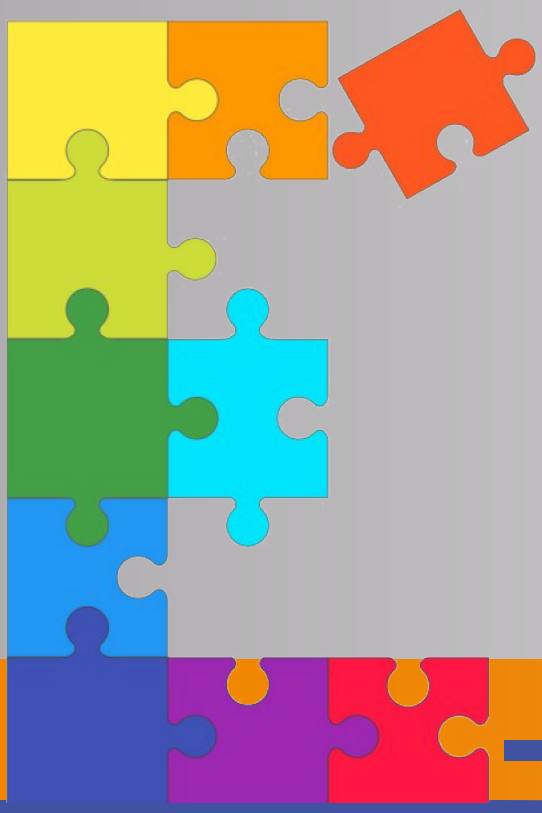


Vincent Heuveline
Nina Bisheh
(Hrsg.)

-Science-Tage 2021

Share Your Research Data

UNIVERSITÄTS-
BIBLIOTHEK
HEIDELBERG

E-Science-Tage 2021

E-Science-Tage 2021

Share Your Research Data

Herausgegeben von

Vincent Heuveline und Nina Bisheh

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 2022.

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-979-9
doi: https://doi.org/10.11588/heibooks.979

ISBN 978-3-948083-55-7 (Softcover)
ISBN 978-3-948083-54-0 (PDF)

Inhaltsverzeichnis

Teil I

Grußworte

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

Theresia Bauer

Daten sind seit jeher eine der tragenden Säulen der wissenschaftlichen Erkenntnis. Mit der digitalen Transformation der Wissenschaft sind sie sowohl in Umfang, Qualität wie auch Wichtigkeit nochmals enorm gewachsen. In der Corona-Pandemie ist die Bedeutung von wissenschaftlichen Datenbeständen wie noch nie zuvor in den Fokus der Öffentlichkeit gerückt – die Leistungsfähigkeit datenorientierter Forschung war gewissermaßen live für alle zu verfolgen.

Die Bedeutung von Forschungsdaten hat in letzter Zeit aber nicht nur wegen der Pandemie immens zugenommen. Alle Zukunftsfelder wie das Maschinelle Lernen und die Künstliche Intelligenz sind auf entsprechende Datensammlungen besonders angewiesen. Deshalb haben Wissenschaftsministerium und Land Baden-Württemberg dieser Entwicklung in den letzten zehn Jahren in besonderem Maße Rechnung getragen. Wir haben in den kontinuierlichen Auf- und Ausbau gemeinsam betriebener und genutzter Forschungsdateninfrastrukturen an den Universitäten und Forschungseinrichtungen des Landes investiert, denn nur so können wir optimale Rahmenbedingungen für eine exzellente Hochschul- und Forschungslandschaft in Baden-Württemberg gewährleisten.

Dies alles gelingt im kooperativen Agieren am besten – ein „Narr auf eigne Hand“, wie es Goethe formulierte, wird in der digital geprägten Wissenschaft keinen Erfolg haben. Diese macht auch nicht an der Bundeslandgrenze halt, weswegen es wichtig ist, die verschiedenen Initiativen auf Landes-, nationaler und internationaler Ebene gemeinsam zusammenzudenken und zu harmonisieren.

Vorhaben wie die Science Data Centers in Baden-Württemberg, die Nationale Forschungsdateninfrastruktur (NFDI) oder die European Open Science Cloud (EOSC) können und dürfen nicht unabhängig voneinander agieren. Erst das Zusammenwirken macht schnelle und skalierbare Fortschritte möglich. Wir haben mit unserer im Land gepflegten Kooperationskultur die besten Erfahrungen gemacht und öffnen uns jederzeit der übergreifenden Zusammenarbeit. Deshalb ist es auch genau richtig, dass die diesjährigen E-Science-Tage unter dem Motto „Share Your Research Data“ stehen. In der Datenagenda BW haben wir festgelegt, dass wir uns am Open Data-Prinzip orientieren, denn mit dem Teilen von Forschungsdaten sind viele Chancen verbunden. So werden Diskurse beflügelt, Erkenntnisgewinne erhöht und die wissenschaftliche Qualität gesichert.

Aber Sie alle wissen am besten: Der geregelte Austausch von Daten ist nicht trivial - technische Fragen etwa der Interoperationalität, Langzeitarchivierung oder Zugriffsverwaltung sind knifflig. Disziplinübergreifender Austausch ergibt sich nicht von allein, sondern es braucht Expertise, Beratung und vor allem viel Initiative.

Es freut mich daher sehr, dass Sie sich als Anwenderinnen und Anwender mit den Informations- und Infrastruktureinrichtungen gemeinsam den Herausforderungen stellen und wünsche Ihnen allen viele anregende Diskussionen zur Zukunft des Managements von Forschungsdaten!

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

Grußwort des Rektors der Universität Heidelberg

Bernhard Eitel

Lieber Herr Heuveline, liebe Kolleginnen und Kollegen,

zu den E-Science Tagen 2021 begrüße ich Sie im Namen der Universität Heidelberg sehr herzlich. Dieses Jahr steht ganz im Zeichen der Covid-19 Pandemie, E-Science gewinnt damit eine besondere Bedeutung. Sie können sich NICHT physisch in Heidelberg treffen, nicht oder nur erschwert informell und ganz zufällig austauschen, nicht den tagungstypischen „Flurfunk“ abhören, nicht abends die Altstadt und ihr besonderes Flair erkunden und ganz nebenbei die schönen Seiten des Wissenschaftlerlebens auskosten… E-Science auf ganz digitalem Grund im virtuellen Raum!

Und dennoch: Die Organisatoren haben zwei Tage Programm auf die Beine gestellt und ein aktuelles Thema wie das Forschungsdatenmanagement in den Mittelpunkt gerückt. Forschungsdaten sind ein Rohstoff, der besondere Aufmerksamkeit verdient, sei es aus Sicht der Datenqualität bis hin zur Sicherung guter wissenschaftlicher Praxis, der Datensicherung, -kompatibilität und -speicherung bis hin zur Verwendung im Rahmen des machine learning oder von KI-Anwendungen. Außerdem rege ich an, verstärkt darüber nachzudenken, wie es eigentlich mit den IP-Rechten an den Daten aussieht, wie weit diese bei der Weiterverwendung reichen usw..

Sie sehen: Auch das diesjährige Kernthema und die begleitenden überwiegend virtuellen Treffen sind vielfältig, komplex und ragen weit über die Mathematik, die Informatik und das Wissenschaftliche Rechnen hinaus. Ökonomische, rechtliche, ethische und fachwissenschaftliche Aspekte durchdringen sich beim Forschungsdatenmanagement gegenseitig. E-Science als Querschnittsdisziplin und methodologisches Forschungsfeld ist aktueller denn je – auch in Heidelberg. Daher ist dieses zweitägige Treffen wiederum an der Universität Heidelberg bestens verortet, die sich ja als Comprehensive Research University versteht.

Ich wünsche Ihnen viel Erfolg, gerade unter den etwas gewöhnungsbedürftigen Bedingungen, die derzeit herrschen. Forschung ist die Grundlage für eine gute Zukunft, mit Erkenntnissen prägen wir ihren Verlauf und übernehmen gesellschaftliche Verantwortung.

Ich danke allen Organisatoren für ihren Einsatz und wünsche Ihnen, dass die in dieses virtuelle Treffen gesetzten Erwartungen erfüllt werden. Nehmen Sie die jungen Nachwuchswissenschaftlerinnen und –wissenschaftler mit, denn diese haben es besonders schwer, sich in die digitalen Netzwerke einzuklicken. Sie werden sehen, beide Seiten werden daraus ihren Vorteil ziehen.

Herzlich heiße ich Sie alle virtuell an der Universität Heidelberg willkommen!

Ihr

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

Vorwort der Herausgeber

Vincent Heuveline und Nina Bisheh

Seit Beginn der letzten Dekade vollziehen etliche Forschungsgebiete einen Paradigmenwechsel durch die Verarbeitung von großen Datenmengen, auf Basis derer neue wissenschaftliche Erkenntnisse gewonnen werden sollen. Dabei geht es nicht nur darum, Rechenkapazitäten bereitzustellen, mit welchen solche Datenmengen effizient analysiert und dadurch sinnvoll verwendet werden können. Ein weiteres Ziel ist es auch, Prozesse zu etablieren, mithilfe derer ein Datenaustausch samt Qualitätssicherung sowohl innerhalb der Communities als auch interdisziplinär umgesetzt werden kann. Inzwischen gehören all diese Aspekte für viele Forschende der guten wissenschaftlichen Praxis an und sind in unserer digitalen Zeit für das wissenschaftliche Arbeiten nicht mehr wegzudenken.

Die E-Science-Tage 2021 und der damit entstandene Tagungsband zielen darauf ab, die Herausforderungen rund um „Share your Research Data“ fachübergreifend von der Konzeptebene hin zur konkreten Umsetzung in den jeweiligen Communities zu beleuchten. Insbesondere soll die Brücke zwischen dem Nutzen des Datenaustausches und den damit verbundenen Risiken geschlagen werden.

Mit Freude haben wir wahrgenommen, dass diese E-Science-Tage einen Rahmen geboten haben, in dem äußerst interessante, kurzweilige Beiträge und Diskussion entstehen konnten. Wir hoffen, dass der Tagungsband die Vielfalt der in diesem Rahmen gewonnenen Erkenntnisse widerspiegelt und bedanken uns sowohl beim Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg als auch bei allen Beteiligten, die eine lebhafte und gewinnbringende Veranstaltung ermöglicht haben.

Unser besonderer Dank gilt weiterhin den überaus engagierten Autorinnen und Autoren, die durch Ihre qualitativ anspruchsvollen Beiträge zur Entstehung dieses außerordentlichen Tagungsbandes beigetragen haben. Zuletzt soll hervorgehoben werden, dass die E-Science-Tage 2021 auch einen Beitrag zur Community-Bildung rund um Forschungsdaten geleistet haben. Es wird angestrebt, diese Entwicklung im Sinne einer offenen Austauschplattform zukünftig im Rahmen des inzwischen gut etablierten Formats E-Science-Tage fortzusetzen.

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

Nina Bisheh, M.Sc.
Organisatorin der E-Science-Tage 2021 Konferenz

Teil II

Wissenschaftliche Beiträge

Schulungskonzept zu Git und GitLab – Gamification zum besseren Lernen

Patrick Jüptner[1], Manuela Dalibor[2], Ute Trautwein-Bruns[1] und Bernhard Rumpe[2]

[1]Universitätsbibliothek, RWTH Aachen University;
[2]Software Engineering, RWTH Aachen University

Im Zuge der Digitalisierung wird der Umgang mit textbasierten Datenformaten aber auch die Nutzung von Codeschnipseln zur Datenerhebung und Analyse zu einer grundlegenden Datenkompetenz für Wissenschaftlerinnen und Wissenschaftler, auch in Fachbereichen, die bisher als weniger IT-affin galten. Git und GitLab sind gut etablierte Werkzeuge zur Entwicklung, Dokumentation, Zusammenarbeit und Veröffentlichung von Softwareprojekten und werden zunehmend auch im Kontext des Forschungsdatenmanagements eingesetzt, da sie sich gut zur Versionsverwaltung von textbasierten Forschungsdaten und zur Unterstützung der Zusammenarbeit und des „Data Sharings“ eignen.

Um auch Nicht-Informatikern den Einstieg in diese Tools zu ermöglichen und die Motivation während des Lernprozesses aufrecht zu halten, setzen wir auf das Konzept der Gamification. In einer spielerischen Umgebung werden die Grundlagen der Versionsverwaltung, essentielle Git Befehle und der Umgang mit GitLab trainiert, was die Motivation der Lernenden steigert und sie zur Mitarbeit anspornt. Aus diesem Ansatz sind zwei Konzepte entwickelt worden, die über einen spielerischen Weg den Umgang und die nötigen Kompetenzen zur Verwendung von Git und GitLab vermitteln.

1 Einleitung

Im Bereich des Forschungsdatenmanagements wird der Umgang mit textbasierten Datenformaten und Codeschnipseln zu einer grundlegenden Datenkompetenz für Wissenschaftlerinnen und Wissenschaftler. Die Zusammenarbeit und das Teilen solcher Daten sind dabei eines der Kernthemen. Git und GitLab sind bereits etablierte Werkzeuge zur Entwicklung, Dokumentation, Zusammenarbeit und Veröffentlichung von Softwareprojekten [1], welche in zunehmendem Maße auch im Kontext des Forschungsdatenmanagements eingesetzt werden [2], da sie sich gut zur Versionsverwaltung von textbasierten Forschungsdaten, wie z.B. Tabellen im csv-Format, und zur Unterstützung der Zusammenarbeit und des „Data Sharings“ eignen. Weil dies auch weniger IT-affine Fachbereiche betrifft, sind Schulungskonzepte erforderlich, die auf einfache Art und Weise den Einstieg in diese Tools ermöglichen.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13715 (CC BY 4.0)

Aus diesem Grund und um die Motivation im Lernprozess aufrecht zu halten, haben wir uns für das Konzept der Gamification entschieden [3]. Kern von Gamification ist die Anwendung von Spielkonzepten bzw. spieltypischen Elementen in einem spielfremden Kontext. Mit der „Git Scavenger Hunt“ und „WordGuess“ wurden zwei Spielkonzepte entwickelt, in denen in spielerischer Umgebung die Grundlagen der Versionsverwaltung, essentielle Git Befehle und der Umgang mit GitLab geübt werden. Dies soll die Motivation des Lernenden steigern und zur Mitarbeit anspornen [4]. Darüberhinaus liegt im WordGuess-Konzept der Fokus auf GitLab als Projektmanagement-Tool. Für den konzeptionellen Rahmen setzen wir auf Scrum als Projektmanagement-Methode, welche hauptsächlich in der agilen Softwareentwicklung eingesetzt wird [5]. Die Rollen und Phasen des Scrum-Modells werden auf die Spielumgebung transferiert und steuern den Spielablauf. So erhalten die Lernenden als Add-On einen oftmals ersten Einblick in ein Projektmanagement-Modell und können sich spielerisch mit diesem vertraut machen.

Die dargestellten Konzepte sind Teil eines Schulungsprogramms zu Git und GitLab, das sich aus einem Einstiegskurs zu Git und Gitlab (elearning), dem Training in den Lernspielen und einem Präsenzworkshop zusammensetzt. Der Moodle-Kurs „Einstieg in die Versionsverwaltung mit Git und GitLab“ führt in die Thematik der Versionsverwaltung mit Git und des Projektmanagements mit Gitlab ein und bietet Tutorials, um die entsprechende Arbeitsumgebung zu installieren und erste Grundbefehle kennenzulernen. Der Präsenzworkshop schließlich dient neben der vertiefenden Übung der Prozesse dem interaktiven Austausch mit Kollegen und der Diskussion von Anwendungsbeispielen.

Initiert wurde die Entwicklung der Lernspiele aufgrund der großen Nachfrage nach Git/GitLab-Schulungen an der RWTH Aachen University auch für individuelle Arbeitsgruppen, dem Wegfall von Präsenzkursen während der Corona-Pandemie und der Beobachtung, dass das Niveau der Gitkompetenzen im Workshop stark variierten und komplette Git-Neulinge davon profitieren würden, ein wenig Training der Befehle und Prozesse vorab zu haben.

2 Git Scavenger Hunt - eine Schnitzeljagd mit Git

2.1 Kursaufbau

In Git Scavenger Hunt gilt es für den Spieler sieben Level mit ansteigender Schwierigkeit zu lösen, um die Schnitzeljagd erfolgreich abzuschließen. Das vorrangige Lernziel dieses Konzeptes ist es, das Verständnis dafür zu erhöhen, was ein Git Repository ist, wie in diesem navigiert wird und wie auf frühere Versionen zugegriffen werden kann.

Git Scavanger Hunt ist eine Einzelspieler-Erfahrung über sieben Level, d.h. jeder Teilnehmende löst die Schnitzeljagd für sich auf dem eigenen Rechner. Die komplette Git Scavanger Hunt-Spielumgebung befindet sich in einem öffentlichen GitLab-Repository (`https://git.rwth-aachen.de/patrick.jueptner/git-hunt-lvl1-7`). Dieses wird zu Beginn von jedem Spieler lokal auf den eigenen Rechner geklont. Das komplette Spiel

findet in diesem lokalen Klon über die Kommandozeile (Terminal) statt. Dies ermöglicht ein direktes reagieren auf Meldungen von Git und fördert so das grundsätzliche Verständnis für die durchgeführten Operationen und Befehle [6]. Nachfolgend werden sowohl das Repository als auch die weiteren Bestandteile des Schulungkonzeptes beschrieben.

Den Einstieg in den Kurs bildet ein RWTHmoodle Raum (Abb. 1) [7], in dem Anleitungen und Tutorials bereitstehen, welche sämtliche benötigen Befehle und Konzepte umfassen, die zur Lösung der Schnitzeljagd - vollständig in der Umgebung der Kommandozeile - benötigt werden. Dies sind neben den verwendeten Git-Befehlen auch die grundlegenden Kommandos für den Umgang mit der Kommandozeile, wie zum Anzeigen von Ordner-/Verzeichnisinhalten, dem Wechseln von Verzeichnissen oder dem Erstellen von neuen Verzeichnissen. Außerdem wird eine Möglichkeit zum Anzeigen von Inhalten einer Textdatei direkt in der Konsole gezeigt. Da Git, z.B. bei einem Merge nach der „Recursive Strategy“, automatisch einen Texteditor für die Eingabe einer Commit-Message startet, wird auch der grundlegende Umgang mit dem Texteditor vim [8] erklärt.

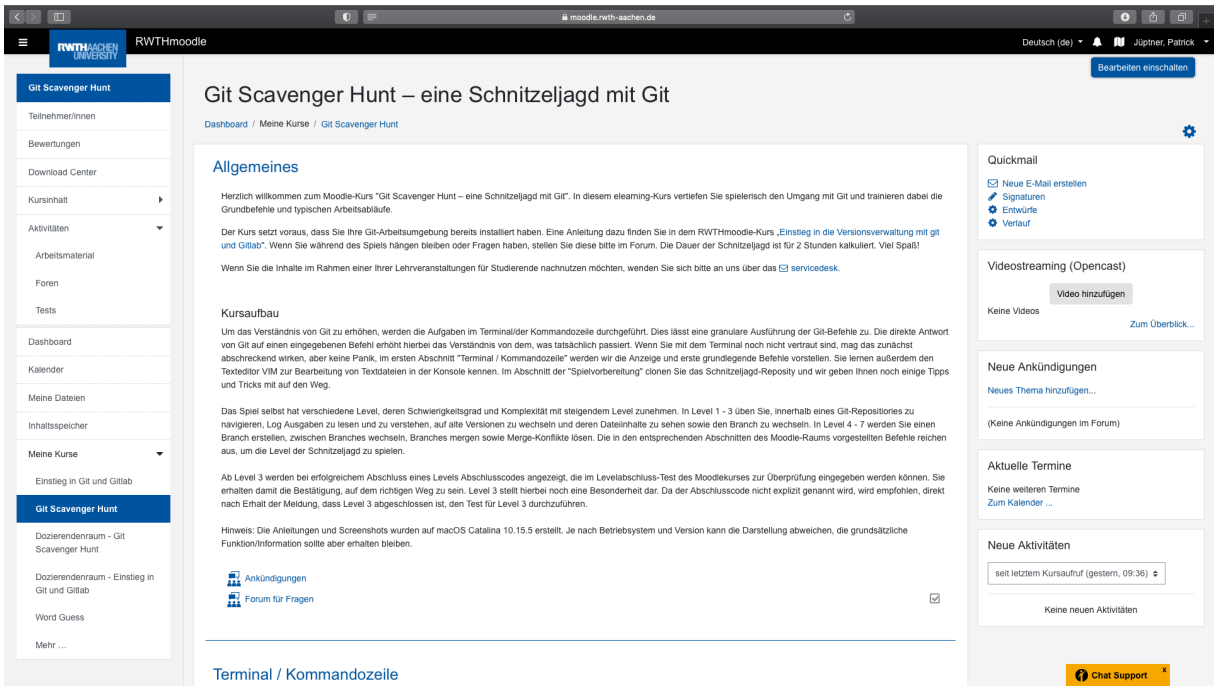


Abbildung 1: Startseite des Moodle Raums Git Scavenger Hunt.

Der Moodle Raum verfügt über ein Forum für Fragen und den Austausch der Teilnehmenden untereinander. Hier können die Spieler Fragen zum Spiel oder den verwendeten Befehlen stellen und sich auch gegenseitig beantworten. Die Kursbetreuer beobachten und moderieren das Forum und beantworten die Fragen nach Bedarf.

