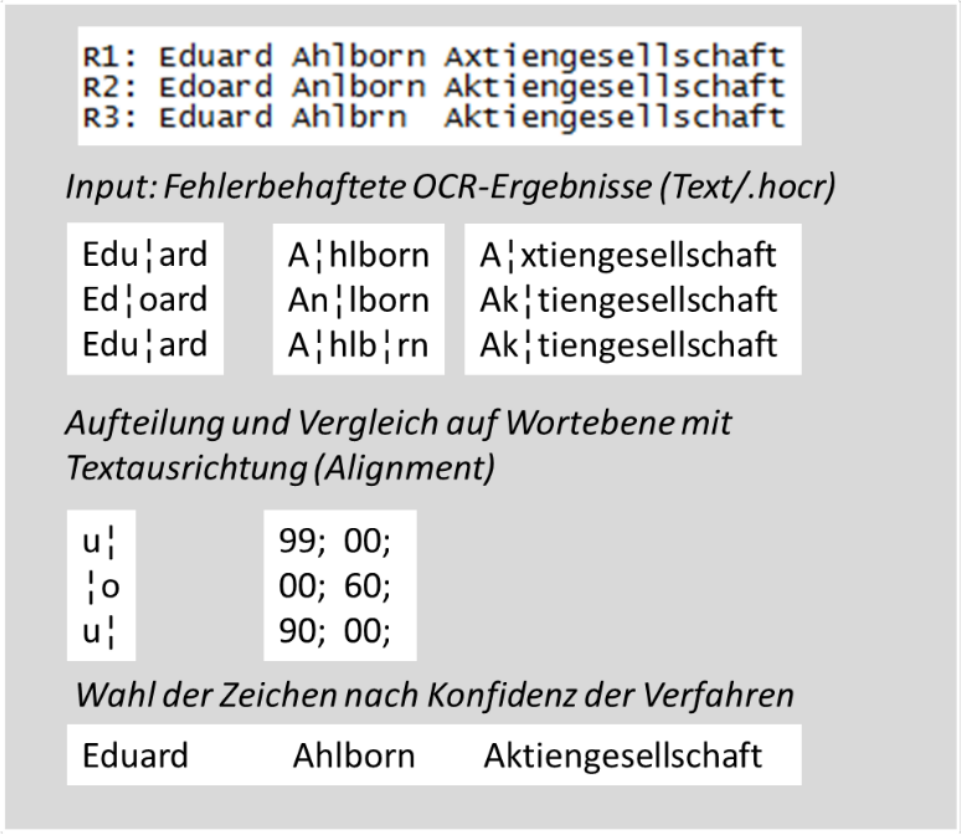


Abbildung 4.: Evaluierung und Zusammenführung der OCR-Ergebnisse durch Ocromore (eigene Darstellung)

Im Ergebnis von Schritt zwei wurde der Text automatisiert erkannt und ist nun im hOCR-Format zur Weiterverarbeitung parat.

Schritt 3 – Datenextraktion und Strukturierung

Mittels des Tools Docxstruct werden die hOCR-Dateien, die das Ergebnis der automatisierten Texterkennung sind, analysiert, kategorisiert und segmentiert und in das JSON-Format überführt. An dieser Stelle wird erkannt, in welche Datenbankbereiche der erkannte Text passt, zu welcher Kategorie die Information also gehört. Ein Beispiel dafür gibt die folgende Abbildung 6.

Man sieht, wie der erkannte Text (origpost) ausgelesen und in einzelne Kategorien (numID, city, street, additional info) aufgelöst wird, die sich dann entsprechend in der Datenbank wiederfinden. Damit gelingt es in Schritt 3, den erkannten Text umfassend zu strukturieren – ihn also den passenden Bereichen der Datenbank zuzuordnen.

Schritt 4 – Befüllen der Datenbank

Nachdem die Daten ausgelesen und strukturiert wurden, können sie in einem letzten Schritt in die in Projektphase I erstellte Datenbank geschrieben werden. Dazu wurde das

OCR-Engine	Aktienführer (AKF)	UNLV
ABBYY	99,35 %	98,46 %
OCRopus (default en-model)	-	92,49 %
OCRopus (trained)	98,76 %	-
Tesseract	99,00 %	98,23 %
ocromore (MSA)	**99,60 %**	**98,65 %**

Abbildung 5.: Erhöhte Erkennungsgenauigkeit durch Ocromore (eigene Darstellung)

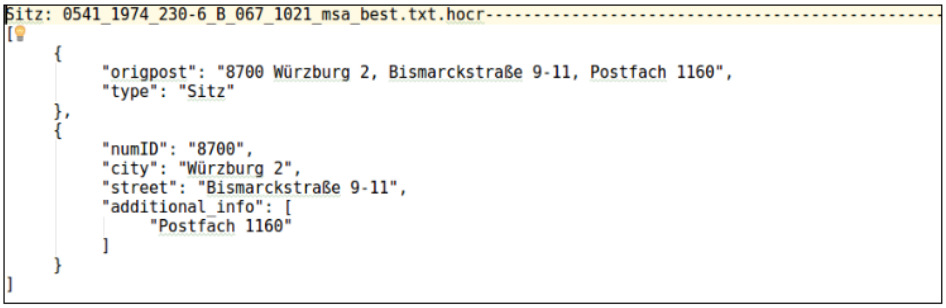


Abbildung 6.: Strukturerkennung mit Docxstruct (eigene Darstellung)

Tool akf-db-Tools geschrieben, eine Sammlung kleinerer Skripte für die Befüllung der Datenbank. Dabei erfolgen auch noch Schritte zur Normalisierung und Deduplizierung der Daten, um die Qualität der Forschungsdaten zu erhöhen.

Im Ergebnis der automatisierten Datenerfassung wurde das Aktienführer-Datenarchiv um 20 Jahrgänge (1956–1975) ergänzt. Zusammen mit den Bänden aus Projektphase I (1976–1999) und den Daten aus den CD-ROMs (2000–2018) ergibt sich eine zusammenhängende Abdeckung über 62 Jahre.[7]

Die Darstellung der automatisierten Datenerfassung zeigt, dass dafür eine Reihe von Schritten und Softwareentwicklungen nötig war. Es wird deutlich, dass für ein solches spezialisiertes Projekt im Moment keine „out-of-the-box“ Lösung verfügbar ist. Nur mit genügend Ressourcen in Form von Zeit und Know-how konnte dieser Weg beschritten werden. Als Gewinn neben den erschlossenen Forschungsdaten erscheinen aber auch die frei nachnutzbaren Software-Tools, die bei zukünftigen, ähnlichen Projekten deutliche Einsparungen versprechen.

Mit Abschluss der zweiten Projektphase steht nun das vollständige Aktienführer-Datenarchiv elektronisch zur Nutzung bereit.[8] Das Aktienführer-Datenarchiv bietet dabei drei Nutzungsmöglichkeiten:

[7] Aus lizenzrechtlichen Gründen können die Jahrgänge 2017 und 2018 allerdings erst 2020 bzw. 2021 zugänglich gemacht werden.

[8] URL des Aktienführer-Datenarchivs: `https://digi.bib.uni-mannheim.de/aktienführer/data/index.php`

- **Schnellsuche** – hier kann die Datenbank (Jahre 1956 – 2018) über die Datenfelder Firmenname, Personen und Wertpapierkennnummer (WKN) durchsucht werden. Es werden dann Treffer für gefundene Unternehmensprofile, Beteiligungen, Personen und WKN angezeigt. Der Zweck der Schnellsuche ist es, einen schnellen Überblick über die verfügbaren Daten zu gewinnen.

- **Datenexport** – hier können große Datenpakete für empirische Analysen zusammengestellt und heruntergeladen werden. Der Nutzer kann dabei die benötigten Unternehmen, Jahre und Datenkategorien wählen und das Ergebnis im CSV-Format herunterladen. Das CSV-Format ist ein Standard im Bereich der quantitativen Datenanalyse und kann direkt in Auswertungsprogramme wie Stata, R oder SPSS aber auch Microsoft Excel oder LibreOffice Calc geladen werden.

- **Scans** – hier stehen die Volltext-Digitalisate der Print-Bände des Hoppenstedt-Aktienführers und seiner Vorgängerpublikationen (Jahre 1870 – 1999) zur Nutzung bereit. Alle Bände sind mit Inhaltsverzeichnissen versehen und können im Volltext durchsucht und heruntergeladen werden. Somit ist eine komfortable und effiziente Nutzung möglich. Falls bei der Nutzung der Datenbank Zweifel an den präsentierten Daten bestehen, können alle Angaben mithilfe der Scans verifiziert werden.

In Abbildung 7 sieht man am Beispiel des Unternehmens Daimler-Benz die Oberfläche zum Export der Aktienkurse aus den Jahren 1970 bis 1980.

Nach dem Export stehen die Daten des Aktienführer-Datenarchivs unmittelbar zur Auswertung bereit. Ein Nutzungsbeispiel könnte die Frage sein, wie sich die Ölkrise der 70er Jahre auf die Firma Daimler-Benz AG ausgewirkt hat. Abbildung 8 präsentiert eine Darstellung der aus dem Aktienführer gewonnenen Daten, die zeigt, dass zwar der Aktienkurs zeitweilig schwankte, Umsatz und Jahresüberschuss aber keine gravierenden negativen Entwicklungen aufwiesen. Größere Datensets dieser Art können von der empirischen Forschung zum Testen von Hypothesen und Theorien verwendet werden.

Das Aktienführer-Datenarchiv hat durch die Ergänzung der Daten der zweiten Projektphase erheblich an Wert für die Forschung gewonnen. Schon bisher zeigten sehr gute Nutzungsdaten die Bedeutung des Angebotes an. So haben sich bisher 142 Einrichtungen für die Nutzung über die Nationallizenz angemeldet. Außerdem gibt es auch noch mehr als 380 persönliche Accounts für Personen, die ein berechtigtes Interesse nachweisen konnten und ohne institutionelle Zugriffsmöglichkeit sind. Die Webseite erhielt im Jahr 2018 etwa 12.500 Besuche. Es ist zu erwarten, dass das deutlich erweiterte Datenangebot zu noch mehr Nutzung führen wird.

Im Endergebnis stehen nun mit dem Aktienführer-Datenarchiv hochrelevante Forschungsdaten für die Wirtschaftswissenschaften bereit. Eine bisher nur gedruckt vorliegende Publikation wurde durch die Anstrengungen der UB Mannheim in Forschungsdaten transformiert, die nach den heutigen Standards der Forschung genutzt werden können. Es steht zu erwarten, dass hieraus viele interessante Publikationen erwachsen.

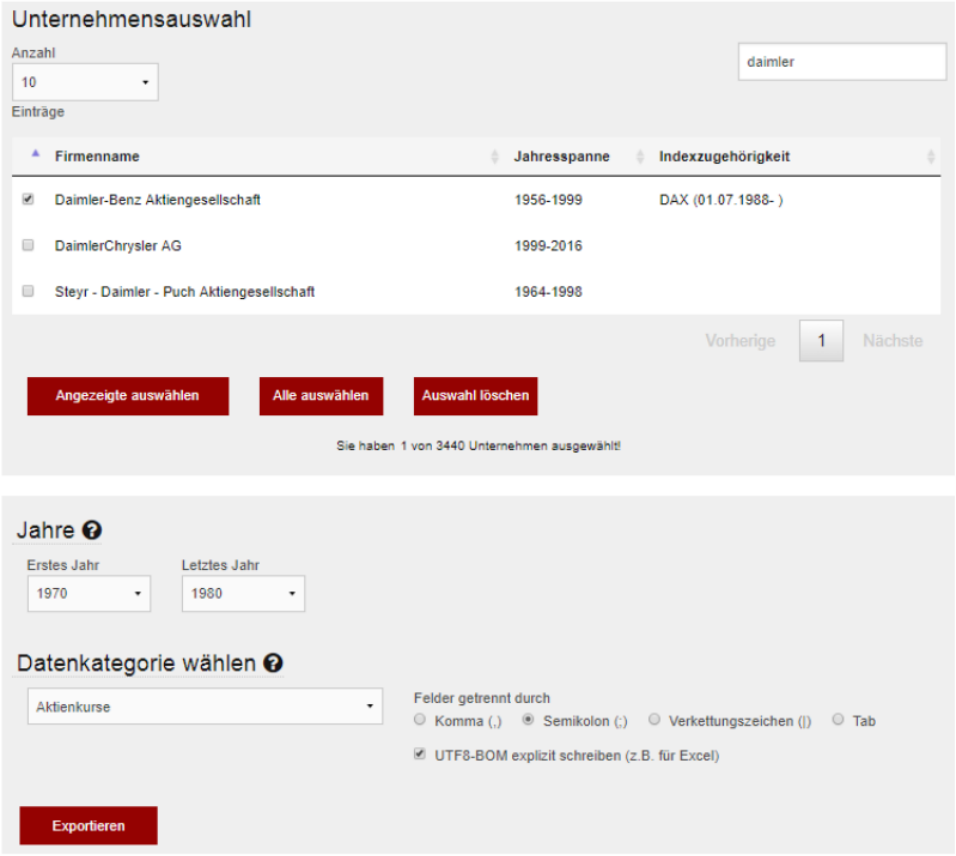


Abbildung 7.: Datenbank-Export für Daimler-Benz (eigene Darstellung)

4. Ausblick – Forschungsdaten an der UB Mannheim

Mit dem Aktienführer-Datenarchiv hat die UB Mannheim einen ersten, richtungsweisenden Schritt im Bereich der wirtschaftswissenschaftlichen Forschungsdaten getan, dem noch weitere folgen werden.

Aufbauend auf dem Aktienführer-Datenarchiv sollen weitere wirtschaftswissenschaftliche Forschungsdaten aus dem Hause Hoppenstedt / Bisnode erschlossen und bereitgestellt werden. Die UB Mannheim hat von dem Anbieter die Digitalisierungsrechte an einer Reihe weiterer fachlich relevanter Publikationen erworben.[9] Auch diese sollen nach dem Vorbild des Aktienführer-Datenarchivs digitalisiert und erschlossen werden. Man kann erwarten, dass hier ein Portfolio hochrelevanter Forschungsdaten zur Entwicklung der deutschen Wirtschaft entsteht.

Diese Forschungsdaten sind der Grundbaustein für das Forschungsdatenzentrum (FDZ) an der Universität Mannheim. Im Rahmen des FDZ werden auch andere Forschungsdaten angeboten, wie zum Beispiel der „Deutsche Reichsanzeiger und Preußische Staatsanzeiger“.

[9] Zu diesen Publikationen gehören: „Leitende Männer und Frauen der Wirtschaft“, „Großunternehmen“, „Mittelständische Unternehmen“, „Handbuch der deutschen Aktiengesellschaften : das Spezial-Archiv der Deutschen Wirtschaft“.

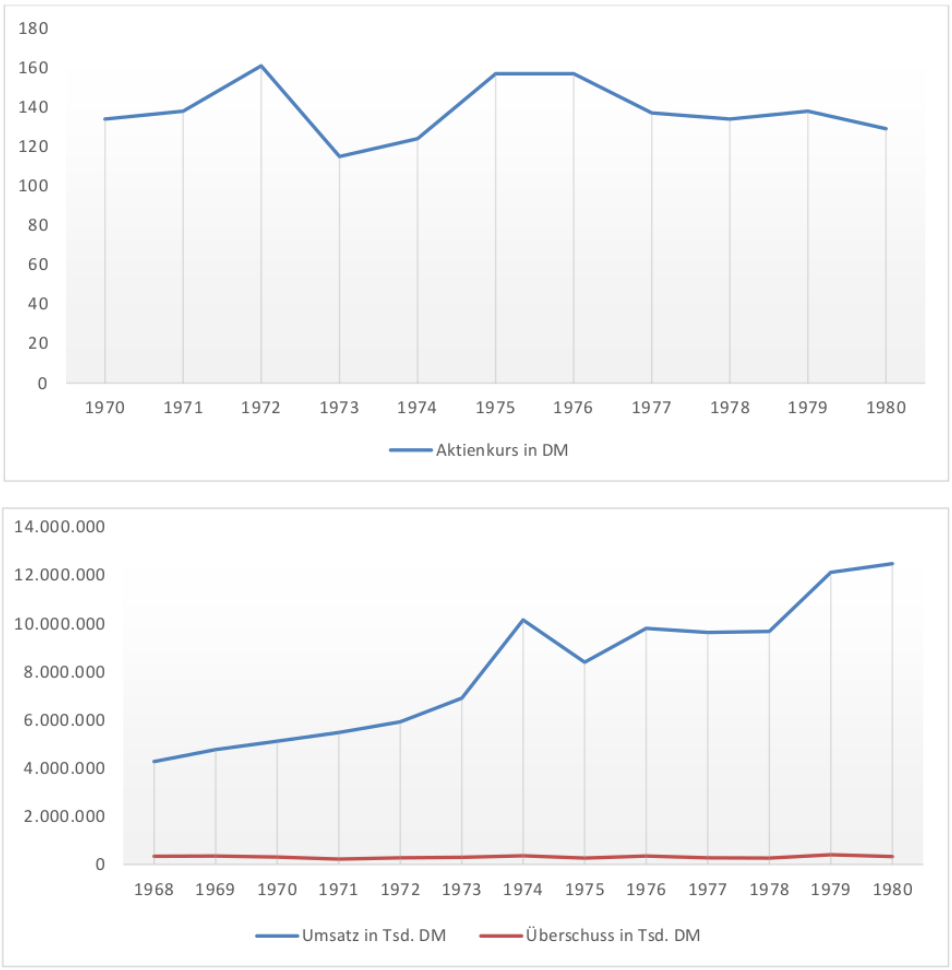


Abbildung 8.: Aktienkurs, Umsatz und Jahresüberschuss der Daimler-Benz AG in den 70er Jahren (eigene Darstellung)

Um die Kompetenz für OCR-Erschließung von vorhandenen Bilddigitalisaten noch breiter an Bibliotheken und anderen Kultureinrichtungen des Landes zu verankern, wird die UB Mannheim gemeinsam mit der Universitätsbibliothek Tübingen zudem in den kommenden beiden Jahren Tools und Beratungsservices bereitstellen, die den Einsatz automatisierter Erschließungsverfahren auch ohne einschlägiges Expertenwissen ermöglichen sollen. Das Projekt OCR-BW schließt an das nationale, DFG-geförderte Projekt OCR-D an, bei dem die UB Mannheim ebenfalls beteiligt ist, und wird durch das Ministerium für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg gefördert.

Ab Mitte 2019 wird die Universitätsbibliothek Mannheim gemeinsam mit Wissenschaftlerinnen und Wissenschaftlern der Universität Mannheim und des Leibniz-Zentrums für Europäische Wirtschaftsforschung (ZEW) ein Kompetenzzentrum für Datenverfügbarkeit und Datenanalyse in den Wirtschaftswissenschaften für Baden-Württemberg aufbauen, das als eines von vier großen Science Data Centers durch das Ministerium für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg gefördert wird. Das Business and Economic Research Data Center (BERD-Center BW) wird den Zugang zu vor-

handenen wirtschaftswissenschaftlichen Datenbeständen verbessern und neue, unstrukturierte Datenquellen (Big Data) erschließen. Damit kann auch ein wichtiger Beitrag für eine künftige Nationale Forschungsdateninfrastruktur (NFDI) und eine European Open Science Cloud geleistet werden, deren Ziel eine umfassende, verlässliche, gut zugängliche und möglichst offene, vernetzte Infrastruktur für Forschungsdaten ist.

Literaturverzeichnis

[1] Breuel, Thomas M. (2007): „The hOCR Microformat for OCR Workflow and Results“. In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2:1063–67. Online:https://doi.org/10.1109/ICDAR.2007.4377078 [02.05.2019].

[2] Kamlah, Jan und Stegmüller, Johannes (2018): Ocromore – Combining multiple OCR-engine results to improve character recognition accuracy. Poster Bibliotheca Baltica Symposium 2018 (BB2018), Rostock, 3.-5.October 2018. Online: http://doi.org/10.5281/zenodo.1493860 [02.05.2019].

[3] item Zumstein, Philipp (2016): Die Bibliothek als Daten-Jongleur. Vortrag beim Bibcast 2016. Online:
http://bibcast.openbiblio.eu/die-bibliothek-als-daten-jongleur-services-fuer-datenzentrierte-forschung/ [02.05.2019].

Open Access für die Mediävistik: das Archivum Medii Aevi Digitale

Aglaia Bianchi und Paul Warner
DFG-Projekt AMAD -Archivum Medii Aevi Digitale, Deutschland

Abstract

Die im DFG-Projekt AMAD angestrebte Koppelung eines Open Access-Fachrepositoriums mit einem Wissenschaftsblog stellt in den Geisteswissenschaften und insbesondere in der Mediävistik ein Novum dar. In den Geisteswissenschaften, deren Publikationskultur noch stärker als in den Naturwissenschaften am gedruckten Buch orientiert ist, sind Fachrepositorien und Open Access-Publikationsplattformen trotz zunehmender Förderung noch nicht in dem Maße etabliert. Kostenpflichtige Angebote der Verlage stellen noch oft die bevorzugte Alternative dar. Die Fachcommunity schätzt dabei das hohe symbolische Kapital der Wissenschaftsverlage, die Qualitätssicherung und die Zuverlässigkeit der Langzeitarchivierung sowie die Sichtbarkeit, die ihre Forschung dadurch erreicht. Bei universalen Repositorien an Hochschulen und Bibliotheken sind Forschungsergebnisse oft frei und schneller zugänglich, die Auffindbarkeit ist jedoch aufgrund der fehlenden Fächerausrichtung und nur wenig spezifischer Suchmöglichkeiten schwierig und mühsam. Zudem fehlt dort meist die Qualitätssicherung, da eine fachredaktionelle Betreuung von institutionellen Repositorien kaum zu leisten ist. Das Archivum Medii Aevi Digitale (AMAD) will mit seinem Angebot und der engen Zusammenarbeit zwischen Vertreter*innen der Fachwissenschaft und von technischen Infrastruktureinrichtungen diese Lücke schließen. Für die Umsetzung der Projektidee haben sich deshalb die Ludwig-Maximilians-Universität München (LMU) und das Akademievorhaben Regesta Imperii (RI) mit der Infrastruktureinrichtung Hessisches BibliotheksInformationsSystem (HeBIS) zusammengeschlossen. Das technische Konzept basiert auf einem DSpace-Repositorium und nutzt die Möglichkeit, über Standardschnittstellen Harvestingprozesse zu konfigurieren. Zu den üblichen Funktionalitäten (Archivierung, Persistent Identifier, Versionierung) werden weitere Instrumente entwickelt, die zur Qualitätssicherung (Redaktion, Kuration, Peer Review) und damit einer breiten Akzeptanz der neuen Plattform beitragen sollen. Von der wissenschaftlichen Seite sollen zum einen die Kriterien (Datenquellen, Vokabular, Zugangsmöglichkeiten etc.) für den Aufbau einer gesicherten und interdisziplinär nutzbaren Datenbasis entwickelt werden, zum anderen die wissenschaftliche Fachcommunity (sowohl die digitale als auch die analoge) erreicht und zu einer aktiven Beteiligung im wissenschaftlichen Austausch auf Repositorium und Blog motiviert werden. Der vorliegende Beitrag informiert über die unterschiedlichen Herausforderungen und die Zusammenarbeit von Technik und Wissenschaft.

Schlagwörter: Mediävistik, Repositorium, Open Access, Blog

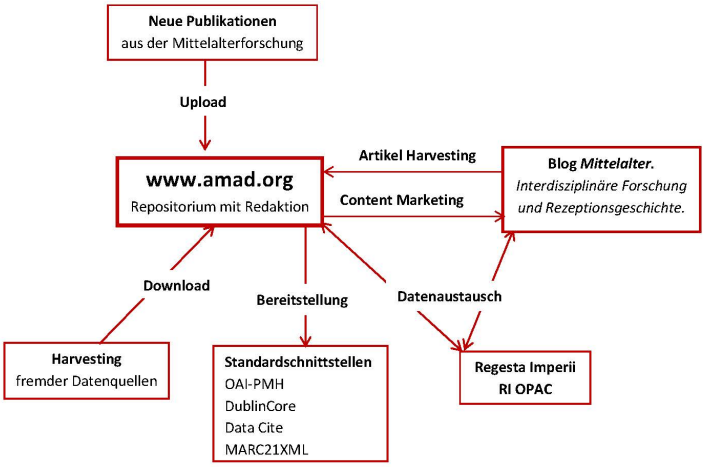


Abbildung 1.: Zusammenspiel der technischen Komponenten im AMAD-Projekt.

1. Das Projekt

Das Projekt „Archivum Medii Aevi Digitale – Mediävistisches Fachrepositorium und Wissenschaftsblog“[1], das seit dem 1. Oktober 2018 von der Deutschen Forschungsgemeinschaft im Förderprogramm „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ (LIS) gefördert wird, stammt aus einer Kooperation zwischen dem Wissenschaftsblog „Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte“[2], das seit 2012 auf dem Blogportal für Geistes- und Sozialwissenschaften hypotheses.org[3], betrieben wird, dem Akademieprojekt Regesta Imperii[4], dem Lehrstuhl für Mittelalterliche Geschichte mit Schwerpunkt Spätmittelalter an der Ludwig-Maximilians-Universität München (Prof. Dr. Claudia Märtl) und dem Hessischen BibliotheksInformationsSystem (HeBIS).

Ziel des Projektes ist die Einrichtung eines Open-Access-Fachrepositoriums, das zusammen mit dem bereits etablierten Wissenschaftsblog „Mittelalter“ die Funktion eines Publikationsorts für Erst- und Zweitveröffentlichungen zur interdisziplinären Mittelalterforschung erfüllt. Es eröffnet dadurch neue Wege der Publikation und des wissenschaftlichen Austausches in der Mediävistik, und zwar in ihrem gesamten Disziplinenspektrum.

[1] `www.amad.org` [zuletzt Zugriff am: 23.08.2019].
[2] `https://mittelalter.hypotheses.org/`[zuletzt Zugriff am: 23.08.2019].
[3] `https://de.hypotheses.org/`[zuletzt Zugriff am: 23.08.2019] .
[4] `www.regesta-imperii.de`[zuletzt Zugriff am: 23.08.2019].

Dieses Ziel ergibt sich aus einer genauen Betrachtung der aktuellen Situation in der Publikationslandschaft der Mediävistik.

Mediävistischen Forschenden, die ihre Werke veröffentlichen wollen, stehen derzeit meist nur zwei Möglichkeiten zur Verfügung: die Veröffentlichung in einem Wissenschaftsverlag oder die Publikation in einem institutionellen Repositorium. Etablierte Wissenschaftsverlage genießen ein hohes Vertrauen der Fachcommunity und bleiben in den meisten Fällen die bevorzugte Variante bei der Publikation. Einer der Hauptgründe hierfür mag wohl darin liegen, dass das Fach selbst immer noch sehr am gedruckten Medium orientiert ist[5]. Gerade für Monografien wird die Verlagspublikation in Buchform als die bessere Option gesehen. Aber auch bei elektronischen Publikationen, insbesondere bei Zeitschriftenartikeln, werden die Angebote der Verlage bevorzugt. Bei der Veröffentlichung im Verlag — sei sie in gedruckter oder elektronischer Form — schätzen die Mittelalterforschenden das Renommee, das die Fachverlage in der Community genießen, und die Verlässlichkeit der Publikation in einem etablierten Verlag. In diesem Fall spielen auch Vorurteile und Befürchtungen, etwa, dass eine Onlinepublikation per se weniger stabil sei, eine Rolle. Darüber hinaus stellt die vom Verlag angebotene Qualitätssicherung ebenfalls einen wichtigen Entscheidungsfaktor für die Autor*innen dar, zum einen weil sie garantiert, dass das Werk in korrekter Form veröffentlicht wird, zum anderen, weil sie auch bei der Rezeption und Akzeptanz des Textes in der Fachcommunity eine entscheidende Rolle spielt. Forschende, die sich für eine Verlagspublikation entscheiden, müssen allerdings auch mit hohen Kosten rechnen. Dabei ist die Publikation in einem Wissenschaftsverlag mittlerweile nicht nur mit Druckkosten verbunden, die oft zumindest teilweise von den Autor*innen zu tragen sind; auch der Zugang für Leser*innen ist oft nicht frei, sondern muss entweder von einer Institution mittels (Zeitschriften-) Abonnement oder von den Nutzer*innen selbst bezahlt werden. Darüber hinaus müssen Forschende oft mit längeren Wartezeiten bis zur Publikation ihrer Ergebnisse rechnen sowie die Rechte am Text weitgehend an den Verlag abtreten[6]. Dieser übernimmt dann unter Umständen sogar nur noch den rein technischen Druckprozess; Korrektorat und Lektorat müssen immer häufiger von den Autor*innen oder Herausgeber*innen selbst unentgeltlich geleistet oder durch auswärtige Beauftragung zusätzlich finanziert werden.

Eine Alternative zur Verlagsveröffentlichung stellen aktuell die digitalen universalen Bibliotheks-und Hochschulrepositorien dar, in denen Publikationen zentral gespeichert und dauerhaft verfügbar gemacht werden. In solchen Repositorien ist sowohl die Publikation für die Autor*innen als auch der Zugang für die Leser*innen häufig kostenfrei und die Rechte sind über gängige Open Access und Open Content-Lizenzen geregelt. Aufgrund des Universalcharakters dieser Repositorien ist allerdings eine Qualitätssicherung, vor allem eine fachliche Betreuung und Redaktion, kaum möglich. Auch die Sichtbarkeit der Forschung leidet unter diesen Umständen, denn es fehlen für die Auffindbarkeit der Publikationen ein gezieltes Content Marketing sowie Vernetzungen mit wichtigen Schnittstellen

[5] Das gilt generell für die Geisteswissenschaften, vgl. Konstanze Söllner: Fachspezifische Perspektive: Geisteswissenschaften. In: Konstanze Söllner, Bernhard Mittermaier (Hrsg.): Praxishandbuch Open Access. Berlin: De Gruyter Saur 2017, S. 247-253, hier S.248f.

[6] Zu den Unterschieden zwischen Verlagen und Repositorien in diesem Bereich vgl. auch Konstanze Söllner, Warum und für wen Open Access?, in: Konstanze Söllner, Bernhard Mittermaier (Hrsg.): Praxishandbuch Open Access. Berlin: De Gruyter Saur 2017, S. 3-11, hier S.4f.

zu Datenbanken, Bibliographien und Katalogen, die nur im fachspezifischen Kontext zu leisten wären. Dies dürfte auch ein Grund dafür sein, dass Repositorien insgesamt weniger etabliert sind und im Vergleich ein geringeres Renommee in der mediävistischen Fachcommunity genießen. In den Naturwissenschaften sind Fachrepositorien schon seit einiger Zeit etabliert und fester Bestandteil der Publikationslandschaft;[7] in den Geisteswissenschaften sind erst in letzter Zeit einige wenige fachspezifische Repositorien entstanden,[8] allerdings fehlt noch ein entsprechendes Angebot in der Mediävistik.

Die Projektinitiator*innen von AMAD haben sich angesichts dieser Situation in der mediävistischen Publikationslandschaft gefragt: „Wie kann man die Vorteile der Verlagspublikation mit den Vorteilen der Open Access-Publikation in Repositorien kombinieren?“

Die Gründer*innen des Blogs „Mittelalter“, Martin Bauch, Karoline Döring und Björn Gebert, haben hierfür führende wissenschaftliche und technische Institutionen, die ihre jeweilige Expertise einbringen, für das Projekt zusammengebracht: die Ludwig Maximilians-Universität München mit der Fachwissenschaftlerin Claudia Märtl und ihrem Lehrstuhl für Mittelalterliche Geschichte, Schwerpunkt Spätmittelalter; die Regesta Imperii, Langzeitvorhaben der Akademie der Wissenschaften und der Literatur | Mainz und Betreiber des RI-OPAC[9], der zentralen internationalen bibliographische Datenbank für die Mediävistik (derzeit 2 Mio. Titel, über 850.000 erfasste Zugriffe bei Steigerungsraten von über 10 % p. A.) und die Verbundzentrale des Hessischen Bibliotheksinformationssystems (HeBIS) als Dienstleistungszentrum der wissenschaftlichen Bibliotheken in Hessen und Teilen von Rheinland-Pfalz.

Aus dieser Kooperation entsteht das Archivum Medii Aevi Digitale, dessen besonderes Kennzeichen die Kombination eines neuen Fachrepositoriums als strukturierendem Speicher mit dem bereits etablierten Wissenschaftsblog „Mittelalter“ als verbreitendem Medium ist.

Das Fachrepositorium erfüllt als Publikationsplattform für Veröffentlichungen gleichzeitig die Grundfunktion der Langzeitarchivierung. Mit einer technischen Lösung für Peer Reviewing-Prozesse, der Vergabe von Persistent Identifiern und der Bereitstellung von Standardschnittstellen werden die Zitier- und Recherchierbarkeit der Publikationen gewährleistet. Ein Beispiel für die Vernetzungsvorhaben ist der angestrebte Datentransfer zwischen dem Repositorium und dem OPAC der Regesta Imperii mittels der Standard-Schnittstellen MARC21 bzw. Dublin Core. Die dadurch erreichte Sichtbarkeit der Publikationen wird durch das Blog gestärkt und ergänzt. Hier finden die Kommentierung und die Verbreitung der Inhalte sowie die Wissenschaftskommunikation und der wissenschaftliche Austausch statt. Dadurch erfolgen schließlich die Vernetzung und die Rückkopplung in der wissenschaftlichen Community. Dabei werden Repositorium und Blog keineswegs

[7] Vgl. den sehr bekannten Preprint-Server arXiv: `https://arxiv.org/`[zuletzt Zugriff am: 23.08.2019].