Für die Level 3 bis 7 bietet der Moodle Raum die Möglichkeit, Lösungscodes einzugeben, um den korrekten Abschluss eines Levels zu überprüfen und sich nicht in einer Sackgasse zu verlaufen. Die Lösungscodes erhalten die Spieler jeweils zum Abschluss eines Levels. Über den Levelabschluss-Test kann zudem überprüft werden, welche Teilnehmer wie weit

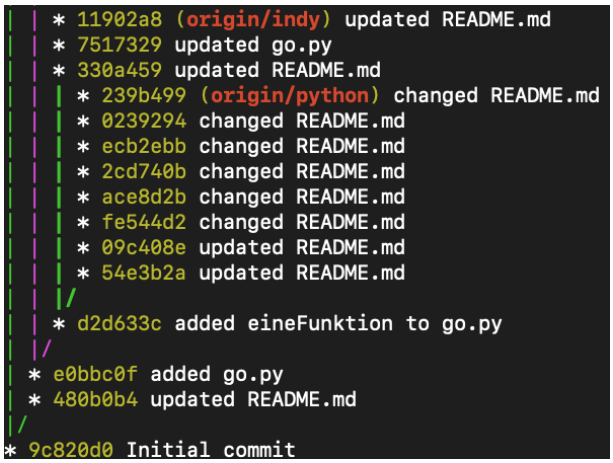


Abbildung 2: Ausschnitt des Git Scavenger Hunt Repositories mittels git log –graph –oneline –all.

im Spiel gekommen sind, um gegebenenfalls individuell auf spezielle Probleme in einem bestimmten Level zu reagieren. Um die Schnitzeljagd lösen zu können, müssen die Git Befehle log, checkout, merge und branch angewendet werden.

Über den Befehl *git log* läßt sich die Commit-Historie eines Repositories anzeigen. In seiner einfachsten Form listet der Befehl alle Commits, inklusive Commit-Message, Commit-ID, Identität der Person, die den Commit durchgeführt hat, und weiteren Informationen auf. Damit lassen sich die Änderungen der Dateien in einem Repository verfolgen.

Mittels *git checkout* wird zwischen Commits und Branches gewechselt. Dies ist im Spiel notwendig, um die benötigten Hinweise zu finden und die gestellten Aufgaben zu lösen. Der Befehl ermöglicht es dem Nutzer, durch das Repository zu navigieren und durch den Wechsel in ältere Commits auf ältere Versionen von Dateien zuzugreifen. Die Übung mit dem Befehl verdeutlicht die Funktionsweise der Versionierung innerhalb von Git und wie der Zugriff und die Weiterarbeit mit älteren Dateiversionen abläuft.

Der Befehl *git merge* wird verwendet, um zwei Entwicklungszweige (Branches) in Git zusammenzuführen. In der Softwareentwicklung ist es überlich, dass neue Features in Softwareprojekten in einem eigenen Branch entwickelt und getestet werden, bevor diese mit dem Features dann über einen Merge in den Master-Branch einfließen.

git branch ist dafür zuständig, einen neuen Branch zu erstellen. Dies geschieht aus dem aktuellen Commit heraus, in dem sich der Anwender befindet, oder über die Angabe einer Commit-ID, von welcher aus der neue Branch erstellt werden soll. Auf dem kurzen Ausschnitt des Repsitories in Abb. 2 sind einige Branches und die Commits, aus denen sie erstellt wurden, zu erkennen.

2.2 Lernziele

Mit der Git Scavenger Hunt lernen die Teilnehmenden, grundlegende Kommandos auf der Kommandozeile kennen und können durch Ordnersysteme auf ihrem Rechner navigieren. Die Teilnehmende verstehen wie ein Git-Repository aufgebaut ist und können mit den Git Befehlen log, checkout, branch und merge umgehen. Sie können die Log-Ausgaben von Git lesen und verstehen diese. Darüberhinaus können sie Branches erstellen und wechseln sowie Merge-Konflikte lösen.

2.3 Spielablauf

Der Spielablauf ist für jedes Level der Schnitzeljagd identisch (Abb. 3). Jedes Level startet in einem Commit.
Beim Start des Spiels ist dies der aktuellste Commit im Master Branch, der beim Klonen eines Git Repositories standardmäßig aktiv ist. Für die weiteren Level erhält der Spieler jeweils einen Hinweis, in welchem Commit das nächste Level startet. Ist der Commit gefunden, gilt es stets die README.md Datei zu lesen. Üblicherweise enthält die README-Datei Informationen über das Projekt oder die Software, zu der die Datei gehört. Sie informiert über wichtige Details und sollte vom Benutzer grundsätzlich gelesen werden. In diesem Spiel enthält die README-Datei die Hinweise bzw. die Aufgabenstellung zum Level. Diese Aufgaben reichen vom Finden von Informationen in Commit-Nachrichten bis hin zur Durchführung von Merge-Operationen auf Branches des Repositories. Die Aufgaben sind im Schwierigkeitsgrad ansteigend. Im Falle der Merge-Aufgaben muss z.B. erst ein „Automatic Merge“ durchgeführt werden, den Git selbstständig vollziehen kann. Im späteren Verlauf müssen Merge Konflikte gelöst werden und diese mit einer entsprechenden Commit-Message ins Repository aufgenommen werden. Die korrekte Lösung der Aufgaben ergibt sich dabei entweder direkt aus dem Dateiinhalt oder aus Textausgaben auf dem Bildschirm oder sie wird in einigen Leveln über ein Skript geprüft, welches der Spieler ausführen muss. Sollte ein Skript zur korrekten Überprüfung der Lösung zur Anwendung kommen, wird der Spieler in der Aufgabenstellung schon darauf hingewiesen. Wurde die Aufgabe erfolgreich gelöst, erhält der Spieler in den Leveln 3 bis 7 einen Levelsabschlusscode zur Eingabe im Moodle Raum.
Dieser Code kann eine bestimmte Zeichenfolge sein, oder aber Teil einer Commit-Nachricht (z.B. Datum), zu der der Spieler in der Schnitzeljagd geführt wurde. Außerdem erhält der Spieler einen Hinweis, in welchem Commit das nächste Level startet. Dieser Hinweis ist dabei oftmals kein direkter Verweis auf den gesuchten Commit, sondern über ein Rätsel zu entschlüsseln. Hierbei kann und muss auf Google oder eine andere Suchmaschine der Wahl bzw. das Internet im allgemeinen zurückgegriffen werden, um an die benötigen Informationen zu gelangen und so den Weg zum richtigen Commit zu finden.

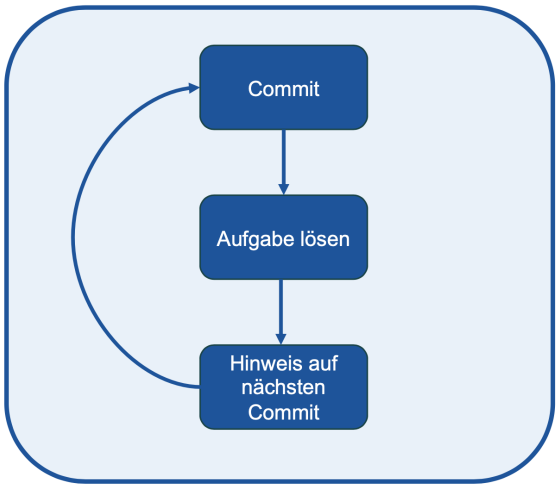


Abbildung 3: Allgemeiner Ablauf der Git Scavanger Hunt.

2.4 Beispiellevel

Bei Abschluss der vorausgehenden Levels hat der Spieler den Hinweis erhalten, dass es im Branch Caesar weitergeht. Der erste Schritt besteht für den Spieler darin, in den Branch Caesar zu wechseln.

Befehle sind im folgenden Fett dargestellt, die folgende Bildschrimausgabe ist nicht-fett.

```
> git checkout Caesar
Your branch is up to date with 'origin/Caesar'.
```

Im nächsten Schritt ist die Date README.md zu lesen, sie enthält die Aufgabenstellung und Hinweise.

```
> cat README.md
Hallo, ich bin Caesar und du hast den richtigen Branch
↪  ausgecheckt!
In diesem Branch werde ich eines meiner Meisterwerke
↪  schreiben, jedenfalls Teile davon.

Es wird der Gallische Krieg heißen und sehr
↪  wahrscheinlich mehrere Bücher umfassen.
In diesem Branch (Caesar) habe ich mein Werk mit Buch
↪  1 begonnen.
Buch 2 habe ich in einem eigenen Branch geschrieben.

Deine Aufgabe:
merge den Branch buch_2, in welchem ich Buch 2
↪  vollendet habe, in diesen Branch.
```

```
Den Befehl dazu solltest du bereits kennen. Eine
 ↪  Kleinigkeit wird diesmal aber anders sein, da
 ↪  diesmal kein automatic/fast forward Merge
 ↪  durchgeführt werden kann, sondern
nur ein Merge nach der recursive Strategy.
Der Merge muss als Commit in das Repository
 ↪  aufgenommen werden und erfordert auch eine Commit
 ↪  Message.
Aber keine Sorge, das meiste davon erledigt Git für
 ↪  dich von allein. Ich denke, dass du das hin
 ↪  bekommst!
Nach dem Merge musst du erneut das Skript checker.sh
 ↪  aufrufen, diesmal aber ohne Parameter. Der Befehl
 ↪  lautet:

./checker.sh

Hast du alles richtig gemacht, erhälst du den nächsten
 ↪  Hinweis!
```

Als nächstes muss der Merge durchgeführt werden.

```
> git merge buch_2
Merge made by the 'recursive' strategy.
Buch2.txt | 9 +++++++++
1 file changed, 9 insertions(+)
create mode 100644 Buch2.txt
```

Nach Ausführung des merge Befehls, muss eine Commit Message über den Texteditor eingegebn werden, bevor der gezeigte Merge durchgeführt wird. Die Überprüfung findet über das Skript checker.sh statt.

```
>./checker.sh

Herzlichen Glückwunsch! Der Merge hat geklappt!
Damit hast du Level 5 abgeschlossen!
Der Levelabschlusscode für Level 5 lautet: 9gt2b7
Gib diesen Code im Moodle Raum ein!
Weiter geht es in einem anderen Branch.
Finde heraus welcher Branch gesucht ist und checke ihn
 ↪  aus.
Der Hinweis:
Trfhpug vfg qre Anpuanzr rvarf Znaarf, qre fvpu vz
 ↪  Cevamvc nhpu zvg Mrvpuraxbqvrehat orfpunrsgvtg
 ↪  ung. Re jheqr 1791 trobera haq vfg 1872 trfgbeora.
```

```
Qvr iba vuz haq rvarz
Xbyyrtra reshaqrar Mrvpuraxbqvrehat jheqr nhpu Ynaq
↪ Yvar Pbqr tranaag. Purpxr qra Oenapu nhf, qrffra
↪ Anzr qrz Anpuanzra qre trfhpugra Crefba ragfcevpug!
```

Der Level ist erfolgreich beendet worden, der Levelabschlusscode und der Hinweis auf den nächsten Level sind dem Spieler angezeigt worden. In diesem Fall ist der Hinweis ein verschlüsselter Text, der dekodiert werden muss. Da das Spiel derzeit im Caesar Branch ist, ist der Code mittels Caesar Verschlüsselung kodiert [9]. Zur Dekodierung muss noch der Verschiebungsfaktor gefunden werden. Diesen findet der Spieler in der README.md Datei in einem früheren Commit des Branches. Hat der Spieler den Text dekodiert, kann er in den genannten Branch wechseln und das Spiel dort fortsetzen.

3 WordGuess – ein kollaboratives Spiel mit Git und GitLab

3.1 Kursaufbau

In WordGuess wird auf spielerische Art und Weise die Arbeit mit Git und GitLab trainiert und gleichzeitig ein Fokus auf das Projektmanagement entsprechend dem Scrum-Modell gesetzt. WordGuess ist dabei als kooperative Mehrspielererfahrung ausgelegt, in der die Teilnehmenden in Gruppen von idealerweise fünf Personen gemeinsam in einem GitLab-Repository als Spielumgebung spielen. Im Spielverlauf werden die Phasen und Rollen von Scrum auf die Spielphasen übertragen und mit einem Issue-Board in GitLab abgebildet, um so den typischen Ablauf einer per Scrum geführten Projektentwicklung mit GitLab nachzustellen. Ziel des Spiels ist es, eine Spielrunde entsprechend eines Sprints erfolgreich abzuschließen, indem ein Suchwort erraten wird. Die Spielidee orientiert sich dabei an JUST ONE, dem Spiel des Jahres 2019. Die Herausforderung besteht nun aber darin, dass die Spielenden nicht zusammen an einem Spieltisch sitzen, sondern verteilt an Ihren Schreibtischen. Die notwendige Kommunikation zum Spiel findet deshalb ausschließlich über Kommentare in GitLab-Issues statt. Aber auch die typischen Arbeitsabläufe im Zusammenspiel von Git und GitLab bei der gemeinsamen Bearbeitung von Dateien werden im Spiel trainiert, da es im Spielverlauf erforderlich ist, gemeinsam an einer Textdatei, dem „Sprint Backlog“ zu arbeiten. Dazu müssen die Spielenden das GitLab-Repository lokal auf den Rechner klonen, die Datei verändern und wieder an das Remote-Repository in GitLab übertragen. Das GitLab-Wiki wird zur Projektdokumentation eingesetzt. Außerdem finden die Teilnehmenden dort auch die kondensierten Informationen zum Spielablauf.

Den Einstieg in den Kurs bildet wieder ein RWTHmoodle Raum [10]. Dort stehen den Teilnehmenden alle Tutorials und Anleitungen bereit, die für das Spiel benötigt werden. Der Fokus dieses Kurses liegt auf GitLab und der Vermittlung von Basiswissen zum Projektmanagement. Die benötigten Git-Kenntnisse werden bereits mit der Git-Schnitzeljagd abgedeckt, so dass sie in diesem Kurs lediglich mit einem Quiz aufgefrischt werden, das gleichzeitig dem Selbsttest der Teilnehmenden dient, ob die Git-Kenntnisse für das Spiel

ausreichen. Die Einteilung in Spielgruppen erfolgt durch Selbstzuordnung in der Gruppenverwaltung von Moodle. Zur initiale Rollenverteilung dient ein Gruppenforum. Neben diesem beinhaltet der Moodle-Raum auch ein Standardforum für generelle Fragen zu den Kursinhalten.

3.2 Lernziele

Nach Absolvieren des WordGuess-Kurses verstehen die Teilnehmenden, wie GitLab die Zusammenarbeit an (textbasierten) Forschungsdaten unterstützt und können Projektmanagementkonzepte mit GitLab umsetzen. Sie kennen mit dem GitLab-Wiki eine Möglichkeit der Dokumentation von Forschungsprojekten und Entwicklungsprozessen und können Wikiseiten erstellen und bearbeiten. Die Teilnehmenden verstehen die verschiedenen Ebenen von Git/GitLab (workspace-local Repository–Remote Repository) und haben den grundsätzlichen Git/GitLab-Workflow mit „add-commit-push“ vertieft. Mit den Befehlen „push“ und „pull“ können sie zwischen lokalem und remote Repository kommunizieren. Sie können Branches erstellen und wechseln sowie Merge-Konflikte auflösen.

Abbildung 4: Ablauf einer WordGuess Runde.

3.3 Umsetzung der Spielidee

In WordGuess nehmen die Spielenden die verschiedenen Rollen nach dem Scrum-Modell ein. Es gibt einen Product Owner, einen Scrum Master, alle übrigen bilden das Entwicklerteam. Der Product Owner trägt die Produktverantwortung. In WordGuess entspricht er dem Rater. Wenn er den Ratebegriff errät, hat das Team ein funktionsfähiges Produkt

Tabelle 1: Scrum Begriffe und ihre Entsprechung in WordGuess.

Scrum	WordGuess
Sprint	Spielrunde
Product Owner	Rater
Scrum Master	Organisiert/moderiert, gibt Suchbegriff vor
Entwicklerteam	Stellen und priorisieren Hinweise
Backlog	Textdatei mit Suchbegriff und Hinweisworten
Scrum Phasen	Scoped Labels in GitLab-Issues
Planning	Suchbegriff bestimmen, Hinweise zusammenstellen
Planning Poker	Max. Anzahl an zu gebenden Hinweisen
Daily Scrum	Hinweis geben - Suchwort raten
Sprint Review	Projektdokumentation erstellen
Sprint Retrospektive	Neue Rollenverteilung

entwickelt. Der Scrum Master ist für die Überwachung und Einhaltung der Scrum-Regeln verantwortlich. Er bestimmt das Suchwort und moderiert den gesamten Prozess. Das Entwicklerteam hat die Aufgabe, die Sprints inhaltlich zu planen und die Pläne dann auszuführen. Sie stellen die Hinweisworte zusammen und diskutieren gemeinsam mit dem Scrum Master die Gestaltung des Sprints. Ziel des Spiels ist es, dass der Product Owner anhand von Hinweisen ein vom Scrum Master präsentiertes Suchwort errät.

Gespielt wird über mehrere Runden, wobei eine Runde einem Sprint in der Scrum-Terminologie entspricht. Ein Sprint teilt sich in die Phasen Planning, Daily Scrum, Sprint Review und Sprint Retrospektive auf, wobei sich das Daily Scrum solange wiederholt, bis der Ratebegriff erraten wurde oder kein weiterer Hinweis im Sprint-Backlog vorgesehen ist. Eine Übersicht über den Ablauf und die Phasen eines Sprints gibt Abb. 4. Eine tabellarische Übersicht der Scrum-Begriffe und ihrer Entsprechung im WordGuess-Spiel ist in Tabelle 1 aufgeführt.

Projektmanagement und Kommunikation werden in GitLab als Spielumgebung mit den Features Issues, Labels, Boards und Milestones umgesetzt:

Issues: Jede Spielrunde (Sprint) verfügt über die GitLab-Issues „Sprint_x“ und „Sprint_x_Entwicklung“ (x=Rundennummer). Diese werden zu Beginn des Sprints vom Scrum Master erstellt und mit den entsprechenden Labels versehen. In diesen Issues findet über Kommentare die gesamte Kommunikation innerhalb des Spiels statt. Der Scrum Master weist die Issues den restlichen Spieler zu, so dass diese eine Benachrichtigung erhalten, sobald im Issue eine Änderung erfolgt ist.

Labels: Die Scrum-Phasen werden über Scoped Labels abgebildet. Diese haben den Vorteil, dass zu jedem Zeitpunkt nur ein Scoped Label des gleichen Typs einem Issue zugeordnet werden kann und die Scrum Phasen so klar voneinander getrennt sind. Zur Kennzeichnung der aktiven Issues und ob ein Sprint erfolgreich war oder nicht, sind im Spiel weitere normale, unscoped Labels vorhanden.

Boards: Zum Planen, Organisieren und Visualisieren des Projektverlaufs können Issue-Boards verwenden werden.
Das Scrum-Board besteht neben der „Open“- und „Closed“-Listen aus Listen für die verschiedenen Scrum-Phasen-Labels. Im Spiel verschiebt der Scrum-Master die Issues entsprechend der Scrum-Phasen über das Board.

Milestones: Milestones werden genutzt, um dem Projekt eine zeitliche Struktur zu geben. In WordGuess wird für jeden Sprint ein Milestone erstellt und dieser den entsprechenden Issues zugewiesen. Wie lange ein Sprint in WordGuess dauert, stimmen die Spielenden im „Planning Poker“ ab.

3.4 Exemplarischer Ablauf eines Sprints

In der Spielvorbereitung muss der erste Scrum-Master die Spielumgebung bereitstellen. Dazu importiert er die über den Moodle-Raum bereitgestellte Exportdatei des WordGuess-Projekts in GitLab. Er legt das Scrum-Board an und lädt seine Mitspieler als Mitglieder zu seinem Projekt ein. Das Spielteam klont das GitLab-Repository und ist damit startklar. Im Folgenden wird exemplarisch für eine Spielrunde (siehe Abb.4) der Sprint 1 ausgeführt. Alle weiteren Sprints verlaufen analog.

Planning: Ein Sprint startet immer mit der Planning-Phase. In dieser Phase erstellt der Scrum Master die zum Sprint gehörenden Issues „Sprint_1“ und „Sprint_1_Entwicklung“. Im Issue „Sprint_1“ vergibt er das Label *In Progress*, um zu kennzeichnen, dass der Sprint begonnen hat, und weist das Issue seinen Mitspielern zu. Das Issue „Sprint_1_Entwicklung“ weist er nur dem Entwicklerteam zu. Auf dem Scrum-Board verschiebt er nun die Issues in die Liste des *Planning*-Labels, um zu kennzeichnen, dass die Planning-Phase läuft.

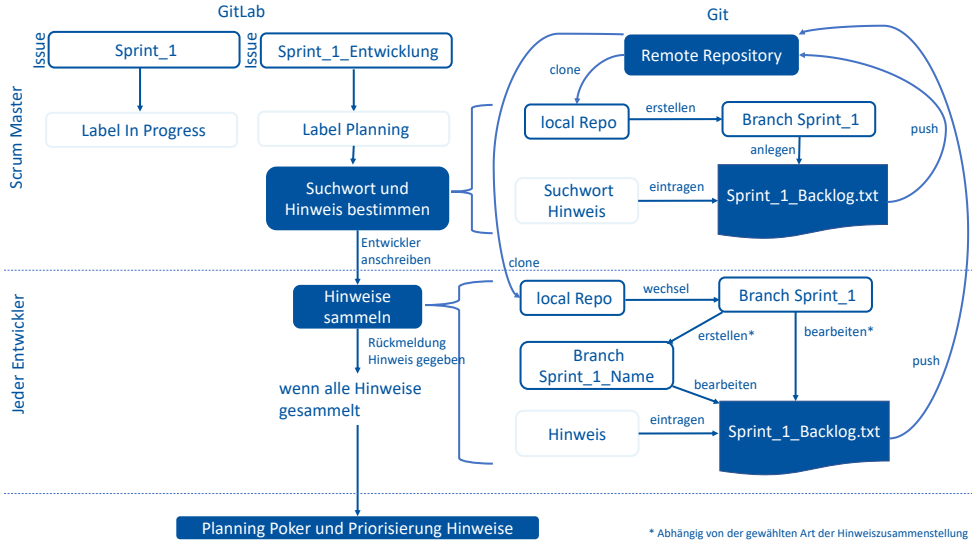


Abbildung 5: Ablauf der Planning Phase nach Erstellung der zugehörigen Issues.

Der weitere Ablauf der Planning-Phase findet sowohl in der Git-Umgebung, als auch im GitLab statt. Ein Überblick der Planning-Phase ist in Abb.5 dargestellt. Der Scrum Master erstellt einen neuen Branch „Sprint_1“ und legt die Datei „Sprint_1_Backlog.txt“ an, in die er das Suchwort sowie ein erstes Hinweiswort einträgt. Er pusht zurück ins Gitlab und fordert das Entwicklerteam auf, Hinweisworte zu ergänzen. Das Entwicklerteam aktualisiert ihr lokales Repository, wechselt in Branch „Sprint_1“ und hat nun zwei Möglichkeiten:

Variante A: Jeder fügt sein Hinweiswort in die Datei „Sprint_1_Backlog.txt“ ein und pusht ins Gitlab. Es werden Merge-Konflikte entstehen, da die Teammitglieder nicht aufeinander warten werden. Dies führt dazu, dass eventuell das lokale Gitlab Repository erneut aktualisiert werden muss, wenn der Stand im Gitlab von einem anderen Teammitglied geändert wurde.

Variante B: Jedes Teammitglied erstellt aus „Sprint_1“ einen weiteren Branch, z.B. „Sprint_1_Name“, fügt dort in der Datei „Sprint_1_Backlog.txt“ ein Hinweiswort hinzu und pusht ins Gitlab. Dabei sollten keine Merge-Konflikte entstehen. In dieser Variante mergt der Scrum Master alle erstellten Unter-Branches in seinen „Sprint_1“ und löst die Merge-Konflikte. In beiden Varianten sind die Merge-Konflikte so zu lösen, dass kein Hinweiswort verloren geht. Die Variante B ist etwas aufwendiger, bietet aber die Möglichkeit den Umgang mit Branches und Merges in Git zu üben.

Nun vergibt der Scrum Master das Label „Planning Poker“ in Issue „Sprint_1_Entwicklung“ und diskutiert über Kommentare mit seinem Entwicklerteam die Anzahl und Reihenfolge der Hinweise sowie die Sprintdauer. Er aktualisiert das „Sprint_1_Backlog.txt“, legt einen Milestone „Sprint_1“ entsprechend der festgelegten Sprintdauer an und ordnet beide Issues diesem Milestone zu. Schließlich entfernt er das Label „Planning Poker“.

Daily Scrum: Der Scrum Master verschiebt die Issues auf dem Scrum-Board in die Liste des *Daily Srcum*-Labels. Da es sich, wie alle Labels der Scrum-Phasen, um ein Scoped Label handelt, sorgt dies automatisch dafür, dass das Label ”Planningäus dem Issue entfernt wird. Weiter schreibt der Scrum Master den Product Owner an, dass ein Suchwort bestimmt wurde und verrät ihm den ersten Hinweis. Der Ablauf der Daily Scrum-Phase ist in Abb.6 dargestellt.

Der Product Owner versucht anhand des Hinweises das Suchwort zu erraten und erhält Feedback. Hat der Product Owner falsch geraten und es sind noch weitere Hinweisworte vorgesehen, startet das Daily Scrum erneut. Hat er richtig geraten, ist das Sprintziel erreicht. Der Sprint ist erfolgreich und das Sprint Review folgt. Sind alle Hinweise gegeben oder die Zeit abgelaufen ohne dass der Product Owner den Suchbegriff erraten konnte, endet der Sprint „erfolglos“ und das Sprint Review folgt ebenso.

Sprint Review: Der Scrum Master verschiebt die Issues auf dem Scrum-Board in die Liste des *Sprint Review*-Labels. Er entfernt das Label „In Progress“ in Issue „Sprint_1“ und vergibt das Label „Sprint erfolgreich“ oder „Sprint fehlgeschlagen“. Er verrät das Suchwort, falls es nicht geraten wurde.

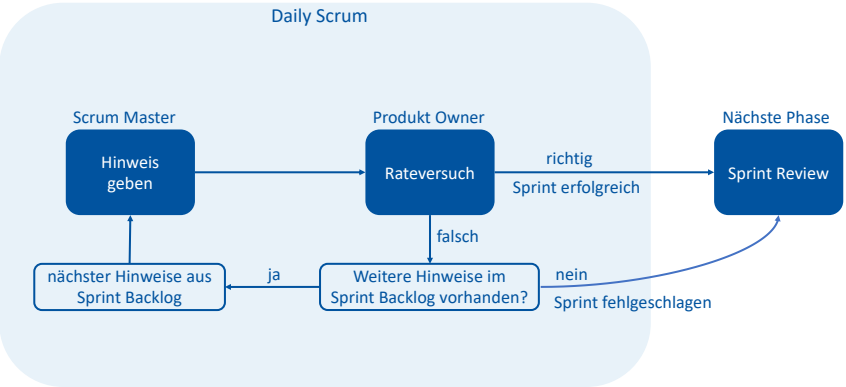


Abbildung 6: Ablauf der Phase Daily Scrum.

Er teilt dem Entwicklerteam über einen Kommentar in Issue „Sprint_1_Entwicklung“ den Ausgang des Sprints mit und schließt das Issue.

Im Sprint Review wird im GitLab-Wiki die Projektdokumentation zum aktuellen Sprint entsprechend einer vorgegebenen Template verfasst. Dies kann durch den Scrum Master erfolgen, oder ein Entwickler meldet sich freiwillig oder wird bestimmt. Nach Erstellung der Dokumentation erfolgt eine Rückmeldung an die Mitspieler und Änderungen können eingebracht werden.

Sprint Retrospektive: Der Scrum Master verschiebt die Issues auf dem Scrum-Board in die Liste des *Sprint Retrospektive*-Labels. In der „Sprint Retrospektive“ werden die Rollen für den nächsten Sprint vergeben. Dazu übergibt der Scrum Master seinen Posten über einen Kommentar in Issue „Sprint_1“ an eine/-n Mitspieler/-in. Der oder die Auserwählte bestätigt dies im Kommentar. Der Scrum Master entfernt das Label *Sprint Retrospektive*, schließt das Issue „Sprint_1“ und wird zum neuen Product Owner im nächsten Sprint. Alle weiteren Teilnehmenden sind automatisch Teil des neuen Entwicklerteams.

4 Fazit/Ausblick

Die Git Scavenger Hunt wird seit Januar 2021 im Kursprogramm für Promovierende und den akademischen Mittelbau an der RWTH Aachen University angeboten [7]. WordGuess wird im Mai 2021 folgen [10]. Zudem wurden die Kurse mit Studierendengruppen getestet und bereits in erste Lehrveranstaltungen integriert.

Das Feedback aus ersten Spielrunden in beiden Konzepten ist überaus positiv. Beispielsweise antworteten einige Studierende der Informatik, dass sie der spielerische Ansatz motiviert habe, die Schnitzeljagd bis zum Ende durchzuspielen, obwohl bereits vorab Git-

Kenntnisse vorhanden waren. An WordGuess wurden nach einem Test mit 16 Studierenden, welche sich gegenseitig Hinweise gaben und in den verschiedenen Rollen abwechselten, noch Verfeinerungen an dem Spiel durchgeführt. Zum einen wurden die Scrum Rollen als Lehrergänzung hinzugefügt, zum anderen die Gruppengröße limitiert.

Beide Kurse wurden weitestgehend im Sinne von Open Educational Ressources (OER), also nachnutzbar für Jedermann, erstellt. Limitierender Faktor ist das RWTHmoodle als Lernplattform, das den Zugang auf Personen mit RWTH-ID einschränkt. Um dies zu relativieren werden die Kurssicherrungsdateien der Moodle-Räume über RWTH Publications zum Download bereit gestellt, so dass sie in anderen Moodle-Instanzen wiederhergestellt und ggf. angepasst werden können. Die GitLab-Repositorien selbst sind öffentlich und enthalten ausreichend Informationen, um die Spielideen auch ohne begleitenden Moodle-Raum umsetzen zu können. Beide Lernspiele sind geeignet, um Data Literacy Kompetenzen an Studierende zu vermitteln. Für Dozierende der RWTH Aachen University werden deshalb die Kursinhalte in einem separaten Moodle-Raum bereitgestellt, von dem aus sie abgeholt und in eigene Vorlesungs- und Übungskonzept eingepflegt werden können.

Aufgrund des großen Interesses und dem generellen Bedarfs an Schulungsangeboten zu Git und GitLab, planen wir auch zukünftig die Konzepte weiterzuentwickeln und zu erweitern. Ideen für weitere Level der Schnitzeljagd liegen bereits vo.[1]

Literaturverzeichnis

[1] Hethey, Jonathan M. “GitLab Repository Management.” Packt Publishing Ltd (2013).

[2] Ayer, Vidya and Herrmann, Fabian and Peil, Vitali and Pietsch, Christian and Rempel, Andreas and Schirrwagen, Jochen and Vompras, Johanna and Wiljes, Cord. “Automatische Qualitätskontrolle von Forschungsdaten durch kontinuierliche Integration mit GitLab CI.” 2019.

[3] Sailer, Michael and Hense, Jan and Mandl, J and Klevers, Markus. “Psychological perspectives on motivation through gamification.” Interaction Design and Architecture Journal, Nr. 19 (2014): 28-37.

[4] Sailer, Michael. “Die Wirkung von Gamification auf Motivation und Leistung.” Springer (2016).

[5] Schwaber, Ken and Sutherland, Jeff. “The scrum guide.” Scrum Alliance, Nr. 21 (2011).

[6] Umali, Rick. “Learn Git in a Month of Lunches.” Manning Publications Co. (2015).

[7] Jüptner, Patrick und Dalibor, Manuela und Trautwein-Bruns, Ute. “Git Scavenger Hunt - Eine Schnitzeljagd mit Git”. DOI: https://doi.org/10.18154/RWTH-2021-03720.

[1] Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder - EXC 2023 Internet of Production

[8] Neil, Drew. “Practical Vim: Edit Text at the Speed of Thought.” Pragmatic Bookshelf. (2015).

[9] Beutelspacher, Albrecht. “Cäsar oder Aller Anfang ist leicht!” Kryptologie (2020): 1-25.

[10] Jüptner, Patrick und Dalibor, Manuela und Trautwein-Bruns, Ute. “WordGuess - ein kollaboratives Spiel mit Git und GitLab”. DOI: https://doi.org/10.18154/RWTH-2021-04607.

Practical Interoperability in the Virtual Observatory

Markus Demleitner[1,2]

[1]Universität Heidelberg, Zentrum fürAstronomie;
[2]German Astrophysical Virtual Observatory GAVO

The Virtual Observatory (VO) is an international effort to run and develop a federated data infrastructure in Astronomy that is held together by a set of data and protocol standards. Consisting of a Registry, some 30000 interoperable services (which roughly correspond to data collections) comprising hundreds of millions of datasets (spectra, images, and the like) and hundreds of billions of table rows, as well as a set of clients and libraries consuming these services, it is widely used in the astronomical community.

In this contribution, I will give a condensed overview of the technologies behind the VO, with a particular emphasis on how the VO is not just a platform but a truly global infrastructure jointly shaped by data providers, software authors, and science users.

1 Introduction

The Virtual Observatory (VO) as a project started in the early 2000s (e.g., [3]) and has grown to become a global data infrastructure for Astronomy and Astrophysics, now encompassing data centers in about 30 countries[1].

The VO is *not* a website ("platform", "portal"), nor some sort of network of websites, nor a programme that tries to do everything astronomers might want to do with a computer. It is the main objective of this contribution to explain what it is instead and why it was designed in this way.

The VO could be defined as

- A few dozen standards for finding, accessing, using, and describing data, authored and agreed upon under the auspices of the International Virtual Observatory Alliance IVOA.
- The astronomy data centers publishing data using these standards; this includes almost all the major players like NASA, ESA, or ESO.

[1]There are 34 top-level domains in the access URLs of VO-registered services in March 2021.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021.
Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13716 (CC BY-SA 4.0)

- Some volunteer institutions running infrastructure services[2], where the design is such that clients can easily be written without dependencies on specific instances of central components.
- Authors of client software, libraries, and web pages making these resources available to astronomers. Of the clients programmes listed on the IVOA's web site[3], non-astronomers might want to look at TOPCAT and Aladin.

The VO Text Treasures[4] service lists worked-out use cases that may give closer insights into what this actually means.

2 Data Discovery in the Virtual Observatory

In the VO, data discovery very typically is a two-step process, in which a client first queries the Registry for services that might have relevant data. The Registry is the collection of the metadata records on the level of data collections and will be looked at in some detail in Sect. 3.7.

The result of the Registry query will in general be a set of access URLs of machine APIs together with an identifier of what standard the API implements. With this information, a client programme can then visit the APIs in turn, running discovery queries in the end yielding references to datasets matching the constraints.

To make this a bit more concrete, consider a researcher looking for images of Barnard's star in X-rays. In this scenario, the researcher will run a client programme that will ask the Registry: What "resources" (services or data collections) are available that:

- serve or contain images
- have data in the X-ray part of the spectrum
- have data around $\alpha = 269.45$, $\delta = 4.693$ (the current position of Barnard's star in ICRS coordinates)?

In a second step, the client visits each service found (provided it supports the advertised communication protocols) and will post one request to each for images that

- cover the position $\alpha = 269.45$, $\delta = 4.693$,
- intersect the spectral range $0.1 \cdots 120\,\mathrm{keV}$ of photon energy.

The on-the-wire serialisation of these constraints depends on the protocol and may even have to happen client-side; for instance, the image discovery protocol SIAP in its version 1 (which is still widely used) cannot express spectral constraints.

[2] This primarily includes components of the service registry, but also the the IVOA web page at `https://ivoa.net` and the associated document repository and collaboration wiki.

[3] `http://ivoa.net/astronomers/applications.html`

[4] `https://dc.g-vo.org/VOTT`

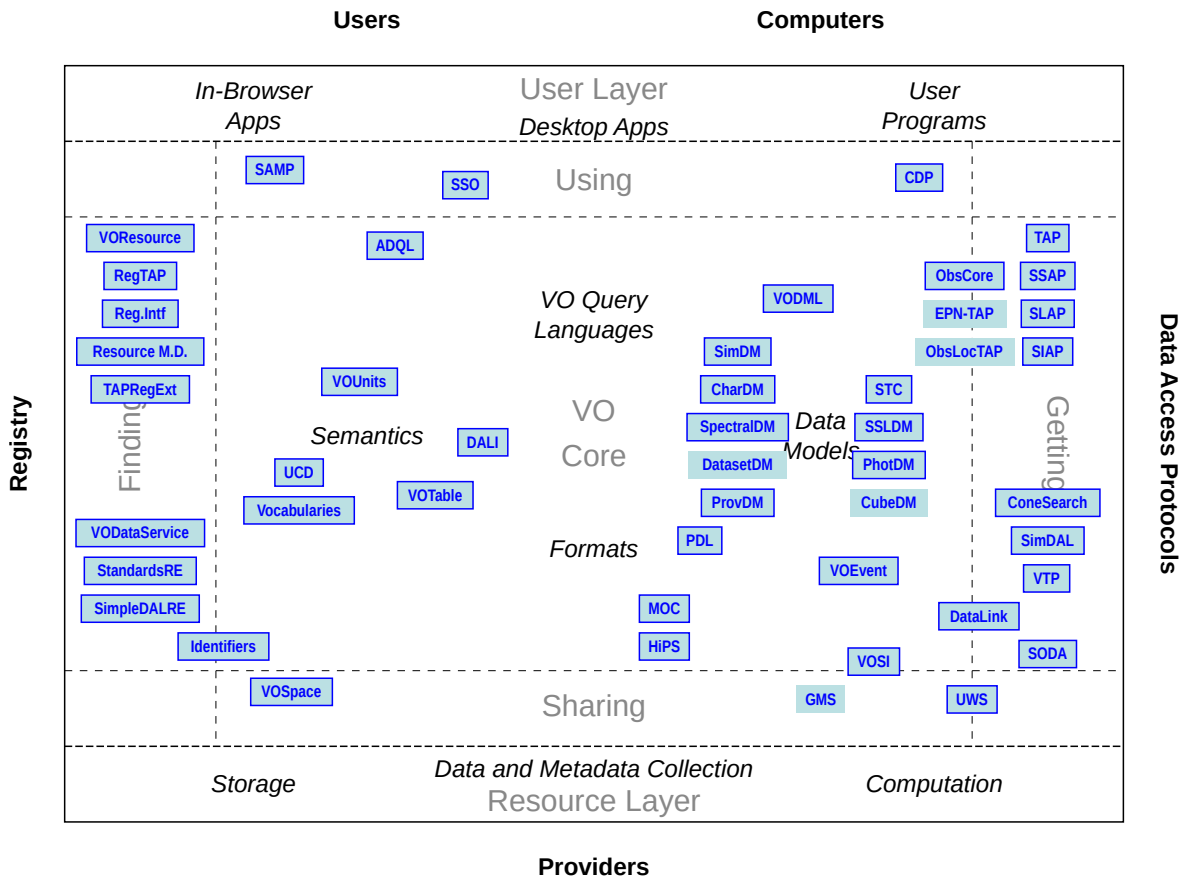


Figure 1: The Virtual Observatory's "Architecture Diagram", showing one box for each standard contributing to the ecosystem, grouped by their function. Some of the standards shown are discussed in the text; all of them are available through the IVOA's document repository.

The client programme then presents the resulting metadata records to the researcher in some suitable form – the Aladin client, for example, will show image footprints on a sky display in addition to a tabular rendering –, who in turn decides which data to retrieve (and in what way).

In practice, this procedure is a lot more natural than it might sound here, mostly because much of the interaction can run behind the scenes thanks to the standardised APIs in use in the VO.

3 Standards

Essentially all VO procedures depend on machines querying each other and understanding the responses without human intervention. This requires a significant effort in standardising protocols and formats.

Ever since the Architecture Note [1], in the Virtual Observatory the standards are visualised as shown in Fig. 1. Since the sheer number of standards involved may be somewhat daunting, let us have a look at some of the more salient ones and their functions. Words written like this correspond to standards readily found in the IVOA document repository[5] In order to keep the bibliography reasonably compact, we do not include full citations for them.

3.1 Finding Data and Datasets

The discovery protocols on the right side of the architecture diagram include "typed" protocols specialised on images (SIAP), spectra (SSAP), objects (SCS), or spectral lines (SLAP). All of these essentially define some query parameters and some basic requirements on the tabular structure of the response.

As the VO matured, tools were in place to define a table schema (ObsCore) that, together with the Table Access Protocol (TAP) to be discussed presently, enables flexible and powerful dataset discovery largely independently of their types. In a re-design of the VO, it is likely that one would think hard whether there actually is a need for the typed protocols – which, on the other hand, were a lot simpler to design than the generic discovery protocols and thus kept up the momentum while the years required to develop something like TAP (and its implementations) went by.

Another standard we should have started with but only introduced relatively late in the VO's evolution is the Data Access Layer Interface DALI giving common patterns for VO standards concerned with finding and accessing datasets.

3.2 Advanced Data Access

Where datasets (such as images, spectra, or time series) are discovered, the services return metadata rather than the datasets themselves because very typically only a small fraction of the discovered datasets turn out to be necessary for a given analysis, and the datasets may be large.

For large and complex datasets, the typical access mode of simply dereferencing an (http) URI may not be suitable; in particular, clients may want to only retrieve parts of the dataset. To cater for such cases, the VO has defined Datalink, which allows data providers to declare relationships between various artefacts (e.g., raw data, calibration files, and reduced data) making up a dataset, and giving separate access to them.

On top of Datalink, SODA defines some standard operations on array-like data (e.g., cutouts), which in many science cases can reduce the amount of data to be transferred by several orders of magnitude.

[5] `https://ivoa.net/documents`.

3.3 Interacting with Databases

The definition of the Table Access Protocol (TAP) in 2010 enabled many interesting use cases; in particular, the built-in table metadata inspection and table upload facilities are central to the protocol's success.

Equally important was the definition of a common language available across all TAP services, the SQL-derived Astronomical Data Query Language ADQL. It is now no longer unusual to see ADQL fragments in scientific publications, and an increasing fraction of astronomers becomes proficient in expressing science questions in SQL-like languages.

3.4 Formats

Rich metadata is a precondition of re-usability of data as well as interoperability of services. Hence, the XML-based table and metadata format VOTable was the first standard defined in the VO and keeps being regularly evolved.

The VO had a head start because FITS [4], a standard for images and several other data types, was already widely accepted within the field and could simply be re-used. Still, some additional formats, for instance for complex spherical geometries (MOC) or for large, multi-resolution, possibly full-sphere images and tables (HIPS), had to be defined.

Related to these data formats are several syntactic aspects, such as agreeing on a well-defined way to serialise physical units into ASCII strings (VOUnit).

3.5 Desktop Interoperability

From the VO's start it was clear that no single software would ever be sufficient to serve the needs of the various communities using VO facilities. Instead, the design called for multiple, independently developable applications that, however, can closely interoperate among each other. Given that no widely adopted standard for cross-platform desktop programme communication existed in the mid-2000 (or, really, exists today), the Simple Application Messaging Protocol SAMP was developed and enjoys great popularity among VO users.