[8] Z.B. das Fachrepositorium GenderOpen zur Geschlechtergeschichte (`https://www.genderopen.de/`) oder das The Stacks, das Fachrepositorium für Amerikastudien, Anglistik, Anglophone Literaturen und Kulturen, Australien- & Neuseelandstudien, Großbritannien- & Irlandstudien und Kanadastudien,`https://thestacks.libaac.de/`) [zuletzt Zugriff am: 23.08.2019]. Zum Unterschied zwischen institutionellen und Fachrepositorien vgl. Björn Gebert: Wissenschaftsblogs und Fachrepositorien. Wege zu Open Access in der Archäologie, in: Mitteilungen des Deutschen Archäologen-Verbandes 49,2 (2018), S. 46–50, hier S. 48f.

[9] `http://opac.regesta-imperii.de/` [zuletzt Zugriff am: 23.08.2019].

als parallel laufende und unabhängig voneinander funktionierende Angebote gesehen. Gerade in der Integration und Zusammenarbeit der beiden Komponenten entfaltet sich das innovative Potential des Angebots von AMAD, vor allem im Hinblick auf die aktive Beteiligung der Community am wissenschaftlichen Diskurs (s. Abbildung 2).

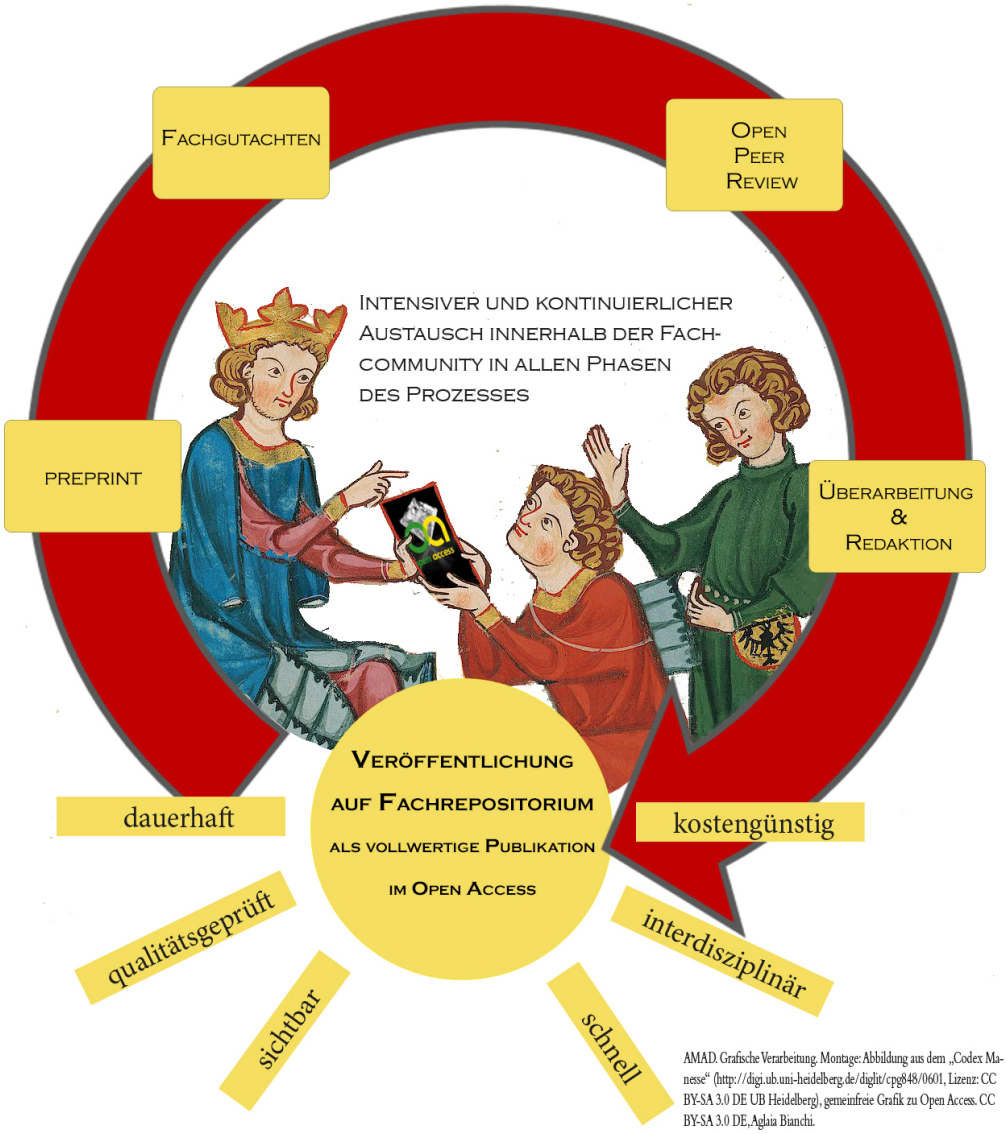


Abbildung 2.: Der Weg des Manuskriptes

AMAD orientiert sich in der Konzipierung und Ausgestaltung des Angebots an der Fachcommunity und ihren Bedürfnissen. Im Folgenden werden deshalb zum einen diese Bedürfnisse, zum anderen die angestrebten technischen Lösungen vorgestellt, die das Projekt Archivum Medii Aevi Digitale dafür vorsieht.

2. Langzeitarchivierung

Die schon erwähnte Langzeitarchivierung, also eine langfristige, stabile und zuverlässige Archivierung der eigenen Publikation und ihre Bereitstellung für die Fachcommunity, spielt für Autor*innen eine zentrale Rolle. Viele in der letzten Zeit entstandene Fachrepositorien sind das Ergebnis einer befristeten Projektförderung. Auch die AMAD-Unterstützung durch die DFG ist zunächst begrenzt.[10] Aus diesem Grund wurde von Beginn an die Kooperation mit einer Partnerinstitution gesucht, deren Fortbestehen gesichert ist und die somit den technischen Support über die Projektlaufzeit hinaus gewährleisten kann. HeBIS verfügt bereits über Erfahrungen mit der Archivierung von Digitalisaten und hat deshalb diese Rolle auch für das AMAD-Repositorium übernommen.[11]

Das DSpace[12]-Repositorium bietet somit einen langfristigen Online-Zugang, mit vielfältigen Rechercheoptionen.

3. Sichtbarkeit und Harvesting

Es reicht allerdings nicht, dass die eigene Publikation langfristig und sicher archiviert ist. Sie soll auch von der wissenschaftlichen Fachcommunity wahrgenommen werden. Ihre Sichtbarkeit ist also fundamental, damit sie eine möglichst breite Interessiertengruppe erreicht.

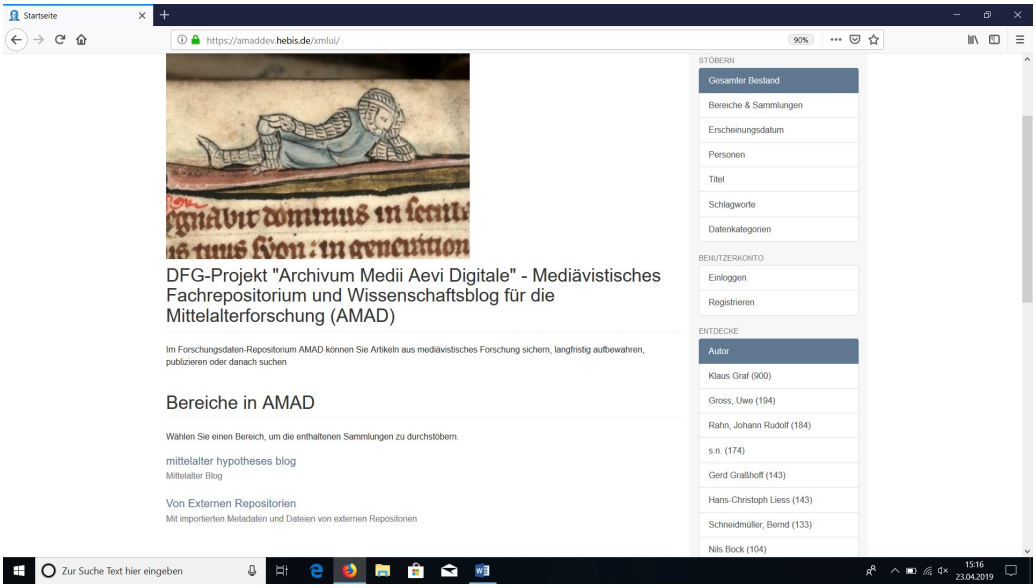


Abbildung 3.: Einblick in die aktuelle Arbeitsoberfläche: Suchmöglichkeiten nach Bereichen und Kategorien

[10] Die Projektlaufzeit beträgt 3 Jahre, von Oktober 2018 bis September 2021.

[11] gl. Helmut Eckardt, Der Erste Weltkrieg im Spiegel hessischer Regionalzeitungen : Ein kooperatives Digitalisierungsprojekt. In: HeBIScocktail. Sonderausgabe Bibliothekartag 2017, S.8. `https://www.hebis.de/de/1cock-tail/pdf/Sonderausgabe2017.pdf` [zuletzt Zugriff am: 23.08.2019].

[12] `https://duraspace.org/dspace/`[zuletzt Zugriff am: 27.08.2019].

Aktuell können Online-Publikationen zur Mediävistik nur mit einigem Suchaufwand in den verschiedenen institutionellen, auch internationalen Repositorien gefunden werden, die dazu einzeln angesteuert werden müssen. Ziel des Projekts ist, ein möglichst umfassendes Angebot an Publikationen vorzuhalten, das gemeinsam leicht durchsuchbar ist und in dem sich nachher auch die neu veröffentlichten Publikationen im Kontext der bereits existierenden, relevanten Forschungslandschaft auffinden lassen.

Aus diesem Grund startete das Projekt mit dem Baustein „Kontrolliertes Harvesting aus Fremddatenquellen". Hierzu werden von den Fachwissenschaftler*innen in der Projektgruppe Merkmale und Kennzeichen geliefert, mit deren Hilfe Filterprozesse für das automatische Harvesten aus anderen Repositorien programmiert werden können. Im ersten Schritt werden dazu die Daten von BASE (Bielefeld Academic Search Engine)[13]abgefragt. Weitere, dort nicht enthaltene digitale Ressourcen, z. B. aus OpenDOAR[14] oder DART-Europe (e-Dissertationen)[15] werden folgen. Es wäre zu einfach, dieses Aussortieren mit der Suche nach Schlagworten, wie z.B. „Mittelalter" gleichzusetzen. Die Metadatenbeschreibung enthält in vielen Fällen keine solch eindeutigen Kennzeichen. Deshalb sind weitere Kriterien erforderlich, die eine saubere Zuordnung zur interdisziplinären Mittelalterforschung ermöglichen. Dort, wo automatische Prozesse dies nicht zuverlässig gewährleisten, erfolgt eine händische Kuratierung der Inhalte durch die Projektmitarbeiter*innen mit manuellen Korrekturen dieser Titelauswahl. Sind die Filterkriterien einmal parametrisiert, kann und soll das Harvesting als regelmäßige Routine installiert werden.

Ein Mittel um die Sichtbarkeit und den Bekanntheitsgrad von AMAD zu erhöhen wird die Verknüpfung mit dem Wissenschaftsblog sein. Alle dort veröffentlichten wissenschaftlichen Artikel werden gleichzeitig im Repositorium abgelegt und sind zusammen mit den hochgeladenen Online-Publikationen suchbar. Über die engen Grenzen der Fachwissenschaft hinaus ist der Direktzugriff auf die Volltexte durch die Vergabe eines zitierfähigen Persistent Identifiers (DOI) und durch Schnittstellen zu einschlägigen Bibliographien wie dem RI-OPAC möglich. Da es sich um ein DSpace-Repositorium handelt, ist damit auch die Durchsuchbarkeit von Google oder Google Scholar relativ leicht konfigurierbar. So erhalten Wissenschaft wie auch die breitere Öffentlichkeit damit gleichermaßen die Möglichkeit auf Fachpublikationen zuzugreifen.

4. Qualitätssicherung

Ein weiteres Bedürfnis der wissenschaftlichen Community ist die Qualitätssicherung. Sowohl die Autor*innen als auch die Leser*innen wünschen sich Professionalität bei der Publikation und möchten an einem Ort veröffentlichen, wo nur qualitativ hochwertige Texte veröffentlicht werden. Dies soll bei AMAD durch unterschiedliche Prozesse und Instrumente gewährleistet werden. Das Einreichen von Texten durchläuft einen Review-Prozess, der vom AMAD-Redaktionsteam gestaltet und moderiert wird. Er soll möglichst offen angelegt sein, ausgewiesene Expert*innen als Gutachter*innen beteiligen und tech-

[13] `https://www.base-search.net/about/de/`
[14] `http://v2.sherpa.ac.uk/opendoar/` [zuletzt Zugriff am: 23.08.2019].
[15] `http://www.dart-europe.eu/basic-search.php` [zuletzt Zugriff am: 23.08.2019].

nisch durch entsprechende Tools unterstützt werden.[16] Auch beim oben beschriebenen Harvestingprozess spielt die Qualitätssicherung eine wichtige Rolle. Ohne den geübten Blick der fachwissenschaftlichen Projektmitglieder auf die aus institutionellen Portalen automatisch gesammelten Metadaten bliebe die Auswahl der Publikationen unzuverlässig und damit der Wert einer Treffermenge im Repositorium gemindert. Um diese Arbeit der Redaktion zu erleichtern und effektive Workflows zu ermöglichen, gibt es ein Administrationstool, mit dessen Hilfe die Filterauswahl korrigiert oder Dubletten aussortiert werden können.

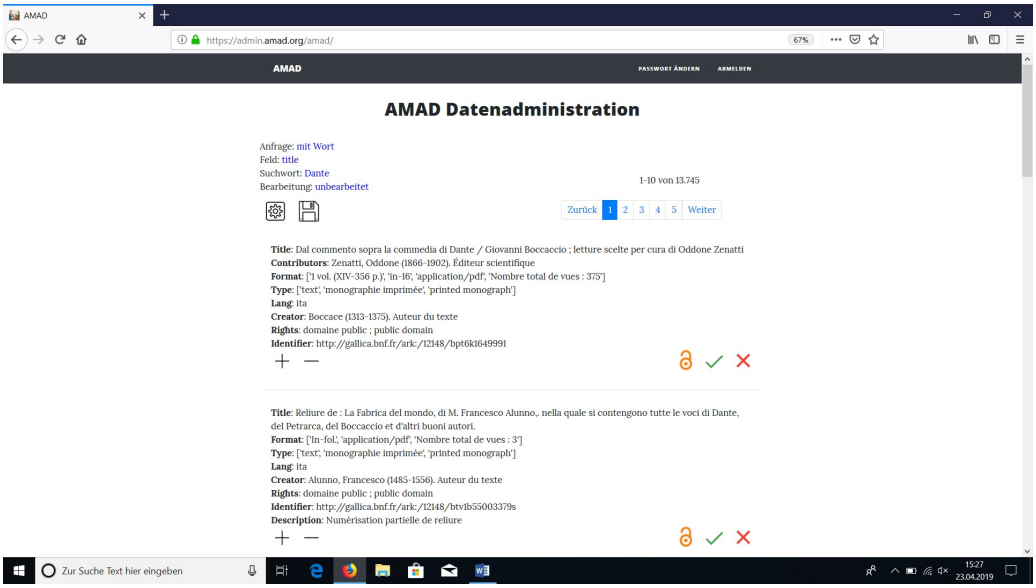


Abbildung 4.: Das AMAD-Datenadministrationstool

5. Kosten

Schließlich spielen auch die Kosten der Publikation vor allem beim wissenschaftlichen Nachwuchs eine Rolle, aber auch allgemein für die Dynamik des wissenschaftlichen Diskurses. Im Sinne des Open Access soll der Zugang zu den Texten für Forschende frei sein. Es sollen auch möglichst keine Publikationskosten für die Autor*innen entstehen. Wie ist dies nun möglich?

Zunächst entfallen natürlich die mitunter hohen Druckkostenzuschüsse bei einer Veröffentlichung in diesem Online-Fachrepositorium. Da es sich außerdem um ein aus öffentlicher Hand gefördertes und betreutes Angebot ohne kommerzielle Absichten handelt,

[16] Um hier ein konkretes Verfahren zu entwickeln, das den Bedürfnissen der Fachwissenschaft entspricht, wird am 9. und 10. September 2019 im München ein Projektworkshop zum Thema „Alles open? Offene Begutachtungsverfahren für offene Publikations- und Informationsinfrastrukturen“ mit Expert*innen mit verschiedenen, nicht nur geisteswissenschaftlichen Hintergründen, die in der Forschung und in Infrastruktureinrichtungen tätig sind, stattfinden. Zum Peer-Review-Verfahren bei Open Access vgl. auch Uwe Thomas Müller: Peer-Review-Verfahren zur Qualitätssicherung von Open-Access-Zeitschriften: systematische Klassifikation und empirische Untersuchung [Dissertation, Humboldt-Universität zu Berlin] 2009.DOI:10.18452/15885 [Zuletzt Zugriff am 27.8.2019]

werden auch keine Article Processing Charges (APC) für Autor*innen erhoben. Für Leser*innen sind die die Inhalte unter kostenfreien Lizenzen verfügbar. Redaktionelle Prozesse übernimmt die Redaktion des Wissenschaftsblogs „Mittelalter“. Die Gutachter*innen beurteilen die Publikationen kostenfrei, was im Übrigen auch bei Wissenschaftsverlagen der Fall ist, die das Begutachten von Zeitschriftenartikeln oder die Qualitätsprüfung von Inhalten für Sammelbände und Monografien innerhalb ihrer wissenschaftlichen Buchreihen in der Regel ebenfalls an unentgeltlich arbeitende Gutachter*innen und Herausgeber*innen aus der Fachwissenschaft abgeben.

Neben der technischen Betreuung, die durch den Projektpartner HeBIS langfristig gewährleistet ist, ist also eine wesentliche Voraussetzung für das Funktionieren dieses Modells das ehrenamtliche Engagement aus den Reihen der Fachwissenschaft selbst. Ohne dieses wären beide Modelle zum Scheitern verurteilt, denn auch Verlage sind auf Beteiligung der Fachwissenschaft angewiesen. Die Institutionalisierung und das Fortbestehen von AMAD können demnach nur mit einem redaktionellen Team, freiwilligen Gutachter*innen und engagierten Autor*innen und Leser*innen aus der Fachcommunity gelingen. Mit dem Unterschied, dass der Selbstregulierung der Fachwissenschaft wieder mehr Raum gegeben wird, indem Rechte bei den Autor*innen verbleiben, Begutachtungsprozesse transparent ablaufen, der wissenschaftliche Diskurs wieder partizipativ und frei von marktökonomischen Erwägungen geführt werden kann. Ein schöner zusätzlicher Anreiz für alle Forschenden, bei künftigen Online-Publikationen die Plattform AMAD.org zu wählen, und wenn es zunächst auch nur für die Zweitveröffentlichung wäre.

Bedarfsgerechte Weiterentwicklung von RADAR als Forschungsdaten-Repositorium für das KIT

Felix Bach[1], Kerstin Soltau[2] und Matthias Razum[2]
[1]Karlsruher Institut für Technologie (KIT);
[2]FIZ Karlsruhe - Leibniz-Institut für Informationsinfrastruktur

Das Karlsruher Institut für Technologie (KIT) baut aktuell ein institutionelles Forschungsdaten Repositorium auf Basis eigener Infrastruktur und des am Steinbuch Centre for Computing (SCC) betriebenen Speicherdienstes bwDataArchive auf. Als Management-Schicht und Benutzeroberfläche soll für die Forschenden die RADAR-Software unter der Bezeichnung RADAR4KIT zum Einsatz kommen und in das Diensteportfolio für das Forschungsdatenmanagement (FDM) der Hochschule integriert werden. Dabei soll das Management von Forschungsdaten sowohl durch die von der RADAR-Software bereitgestellten Dienstleistungen und -merkmale als auch durch die optimale Integration in die Dienstlandschaft des KIT vereinfacht werden.

Das KIT hat jedoch, neben der effektiven Einbindung der KIT-eigenen Infrastruktur, noch weitere institutionsspezifische Anforderungen an RADAR, die dessen bisheriger Funktionsumfang nicht abdeckt. Der Beitrag beschreibt die bedarfsgetriebene Weiterentwicklung der RADAR-Software und die Integration von RADAR für das KIT. Alle neu entwickelten RADAR-Funktionalitäten sind dabei nicht KIT-spezifisch ausgerichtet, sondern werden auch anderen Institutionen zur Verfügung stehen. Die beschriebenen Anpassungen stellen eine Öffnung RADARs für neue, alternative Einsatzszenarien dar und bedingen eine Weiterentwicklung des RADAR-Geschäftsmodells, der Dienstleistungsverträge und der Leistungsbeschreibung.

1. Einleitung

Das KIT[1] baut aktuell ein institutionelles Forschungsdaten-Repositorium auf Basis eigener Infrastruktur und des am SCC[2] betriebenen Speicherdienstes bwDataArchive[3] auf. Als Management-Schicht und Benutzeroberfläche soll für die Forschenden die RADAR-Software unter der Bezeichnung RADAR4KIT zum Einsatz kommen und in das FDM-Diensteportfolio der Hochschule integriert werden. Das Ziel dabei ist es, das Management von Forschungsdaten für die Wissenschaftlerinnen und Wissenschaftler am KIT durch optimale Integration in die vorhandene Dienstelandschaft des KIT zu vereinfachen. Hierzu

[1] https://www.kit.edu
[2] https://www.scc.kit.edu
[3] https://www.rda.kit.edu

Felix Bach, Kerstin Soltau und Matthias Razum: Bedarfsgerechte Weiterentwicklung von RADAR als Forschungsdaten-Repositorium für das KIT. In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge, Heidelberg: heiBOOKS, 2020, 162–172. DOI: https://doi.org/10.11588/heibooks.598.c8425

können Diensteigenschaften von RADAR wie beispielsweise das Rollen- und Rechtekonzept, das für die Langzeitarchivierung am KIT geeignete Datenmanagement, die Möglichkeit der Datenpublikation mit DOI-Vergabe, die Integrationsmöglichkeiten via API und das disziplinübergreifende Metadatenschema einen entscheidenden Beitrag leisten.

Das KIT hat neben der effektiven Einbindung der eigenen Speicherinfrastruktur noch weitere institutionsspezifische Anforderungen an RADAR, wie z.B. die Verwendung eines eigenen DOI-Prefix, die Möglichkeit einer institutionellen Sicht auf eigene Forschungsdatensätze und Anpassungsmöglichkeiten der Plattform an das Corporate Design des KIT.

Diese Anforderungen werden bisher noch nicht vollständig durch den bestehenden Funktionsumfang von RADAR und dessen Betriebsmodell abgedeckt. Die notwendigen Anpassungen und Erweiterungen der RADAR-Software, die zukünftig auch allen anderen Institutionen zur Verfügung stehen, werden in diesem Beitrag beschrieben.

2. Anforderungen des KIT

2.1. Integration eines Forschungsdaten-Repositoriums in die FDM-Dienste des KIT

Das KIT hat 2016 eine Forschungsdaten-Policy[4] verabschiedet, in der es sich zu einem verantwortungsvollen und nachhaltigen Umgang mit Forschungsdaten verpflichtet. Wissenschaftler und Wissenschaftlerinnen sollen durch geeignete Infrastrukturen und Dienste beim FDM unterstützt werden. Hierzu wurde das Serviceteam RDM@KIT[5] ins Leben gerufen, das Forschenden Support entlang des Forschungsdaten-Zyklus bietet - insbesondere zu FDM-Diensten, Metadaten, Forschungsdaten (FD)-Archivierung und dem Aufbau von Repositorien und elektronischen Laborjournalen (engl. electronic lab notebooks, ELN).

Die bislang am KIT verfügbaren FDM-relevanten Dienste (siehe Abbildung 1) decken bereits die Bedarfe der Forschenden in den meisten Phasen des FD-Zyklus gut ab. So unterstützt bereits bei der Planung des FDM neuer Projekte der sog. Reseach Data Management Organizer (RDMO)[6] und es gibt ein breites Informationsangebot auf den Webseiten des Serviceteams RDM@KIT und auf forschungsdaten.info[7]. Fachspezifische Labore, ELN und Virtuelle Forschungsumgebungen (VFU) unterstützen das Sammeln und die Analyse von Daten. Speicher- und Archivsysteme wie bwDataArchive stehen zur Verfügung, das Teilen und Zusammenarbeiten an Dokumenten wird zumindest für kleinere Dokumente und Programm-Quellcode durch Dienste wie bwSync&Share[8] und GitLab[9] möglich. Außerdem stehen einige fachspezifische Repositorien mit erweitertem Funktionsumfang bereit (z.B. Chemotion[10]), die in aktuellen Projekten (z.B. Scientific Data Center MoMaF)[11]

[4] https://www.rdm.kit.edu/downloads/KIT-FDM-Policy.pdf
[5] https://www.rdm.kit.edu
[6] https://rdmo.forschungsdaten.info
[7] https://www.forschungsdaten.info
[8] https://bwsyncandshare.kit.edu
[9] https://git.scc.kit.edu/KIT
[10] https://www.chemotion.net
[11] https://www.kit.edu/kit/pi_2019_021_molekuel-und-materialforschung-daten-leicht-teilen.php

für eine Nutzbarkeit durch benachbarte Fachbereiche weiterentwickelt werden. Des Weiteren gewährleisten zentrale Nachweis- und Suchsysteme den Zugriff auf archivierte FD, die zusammen mit Text-Publikationen[12] abgelegt werden. Was dem KIT allerdings momentan noch fehlt, ist eine Möglichkeit für Forschende, Forschungsdaten - auch große Datenmengen - mit anderen zu teilen oder zentral zu publizieren, zusammen mit entsprechenden Metadaten und referenzierbar über eine DOI.

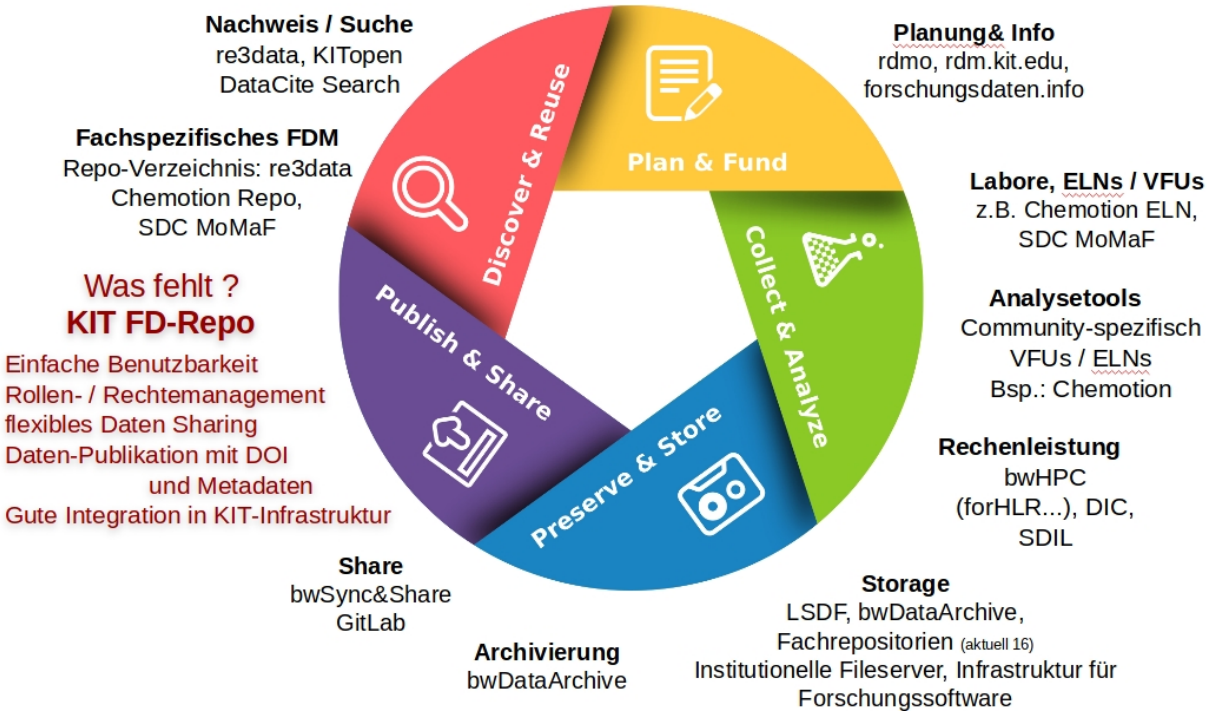


Abbildung 1.: Die FDM-Dienstlandschaft am KIT entlang des FD-Zyklus.

Diese fehlende Funktionalität bietet das generische Forschungsdaten-Repositorium RADAR[13]. Jedoch benötigt das KIT den Dienst RADAR nicht genau so, wie er aktuell durch FIZ Karlsruhe - Leibniz-Institut für Informationsinfrastruktur angeboten wird (siehe Leistungsbeschreibung[14]). Dies hat im Wesentlichen zwei Gründe: Das KIT möchte zum einen ein eigenes, an das Corporate Design angepasstes fachbereichübergreifendes Forschungsdaten-Repositorium, das nur Datensätze von KIT-Wissenschaftlern enthält und präsentiert. Und zum anderen betreibt das KIT eigene, groß angelegte Speicher-Infrastrukturen und -Dienste, die in einem institutionellen Forschungsdaten-Repositorium so genutzt und integriert werden sollen, dass einerseits effiziente Datenflüsse ermöglicht werden und andererseits die Daten ausschließlich am KIT gespeichert werden.

[12] https://www.bibliothek.kit.edu/cms/kitopen.php
[13] https://www.radar-service.eu
[14] https://www.radar-service.eu/sites/default/files/Dienstbeschreibung_RADAR.pdf

2.2. Anbindung von RADAR4KIT an die KIT-Infrastruktur

RADAR bietet bislang als "All-In-One" Cloud-Dienst für Hochschulen und außeruniversitäre Forschungseinrichtungen eine generische Infrastruktur zur langfristigen Archivierung und Publikation digitaler Forschungsdaten, ohne dass diese eigene Infrastruktur betreiben müssen. Dies ist vor allem für kleinere Einrichtungen von Vorteil, für die sich oft der Betrieb von Repositorien an eigenen Rechenzentren nicht lohnt, passt jedoch nicht optimal für das KIT, das mit dem SCC ein vollständiges Informationstechnologiezentrum betreibt, welches den Ansprüchen der zahlreichen Großgeräte, die am KIT betrieben werden und den großen Datenmengen, die am KIT verarbeitet werden müssen, gerecht wird.

Die RADAR-Software wurde in einem DFG-Projekt[15] (2013-2016) von einem interdisziplinären Projektkonsortium bestehend aus fünf Forschungseinrichtungen entwickelt, das sowohl Community-Vertreter als auch Infrastruktureinrichtungen umfasste. Das KIT war durch das SCC selbst vertreten, wodurch die Bedarfe des KIT schon bei der Entwicklung berücksichtigt werden konnten. Des Weiteren hostet das SCC den größten Teil des RADAR-Dienstes von FIZ Karlsruhe, so dass die primäre IT-Infrastruktur für dessen Betrieb durch das SCC gestellt wird und optimal an die dortigen Speichersysteme angebunden ist. Damit lag das Aufsetzen und Betreiben einer eigenen KIT-Instanz von RADAR (RADAR4KIT) nahe, die ausschließlich eigene technische Infrastruktur verwendet, jedoch in Sachen Nachhaltigkeit davon profitiert, dass eine Weiterentwicklung der RADAR-Software durch FIZ Karlsruhe gewährleistet ist.

Um zu verdeutlichen, wie RADAR4KIT sich in die Speicherinfrastrukturen am KIT integriert, ist in Abbildung 2 die IT-Architektur am SCC zu sehen, auf der u.a. der Dienst bwDataArchive basiert. Der Landesdienst bwDataArchive[16] bietet Wissenschaftlern und Wissenschaftlerinnen Zugang zu einer technischen Infrastruktur zur langfristigen Datenarchivierung, die insbesondere für Universitäten und öffentliche Forschungseinrichtungen aus Baden-Württemberg zur Verfügung gestellt wird. Die Datenarchivierung erfolgt am KIT und umfasst eine verlässliche Speicherung auch großer Datenbestände für einen Zeitraum von zehn oder mehr Jahren. Der Dienst ermöglicht eine qualifizierte Umsetzung der Empfehlungen der Deutschen Forschungsgemeinschaft (DFG) zur Sicherung und Aufbewahrung von Forschungsdaten.

Der Aufbau von bwDataArchive wurde in einem Landesprojekt vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK) gefördert und arbeitete im Rahmen des Helmholtz-Projekts LSDMA eng mit verschiedenen wissenschaftlichen Communities sowie den internationalen Projekten EUDAT, Human Brain Project (HBP) und dem World Wide LHC Computing Grid (WLCG) zusammen.