3.6 Semantics

Early on, a relatively complex labeling scheme called UCD was devised to enable machine-readable annotation of table columns with what sort of physics they represent (e.g., "radio flux" versus "distance").

Later, RDF-compliant, hierarchical vocabularies were defined for many different purposes, from the relationships between different artefacts in Datalink to time scales to content levels of resources (Vocabularies).

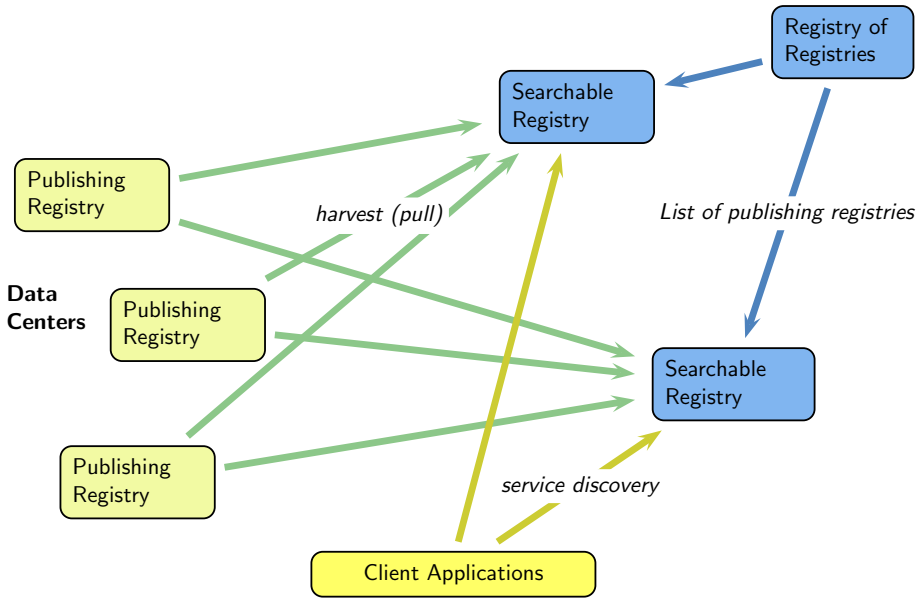


Figure 2: A sketch of the registry architecture of the VO: Data centers operate OAI-PMH endpoints ("publishing registries"), which are harvested by searchable registries, which in turn are what client applications talk to. The searchable registries learn what publishing registries to harvest from a central list kept at the registry of registries.

3.7 Registry

The part of the VO that is probably most readily re-usable in large parts in other disciplines is the Registry (cf. [2]), the distributed data collection metadata store for the VO's resources. Let us therefore take a closer look at its architecture.

The basic scheme, as defined in Registry Interfaces, is shown in Fig. 2. It again follows the fundamental VO principle that no single central component should be required for client applications. The searchable registries – all of them ideally serving identical data – can be operated by anyone and liberally scaled, thus providing substantial resilience against failures. In practice, three major searchable registries are operated in the VO, one each by NASA, ESA, and GAVO.

The components not easily reproducible, the Registry of Registries and the publishing registries, can in principle be down for days without VO users noticing; it is only registry updates – to the list of publishing registries and the records published by a single publisher, respectively – that will not happen during the downtime.

Standards contributing to this system are a lightweight identifier scheme, IVOA identifiers, a set of rules for the VO-specific application of OAI-PMH, Registry Interfaces, and the metadata schema based in VOResource and extended to various resource types in VODataService, SimpleDALRegExt, and TAPRegExt.

The now-dominant registry client interface RegTAP builds on TAP and essentially just defines a relational mapping of the metadata scheme.

This mapping yields moderately-sized tables ranging from $\sim$ 30000 records for the table of resources to $\sim$ 1.2 million records for the table of columns in published tables.

4 Concluding Remarks

Compared with just writing some custom web page, building a system like the Virtual Observatory may seem like a daunting task. On the other hand, many publishers rebuilding the same kinds of tools on their custom web pages over and over is substantially more work on the long run, and the benefits of being able to share the load of developing client software provides a large benefit to both data producers and data consumers.

An alternative to our current design – that predates the "platform economy" – would be a single, giant, central platform keeping essentially all astronomical data. This is a model that in astronomy works quite well for literature in the form of NASA's Astrophysics Data System ADS[6]. For data, with its much greater variety, volume (at least when measured in bytes), and demands on machine-readability, it would probably be very difficult to work out a global funding scheme for such and establishment.

More importantly, however, with multiple interoperable data centers, the VO can "grow from the edges", much like the internet (at least in its early days). This means that everyone is free to run services and to improve the tools and, within reason, also the standards – there is no single entity determining what can and cannot be done.

For users, this means that they are free to choose whatever tools they want to use, something a platform would severely limit. Having data uniformly presented and described also not only greatly facilitates working with cross-instrument (and hence multi-wavelength, possibly even multi-messenger) data, it also significantly increases the chances that workflows will be reproducible (again, within reason) years down the road: Our API endpoints have proved to be a lot more stable than web pages.

Acknowledgements

The German part of the Virtual Observatory, GAVO, has been supported by BMBF under several grants. Current grant: e-inf-astro, FKZ 05A20VH5.

Bibliography

[1] Christophe Arviset, Severin Gaudet, and IVOA Technical Coordination Group. IVOA Architecture Version 1.0. IVOA Note 23 November 2010, November 2010.

[6] `https://ads.harvard.edu`

[2] M. Demleitner, G. Greene, P. Le Sidaner, and R. L. Plante. The virtual observatory registry. *Astronomy and Computing*, 7:101–107, November 2014.

[3] R. J. Hanisch. The Virtual Observatory: I. *Astronomy and Computing*, 7:1–2, November 2014.

[4] D. C. Wells, E. W. Greisen, and R. H. Harten. FITS - a Flexible Image Transport System. *Astronomy and Astrophysics Supplement*, 44:363, June 1981.

Lokal betrieben, remote gepflegt – Software für ein Datenrepositorium in Kooperation implementieren

Matthias Landwehr [1], Gabriel Schneider [1], Stefan Hofmann [2], Matthias Razum [2] und Kerstin Soltau [2]

[1]Kommunikations-, Informations-, Medienzentrum (KIM), Universität Konstanz
[2]FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur

Die Universität Konstanz deckt ihren Bedarf nach einem institutionellen Forschungsdatenrepositorium mit der von FIZ Karlsruhe angebotenen Lösung „RADAR Local“. Als Alternative zu einer Eigenentwicklung wurde das Datenrepositorium als hybrides Modell mit Repositorien-Software und Archivierung auf lokaler Infrastruktur implementiert. Dabei stellt FIZ Karlsruhe die etablierte Repositorien-Software RADAR zur Verfügung, wartet und betreibt sie aus der Ferne und passt sie nach Kundenwunsch an. Um die parallele Installation und Pflege der RADAR-Software auf mehreren lokalen Instanzen effizient bewältigen zu können, hat FIZ Karlsruhe vorab den Automatisierungsgrad der betroffenen Prozesse in der Software-Entwicklung, in der Systemkonfiguration und im Deployment erhöht. Dies wurde durch den Einsatz von Container-Virtualisierung wie Docker und Docker Swarm sowie mit Orchestrierungswerkzeugen wie Ansible erreicht.

Der zeitliche Aufwand und der personelle Ressourcenbedarf reduzieren sich dadurch für die Universität Konstanz und als Ergebnis erhält sie ein gepflegtes Repositorium auf dem aktuellen Stand der Technik. Gleichzeitig erfordert diese Betriebsvariante eine intensive Auseinandersetzung mit dem jeweiligen Geschäftsmodell und den technischen Rahmenbedingungen des Anbieters, eine genaue Kostenkalkulation sowie möglicherweise Kompromisse oder Abstriche bei individuellen Wünschen.

1 Einleitung

Universitäten und andere wissenschaftliche Einrichtungen sehen sich bei der Entwicklung und Einführung von Datenrepositorien vielschichtigen Herausforderungen gegenüber. Am Anfang steht der Bedarf nach einem institutionellen Repositorium, die Herausforderung liegt in der Umsetzung.

Neben der Wahl einer geeigneten technischen Basis ist der Aufwand einer individuellen Implementierung und langfristigen Pflege nicht zu unterschätzen.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13717 (CC BY 4.0)

Erschwerend hinzu kommt der aktuelle IT-Fachkräftemangel, der sich gerade in öffentlichen Einrichtungen zunehmend niederschlägt und zu großen Verzögerungen bei Eigenentwicklungen im IT-Bereich führt. Diesen Problemen sieht sich auch die Universität Konstanz gegenüber.

2 Bedarf

Am Ausgangspunkt der Kooperation stand der Bedarf der Universität Konstanz nach einem eigenen institutionellen Datenrepositorium. Dieser Bedarf bildete sich in den vergangenen Jahren aus verschiedenen Gründen heraus. Die Bedeutung von Forschungsdaten und deren Management hat in den letzten Jahren stetig zugenommen. Nicht zuletzt durch die Richtlinien von nationalen und internationalen Forschungsförderern, ist die eigenständige Publikation von Forschungsdaten in zitierfähiger Form eine Aufgabe geworden, mit der sich Forschende beschäftigen müssen und auch wollen. Zusätzlich stellt das Thema Open Science einen Schwerpunkt der Universität Konstanz dar und das Anbieten eines institutionellen Datenrepositoriums kann einen Beitrag zur Öffnung der Wissenschaft darstellen.

Die Universität Konstanz wurde im Jahr 1966 als Reformuniversität gegründet und ist seit 2007 dauerhaft in den Exzellenzwettbewerben des Bundes erfolgreich. Das Kommunikations-, Informations-, Medienzentrum (KIM) ist der Zusammenschluss von Bibliothek und Rechenzentrum und betreibt und pflegt die IT-Angebote für die Universität. Das Thema Open Science ist in der Exzellenzstrategie prominent verankert und das KIM unterstützt die Ziele mit einem gut aufgestellten Team Open Science, das sich u.a. um die Themen Open Access und Forschungsdatenmanagement kümmert.

2.1 Ausgangslage

Die erste Empfehlung für die Datenpublikation ist immer die Nutzung eines fachspezifischen Repositoriums. Leider findet sich in manchen Fällen kein geeigneter Dienst und dann sieht man sich als Universität und Infrastruktureinrichtung mit der Forderung konfrontiert, eine dauerhafte Heimat für diese Daten bereitzustellen. Repositorien können aufgrund von Veröffentlichungskosten, mangelnder Einschlägigkeit, Limitierungen in Bezug auf die Größe der Datensätze oder anderen Hürden in den Richtlinien unpassend für die Bedarfe der Wissenschaftler*innen sein. In einigen Fällen ist es auch gewünscht, dass die publizierten Daten „physisch“ im Haus bleiben. Dann fallen internationale Repositorien mit außereuropäischen Servern z.B. aus Gründen des Datenschutzes heraus. Zudem hat sich auch das Selbstverständnis vieler Einrichtungen dahingehend verändert, dass ein institutionelles Datenrepositoriums inzwischen zum Serviceangebot gehört.

Bei jedem Dienst, den man als Universität anbieten möchte, muss man die Ressourcenlage betrachten. Während technische und finanzielle Ressourcen das geringere Problem darstellen, steht das Thema der personellen Ressourcen im Vordergrund.

Der Fachkräftemangel schlägt - insbesondere in der IT - auch an Universitäten durch. In Konstanz kommt als zusätzlicher Faktor noch die Nähe zur Schweiz hinzu, die den Jobmarkt nochmal verengt.

Ebenfalls zu berücksichtigen ist die Tatsache, dass es bereits zahlreiche institutionelle Datenrepositorien gibt und es nicht angebracht ist, dass überall „das Rad neu erfunden“ werden muss. Im akademischen Umfeld gab es in der Vergangenheit viele Fälle von Einrichtungen mit nach einigen Jahren schlecht wartbarer Individualsoftware, die nicht nachhaltig gepflegt und betrieben werden konnte Aus dieser Ausgangslage entwickelte sich die Anforderung, eine geeignete Lösung mit ausreichender Funktionalität zu finden, die mit geringem Aufwand zu realisieren ist.

2.2 Anforderungen

Der erste Schritt ist immer die Aufstellung der funktionalen Anforderungen, die ein Produkt erfüllen muss. In gemeinsamer Arbeit mit allen beteiligten Personen in der Universität wurde ein Kriterienkatalog erstellt. Dabei kooperierten vor allem Mitarbeiter*innen des Bereichs Open Science und Software-Entwickler*innen des KIM mit den verantwortlichen Personen aus der technischen Infrastruktur. Ebenfalls einbezogen wurde die bibliothekarische Kompetenz in den Bereichen Metadaten und Publikationen. Die wichtigsten Punkte des Katalogs werden hier dargestellt. Das gewünschte Datenrepositorium soll:

- ein generisches Repositorium für alle Arten von Datensätzen sein
- über ein ausreichendes Rollen- und Rechtemanagement verfügen
- eine flexible Metadatenverwaltung mitbringen
- bei der Publikation einen DOI vergeben
- verschiedene Lizenzen anbieten, Embargo und Closed Access unterstützen
- intuitiv benutzbar sein
- sich in die bestehende IT-Landschaft der Universität integrieren lassen
- die Authentifizierung, Autorisierung und die Bereitstellung der Datensätze über etablierte Schnittstellen ermöglichen
- leicht wartbar sein und aktiv weiterentwickelt werden

Bevor im nächsten Schritt verschiedene Systeme grundlegend gegen diesen Kriterienkatalog evaluiert wurden, wurde die Frage gestellt, ob man ein Repositorium nicht als „Software / Infrastructure as a service“ komplett einkaufen kann. Bei der Bereitstellung eines Repositoriums durch einen externen Dienstleister in einer Cloud werden Wartung, Pflege und Entwicklung komplett abgetreten, der Kunde nutzt ein fertiges Produkt. Dies ermöglicht den Betrieb mit einem wesentlich geringeren Personalaufwand.

Statt Ressourcen für Serveradministration und Softwareinstallation zu binden, kann eine Konzentration auf die bibliothekarische Kernkompetenz stattfinden und den Nutzenden eine umfangreiche Unterstützung bei der Kuration und der Publikation der Daten zur Verfügung gestellt werden.

Ein Hindernis bei einer Cloud-Lösung ist die Tatsache, dass die Daten in der Cloud liegen und damit die lokale Infrastruktur der Universität verlassen. Bei der Recherche und Marktanalyse stieß die Universität Konstanz dann auf das Produkt „RADAR Local“ von FIZ Karlsruhe. Das Leibniz-Institut für Informationsinfrastruktur ist eine GmbH mit anerkannter Gemeinnützigkeit und hat als eine der großen Informationsinfrastruktureinrichtungen in Deutschland den öffentlichen Auftrag, Wissenschaft und Forschung mit wissenschaftlicher Information zu versorgen und entsprechende Produkte und Dienstleistungen zu entwickeln.

Mit RADAR bietet FIZ Karlsruhe seit 2017 Forscherinnen und Forschern an Hochschulen und außeruniversitären Forschungseinrichtungen in Deutschland ein flexibles, kostengünstiges und einfach zu nutzendes Repository für die Archivierung und Publikation von Forschungsdaten. Gemeinsam mit den nutzenden Einrichtungen wird RADAR kontinuierlich weiterentwickelt und an neue Nutzungskontexte angepasst. Mit der Betriebsvariante RADAR Local wurde ein Produkt geschaffen, das auf der Infrastruktur des Kunden läuft, aber von extern betrieben und gepflegt wird. RADAR Local ging in der Gesamtbetrachtung als Sieger aus der Evaluation verschiedener Produkte hervor.

3 Technische Umsetzung

3.1 Motivation

Das Forschungsdatenrepositorium RADAR[1] entstand als Cloud-Lösung im Rahmen eines DFG-Projekts[2] und ging 2017 online. Mittlerweile setzen zehn Hochschulen und außeruniversitäre Forschungseinrichtungen „RADAR Cloud“ ein. Die Einführung und Nutzung von RADAR Cloud ist für diese Institutionen sehr einfach, allerdings fallen volumenabhängige Speicherkosten an.

Auch gibt es vertragliche Beschränkungen bezüglich der Archivierung personenbezogener Forschungsdaten. In Gesprächen mit interessierten Institutionen entstand so die Idee, RADAR in einer weiteren Betriebsform anzubieten: die Einrichtung stellt die notwendige IT-Infrastruktur für den Betrieb des Systems, FIZ Karlsruhe stellt die Software und betreibt RADAR auf den lokalen Systemen der Einrichtung.

Aktuell läuft die RADAR-Software - eine verteilte, mehrschichtige Anwendung, die sich in eine Vielzahl von Diensten und Schnittstellen gliedert (siehe Abbildung 1) - nur auf

[1] https://www.radar-service.eu

[2] https://radar.products.fiz-karlsruhe.de/sites/default/files/radar/docs/info/Abschlussbericht_RADAR_DFG-Projekt_Publikation.pdf

einer zentralen Cloud-Instanz. Für RADAR Local kommen weitere Instanzen hinzu. Alle Instanzen nutzen dabei den identischen Software-Stack. Damit stellt sich die grundsätzliche Frage, wie FIZ Karlsruhe das komplexe System nicht nur effizient für RADAR Cloud installieren kann, sondern zukünftig parallel auch auf den Systemen aller Einrichtungen, die RADAR Local nutzen.

Die beiden Betriebsvarianten unterscheiden sich hauptsächlich in der Storage Layer, denn hier findet die Verknüpfung mit dem institutseigenen Archivspeicher statt. Um den Aufwand für Installation und Pflege möglichst gering zu halten, war das Ziel eine weitestgehende Automatisierung des Deployments. Die dafür notwendigen Schritte werden in diesem Kapitel beschrieben. Sie greifen nicht nur für RADAR Local, sondern betreffen ausdrücklich beide Angebotsformen.

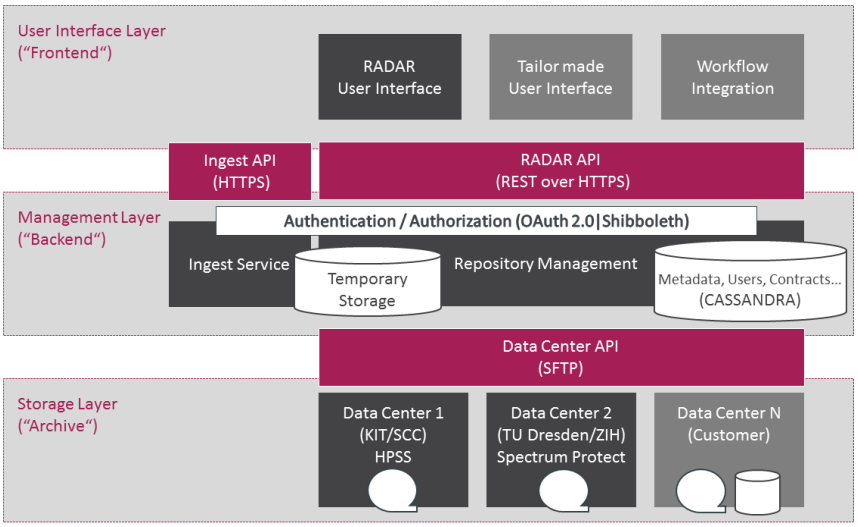


Abbildung 1: RADAR Systemarchitektur.

Das Aufsetzen und der Betrieb einer RADAR-Instanz sind komplexe Prozesse. Für die Bereitstellung von RADAR Cloud sind sowohl verschiedene interne Abteilungen bei FIZ Karlsruhe als auch bei externen Partnern (Steinbuch Centre for Computing[3] des KIT,[4] Zentrum für Informationsdienste und Hochleistungsrechnen der TU Dresden,[5] DataCite[6]) erforderlich. Die anfallenden Aufgaben umfassen grob die Bereiche Betriebssystem, Datenhaltung, Anwendung, Netzwerke und Monitoring. Jeder Bereich umfasst eine Vielzahl an einmaligen und wiederkehrenden Aufgaben (siehe Abbildung 2), die jeweils spezielles Domänenwissen benötigen. Jeder Bereich muss schnell auf ändernde Anforderungen bzw. Situationen reagieren können. Man denke zum Beispiel an das Einspielen von Sicherheitspatches und Bugfixes oder das Zurückspielen eines Backups. Diese Komplexität wird sich mit RADAR Local weiter erhöhen, denn mit der Installation jeder lokalen Instanz kommen weitere Parteien mit ins Spiel.

[3] https://www.scc.kit.edu/
[4] https://www.scc.kit.edu/
[5] https://tu-dresden.de/zih
[6] https://datacite.org/

Um Installationen und Betrieb für RADAR Cloud und mehrere Instanzen von RADAR Local parallel bewältigen zu können, benötigt FIZ Karlsruhe standardisierte und konfigurierbare Automatisierungswerkzeuge, die komplexe Aufgaben kapseln und so abteilungs- und institutionenübergreifende Arbeiten zusammenfassen und einfach bedienbar machen.

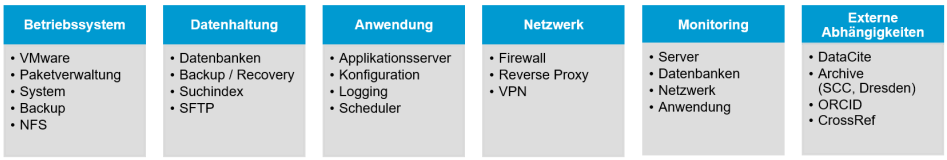


Abbildung 2: Notwendige Software-Komponenten für den Betrieb von RADAR Local.

3.2 Automatisierungswerkzeuge

Für RADAR Local liegt der Fokus der Automatisierung auf den Bereichen Systemkonfiguration, Deployment sowie Container-Virtualisierung und Container-Orchestrierung. Bei der Auswahl geeigneter Werkzeuge spielten Faktoren wie bereits vorhandene Kenntnisse und die Lernkurve eine wichtige Rolle. Die Entscheidungsfindung und untersuchte Alternativen zu den ausgewählten Werkzeugen sollen hier jedoch nicht weiter vertieft werden.

Ansible[7] ist ein Open-Source Automatisierungs-Werkzeug zur Orchestrierung, allgemeinen Konfiguration und Administration von Computern. FIZ Karlsruhe nutzt es in den Bereichen Systemkonfiguration und Deployment, spezifisch zum Aufsetzen und zur Wartung der eingesetzten Software-Infrastruktur. Das Tool zeichnet sich durch seine Einfachheit und die Unterstützung durch eine große Community aus. Ansible hat einen modularen Aufbau und unterstützt diverse Betriebssysteme. Eine wichtige Eigenschaft ist seine Idempotenz, d.h. auch bei mehrfacher Ausführung einer Aufgabe ist das gleiche Ergebnis wie bei einmaliger Ausführung erwartbar. Mit den Modulen für die CentOS-Paketverwaltung und Docker-Laufzeitumgebung können die Mitarbeiterinnen und Mitarbeiter von FIZ Karlsruhe einen Großteil der geplanten Aufgaben umsetzen. Ansible setzt auf der Zielinfrastruktur lediglich einen SSH Zugang voraus. Es integriert sich zudem in den Entwicklungsprozess und folgt der DevOps-Philosophie, die eine effektive und effiziente Zusammenarbeit zwischen Softwareentwicklung und IT-Betrieb anstrebt.

Automatisierungsaufgaben werden in sogenannten Ansible Playbooks gut lesbar formuliert und dienen so auch gleich der Dokumentation. Die dazugehörigen Host-Konfigurationen werden in einem Git-Repository für unterschiedliche Instanzen (Entwicklung, Test, RADAR Cloud, Universität Konstanz) abgelegt. Tabelle 1 zeigt die im Rahmen von RADAR Local entwickelten Playbooks mit deren Hilfe es nun möglich ist, innerhalb kurzer Zeit einen Cluster für den Betrieb von RADAR Local vorzubereiten. setup-vm.yml Installation von Basispaketen (chronyd, net-tools) Installation von Docker, Docker-Compose

[7] `https://www.ansible.com/`

Tabelle 1: Ansible Playbooks für RADAR.

setup-vm.yml	Installation von Basispaketen (chronyd, net-tools) Installation von Docker, Docker-Compose Initialisierung http-proxy
init-docker-swarm.yml	Initialisierung des Docker Swarms
init-services,yml	Erstellen der Verzeichnisstruktur Kopieren von Konfigurationsdateien Starten der Dienste

Initialisierung http-proxy init-docker-swarm.yml Initialisierung des Docker Swarms init-services,yml Erstellen der Verzeichnisstruktur Kopieren von Konfigurationsdateien Starten der Dienste

Das Deployment der einzelnen Softwarekomponenten erfolgt in Containern. Container isolieren Anwendungen mit Hilfe von Container-Virtualisierung. Als Virtualisierungsumgebung kommt Docker[8] zum Einsatz, eine weit verbreitete freie Software, die auf vielen Linux-Distributionen eingesetzt werden kann und einfach zu erlernen ist. Für jede Anwendung oder Komponente von RADAR (siehe auch Abbildung 1) wird ein Docker-Image (die Basis eines Containers) verwendet. Ein Docker-Image kapselt den Dienst und alle vom Betriebssystem benötigten Ressourcen. Ein wichtiges Ziel beim Aufsetzen einer Docker-Imagedatei ist es, die Konfiguration von der Software sauber zu trennen. Ebenso werden notwendige persistente Speicherbedarfe externalisiert. Dadurch lassen sich Images einfach ersetzen oder unverändert auf unterschiedlichen Systemen nutzen. Der Einsatz von Docker-Images spart zeitaufwendige Arbeitsschritte wie das Aufsetzen von Webservern, Datenbanken oder Suchindizes ein. Zudem stehen für viele der in RADAR eingesetzten Komponenten fertige und von der Community bereits umfänglich getestete Docker-Images bereit.

Docker Swarm[9] verknüpft die eingesetzten Container zu einer Gesamtanwendung und verteilt diese auf den Rechnern der Zielinfrastruktur. Docker Swarm hilft zudem bei der Skalierung und Lastverteilung (Load Balancing) von Applikationen.

3.3 Laufzeitumgebung

Für den Betrieb von RADAR Local stellt die Universität Konstanz via VMware mehrere virtuelle Maschinen zur Verfügung. Auf diesen wird seitens FIZ Karlsruhe mit Ansible die Docker-Laufzeitumgebung installiert. Anschließend wird der Docker Swarm initialisiert, indem eine Maschine zum Manager bestimmt und den restlichen Maschinen die Rolle

[8] `https://www.docker.com/`
[9] `https://docs.docker.com/engine/swarm/`

von Workern zugeteilt wird (siehe Abbildung 3). In einer Konfigurationsdatei (docker-compose file) sind die Beziehungen der einzelnen Dienste untereinander und damit die Struktur der Gesamtanwendung definiert. Der Swarm Manager liest diese Datei ein und verteilt daraufhin die einzelnen Dienste als Container auf den vorhandenen virtuellen Maschinen (Swarm Nodes).

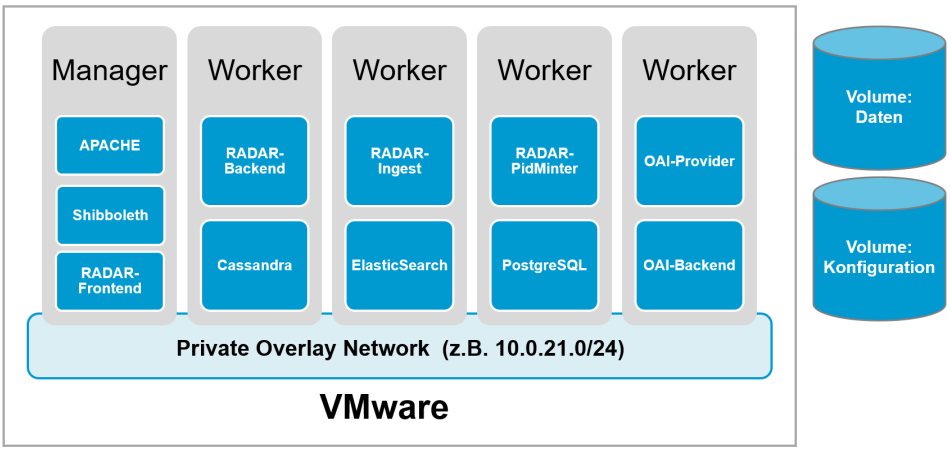


Abbildung 3: Docker Swarm-Umgebung.

Die Kommunikation zwischen den Services läuft in einem privaten Overlay-Netzwerk. Dieses virtuelle Netzwerk ist besonders geschützt und die gesamte interne Kommunikation erfolgt verschlüsselt. Dabei können gezielt einzelne Ports eines Containers, wie z.B. der Port 443 vom Webserver, für die VM freigegeben werden. Erst diese Portfreigabe ermöglicht Anfragen aus dem Internet via SSL.

Neben der Anwendung existieren noch zwei persistente Datenbereiche, die als Volumes in den Containern gemounted werden. Im Volume „Daten“ liegen die Datenbanken, der Suchindex, die Logdateien und die eingehenden Forschungsdaten. Im Volume „Konfiguration“ finden sich die Einstellungen für die einzelnen Dienste, wie z.B. die Webserverkonfigurationen oder Zertifikate.

3.4 Deployment Pipeline

Bei der Entwicklung und dem Betrieb von RADAR baut FIZ Karlsruhe auf die Prinzipien CI (Continuous Integration – mit jeder abgeschlossenen Änderung am Programmcode wird die Applikation neu kompiliert) und CD (Continuous Delivery – automatisierte Testung und Installation neuer Softwareversionen). Hierbei spielen die Entwickler regelmäßig ihre Quellcodeänderungen in das Software-Repositorium Git ein. Auf einem Bamboo Build-Server werden die Komponenten automatisch gebaut und mit verschiedenen Tests und Qualitätssicherungswerkzeugen wie z.B. Sonar Cube oder DependencyCheck überprüft. Laufen diese Prozesse fehlerfrei durch, erfolgt anschließend die Installation der neu gebauten Software auf einer internen und einer externen Testumgebung.

Der Einsatz von Docker ermöglicht es, die Frequenz der produktiven Deployments von bisher ca. vier pro Jahr zu steigern und so näher an das Ziel einer echten Continuous Delivery zu kommen. Jede Softwareänderung (im Sinn eines neuen oder überarbeiteten Features oder eines Bugfixes) soll dabei automatisiert und schnell durch die Pipeline auf den Produktionsmaschinen installiert werden. Dies wird erst durch den Einsatz von Containertechnologie möglich, da nun in den Container-Images die benötigte Infrastruktur definiert und fertig getestet ausgeliefert werden kann und aufwändige manuelle Prozessschritte entfallen.

Die grün hervorgehobenen Zeilen in Abbildung 4 zeigen, in welchen Elementen der Deployment Pipeline FIZ Karlsruhe mit den vorbereitenden Schritten zur Einführung von RADAR Local Automatisierungswerkzeuge integrieren konnte. Die schwarz gesetzten Komponenten sind bereits seit längerer Zeit bei FIZ Karlsruhe im Einsatz.

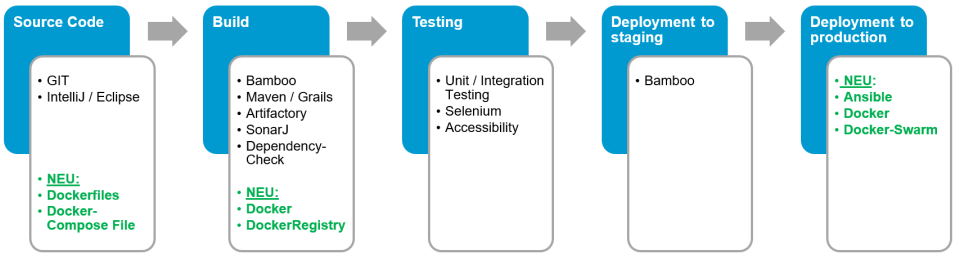


Abbildung 4: Continuous Integration / Continuous Delivery Pipeline.

3.5 Sicherheit

RADAR Local ist eine öffentlich zugängliche Webanwendung mit vielfältigen Schnittstellen zu anderen Systemen, zum Beispiel den eingebundenen Speichersystemen oder externen Diensten wie DataCite, CrossRef[10] und ORCID[11]. Neben der Härtung der Applikationskomponenten mittels automatisierter Tests mit OWASP ZAP (einem Open-Source Penetration Testwerkzeug) und der strikten Beschränkung der Kommunikation mit dem Internet auf HTTPS bzw. auf SFTP mit den Speichersystemen kapselt die Universität Konstanz ihre Instanz von RADAR Local in einem eigenen Subnetz, das vom Rest des Netzwerks der Universität und nach außen über Firewalls abgesichert ist.

Besonders sensibel ist der Zugriff der Administratoren von FIZ Karlsruhe über öffentliche Netze auf die Systeme der Universität Konstanz. Die gesamte Kommunikation zwischen FIZ Karlsruhe und den Systemen der Uni Konstanz erfolgt daher in einem verschlüsselten virtuellen privaten Netzwerk (VPN). Die Universität stellt darüber hinaus dedizierte Accounts für die Softwareinstallation und -pflege zur Verfügung.

[10] `https://www.crossref.org/`
[11] `https://orcid.org/`

Sämtliche Leistungen von FIZ Karlsruhe sowie die Mitwirkungspflichten einer RADAR Local nutzenden Einrichtung werden im „Dienstleistungsvertrag über die Nutzung von RADAR in kundeneigener Ablaufumgebung“[12] geregelt.

Informationen zu technischen Voraussetzungen und administrativen Vorarbeiten auf Institutionsseite sind als Checkliste[13] auf der RADAR-Website verfügbar.

4 Ausblick

Als Ergebnis der gemeinsamen Kooperation kann die Universität Konstanz mit dem fertigen Repositorium KonDATA zügig den gewünschten Service anbieten. Mit der reinen Bereitstellung von virtuellen Maschinen und Storage ist der Ressourcenaufwand auf der technischen Seite sehr gering. FIZ Karlsruhe kümmert sich um den Betrieb, die Wartung, die Pflege und die Weiterentwicklung. Als Dienstanbieter fungiert das KIM, das die IT-Infrastruktur bereitstellt, die Nutzenden unterstützt und die Daten kuratiert, Schulungen anbietet und Marketing für KonDATA betreibt.

Im Vergleich zu einer Eigenentwicklung wurde das gewünschte Ziel schneller und günstiger erreicht. Auch der dauerhafte Pflegeaufwand ist reduziert. Spannend bleibt die Frage, wie sich die Weiterentwicklung in Zukunft gestalten wird. Die Universität Konstanz ist Mitglied im Nutzerbeirat von RADAR geworden und kann somit aktiv Einfluss auf zukünftige Features und Verbesserungen nehmen. Ebenfalls herausfordernd wird die weitere Integration in die Dienstelandschaft an der Universität, da die Kontrolle über das System nicht vollständig in der eigenen Hand liegt. In der Gesamtschau überwiegen aber eindeutig die Vorteile. Die nicht vollständige Individualisierbarkeit wird durch ein nach dem Stand der Technik und mit geringem Aufwand betriebenes Repositorium mehrfach aufgewogen.

ORCID IDs

- Matthias Landwehr https://orcid.org/0000-0001-9274-2578
- Gabriel Schneider https://orcid.org/0000-0001-6573-3115
- Stefan Hofmann https://orcid.org/0000-0003-0790-112X
- Matthias Razum https://orcid.org/0000-0002-5139-5511
- Kerstin Soltau https://orcid.org/0000-0002-6368-1929

[12] https://www.radar-service.eu/index.php/de/nutzungshinweise
[13] https://www.radar-service.eu/sites/default/files/RADAR_Local_Checkliste.pdf

A Cloud-based Infrastructure for Interactive Analysis of RNFLT Data

Thomas Peschel[1,2], Mengyu Wang[1,4], Toralf Kirsten[1,2,3,5], Franziska G Rauscher[1,2,*] and Tobias Elze[1,4]

[1]Leipzig Research Centre for Civilization Diseases (LIFE), Leipzig University, Germany
[2]Institute for Medical Informatics, Statistics, and Epidemiology, Leipzig University, Germany
[3]Institute for Medical Data Science, Leipzig University Medical Center, Germany
[4]Schepens Eye Research Institute, Harvard Medical School, Boston, MA, USA
[5]Applied Computer Science and Biosciences, University of Applied Sciences Mittweida, Mittweida, Germany
[*]Corresponding author: Franziska G. Rauscher, franziska.rauscher@medizin.uni-leipzig.de

Good functional vision until old age benefits from early detection of eye diseases. Investigation of retinal structure enables monitoring of eye health. To detect early changes potentially leading to optic neuropathies, such as Glaucoma, retinal nerve fiber layer thickness (RNFLT) is measured with optical coherence tomography (OCT) as an early marker. Different manufacturers provide OCT devices, most implement a 'normative database' allowing to interpret RNFLT immediately after the measurement by the ophthalmologist.

However, vendor-specific normative data is often neither published nor publicly available. Moreover, most OCT devices provide vendor-specific software for measurement and basic analysis but do not allow to extend or exchange normative data. We address both aspects by a) reusing already created published normative data and b) by designing and providing the RNFLT(D)-Visualizer. The published normative data for RNFLT rely on investigations in a large population-based sample taken from the LIFE Adult study at the Leipzig Research Centre for Civilization Diseases. These normative data represent a reference population, with nearly balanced subsets along dimensions such as age, sex and ocular laterality, against which the RNFLT measurement is compared to detect retinal changes.

The RNFLT(D)-Visualizer is part of the Leipzig Health Atlas (LHA), a larger cloud-based infrastructure. The LHA is a platform providing publication data, novel phenotypes, algorithms, and - as in our case - applications for novel normative data. The RNFLT(D)-Visualizer aims to compare individual patient RNFLT data and their differences to our normative database. In this way, the RNFLT-(D)-Visualizer takes the OCT device read-out as input and visualizes the patient measurement regarding both the vendor-specific normative data and those that are already published by us. Most importantly, the comparison of individual measurements to our normative data allows evaluation with age and sex, as well as specific to eye laterality. Additionally, we en-

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13718 (CC BY-SA 4.0)

able longitudinal data visualization by taking reports from multiple measurements (at different time points) showing the ongoing retinal changes. Moreover, this application is designed to add new RNFLT normative data as they are published and made available in future. The RNFLT(D)-Visualizer is designed as R Shiny application. The application never saves any data that has been imported. Instead it visualizes and allows download of the plotted information. The RNFLT(D)-Visualizer is accessible via an internet browser with no access restriction. It is freely usable for research purposes but presently not yet approved for clinical applications.

1 Introduction

For many diseases of the eye, early diagnosis and accurate monitoring over time is essential to initiate or adjust treatment to prevent the onset or progression of vision loss. Optical coherence tomography (OCT) is a three-dimensional imaging technology frequently applied to visualize the retina for this purpose. Particularly for the diagnosis of optic neuropathies like glaucoma, one of the leading causes of blindness in the developed world, OCT is used to determine retinal nerve fiber layer thickness (RNFLT), as the thinning of the nerve fiber layer indicates disease onset or progression. To determine if RNFLT is abnormally thin, the normal range of RNFLT needs to be appropriately determined and available at the time of the clinical assessment of the eye. For this purpose, OCT manufacturers provide their machines with custom normative databases to highlight those retinal locations which present as abnormally thin during an ophthalmic measurement.

The manufacturer RNFLT databases, however, are typically based on only a few hundred healthy subjects over a large age range (usually between around 20 to 80 years) and only consider age as a normative parameter, which might be insufficient to represent the natural variation in healthy eyes, which may result in false positive or false negative abnormality marks. By analyzing large datasets of clinical OCT measurements, we could previously show, for example, that individual anatomical characteristics associated with myopia were associated with machine induced abnormality patterns related to overdiagnosis or missing of glaucoma [1, 2].

The availability of more personalized RNFLT norms would therefore improve the clinical diagnosis of eye diseases, which requires OCT measurements of larger populations of healthy individuals. The population-based Leipzig Research Centre for Civilization Diseases – LIFE Adult study [3] acquired retinal OCT scans of about 10,000 randomly chosen, age and sex stratified participants from the medium-sized city of Leipzig, Germany. We previously used this large dataset to investigate additional parameters to explain RNFLT variance and to generate novel norms based not only on age but also ocular magnification [4], sex [5], or ocular laterality [6].

While our novel normative data are publicly available and accessible to researchers, practicing clinicians might be challenged by the effort to work with raw data from repositories and would typically still rely on machine generated printouts based on the insufficient de-

vice manufacturer norms. The usability of our new norms would be strongly facilitated by an interface that accepts existing machine printouts as an input and generates a graphical representation very similar to the original printout but replaces the machine norms with our new normative data.

Here, we provide such an application as part of a larger cloud-based infrastructure, the so-called Leipzig Health Atlas (LHA), `https://www.health-atlas.de`. The LHA is a platform that provides the publication and visualization of raw data, novel phenotypes, algorithms, or - as in our case - novel normative data to provide more personalized abnormality marks for clinical retinal OCT scans.

The LHA provides an environment for our app, termed RNFLT(D)-Visualizer, that makes the aformentioned normative data accessible to clinicians and scientists. The application is specifically aimed at making the new norms available for individual comparison of measurement data. In the following, we describe the input data that clinicians can incorporate into the RNFLT(D)-Visualizer. Furthermore, we detail the cloud environment for running this application.

2 Materials and methods

Investigation of retinal structure enables monitoring of eye health. To detect early changes potentially leading to optic neuropathies, such as Glaucoma, retinal nerve fibre layer thickness is measured with optical coherence tomography (OCT) as an early marker. Different manufacturers provide OCT devices, these implement a normative database allowing to interpret RNFLT immediately after the measurement by the ophthalmologist. The delivered normative data should rely on a large cohort, ideally a population-based sample. Therefore, we have created new normative data for RNFLT based on investigations in the large population-based Leipzig Research Centre for Civilization Diseases - LIFE Adult study [4], [5], [6]. Our normative data establishes a reference population, with nearly balanced subsets along dimensions such as age, sex and ocular laterality, against which the RNFLT measurement is compared to detect retinal changes.

2.1 Input data

Acquired raw data is typically measured for 768 equally distant points arranged in a circle around the optic nerve head. OCT devices provide a read-out of such data in a proprietary report customized for each patient that can be exported into a PDF file. The benefit of the app is the provision and further processing of measurements of retinal nerve fiber layer thickness routinely determined with an OCT device and available in digital form as a PDF document. The patient data prepared in such documents enable the eye specialist to identify early signs of disease via the graphic presentations, which are provided in front of a background of manufacturer-specific normative data.

Further processing of such patient data, such as comparison with other normative reference values, is thus not possible. This problem is solved by the RNFLT(D)-Visualizer.

The RNFLT(D)-Visualizer can directly import such PDF files of a specific patient, extract the measured data from diagrams included in the PDF file and visualizes measurements in comparison to normative data.