Im unteren Bereich von Abbildung 2 ist das Speicherbackend zu sehen, welches das High Performance Storage System (HPSS)[17] verwendet, das von der HPSS Collaboration und IBM entwickelt wurde. HPSS ist ein skalierbares, hierarchisches Storage Management (HSM) System, das eine Kombination aus einem schnellen, zentralen Plattenspeicher und langsameren, örtlich verteilten Bandspeichersystemen nutzt. Dieses Backend wird

[15] https://www.radar-service.eu/sites/default/files/publications/Abschlussbericht_DFG-Projekt_RADAR_Vero%CC%88ffentlichung.pdf

[16] https://www.rda.kit.edu

[17] http://www.hpss-collaboration.org

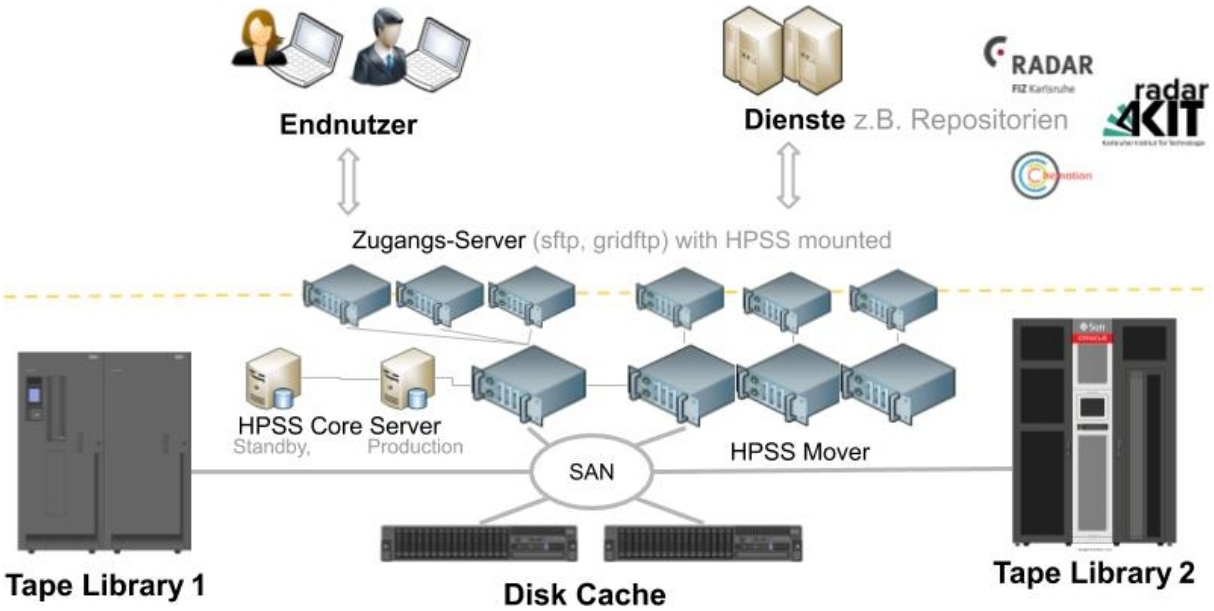


Abbildung 2.: Die bwDataArchive-Infrastruktur und aufsetzende Dienste.

auf mehreren Zugangsservern gemountet, die den Speicher über die Protokolle SFTP und GridFTP zugänglich machen. Hierüber können nun einerseits Endnutzer über den Dienst bwDataArchive zugreifen, als auch eigenständige Dienste wie etwa Repositorien. Momentan zählt zu diesen Diensten neben RADAR auch das Chemie-Repositorium Chemotion und zukünftig auch RADAR4KIT.

3. Forschungsdaten-Repositorium RADAR

Seit 2017 bietet RADAR akademischen Institutionen und Forschenden eine generische Infrastruktur zur langfristigen Archivierung und Publikation digitaler Forschungsdaten. RADAR wird von FIZ Karlsruhe - Leibniz-Institut für Informationsinfrastruktur[18] als disziplinübergreifender "All-in-One" Cloud-Dienst angeboten und wendet sich derzeit primär an Hochschulen und außeruniversitäre Forschungseinrichtungen, die keine eigene Forschungsdateninfrastruktur betreiben oder die RADAR ergänzend zu existierenden disziplinspezifischen Angeboten nutzen möchten.

Die technische Infrastruktur von RADAR wird von FIZ Karlsruhe, dem SCC und dem Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) der TU Dresden[19] bereitgestellt. Die Speicherung der Forschungsdaten erfolgt in drei Kopien an geographisch getrennten Standorten in den Rechenzentren des SCC (zwei Kopien) und des ZIH (zusätzliche Kopie). Die TIB Hannover[20] registriert DOIs für sämtliche über RADAR publizierte Forschungsdatensätze. Der komplette RADAR-Service und seine Infrastruktur unterliegen damit den rechtlichen Bestimmungen der Bundesrepublik Deutschland.

[18] `https://www.fiz-karlsruhe.de`
[19] `https://tu-dresden.de/zih`
[20] `https://www.tib.eu`

3.1. RADAR Dienstleistungen und Geschäftsmodell

Das RADAR-Angebot umfasst drei zentrale Dienstleistungen:

Die **Datenarchivierung** dient der sicheren und formatunabhängigen Aufbewahrung von Forschungsdaten über flexibel wählbare Haltefristen (5, 10, 15 Jahre). Die Forschungsdaten werden dabei in Form paketierter Zusammenstellungen gesichert und erhalten einen eindeutigen Identifier. Sofern von den Datengebenden nicht anders vorgesehen, werden archivierte Forschungsdaten und zugehörige Metadaten nicht veröffentlicht. Das Teilen archivierter Datensätze mit anderen Nutzerinnen und Nutzern ist jedoch über eine flexible Zugriffsverwaltung möglich.

Bei der **Datenpublikation** werden die Datensätze für mindestens 25 Jahre gesichert. Jedes publizierte Datenpaket erhält einen persistenten Identifier (DOI), wird automatisch bei DataCite indexiert und über standardisierte Protokolle (OAI-PMH) zum Harvesting angeboten. Dies sorgt für maximale Verbreitung und Auffindbarkeit der Forschungsdaten. Über die DOI ist der Datensatz eindeutig und dauerhaft identifizierbar, zitierfähig und kann mit wissenschaftlichen Publikationen verknüpft werden. Bei Bedarf können DOIs bereits vor der Datenpublikation reserviert werden. Optional kann die Datenpublikation über eine Embargofrist von bis zu einem Jahr verzögert werden. Für jedes publizierte Datenpaket muss die Datengeberin bzw. der Datengeber eine Lizenz (z.B. Creative Commons 4.0) vergeben, welche dessen Nachnutzung regelt.

Vor der Datenpublikation können im Rahmen eines **Peer-Review-Prozesses** externe Gutachterinnen und Gutachter, beispielsweise von Verlagen, die Forschungsdaten zudem über einen sicheren Link begutachten.

RADAR ist auf langfristigen Betrieb ausgelegt und operiert nicht gewinnorientiert. Bereits während der Projektphase wurde ein nachhaltiges **Geschäftsmodell** erarbeitet, das - nach einer fünfjährigen Anlaufphase - den Betrieb ohne Projektförderung sichern soll. FIZ Karlsruhe als Betreiber sieht die Archivierung und Publikation wissenschaftlicher Forschungsdaten als wichtigen Teil seines öffentlichen Auftrags und in Übereinstimmung mit dem eigenen Leitmotiv „Advancing Science“ - den gesamten wissenschaftlichen Wertschöpfungsprozess in allen Stufen, in denen Daten anfallen und Information und Wissen relevant sind, zu unterstützen. Daher übernimmt FIZ Karlsruhe die Hälfte der operativen Fixkosten aus der eigenen Grundfinanzierung. Der verbleibende Fixkostenanteil und alle variablen Betriebskosten sollen über Nutzungsgebühren eingenommen werden. Die Nutzung als Forschungsdaten-Repositorium für Institutionen setzt deshalb den Abschluss eines Dienstleistungsvertrags[21] voraus, für den eine jährliche Grundgebühr anfällt. Die in Anspruch genommenen Dienstleistungen Datenarchivierung und Datenpublikation werden darüber hinaus nutzungsbasiert je nach angefallenem Datenvolumen in Rechnung gestellt[22]. Während für archivierte Daten innerhalb der gewählten Haltefrist jährlich Kosten berechnet werden, fallen für publizierte Daten nur im Jahr der Datenpublikation Kosten in Form einer Einmalzahlung an. Neben der Preistransparenz gewährt dieses Abrechnungsmodell nutzenden Einrichtungen zum einen die Möglichkeit, archivierte Daten nach Vertragsende flexibel an andere Dienstleister weiterzugeben und zum anderen eine von

[21] `https://www.radar-service.eu/sites/default/files/Dienstvertrag_RADAR.pdf`
[22] `https://www.radar-service.eu/de/preise`

der Vertragssituation unabhängige garantierte Haltefrist von 25 Jahren bei publizierten Daten.

3.2. RADAR Dienstmerkmale

RADAR unterstützt Einrichtungen beim Forschungsdatenmanagement, indem sich das System flexibel an institutionelle FDM-Workflows anpassen lässt. Zu den in diesem Zusammenhang wichtigsten Dienstmerkmalen zählen das disziplinübergreifende Metadatenschema, die offene Systemarchitektur sowie das Rollen- und Rechtekonzept, welches die delegierte Administration durch eine nutzende Einrichtung erlaubt.

Das **RADAR Metadatenschema**[23] ist disziplinübergreifend angelegt, kompatibel mit dem DataCite Metadata Schema[24] sowie DublinCore[25] und fördert die Umsetzung der FAIR Prinzipien.[26] Es definiert 10 Pflichtfelder, die für die DOI-Registrierung des Forschungsdatensatzes notwendig sind, und 13 optionale Felder zur Beschreibung der Datenerhebung und -aufbereitung. Das Schema erlaubt die Verwendung von Normdaten für Personen (ORCID iD[27]) und Förderorganisationen (CrossRef Open Funder Registry). Durch die Kombination aus kontrollierten Vokabularen und Freitext-Einträgen ermöglicht RADAR die Interoperabilität der beschriebenen Forschungsdaten und trägt gleichzeitig der Heterogenität der Daten aus einer Vielzahl von Disziplinen Rechnung. Neben der Beschreibung auf Ebene des Forschungsdatensatzes erlaubt RADAR auch die Beschreibung mit Metadaten auf Datei- und Verzeichnisebene.

Die **RADAR-Systemarchitektur** (Abbildung 3) ist modular aufgebaut und besteht aus dem User Interface (Frontend), der Management-Schicht (Backend) und der Speicherschicht (Archiv), welche das OAIS-konforme Langzeitarchivierungssystem implementieren. Die Schichten kommunizieren über Application Programming Interfaces (API) miteinander. Dieser offene Aufbau ermöglicht die Integration von RADAR in bestehende Systeme und Arbeitsprozesse, wobei einzelne Komponenten von RADAR gegen eigene Lösungen ausgetauscht beziehungsweise parallel betrieben werden können. Anwendungsfälle umfassen beispielsweise den Betrieb eines institutionseigenen Frontends, das automatisierte Hochladen von Forschungsdaten oder die Übertragung beziehungsweise den Abruf von (Meta-)Daten aus anderen Anwendungen. Über das Frontend können Benutzer und Rollen verwaltet, die institutionseigene RADAR-Umgebung adminstriert und gestaltet, Arbeitsbereiche eingerichtet, Daten hochgeladen, zu Datenpaketen zusammengestellt und mit Metadaten versehen sowie archiviert bzw. publiziert werden. Die Management-Schicht umfasst die gesamte Geschäftslogik von RADAR. Sie unterteilt sich in den Ingest Service und das Repository Management. Der Ingest Service nimmt einzelne Dateien und gepackte Archive (z.B. ZIP, gZIP, TAR) entgegen, entpackt sie und erzeugt ein neues Datenpaket. Eine eventuell im Dateiarchiv bereits vorhandene Verzeichnisstruktur wird dabei auch im neuen Datenpaket abgebildet. Das Repository Management verwaltet die für RADAR

[23] https://www.radar-service.eu/de/radar-schema
[24] https://schema.datacite.org
[25] http://dublincore.org
[26] https://www.force11.org/group/fairgroup/fairprinciples
[27] https://orcid.org

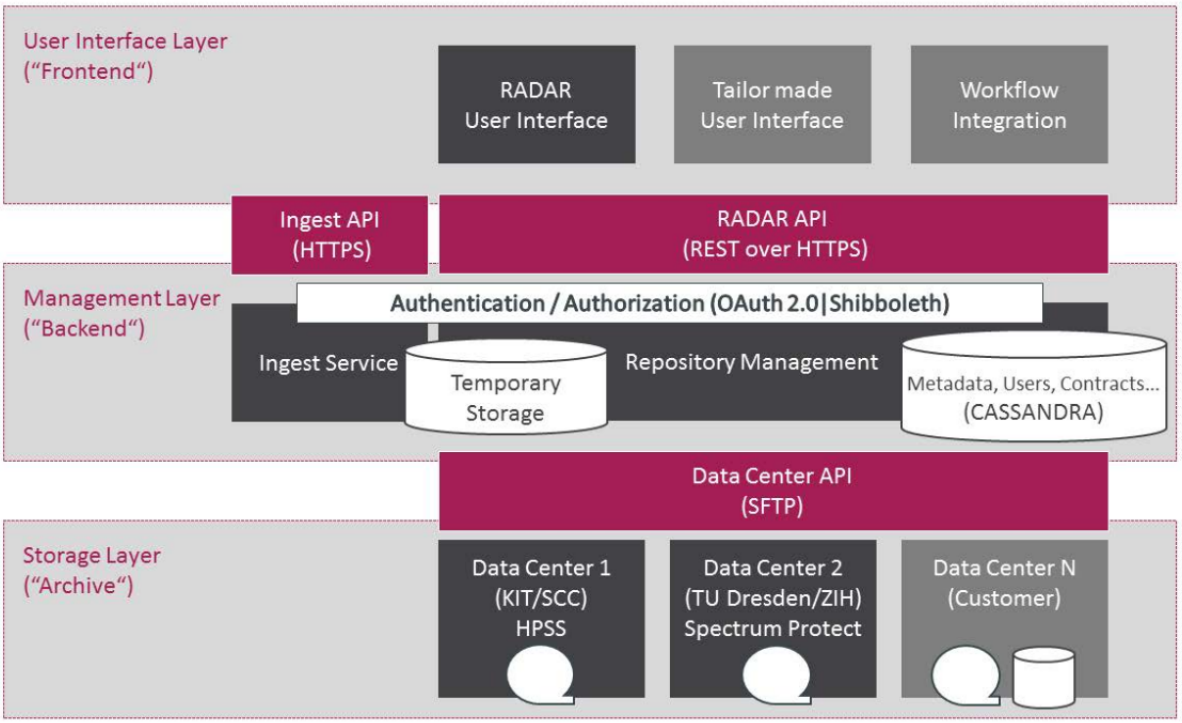


Abbildung 3.: Die RADAR Systemarchitektur.

wichtigen Entitäten wie Verträge, Arbeitsbereiche, Datenpakete, Verzeichnisse und Dateien. Darüber hinaus enthält es Funktionen zum Validieren von Metadaten, zum Archivieren bzw. Publizieren von Datenpaketen, zur Erstellung von Statistiken und zur Überwachung von Quotas. Auch der Lebenszyklus der Daten innerhalb des RADAR-Workflows mit den vier möglichen Stati "in Bearbeitung", "in Begutachtung", "Archiviert" und "Publiziert" wird von der Management-Schicht implementiert. Das **RADAR Rollen- und Rechtemodell** (Abbildung 4) ermöglicht die delegierte Administration durch die nutzende Einrichtung. Entsprechend der institutionellen Bedürfnisse können Arbeitsprozesse strukturiert, Aufgaben verteilt und interne Verantwortlichkeiten definiert werden. Von der Einrichtung eingesetzte Administratorinnen und Administratoren verwalten die RADAR-Arbeitsbereiche, die als zentrale Einstiegspunkte für Forschende eines Projekts oder einer Arbeitsgruppe dienen. Administratorinnen und Administratoren können Kuratorinnen und Kuratoren bestimmen, die Forschungsdaten in einem Arbeitsbereich ablegen, mit Metadaten beschreiben und nach qualitätssichernden Maßnahmen archivieren oder publizieren. Optional können Subkuratorinnen und Subkuratoren benannt werden, die Forschungsdaten bearbeiten und beschreiben, jedoch weder archivieren noch publizieren können. Über eine integrierte Nutzerregistrierung wird die Nutzung von RADAR durch Forschende administrativ erleichtert. Die Authentifizierung in RADAR ist sowohl über eine lokale Datenbank als auch über delegierte Verfahren wie z.B. Shibboleth möglich. Auch ein gemischter Betrieb ist denkbar. Sofern die nutzende Einrichtung an DFN-AAI[28] teilnimmt, erfolgt die Authentifizierung mit der institutionellen Nutzerkennung.

[28] `https://www.aai.dfn.de`

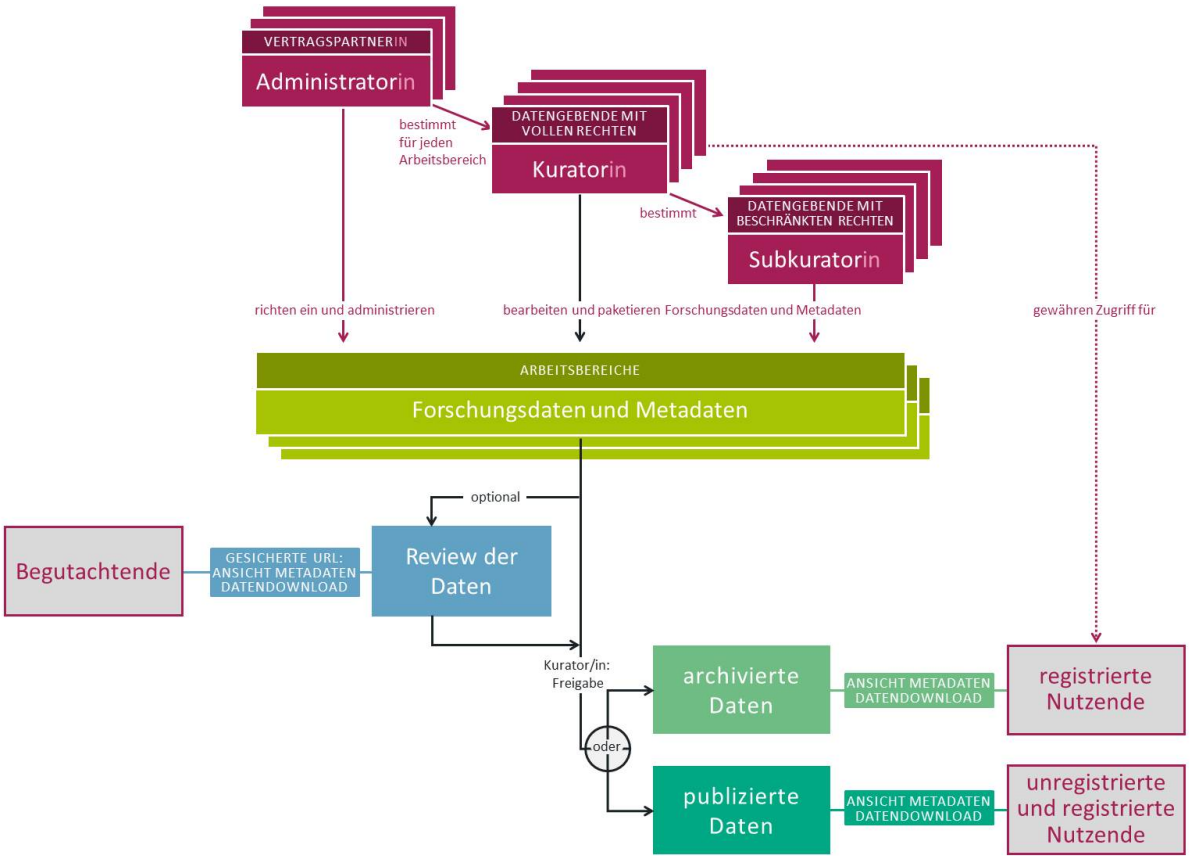


Abbildung 4.: Das RADAR Rollen- und Rechtemodell.

4. Anpassungen für RADAR4KIT

Die genannten Dienstmerkmale sind – neben dem besonderen Verhältnis von KIT und FIZ Karlsruhe als ehemalige RADAR-Projektpartner sowie der Rolle des SCC als Bereitsteller der primären IT-Infrastruktur für RADAR – ausschlaggebend für das KIT, sich beim Aufbau des eigenen Forschungsdaten-Repositoriums für die RADAR-Software zu entscheiden. Die bisherigen Möglichkeiten zur einrichtungsspezifischen Anpassung der Software decken jedoch nicht alle vom vom KIT gewünschten Aspekte ab.

4.1. Neue Anforderungen für einrichtungsspezifische Anpassungen

Aus den oben genannten Gründen sollen weitere institutionsspezifische Anpassungen im engen Austausch miteinander realisiert werden, die ein passgenaues Forschungsdaten-Repositorium für das KIT auf Basis eigener Infrastruktur ermöglichen. Dazu zählen:

- die automatische Vergabe von DOIs für publizierte Datensätze mit einem KIT-eigenen Präfix

- Branding-Optionen, die die Anpassung der Benutzeroberfläche an das KIT-Corporate Design ermöglichen, beispielsweise die Einbindung eines KIT-eigenen Logos sowie die Definition einer alternativen primären Schmuckfarbe
- der Betrieb der RADAR4KIT-Software unter einer eigenen URL auf KIT-eigener Domain
- die Möglichkeit, von der RADAR4KIT-Benutzeroberfläche auf verschiedene KIT-eigene Unterstützungsangebote zu verweisen
- die Möglichkeit einer institutionellen Sicht, über die standardmäßig ausschließlich die vom KIT publizierten Datensätze angezeigt werden
- die Verfügbarkeit der Metadaten als eigenes Set über den OAI-Provider von RADAR zum Harvesting durch Dritte
- die ausschließliche Verwendung KIT-eigener Speicherinfrastruktur, wie in 4.2 beschrieben.

4.2. Anpassungen auf Ebene der RADAR-Speicherschicht

Die technische Infrastruktur der Speicherschicht von RADAR als "All-In-One" Cloud-Dienst wird aktuell von den beiden Rechenzentren bereitgestellt, die derzeit RADAR-Daten hosten. Die Forschungsdaten sind dabei redundant in mehreren Kopien (zwei am SCC, eine weitere am ZIH) an unterschiedlichen, geographisch verteilten Standorten abgelegt, was die Zuverlässigkeit von RADAR erhöht. RADARs Speicherschicht ist über die Data Center API gekapselt und "verbirgt" so die von den Rechenzentren eingesetzten Technologien zur dauerhaften Speicherung der Datenpakete vor der Management-Schicht. Die RADAR-Systemarchitektur garantiert dadurch nicht nur die Unabhängigkeit von einer speziellen Speichertechnologie, sondern schafft gleichzeitig auch die Möglichkeit, weitere Rechenzentren in die technische Infrastruktur einzubinden oder bestehende zu ersetzen. Die Data Center API ermöglicht es somit auch, pro Kunde die zu verwendenden Rechenzentren festzulegen und eine oder alle Datenkopien im eigenen Rechenzentrum (sei es als Tape oder als Disk) zu verwahren.

Für RADAR4KIT wird auf dieser Ebene das Archiv des ZIH als Backup-Speicher deaktiviert. Das KIT wird seine Forschungsdaten somit ausschließlich auf dem vor Ort betriebenen Speicherdienst bwDataArchive in zwei Kopien an unterschiedlichen Standorten verwahren, daneben jedoch das RADAR-Frontend, die RADAR-Managementschicht und die RADAR API regulär nutzen.

4.3. Auswirkungen auf Leistungsbeschreibung, Vertrag und Vergütung

Vor allem die beschriebene Modifikation der RADAR-Systemarchitektur für RADAR4KIT beinhaltet nicht nur technische Herausforderungen. Das neue, alternative Betriebszenarium hat auch Auswirkungen auf die Leistungsbeschreibung, den RADAR-Dienstleistungsvertrag inklusive seiner Haftungsregelungen und die Vergütung.

So entfällt beispielsweise im Gegensatz zum gehosteten "All-In-One" Cloud-Dienst bei RADAR4KIT die dritte Kopie der Daten am ZIH, aber auch die Haftung von FIZ Karlsruhe für die gespeicherten Daten. Aus diesem Grund werden Dienstvertrag und Preisgestaltung, basierend auf einer jährlichen Pauschalvergütung, individuell zwischen FIZ Karlsruhe und KIT verhandelt. Die Vergütung bezieht sich dabei auf die Leistungen Softwareentwicklung, den Betrieb und die Pflege der RADAR4KIT-Instanz durch FIZ Karlsruhe sowie das Repository-Management.

5. Fazit und Ausblick

Alle unter 4.1 und 4.2 beschriebenen Anpassungen für RADAR4KIT werden aktuell von FIZ Karlsruhe – in enger Abstimmung mit dem Service-Team RDM@KIT – spezifiziert, entwickelt und umgesetzt. Für die institutionelle Lösung kommen dabei ausschließlich eigene Server und Speicherdienste des KIT zum Einsatz und bwDataArchive wird als Archivierungslösung verwendet. Das Forschungsdaten-Repositorium RADAR4KIT wird von FIZ Karlsruhe im Auftrag administriert und weiterentwickelt. Gleichzeitig kann das KIT jedoch eine eigene Plattform im Corporate Design anbieten und Datensätze über RADAR4KIT mit einem eigenen DOI Prefix publizieren.

RADAR4KIT wird im Anschluss sukzessive in die Dienste des KIT integriert werden, v.a. an die bereits etablierten Nachweis- und Speichersysteme. Langfristig soll z.B. ein effizientes Management von Datensätzen ermöglicht werden, die bereits in bwDataArchive archiviert wurden. Das RDM@KIT-Team wird außerdem in Schulungen und Support den Nutzerinnen und Nutzern am KIT RADAR4KIT näherbringen.

Alle vom KIT gewünschten funktionalen Anpassungen wurden so in die Software integriert, dass sie zukünftig auch anderen interessierten Einrichtungen zur Verfügung stehen und sich in Teilen auch im "All-In-One" Cloud-Dienst finden. Hiervon können Institutionen profitieren, die weitergehende einrichtungsspezifische Anpassungen oder eine Anbindung eigener Infrastruktur bzw. eine Integration mit Diensten wie bwDataArchive benötigen. Auch für Forschende an Einrichtungen, die bereits bwDataArchive nutzen, könnte sich das Datenmanagement vereinfachen. Die Einbindung des eigenen Rechenzentrums in die RADAR-Speicherschicht kann zukünftig entweder exklusiv oder ergänzend zu einem der bestehenden RADAR-Datenzentren realisiert werden. Die neue Option der Unterstützung lokaler Datenkopien könnte für Institutionen oder Konsortien attraktiv sein, die bereits in eigene Speicherkapazität investiert haben und daher deren Einbindung in Repositorien anstreben. Dies räumt Einrichtungen mehr Kontrolle über die eigenen Forschungsdaten ein und könnte, aufgrund des Wegfalls variabler Gebühren, zu einer Kostensenkung des gesamten Forschungsdatenmanagements führen und dieses somit befördern.

bwVisu: A Scalable Remote Visualization Service and its Application to Flow Visualization

Aksel Alpay[1], Karsten Hanser[2], Egzon Miftari[1], Dennis Schridde[1], Sabine Richling[1], Martin Baumann[1], Filip Sadlo[2] and Vincent Heuveline[1]

[1]Heidelberg University Computing Centre

[2]Interdisciplinary Center for Scientific Computing, Heidelberg University

bwVisu combines an intuitive web frontend with well-established HPC technologies providing an easy-to-use and powerful remote visualization solution. We present bwVisu at the example of the finite-time Lyapunov exponent, a contemporary visualization method for the analysis of time-dependent flow.

1. Introduction

As computational power grows, so does the size and complexity of scientific data. In addition, the transfer of large data is cumbersome, and usually requires centralized services for the secure storage of scientific data. The visualization of this data is therefore increasingly difficult and can hardly be handled by local workstations with their limited computing power. Remote visualization, i.e., bringing the visualization techniques to the data and transferring only the derived visual representations to the user, is therefore steadily gaining importance.

Remote visualization services such as bwVisu stand in between data acquisition and data archiving in the data life cycle: They allow for a convenient way of analyzing and exploring the data after it has been created and is still located on hot storage.

bwVisu aims at providing scientists from Baden-Württemberg with a scalable service for the remote visualization of scientific data, while focusing on the user experience and ease of use. Section 2 describes the underlying hardware of bwVisu. Section 3 details the middleware that forms the interface between the bwVisu hardware and bwVisu Web. In Section 4, we go into more detail on bwVisu Web and its functionality, while Section 5 exemplifies bwVisu at the example of time-dependent flow visualization.

2. Cluster

For bwVisu, a dedicated visualization cluster is available. The cluster consists of eight compute nodes and two login nodes. Each compute node houses 28 CPU cores and two specialized GPUs, which are particularly suited to serving many concurrent users by means of dedicated virtualization functions. The nodes are connected with an Infiniband

Aksel Alpay et al.: bwVisu: A Scalable Remote Visualization Service and its Application to Flow Visualization.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 173–184. DOI: https://doi.org/10.11588/heibooks.598.c8426

network, and two high-bandwidth uplinks are available for the login nodes. This guarantees fast access for jobs running on the bwVisu visualization cluster to data on other nearby systems, such as the bwForCluster MLS+WISO [19] or SDS@hd [3]. The system is running CentOS and uses the widespread HPC job scheduler Slurm [14].

3. Middleware

A custom middleware provides an interface for the web frontend to manage the visualization jobs. The middleware, called bwVisu *runner*, is written in Python and exposes its features via a flask-based [18] HTTP REST API to the web frontend.

The middleware is designed to be minimally invasive and can be deployed either on nodes of the cluster or on an external system. Figure 1 illustrates the architecture of the bwVisu stack. For job management, the middleware communicates via `ssh` with a set of *command nodes*, where it starts, stops, and monitors jobs with the Slurm commands `sbatch`, `squeue`, and `scancel`. Additionally, the middleware uses `ssh` to access a set of *gateway nodes*, where it establishes forwarding rules of network traffic from ports to the individual compute nodes where the user application runs. This allows the user to connect to his job simply by connecting to a specific network port on one of the gateway nodes. Authentication is handled by the application of the user (e.g., Xpra). In order to scale to many concurrent jobs, the middleware automatically distributes the network load of the jobs across the available gateway nodes.

The state of jobs is continuously monitored from the gateway nodes. This allows the middleware to provide information to the web frontend whether the job is ready to interact with the user, or whether the user has to wait before the job can be accessed (e.g., because it is queued and not yet running, or because it is running but still in an initial start-up phase where network connections are not yet accepted).

All state of the middleware is stored in an `etcd` database, which is a distributed key-value store. Because of the distributed design of `etcd`, it is easy to operate in a redundant manner, and hence allows for a resilient, highly-available deployment of the middleware.

The requirements that the bwVisu middleware imposes on the cluster are designed to be as low as possible: The middleware must be granted ssh access to a dedicated user account, and this account must be able to run Slurm commands on the command nodes and iptables on the gateway nodes[1]. Since the installation of the bwVisu middleware is the only requirement to use a given HPC cluster for bwVisu, extending an existing HPC cluster with bwVisu's remote visualization capabilities is easy and does not necessitate any changes to the existing cluster configuration.

4. bwVisu Web

bwVisu includes the web user interface "bwVisu Web", which allows researchers to start their individual visualization software easily and connect to it using nothing more than

[1] Note that the gateway nodes can also be virtual machines if it is not desired that iptables rules of physical machines in the cluster are modified by the bwVisu software stack.

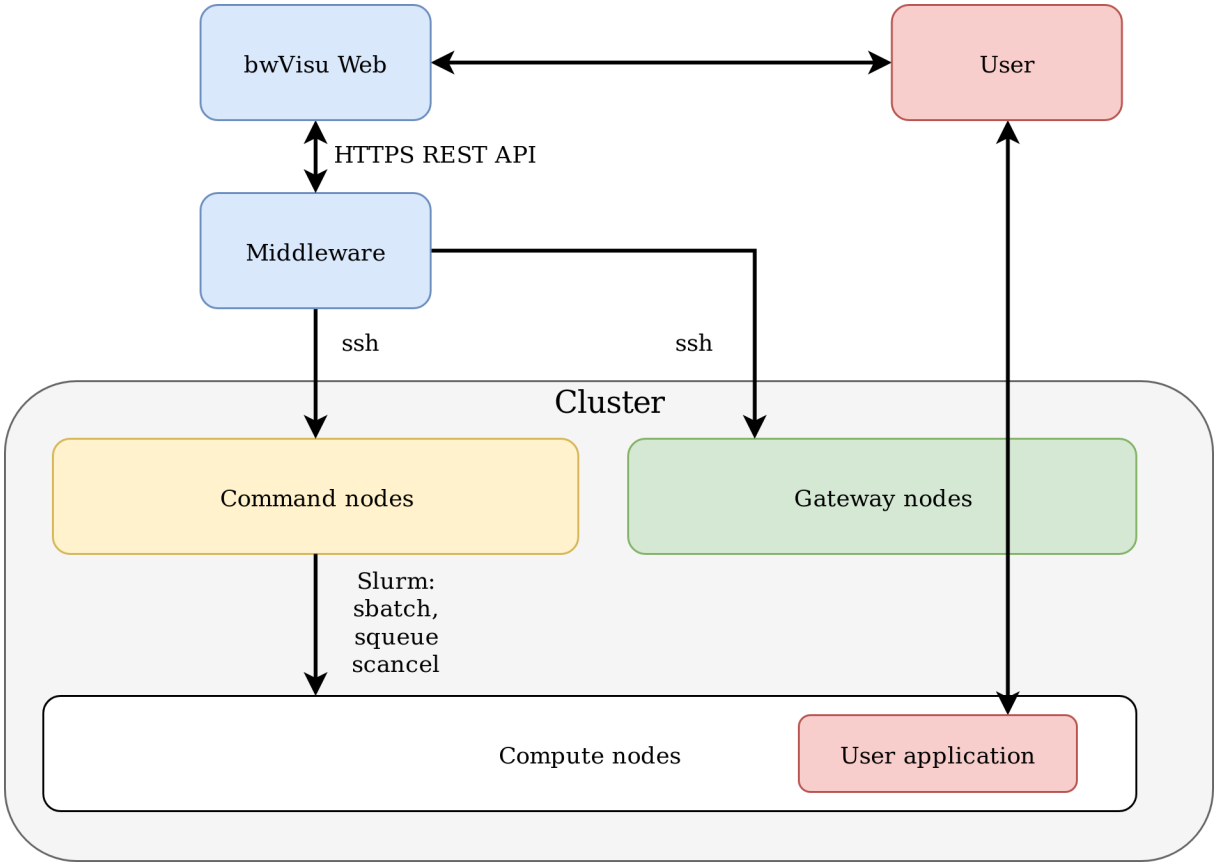


Figure 1.: The architecture of the bwVisu stack.

their web browser (Figure 2). It consists of two parts, the frontend or user interface (UI), and the backend, which supplies the frontend with data and implements the essential "business" logic (Figure 3). The user-visible part of this has only changed marginally from the mockups presented during the last E-Science-Tage conference in 2017 [20], but the architecture was revamped significantly, and the system has matured into a productive system.

While the previous system was built around Kubernetes [1], we now utilize Slurm [14] instead to allow an easier migration path from HPC clusters that previously did not use application containers. Since the latter does not provide an HTTP API that can be used by web applications to interface with it, we no longer interface directly with the HPC scheduler or orchestration engine, but use a custom-made middleware instead, as described in Section 3.

The fact that we were able to reduce the requirements for researchers to just a web browser is an important milestone made possible by the utilisation of the Xpra HTML5 client by Martin and contributors [15], which is embedded into our application container images.

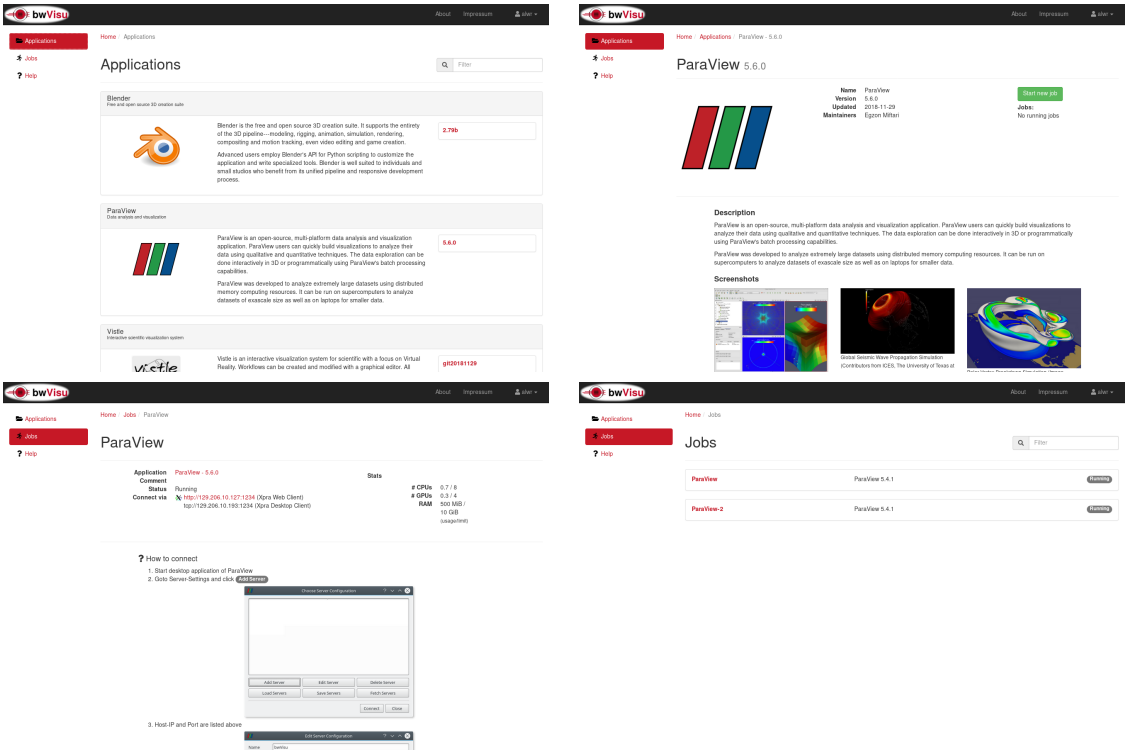


Figure 2.: bwVisu Web aims to be a user-friendly, straightforward web interface allowing users to browse available applications, learn more about them, submit them as jobs to the HPC cluster, as well as manage, connect to, and stop running jobs.

4.1. System architecture

The frontend of bwVisu Web is implemented as a light-weight single page application (SPA). "Light weight" does not only describe the low implementation complexity, but translates also to a clean visual design that is reduced to the essential functions required by users, allowing them to quickly reach their goals without distractions. The frontend application is implemented using a functional data-oriented architecture based on re-frame[2] by Thompson [22]. Rendering and updates of the HTML view is handled by Reagent[3] courtesy of Holmsand and contributors [9].

The internal interface between this frontend and the backend of the bwVisu Web system is realized using GraphQL, again popularised by Inc. [11], which defines an easy-to-use, flexible, well-defined, language-independent query language based on HTTP / WebSockets and JSON that also supports, subscription (live updates) and mutation (write access). bwVisu Web's GraphQL queries can be broken down into three categories:

1. user log in and authentication (GraphQL mutation),
2. queries for available applications and jobs currently running (GraphQL queries and subscriptions),

[2] It shares similarities with the Flux architecture popularised by Inc.[10]
[3] Which itself is based on React by Inc. [12]

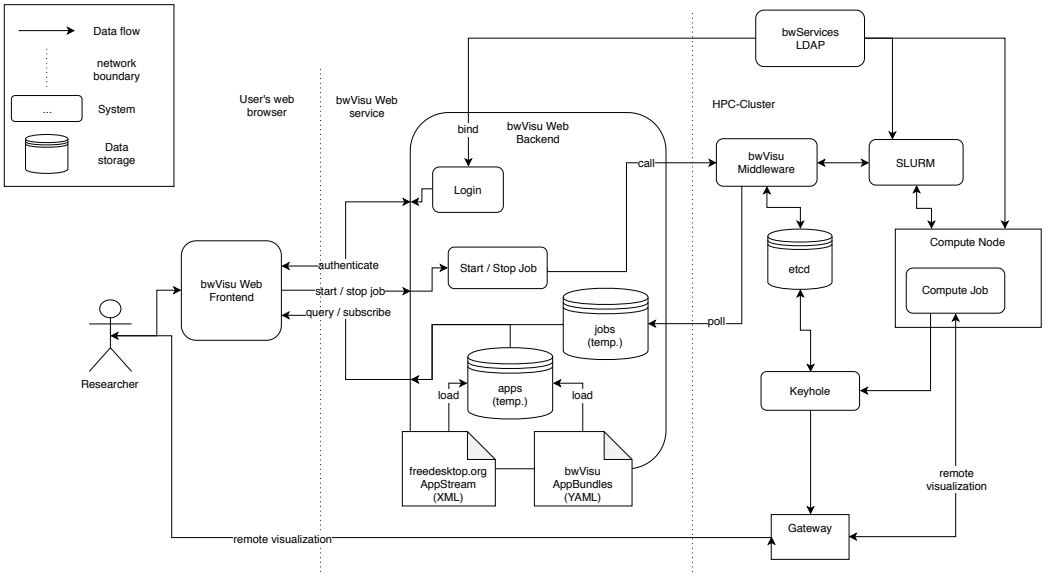


Figure 3.: bwVisu Web consists of two parts, in the user's browser and at the service provider's site, communicating via GraphQL with each other. It shares an LDAP database for user accounts with the HPC-Cluster and communicates via HTTP REST with the Infectoid middleware.

3. commands to start or stop a visualisation job (GraphQL mutations).

In exchange for a username / password pair, the authentication interface provides the bwVisu Web frontend with a JSON Web Token (JWT), which contains information about the user and will be used to authenticate all other queries. This initial log in step is performed against an LDAP database provided by the bwRegApp service developed within the project for the federal state of Baden-Württemberg [17].

The list of applications available in bwVisu is provided to the frontend as a GraphQL query and subscription (and hence supports long running user sessions). All this information is tracked in simple text files (XML or YAML), so it can easily be maintained without special tools and be stored in version control systems, e.g., directly alongside build instructions for the images themselves (e.g., `Dockerfile`s or more sophisticated build systems using `buildah`). Klumpp et al. of freedesktop.org created the AppStream 0.12 [13] infrastructure, in particular the `appdata.xml` file format, to allow developers to provide information about their applications to distributors. Using these facilities has the benefit for bwVisu that accurate and up-to-date application descriptions, screenshots and the likes are made available by the application developers themselves, which reduces the maintenance burden for bwVisu administrators. This choice poses no limitation, because this information can easily be added, should an application not provide it. In custom YAML files, which we call "appbundles", that information is extended with bwVisu-specifics, like available version numbers, corresponding container image names, and information needed by the bwVisu Web middleware in order to submit jobs to the HPC scheduler. Not merging these two into one file provides the benefit of separating

concerns and allows an easier intake of (updated) information from upstream application developers.

In the bwVisu Web frontend, this application information is used to present the list of available applications and their versions/flavours to the user, as well as to show them details, descriptions, and screenshots of individual applications. Finally, it also defines the possible settings users can define when they want to start applications and submit their jobs. These settings are being send to the bwVisu Web backend, where they are merged with any immutable data (e.g., the image name), before being passed on to the middleware. Basic structural conformance checks will be carried out in each of these steps, but actual authorization is done via user account file system permissions by the operating system of the compute nodes when the job actually starts.

Information about currently running jobs is acquired directly from the bwVisu middleware using polling and passed on to the bwVisu Web frontend via GraphQL queries and subscriptions, just like application data. bwVisu Web is equipped to handle dynamic information, such as network addresses, which are only known after job submission, when the application actually starts. Any changes compared to the previous architecture are encapsulated by bwVisu Web and should be opaque to the user.

The only external systems bwVisu Web interfaces with are an LDAP service containing the user database and the bwVisu middleware that abstracts the HPC scheduler. Additionally we avoid any persistent state within bwVisu Web itself (e.g., job databases or user session data), which follows the design proposed by Moseley and Marks in their paper "Out of the Tar Pit" [16]. We confirm their findings that a stateless design simplifies reasoning about the system, which in turn leads to less bugs and increased developer productivity.

4.2. Deployment

bwVisu Web is meant to be easily deployed onto the target HPC cluster. For this reason we selected deployment methods that can be adapted to different target systems. We utilise *Packer* by HashiCorp to build virtual machine images on common platforms like *OpenStack*, which is e.g. used by the Heidelberg *heiCLOUD* platform. The flexibility of Packer allows us to easily build images for other systems, including QEMU/libvirt for consumption on individual non-cloud machines. These virtual machine images are then deployed onto the *OpenStack* platform using *Terraform*, also maintained by HashiCorp. Terraform also allows us to target bare-metal machines and many other types of environments.

5. Visualization Applications

The applications provided in bwVisu are containerized using the well-known container platform Docker and are stored as images in a Docker Container Registry integrated into GitLab. Furthermore the CI/CD service provided by GitLab automates software development and deployment. Since bwVisu builds on well-established HPC technologies and was developed with HPC environments in mind, there is an additional conversion

step of the Docker images to the HPC-focused container platform Singularity. Unlike Docker containers, Singularity containers do not allow privilege escalation inside of them. We provide developers with base images to ensure applications are working properly and all the required dependencies are present at runtime. These base images include MPI and drivers for full hardware acceleration on Nvidia and AMD graphics nodes. Each application image is uniquely labeled and can be archived by the user, thus increasing the degree of reproducibility and portability. In addition, developers will be able to restrict access and visibility of their applications on bwVisu Web if needed. bwVisu already offers a wide range of visualization applications such as ParaView, VMD, VisIt, Blender, Vistle, and many more.

5.1. FTLE-Based Flow Visualization

We exemplify flow visualization of time-dependent flow fields in bwVisu by means of the finite-time Lyapunov exponent (FTLE). The FTLE field [6] measures the divergence of time-dependent trajectories with respect to finite-time advection, and is able to reveal the topology of time-dependent vector fields, i.e., their overall structure with respect to finite-time transport.

Given a time-dependent vector field $\mathbf{u}(\mathbf{x}, t) \in \mathbb{R}^n$ defined at each point $\mathbf{x} \in \Omega \subseteq R^n$ of the domain Ω, a trajectory (i.e., a pathline) is given as the solution to the initial value problem

$$\boldsymbol{\xi}_{\mathbf{x}_0}^{t_0}(t) := \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{u}\left(\boldsymbol{\xi}_{\mathbf{x}_0}^{t_0}(\tau), \tau\right) d\tau, \tag{0.1}$$

with $\mathbf{x}_0$ being the seed point of the pathline, t_0 its seeding time, and t its integration time. From this, the *flow map* is obtained by seeding a pathline at each point $\mathbf{x}$ at time t_0, integrating it for time T, and storing the coordinates of the endpoint of the respective pathline in the map:

$$\boldsymbol{\phi}_{t_0}^{T}(\mathbf{x}) := \boldsymbol{\xi}_{\mathbf{x}}^{t_0}(T). \tag{0.2}$$

Consequently, pathline divergence can be quantified by the gradient of the flow map, leading to the definition of the FTLE:

$$\sigma_{t_0}^{T}(\mathbf{x}) := \frac{1}{|T|} \ln \|\nabla \boldsymbol{\phi}_{t_0}^{T}(\mathbf{x})\|, \tag{0.3}$$

with $\|A\|$ being the spectral norm of matrix A, i.e., the largest eigenvalue of $A^\top A$.

Apparently, the computation of the FTLE field $\sigma_{t_0}^{T}(\mathbf{x})$ is very costly, since integration of a pathline is required for each spatial sample $\mathbf{x}$, and also for each seeding time t_0 (for time-dependent FTLE visualizations, t_0 is varied to produce animations). Even worse, the structure in the FTLE field become more detailed as the advection time T is increased, requiring very high spatial resolutions. On the other hand, the computation of the pathlines is straightforward to parallelize due to their independent computation. As such, FTLE-based visualization is a perfect candidate for demonstrating and evaluating

bwVisu, also because the resulting FTLE data is a scalar field, which is, compared to the huge amount of trajectory data, comparably small. Common approaches to FTLE visualization are direct volume rendering, and extraction of ridge surfaces [5]. In this example, we follow both approaches. We distribute the entire time-dependent vector field to all compute instances, but each instance computes only a part of the FTLE field, i.e., we employ domain decomposition of the FTLE field domain. While parallel volume rendering by bricking (i.e., of the FTLE parts) is somewhat involved, ridge extraction lends itself well for distributed computation. That is, each compute instance extracts ridges from the FTLE part it computed, and sends the result to the visualization entity for interactive exploration.

We start our exploration with a time-dependent vector field obtained from a heat-driven convective 3D ($n = 3$) flow simulation in a closed container. For overview, we have a look at the velocity magnitude at time $t_0 = 5.4\,\mathrm{s}$, and vortex core lines [21], as well as instantaneous streamlines at time $t_0 = 5.4\,\mathrm{s}$ and pathlines seeded at time $t_0 = 5.4\,\mathrm{s}$ and integrated for $T = 0.15\,\mathrm{s}$ (Figure 4). Of course, such basic visualization could be accomplished without parallelization, and as such provides only limited insights in the time-dependent dynamics of this flow.

Our FTLE-based visualization, on the other hand, captures the full time-dependent dynamics. Figure 5 shows the volume rendering of the resulting FTLE field, the height ridge surfaces extracted from the FTLE field, as well as the height ridges surfaces together with a subset of the pathlines that have been used for FTLE computation, to provide flow dynamics context. One can observe that the FTLE ridge surfaces indeed separate regions of qualitatively different flow dynamics and thus provide a topological overview.

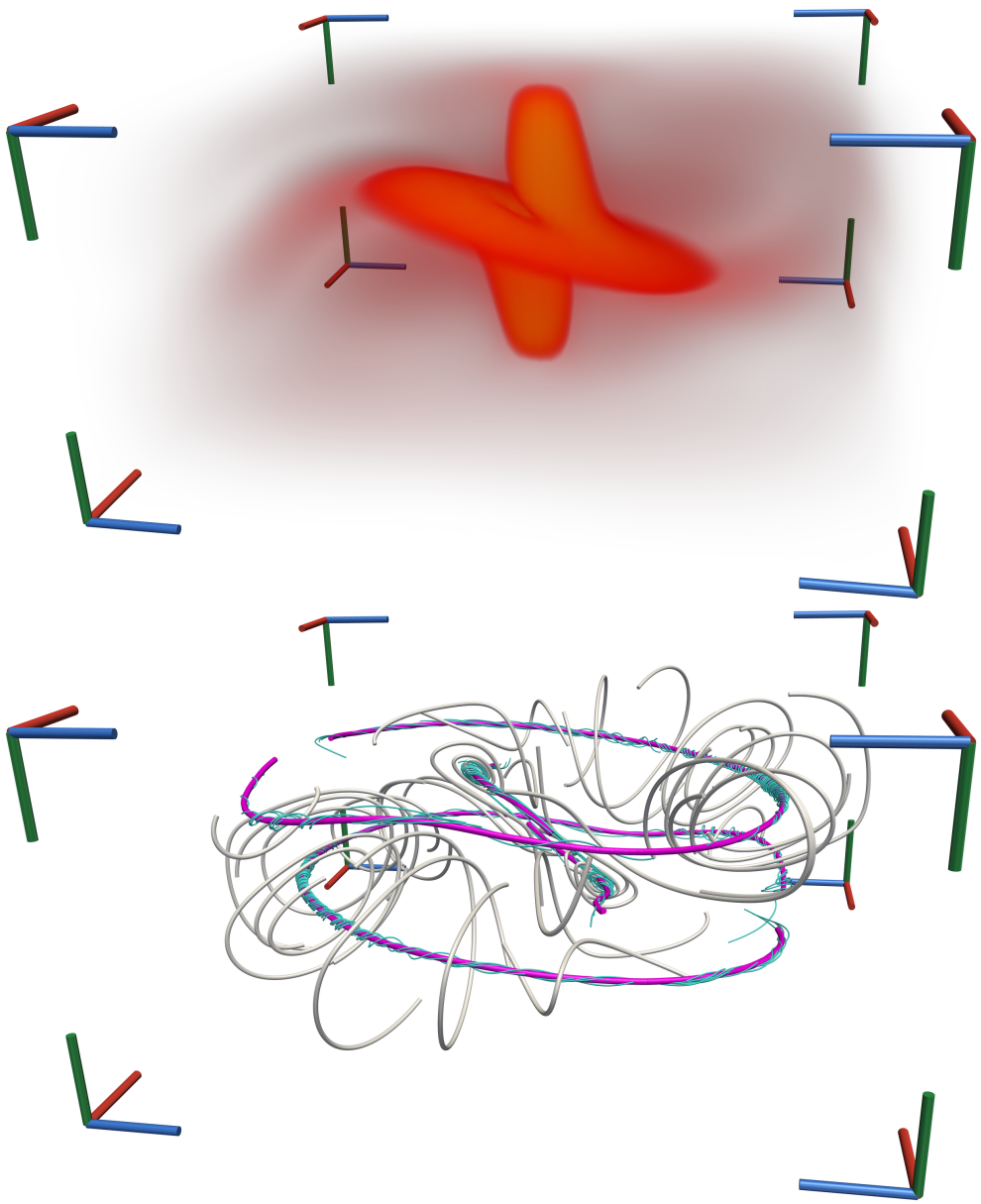


Figure 4.: Overview of Convective Flow example, at time 5.4 s. Top: Velocity magnitude by volume rendering (higher magnitude by higher opacity / brighter colors). Bottom: Vortex core lines (purple), with instantaneous streamlines along vortex core lines (blue), and pathlines seeded at $t_0 = 5.4\,\mathrm{s}$ integrated for $T = 0.15\,\mathrm{s}$ (white).

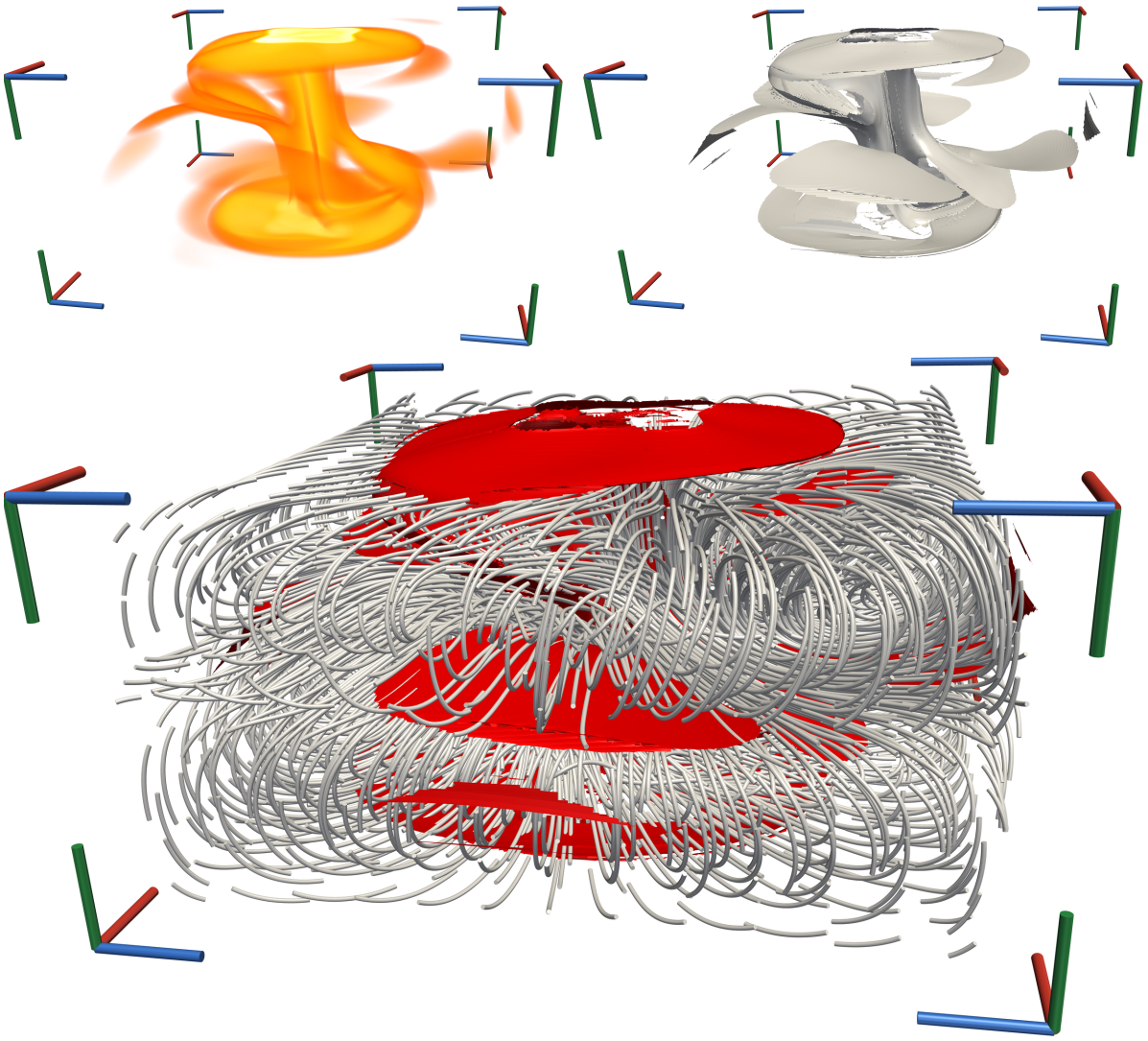


Figure 5.: FTLE-based visualization of Convective Flow example, with $t_0 = 5.4\,\mathrm{s}$ integrated for $T = 0.05\,\mathrm{s}$. Top left: Volume rendering of the FTLE field. Top right: Height ridge surfaces extracted from the FTLE field. Bottom: Height ridges with some of the trajectories used for FTLE computation, for space-time context.

6. Conclusion

We presented bwVisu, a scalable service for remote parallel visualization. We described the underlying hardware, the middleware, which has been designed with a particular focus on ease of use, the web frontend that enables effective and simple visualization, and exemplified its usage at the example of time-dependent flow visualization using the finite-time Lyapunov exponent.

Bibliography

[1] The Kubernetes Authors. Kubernetes. Production-Grade Container Orchestration. Cloud Native Computing Foundation and The Linux Foundation. URL: https://kubernetes.io/ (visited on 04/10/2019).

[2] Software: The OpenStack Authors. OpenStack. URL: https://www.openstack.org/ (visited on 04/10/2019).

[3] Martin Baumann et al. "SDS@hd – Scientific Data Storage". In: Universität Tübingen, 2017. doi: 10.15496/publikation-25204. URL: http://dx.doi.org/10.15496/publikation-25204.

[4] Software: Heidelberg University Computing Centre. heiCLOUD. Infrastructure-as-a-Service.URL: https://heicloud.uni-heidelberg.de/ (visited on 04/10/2019).

[5] David Eberly. Ridges in Image and Data Analysis. Computational Imaging and Vision. Kluwer Academic Publishers, 1996.

[6] G. Haller. "Distinguished material surfaces and coherent structures in three-dimensional fluid flows". In: Physica D 149 (2001), pp. 248–277.

[7] Software: HashiCorp. Packer. URL: https://www.packer.io/(visited on 04/10/2019).

[8] Software: HashiCorp. Terraform. URL: https://www.terraform.io/ (visited on 04/10/2019).

[9] Dan Holmsand and Reagent contributors. Reagent. Minimalistic React for ClojureScript. URL: https://reagent-project.github.io/ (visited on 04/10/2019).

[10] Facebook Inc. Flux. Application Architecture for Building User Interfaces. Facebook Inc. URL: https://facebook.github.io/flux/(visited on 04/10/2019).

[11] Facebook Inc. GraphQL. A query language for your API. Facebook Inc. url: https://graphql.org/ (visited on 04/10/2019).

[12] Facebook Inc. React. A JavaScript library for building user interfaces. url: https://reactjs.org/ (visited on 04/10/2019).

[13] Matthias Klumpp et al. AppStream 0.12. Infrastructure for distro-agnostic software-centers and universal software component metadata.https://www.freedesktop.org/ software/appstream/docs/ (visited on 04/10/2019).

[14] SchedMD LLC. Slurm workload manager.url:https://slurm.schedmd.com/ (visited on 04/10/2019).

[15] Software: Antoine Martin and Xpra contributors. Xpra. multi-platform screen and application for- warding system. URL: https://xpra.org/ (visited on 04/10/2019).

[16] Ben Moseley and Peter Marks. "Out of the Tar Pit". In: Software Practice Advancement (SPA) 2006 (Feb. 6, 2006).

[17] bwHPC project. bwRegApp. bwForCluster registration application. url:https://wiki.bwhpc.de/e/BwForCluster_User_Access#Personal_registration_at_bwForCluster(visited on 04/10/2019).

[18] The flask project. flask – a microframework for python. url:http://flask.pocoo.org/ (visited on 04/10/2019).

[19] Sabine Richling, Martin Baumann, and Vincent Heuveline. "bwForCluster MLS & WISO". In: Proceedings of the 3rd bwHPC-Symposium. Heidelberg: heiBOOKS, 2017. isbn: 978-3- 946531-70-8. doi: 10.11588/heibooks.308.418.10

[20] Dennis Schridde, Martin Baumann, and Vincent Heuveline. "Skalierbare und flexible Arbeitsumgebungen für Data-Driven Sciences". In: E-Science-Tage 2017: Forschungsdatenmanagen. Ed. by Jonas Kratzke and Vincent Heuveline. Heidelberg: heiBOOKS, 2017, pp. 153–166. isbn: 978-3-946531-75-3. doi: 10.11588/heibooks.285.377.

[21] David Sujudi and Robert Haimes. "Identification of swirling flow in 3-D vector fields". In:12th Computational Fluid Dynamics Conference. 1995, p. 1715.

[22] Michael Thompson. Re-Frame: A Reagent Framework For Writing SPAs, in Clojurescript. Mar. 2015. doi: 10.5281/zenodo.801613.

RDMO4Life und Fachrepositorium Lebenswissenschaften im Projekt „Emissionsminderung Nutztierhaltung“ EmiMin – Datenmanagementplan und Publikation von Forschungsdaten in der Agrartechnik

Birte Lindstädt[1] und Katrin Wagner[2]
[1]ZB MED Informationszentrum Lebenswissenschaften, Deutschland;
[2]KTBL Kuratorium für Technik und Bauwesen in der Landwirtschaft

Die Kooperation im Projekt EmiMin besteht aus fünf Instituten aus dem Forschungsfeld der landwirtschaftlichen Verfahrenstechnik in der Nutztierhaltung und dem ZB MED Informationszentrum Lebenswissenschaften als Informationsinfrastruktureinrichtung.

Das Kuratorium für Bauwesen und Technik e.V.(KTBL) als einer der fachlichen Partner hat die Projektkoordination inne und erstellt und betreibt die zentrale Datenbank, in der die gemessenen Emissionswerte sowie die Begleitdaten gespeichert und verarbeitet werden. Aus den Daten sollen in erster Linie Emissionsfaktoren bzw. Minderungsgrade bei Verwendung neuer baulich- technischer Maßnahmen zur Emissionsminderung abgeleitet werden. Die Daten sollen darüber hinaus für weitere Forschungsvorhaben nutzbar sein.

Die entstehenden Forschungsdaten werden von Projektbeginn an in einen Managementprozess einbezogen. Der Managementprozess umfasst die Standardisierung der Messungen über ein vorgegebenes Protokoll, die Dokumentation der Messdaten durch einheitliche Metadaten in der Datenbank, sowie die Publikation und die Nachnutzbarkeit der Ergebnisse sowohl über die KTBL- Datenbank als auch publizierter Messwerte. Ein Ziel dabei ist, Emissionsdaten im Fachrepositorium Lebenswissenschaften (FRL) zu veröffentlichen. Hierzu werden die Forschenden bei dem Veröffentlichungsprozess projektbegleitend beraten und das FRL wird um die zusätzlich erforderlichen Strukturen zur Datenhaltung und Datenbeschreibung erweitert.

Die Erstellung eines Datenmanagementplans wird als Möglichkeit des projektbegleitenden Forschungsdatenmanagements verstanden. Dabei kommt das DFG-geförderte Tool Research Data Management Organizer (RDMO) zur Erstellung und Bearbeitung von DMP zum Einsatz. „Mit dem Research Data Management Organiser (RDMO) können Institutionen und Forschende das Forschungsdatenmanagement ihrer Projekte strukturiert planen und durchführen. Es erlaubt das Erfassen aller relevanten Planungsinformationen in Datenmanagementplänen und die Verwaltung aller Datenmanagementaufgaben über den gesamten Datenlebenszyklus.[1]

[1] `https://rdmorganiser.github.io/`(Zugriff 14.08.2019)

Birte Lindstädt und Katrin Wagner: RDMO4Life und Fachrepositorium Lebenswissenschaften im Projekt „Emissionsminderung Nutztierhaltung“ EmiMin – Datenmanagementplan und Publikation von Forschungsdaten in der Agrartechnik.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 185–186. DOI: https://doi.org/10.11588/heibooks.598.c8427

Aufbauend darauf, dass RDMO als generisches Werkzeug lokal implementiert und weiterentwickelt werden kann, verfolgt ZB MED das Ziel, dieses an fachliche Gegebenheiten der Lebenswissenschaften anzupassen. Dafür wurde RDMO4Life[2] als Tool für die Erstellung und Fortschreibung von Datenmanagementplänen in das Angebot von PUBLISSO, dem Open-Access-Publikationsportal von ZB MED, integriert. RDMO4Life verfolgt in einem ersten Schritt das Ziel, RDMO in seiner generischen Form durch fachspezifische Anpassungen zu einer agrarwissenschaftlich orientierten Software weiter zu entwickeln. Dazu werden u.a. in Workshops mit den Forschenden die Möglichkeiten eines aktiven Forschungsdatenmanagements erörtert sowie projektspezifische Fragebögen für EmiMin entwickelt, getestet und ggf. angepasst. Sie dienen dabei als wesentlicher Bestandteil des Datenmanagements zur Begleitung der Forschung von Beginn an bis zur Publikation von Ergebnissen. RDMO4Life hat damit im Rahmen des Projektes EmiMin Pilotcharakter für zukünftige Tätigkeitsfelder in den Agrarwissenschaften und darüber hinaus langfristig für die Lebenswissenschaften. Die Verbundpartner aus den Fachinstituten bekommen Werkzeuge für das Forschungsdatenmanagement in künftigen Forschungsvorhaben an die Hand.

Wichtig hierbei ist, dass die Schritte gemeinsam entwickelt werden und ein enger Abstimmungsprozess – beispielsweise über die zu publizierenden Daten - erfolgt. Die Motivation für diese Kooperation liegt in der Bündelung von fachlicher und infrastruktureller Expertise zur Schaffung von Mehrwerten für die Forschenden in Form von Transparenz über die Datenstruktur, Zugänglichmachung der erzeugten Daten und Reputationsgewinn. Somit entsteht zum einen Wissen und Kompetenzgewinn bei den Forschenden, zum anderen werden die Forschungsdaten zugänglich und nachnutzbar gemacht, um Messmethoden an weiteren Standorten zu vereinheitlichen, Daten abzugleichen und ggf. zu vergleichen. Die Bewertung verfahrensintegrierter, baulich-technischer Maßnahmen zur Emissionsminderung wird dadurch sowohl standortübergreifend als auch international möglich.

[2] `https://rdmo.publisso.de/` (Zugriff 14.08.19)

RePlay-DH: Ein Werkzeug für Wissenschaftler, um Wissen zu erhalten und zu teilen

Sibylle Hermann and Markus Gärtner
Universität Stuttgart, Deutschland

Der im Projekt RePlay-DH entwickelte Client dient zur Unterstützung der einheitlichen Dokumentation der Arbeitsergebnisse im wissenschaftlichen Workflow. Der Client basiert auf der Versionsverwaltungssoftware git.