There are different sources for normative data. First, OCT device manufacturers typically provide such data with the read-out PDF file. Our application extracts this kind of information for visualizations. Secondly, we employ normative data that has been created based on the large population-based LIFE Adult study. For the RNFLT(D)-Visualizer, we refer to the Wang and Elze et al. normative data [4] and their inter-ocular differences [6], sex-specific differences can also be accounted for, this will be implemented into the app in due course [5]. Therefore, the RNFLT(D)-Visualizer can provide individual comparison to different normative data, taking into account refractive error (i.e. measurement diameter) and allowing evaluation with age and sex, specific for eye laterality.

2.2 Leipzig Health Atlas

The RNFLT(D)-Visualizer is part of the Leipzig Health Atlas which provides publications, corresponding data, (e.g. supplements), and methods. Such entities can be uploaded by registered scientists. The LHA structure contains projects, where publications and applications and their meta-data are linked to each other. Apps such as the RNFLT(D)-Visualizer are accessed via an URL that is managed and provided by the LHA. These apps, as well as associated data and documents, can also be found through a search on the LHA website `https://www.health-atlas.de` if the user only has topic keywords. The RNFLT(D)-Visualizer can be accessed via `https://apps.health-atlas.de/rnflt-visualizer`. Typically, applications are free to use, only a few are restricted to a pre-specified user group. Due to medical device regulations applications are aimed for research purposes. For clinical settings we provide a disclaimer at the moment.

The LHA environment with all linked applications is managed by a cloud-based Kubernetes infrastructure which is physically hosted by Leipzig University computing centre. This infrastructure requires to package each new application using up-to-date containerization technology, such as Docker. We use the University GitLab system to manage the application source code from which a Docker image is automatically created and shipped to the Kubernetes infrastructure (continuous deployment) whenever changes to the source code arrive. This speeds up the process bringing the latest development to the public. Moreover, the app is provided by a R Shiny server that allows to scale up / down when new requirements, e.g., a higher number of concurrent users, arise.Additionally, the app's environment in the Docker container can be customized to meet other requirements such as the R version or package dependencies.

2.3 RNFLT(D)-Visualizer application

The RNFLT(D)-Visualizer is developed as a R Shiny app. Shiny `https://shiny.rstudio.com` is a R package enabling quick and easy web application development in R for visualizing statistical data.

A R Shiny web application is made available via a Shiny server and, thus, allows multiple clients to use the app in parallel. A modern responsive design was realized by the package bs4Dash [7] through its bootstrap functionality. Moreover, the R packages ShinyWidgets [8] and ShinyJS [9] are used to improve the visual impact. All plots are made with the Plotly-Package [10]. The R-package used for statistical computations is GAMLSS [11]

The RNFLT(D)-Visualizer is structured in sections e.g. 'Import Data' section or 'Analysis' section, accessible on the left panel, see Figure 1.

3 Results

In the following, we describe and discuss different steps along the analysis workflow of a typical usage of the RNFLT(D)-Visualizer.

3.1 User specific data import

A document is uploaded by the user and stored only for the period of the session. For demonstration purposes, we also provide some sample files which can be directly used to feed the application. After uploading a PDF file, the RNFLT(D)-Visualizer provides an overview in tabular form, see Figure 1. This table provides information of the measurement data, the patient's age, the eye side, and the available norms the data can be analyzed with. Moreover, some analyses require further data, for instance, the diameter of the RNFLT measurement ring and Bruch's Membrane Opening (bmo). The user can add and modify this data directly in this table. The position of graphics in the PDF file is arbitrary. This is caused by the OCT device software, since there are a number of different document types depending on the measurement and the type of generated print-out. In a first step, the app extracts meta-data about the examination, (red framed box in Figure 1), such as specified patient birth date, date of measurement, eye side, and sex. This information extraction is realized by the help of the R package 'pdftools' [12]. It extracts the complete text of the PDF document and provides it as a single string. Finding and extracting relevant data is easily implemented by using regular expressions.

After uploading patient data into the application in the Import Data Section, the user can edit missing meta data, as illustrated in Figure 1. These meta data are measurement diameter, specific eye (ocular laterality), sex, age, and the Bruch's Membrane Opening (for further invesigations). After checking that all necessary data were correctly uploaded or entered, the user stores the PDF with its (amended) meta data in the Data Store

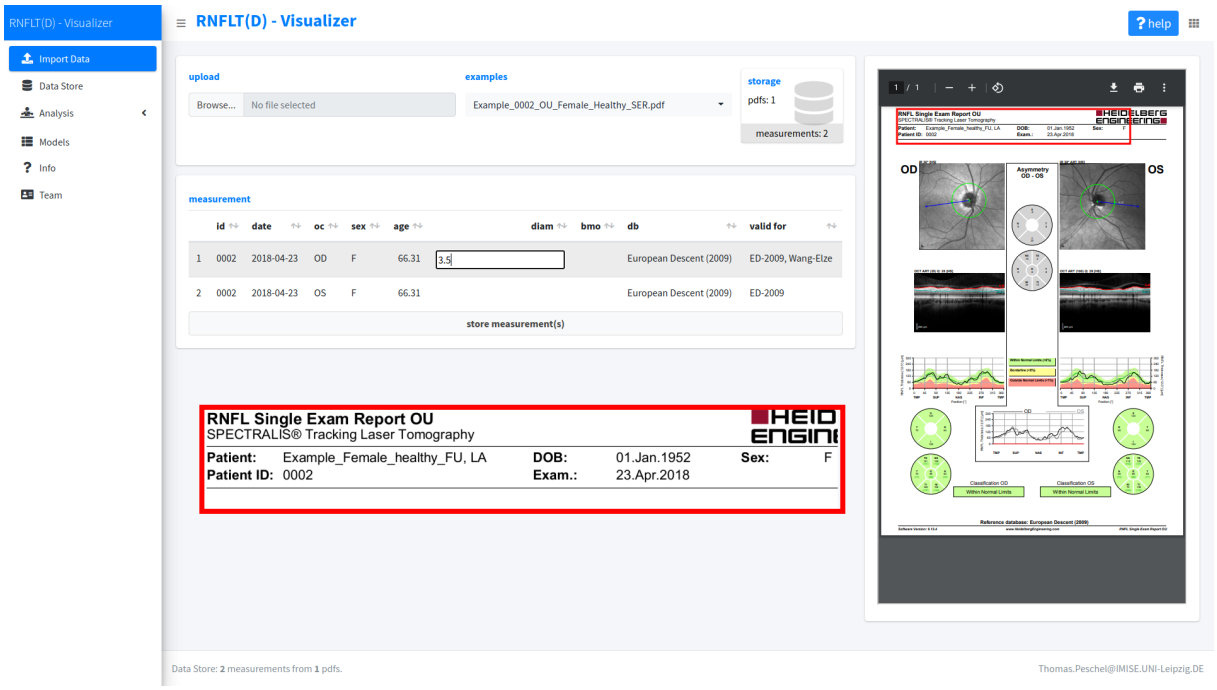


Figure 1: RNFLT(D)-Visualizer during data import. This scene shows the app directly after an upload of an OCT PDF. As shown, the user here adds the measurement diameter value in the presented table. After a click on the button 'store measurement(s)', the data will be scanned and stored with the meta data. The red frame in the middle is only used to highlight this zoom of the PDF's header to show how meta data are extracted from the document's header and copied to the table.

Section by pressing the button 'store measurement(s)'. Now the application extracts the measurement data by searching and scanning the graphs inside the document. That is why it also works for unusual formats, e.g. PDF with notes footer.

Next, measurements are extracted. Both, the patient-specific measurements and the related normative data, are include graphically in the PDF file in plots. And also some meta-data are placed in the header and additionally in the footer sometimes. Therefore and depending on the document type, the positions of graphics need to be determined in a second step before data values can be read out. In order to detect the graphics' positions, the PDF file is rendered at a low resolution of 72 DPI. This allows a approximate detection of position of the plots.

In step three, the imported PDF document is then rendered at a resolution of 312 DPI and the measurement data and background percentile data are scanned and internally managed in a tabular form (R data type Data Frame). In case the user missed to properly edit meta data, further information can be added in the Data Store Section. So it is possible to upload data in an unusual format. For example, a PDF with missing header can be analyzed. Additional information with relevance to norms, e.g. age and sex, can be manually entered into the application table of meta data.

Having tabular measurement data available, it can either be downloaded for the user's own investigations or subsequently analyzed with this app, in particular, the RNFLT-related analyses and inter-ocular RNFLT differences. The app allows to upload multiple PDF files for the same or different patients. All data is managed within the app but only saved temporarily while the session runs. The user can select the data set in the 'Analysis' section by selecting the file name and the meta-data of the specific measurement.

The imported data sets can be analyzed in two different ways, the retinal nerve fiber layer thickness and inter-ocular retinal nerve fiber layer thickness differences. We describe both in the next two subsections.

3.2 Retinal nerve fiber layer thickness analysis

The retinal nerve fiber layer thickness analysis is comparable to that already available on the OCT device but the app supplies improved analysis methods. In particular, the app provides multiple normative reference values (e.g. Wang and Elze et al. norm) allowing to compare patient measurements with OCT device built-in normative data (e.g. 'European-Descent-2009'). The app is constructed to enable implementation of future normative data if and when they become available to extend the opportunities of analysis.

For all visualizations, the user can select the normative data-set for analysis. Values of European-Descent-2009 norm do not factor in age, sex and measurement diameter. In contrast, the normative data by Wang and Elze et al. specifically take these two variables into account. As an extension, sex will be included as a further factor.

Based on imported patient measurements, the app provides different visualizations, each of them is available by tabs on the top. Figure 2 shows four different visualizations as a montage. The top two graphs represent visualizations of the sector analysis and the lower two graphs represent the continuous analysis (768 individual equally distributed measuring points of the 360° ring scan). Both use normative reference values, selectable by radio buttons in the menu seen on the top of the montage, as comparative data. Second and fourth graph are the z-standardized representation of first and third graph. Each visualization consists of two parts, a polar plot on the left and a colored line chart on the right, representing the same data.

The polar plot shows data taken from 768 measuring points on a circle around the optic nerve head of the human eye together with the selected normative data. In the sector analyses, the 360° ring scans are divided into sectors of either 45° or 90° as common for such RNFLT data analysis in Ophthalmology. Parameters are averaged in each sector. Such sectors are employed to analyse directions of retinal nerve fiber layer entering the optic nerve of the human eye.

In the RNFLT analysis, the measurement is seen against a percentile background corresponding to the selected reference standard data (European-Descent-2009 or Elze-Wang). In the z-space analysis, the measurement is shown as a standard deviation score or z-score against a z-space background. Unlike the percentile plot, which is essentially the same

as the plot from the PDF document, the z-score plot provides easier access to what is happening. The polar plots are generally more intuitive. To account for eye side, direction of the plots can be selected by the check boxes 'direction'.

The black line in all plots represents the patient measurement whereas the dark green line depicts the 50% percentile, and the green (yellow, red) colored area symbolizes the 95% (5%, 1%) percentiles as taken from normative data. Since the normative data by Wang and Elze et al. account for confounding variables described above, the diameter and the age must be given at import time. In general, age is implicitly given as part of the uploaded PDF document by dates of birth and measurement or can be added afterwards.

3.3 Inter-ocular retinal nerve fiber layer thickness differences

Similar to the RNFLT analysis, for the inter-ocular retinal nerve fiber layer thickness difference analysis we provide four visualizations combining sector and continuous analyses with normative percentiles and z-scores. The sector analysis utilizes a parameter-free Box-Cox-T model over age and the measurement diameter for the absolute magnitude of the nerve fiber layer thickness differences. The Box-Cox-T model models the first four moments (i.e., mean μ, standard deviation σ, skewness ν, and kurtosis τ) of the distribution of absolute thickness values. Conversely, continuous analysis assumes a normal distribution of thickness differences.

3.4 Patient-specific analysis result

Besides the shown visual representation, the RNFLT(D)- Visualizer provides all data in tabular format (see Fig 4) for all analyses. The user can freely access this data and download it in different formats, e.g. xslx, csv, tsv and RData format. All plots can be downloaded in png-format via press on the camera icon inside the plot as seen in the lower graph in Figure 2.

3.5 Patient-specific normative data

Additional to the visual representation described in the last two sub-sections, the RNFLT(D)-Visualizer provides Elze-Wang normative data for both, RNFL thickness and inter-ocular RNFLT differences. The normative data for RNFLT are represented by the parameters μ, standard deviation σ of a normal distribution in the sectors and as well as in the continuum analysis. The sectors of the RNFLTD analysis are described with the Box-Cox-T model with the parameters mean μ, standard deviation σ, skewness ν, and kurtosis τ, and the continuum analysis of RNFLTD is also modelled by a normal distribution.

These parameters are dependent on age, measuring diameter, and measuring diameter differences respectively, which one can choose in the menu of the 'Models' section. All

Figure 2: Montage of the 4 provided RNFLT-Visualization. All pictures show the patient's data against percentiles background (1%, 5% and 95%) related to the selected normative data. The sector analyses use norms averaged on the sectors. The continuum analyses use normative data for each of the 768 measure points. The graphs in the pictures 1 and 3 are plotted in real space where the graphs 2 and 4 are plotted in z-space. In the left column we have polar plots and in the right one filled line charts. The visualizations can be selected by tabs on the top. RNFLT is evaluated in six sectors (T: temporal, TS: temporal-superior, TI: temporal-inferior, N: nasal, NS: nasal-superior, NI: nasal-inferior) and overall average (G; used in table only).

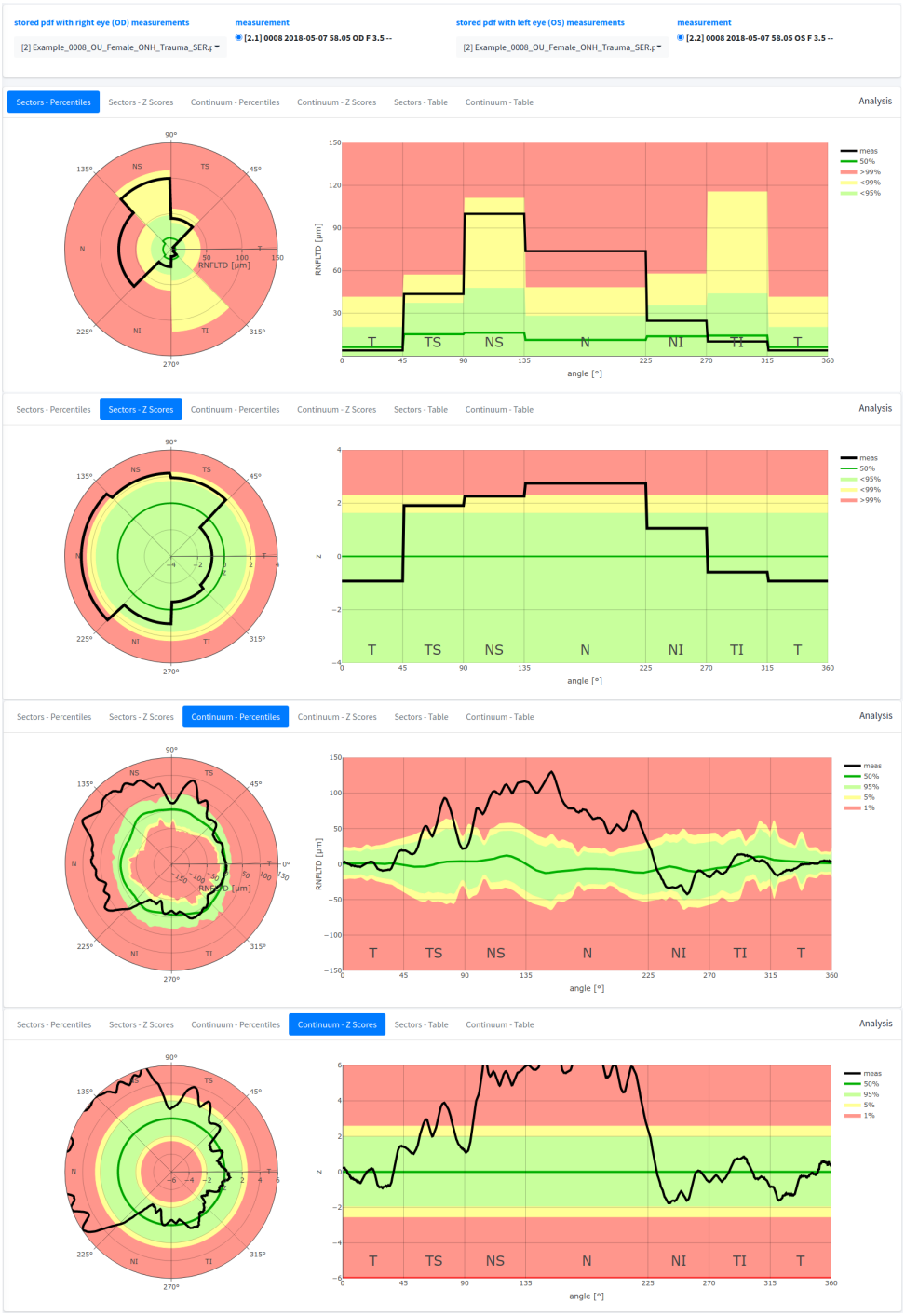


Figure 3: Montage of the four provided RNFLTD presentations. All graphs show the patient's data against percentiles background (1%, 5% and 95%) related to the selected normative data. The sector analyses use Elze-Wang Box-Cox-T normative data computed for each sector. The continuum analysis employs normative data for the difference of each of the 768 measuring points of both eyes. The first and third graphs are plotted in real space whereas the second and fourth graphs are plotted in z-space. Similar to the RNFLT plots, the left column depicts polar plots and the right column depicts filled line charts.

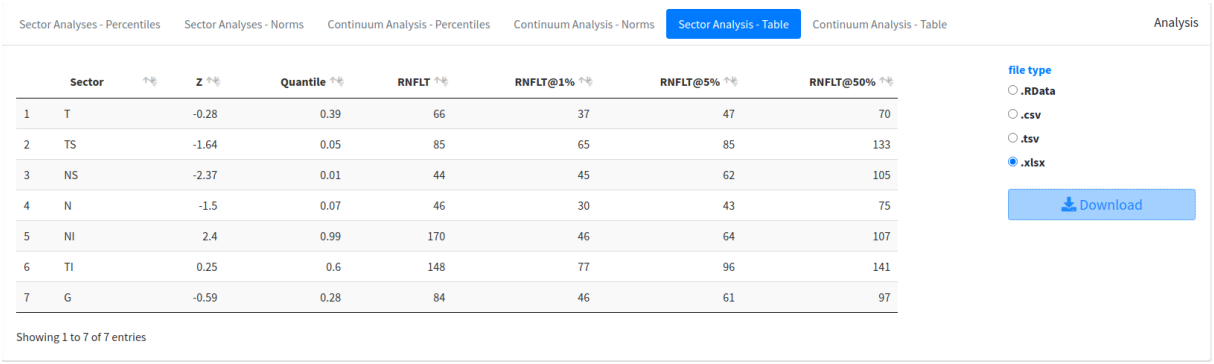


Figure 4: Table of data for a RNFLTD sector analysis ready for download. The columns are sector id, z-score and quantile of the patient's measurement followed by the thickness difference values related to Elze-Wang normative data that includes 1%, 5% and 50% of measurements assuming a Box-Cox-T-Distribution.

plots and their related tables are therefore interactive. For the thickness we also provide a plot that shows the differences between the models of Elze-Wang and the European-Descent-2009. This can be found in the section 'Models', see Figure 5.

The parameters of normative data are visualized either as bar charts for sectors or in a line chart for continuum analysis. All graphics and tabular data can be downloaded (see Fig. 5) by the user for own examinations and further analysis purposes.

3.6 Help

We include a context specific help button for every section of the app. This is a specific help window as seen in Figure 6. It can be accessed by pressing the help button in the top right corner of the app as seen in Figure 6.

4 Conclusions

Our app employs the latest technique using R Shiny app and Docker/Kubernetes with continuous integration via Gitlab. This set of software offer a solution for fast and easy web application development and role out. This has lead to a very user friendly graphical user interface with direct access to new reference values for eye specialists. The RNFLT(D)-Visualizer app offers the latest normative data, which is not accessible directly otherwise.

For the first time the app offers new opportunities for retinal nerve fiber layer analysis. Furthermore, the eye specialist or scientist can directly and easily apply the new normative data to analyse own patient measurements on the basis of the latest normative reference values.