Forschungsdaten werden oft nicht veröffentlicht, da sie am Ende eines Projektes nur unstrukturiert vorliegen. Deshalb ist der Ansatz im Projekt RePlay-DH, von Beginn an eine lückenlose Dokumentation zu gewährleisten und somit eine mögliche Datenpublikation zu vereinfachen. Darüber hinaus wird so sichergestellt, dass die Daten auch nach einiger Zeit noch verstanden und nachvollzogen werden können. Der mit einer Prozess-Dokumentation einhergehende Aufwand neben der eigentlichen Arbeit wird von Forschenden häufig als einer der hemmenden Faktoren angegeben, wenn es um das Dokumentieren von Forschungsarbeit geht. Dieser Zusatzaufwand kann hierbei manigfaltige Formen annehmen:

die Einarbeitung in technisch anspruchsvolle Systeme, aufwändige manuelle Erstellung von Metadaten zur Dokumentation, oder selbst die komplette Umstrukturierung der eigenen Arbeitsprozesse. Für eine Akzeptanz von Seiten der Forschungscommunity ist es daher unabdingbar, dass sich angebotene Lösungen möglichst nicht-invasiv und ohne großen Mehraufwand, in etablierte Arbeitsabläufe integrieren lassen. Der von RePlay-DH entwickelte Client bietet hier einen auf die Bedürfnisse der Wissenschaftlerinnen und Wisesnschafler angepassten Kompromiss hinsichtlich einfacher Nutzbarkeit und den technischen Anforderungen an die Anwenderinnen und Anwender.

Das Werkzeug kommt ohne komplizierte Einrichtung aus und erfordert praktisch keine Umstellung von bestehenden Abläufen. Ideal ist die Verwendung für Prozesse, bei denen Daten lokal anfallen oder manipuliert werden. So verhindert die automatische Erkennung von Änderungen an lokalen Dateien im Arbeitsbereich, dass die Dokumentation potentiell wichtiger aber leicht zu übersehender Details ausbleibt.

Die Funktion des Clients, Metadaten zur Beschreibung eines Arbeitsschrittes inkrementell „on the fly“ während dessen Ausführung aufnehmen zu können, ermöglicht eine einfachere Integration in den Forschungsalltag, da der anfallende Zusatzaufwand zeitlich frei verteilt werden kann. Dazu wurde gemeinsam mit Wissenschaftlerinnen und Wissenschaftlern aus der Computer-Linguistik ein Metadatenschema entwickelt, das die Arbeitsschritte in deren Arbeitsalltag möglichst genau und mit geringem Aufwand abbildet. Es besteht die Möglichkeit, pro Workspace ein anderes auf die Bedürfnisse der Wissenschaftsdisziplin angepasstes Metadatenschema zu implementieren.

Sibylle Hermann und Markus Gärtner: RePlay-DH: Ein Werkzeug für Wissenschaftler, um Wissen zu erhalten und zu teilen.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 187–188. DOI: https://doi.org/10.11588/heibooks.598.c8428

Im Client ist ein Ressourcenverwaltungsprogramm integriert, das ermöglicht, Ressourcen wie Textkorpora und Software einmalig mit Metadaten im Client zu dokumentieren, um somit immer wieder auf diese verweisen zu können. Des Weiteren besteht die Möglichkeit, Teil- und Zwischenergebnisse im institutionellen Repositorium zu veröffentlichen oder in einer git-Instanz zu archivieren. Der Replay-Client ist im Projektkontext auf die Bedürfnisse der Computerlinguistik zugeschnitten worden, aber auch andere Textbasierte Wissenschaftsdisziplinen können das Werkzeug verwenden. Der Replay-Client bietet dazu die Möglichkeit eigene Metadaten anzulegen. Des Weiteren können Anpassungen und Weiterentwicklungen des Replay-Clients über die öffentliche git-Instanz (`https://github.com/RePlay-DH/replay-dh-client`) vorgenommen werden.

Forschungsdaten aus Digitalisaten

Stefan Weil und Jan Kamlah
Universitätsbibliothek Mannheim, DE

Bibliotheken tragen einen wichtigen Teil zur Digitalisierung des kulturellen Erbes bei und ermöglichen Forschenden den Zugang zu diesen Werken weltweit. Digitalisierte Dokumente werden immer öfter nicht nur als Bilder zum Lesen angeboten, sondern zusätzlich durch OCR (optical character recognition) mit dem erkannten Volltext aufgewertet. Dies erlaubt eine Suche nach Begriffen im gesamten Inhalt sowie weitere Analysemöglichkeiten. Während bei der Suche gewisse Fehlerraten toleriert und eine mäßige Layouterfassung akzeptiert werden kann, ist beispielsweise für die Extraktion von Forschungsdaten aus Digitalisaten eine hohe Zeichengenauigkeit und eine eindeutige Erfassung des Seitenaufbaus zwingend.

Gleich mehrere Projekte der Universitätsbibliothek Mannheim beschäftigen sich derzeit mit dieser Thematik. Zwei dieser Projekte werden näher vorgestellt.

Die automatische Texterkennung (OCR) mit Hilfe von Tesseract ist das Ziel des DFG-geförderten Projektes Tesseract als Komponente im OCR-D-Workflow, eines von acht Modulprojekten der DFG-Initiative OCR-D zur Verbesserung der Texterkennung historischer Drucke aus dem 16. bis 19. Jahrhundert. Hier wird die freie OCR-Software Tesseract praxistauglicher und anwendungsfreundlicher gemacht, z. B. durch Dokumentation, Korrektur von Softwarefehlern oder durch verbesserte Performance.

Schwerpunkte im DFG-Projekt Aktienführer-Datenarchiv sind Zeichengenauigkeit und Strukturerfassung zur Extraktion von Forschungsdaten. Der jährlich von 1956 bis 1999 im Buchformat erschienene Aktienführer beinhaltet Daten von börsennotierten Firmen und dient als Referenzwerk für die Forschung in den Bereichen BWL und VWL. Die zusammengestellten Firmenprofile enthalten allgemeine sowie geschäftsjahrspezifische Informationen, unterteilt in einzelne Abschnitte und Tabellen Ein Ziel des Projektes ist die automatisierte Verarbeitung der digitalisierten Daten und Speicherung in strukturierter Form in einer Datenbank. Dafür wurden gescannte Seiten vorverarbeitet, die OCR-Ergebnisse durch die Kombination mehrerer OCR-Ausgaben verbessert, eine automatisierte Strukturerfassung inklusive Tabellen entwickelt und die Daten extrahiert.

Abschließend wird noch ein Ausblick auf zukünftige Projekte mit verwandter Thematik gegeben.

Alle vorgestellte Software ist frei nachnutzbar für eigene Projekte und wird von der Universitätsbibliothek Mannheim auf GitHub unter freien Lizenzen veröffentlicht.

Stefan Weil und Jan Kamlah: Forschungsdaten aus Digitalisaten.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 189. DOI: https://doi.org/10.11588/heibooks.598.c8429

coastDat - von Big Data zu Smart Data

Elke Meyer[1], Heinke Höck[2] und Hannes Thiemann[2]
[1]Helmholtz-Zentrum Geesthacht Zentrum für Material- und Küstenforschung GmbH;
[2] Deutsches Klimarechenzentrum GmbH;

Ozeane und Randmeere sind noch immer größtenteils unerforschte Regionen. Deren Umweltbedingungen zu erfassen und einzuordnen, ist weiterhin eine Herausforderung für die Wissenschaft und Gesellschaft. Kenntnisse über die vergangenen, derzeitigen und zukünftigen Umweltbedingungen der Meere sind von existenzieller Bedeutung für die Menschheit, da sie am und vom Meer lebt. Zum Beispiel sind Fragen zum Meeresspiegelanstieg oder Deckung des Energiebedarfes durch Offshore-Windkraftanlagen von hoher gesellschaftlicher Relevanz.

Vor ca. 20 Jahren wurde am Institut für Küstenforschung am Helmholtz-Zentrum Geesthacht damit begonnen, mit einem Regionalmodell Simulationen für die Atmosphäre über den Nordostatlantik und Europa für lange Zeiträume zu rechnen (Weisse et al., 2008). Als Antriebsdaten wurden die Ergebnisse aus den damaligen Globalmodellen verwendet, die räumlich auf einem 2° Gitter in sechsstündiger Auflösung vorlagen (Kalnay et al., 1996). Regionalmodelle haben eine höhere zeitliche und räumliche Auflösung, d.h. die Küste und Gebirge können besser aufgelöst werden. Durch die höhere zeitliche Auflösung der Modellausgabe können z.B. Sturm- und andere Wetterereignisse besser simuliert werden. Mit diesen atmosphärischen Modelldaten (Wetterdaten) werden z.B. Strömungs- und Seegangsmodelle angetrieben. Diese Modellergebnisse (coastDat) werden genutzt, um die aktuellen und die vergangenen marinen Umweltbedingungen und deren Variabilität zu bestimmen, wo keine ausreichenden Beobachtungsdaten zur Verfügung stehen (Weisse et al., 2015 & 2018). Wissenschaftliche Fragestellungen sind z.B., wie hat sich das Sturmklima in den letzten Dekaden verändert, oder wie verhält sich der Seegang für bestimmte atmosphärische Situationen.

Für diese Anwendungen werden entsprechende Hochleistungsrechner und Datenspeicherkapazitäten am Deutschen Klimarechenzentrum (DKRZ) genutzt. Neben einem Hochleistungsrechner betreibt das DKRZ das World Data Center for Climate (WDCC), ein zertifiziertes Langzeitarchiv für die Klimaforschung, in dem die coastDat-Daten archiviert sind. Das WDCC ist Teil des World Data Systems (WDS) und bietet als Service die Langzeitarchivierung, Katalogisierung, Kuration und Publikation klimarelevanter (Modell-)Daten für die internationale Nutzergemeinschaft.

Geleitet von den FAIR-Prinzipien beinhaltet der Archivierungsprozess eine Beratung für die Standardisierung und Datenlizenzvergabe. Hierbei wird vor allem das NetCDF4 Format für die Daten in Verbindung mit den CF-Konventionen (Climate Forecast conventions) für die Metadaten (`http://cfconventions.org`) empfohlen. Weiterhin sind die Qualitätsprüfung der Daten und Bereitstellung von Informationen für die interdisziplinäre Nachnutzung Teil des Prozesses. Mit der Archivierung erfolgt eine Zuweisung von DataCite DOIs, die eine eindeutige Auffindbarkeit und Zitierbarkeit ermöglichen. Verbunden

Elke Meyer, Heinke Höck und Hannes Thiemann: coastDat - von Big Data zu Smart Data.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 190–191. DOI: https://doi.org/10.11588/heibooks.598.c8430

mit der Archivierung im WDCC ist eine Publikation der Metadaten in verschiedenen weiteren Portalen, die eine noch höhere Sichtbarkeit der Daten schafft. Mit coastDat ist eine Datenbank entstanden, die Daten für diese Art von Fragestellungen zur Verfügung stellt. Mehr als 100 unterschiedliche internationale Nutzer aus Wissenschaft, Administration und Industrie haben bisher die Daten genutzt. Über die Informationen im WDCC hinaus, sind weitere Details auf coastdat.de zu finden. Die Infrastruktur vom DKRZ erlaubt, dass die großen Datenmengen von coastDat weltweit heruntergeladen werden können.

Weiterführende Informationen:

coastDat:	coastdat.de
WDCC:	`https://www.dkrz.de/up/systems/wdcc`
WDCC Portal:	`https://cera-www.dkrz.de/WDCC/ui/cerasearch/`

Short Overview:

`https://www.dkrz.de/kommunikation/pub/dm-stories/de-DM-stories-at-DKRZ_coastDat.pdf`

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K.C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph,1996:The NCEP/NCAR 40-Year Reanalysis Project.Bull. Amer. Meteor. Soc.,77,437–472,`https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2`

Weisse, R., H. v. Storch, U. Callies, A. Chrastansky, F. Feser, I. Grabemann, H. Guenther, A. Pluess, Th. Stoye, J. Tellkamp, J. Winterfeldt and K. Woth, (2009): Regional meteo-marine reanalyses and climate change projections: Results for Northern Europe and potentials for coastal and offshore applications. Bull. Amer. Meteor. Soc., 90, 849–860. doi: ?url http://dx.doi.org/10.1175/2008BAMS2713.1

Weisse, R, Bisling, P., Gaslikova, L., Geyer, B., Groll, N., Hortamani, M., Matthias, V., Maneke, M., Meinke, I., Meyer, E.M.I., Schwichtenberg, F., Stempinski, F., Wiese, F. and Wöckner-Kluwe, K. (2015) 2:3: Climate services for marine applications in Europe, Earth Perspectives doi:`http://dx.doi.org/10.1186/s40322-015-0029-0`

Weisse, R., Gaslikova, L., Geyer, B. Groll, N. und Meyer, E (2018): coastDat: Modelldaten für Wissenschaft und Industrie, Die Küste, 86, 5-1

Nachhaltige Infrastruktur zur Integration von Forschungssoftware in Forschungsdatenrepositorien

Anett Seeland[1], Timo Koch[2], Sibylle Hermann[3] und Bernd Flemisch[2]
[1]Technische Informations- und Kommunikationsdienste (TIK), Universität Stuttgart;
[2]Institut für Wasser- und Umweltsystemmodellierung (IWS), Universität Stuttgart;
[3]Universitätsbibliothek, Universität Stuttgart

Komplexe Forschungssoftware hat häufig hohe Einstiegshürden: Bei der Installation sind Kenntnisse des Betriebssystems nötig, außerdem erfordern oft selbst einfache Änderungen, z.B. von Parametern, Programmierkenntnisse. Architekturabhängige Installationsskripte, die oft auch noch von speziellen Paketversionen ausgehen, erschweren die Reproduzierbarkeit von Forschungsergebnissen. Das Projekt „Sustainable infrastructure for the improved usability and archivability of research software on the example of the porous-media-simulator DuMux" (SusI) hat zum Ziel, die Benutzbarkeit der freien Forschungssoftware DuMux (`https://dumux.org/`) zu erhöhen. Zielgruppe sind dabei sowohl Wissenschaftlerinnen und Wissenschaftler, die zum ersten Mal mit der Software arbeiten, als auch erfahrene Entwickler. Darüber hinaus soll die Archivierbarkeit von Softwareanwendungen mit Verlinkung zu einer Publikation ermöglicht werden. Dabei besteht der nachhaltige Lösungsansatz darin, bereits vorhandene Services der Universität (ViPLab, Datenrepositorium) so zu erweitern und zu vernetzen, dass eine von der verwendeten Forschungssoftware unabhängige Infrastrukturlösung geschaffen wird.

ViPLab (`https://www.tik.uni-stuttgart.de/forschung/projekte/vip`) ist eine virtuelle Programmierumgebung mit JavaScript-basiertem Editor zur Programmierung und Entwicklung. Der im Browser geschriebene Code wird dabei zu einem Server geschickt, dort ausgeführt und Ergebnisse (Text, 2D- und 3D-Grafiken) zurückgesandt. Bisher vor allem in der Lehre eingesetzt, hätte ViPLab für Forschungssoftware den Vorteil, dass Forschende die Software, zunächst ohne sie lokal installieren zu müssen, verwenden und konfigurieren können. Die Installation auf dem Server kann durch Container-Virtualisierung (z.B. mit Docker) robust realisiert werden. Verallgemeinert sollen aus den Erfahrungen mit DuMux Richtlinien zum Containerisieren von Software entwickelt werden, damit diese in ViPLab automatisiert ausgeführt werden kann.

Zentraler Interaktionsort, um aus Forschungsdaten Wissen zu gewinnen, ist ein Repositorium. Hier kann nach Daten gesucht sowie ihre Verarbeitung nachvollzogen und verstanden werden. Deshalb ist es naheliegend, auch Forschungssoftware in ein Repositorium zu integrieren. Dies ist in Form eines ViPLab-Plugins für die Repositoriensoftware Dataverse geplant. Das Webfrontend für ViPLab wird dabei anhand der Bedürfnisse der Forschenden erweitert. So soll es neben dem Editor weitere durch den Forscher spezifizierbare GUI-Elemente geben, die es gezielt erlauben die Komplexität der Interaktionsmöglichkeiten zu

Anett Seeland et al.: Nachhaltige Infrastruktur zur Integration von Forschungssoftware in Forschungsdatenrepositorien.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 192–193. DOI: https://doi.org/10.11588/heibooks.598.c8431

reduzieren (von Code, über Inputfelder für bestimmte Parameter bishin zu Auswahllisten und Checkboxen). Dies ermöglicht außerdem die Konfiguration einfacher GUIs für Forschungssoftware die selbst keine GUI bereitstellt. Durch das Repositorium soll dabei die dauerhafte Verfügbarkeit und eindeutige Zitation – auch von verschiedenen Versionen der Software – sichergestellt werden. Die Beschreibung der veröffentlichten Software im Repositorium wird mit geeigneten Metadaten unterstützt. Diese sollen teilweise automatisiert aus der Software extrahiert und von den Softwareentwicklern ergänzt werden.

Die geplanten Maßnahmen gehen in ihrer Gesamtheit über bestehende Projekte hinaus. So eignen sich Jupyter-Notebooks (`http://jupyter.org`) eher für skriptbasierte Software und weniger für komplexe objektorientierte Programme. Andererseits berücksichtigen Archivierungslösungen (wie z.B. bwFLA) nicht vergleichbar die Interaktion der Software in der Nachnutzung. Zusammenfassend gewinnt die Forschungssoftware DuMux, auch durch erleichterten Zugang, an Benutzbarkeit. Gleichzeitig wird die Reproduzierbarkeit von Forschungsdaten erleichtert und nachhaltig gewährleistet.

re3data - Advancing Services for Open Science

Robert Ulrich[1], Heinz Pampel[2], Maxi Kindling[3], Paul Vierkant[2], Frank Scholze[1], Michael Witt[4], Martin Fenner[5], Kirsten Elger[6] and Gabriele Kloska[1]

[1]Karlsruhe Institute of Technology (KIT);
[2]Helmholtz Association ;
[3]Humboldt-Universität zu Berlin;
[4]Purdue University, United States;
[5]DataCite e.V., Germany;
[6]GFZ German Research Centre for Geosciences;

re3data is the global registry for research data repositories [2]. With January 2019 the service lists over 2250 digital repositories and provides an extensive description based on a detailed metadata schema [3]. A variety of funders, publishers and scientific organizations around the world refer to re3data within their guidelines and policies, recommending the service to researchers looking for appropriate repositories for storage and search of research data.

Starting with an introduction and overview to re3data and its current status under the auspices of DataCite, the talk will outline the recent and upcoming development in a heterogeneous and highly dynamic research data infrastructure landscape. The diverse requirements of the institutional stakeholders as well as the scientific communities impose demanding challenges on the architecture, networking with other services and technical implementation. The presentation will illustrate that with recent examples, like the integration and reuse of re3data in the American Geophysical Union's (AGU) 'Repository Finder'[1] , landscape analysis of data repositories for the Swiss National Science Foundation (SNSF) and a planned cooperation with B2FIND, RADAR and GeRDI on the subject classification. Fostering Open Science and FAIR data, the talk will close with a prospect on the planned next steps towards an open and linked data service matching the demands of researchers and organization.

Literaturverzeichnis

[1] Dasler, R. (2018). Data sharing made easier: use Repository Finder to find the right repository for your data. DataCite Blog, 19.12.2018. DOI:https://doi.org/10.5438/wday-8958.

[2] Pampel, H. et al. (2013). Making Research Data Repositories Visible: The re3data.org Registry. PLOS ONE, 8(11), e78080. DOI: https://doi.org/10.1371/journal.pone.0078080.

Robert Ulrich et al.: re3data - Advancing Services for Open Science.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 194–195. DOI: https://doi.org/10.11588/heibooks.598.c8432

[3] Rücknagel, J. et al. (2015). Metadata Schema for the Description of Research Data Repositories. Version 3.0. DOI: https://doi.org/10.2312/re3.008.

[4] Witt M. et al. (2019) Connecting Researchers to Data Repositories in the Earth, Space, and Environmental Sciences. In: Manghi P., Candela L., Silvello G. (eds) Digital Libraries: Supporting Open Science. IRCDL 2019. Communications in Computer and Information Science, vol 988. Springer, Cham DOI: https://doi.org/10.1007/978-3-030-11226-4_7

Teil II.

Artikel zu den Postern

V-FOR-WaTer – the virtual research environment to discover and analyse environmental data

Jörg Meyer[1], Elnaz Azmi[1], Sibylle K. Hassler[2], Mirko Mälicke[2], Marcus Strobl[1] and Erwin Zehe[2]

[1]Steinbuch Centre for Computing, Karlsruhe Institute of Technology;
[2]Institute for River and Basin Management, Karlsruhe Institute of Technology

1. Introduction

The extent and diversity of environmental data continuously increase due to more and new sensors with higher spatial and temporal resolution and due to the growth and automation of observational networks. The observed data form the basis for a better understanding of ecological systems either by data driven methods or by comparisons of the data with model predictions. However, a considerable amount of these data are difficult to access or even still stored on local data storage devices making it difficult or even impossible to find, access and re-use the data. In addition the data lack a proper metadata description required for an interoperable analysis. This results in very time consuming preparation and pre-processing of data, especially when datasets from different sources are combined.

2. Objectives

The main objectives of V-FOR-WaTer are to simplify data access for environmental sciences, foster data publications, and facilitate preparations of data and their analyses with a comprehensive toolbox allowing pre-processing of data from diverse sources. Also, bringing data and tools together in a single shared environment maximises the reproducibility of analyses and models. V-FOR-WaTer evolves towards a community data repository obeying the FAIR principles. Data owners may grant individual access with the fine-grained access management.

3. Web Portal

The web portal is still under development. The open source project written in the Python web framework Django is published at [1]. The main tool to browse and select data is the map which integrates features known from geographic information systems (GIS). In addition a filter menu allows to refine the selection based on various metadata fields

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026733` veröffentlicht.

and to add the data to the *datastore* for further processing. Users lacking access rights for certain datasets may send a request to the data owner. The identity management is delegated to the external tool B2ACCESS [2], i.e. the V-FOR-WaTer web portal reuses existing federated accounts of researchers.

4. Data and Tools

he incorporated datasets include heterogeneous data from the Catchments as Organized Systems (CAOS) project [3] that has not yet been publicly available, data from university projects, and from state offices. The metadata scheme of V-FOR-WaTer was designed to describe the heterogeneous datasets. The focus of the schema is flexibility and compatibility with existing international standards (INSPIRE, ISO19115). V-FOR-WaTer realizes online tools as Web Processing Services (WPS) [4] -– a standard of the Open Geospatial Consortium. A WPS server provides a RESTful interface to query the description of a WPS and to trigger the execution of a WPS. Users interact with the WPS server only indirectly via the V-FOR-WaTer web portal. With the already implemented toolbox users may run first steps of a geostatistical analysis. An example is a simple variogram analysis to determine the spatial dependencies of distributed measurements. Another example is the visualization of a flow duration curve. In order to run complex workflows consisting of several WPS a graphical workflow editor will be integrated in the portal.

5. Conclusion

The virtual research environment V-FOR-WaTer addresses challenges of environmental scientists in terms of data discovery, data management, and data analysis. The V-FOR-WaTer portal offers features of a data repository, a geographic information system, and online tools for repro- ducible data analysis.

Acknowledgements

V-FOR-WaTer has been funded by the ministry of science, research and arts of the state of Baden-Wuerttemberg, Germany.

Bibliography

[1] V-FOR-WaTer github project https://github.com/VForWaTer/vforwater-portalRe-trieved06-05-2019

[2] TB2ACCESS Service https://eudat.eu/services/Retrieved06-05-2019.

[3] E. Zehe, U. Ehret, L. Pfister, T. Blume, B. Schröder, M. Westhoff, C. Jackisch, S. J. Schy- manski, M. Weiler, K. Schulz, N. Allroggen, J. Tronicke, L. van Schaik, P. Dietrich, U. Scherer, J. Eccard, V. Wulfmeyer, and A. Kleidon. 2014. "HESS Opinions: From response units to functional units: a thermodynamic reinterpretation of the HRU concept to link spatial or- ganization and functioning of intermediate scale catchments." Hydrology and Earth System Sciences 18, 11 (2014), 4635–4655.https://doi.org/10.5194/hess-18-4635-2014

[4] Web Processing Service https://www.opengeospatial.org/standards/wpsRetrieved06-05-2019

Das Konzept für ein FDM-Kompetenznetzwerk an der Universität zu Köln

Monika Linne[1], Constanze Curdt[2], Jens Dierkes[1] und Sonja Kloppenburg[3]
[1]Universitäts- und Stadtbibliothek Köln, Universität zu Köln;
[2]Regionales Rechenzentrum Köln, Universität zu Köln;
[3]Dezernat Forschungsmanagement, Universität zu Köln

Forschungsdatenmanagement (FDM) hat sich innerhalb der letzten Jahre zunehmend als integraler Bestandteil von Forschungsprozessen etabliert. Durch diesen Umstand stellt es eine gemeinschaftliche Herausforderung in einem komplexen sozio-technologischen Umfeld dar. Innerhalb deutscher Hochschulen sind an der Umsetzung von übergreifenden FDM-Maßnahmen beispielsweise Informationsinfrastrukturanbieter (Rechenzentrum, Bibliothek), Universitätsleitungen, FDM-Stabstellen, Abteilungen Forschungsmanagement, Forschungsförderung und FDM- Expert*innen von Fakultäten beteiligt. Hierbei trägt eine koordinierte Vernetzung auf dem Campus zur Bildung einer lokalen FDM Gemeinschaft bei und hilft Best Practices in den Forschungsalltag zu integrieren. Neben der Vernetzung bleibt die entscheidende Frage, welches Kooperationsmodell und welcher Grad der Zentralisierung mit vielen Stakeholdern gewählt werden sollte. Hierbei sind deutliche Unterschiede zwischen den einzelnen Universitäten festzustellen.

Das Cologne Competence Center for Research Data Management (C^3RDM) wurde im Jahr 2018 an der Universität zu Köln (UzK) durch die drei Einrichtungen der Universitäts- und Stadtbibliothek (USB), dem Regionalen Rechenzentrum (RRZK) und dem Dezernat Forschungsmanagement (D7) gegründet. Bei diesem Bündnis handelt es sich um eine natürliche Kooperation, welche die Kuratierung und das Management von Wissen durch die USB und D7 auf der einen Seite und die technische Expertise des RRZK auf der anderen bündelt. Das Kompetenzzentrum hat im Wesentlichen zwei übergeordnete Ziele im Fokus seiner Handlungsfelder:

1. Der Aufbau von umfangreichen FDM-Unterstützungsangeboten und eine individuelle Beratung in allen Phasen des Forschungsvorhabens.

2. Der Aufbau eines FDM-Expertise-Netzwerkes und ein Forum zur Vernetzung der gewachsenen FDM-Strukturen bzw. verschiedener Stakeholder an der UzK.

Insbesondere das zweite Ziel einer stärkeren Vernetzung der FDM-Akteur*innen auf dem Campus erscheint notwendig, da FDM an vielen Stellen der Universität in einer nur sehr lose gekoppelten Art und Weise stattfindet, bspw. in Forschungsprojekten, in Instituten,

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026848` veröffentlicht.

Monika Linne et al.: Das Konzept für ein FDM-Kompetenznetzwerk an der Universität zu Köln.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 202–204. DOI: https://doi.org/10.11588/heibooks.598.c8435

an Fakultäten usw. So sind dezentrale FDM-Aktivitäten sowohl auf Fakultätsebne (z.B. am Data Center for the Humanities der Philosophischen Fakultät oder am Cologne Center for eHumanities) als auch innerhalb der Forschungskonsortien (z.B. in den Exzellenzclustern oder innerhalb der Sonderforschungsbereiche) festzustellen. Der im C^3RDM verfolgte Ansatz geht initial von den zentralen Infrastruktureinrichtungen D7, USB und RRZK aus. Diese drei Einrichtungen kooperieren im C^3RDM und konzipieren gemeinsam ein strukturelles Modell für ein universitäts- weites FDM, welches sowohl Policies, Entwicklungskonzepte, als auch ein Service-Portfolio zum FDM umfasst. Die bereits vorhandenen und noch aufzubauenden Services auf Seiten des C^3RDM sollen möglichst generisch sein und in die Breite wirken. Beim Aus- und Aufbau der Dienste für den Campus orientiert sich die Rollenverteilung zwischen den drei Institutionen anhand der Expertise der Einrichtungen. Beispielsweise erfolgt die technische Implementierung eine DMP-Tools im RRZK, während die inhaltliche Ausgestaltung des Fragenkataloges wesentlich in D7 passiert. Für den weiteren Aufbau ist es wichtig, eine gezielte Vernetzung zwischen allen Stakeholdern voranzutreiben, um fachspezifische Besonderheiten und lokale Strukturen zu berücksichtigen. Zur systematischen Förderung des fakultätsübergreifenden Erfahrungsaustausches zwischen den FDM-Agierenden gibt das C^3RDM zukünftig Impulse, um Synergien zu identifizieren und nachhaltige FDM-Strukturen aufzubauen. Dies findet u. a. im Rahmen von Informations- und Vernetzungsveranstaltungen statt, wie z.B. der Veranstaltung zur Nationalen Forschungsinfrastruktur (NFDI) im Juli 2019. Hier konnten sich Forschende der verschiedenen Fakultäten über ihre geplanten Beiträge zur NFDI gegenseitig informieren und Querschnittsthemen identifizieren.

Ausgehend von einem generischen FDM-Service-Portfolio [1] handelt es sich um ein initiales Konzept für die Vernetzung an der UzK und der Sammlung erster Erfahrungen bei deren Umsetzung. Bei der Größe des Standortes und der Komplexität der Aufgabe ist eine wichtige Charakteristik in diesem Zusammenhang der experimentelle Charakter bzw. die iterative Erprobung und Anpassung an spezifische Bedingungen (lokal oder fachspezifisch). Die Entwicklung eines mehrteiligen Maßnahmenpaketes fand deshalb unter Berücksichtigung dieser lokalen und fachspezifischen Charakteristika einerseits und unter Einbezug von konkreten Anforderungen der Forschenden andererseits statt. Hierbei wurden sowohl Basisdienstleistungen als auch weiterführende, (fach-)spezifischere Dienstleistungen berücksichtigt, wobei die Einführung der Basisdienste zu Beginn des Projektes im Vordergrund stehen sollte. Alle Maßnahmen haben das übergeordnete Ziel, die Ressourcen und Kompetenzen zum FDM innerhalb der UzK zu bündeln und die Forschenden im Kontext von FDM-Belangen zu unterstützen.

Unter Berücksichtigung der vorhandenen Expertise der drei beteiligten Einrichtungen und infrastrukturellen Vorbedingungen an der Universität soll zunächst eine Sensibilisierung und eine Awarenessschaffung auf dem Kölner Campus hinsichtlich FDM betrieben werden. In diesem Zusammenhang sollen Beratungs-, Schulungs- und Informationsangebote für mehrere Zielgruppen entwickelt werden. Im Rahmen der individuellen Beratung erhalten die Forschenden an der UzK Unterstützung für sämtliche Schritte entlang des gesamten Forschungszyklus, da es in allen Phasen eines Forschungsprojektes Beratungsbedarf zum FDM gibt. Von besonderer Relevanz sind diesbezüglich die Beratung innerhalb der Projektplanung und Unterstützung bei der Antragstellung von Drittmittelprojekten,

aber auch die Erhebung, Auswertung, Archivierung und gegebenenfalls die Veröffentlichung von Forschungsdaten. Das Kompetenzzentrum soll hierbei als erste und zentrale Anlaufstelle dienen. Die fachspezifischen Anforderungen zum FDM sollen – wenn möglich – über das sich im Aufbau befindliche Expert*innennetzwerk aufgefangen werden. Optimalerweise können in solchen Fällen fachspezifische Anfragen, zu denen im C^3RDM die notwendige Expertise nicht (vollständig) vorhanden ist, Informationen aus den Infrastrukturen und/oder den Fachbereichen eingeholt werden. Ergänzend zur persönlichen Beratung wird darüber hinaus ein webbasiertes Informationsportal aufgebaut. Interessierte Forschende sollen auf dem Webportal des C^3RDM vielfältige, modular aufgebaute Informationsangebote (u.a. in Form von Tutorials oder Webinars) zu allen Aspekten des FDM finden können, welche sie zielgerichtet und just-in-time bei ihren spezifischen Fragestellungen unterstützen. Darüber hinaus wird zum Zwecke von Schulungs- und Informationsveranstaltungen ein umfassendes Programm entwickelt, welches sich an die breite Zielgruppe von jungen Forschenden, Promovierenden, Studierenden vor Abschlussarbeiten, aber auch an alle anderen Forschenden wendet.

Nicht zuletzt wird das C^3RDM System-Dienste in Form einer digitalen Infrastruktur entwickeln, welche z.B. die Bereitstellung und Verwaltung von Speicher für verschiedene Domänen beinhaltet oder auch die Bereitstellung und Betreuung von kooperativen Arbeitsplattformen für Verbundprojekte. Hierzu gehören auch die Archivierung von Forschungsdaten und der Aufbau einer zentralen Nachweisdatenbank für Forschungsdaten. Zusammenfassend kann festgehalten werden, dass institutionelles FDM an der Universität zu Köln eine gemeinschaftliche Herausforderung in einem komplexen sozio-technologischen Umfeld mit vielen Stakeholdern ist, welche unter Berücksichtigung und Anerkennung gewachsener, bestehender FDM-Strukturen der UzK (zentral und dezentral) stattfinden muss. Neben generischen Diensten sind die Vernetzung und Etablierung von Kooperationen mit den FDM-Aktivitäten in den einzelnen Fachbereichen zentrale Anliegen des C^3RDM. Hierbei wird eine dialogische Vorgehensweise zusammen mit Pilotprojekten verfolgt. Die Herangehensweise in Form einer Netzwerk-/Kooperationsstruktur ist der erste Schritt zur Herstellung eines breiteren Konsenses über die Anforderungen an eine nachhaltige FDM-Infrastruktur an der Universität zu Köln.

Literaturverzeichnis

[1] Jens Dierkes, Constanze Curdt: Von der Idee zum Konzept - Forschungsdatenmanagement an der Universität zu Köln,o-bib, 2018, 5(2), 28, https://doi.org/10.5282/o-bib/2018H2S28-46

SERVICEVERZEICHNIS FORSCHUNGSDATEN

Judith Erven[1], Jens Dierkes[1], Alvaro Aguilera[2], Ortrun Brand[3], Jens Ludwig[4], Ralph Müller-Pfefferkorn[2], Paul Schubert[2] und Paul Sutter[3]

[1]Universitäts- und Stadtbibliothek Köln, Universität zu Köln;
[2]Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH), Technische Universität Dresden;
[3]Stabsstelle Forschungsdatenmanagement, Philipps-Universität Marburg;
[4]Stiftung Preußischer Kulturbesitz

Der Bereich Forschungsdatenmanagement (FDM) entwickelt sich rasant. FDM hat an vielen Universitäten und außeruniversitären Forschungseinrichtungen strategische Bedeutung erlangt. Standorte müssen nicht von Grund auf neu anfangen, um Dienstleistungen zum FDM anzubieten, sondern es besteht das Potenzial von Synergien, zumindest aber von gegenseitigen Lernen, wenn bekannt ist, wo welche Dienste angeboten werden und spezifische Erfahrungen vorhanden sind. So ist es möglich, besser als bisher voneinander zu lernen und auf Diensten und Expertise von anderen aufzubauen. Gerade im Kontext des Aufbaus der Nationalen Forschungsdateninfrastruktur in Deutschland als einem kooperativen Verbund von Diensten und Expertinnen und Experten quer durch alle Forschungsdisziplinen, ist eine zentrale Informationsstruktur zur Vernetzung essentiell.
Das Projekt „Serviceverzeichnis Forschungsdaten" (SVF) ist 2016 aus dem 6. DINI/nestor-Workshop „Kooperationstreffen Forschungsdaten" in Göttingen hervorgegangen[1]. Hier hat sich ein großer Informationsbedarf über vorhandene nationale FDM-Angebote/Initiativen und vorhandene Expertise gezeigt. Die Arbeitsgruppe des SVF hat sich das Ziel gesetzt, eine systematische Sammlung von Diensten und Expertinnen und Experten beziehungsweise Ansprechpartnerinnen und Ansprechpartnern in dem Bereich Management von Forschungsdaten zu erstellen um Synergieeffekte zu nutzen und aus den Erfahrungen von anderen zu lernen. Diese Informationen sollen offen und interaktiv über eine Webplattform präsentiert werden. Primäre Zielgruppe sind alle diejenigen, die sich mit dem Aufbau von Informationsinfrastruktur beschäftigen, aber auch Forschende, die Dienste nutzen möchten oder Expertise zu bestimmten Fragestellungen im Datenmanagement suchen. Das SVF steht aber auch allen anderen Interessierten offen[2].

Die Inhalte sollen sowohl durch die Community der FDM-Expertinnen und Experten als auch durch die Dienstanbieter selber gepflegt werden. Die Projektgruppe hat hierfür ein Datenmodell und eine Plattform prototypisch entwickelt, die wir in unserem Beitrag

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI https://doi.org/10.11588/heidok.00026850 veröffentlicht.

[1] https://www.forschungsdaten.org/index.php/DINI-nestor-WS6

[2] Das Poster zu diesem Artikel mit Abbildungen und Screenshots des SVF ist zu finden unter https://doi.org/10.11588/heidok.00026850

vorstellen möchten. Der Prototyp basiert auf dem Web-CMS Contao, welches um eigene Module erweitert wurde und auf einem Apache Webserver läuft. Die Suche läuft über eine MariaDB-Datenbank. Gehostet wird der Prototyp beim Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) der TU Dresden. Der Prototyp ist unter folgendem Link zu erreichen:`https://serviceverzeichnis-forschungsdaten.org/`. Das Datenmodell besteht aus drei zentralen Klassen, die einen Service beschreiben: Service. Anbietende und Kontakt. Die drei Klassen stehen jeweils in 1-zu-n-Beziehungen zueinander. Ein Service hat genau einen Anbietenden, Anbietender und Service haben genau einen Kontakt. Ein Kontakt wiederum kann mehreren Services zugeordnet sein, ebenso wie ein Anbietender mehrere Services anbieten kann. Einem Kontakt können wiederum mehrere Anbieter zugeordnet sein. Die Services werden anhand verschiedener Informationen beschrieben, die über ein Formular abgefragt werden. Das Eingabeformular für einen neuen Dienst enthält folgende Felder: Name, Anbieter, Kontakt, Fachdisziplin nach DFG, Zielgruppe, Datentyp, Service-Art, Laufzeit, und URL. Das Freitextfeld Beschreibung ermöglicht die freie Beschreibung des Service inklusive der Ergänzung von Logos, weiteren Links oder ähnlichem. Der Prototyp für das SVF wurde anhand von User-Stories entwickelt, die die unterschiedlichen Anforderungen an den Prototypen beschreiben. Um die User Stories zu erstellen wurden folgende Rollen definiert, die für den Prototypen relevant sind: Informationssuchende, Informationsanbietende, Redakteur*in und technische Betreibende. Beispiele für die erstellten User Stories sind:

- Informationssuchende: „Als Informationssucher möchte ich Trefferlisten nach unterschiedlichen Kriterien filtern, um die Liste systematisch zu untersuchen."
- Informationsanbietende: „Als Informationsanbietende möchte ich neue Einträge einfügen, damit ich und mein Service von der Community gefunden werden."
- Technischer Betreiber: „Als technischer Betreiber des Angebots möchte ich sicherstellen, dass der Dienst auch in den nächsten Jahren noch Bestand hat, damit die Nachhaltigkeit des Angebots gesichert ist."

Ein schlankes Redaktionskonzept soll die Qualität des SVF sicherstellen. Neue Einträge im SVF durchlaufen den Redaktionsprozess, der sich wie folgt gestaltet: Neue Einträge sind zunächst standardmäßig freigeschaltet. Das Redaktionsteam wird bei Änderungen und Neueinträgen benachrichtigt und prüft, ob es sich um Spam handelt. Alle Anbieter werden regelmäßig per E-Mail daran erinnert, ihre Einträge zu überprüfen und gegebenenfalls Kontakte oder andere Parameter zu aktualisieren. Ein bestehender Kontakt kann seinem Service einen neuen Kontakt zuweisen. So soll sichergestellt werden, dass die Einträge auch beim Wechsel des Kontakts Bestand haben. Das Redaktionsteam kann Services ebenfalls einen neuen Kontakt zuweisen. Kann kein neuer Kontakt ermittelt werden, wird der Eintrag gesperrt.

In zwei Anwender-Tests mit konkreten, praxisnahen Aufgabenstellungen und insgesamt elf Teilnehmer*innen aus der FDM-Community sollte das SVF auf Praxistauglichkeit getestet werden und ein Feedback der FDM-Community eingeholt werden, um herauszufinden, ob der Prototyp die Informationen abfragt und liefert, die gesucht und erwartet

werden. Der erste Anwender-Test inklusive anschließender Diskussionsrunde fand im September 2018, der zweite Anwender-Test im März und April 2019 statt. Die zu lösenden Aufgaben beschäftigten sich mit der Startseite, der Suche, dem Eingabeformular für einen neuen Dienst sowie der Ergebnisseite. Die Rückmeldungen der Testenden lieferten wertvolle Hinweise auf Stärken und Schwächen des Prototyps. Im Anschluss an die Tests wurden die Ergebnisse jeweils in der Projektgruppe ausgewertet, diskutiert und in neue Anforderungen umgewandelt. So führte der erste Anwender-Test dazu, dass das Eingabeformular für einen neuen Dienst verschlankt wurde und die Bearbeitungsmöglichkeiten für einen bestehenden Dienst vereinfacht wurden. Die Such- und Filtermöglichkeiten der Suchergebnisse wurden optimiert. Derzeit wird der zweite Anwender-Test ausgewertet. Sobald die Auswertung abgeschlossen und die daraus resultierenden und priorisierten Anforderungen umgesetzt sind, steht der Go-live des SVF unter der oben genannten URL an.

Abschließend lässt sich festhalten, dass sowohl der 6. DINI/nestor-Workshop „Kooperationstreffen Forschungsdaten" im Jahr 2016 als auch die im Zuge der Entwicklung des Prototyps durchgeführten Anwender-Test den Bedarf und den Wunsch nach Vernetzung im Bereich Forschungsdatenmanagement ergaben. Die Vernetzung kann jedoch nur gelingen, wenn Angebot wie das Serviceverzeichnis Forschungsdaten von der Community angenommen, also aktiv genutzt und mit Inhalten befüllt werden. Ob dies gelingt, werden die nächsten Wochen und Monate zeigen.

SARA: Open Source Projekt zur langfristigen Verfügbarkeit und Zitierbarkeit von Software

Franziska Rapp[1], Daniel Scharon[2], Matthias Fratz[2], Stefan Kombrink[1], Volodymyr Kushnarenko[1], Pia Schmücker[1], Marcel Waldvogel[2] und Stefan Wesner[1]

[1] Universität Ulm;
[2] Universität Konstanz

1. Motivation

Softwareartefakte (von kleinen Skripten bis hin zu komplexen Software-Frameworks) spielen im wissenschaftlichen Forschungsprozess eine tragende Rolle. Sie werden z.B. zur Erhebung, Verarbeitung, Visualisierung oder Modellierung von Forschungsdaten eingesetzt, sollten langfristig verfügbar sein und nach den FAIR Prinzipien[1] behandelt werden. Das Projekt SARA zielt darauf ab, Softwareartefakte langfristig verfügbar [1] und referenzierbar [2] zu machen. Dabei richtet sich SARA disziplinübergreifend an sämtliche Forschende, die Software selbst entwickeln oder bestehende Software nachnutzen und weiterentwickeln.

2. Der SARA Webservice

Im Bereich der Softwareentwicklung hat sich das Versionierungssystem Git als de-facto-Standard etabliert. Zur Veröffentlichung von Daten und Software sind Repositorien bewährte, nachhaltige Infrastrukturen. Dabei handelt es sich um Publikationsplattformen, die zentral für Mitglieder und Angehörige einer wissenschaftlichen Einrichtung oder einer Fachcommunity angeboten werden. Sie verfügen in der Regel über eine Landing Page mit beschreibenden Metadaten wie Titel, Autoren, Abstract, Lizenz etc. Zur eindeutigen Referenzierbarkeit und Zitierbarkeit wird ein persistenter Identifier, wie z.B. ein Digital Object Identifier (DOI), vergeben. Die Auffindbarkeit von in Repositorien beschriebenen Objekten wird durch die Indexierung in Suchmaschinen und -portalen wie Google und OpenAIRE[2]erhöht.

Im Projekt wurde ein Webservice entwickelt, der diese bestehenden Infrastrukturen verbindet. Der SARA Webservice extrahiert Metadaten und Dateien aus Projekten in GitHub und angebundenen GitLab-Installationen. Über eine intuitive graphische Oberfläche können Forschende den Umfang der extrahierten Dateien nach ihren Vorstellungen

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI https://doi.org/10.11588/heidok.00026854 veröffentlicht.

[1] https://www.go-fair.org/fair-principles/,zuletztaufgerufenam23.04.2019

[2] https://www.openaire.eu/,zuletztaufgerufenam23.04.2019

Franziska Rapp et al.: SARA: Open Source Projekt zur langfristigen Verfügbarkeit und Zitierbarkeit von Software. In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge, Heidelberg: heiBOOKS, 2020, 208–210. DOI: https://doi.org/10.11588/heibooks.598.c8437

festlegen, indem sie z.B. nur bestimmte Branches für die Archivierung auswählen. Die Entwicklungshistorie kann komplett oder in gekürzter Version mitarchiviert werden. Für die Nachnutzung der Software wird vom Forschenden eine Lizenz festgelegt. Ist bereits eine LICENSE Datei vorhanden, werden die gängigen Lizenzen von SARA automatisch erkannt. Die Dateien werden anschließend in ein Git-Archiv (read-only) überführt, das die langfristige Verfügbarkeit nach den Empfehlungen der DFG sicherstellt. Die aus den Projekten extrahierten Metadaten können durch den Forschenden editiert und ergänzt werden, bevor sie an das Repositorium übergeben werden. Im SARA Projekt wurden als Repositorien exemplarisch DSpace-Instanzen angebunden. Die konkrete Anbindung ist hierbei flexibel konfigurierbar und passt sich den Anforderungen von Repositorienbetreibern an [3]. Über die Registrierung einer DOI werden die archivierten Softwareartefakte referenzierbar und zitierbar und durch die Indexierung in Suchmaschinen wird schließlich die Auffindbarkeit und Sichtbarkeit erhöht.

3. Vorteile

3.1. Wissenschaftler

- Wissenschaftler können den Empfehlungen der DFG zur Aufbewahrung der Softwareartefakte für mind. 10 Jahre nachkommen (gute wissenschaftliche Praxis)
- Zitierbarkeit der eigenen Arbeit, bessere Auffindbarkeit
- Nachnutzung von Softwareartefakten
- Zugriffsgeschützte Archivierung möglich (dark archive)

3.2 Infrastrukturanbieter

- Individueller oder kooperativer Betrieb
- Git mit dem eigenen Repositorium verbindbar
- Strategische Erweiterung der eigenen FDM-Infrastruktur

4. Nachnutzung

Der Quellcode des SARA Webservice ist auf GitHub frei verfügbar[3]. Es sind verschiedene Einsatzszenarien denkbar. SARA kann beispielsweise an einer Einrichtung betrieben und mit den dortigen lokalen GitLab-Installationen, GitHub und dem institutionellen Repositorium verbunden werden. Auch ein kooperativer Einsatz ist denkbar, bei dem eine Einrichtung den SARA Webservice hostet und GitLab-Installationen verschiedener Einrichtungen sowie deren jeweilige institutionelle Repositorien angebunden sind. Dasselbe Szenario ist für Fachcommunities statt für einzelne Einrichtungen denkbar, was SARA

[3] `https://github.com/sara-service,zuletztaufgerufenam16.04.2019`

auch in den Kontext der sich bildenden Konsortien der Nationalen Forschungsdateninfrastruktur[4] setzt.

Literaturverzeichnis

[1] Deutsche Forschungsgemeinschaft. Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“. Ergänzte Auflage. Weinheim: Wiley-VHC, 2013.https://doi.org/10.1002/9783527679188.oth1.

[2] Smith, Arfon M., Daniel S. Katz, Kyle E. Niemeyer und FORCE11 Software Citation Working Group. „Software Citation Principles.“ PeerJ Computer Science 2 (2016): e86. https://doi.org/10.7717/peerj-cs.86.

[3] Rapp, Franziska; Kombrink, Stefan; Kushnarenko, Volodymyr; Fratz, Matthias; Scharon, Daniel (2018): SSARA-Dienst: Software langfristig verfügbar machen", o-bib - Das offene Bibliotheksjournal, Bd. 5/Nr. 2 (2018). https://doi.org/10.5282/o-bib/2018H2S92-105.

[4] https://www.dfg.de/nfdi,zuletztaufgerufenam24.04.2019

Das Kompetenzteam Forschungsdaten an der JGU – Ein kooperatives Angebot

Anne Vieten[1], Karin Eckert[1], Anne Klammt[2], Elisabeth Klein[1] und Jörg Steinkamp[1]
[1]Johannes Gutenberg-Universität Mainz;
[2]Georg-August-Universität Göttingen; Niedersächsische Staats- und Universitätsbibliothek

1. Das Kompetenzteam Forschungsdaten

Im Zusammenhang mit dem Pilotprogramm zu Open Data in Horizon 2020 wurde an der Johannes Gutenberg-Universität Mainz (JGU) nach Möglichkeiten gesucht, die Wissenschaftler/-innen in Bezug auf die gestiegenen Anforderungen im Umgang mit Forschungsdaten zu unterstützen. Hierzu wurde im Jahr 2016 eine Zielgruppenbefragung durchgeführt. Diese gliederte sich in qualitative Interviews mit Wissenschaftler/-innen der verschiedenen Fachbereiche und eine quantitative Umfrage der Wissenschaftler/-innen der JGU mithilfe einer Online-Befragung. Gerade die Verbindung der beiden Befragungsverfahren hat sich als sehr sinnvoll erwiesen, da die qualitativen Befragungen eine größere Tiefe ermöglichen und in der Online-Befragung eine größere Anzahl an Befragten als in der qualitativen Analyse möglich ist, so werden mögliche Nachteile der Befragungsarten ausgeglichen (`https://www.forschungsdaten.org/images/3/3f/04-Klein-mixed_methods.pdf`).

Als Konsequenz aus den Befragungen und um den Wissenschaftlerinnen und Wissenschaftlern an der JGU, vielfältige Services zum Forschungsdatenmanagement (FDM) strukturiert anbieten zu können, wurde im Sommer 2018 das Kompetenzteam Forschungsdaten gegründet.

Zu diesem Zeitpunkt bestanden bereits punktuell FDM-Angebote an verschieden Einrichtungen. Dies betraf Services der Stabsstelle Forschung und Technologietransfer (FT), der Universitätsbibliothek (UB), dem Zentrum für Datenverarbeitung (ZDV), sowie dem Zentrum für Digitalität in den Geistes- und Kulturwissenschaften (mainzed). Zur besseren Sichtbarkeit und Vereinfachung wurden die Services über eine einzige Anlaufstelle zusammengeführt. Fünf Mitarbeiter/-innen der genannten Einrichtungen sind derzeit mit verschiedenen Schwerpunkten und in unterschiedlichem Umfang im Kompetenzteam Forschungsdaten aktiv. Das Team ist über eine Funktions-E-Mailadresse `forschungsdaten@uni-mainz.de` erreichbar. Der Erstkontakt erfolgt über die Stabsstelle Forschung und Technologietransfer. Die genauen Services, Veranstaltungen sowie Informationen zum Forschungsdatenmanagement sind über einen neu geschaffenen Webauftritt `https://www.forschungsdaten.uni-mainz.de/` auffindbar.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026430` veröffentlicht.

Oberhalb der Serviceebene wird das Kompetenzteam bzw. das Thema Forschungsdatenmanagement an der JGU durch eine Lenkungsrunde mit den Leitungen der beteiligten Einrichtungen/Abteilungen der JGU gesteuert.

2. Beratungsservice/-workflow

Durch die Ergebnisse der 2016 durchgeführten Zielgruppenbefragung und in Abstimmung mit den Mitgliedern der Lenkungsrunde zum FDM wurde die Stabsstelle Forschung und Technologietransfer als optimale Erstanlaufstelle in Bezug auf das FDM an der JGU identifiziert, da die Wissenschaftler/- innen dort Unterstützung zu ihren Drittmittelanträgen suchen und somit bereits vor Projektbeginn zum FDM beraten werden können. Vergleichbar mit der Lösung an der JGU sind auch an den Universitäten Hannover und Göttingen Referent/-innen für das Forschungsdatenmanagement in den Forschungsabteilungen angesiedelt (`https://www.fdm.uni-hannover.de/de/team/` und `https://www.uni-goettingen.de/de/team+%26+kontakt/582038.html)` und arbeiten mit weiteren Einrichtungen der Universitäten, wie z.B. den Universitätsbibliotheken, Rechenzentren usw. zusammen.

Die Kooperation verschiedener universitärer Einrichtungen zum Forschungsdatenmanagement ist aufgrund der Komplexität des Themas sehr sinnvoll und erlaubt eine optimale Unterstützung der Wissenschaftler/-innen durch die zum Teil sehr unterschiedliche Expertise der beteiligten Einrichtungen. Das Kompetenzteam Forschungsdaten hat für die JGU den nachfolgenden Beratungsworkflow entwickelt (siehe Abbildung). Der Erstkontakt läuft über die Stabsstelle Forschung und Technologietransfer. Diese bietet zunächst Beratung zu den Grundlagendes Forschungsdatenmanagements, zu organisatorischen Fragen sowie zu den Anforderungen der Drittmittelgeber zum Forschungsdatenmanagement und Open Data, wie zu den notwendigen Inhalten eines Datenmanagementplans etc. an. Ausgehend von der Stabsstelle Forschung und Technologietransfer erfolgt dann, je nach Bedarf und Thema, eine spezialisierte Unterstützung durch die Partner im Kompetenzteam z.B. am Zentrum für Datenverarbeitung, an der Universitätsbibliothek und dem mainzed. Die Universitätsbibliothek übernimmt hier z.B. die Themen Datenpublikation oder Metadatenstandards und das Zentrum für Datenverarbeitung bietet Beratung zu technischen Fragen im Forschungsdatenmanagement an. Im mainzed finden die Wissenschaftler/-innen Unterstützung zu digitalen Methoden in den Geistes- und Kulturwissenschaften.

3. Schulungsangebote/Veranstaltungen

Grundlagenschulungen zum FDM werden sowohl über hochschulinterne Weiterbildungsprogramme (Personalfortbildung, Allgemeines Promotionskolleg, Fortbildungsprogramm des ZDV) als auch auf Anfrage, z.B. für Graduiertenkollegs oder Sonderforschungsbereiche angeboten.

Spezialisierte Schulungen, wie zu digitalen Methoden in den Geisteswissenschaften oder zu Metadaten werden ausschließlich auf Anfrage durchgeführt, da so besser auf die spezifischen Bedürfnisse einzelner Gruppen eingegangen werden kann.

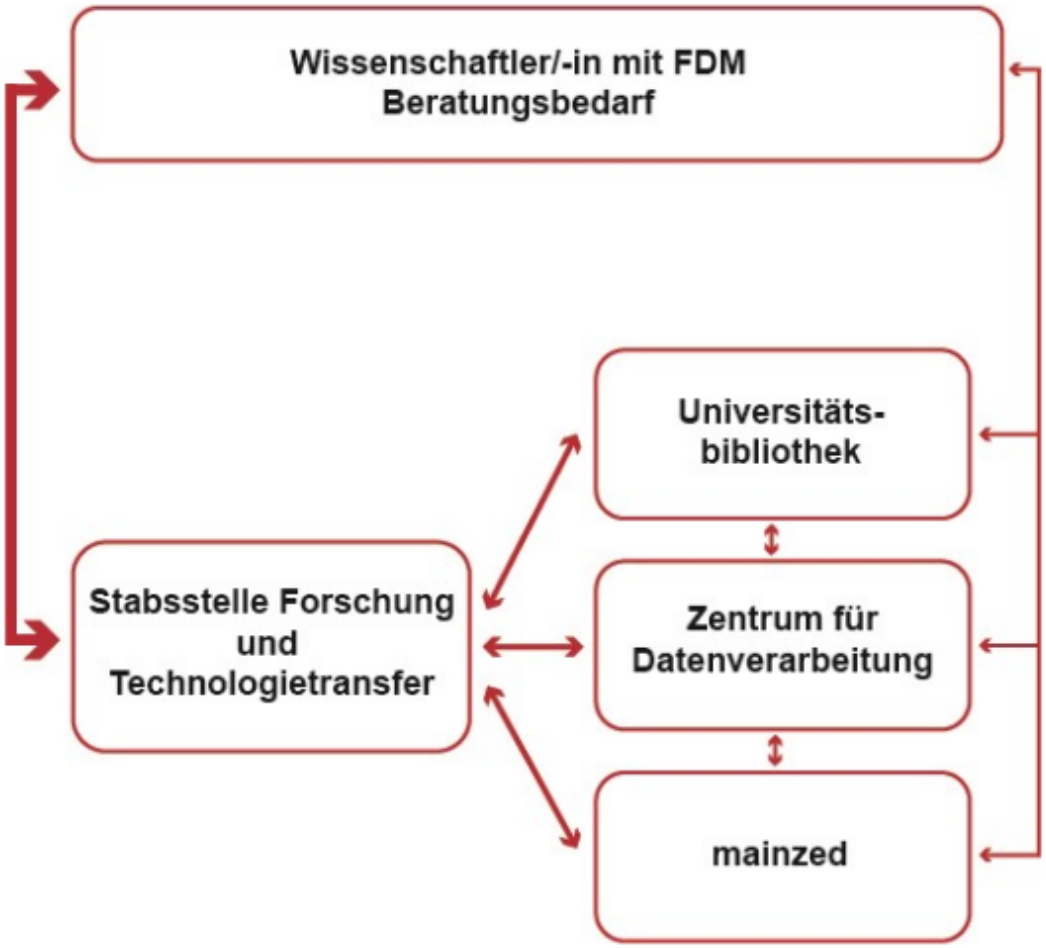


Abbildung 1.: Beratungsworkflow an der JGU

Die Zusammenarbeit mit Graduiertenkollegs an der JGU soll zukünftig noch dadurch verstärkt werden, dass das Kompetenzteam Forschungsdaten bereits in die Planung der Kollegs einbezogen wird.

Einmal jährlich wird an der JGU ein „Aktionstag Forschungsdaten“ durchgeführt. Dieser bietet anwendungsorientierte Vorträge und Workshops zum Forschungsdatenmanagement für die Wissenschaftler/-innen am Standort Mainz an. Zu den Vorträgen werden u.a. externe Referent/- innen eingeladen. Ziel ist die allgemeine Information und Sensibilisierung zum Thema Forschungsdatenmanagement.

4. Technische Infrastruktur zur Publikation und Archivierung von Forschungsdaten

An der JGU befindet sich die technische Infrastruktur zum Forschungsdatenmanagement derzeit im Aufbau. Das bestehende Publikationsrepositorium der Universitätsbibliothek wird basierend auf D- Space zu einem Repositorium für Forschungsdaten erweitert. Ziel ist

die Veröffentlichung und dauerhafte Bereitstellung von zitierfähigen, frei nachnutzbaren Forschungsdaten sowie deren Verlinkung mit Publikationen und die Distribution der Metadaten in übergreifende Nachweissysteme. Im Zentrum für Datenverarbeitung wird ein auf iRODs basierendes Forschungsdatenarchiv für die Wissenschaftler/-innen der JGU entwickelt, das große Datenmengen dauerhaft speichern soll und direkt in die Arbeitsworkflows innerhalb der Projekte eingebunden werden kann. Daten können aus iRODS exportiert und über das Repositorium der Universitätsbibliothek veröffentlicht werden.

Patienten-Apps sammeln Forschungsdaten: IMeRa – Integrated Mobile Health Research Platform

Heinrich Lautenbacher[1], Verena Bizu[2] und Michael Thiede[2]
[1]Universität Tübingen;
[2]Universitätsklinikum Tübingen

1. Motivation

Wenn Patienten medizinische Informationen mit ihren Smartphones und Tablets für Forschungszwecke sammeln, bietet das viele Vorteile um z.B. klinische Studien mit angemessenem Aufwand durchzuführen. Mobile Apps ermöglichen – verglichen mit den herkömmlichen Papierfragebögen – einen höheren **Datenumfang** (mehr Studienpatienten sind einbindbar) und **Datendichte** (z.B. durch tägliche Befragungen der Patienten) sowie eine verbesserte **Datenqualität** (z.B. durch Plausibilitätsprüfungen in den Dialogen). Die Einbindung von Vitalparametern aus Aktivitätssensoren erweitert zudem auch das **Datenspektrum**. Mit der Entwicklung von Apps alleine ist es allerdings nicht getan, denn die so gewonnenen, hochsensiblen Daten müssen sicher übertragen, in medizinischen Datenbanken sicher gespeichert und mit weiteren Forschungsdaten sowie klinischen Behandlungsdaten verknüpft werden, so dass Forscher daraus Erkenntnisse generieren können.

Das Projekt **I**ntegrated **M**obile H**e**alth **R**esearch Pl**a**tform (abgekürzt mit IMeRa) ist ein vom Ministerium für Wissenschaft, Forschung und Kunst, Baden-Württemberg im Rahmen des Förderprogramms eScience gefördertes Vorhaben zur Unterstützung medizinischer Forschung mit mobilen Geräten. Es beinhaltet eine digitale Forschungsplattform für forschungs- und patientenbezogene Daten, die z.B. über Smartphones, Tablets, aber auch mit Aktivitätssensoren datenschutzkonform erhoben, weiterverarbeitet und mit anderen klinischen Daten zusammengeführt werden.

2. Kernziel

Kernziel des Entwicklungsprojekts IMeRa ist die Bereitstellung einer datenschutzgerechten digitalen Forschungsplattform für forschungs- und patientenbezogene Daten, die mit mobilen Endgeräten erhoben werden.

Die Plattform erleichtert das Handling von Forschungsdaten und ist Basis einer disziplin- und lokationsübergreifend verfügbaren IT-Lösung für kollaborative Forschungsvorhaben

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026857` veröffentlicht.

(vor allem klinische Studien). Die Plattform lässt sich deshalb in weitere, zukünftige nationale und internationale Netzwerke interoperabel einbinden.

Wichtige Anwendungsfelder sind z.B. das forschungsbezogene Patientenmonitoring (Remote-Messungen von Vitalparametern und Datenübermittlung an das Klinikum), klinische Studiendokumentationen, aber auch Forschungsvorhaben zur Verbesserung der Nachsorge von Patienten z.B. mit kinder- und jugendpsychiatrischen oder onkologischen Erkrankungen durch effizientere Outcome-Messungen.

Durch die im Projekt entwickelte IT-Infrastruktur sowie IT-Anwendungen ist es nicht nur den Forschern, sondern auch Patienten einer klinischen Studie möglich, effizienter mittels mobiler Endgeräte mit forschungsbezogenen Datenbanken zu kommunizieren. Außerdem können Messdaten von Aktivitätssensoren der Patienten übernommen werden. Die damit verbesserte Informationsverarbeitung unterstützt zudem durch einen höheren Datenumfang und Datendichte sowie eine verbesserte Datenqualität die Entwicklung entsprechender Studien.

3. Erreichtes

Verschiedene mobile Forschungsanwendungen laufen bereits als Pilotprojekte mit der IMeRa-Plattform (z.B. für Zwangserkrankungen bei Kindern, für Patient Reported Outcome Messungen in der Strahlen- und Chemotherapie), sie belegen die Translation der Projektergebnisse in die praxisnahe Anwendung. Auch ist mit dem Universitätsklinikum Ulm der erste externe Partner beim IMeRa aktiv beteiligt. Die wichtigsten erreichten Ziele sind:

1. Entwicklung eines konzeptionellen Datenmodells (Basis Data Set for Research als Referenz-Datenmodell) und Aufbau einer Datenbanklösung für ein Mobile Data Repository for Research, das als Forschungs-Datenquelle dient

2. Realisierung sicherer Web-Services für die verschlüsselte Kommunikation mit den mobilen Geräten (siehe unten nach der Aufzählung)

3. Schnittstellen für verschiedene mobile Aktivitätssensoren, die in Studien der Kinder- und Jugendpsychiatrie sowie der Radioonkologie bereits erfolgreich im Einsatz sind

4. Softwareentwicklung für das Handling von Datenströmen (z.B. im XML- oder JSON-Format) der mobilen Applikationen, Aufbereitung, Konsolidierung und Zusammenführung der Daten

5. Schnittstellenkonzeption für die Datenübermittlung an die zentrale Tübinger Forschungsdatenbank CentraXX und das im Aufbau befindliche Tübinger Datenintegrationszentrum im Rahmen des BMBF-geförderten Projekts DIFUTURE

6. Studiendaten-Auswertungsportal für Forscher

7. Nutzungkonzept für externe Partner (Beispiel Ulmer Quality App: IMeRa-Plattform für Apps zur forschungsorientierten Qualitätssicherung des Universitätsklinikums Ulm)

Besonders sensibel sind **Datenschutz** und **Datensicherheit**: Die mobilen Geräte kommunizieren über verschlüsselte, REST-konforme Webservices mit dem IMeRa-System, um Angriffsmöglichkeiten zu minimieren (unterstützt durch sichere Nutzerauthentifizierungen). Das IMeRa Data Repository wird in der IT-Infrastruktur des Klinikums (als sicherheitskritische Infrastruktur BSI-auditiert) u.a. durch kaskadierte Firewalls geschützt. Wichtigste Datenschutzmaßnahme ist die Pseudonymisierung aller personenidentifizierenden Merkmale (z.B. keine Klartext-Namen), so dass illegal abgegriffene Daten wertlos sind.

Erfolgreiche Software-Entwicklungen für klinische Studien in Tübingen auf IMeRa-Basis sind:

1. iCBT OCD: Prototyp der IMeRa-Plattform für die internetbasierte Studienbegleitung von Kindern mit einem Zwangssyndrom (Obsessive Compulsive Disorder); Kinder- und Jugendpsychiatrie (Studienstart 2017)

2. PROMetheus: Patient Reported Outcome Measurement (PROM) in der gastrointestinalen onkologischen Nachsorge von Patienten; Radioonkologie und Medizinische Klinik (Vorstudie 2018)

3. RadioCareApp: Untersuchung des Einflusses moderaten körperlichen Trainings auf die Rekonvaleszenzphase onkologischer Patienten der Radioonkologie mittels Messung durch Aktivitätsensoren (Vorstudie 2018)

4. iCBT II: Vom Sozialministerium Baden-Württemberg geförderte Weiterentwicklung des iCBT OCD für psychische Störungen des Kindesalters, die mit einer internetunterstützten kognitiven Verhaltenstherapie ambulant versorgt werden (Studienstart 2019)

5. arcTMobile: Patient Reported Outcome Measurement in der Versorgung von Kindern mit autoinflammatorischen Erkrankungen (Studienstart Ende 2019 geplant)

4. Ausblick

Die Vielgestaltigkeit schon entstandener und noch geplanter mobiler Anwendungsbereiche erforderte einen anpassbaren und modularen Aufbau der Plattform, der durch eigene Softwareentwicklung erzielt wurde. Wir sind auch in der Lage, extern erstellte Apps (so z.B. die Ulmer Quality App) in IMeRa einzubinden. Damit ist gewährleistet, dass die erzielten Lösungsansätze übertragbar, erweiterbar und somit für Folgeprojekte anwendbar sind.