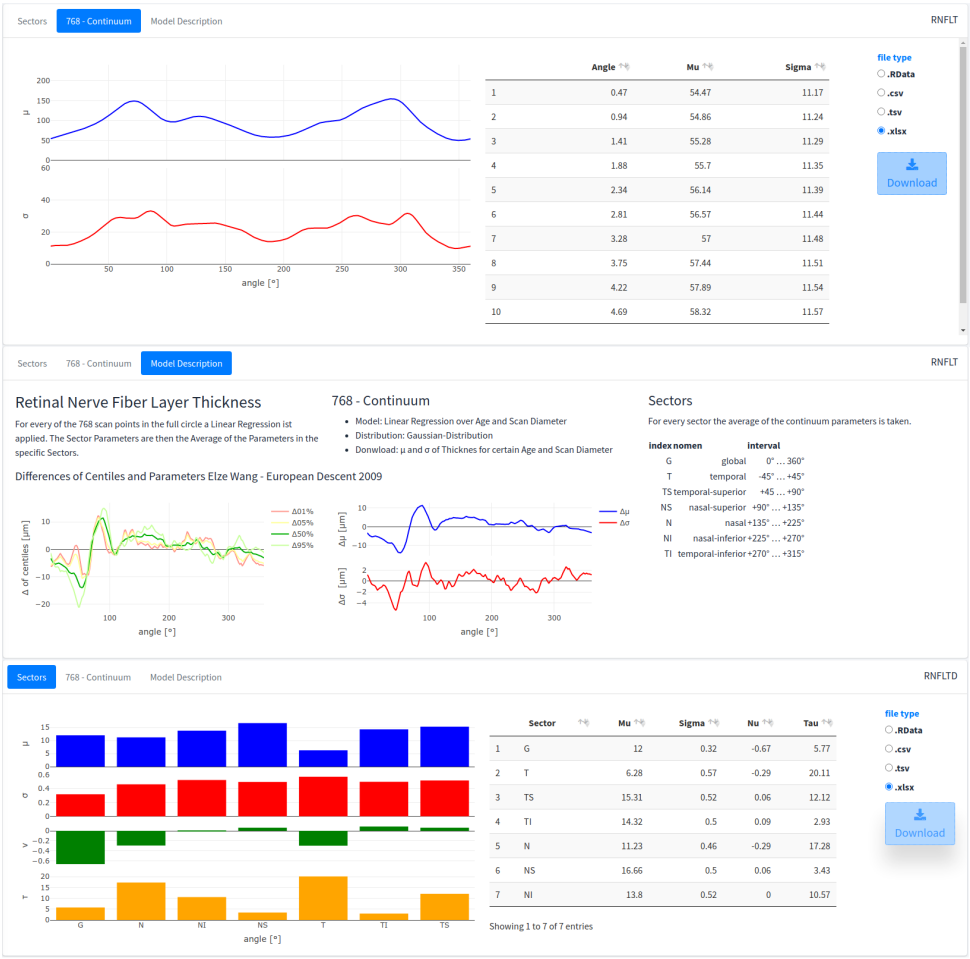


Figure 5: Montage of screenshots of the 'Models' section. The two upper graphs show the RNFLT model and the lower graph presents the Box-Cox-T model for the magnitude of thickness differences in the specific sectors.

This enables new ways of detection of early changes of the retinal nerve fiber layer thickness and thickness differences.

5 Outlook

In future, the app will implement our current work on sex [5] and ocular laterality [6], taking into account these influencing factors to enable finer differentiation of early signs of disease. Furthermore, reports will be offered for download in PDF and HTML format. Additionally, we will enable longitudinal data visualization by taking reports from multiple measurements (at different time points) showing the ongoing retinal changes.

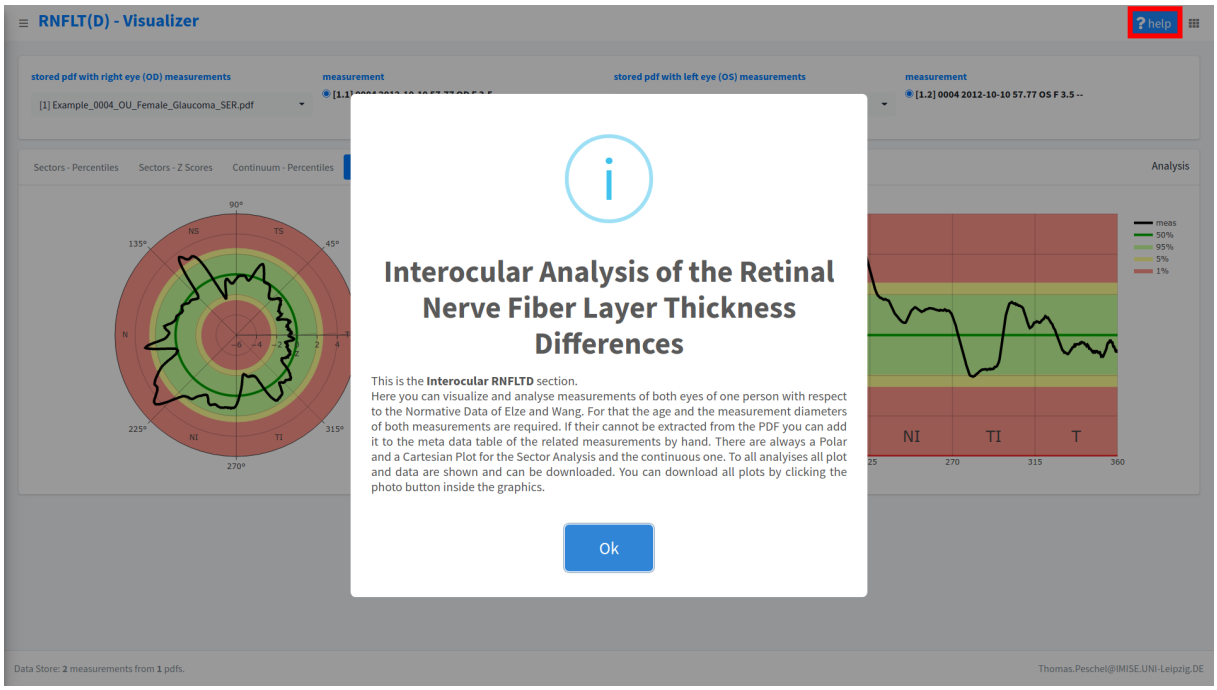


Figure 6: Example of help window for 'inter-ocular RNFLTD' section. In this context the help button highlights the necessary steps for the user.

Acknowledgements

The authors wish to thank the LIFE Adult study participants for their time and furthermore we gratefully acknowledge the LIFE Adult study team for their commitment to the eye investigation and corresponding exams to make this analysis possible.

Funding

This research was supported by LIFE Leipzig Research Center for Civilization Diseases, Leipzig University (LIFE is funded by the EU, the European Social Fund, the European Regional Development Fund, and Free State Saxony's excellence initiative; project numbers: 713-241202, 14505/2470, 14575/2470); Lions Foundation; Grimshaw-Gudewicz Foundation; Research to Prevent Blindness; BrightFocus Foundation; Alice Adler Fellowship; NIH K99EY028631 to MW; NEI Core Grant P30EYE003790; NIH R21 EY030142; R21EY030631; R01EY030575; German Federal Ministry of Education and Research i:DS-em - Integrative data semantics in systems medicine (031L0026).

The sponsor or funding organization had no role in the design or conduct of this research.

Bibliography

[1] Tobias Elze, Neda Baniasadi, Qingying Jin, Hui Wang, and Mengyu Wang. Ametropia, retinal anatomy, and OCT abnormality patterns in glaucoma. 1. impacts of refractive error and interartery angle. *Journal of Biomedical Optics*, 22(12):121713, 2017.

[2] Neda Baniasadi, Mengyu Wang, Hui Wang, Qingying Jin, and Tobias Elze. Ametropia, retinal anatomy, and OCT abnormality patterns in glaucoma. 2. impacts of optic nerve head parameters. *Journal of Biomedical Optics*, 22(12):121714, 2017.

[3] Markus Loeffler, Christoph Engel, Peter Ahnert, Dorothee Alfermann, Katrin Arelin, Ronny Baber, Frank Beutner, Hans Binder, Elmar Braehler, Ralph Burkhardt, Uta Ceglarek, Cornelia Enzenbach, Michael Fuchs, Heide Glaesmer, Friederike Girlich, Andreas Hagendorff, Madlen Haentzsch, Ulrich Hegerl, Sylvia Henger, Tilman Hensch, Andreas Hinz, Volker Holzendorf, Daniela Husser, Anette Kersting, Alexander Kiel, Toralf Kirsten, Juergen Kratzsch, Knut Krohn, Tobias Luck, Susanne Melzer, Jeffrey Netto, Matthias Nuechter, Matthias Raschpichler, Franziska G. Rauscher, Steffi G. Riedel-heller, Christian Sander, Markus Scholz, Peter Schoenknecht, Matthias L. Schroeter, Jan-Christoph Simon, Ronald Speer, Julia Staeker, Robert Stein, Yve Stoebel-richter, Michael Stumvoll, Attila Tarnok, Andrej Teren, Daniel Teupser, Francisca S. Then, Anke Toenjes, Regina Treudler, Arno Villringer, Alexander Weissgerber, Peter Wiedemann, Silke Zachariae, Kerstin Wirkner, and Joachim Thiery. The LIFE-adult-study: Objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in germany. *BMC Public Health*, 15:691, 2015.

[4] Mengyu Wang, Tobias Elze, Dian Li, Neda Baniasadi, Kerstin Wirkner, Toralf Kirsten, Joachim Thiery, Markus Loeffler, Christoph Engel, and Franziska G. Rauscher. Age, ocular magnification, and circumpapillary retinal nerve fiber layer thickness. *Journal of Biomedical Optics*, 22(12):121718, 2017.

[5] Dian Li, Franziska G. Rauscher, Eun Young Choi, Mengyu Wang, Neda Baniasadi, Kerstin Wirkner, Toralf Kirsten, Joachim Thiery, Christoph Engel, Markus Loeffler, and Tobias Elze. Sex-specific differences in circumpapillary retinal nerve fiber layer thickness. *Ophthalmology*, 127(3):357–368, 2020. doi:https://doi.org/10.1016/j.ophtha.2019.09.019.

[6] Neda Baniasadi, Franziska G. Rauscher, Dian Li, Mengyu Wang, Eun Young Choi, Hui Wang, Thomas Peschel, Kerstin Wirkner, Toralf Kirsten, Joachim Thiery, Christoph Engel, Markus Loeffler, and Tobias Elze. Norms of interocular circumpapillary retinal nerve fiber layer thickness differences at 768 retinal locations. *Translational Vision Science & Technology*, 9(9):23, 2020. doi:https://doi.org/10.1167/tvst.9.9.23.

[7] David Granjon, RinteRface, Almasaeed Studio (AdminLTE3 theme for Bootstrap 4), Winston Chang (Utils functions from shinydashboard), and Thomas Park (Bootswatch Sketchy theme CSS). *bs4Dash: A 'Bootstrap 4' Version of 'shinydashboard'*, 2019. R package version 0.5.0.

[8] Victor Perrier, Fanny Meyer, David Granjon, Ian Fellows, Wil Davis, and Spencer Matthews. *shinyWidgets: Custom Inputs Widgets for Shiny*, 2021. R package version 0.6.0.

[9] Dean Attali. *shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds*, 2020. R package version 2.0.0.

[10] Carson Sievert, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec, Pedro Despouy, and Salim Brüggemann. *plotly: Create Interactive Web Graphics via 'plotly.js'*, 2021. R package version 4.9.4.1.

[11] Mikis Stasinopoulos and Bob Rigby. *gamlss: Generalised Additive Models for Location Scale and Shape*, 2021. R package version 5.3-4.

[12] Jeroen Ooms. *pdftools: Text Extraction, Rendering and Converting of PDF Documents*, 2021. R package version 3.0.1.

MoveApps - Etablierung eines Dienstes zur Entwicklung, Veröffentlichung und langfristigen Nachnutzung fachspezifischer Forschungssoftware

Gabriel Schneider[1], Andrea Kölzsch[2] und Kamran Safi[2]

[1]Kommunikations-, Informations-, Medienzentrum, Universität Konstanz
[2]Max-Planck-Institut für Verhaltensbiologie

Der Beitrag stellt MoveApps, einen Software-Dienst zur Verfügbarmachung, Nutzung und langfristigen Speicherung von Forschungssoftware vor, welcher gemeinsam unter der Leitung des Max-Planck-Instituts für Verhaltensbiologie (MPIAB) in Kooperation mit dem Kommunikations-, Informations-, Medienzentrum (KIM) der Universität Konstanz entwickelt wird. MoveApps ermöglicht es seinen Nutzer*innen, Applikationen (kurz Apps) zur Analyse fachspezifischer Forschungsdaten selbst zu programmieren und zur Verfügung zu stellen oder auf die Apps anderer Nutzer*innen zurückzugreifen. Die Apps sind dabei als modulare Bausteine konzipiert, die in verschiedenen Kombinationen, abhängig von ihren Ein- und Ausgabewerten, hintereinandergeschaltet werden können, um komplexe Analyseabläufe (sogenannte Workflows) zu realisieren. Durch den modularen Ansatz wird sowohl der Aufwand der Software-Entwicklung für einzelne Entwickler*innen gesenkt als auch die Konfiguration komplexer Workflows für Wissenschaftler*innen mit geringen Programmierkenntnissen ermöglicht.

Um die Workflows langfristig nachvollzieh- und zitierbar zu machen, werden sie mit umfangreichen Metadaten angereichert, exportiert und persistent in einem Repositorium gespeichert. In Kombination mit den dort ebenfalls veröffentlichten Forschungsdaten machen Wissenschaftler*innen ihre Forschungsergebnisse so nachvollzieh- und nachnutzbar für Dritte. Gleichzeitig werden durch die Zitierbarkeit von Workflows (sowie für Apps über MoveApps selbst) auch die Entwickler*innen von Forschungssoftware sichtbar. Der Beitrag beschreibt das Konzept von MoveApps, dessen Aufbau und wie damit die Bedarfe einer fachspezifischen Wissenschaftsdisziplin bedient werden können. Dazu werden Use-Cases aus der Praxis vorgestellt. Weiterhin wird beleuchtet, wie die Apps entwickelt, kombiniert, mit Metadaten beschrieben und als Workflows veröffentlicht werden können. Dabei wird die gemeinsame Arbeit des MPIAB mit dem KIM der Universität Konstanz herausgestellt und abschließend die daraus gewonnenen Mehrwerte präsentiert.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13719

1 Einleitung

Durch den gleichzeitigen Zugang zu Daten, Software und deren Dokumentation kann die Nachvollziehbarkeit von Forschungsergebnissen gesteigert werden. Während für die Veröffentlichung von Forschungsdaten durch die Bereitstellung von Datenrepositorien in letzter Zeit vielfältige Möglichkeiten geschaffen wurden, fehlen im Bereich der Forschungssoftware oftmals noch Angebote, mit denen die Fachcommunities Software sammeln und zugänglich machen können. Zusätzlich steigen aufgrund immer datenintensiverer Forschungsmethoden die benötigten Programmierkenntnisse zur Analyse von Forschungsdaten in vielen Fachdisziplinen an. Dies macht eine stärkere Vernetzung zwischen Personen mit hohen Programmierkenntnissen und praxisorientierten Forscher*innen nötig.

Mit „Movebank 2.0“ fördert das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg von 2019 bis 2023 ein Projekt, das die Aktualisierung der Tierbewegungsdatenbank Movebank (movebank.org)[1]zum Ziel hat. Dabei arbeiten das Max-Planck-Institut für Verhaltensbiologie (MPIAB), das Kommunikations-, Informations-, Medienzentrum (KIM) der Universität Konstanz und der Lehrstuhl für Informatik in den Lebenswissenschaften der Universität Konstanz in verschiedenen Bereichen an der Weiterentwicklung der Plattform. Zu den Zielen gehört unter anderem die Erweiterung des Metadatenschemas, um Daten zu unterstützen, die durch neue Sensoren aufgenommen werden. Weiterhin sollen Bürgerwissenschaftler*innen, die Apps wie den „Animal Tracker“[2] nutzen, bei der Arbeit mit Movebank unterstützt werden.

Um Movebank herum hat sich eine aktive Open Source-Community gebildet. Diese entwickelt Anwendungen zur Analyse der verfügbaren Tierbewegungsdaten und stellt sie bereit. Um diese Arbeiten zu fördern und mehr Nutzer*innen einen vereinfachten Zugang zu den Anwendungen zu gewähren, entsteht im Kontext des Projekts der Dienst MoveApps.

MoveApps (moveapps.org) ist eine No-Code-Plattform, die es Wissenschaftler*innen ermöglicht, Tierbewegungsdaten von Movebank zu analysieren, ohne selbst ausgeprägte Programmierkenntnisse besitzen zu müssen. Auf der Oberfläche des Dienstes können modulare Analyseprogramme (Apps genannt) ausgewählt, konfiguriert und hintereinandergeschaltet werden, um komplexe Analyseabläufe durchzuführen. Diese kombinierten Abläufe werden Workflows genannt. Die Berechnung der Analyse läuft auf Infrastruktur der Max-Planck-Gesellschaft ab, Nutzer*innen benötigen lediglich einen Internetzugang, um auf die Weboberfläche des Dienstes zuzugreifen. Dabei läuft jede ausgewählte App in einer eigenen Container-Umgebung (Docker), die nach der Berechnung die Ergebnisdaten an den nächsten Container im Workflow weiterreicht.

Sollten die Nutzer*innen bei der Nutzung von MoveApps Workflows erstellen, die in der Veröffentlichung einer wissenschaftlichen Publikation münden, können sie die Workflows im Movebank Data Repository (www.datarepository.movebank.org) veröffentlichen. Das Repositorium ist ein Dienst, den das KIM der Universität Konstanz seit 2012 in enger Kooperation mit dem MPIAB betreibt. Auf der Plattform werden bisher ausgewählte

[1]https://www.movebank.org
[2]https://www.icarus.mpg.de/29143/animal-tracker-app

Datensätze von Movebank veröffentlicht, mit einem persistenten Identifikator versehen sowie langfristig und frei zugänglich gespeichert. In Zukunft werden auch die Workflows aus MoveApps hier veröffentlicht und persistent nachgewiesen.

2 Motivation hinter MoveApps

Die Grundlage für Tierbewegungsforschung stellt das Analysieren von Tierbewegungsdaten dar, die über Sensoren an Tieren gesammelt werden. Dazu werden die Tiere in der Natur kurzzeitig gefangen, mit einem Sensor versehen und anschließend wieder frei gelassen. Der angebrachte Sensor überträgt fortlaufend Daten, die Wissenschaftler*innen auswerten können, um beispielsweise darin Muster zu finden, die zur Beantwortung von Forschungsfragen genutzt werden können[3]. Eine Plattform zur Standardisierung, Speicherung und Veröffentlichung der Daten ist Movebank.

Innerhalb dieser Forschungscommunity gibt es verschiedene Ausrichtungen. Auf der einen Seite stehen eher praktisch orientierte Forscher*innen, deren Hauptaugenmerk auf der Besenderung der Tiere liegt. Sie führen Feldforschung durch und bereiten mit dem Anbringen der Sensoren die Grundlage für die spätere Analyse der Daten vor. Ergänzend zu diesen Forscher*innen arbeitet eine andere Gruppe von Wissenschaftler*innen an den benötigten Werkzeugen für die Analyse, indem sie Analysesoftware entwickeln und diese bereitstellen. MoveApps soll diese beiden Nutzergruppen zusammenbringen.

Die praktisch orientierten Forscher*innen können auf MoveApps die Werkzeuge finden, die sie für die Analyse der von ihnen gesammelten Daten benötigen. Durch eine intuitive Web-Oberfläche reichen IT-Anwenderkenntnisse aus, um die bereitgestellten Programme zu nutzen. Der modulare Charakter der Apps senkt die Komplexität für die Einzelbestandteile, dennoch lassen sich durch die Kombination einzelner Apps detaillierte Analyseabläufe realisieren. Dadurch, dass die eigentliche Berechnung der Ergebnisse auf Infrastruktur der Max-Planck-Gesellschaft durchgeführt wird, benötigen Nutzer*innen selbst keine leistungsstarke Hardware vor Ort. Somit sind auch Analysen von Daten möglich, wenn sich Forscher*innen gerade im Feld befinden und ihnen somit nicht die gewohnte IT-Infrastruktur ihres Instituts zur Verfügung steht.

Für die Entwickler*innen von Analyse-Software bietet MoveApps den Vorteil, dass ihnen die Plattform eine Veröffentlichung ermöglicht, ohne dass sie dafür die Rechte an ihrer Software abgeben müssen. Sie stellen die Software bei einer Veröffentlichung zwar unter einer offenen Lizenz bereit, die anderen eine Nachnutzung ermöglicht, jedoch wird ihre Arbeit als wissenschaftliches Produkt durch automatisch erstelle Zitationsangaben gewürdigt. Durch die Modularität können sie sich auf die Optimierung ihrer App konzentrieren, da der Funktionsumfang jeder einzelnen App gering ist.

[3] Kays, R., Crofoot, M. C., Jetz, W., & Wikelski, M.: Terrestrial animal tracking as an eye on life and planet. Science, Band 348, Heft 6240 aaa2478–aaa2478 (2015). https://doi.org/10.1126/science.aaa2478.

Abbildung 1: Ausgangssituation für Workflows.

In den nachfolgenden Abschnitten wird die Funktionsweise von MoveApps genauer beleuchtet und dabei auf das Nachnutzen, Entwickeln und Veröffentlichen von Apps und Workflows, eingegangen.

3 Daten mit MoveApps analysieren

Sollte ein/e Nutzer*in keine ausreichenden Programmierkenntnisse besitzen oder sich auf das Durchführen von Analysen fokussieren wollen, kann er/sie die von anderen Nutzer*innen veröffentlichten Apps dafür nutzen. Über eine grafische Oberfläche können Apps hintereinandergeschaltet werden, um so komplexe Analyseabläufe zu realisieren. Ein neuer Workflow kann dabei über das Dashboard auf MoveApps zusammengestellt werden, sobald ein/e Nutzer*in einen kostenfreien Account anlegt.

Zu Beginn eines jeden Workflows steht der Import der gewünschten Datengrundlage. MoveApps bietet hier mehrere Möglichkeiten. Mit der App „Movebank“ können Daten über eine Schnittstelle direkt aus der Datenbank von Movebank heruntergeladen und genutzt werden. Die Voraussetzung für die Nutzung der Daten ist ein Movebank-Account und dort entsprechend konfigurierte Zugriffsrechte für die gewünschten Datensätze. Beim Import der Datenquelle wählen Nutzer*innen aus, welche Teile der Daten sie jeweils für den Workflow benutzen wollen. Dabei können die Datensätze bis auf Einzeltiere eingeschränkt werden, je nachdem, was die jeweilige Forschungsfrage erfordert. Weiterhin können Nutzer*innen die Daten direkt aus einer Cloud beziehen. Aktuell unterstützt MoveApps an dieser Stelle eine Anbindung an Dropbox und Google Drive. Werden die Daten über einen Cloud-Anbieter importiert, müssen sie vorher in das von Movebank verwendete Format[4] überführt werden.