IMeRa wird kontinuierlich weiterentwickelt und in Studien und forschungsorientierten Anwendungen eingesetzt. Derzeitiger Schwerpunkt der Entwicklung ist ein erweiterter Datenaustausch mit Forschungsdatenbanken und klinischen Systemen.

5. Weitere Informationen

`https://www.medizin.uni-tuebingen.de/nfmi/imera/index.html`

ViCE – Creating Uniform Approach to Large-Scale Research Infrastructures

Dirk von Suchodoletz und Jonathan Bauer
eScience department, Computer Center, University of Freiburg

Das Ende 2018 abgeschlossene Projekt ViCE – Virtual Open Science Collaboration Environment – beschäftigte sich im Rahmen der eScience-Initiative des Landes Baden-Württemberg mit der Bereitstellung von Virtuellen Forschungsumgebungen (VFU) [1, 2]. Das Projekt bot die Chance, neue Betriebsmodelle und Containerisierungslösungen in verschiedenen Kombinationen und Setups zu evaluieren. Die Ergebnisse wurden in einer Reihe von Publikationen sukzessive veröffentlicht [1, 2, 3, 4, 5, 6, 7, 8]. Die im Projekt bearbeiteten Problemfelder umfassten die Anbindung von VFUs an zentrale Speicherinfrastrukturen, die Anforderungen, die an zukünftige Speichersysteme erwachsen [2, 12], der Transport größerer Datenmengen und Caching [5], das Scheduling von virtualisierten Umgebungen auf einem HPC-System [3], Bedarf nach einer Austauschplattform [11], Betriebs- und Geschäftsmodelle [14, 15] oder das Durchreichen spezieller Hardware in eine VFU. Aus dem Projekt sind weitere Kooperationen wie die Zusammenarbeit mit dem de.NBI (deutsches Netzwerk Bioinformatik) und die Antragstellung für das Science Data Center "BioDATEN" im Bereich Bioinformatik hervorgegangen.[1] Das bildet gleichzeitig die Grundlage für eine Beteiligung an den Aktivitäten zur Nationalen Forschungsdateninfrastruktur.

Aus Sicht der Betreiber wurde zudem bei den verschiedenen Workshops und Schulungen klar, dass das Thema "sichere bzw. datengeschützte Compute-Infrastrukturen" zunehmend an Bedeutung gewinnt. Spezielle Herausforderungen ergeben sich beim Umgang mit personenbeziehbaren Daten sowohl im Bereich der Nutzung in HPC- als auch in Cloud-Umgebungen. Das gilt ebenfalls für eine sichere (langfristige) Speicherung solcher Datensätze im Sinne des Forschungsdatenmanagements. Inzwischen liegen diesbezügliche Anfragen aus verschiedenen Disziplinen vor. Ein Ergebnis von ViCE ist die begonnene Zertifizierung von Speicher- und Compute-Infrastrukturen am Beispiel der de.NBI-Cloud an den Standorten Freiburg und Tübingen. ViCE befasste sich weiterhin mit der reproduzierbaren Erstellung von VFUs durch automatisierte Prozesse (Abb. 1) als auch mit der Bereitstellung und Konfiguration der vorhandenen Forschungsinfrastrukturen für die Nutzung durch Virtuelle Forschungsumgebungen [9].

Im technischen Bereich wurde begleitend zum ViCE-Projekt ein gemeinsames zustandsloses Netzwerkboot-Konzept für die Bereitstellung der Basisinfrastrukturen HPC und Cloud entwickelt. Dieses versetzt reproduzierbar auf Basis des *Boot Selection Service* [7] Compute-Knoten in den für die jeweilige Infrastruktur notwendigen Betriebsmodus. Auf

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026858` veröffentlicht.

[1] Vgl. hierzu die Projekt-Homepage `www.biodaten.info`.

Dirk von Suchodoletz und Jonathan Bauer: ViCE – Creating Uniform Approach to Large-Scale Research Infrastructures.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 218–222. DOI: https://doi.org/10.11588/heibooks.598.c8440

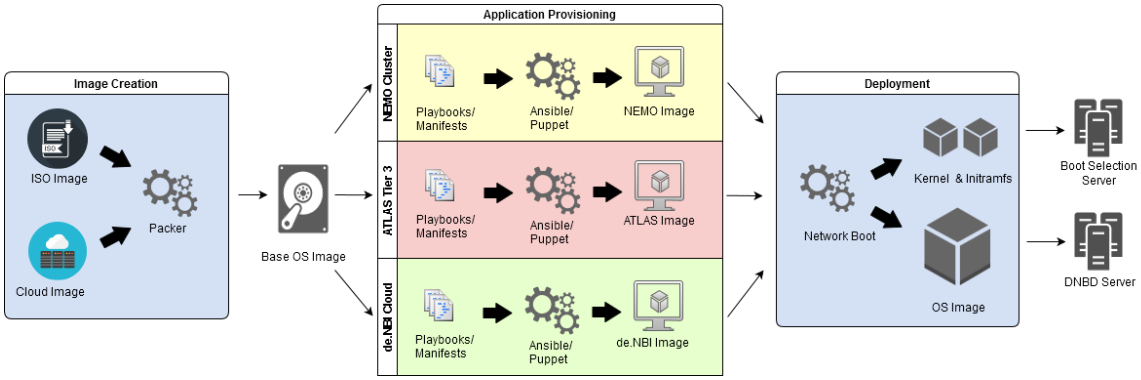


Abbildung 1.: Reproduzierbares Software Provisioning und Deployment.

diese Weise komplementiert es die durch VFU geschaffene Flexibilität auf Ebene der Rechenressourcen, je nach Nutzung und Auslastung der jeweiligen Forschungsinfrastruktur. Dieses gemeinsame Bare-Metal Provisionierungskonzept basiert auf dem in Freiburg entwickelten *Distributed Network Block Devices* (DNBD3) [8] – eine verteilte, redundante und performanceorientierte Speicherlösung.

Im Zuge des Projekts erfolgte eine breit angelegte Information für die beteiligten und später über die ursprünglichen Wissenschafts-Communities hinaus mit dem Ziel, neue Wege der Nutzung bestehender (Groß-)Forschungsinfrastrukturen aufzuzeigen und Hürden beim Einstieg abzubauen. Dieses schloss Aufklärung und Hinweise zum Forschungsdatenmanagement ein. ViCE diskutierte durch die Containerisierung oder Paketierung erste Grundlagen für den langfristigen Zugriff auf Forschungsumgebungen, die durch Daten und Prozesse gekennzeichnet sind.

Das Projekt beschäftigte sich desweiteren mit der breit angelegten Unterstützung von General Purpose GPU-Beschleunigerkarten sowohl für den Zugriff aus einer VFU als für die Bereitstellung in den Forschungsinfrastrukturen bwHPC, bwCloud und bwLehrpool.[2] Es entwickelte für letzteres eine allgemeine Lösung auf Basis des gemeinsamen Netzwerk-Bootkonzeptes. Die Problemstellung lag hierbei in der Umsetzung einer generischen Lösung, die installierte GPU-Karten erkennt und dynamisch den notwendigen Hersteller-Treiber bereitstellt. Dies ist eine Herausforderung, da in Linux-Systemen oft Bibliothekenkonflikte zwischen GPU-Karten-Treiber verschiedener Hersteller bzw. Versionen, die für unterschiedlichen Modelle oder Anwendungszwecke benötigt werden, vorliegen. Daher können diese nicht gleichzeitig in die Systemabbilder installiert werden, sondern müssen beim Netzwerkboot erst zur Bootzeit entsprechend bereitgestellt werden. Systemabbilder mit verschiedenen Treiberversionen abzuwandeln und die treiberspezifischen Abbilder nach der Erkennung der GPU-Karten auszuliefern, ist jeodch aus administrativer Sicht suboptimal.[3]

Das in den Forschungsinfrastrukturen eingesetze Netzwerkboot-Framework wurde so

[2] Landesdienste für die Bereitstellung großer Forschungsinfrastrukturen, siehe `www.bwhpc.de` und `www.bw-cloud.org` bzw. für die Provisionierung und flexible Nutzung von PC-Pools, `www.bwlehrpool.de`.

[3] Die Verwaltung mehrerer Abbilder, die sich im Grund kaum unterscheiden, führt zu unnötigem Mehraufwand und Platzbedarf. Die Unterstützung spezieller Hardware schafft Herausforderungen für den langfristigen Zugriff auf VFUs, die jedoch in diesem Projekt nicht angegangen werden konnten.

erweitert, dass die durch das Installieren bestimmter GPU-Treiber entstandene Änderungen an Systemabbilder als eigenständige "Addons" zusammenfügt werden. Diese werden innerhalb der Systemabbilder abgelegt, sind aber beim Systemstart zu Beginn nicht aktiv. Komplementär zu den Addons wird bei der Einbindung und Vorbereitung des Systemabbilds während der Initiierung des Bootvorgangs die vorhandene GPU-Karte ermittelt und, falls notwendig, das dafür notwendige Addon aktiviert. Anders als bei direkt installierten Treibern ist es hiermit auch denkbar, unterschiedliche Treiberversionen parallel mitzuführen und nicht-destruktive Updates der Treibersoftware auszurollen.

Aus Sicht der Dienst- und Infrastrukturanbieter ging es im Projekt um die Schaffung geeigneter Rahmen für die Austarierung der Bedürfnisse der verschiedenen Nutzergruppen in Hinblick auf zukünftige Betriebsmodelle [4]. Auf diese Weise können neue Nutzergruppen in bestehende föderierte Großinfrastrukturen integriert werden, die vorher mit hohem Aufwand (und potenziell schlechteren Ergebnissen) eigene Forschungsinfrastrukturen betrieben haben. Forschende können sich auch mit kleineren Summen an größeren Infrastrukturprojekten beteiligen und dadurch deutlich schneller starten, als wenn sie selbst den kompletten Software-Hardware-Stack definieren, ausschreiben, installieren und administrieren müssen. Solcherart Konsolidierungen helfen dem Gesamtsystem Universität durch effizienteren Mitteleinsatz und Verbreitung moderner Betriebsmodelle und Organisationskonzepte [14, 15]. Eine weitere Nachnutzung von ViCE-Erkenntnissen deutet sich für die Unterstützung von Aus- und Weiterbildung an den Hochschulen für Angewandte Wissenschaften im Land Baden-Württemberg an, "Machine Learning Lab as a Service", die Nutzung von durchgereichten GPUs in virtuellen Lern- und Arbeitsumgebungen für die Lehre in bwLehrpool und bwCloud.

Danksagung

An dieser Stelle möchten die Autoren dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg für die hervorragende Unterstützung der Projekte ViCE und bwHPC-S5 danken.

Literaturverzeichnis

[1] Konrad Meier, Björn Grüning, Clemens Blank, Michael Janczyk, and Dirk von Suchodoletz. Virtualisierte wissenschaftliche Forschungsumgebungen und die zukünftige Rolle der Rechenzentren. In *10. DFN-Forum Kommunikationstechnologien, 30.-31. Mai 2017, Berlin, Gesellschaft für Informatik eV (GI)*, pages 145–154, 2017. `https://dl.gi.de/bitstream/handle/20.500.12116/473/paper13.pdf`.

[2] Jonathan Bauer, Dirk von Suchodoletz, Jeannette Vollmer, and Helena Rasche. Game of Templates. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 245–262. TLP, Tübingen, 2019. doi:10.15496/publikation-29057.

[3] Felix Bührer, Frank Fischer, Georg Fleig, Anton Gamel, Manuel Giffels, Thomas Hauth, Michael Janczyk, Konrad Meier, Günter Quast, Benoît Roland, Ulrike Schnoor, Markus Schumacher, Dirk von Suchodoletz, and Bernd Wiebelt. Dynamic Virtualized Deployment of Particle Physics Environments on a High Performance Computing Cluster. *Computing and Software for Big Science*, 2018.

[4] Dirk von Suchodoletz, Jonathan Bauer, Susanne Mocken, Oleg Zharkov, and Björn Grüning. Lessons learned from Virtualized Research Environments in today's scientific compute infrastructures, 2019. *E-Science-Tage 2019, Heidelberg* 2019.

[5] Christoph Heidecker, Matthias J. Schnepf, Florian von Cube, Manuel Giffels, and Günter Quast. Dynamic Resource Extension for Data Intensive Computing with Specialized Software Environments on HPC systems. In *Proceedings of the 5th bwHPC Symposium*, pages 161–172. TLP, Tübingen, 2019. doi:10.15496/publikation-29051.

[6] Felix Bührer, Anton J. Gamel, Benoît Roland, Benjamin Rottler, Markus Schumacher, and Ulrike Schnoor. Integration of NEMO into an existing particle physics environment through virtualization. In *Proceedings of the 5th bwHPC Symposium*, pages 187–200. TLP, Tübingen, 2019. doi:10.15496/publikation-29053.

[7] Jonathan Bauer, Manuel Messner, Michael Janczyk, Dirk von Suchodoletz, Bernd Wiebelt, and Helena Rasche. A Sorting Hat for Clusters. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 217–229. TLP, Tübingen, 2019. doi:10.15496/publikation-29055.

[8] Simon Rettberg, Dirk von Suchodoletz, and Jonathan Bauer. Feeding the Masses: DNBD3. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 231–243. TLP, Tübingen, 2019. doi:10.15496/publikation-29056.

[9] Michael Janczyk, Bernd Wiebelt, and Dirk von Suchodoletz. Virtualized Research Environments on the bwForCluster NEMO. In *Proceedings of the 4th bwHPC Symposium October 4th, 2017, Alte Aula Eberhard Karls Universität Tübingen*, pages 37–40. TLP, Tübingen, 2017. doi:10.15496/publikation-25205.

[10] Lev Lafayette, Bernd Wiebelt, Dirk von Suchodoletz, Helena Rasche, Michael Janczyk Janczyk, and Daniel Tosello. The Chimera and the Cyborg – Invivo Hybrid Compute: HPC, Cloud, and Container Implementations. In *Advances in Science, Technology and Engineering Systems Journal – Special Issue on Multidisciplinary Sciences and Engineering*, pages = 1–7, 2019. doi:10.25046/aj040201.

[11] Christopher B. Hauser and Jörg Domaschka. ViCE Registry : An Image Registry for Virtual Collaborative Environments. In *2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, (2):82–89, 2017. doi:10.1109/CloudCom.2017.11.

[12] Dirk von Suchodoletz, Ulrich Hahn, Bernd Wiebelt, Kolja Glogowski, and Mark Seifert. Storage infrastructures to support advanced scientific workflows. In Michael Janczyk, Dirk von Suchodoletz, and Bernd Wiebelt, editors, *Proceedings of the 5th bwHPC Symposium*, pages 263–279. TLP, Tübingen, 2019. doi:10.15496/publikation-29058.

[13] Felix Bartusch, Kolja Glogowski, Ulrich Hahn, Michael Janczyk, Steve Kaminski, Jens Krüger, Volker Lutz, Gerhard Schneider, Dirk von Suchodoletz, Thomas Walter and Bernd Wiebelt. Defining the future scientific data flow for multi-disciplinary research data. *E-Science-Tage 2019, Heidelberg* 2019.

[14] Dirk von Suchodoletz, Janne Chr. Schulz, and Jan Leendertse. Abstraktion erlaubt neue Aufgabenverteilung – Virtualisierung, Clouds und die zukünftige Rolle wissenschaftlicher Rechenzentren. *Wissenschaftsmanagement*, (4):31–35, 2017.

[15] Dirk von Suchodoletz, Janne Chr. Schulz, and Jan Leendertse. Vom wissenschaftlichen Rechenzentrum zum Rechenzentrum für die Wissenschaft – Überlegungen zur Rekalibrierung von IT-Strategien an Universitäten und Hochschulen. *Wissenschaftsmanagement*, (5/6):30–35, 2017.

Projekt UNEKE: Roadmap zu passgenauen Infrastrukturen für Forschungsdatenspeicherung

Bela Brenger[1], Ania López[2], Stephanie Rehwald[3], Stefan Stieglitz[3] und Konstantin Wilms[2]

[1]IT Center RWTH Aachen;
[2]Universitätsbibliothek Duisburg-Essen;
[3]Universität Duisburg-Essen, Abteilung für Informatik und Angewandte Kognitionswissenschaft

Das BMBF- geförderte Projekt UNEKE („Vom USB-Stick zur NFDI – Entwicklung eines Kriterien geleiteten Entscheidungsmodells für den Aufbau von Forschungsdateninfrastrukturen“) hat eine Roadmap entwickelt, wie wissenschaftliche Bedarfe und zur Verfügung stehende Lösungen zur Forschungsdatenspeicherung in Einklang gebracht werden können.

Anschlussfähige Infrastrukturen zur Forschungsdatenspeicherung Zur nachhaltigen Aufbewahrung von Forschungsdaten sind Hochschulen bestrebt ihren Forschenden die dafür notwendigen Infrastrukturen zur Verfügung stellen [1]. Viele stehen dabei vor dem Problem weder die Bedarfe der eigenen Wissenschaftler zu kennen noch verfügbare Repositorien-Lösungen hinsichtlich der Rahmenbedingen und Bedarfe vor Ort bewerten zu können. Hochschulen befinden sich zusätzlich in dem Dilemma, einerseits zwar lokale Strukturen aufbauen zu müssen, um aktuelle Bedarfe zu decken. Anderseits sollen lokale Insellösungen vermieden werden und langfristig vernetzte Strukturen auf (inter)nationaler Ebene wie der NFDI etabliert werden.

Die Entscheidung, ob und welchem Umfang eine Infrastruktur zur Speicherung von Forschungsdaten an einer Hochschule aufgebaut wird, bewegt sich in einem Spannungsfeld, das durch die drei Eckpunkte „Bedarfe der Wissenschaflter*innen“, „Rahmenbedingungen“ und „Verfügbare Lösungen“ abgesteckt ist. Ein Entscheidungsprozess sieht zunächst eine Bestandaufnahme aller Eckpunkte vor und diese aufeinander abzustimmen.

1. Bedarfe der Forschenden

Zur Bestimmung der Praxis und Herausforderungen im Umgang mit Forschungsdaten wurden in den letzten Jahren an mehreren deutschen Hochschulen Wissenschaftler*innen mit online-gestützten Fragebögen befragt, um möglichst übergreifende Bedarfe zu identifizieren und einen guten Überblick über den Status Quo des Forschungsdatenmanagements zu erfassen [2]. Die einzelnen Umfragen wurden meist lokal an einer Hochschule durchgeführt und lassen sich trotz ähnlicher Fragebögen nicht direkt vergleichen. UNEKE ver-

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI https://doi.org/10.11588/heidok.00026859 veröffentlicht.

folgte den Ansatz eine gleichbleibende generische Umfrage an vielen Hochschulen durchzuführen, um übergreifende Aspekte und standortspezifische Unterschiede herausarbeiten zu können: zwischen März 2018 und Januar 2019 wurden in der UNEKE-Umfrage an dreizehn Hochschulen in NRW mit über 1600 Datensätzen die Bedarfe der Wissenschaftler*innen erfasst. Die Analyse ergab trotz disziplinspezifischer als auch hochschultypischer Unterschiede bei der angestrebten und praktizierten Aufbewahrung von Forschungsdaten als übergreifende Gemeinsamkeit die hohe Nachfrage nach Langzeitverfügbarkeit und Archivierung von Forschungsdaten [3]. Die Praxis und angestrebte Speicherung von Forschungsdaten klafft allerdings auseinander, sodass zu vermuten ist, dass der Bedarf durch die bestehenden institutionellen Angebote momentan nicht gedeckt werden kann. Des Weiteren wurde deutlich, dass der Großteil der gespeicherten Forschungsdaten entweder gar nicht oder nicht systematischen mit Metadaten beschrieben wird. Daraus könnte sich ein guter Anknüpfungspunkt für Beratungsangebote zur Archivierung und zu Repositorien ergeben.

Darüber hinaus hat sich gezeigt, dass neben den konkreten technischen Anforderungen die Einstellung der Forscher/innen zu open data und zu einem offenen Umgang mit Forschungsdaten eine entscheidende Rolle spielen [4] und entsprechend bei der Konzeption einer Infrastruktur zur Forschungsdatenspeicherung berücksichtigt werden sollte.

2. Bewertung von Repositorienlösungen

In einem zweiten Schritt wurden die derzeit zur Verfügung stehenden verschiedenen Software- Lösungen zur Umsetzung eines institutionellen Repositoriums verglichen. Es existieren generische, anwendungsfertige Out-Of-The-Box-Systeme, aber auch Software, die umfangreiche individuelle Ausgestaltungen z. B. des Datenmodells oder der physischen Speicherschicht ermöglichen. Die Auswahl ist groß und ein direkter Marktführer zeichnet sich nicht ab. UNEKE hat daher in einer deutschlandweiten Kurzumfrage den Stand von institutionellen Repositorien an deutschen Universitäten abgefragt und typische Umsetzungsszenarien herausgearbeitet.

3. Rahmenbedingungen

Die Einrichtung und Nutzung eines Repositoriums sind eng verknüpft mit dem Gesamtentwicklungsstand des FDM an einer Hochschule. Die erfolgreiche Bereitstellung eines Repositoriums kann nicht losgelöst von Angeboten wie FDM-Beratung, Publikationsberatung, Datenmanagementplänen und Speicherstrukturen für aktive Daten gesehen werden. Darüber hinaus bestimmt der Umfang an zur Verfügung stehenden Personal die Auswahl an möglichen Repositorien- Lösungen in erheblichen Maße. Hier sind nicht nur die technischen Aspekte der Umsetzung eines Repositoriums relevant, sondern auch die Verknüpfung mit FDM-Services allgemein. Nicht zuletzt benötigt die Etablierung einer nachhaltigen Speicherpraxis einen verbindlichen Rahmen in Form von beispielsweise Leitlinien.

4. Roadmap

Obwohl eindeutig Bedarfe seitens der Forschenden bestehen, liegen diese oftmals nicht gebündelt oder als konkrete Forderung formuliert vor. Zusammen mit einer niedrigen politischen Relevanz innerhalb der Hochschule für das Thema führt dies zu einer abwartenden Position vieler Hochschulen. Es besteht kein starker Handlungsdruck. Aus der Kurzumfrage wurde deutlich, dass der Frage nach einer geeigneten Ablage für Forschungsdaten erst nachgegangen wird, wenn politischer und lokaler Handlungsdruck in Summe eine bestimmte Schwelle überschreiten. Bei der Entscheidungsfrage, welches System gewählt werden soll, können keine klaren Kriterien gegeben werden – weder die in Umfragen erfassten Bedarfe der Wissenschaftler noch die technischen Unterscheidungsmerkmale der Repositorienlösungen führen für sich zu klaren und übertragbaren Entscheidungskriterien und Handlungsempfehlungen. Vielmehr entscheidend bei der Frage nach Infrastruktur ist das Anstoßen eines Prozesses, der die „Repositorienentscheidung“ in einen FDM Gesamtprozess integriert und die Komponenten des Spannungsfelds in ihrer Dynamik mitbedenkt und in Einklang bringt. Dieser Prozess wird üblicherweise von „Triggern“ angestoßen – initialen Ereignissen und Konstellationen, die sowohl eine politische oder lokale Dimension haben können und häufig gemeinsam auftreten.

Diese initiale „Trigger“ führen verknüpft mit dem Status Quo der drei Eckpunkte des Spannungsfelds zu passenden Anwendungsszenarien, die vom Aufsetzen eines individuellen Forschungsdatenrepositoriums über die Nutzung eines externen Dienstes wie RADAR bis hin zum gemeinsamen Betrieb innerhalb einer Landeslösung oder eines Konsortiums reichen. Gibt es ein richtig oder falsch?! Entscheidend ist der Prozess und weniger die konkrete Designentscheidung.

Literaturverzeichnis

[1] Dierkes, J., & Curdt, C. (2018). Von der Idee zum Konzept – Forschungsdatenmanagement an der Universität zu Köln. O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB, 5(2), 28-46. https://doi.org/10.5282/o-bib/2018H2S28-46

[2] Tristram, Frank; Bamberger, Peter; Uğur Çayoğlu; Hertzer, Jörg; Knopp, Johannes; Kratzke, Jonas; Rex, Jessica; Schwabe, Fabian; Shcherbakov, Denis; Svoboda, Dieta-Frauke; Wehrle, Dennis; „Öffentlicher Abschlussbericht von bwFDM-Communities - Wissenschaftliches Datenmanagement an den Universitäten Baden-Württembergs“; KITopen 2018; doi:DOI: 10.5445/IR/1000083272

[3] Brenger, Bela et al. (2019): UNEKE: Forschungsdatenspeicherung - Praxis und Bedarfe: Online- Survey 2019. DuEPublico: Duisburg-Essen Publications online, University of Duisburg-Essen, Germany. Online unter: https://duepublico2.uni-due.de/receive/duepublico_mods_00070259.

[4] Wilms, K., Brenger, B., López, A., Rehwald, S., „Open Data in Higher Education - What Prevents Researchers from Sharing Research Data?”, ICIS 2018 Proceedings, ISBN 978-0-9966831-7-3

Implementing a Data Sharing Agreement within a biomedical research consortium

Jonas Narchi[1], Christian Deisenroth[1] and Christoph Schickhardt[2]

[1]National Center for Tumor Diseases, University Hospital, Heidelberg;
[2]National Center for Tumor Diseases, German Cancer Research Center, Heidelberg

1. Background

Sharing research data is pivotal for efficient and successful biomedical research and a crucial aspect of open science. Although many researchers aware of the potential to increase pace and transparency in science through sharing data, data sharing is still far from being common practice. This somewhat paradox situation is due to several hindering factors. One key factor is the researchers' fear of giving away opportunities by sharing data with other researches in their field. Data sharing in medical sciences is reported to be less likely than in other disciplines. It has been argued that this might be the case due to the higher sensibility of the data which often includes personal data from patients. However, if restrictions can be put on the possible use of data, the willingness to share increases. Some actions have already been taken to improve data sharing in biomedical sciences like requirements for publishing a medical trial in ICMJE-Journals or the release of new data sharing guidelines of funding organisations (EU Horizon 2020, NIH, MRC). There are some ideas how sharing of medical research data could be encouraged but most of them are of a theoretical kind. What is still missing, are use cases that may serve as best practice examples for the implementation of data sharing in the field of biomedicine.

The coNfirm project (Systems Medi**c**ine **o**f a Heart Disease **N**etwork **f**or **i**mproving **m**ultilevel heart health) is part of the e:Med funding program by the Federal Ministry of Education and Research which aims to establish systems medicine in Germany. Systems medicine is one among several current developments in biomedicine which are all based upon the gathering and processing of very large data sets (Big Data). The interdisciplinary coNfirm project consists of researchers in the areas medicine, bioinformatics and ethics from different German university hospitals and research institutions. While the biomedical subprojects mainly aim for understanding of cardiac health conditions, the ethical subproject organises, monitors, and evaluates data sharing between these researchers. The ethical subproject aims at implementation of data sharing concepts and processes in the consortium. The hands-on experiences will be useful to other researchers and consortia in the biomedical field as well. In this abstract we briefly illustrate the development and implementation of the data sharing agreement within the coNfirm consortium.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026853` veröffentlicht.
Corresponding author:Christoph.Schickhardt@med.uni-heidelberg.de

Jonas Narchi, Christian Deisenroth und Christoph Schickhardt: Implementing a Data Sharing Agreement within a biomedical research consortium. In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge, Heidelberg: heiBOOKS, 2020, 226–228. DOI: https://doi.org/10.11588/heibooks.598.c8442

2. Development of the Data Sharing Agreement

We started with a systematic online research on already existing guidelines for and studies on data sharing published by researchers and international research organisations. The results of this literature research informed the compilation of a survey that we carried out among the members of the coNfirm consortium using questionnaires with open and multiple-choice questions. Targeting both negative factors that impede and positive factors that facilitate data sharing we aimed at gaining insights into the motivational and organisational aspects of data sharing that could then be used as a starting point in developing a data sharing agreement within the consortium.

The results of our survey showed that the major issues a data sharing agreement would have to address are: authorship, questions about the publication of results, the kind of metadata needed to understand a dataset, and rules and guidelines for the sharing of data in general. Also, the members considered gathering and preparation of data at the one hand and analysis and evaluation of data on the other hand as the main work tasks to be valued in authorship decisions. "Agreement on authorships before the start of a project" and "mediation by the coordinator" were judged to be the most important factors for a successful process of data sharing within the consortium.

We then used the results of the literature review and the intraconsortial survey to develop a data sharing agreement for the consortium. We integrated paragraphs concerning authorship, metadata and the division/sharing of data and workload as mandatory parts to be addressed by every cooperation within the consortium. Among others, the data sharing agreement addresses the following issues:

- Introduction and background
- Underlying principles of data sharing (transparency, cooperation, mutual trust, FAIR data principles, good scientific practice, data protection)
- Hints to the Legal framework
- Principles of data protection
- Handling of data
- Recognition of data producers
- Regulations for the implementation of project-specific data usage agreements
- Legal liabilities
- The role of the data usage consultants
- Procedures to cope with conflicts among the consortium members

3. Development of the Data Sharing Agreement

The last step of the development of the data sharing agreement was already the first step of its implementation: In a consortium workshop dedicated to data sharing we discussed the first draft of the agreement with the consortium members. The data sharing agreement was revised according to the discussions and feedback loops, and then accepted in consensus and finally signed by all members of the consortium. The function of the data sharing agreement is to build a general framework to raise awareness and induce trust to enable and encourage safe and successful data sharing.

The general data sharing agreement determines that concrete research and data sharing undertakings within the consortium are being planned and outlined in specific data usage agreements. We thus developed a template for such specific data usage agreements that will serve as a way of applying and implementing the general principles of the data sharing agreement to concrete and individual research undertakings. We, as the data sharing consultants, accompany and moderate discussions and the elaboration of specific data usage agreements between the consortium members for every concrete research undertaking. While e.g. the data sharing agreement establishes that authorships should be distributed fairly among researchers according to their respective tasks and work load, the template for individual data usage agreements provides a concrete list of planned work contributions and authorship order. Once the participating researchers have reached consensus on a preliminary authorship list in accordance with their respective tasks and workload, the data sharing process can be initiated.

4. Conclusion

The development of a data sharing agreement bears the possibility to induce trust and awareness within a group of researches. By pro-actively addressing and moderating the main questions that could hinder the exchange of data (authorship, distribution of tasks,. . .) it aims to support data sharing and raise awareness for its substantial benefits. The publication of our hands-on experiences will help other researchers and research consortia share data and establish a culture of data sharing.

Koordiniertes Forschungsdatenmanagement in Baden-Württemberg: Die Projekte bwFDM-Info und bw2FDM

Fabian Gebhart[1], Jan Kröger[2], Kerstin Wedlich-Zachodin[3] und Frank Tristram[3]

[1]Universität Heidelberg;
[2]Universität Konstanz;
[3]Karlsruher Institut für Technologie

Das Poster stellt die Tätigkeiten und Entwicklung der baden-württembergischen Landesprojekte bwFDM-Info I und II vor sowie die Pläne für das künftige Projekt bw2FDM.

Ziel des Projekts bwFDM-Info I war es, freies Material zum Forschungsdatenmanagement für Forschende auf einem Informationsportal zur Verfügung zu stellen. Sie sollten sich auf forschungsdaten.info über Forschungsdatenmanagement informieren und an Best Practices orientieren können. Das Angebot stützte sich dabei auf den im Landesprojekt bwFDM- Communities identifizierten Bedarfen.

Im Nachfolgeprojekt bwFDM-Info II verschob sich der Fokus hin zum Bekanntmachen der Plattform und Etablieren ihres nachhaltigen Betriebs. Die Universitäten Heidelberg, Hohenheim, Konstanz, Tübingen und das Karlsruher Instituts für Technologie (KIT) vereinbarten, den langfristigen Betrieb der Plattform sicherzustellen. Im Rahmen eines Beteiligungsmodells konnten weitere institutionelle und individuelle Partner aus Hessen, Nordrhein-Westfalen, Niedersachsen, Thüringen und Sachsen gewonnen werden, die die Plattform inhaltlich und strukturell weiter ausgebaut haben. Daneben haben die Projektpartner eine Instanz des DMP-Tools Research Data Management Organiser (RDMO) eingerichtet und auf der Webseite integriert.

Vernetzung und Outreach waren weitere wichtige Themen von bwFDM-Info I und II. Mit dem Arbeitskreis der Forschungsdatenverantwortlichen aller Universitäten im Land (AK FDM) wurde ein Gremium etabliert, in dem sich die Beteiligten regelmäßig über ihre jeweiligen FDM- Aktivitäten austauschen und offene Fragen klären. Die E-Science-Tage 2017 in Heidelberg haben als erfolgreiche Fachkonferenz maßgebliche FDM-Akteure bundesweit zusammengebracht. Mit der Koordination der baden-württembergischen E-Science-Projekte aus den Bereichen Forschungsdatenmanagement und Virtuelle Forschungsumgebungen hat das Projekt bwFDM-Info deren Austausch gefördert und Synergien in der Entwicklung geweckt. Dieses erfolgreiche Konzept soll 2019 bis 2023 mit dem neuen Projekt bw2FDM weiterentwickelt werden. Ein Hauptziel ist es, die Aufbauaktivitäten der baden-württembergischen Science Data Centers (SDC) koordinierend zu begleiten, deren

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026852` veröffentlicht.
Corresponding author:jan.kroeger@uni-konstanz.de

Vernetzung untereinander zu intensivieren und ihre Sichtbarkeit für die Forschungscommunities im In- und Ausland zu steigern. In den SDCs arbeiten Wissenschaftler/innen und Mitarbeiter/innen von Rechenzentren und Bibliotheken zusammen, um

1. den Zugang zu und die Nutzung von digitalen Datenbeständen zu ermöglichen,
2. Daten für Forschung und Innovationen mit den Mitteln der Big-Data-Analyse zu erschließen und
3. Aus- und Weiterbildungsangebote für die digitale datengetriebene Forschung und Entwicklung zu erstellen.

Damit können die Science Data Center einen Beitrag zur Ausgestaltung der künftigen Nationalen Forschungsdateninfrastruktur leisten.

Ziel ist u.a. die Nachnutzung der von den SDCs entwickelten Produkten für die Forschungscommunity. Die Weiterentwicklung des Info-Portals sowie die Fortführung der Konferenzreihe „E-Science-Tage“ werden Fachkompetenzen zum Forschungsdatenmanagement auf nationaler Ebene weiter bündeln.

Die Projekte zum Forschungsdatenmanagement bwFDM-Info I und II sowie bw2FDM wurden bzw. werden gefördert vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg.

SDS@hd – Scientific Data Storage

Martin Baumann, Oliver Mattes, Sabine Richling, Sven Siebler and Alexander Balz
Computing Centre (URZ), Heidelberg University, Germany

SDS@hd is a central storage service for hot large-scale scientific data that can be used by researchers from all universities in Baden-Württemberg. It offers fast and secure file system storage capabilities to individuals or groups, e.g. in the context of cooperative projects. Fast access is possible from data generating facilities like microscopes as well as from data analysis systems like HPC systems. Data protection requirements can be fulfilled by data encryption and secure data transfer protocols. The service is operated by the Heidelberg University Computing Centre.

1. Introduction

SDS@hd [1] is a service that provides large-scale storage for scientific data and is meant to be used for "hot data" – data that is frequently accessed and worked with. The service can be used by researchers at most public higher education institutions in Baden-Württemberg. User authentication and authorization are implemented in terms of the federated identity management in Baden-Württemberg bwIDM [2] allowing researchers to use the existing ID of their home institution transparently for this service. The SDS@hd service website [1] provides further information about the technical and institutional requirements and the registration process.

2. Features of SDS@hd

Tailored to the safety of research data Research data is a precious resource and should be stored using a trustworthy service to keep it safe from prying eyes. Data handled via SDS@hd is stored in an appropriate environment at the Heidelberg University Computing Centre (URZ). It is protected by state-of-the-art technologies, encryption as well as restrictive access and data policies.

Ideal for collaborations SDS@hd is useful for researchers from different departments or institutions who want to work together. They can join a collaborative storage project and store their research data at a single spot. Using a web interface, the storage project owner can manage user groups and user roles and can thus determine who is allowed to access which parts of the data storage.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026851` veröffentlicht.

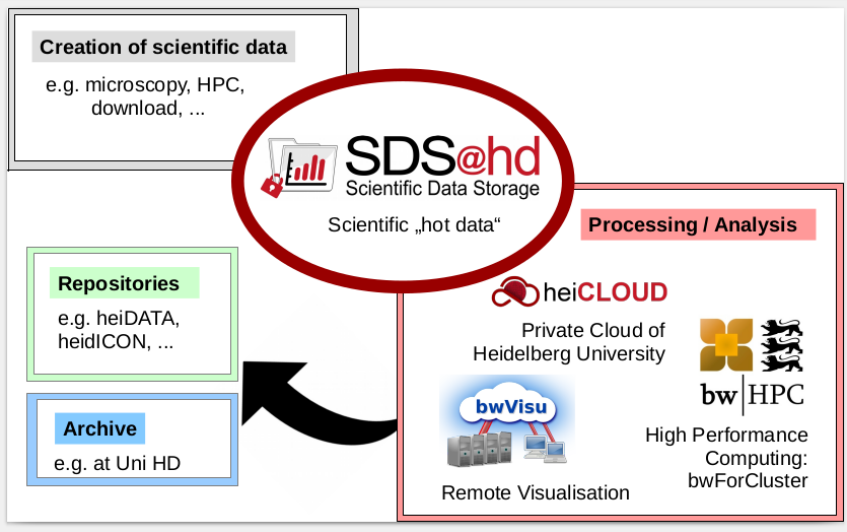


Figure 1.: SDS@hd supports the entire scientific data workflow.

Storing HPC and visualisation data The SDS@hd service is connected to a range of data-intensive computing resources also available through the Heidelberg University Computing Centre. These resources – such as the bwForCluster MLS&WISO or the bwVisu visualisation service – can automatically obtain or save data to and from the SDS@hd storage infrastructure. General access to SDS@hd, e.g. from the bwUniCluster and other bwForClusters or from your notebook, is possible via the protocols SMB (2.x/3.x), NFSv4 (Kerberos) or SFTP. Users no longer need to find their own large-scale storage solutions.

Support for the whole data life cycle SDS@hd aims to serve the entire scientific data workflow (see Fig. 1). When research projects are finished, support is provided for the transfer of data to general or community-specific archives and repositories.

3. Hardware and Software

The data is stored on the **Large Scale Data Facility (LSDF2)**, a state-of-the-art storage system located at the URZ's main server rooms in Heidelberg. The system is associated to a project between URZ and the Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT). The project aims to provide modern and valuable storage services for researchers.

The system is gradually extended and will provide up to 25 petabytes of storage space in 2020. A range of the newest HDDs interspersed with several solid-state drives (SSDs) guarantees high access speeds. Even large numbers of smaller files – usually a time-consuming challenge for storage systems – can be swiftly processed.

LSDF2 is made for high availability: An efficient RAID concept prevents any data loss, and a backup power system weathers any potential power outages. There is no single point of failure in the system. Additionally a backup concept allows to access the data, even in a disaster situation without any time consuming data restore needed.

On LSDF2, the high-performance clustered file system software "Spectrum Scale" (formally known as "GPFS") developed by IBM is used. It provides concurrent and fast file access to applications via various protocols including NFS and SMB. It offeres several features like high availability, disaster recovery and shared access to file systems from remote Spectrum Scale clusters. Additionally, tools for management and administration of storage clusters are contained. In total, Spectrum Scale is a solid and flexible basis for high-quality storage services for research.

Acknowledgements

The LSDF2 infrastructure was built as part of the bwDATA initiative that fosters collaboration in the field of data-intensive computing between higher education institutions in Baden-Württemberg [3]. The system is funded by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG.

Bibliography

[1] SDS@hd – Scientific Data Storage. `https://sds-hd.urz.uni-heidelberg.de`

[2] Föderiertes Identitätsmanagement der baden-württembergischen Hochschulen. `http://bwidm.de`

[3] Hartenstein, H., T. Walter, and P. Castellaz. “Aktuelle Umsetzungskonzepte der Universitäten des Landes Baden-Württemberg für Hochleistungsrechnen und datenintensive Dienste.” Praxis der Informationsverarbeitung und Kommunikation, Band 36, Heft 2 (2013): 99-108.

Community-spezifische Forschungsdatenpublikation (CS-FDP)

Fabian Gebhart[1], Jochen Apel[2], Martin Baumann[1], Jeromin Fest[1], Benjamin Scherbaum[1], Leonhard Maylein[2] und Georg Schwesinger[2]

[1]Unversität Heidelberg, Universitätsrechenzentrum (URZ), Deutschland;
[2]Unversität Heidelberg, Universitätsbibliothek (UB), Deutschland

Mit der Einrichtung des Kompetenzzentrums Forschungsdaten (KFD) als zentralem Beratungs- und Dienstleistungsanbieter verfolgt die Universität Heidelberg das Ziel, Wissenschaftlerinnen und Wissenschaftler während des gesamten Lebenszyklus von Forschung-sdaten konstruktiv zu begleiten. Im Projekt Community-spezifische Forschungsdatenpublikation (CS-FDP) werden die Infrastruktur- und Serviceangebote des KFD systematisch weiterentwickelt. Dies geschieht durch den Aufbau eines Pools von generischen Softwarewerkzeugen zur Erstellung dynamischer Forschungsdatenportale sowie ein Konzept zur nachhaltigen Integration der Heidelberger Forschungsdaten in übergreifende Archivierungskonzepte. Darüber hinaus wird die Professionalisierung des Datenmanagements an der Universität durch die Verankerung des Themas in Forschung und Lehre befördert.

Die zentrale Fragestellung des Projekts lautet dabei: Wie können institutionelle Infrastruktureinrichtungen – in diesem Fall die Universitätsbibliothek und das Universitätsrechenzentrum – mit ausgewählten, generischen Lösungen Möglichkeiten zur Publikation von Forschungsdaten schaffen, die dennoch eine auf spezifische fachliche Bedürfnisse zugeschnittene individuelle Präsentation und Vernetzung der Daten sowie eine nachhaltige und verlässliche Langzeitarchivierung derselben gewährleisten?

Mit dem Angebot von flexiblen, generischen Präsentationsoberflächen für veröffentlichte Forschungsdaten werden Forschungsdisziplinen adressiert, denen bislang noch keine im Fach etablierten, auf dessen spezifische Anforderungen zugeschnittenen Datenpublikationsplattformen zur Verfügung stehen und deren Bedarf im Hinblick auf die Webpräsentation der Daten über bloße Downloadoptionen für die Rohdaten hinausgeht. In enger Zusammenarbeit mit Pilotanwendern aus unterschiedlichen Disziplinen wird ein Portfolio an cloudbasierten Präsentationsoberflächen bereitgestellt, das es ermöglicht, Daten in ihrem jeweiligen fachspezifischen Kontext darzustellen und zu vernetzen.

Ziel dabei ist es, eine handhabbare Auswahl von generischen Softwareangeboten aufzubauen, mit deren Hilfe entsprechende Portale erstellt werden können, deren Betrieb inklusive der dauerhaften Pflege der Portale mit den Personalressourcen des KFD sichergestellt werden kann. Parallel wird ein Konzept für die Langzeitarchivierung der unterschiedlichen Forschungsdatenbestände der Heidelberger Universität erarbeitet, auf dessen Grundlage die Aufgabe der dauerhaften Sicherung entsprechender Datenbestände systematisch verfolgt wird. Hierbei ist sowohl eine campusinterne Archivinfrastruktur prototypisch zu

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026741` veröffentlicht.

Fabian Gebhart et al.: Community-spezifische Forschungsdatenpublikation (CS-FDP).
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 234–235. DOI: https://doi.org/10.11588/heibooks.598.c8445

entwickeln, die der Heterogenität der Gesamtheit der Forschungsdaten Rechnung trägt, als auch deren Anbindung an übergeordnete Archivlösungen zu berücksichtigen.

OpenDACHS: Ein Citation Repository zur nachhaltigen Archivierung zitierter Online-Quellen

Matthias Arnold[1], Hanno Lecher[2] und Sebastian Vogt[2]
[1]Heidelberg Research Architecture - Heidelberg Centre for Transcultural Studies, Universität Heidelberg;
[2]Zentrum für Ostasienwissenschaften, Universität Heidelberg

Das Internet ist seit dem Ende des vorigen Jahrhunderts eine zunehmend wichtige Plattform verschiedenster Publikationen sowie ganz allgemein des sozialen Diskurses. Die hier veröffentlichten Inhalte sind jedoch extrem flüchtig, da sie jederzeit inhaltlich verändert werden können oder ganz aus dem Netz verschwinden. Gleichzeitig sind sie jedoch für unterschiedliche Forschungsfragen von großer Bedeutung, und das Zitieren von Online-Quellen ist mittlerweile Alltag in der akademischen Forschung. Die Flüchtigkeit dieser Quellen und damit deren Nachprüfbarkeit ist zwar durch verschiedene Untersuchungen gut dokumentiert und belegt, die daraus resultierenden Folgen für die Forschungspraxis wurden aber bislang weitgehend ignoriert. Zwei Beispiele verdeutlichen die Problematik:

1. In Artikeln der Ausgabe 1.2010 der vom Exzellenzcluster Asia and Europe veröffentlichten Zeitschrift „Journal of Transcultural Studies“ wurden insgesamt 6 Online-Ressourcen zitiert, von denen heute nur noch 3 funktional sind. Die anderen 3 zitierten Quellen sind aus dem Netz genommen und nicht mehr nachprüfbar.

2. Für die Monographie A Continuous Revolution von Barbara Mittler (Cambridge, 2012) wurde im Webarchiv DACHS (Digital Archive for Chinese Studies) des Instituts für Sinologie ein Citation Repository angelegt. Obwohl von den zitierten 76 Internetquellen heute über 50% nicht mehr erreichbar sind, können via DACHS alle zitierten Quellen in ihrer ursprünglichen Form eingesehen und nachgeprüft werden.

In diesem Poster stellen wir eine Erweiterung des DACHS-Projekts vor: OpenDACHS. Das bisher auf sinologische Inhalte konzipierte DACHS wird derzeit in einer Kooperation zwischen dem Institut für Sinologie am Zentrum für Ostasienwissenschaften (ZO) und der Heidelberg Research Architecture (HRA) des Heidelberg Centre for Transcultural Studies (HCTS) zu OpenDACHS umgestaltet. Es wird als Service aufgebaut, funktional erweitert und zunächst für das neue Centre for Asian and Transcultural Studies (CATS) geöffnet. Außerdem werden Arbeitsabläufe etabliert, die auch Fragen der Katalogisierung und der Verwaltung von Speicherplatz einbeziehen.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026743` veröffentlicht.

Matthias Arnold, Hanno Lecher und Sebastian Vogt: OpenDACHS: Ein Citation Repository zur nachhaltigen Archivierung zitierter Online-Quellen. In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge, Heidelberg: heiBOOKS, 2020, 236–237. DOI: https://doi.org/10.11588/heibooks.598.c8446

DACHS nutzt das Datenformat WARC, das als ISO Standard 28500:2017 veröffentlicht ist. Als Webcrawler kommt die Open-Source Software Heritrix zum Einsatz, die vom Internet Archive geschrieben wurde und heutzutage vielfach im Einsatz ist. Archivdateien können über Tools wie den Betrachter der Wayback Machine angesehen werden.

Zwar bieten frei zugängliche Services wie WebCite oder das Internet Archive an, einzelne Webseiten zu archivieren. Der Funktionsumfang dieser Services ist allerdings stark eingeschränkt. Open DACHS bietet dagegen die Möglichkeit, eine ganze Website oder mehrere Teile davon zu archivieren sowie regelmäßig wiederkehrende Archivierungen einzurichten, die Pfadtiefe frei zu definieren, sowie Seiten des “Deep Web” oder dynamische Inhalte zu berücksichtigen.

SuLMaSS - Sustainable Lifecycle Management for Scientific Software

Axel Loewe[1], Gunnar Seemann[2], Eike Moritz Wülfers[2], Yung-lin Huang[2], Jorge Sánchez[1], Felix Bach[1], Robert Ulrich[1] and Michael Selzer[1]
[1]Karlsruhe Institute of Technology (KIT);
[2] Universitäts-Herzzentrum Freiburg

The SuLMaSS project [1] will advance, develop, build, evaluate, and test infrastructure for sustainable lifecycle management of scientific software. The infrastructure is tested and evaluated by an existing cardiac electrophysiology simulation software project, which is currently in the prototype state and will be advanced towards optimal usability and a large and active user community. Thus, SuLMaSS is focused on designing and implementing application-oriented e-research technologies and the impact is three-fold:

- Provision of a high quality, user-friendly cardiac electrophysiology simulation software package that accommodates attestable needs of the scientific community.
- Delivery of infrastructure components for testing, safe-keeping, referencing, and versioning of all phases of the lifecycle of scientific software.
- Serve as a best practice example for sustainable scientific software management.

Scientific software development in Germany and beyond shall benefit through both the aforementioned best practice role model and the advanced infrastructure that will, in part, be available for external projects as well. With adding value for the wider scientific cardiac electrophysiology community, the software will be available under an open source license and be provided for a large share of people and research groups that can potentially leverage computational cardiac modeling methods. Institutional infrastructure will be extended to explore, evaluate and establish the basis for research software development regarding testing, usage, maintenance and support. The cardiac electrophysiology simulator will drive and showcase the infrastructure formation, thus serving as a lighthouse project.

The developed infrastructure can be used by other scientific software projects in future and aims to support the full research lifecycle from exploration through conclusive analysis and publication, to archival, and sharing of data and source code, thus increasing the quality of research results. Moreover it will foster a community-based collaborative development and improve sustainability of research software.

SuLMaSS will provide a web platform to the community which integrates Gitlab and a question and answering system for collaboration. Providing the foundation for a modern software development, the system will be extended by a scientific test framework to

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI https://doi.org/10.11588/heidok.00026843 veröffentlicht.

Axel Loewe et al.: SuLMaSS - Sustainable Lifecycle Management for Scientific Software.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 238–239. DOI: https://doi.org/10.11588/heibooks.598.c8447

indicate unexpected changes in the outcome of the simulations and therefore ensure the correctness for each software build. Jointly with the software platform, related information like documentation, metadata, scientific setup of the simulation etc. will be extracted and joined into one package following the open archival information system model. This package will be moved to a long-time archive and will get a persistent identifier assigned for referencing. By adding context information to build a self contained software package, it will improve the verification and reusability independent from community platform as well as ease sharing of scientific software.

Literaturverzeichnis

[1] http://www.dfg.de/dfg_magazin/aus_der_wissenschaft/impulse_fuer_das_digitale_lis_jb17/02_aus_der_foerderung/index.html

heiMAP – Virtual Research Environment for collaborative spatio-temporal research in the Humanities

Martin Baumann, Dirk Eller, Vincent Heuveline, Mohammed Rizwan Khan, Lukas Loos, Leonhard Maylein, Jörg Peltzer, Michelle Pfeiffer, Benjamin Scherbaum, Kilian Schultes, Amon Veiga Santana, Armin Volkmann, Mohammed Zia and Alexander Zipf

Universität Heidelberg, Deutschland

heiMAP is a Virtual Research Environment (VRE) for research in the Humanities that facilitates collaborative work in spatio-temporal contexts. The differentiating characteristic of heiMAP is its holistic approach, representing the entire scientific data lifecycle. This includes the generation of, as well as discourse about, spatio-temporal data, their publication, archiving and sustainable reuse. The VRE consists of an Open Source Content Management System (CMS) that handles data and research projects, and an integrated Web Geographic Information Systems (GIS) application used to contextualize vector and raster data, especially maps. By sticking to international standards (OGC, CIDOC CRM among others), we strive for a maximum of data interoperability and reusability of the data produced within heiMAP. Close cooperation with the Heidelberg University Library and Heidelberg Computing Center allows for the publication, referenced by a Digital Object Identifier (DOI), as well as the long-term archiving of individual research projects and their related outcomes.

One major challenge in building an environment for collaborative research is to ensure that users retain control over their data where necessary but are also enabled to share it with others when they wish to. Therefore, heiMAP possesses a sophisticated system for user and rights management as well as fine-grained control of user rights based on the CMS. To allow a broad range of users from different parts of the Humanities (including both University and Citizen Science projects) with varying degrees of expertise in spatial research and GIS, to effectively use heiMAP, the platform focuses on a number of core functions while allowing export of research data into more sophisticated desktop-based GIS programs when necessary.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026845` veröffentlicht.

Martin Baumann et al.: heiMAP – Virtual Research Environment for collaborative spatio-temporal research in the Humanities.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 240. DOI: https://doi.org/10.11588/heibooks.598.c8448

Dokumentation von Forschungsprozessen mit dem RePlay-DH-Client

Sibylle Hermann[1], Uli Hahn[2], Markus Gärtner[2,3], Florian Fritze[1] und Volodymyr Kushnarenko[4]

[1]Universität Stuttgart, Universitätsbibliothek;
[2]Universität Ulm, Kommunikations- und Informationszentrum (kiz);
[3]Universität Stuttgart, Institut für maschinelle Sprachverarbeitung (IMS);
[4]Universität Ulm, Institut für Organisation und Management von Informationssystemen (OMI)

Im Projekt RePlay-DH wurde ein Werkzeug entwickelt, das Wissenschaftlerinnen und Wissenschaftler bei der nachhaltigen Dokumentation und Archivierung der Forschungsdaten bereits im Forschungsprozess unterstützt. Die einzelnen Prozessschritte werden meist nicht ausreichend dokumentiert, um selbige nachvollziehen zu können.

Die Kernidee liegt deshalb auf der Wiederauffindbarkeit und Nachvollziehbarkeit von Forschungsprozessen. Im Projekt wurden hierfür exemplarisch Workflows aus der Computerlinguistik analysiert, um praxisnahe Lösungen zu schaffen. Die Software wurde in einem iterativen Prozess gemeinsam mit Wissenschaftlerinnen und Wissenschaftlern entwickelt und die Ergebnisse einer begleitenden Nutzerstudie direkt umgesetzt. Mit der jetzt verfügbaren Software gibt es nun die Möglichkeit, den Arbeitsprozess der Forscher mit einem geringen Mehraufwand zu dokumentieren, mit Metadaten anzureichern und damit für Dritte nachvollziehbar und verfügbar zu machen. Der Client abstrahiert die Versionsverwaltungssoftware *Git* und ermöglicht dadurch intuitiv, den Workflow abzubilden und damit zu interagieren.

In komplexen Projekten soll dieses Vorgehen dabei helfen, verschiedene Optionen während des Forschungsprozesses auszuloten und zu dokumentieren. Dabei bleibt es stets den Wissenschaftlern überlassen, welche Teile des eigenen Projekts archiviert, mit Kollegen geteilt oder veröffentlicht werden sollen.

Sibylle Hermann et al.: Dokumentation von Forschungsprozessen mit dem RePlay-DH-Client.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 241. DOI: https://doi.org/10.11588/heibooks.598.c8449

Spacialist – Virtual Research Environment for the Spatial Humanities

Matthias Lang[1], Michael Derntl[1], Benjamin Glissmann[1], Vinzenz Rosenkranz[1] and Dirk Seidensticker[2]

[1]Universität Tübingen, Deutschland;
[2]Universität Gent, Belgien

Research projects in the humanities generate data and tools that are often abandoned after the project funding ends. Moreover, research data handling and the deployed tools are often highly specific for single projects. This insustainable practice leads to solutions that are incompatible with other tools, projects and infrastructures, and they often do not rely on accepted standards.

To close this gap the project Spacialist, which was funded by the Ministry of Science, Research and the Arts Baden-Württemberg in the "E-Science" program, set out to develop a modular virtual research environment that offers an integrated, web-based user interface to record, browse, analyze, and visualize all spatial, graphical, textual and statistical data from archaeological or cultural heritage research projects.

To address the highly heterogeneous requirements of such projects, Spacialist was developed as a software platform that is instantiated, customized and deployed separately for each project. The data model was designed as a meta model that defines entities with their properties and relationships. During the customization of the data model for a particular project, these abstract entities need to be instantiated for the project's domain. For representing domain-specific concepts Spacialist uses controlled vocabularies (thesauri) based on the XML-based standard SKOS (Simple Knowledge Organization System), thus facilitating data analysis and interoperability.

Core functionality such as the thesaurus and the creation and editing of entities is available out of the box for each project. Additional functionality is implemented in plugin modules that can be added on demand. These include file management, data analysis, geographical maps, and others.

The development of Spacialist's open-source software was driven by an interdisplinary team of software developers, geographers, ethnologists, archaeologists and librarians in collaboration with pilot projects in various areas like mediterranean archaeology and cultural heritage preservation.

To address the challenge of creating a sustainable business model beyond the initial funding, Spacialist was integrated into the service offered by the eScience-Center Tübingen, which has the necessary infrastructure and staff to provide Spacialist instances initially free to projects. The initial deployment and custom data model are covered by permanent staff. If the client project decides to adopt Spacialist as their research environment,

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026845` veröffentlicht.

Matthias Lang et al.: Spacialist – Virtual Research Environment for the Spatial Humanities.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 242. DOI: https://doi.org/10.11588/heibooks.598.c8450

the project is charged with a fee that covers hosting and maintenance of their Spacialist instance, and they have to enter a contractual agreement with eScience-Center defining usage and data privacy issues.

To support the full research project lifecycle even after the projects end, the platform is being integrated with our University's research-data archive, which guarantees the long-term availability and reusability of project data.

bwScienceToShare: Erschließung und Vernetzung von Forschungsdaten der Universitäten und Hochschulen in Baden-Württemberg

Saher Semaan
Albert-Ludwigs-Universität Freiburg, Deutschland

Aufgrund zahlreicher nationaler und internationaler Initiativen haben Forschungsdaten und ihre Bereitstellung ganz im Sinne von Open Science stark an Bedeutung gewonnen. Die Erschließung, die Sichtbarkeit und die Nachnutzbarkeit von Forschungsdaten zu verbessern ist ein wichtiges Ziel von bwScienceToShare. Das bwScienceToShare-Portal, ein zentrales Registrations- und Suchwerkzeug für die Metadaten von Forschungsdaten an den Universitäten und Hochschulen im Land Baden-Württemberg, kann dazu entscheidend beitragen.

Zusätzlich wird eine individualisierbare Publikationsplattform basierend auf FreiDok plus (Forschungsdokumentationssystem der Albert-Ludwigs-Universität Freiburg, `https://freidok.uni-freiburg.de`) für Einrichtungen des Landes, die über keine eigene Forschungsdateninfrastruktur verfügen, bereitgestellt.

Im bwScienceToShare-Portal (`https://bwsciencetoshare.ub.uni-freiburg.de`) werden Metadaten zu Forschungsdatensätzen automatisiert gebündelt und mit Hilfe einer modernen Suche sichtbar gemacht. Über eine Registrierungsschnittstelle, die zentral von der Universitätsbibliothek der Albert-Ludwigs-Universität Freiburg verwaltet wird, können weitere Repositorien über das Portal selbst (`https://bwsciencetoshare.ub.uni-freiburg.de/about`) vorgeschlagen und nach der Überprüfung durch die Universitätsbibliothek angemeldet werden, um ihre Metadaten über ihre OAI-PMH- Schnittstelle einzusammeln und im Portal durchsuchbar zu machen.

Die Publikationsplattform FreiDok plus der Universität Freiburg wurde ausgebaut und zu einer multidomänenfähigen Publikationsplattform weiterentwickelt. Das System steht für die Nachnutzung bereit. Dazu wurde das Deployment der Software vereinfacht und konfigurierbar gemacht. Eine Schlüsselfunktion der Publikationsplattform ist die Verwendung von Normdaten für Personen, Institutionen und Schlagwörter, die automatisiert mit den Angaben der Gemeinsamen Normdatei (GND) und ORCiD angereichert werden. Die strukturierten Metadaten und die Normdaten bilden die Grundlage für die Verknüpfung von Publikationen, Personen, Institutionen und Projekten innerhalb der Publikationsplattform sowie zu anderen Nachweissystemen. Zudem ermöglichen sie exakte statistische und bibliometrische Auswertungen.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026856` veröffentlicht.

Saher Semaan: bwScienceToShare: Erschließung und Vernetzung von Forschungsdaten der Universitäten und Hochschulen in Baden-Württemberg. In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge, Heidelberg: heiBOOKS, 2020, 244–245. DOI: https://doi.org/10.11588/heibooks.598.c8451

Die Verwendung von Open Source Software bei der Entwicklung und der stetige Ausbau der Publikationsplattform stellen ebenso wie die Langzeitarchivierung der Daten die Nachhaltigkeit des Systems sicher. Der Einsatz offener Standards garantiert die Nachnutzbarkeit der Metadaten.

Das Computational Science Lab Hohenheim

Vincent Dekker
Computational Science Lab, Universität Hohenheim, Deutschland

Das Computational Science Lab (CSL) ist eine Initiative von 16 Fachgebieten an der Universität Hohenheim, die im Bereich der Verarbeitung großer Datenmengen tätig sind. Die Fachgebiete kommen aus allen drei Fakultäten der Universität. Das Ziel der Initiative ist es, im engen Austausch über Methoden und Kompetenzen Synergieeffekte aus der transdisziplinären Forschung zu ziehen. Wesentliches Element der geplanten Zusammenarbeit ist der Umzug aller Fachgebiete unter ein gemeinsames Dach, in dem neben einem großzügigen Open-Space-Bereich, einem Kreativlabor und PC-Pool-Räumen auch Platz zum Austausch mit externen Wissenschaftlerinnen und Wissenschaftlern geschaffen werden soll. Unterstützt wird die Initiative auch vom Kommunikations-, Informations- und Medienzentrum (KIM) Hohenheim, welches ebenfalls zum Teil in das neue Gebäude einziehen wird. Aufgegliedert ist das CSL in die vier Bereiche:

- Mathematische und statistische Methoden der Datenanalyse,
- Verarbeitung und Analyse großer Datenmengen,
- Modellierung und Simulation komplexer Systeme, und
- Bioinformatics.

Speziell in diesen Bereichen sollen gemeinsame Forschungsprojekte über Fakultätsgrenzen hinweg durchgeführt sowie gemeinsame Lehrveranstaltungen konzipiert werden. Hierbei wird auch explizit die Nähe zum Core Facility Modul „Data and Statistical Consulting“ gesucht, das ein Beratungsangebot für Wissenschaftlerinnen und Wissenschaftler der Universität Hohenheim zum Thema statistische Datenanalyse anbietet.

Das CSL ist im Format eines *Labs* organisiert und damit quer zur an der Universität Hohenheim üblichen Organisationsstruktur von Fakultäten und Instituten konzipiert. Geleitet wird das Lab von einem gewählten Sprecher, der von Forschungsdirektoren aus den vier Themenbereichen des CSL sowie einem Verantwortlichen für das gemeinsame Lehrprogramm unterstützt wird.

Derzeit befindet sich das CSL noch im Aufbau und ist aktiv auf der Suche nach Kooperationspartnern. Das CSL ist dabei offen für jede Art von wissenschaftlicher Zusammenarbeit auf den oben aufgeführten Gebieten. Bedingt durch die unterschiedlichen Hintergründe der einzelnen beteiligten Fachgebiete, die sich aber sehr eng verwandter Methoden bedienen, will das CSL zur Lösung komplexer Probleme und Fragestellungen auf diesen

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026841` veröffentlicht.

Vincent Dekker: Das Computational Science Lab Hohenheim.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 246–247. DOI: https://doi.org/10.11588/heibooks.598.c8452

Gebieten beitragen, sowie neue, innovative Ansätze und Konzepte in der Theorie entwickeln und dann auch in der Praxis umsetzen.

Die E-Science Tage in Heidelberg sind eine ideale Gelegenheit, diese neue Initiative einem breiten und fachkundigen Publikum vorzustellen, Kontakte zu anderen Instituten und Initiativen zu knüpfen.

bwDIM - Data In Motion

Felix Bach und Robert Ulrich
Karlsruher Institut für Technologie (KIT), Deutschland

Im Landesprojekt bwDIM wurden Schnittstellen und Prozesse analysiert, modelliert und optimiert, welche für ein niederschwelliges, nutzerfreundliches Forschungsdatenmanagement (FDM) und das Verknüpfen einzelner Infrastrukturkomponenten bedeutsam sind. Das Ermöglichen effizienter Datenflüsse zwischen verschiedenen am FDM beteiligten Systemen und Infrastrukturen wird Wissenschaftler dabei unterstützen, den langfristigen Erhalt und den gesicherten Zugang zu ihren wissenschaftlichen Daten zu gewährleisten. Dabei wird auf allen Ebenen die Umsetzung der FAIR-Prinzipien und die Bereitstellung als Open Data gefördert.

Zur Unterstützung der Kommunikationsbeziehungen zwischen den unterschiedlichen am Forschungsdatenmanagement beteiligten Datenquellen und -zielen, wie z.B. Laboren, Archiv, Datenzentren, Repositorien sowie Systemen zur Nachnutzung und Verarbeitung, wurden Mechanismen für einen effizienten und teilautomatisierbaren Datenaustausch konzipiert. Hierbei kommen je nach Datenmenge unterschiedliche Verfahren zum Einsatz. Die spezifische Modellierung der Workflows soll es den Forschenden ermöglichen, in den eingesetzten Datenmanagementsystemen Zugriffe auf die Daten effizient und asynchron durchführen zu können, so dass der Zugang zu großen Datenmengen für einzelne Forschende auch ohne detaillierte Kenntnisse über komplexe technische Infrastrukturen möglich wird und diese leicht in den wissenschaftlichen Publikationsprozess eingebunden werden können.

Die Anbindung der FDM-Systeme an die föderierten IDM-Systeme wie z.B. bwIDM und DFN-AAI alleine genügt nicht den speziellen Anforderungen im Kontext Forschungsdatenmanagement. In Hinblick auf die Lebensdauer von Nutzeraccounts an Universitäten und den Zugriff über Föderationsgrenzen hinweg, wie etwa in internationalen Forschungsprojekten, ist ein einfaches und einheitliches Nutzermanagement zu gewährleisten. Daher wurden international verbreitete Ansätze für eine zentrale Identitätsverwaltung wie z.B. diejenigen von ORCID untersucht, die versprechen, auch über Ländergrenzen und Organisationszugehörigkeiten hinweg langfristig zu funktionieren. Das Archivsystem bwDataArchive wurde dazu exemplarisch an ORCID angebunden.

Des Weiteren wurden in Zusammenarbeit mit KITopen die nötigen Grundlagen für die Verknüpfung wissenschaftlicher Artikel mit Forschungsdaten erarbeitet und die nötigen technischen Voraussetzungen für den Austausch von Zugangsinformationen und Metadaten mit Publikationssystemen skizziert. Das Konzept wurde in Zusammenarbeit mit Chemotion und Beilstein weiter ausgearbeitet mit dem Ziel, Reviewern und Journals nicht nur die für das Review benötigten Forschungs- und Metadaten bereitzustellen, sondern

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026842` veröffentlicht.

Felix Bach und Robert Ulrich: bwDIM - Data In Motion.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 248–249. DOI: https://doi.org/10.11588/heibooks.598.c8453

diese mit dem Repositorium und der Fachcommunity zu vernetzen. Durch die engere Verknüpfung mit der wissenschaftlichen Plattform soll die Interpretation der Daten durch die Reviewer erleichtert und die Qualität des Reviewprozesses verbessert werden.