Ausgehend von der Datengrundlange fügen Nutzer*innen nachfolgend weitere Apps zu ihrem Workflow hinzu (siehe Abbildung 1). Dabei stehen ihnen alle Apps zur Verfügung, die zum aktuellen Zeitpunkt für die Plattform freigeschaltet wurden und die von den Ein- und Ausgabe-Werten kompatibel mit der jeweils vorherigen App des aktuellen Workflows sind.

[4]Kranstauber, B. et al.: The Movebank data model for animal tracking. Environmental Modelling & Software, Band 26, Heft 6 (2011): 834-835. https://doi.org/10.1016/j.envsoft.2010.12.005.

Jede App, die dem Workflow angefügt wird, kann anschließend über das entsprechende Menü konfiguriert werden. Zusätzlich können Nutzer*innen weitere Informationen über jede App abrufen. Dazu zählt eine Zitationsangabe, ein Link auf das Git-Repositorium, das den Quellcode der App enthält und die Dokumentation, die der/die Entwickler*in bereitgestellt hat, um die App nachvollziehbar zu machen.

Sobald der/die Nutzer*in alle gewünschten Apps zum Workflow hinzugefügt hat, kann er/sie den Workflow starten und die Analyse wird durchgeführt. Je nach gewählter App, produziert der Workflow anschließend ein Endergebnis. Dies kann zum Beispiel eine Visualisierung der Daten sein oder eine PDF-Datei, die aggregierte Informationen über die beobachteten Tiere beinhaltet.

Sollte ein Workflow relevant für andere Nutzer*innen sein, kann er freigegeben werden. Er kann als öffentlicher Workflow mit allen Nutzer*innen der Plattform geteilt oder nur für Einzelpersonen freigegeben werden. Sobald die Freigabe abgeschlossen ist, wird der Workflow auf dem Dashboard der anderen Nutzer*innen angezeigt. Diese können sich nun eine lokale Kopie des Workflows erstellen und diese nach Bedarf verändern. Die Funktion kann genutzt werden, um gängige und validierte Auswertungsvorgänge leichter zu verbreiten, sei es innerhalb von Arbeitsgruppen oder generell für das gesamte Forschungsfeld.

4 Software für MoveApps entwickeln

Falls Entwickler*innen Software für die Analyse von Tierbewegungsdaten entwickeln und für die Community bereitstellen möchten, können sie MoveApps dafür nutzen. Aktuell unterstützt die Plattform die Programmiersprachen „R“ und „R Shiny“, da diese sich im Bereich der Tierbewegungsforschung durchgesetzt haben. Die Implementierung weiterer Programmiersprachen ins System ist für die Zukunft geplant.

Um Software für die Plattform zu entwickeln und auf ihr zu veröffentlichen, nutzen Entwickler*innen ihre gewohnte lokale Entwicklungsumgebung und können sich dadurch auf die Programmierung fokussieren. MoveApps unterstützt kleine Softwareanwendungen, die jeweils eine bestimmte Funktion erfüllen. Komplexe Analysen lassen sich anschließend durch die modulare Kombination solcher Anwendungen realisieren. Anstatt große Softwareprojekte umzusetzen, können Entwickler*innen sich also darauf konzentrieren ihre Anwendungen möglichst kompakt zu entwickeln und zu optimieren.

Sobald der/die Entwickler*in den Quellcode für das Softwaremodul geschrieben und getestet hat, kann er/sie die App auf MoveApps einreichen. Dazu verweist er/sie mit einem Link auf ein Git-Repositorium, in dem er/sie den Quellcode und weitere Materialien bereitstellt. Diese Materialien dienen dazu, die App verständlich und somit nachnutzbar für andere Nutzer*innen zu machen. Sie setzen sich aus einer Dokumentation sowie Metadaten zusammen. Letztere erstellt der/die Entwickler*in mithilfe eines Settings Editors (siehe Abbildung 2). Dazu gehören administrative Angaben, wie der Name des/der Entwickler*in, weitere beteiligte Personen, Keywords mit denen Nutzer*innen die App auf der Plattform finden können und die Angabe einer Lizenz, unter der die App veröffentlicht

wird. Für den letzten Punkt steht eine Auswahl von unterschiedlichen Lizenzen zur Verfügung, die alle einen gewissen Grad der Offenheit unterstützen. MoveApps erhebt diese administrativen Informationen unter anderem, um Zitationsvorschläge für einzelne Apps bereitzustellen. Diese werden in der Suchoberfläche dargestellt und führen dazu, dass die entwickelten und veröffentlichten Apps auf der Plattform als wissenschaftliches Produkt sichtbar sind, gewürdigt werden und die Plattform somit attraktiver für Entwickler*innen wird.

Das Metadatenschema zur Beschreibung der Apps folgt dabei einer hierarchischen Struktur und orientiert sich wo möglich am DataCite-Metadatenschema[5] Für Teile der Angaben, beispielsweise Rollen von Autoren, Lizenzen, Sprachen etc., werden kontrollierte Vokabulare verwendet, um Fehlern vorzubeugen und die Qualität der Metadaten zu erhöhen. Mithilfe geführter Schaltflächen beschreibt der/die Entwickler*in zusätzlich die einzelnen Funktionen und die zugehörigen Parameter der App. Dafür stehen Optionen zur Verfügung, die jeweils unterschiedliche Arten der Konfiguration ermöglichen. Beispiele sind hier Dropdown-Menüs, Integer-Werte oder Zeichenketten.

Abbildung 2: Der “Settings Editor” von MoveApps.

Sobald der/die Entwickler*in diese Schritte absolviert hat, kann die App zur Begutachtung eingereicht werden. Ein/e Administrator*in von MoveApps prüft die App. Sind alle nötigen Kriterien erfüllt, wird sie anschließend freigeschaltet und ist für alle Nutzer*innen der Plattform verfügbar.

MoveApps unterstützt die Versionierung von Apps. Dies bedeutet, dass Apps, die bereits eingereicht und zur Verfügung gestellt wurden, mit einer neuen Version ersetzt werden kön-

[5]DataCite Metadata Working Group: DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs. (2021). Version 4.4. DataCite e.V. `https://doi.org/10.14454/fxws-0523`.

nen. Dazu muss der/die Entwickler*in einen neuen Tag in ihrem Git-Hub-Repositorium erstellen, den Quellcode entsprechend anpassen und in der MoveApps-Oberfläche auf den neuen Tag verweisen. Neue Versionen müssen erneut von Administrator*innen akzeptiert werden. Sollte eine veraltete Version einer App in einem Workflow verwendet werden, kann diese automatisiert gegen die aktualisierte Version ausgetauscht werden.

5 Veröffentlichung von Workflows

Sollten Nutzer*innen einen oder mehrere Workflows in einer wissenschaftlichen Publikation verwenden wollen, können sie diese persistent und langfristig veröffentlichen. Dadurch werden Metadaten zum Workflow sowie die Metadaten und der Quellcode der verwendeten Apps von MoveApps an das Movebank Data Repository übertragen und dort veröffentlicht. Das Repositorium ist im Zuge des DFG geförderten Projektes „MoveVRE“ (2010 – 2012) entstanden und hat seitdem 241 ausgewählte Datensätze aus Movebank veröffentlicht. Ein Großteil der Datensätze bildet die Grundlage für Zeitschriftenartikel, die ein Peer Review durchlaufen haben. Zusätzlich werden alle Datensätze in enger Zusammenarbeit mit den Datengeber*innen kuratiert.

Um die Veröffentlichungsfunktion auf MoveApps zu nutzen, müssen zwei Voraussetzungen erfüllt sein: Die Daten, auf denen die Analyse durch den Workflow ausgeführt wird, müssen im Movebank Data Repository veröffentlicht sein. Movebank stellt die Datengrundlage für MoveApps dar und es bestehen bereits etablierte Workflows zur Veröffentlichung der Daten von Movebank im Movebank Data Repository. Als zweite Voraussetzung müssen Nutzer*innen nachweisen, dass sie an einer Textveröffentlichung, wie einem wissenschaftlichen Artikel, arbeiten, für den Analysen mithilfe von MoveApps durchgeführt werden. Dazu kann zum Beispiel ein Manuskript des Artikels dienen.

Um die Veröffentlichung im Repositorium einzuleiten, steht eine entsprechende Oberfläche auf MoveApps bereit. Nutzer*innen müssen im ersten Schritt den Workflow mit weiteren Metadaten beschreiben. Wie schon bei den Metadaten für die Apps strebt MoveApps hier eine hohe Konformität zum DataCite-Metadatenschema an. Innerhalb der Oberfläche können Nutzer*innen außerdem festlegen, welche einzelnen Instanzen ihres Workflows sie veröffentlichen möchten. Instanzen eines Workflows bestehen jeweils aus den gleichen Apps, die jedoch unterschiedlich konfiguriert werden.

Sobald die Beschreibung des Workflows hinreichend abgeschlossen ist, wird eine XML-Datei generiert, die sich aus unterschiedlichen Bestandteilen (siehe Abbildung 3) zusammensetzt und an das Team des Movebank Data Repository versendet wird.

Neben den Workflow-Metadaten, die den Workflow selbst beschreiben, werden die Metadaten jeder einzelnen App übermittelt, die im Workflow verwendet wurde. Dies enthält neben den bereits erwähnten administrativen Informationen auch Angaben darüber, auf welchen Zusatzpaketen der Programmiersprache (derzeit R und R Shiny) die App selbst aufbaut. Zusätzlich zu diesen von den Entwickler*innen bereitgestellten Metadaten extrahiert MoveApps automatisiert Informationen über die Konfigurationsparameter der

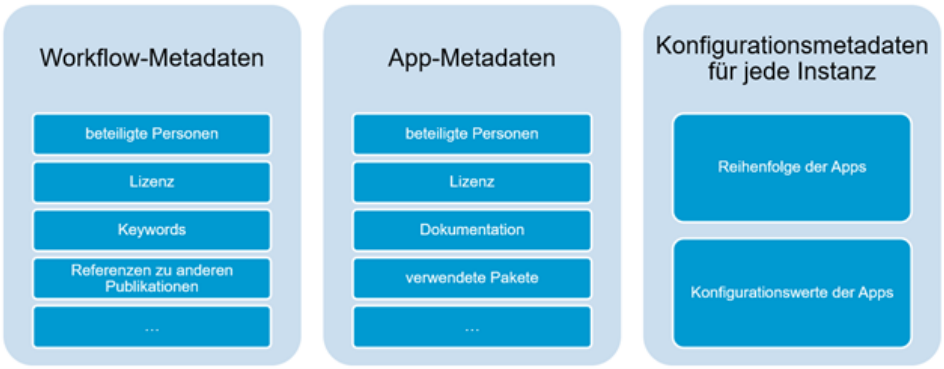


Abbildung 3: Zusammenführung der Metadaten-Bestandteile eines Workflows.

einzelnen Apps. Diese beschreiben wie die einzelnen Ein- und Ausgabewerte der App in diesem spezifischen Fall konfiguriert wurden. Durch ein Speichern und Veröffentlichen dieser Werte ist eine genaue Dokumentation und somit Nachvollziehbarkeit der durchgeführten Analyseschritte möglich. Falls Nutzer*innen verschiedene Versionen eines Workflows in Form von unterschiedlichen Workflow-Instanzen durchführen, können sich diese Konfigurationsparameter unterscheiden. Bei der Veröffentlichung können Nutzer*innen entscheiden, welche Instanzen jeweils veröffentlicht werden sollen. Sobald diese XML-Datei an das Movebank Data Repository übertragen wurde, prüft das Team des Repositoriums die Informationen auf Vollständigkeit und Korrektheit. Anschließend wird der Datensatz, bestehend aus den zusammengesetzten Metadaten, dem Quellcode der einzelnen Apps und deren Dokumentation im Repositorium hochgeladen und erhält einen Digital Object Identifier[6] (DOI), der den Workflow eindeutig global kennzeichnet und somit zur Zitation in Textveröffentlichungen genutzt werden kann. Dabei wird ein DOI für den gesamten Workflow vergeben. Im Fall von mehreren veröffentlichten Workflow-Instanzen werden diese also alle unter dem gleichen DOI referenziert. Dadurch, dass der Quellcode und die Dokumentation öffentlich verfügbar ist, können Wissenschaftler*innen den Workflow auch lokal reproduzieren, unabhängig vom Zustand der Move-Apps-Plattform. Des Weiteren können auf diese Weise mögliche Speicherbegrenzungen auf MoveApps, z.B. für sehr große Datensätze, umgangen werden.

Das gesamte Metadatenschema, sowohl für Apps als auch für Workflows, basiert auf XML und ist hierarchisch strukturiert (siehe Abbildung 4). Dies bedeutet, dass für einzelne Metadatenelemente Unterfelder bestehen, die zusammengesetzt detaillierte Informationen über den jeweiligen Sachverhalt enthalten. Für einzelne Personen können so z.B. mehrere unterschiedliche Rollen zugewiesen werden, die sich an den MARC 21 Relator-Codes[7]orientieren. Sowohl für die Apps als auch für die Workflows können Nutzer*innen Referenzen zu Publikationen oder digitalen Objekten angeben. Dadurch setzen sie die App/den Workflow in einen Kontext und reichern sie/ihn mit Zusatzinformationen an, die für bei der Nachnutzung hilfreich sein können. Um diese Referenz-Angaben in einer standardisierten Form vorliegen zu haben, baut das MoveApps-Metadatenschema an

[6] https://www.doi.org/
[7] https://www.loc.gov/marc/relators/relaterm.htm

```
<workflowPerson>
   <workflowPersonName>
      <workflowPersonNameFirstName>Max</workflowPersonNameFirstName>
      <workflowPersonNameLastName>Mustermann</workflowPersonNameLastName>
   </workflowPersonName>
   <workflowPersonID></workflowPersonID>
   <workflowPersonRoles>
      <workflowPersonRole>contributor</workflowPersonRole>
   </workflowPersonRoles>
</workflowPerson>
```

Abbildung 4: Personenbeschreibung in den Workflow-Metadaten.

dieser Stelle auf die „relationTypes“ des DataCite-Metadatenschemas. Das Metadatenschema ist in diesem Bereich also interoperabel, was den Austausch der Metadaten mit anderen Diensten verstärkt und die Verständlichkeit erhöht.

Dabei unterstützt das Metadatenschema von MoveApps etablierte Identifikatoren. Für die eindeutige Bestimmung von Personen, können Nutzer*innen über die MoveApps-Oberfläche ihre ORCID-Kennung[8] angeben. Dadurch können Namensambiguitäten vermieden werden und Nutzer*innen die veröffentlichten Workflows oder Apps zu ihren wissenschaftlichen Publikationen hinzufügen. Um Angaben über die Zugehörigkeit zu Institutionen zu normieren, setzt MoveApps auf die Research Organization Registry-ID[9] (kurz ROR-ID). Das Metadatenschema und der Publikationsprozess befinden sich aktuell in einer Testphase. Im Laufe des Jahres soll die Implementierung des Features auf Seiten des Repositoriums abgeschlossen und in den Regelbetrieb überführt werden.

6 Zusammenarbeit zwischen MPIAB und KIM Universität Konstanz

Das gesamte Projekt Movebank 2.0 und MoveApps im Speziellen stellt ein Beispiel für die Vorteile einer engen Zusammenarbeit zwischen einem Forschungsinstitut und einem Infrastrukturanbieter dar. Die daraus entstehenden Mehrwerte helfen beiden beteiligten Partnern auf verschiedene Art und Weise. Diese spezielle Kooperation besteht bereits seit Jahren, beispielsweise in der Zusammenarbeit am Movebank Data Repository, bei dem das KIM den technischen Betrieb gewährleistet und das MPIAB die Datenkuration und die Betreuung der Nutzer*innen übernimmt.

Für das KIM birgt die Zusammenarbeit mit dem MPIAB, einem renommierten Forschungsinstitut, die Möglichkeit die Bedarfe einer wissenschaftlichen Fachcommunity aus nächster Nähe zu erfahren. Dadurch kann das KIM lernen, mit welchen Services es Wissenschaftler*innen und deren Spitzenforschung am besten unterstützen kann. Im Falle von Movebank 2.0 kann das KIM das Projekt mit seiner Expertise im Bereich der Metadaten

[8] https://orcid.org/
[9] https://ror.org/

und der langfristigen Verfügbarmachung von Daten und dem Forschungsdatenmanagement im Allgemeinen unterstützen. Für das MPIAB stellt das KIM einen Partner dar, der langfristig für den Erhalt der Workflows und den zugehörigen Dokumentationsmaterialien garantieren kann.

Das Projekt zeichnet sich dadurch aus, dass die beteiligten Partnerinstitutionen sich auf ihre Kernkompetenzen in ihren jeweiligen Fachgebieten konzentrieren und Innovationen vorantreiben. So erweitert das MPIAB mit MoveApps das Umfeld um Movebank und unterstützt damit die Bedarfe der eigenen Fachcommunity. Das KIM arbeitet hier an der langfristigen Erhaltbarkeit und dem Nachweis wissenschaftlicher Software als Teil des Forschungsdatenmanagements. In enger Kooperation konnten hier Herausforderungen wie die Non-Trivialität der Datenkonsistenz und die Komplexität der Reproduzierbarkeit von Software von beiden Blickwinkeln aus analysiert und Lösungsansätze erarbeitet werden.

7 Aktueller Stand und weitere Entwicklungen

MoveApps befindet sich seit Februar 2021 in einer Beta-Version. Dies bedeutet, dass die grundlegenden Funktionen des Dienstes vorhanden sind und getestet werden können. Dazu zählen das Entwickeln und Einreichen von Apps inklusive Versionierung, und das Kombinieren von Apps zu Workflows. Aktuell stehen hier circa 20 geprüfte Apps zur Verfügung. Aus diesen können registrierte Nutzer*innen Workflows erstellen und ihre Daten analysieren. Um das System zu testen wurden Wissenschaftler*innen aus der Tierbewegungsforschungscommunity eingeladen und in enger Kooperation werden deren Rückmeldungen evaluiert und darauf basierend Verbesserungen in der MoveApps-Plattform vorgenommen.

Für die Zukunft ist die Implementierung weiterer Funktionen geplant. Bisher ist die Publikations-Funktion für Workflows auf Seiten von MoveApps implementiert. Nutzer*innen können ihre Workflows mit den benötigten Metadaten beschreiben und diese an das Movebank Data Repository versenden. Für die anschließende Veröffentlichung, inklusive DOI-Vergabe, muss das Repositorium noch vorbereitet werden. Dazu wird es im Laufe des Sommers 2021 auf eine neue Software migriert. Zusätzlich wird das finalisierte Metadatenschema in Form eines XML-Schemas veröffentlicht.

Aktuell unterstützt MoveApps die Verarbeitung von „move“-Objekten, einem Datenformat, das von vielen Movebank-Nutzer*innen verwendet wird. Hier ist auch die Unterstützung weiterer Objekt-Arten denkbar. Abhängig von den Bedarfen der Nutzer*innen werden hier zukünftig weitere Formate unterstützt, um so Analysen unter Einbindung anderer Arten von Daten durchführen zu können.

In der Community rund um Movebank hat sich R als Standard-Programmiersprache etabliert. Nichtsdestotrotz bestehen Bedarfe, die Nutzung anderer Programmiersprachen in MoveApps zu ermöglichen. Hier wurden Python und Java als nächste mögliche Erweiterungen identifiziert. Die Implementierung neuer Programmiersprachen bringt technische Anforderungen an die Plattform mit sich und benötigt Erweiterungen des Metadatenschemas, damit Apps in den neuen Sprachen bestmöglich beschrieben werden können.

Neben der funktionalen Erweiterung der Plattform arbeitet das Team von MoveApps an Schulungsmaterialien, die Nutzer*innen im Umgang mit der Plattform unterstützen. Aufgrund der internationalen Ausrichtung von Movebank und MoveApps werden diese auf Englisch angefertigt. Geplant sind neben asynchronen Lehrmaterialien, wie Screencasts, auch Schulungen.

Danksagungen

Die Autor*innen möchten dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg für die Förderung des Projekts „Movebank 2.0“ danken.
Des Weiteren bedanken wir uns bei der Knobloch Family Foundation für die Förderung der Entwicklung von MoveApps. An der Entwicklung der Plattform und ihrer Funktionalität waren weitere Mitglieder des KIM, MPIAB, der Max Planck Computing und Data Facility, des North Carolina Museum of Natural Sciences und die couchbits GmbH beteiligt.

Literaturverzeichnis

[1] DataCite Metadata Working Group: DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs. (2021). Version 4.4. DataCite e.V. https://doi.org/10.14454/fxws-0523.

[2] Digital Object Identifier.https://www.doi.org/.Abgerufenam22.04.2021.

[3] Icarus. Animal Tracker App.https://www.icarus.mpg.de/29143/animal-tracker-app. Abgerufen am 22.04.2021.

[4] Kays, R., Crofoot, M. C., Jetz, W., & Wikelski, M.: Terrestrial animal tracking as an eye on life and planet. Science, Band 348, Heft 6240 aaa2478–aaa2478 (2015). https://doi.org/10.1126/science.aaa2478.

[5] Kranstauber, B. et al.: The Movebank data model for animal tracking. Environmental Modelling & Software, Band 26, Heft 6 (2011): 834-835. https://doi.org/10.1016/j.envsoft.2010.12.005.

[6] MARC Code List for Relators. https://www.loc.gov/marc/relators/relaterm.html.Abgerufenam22.04.2021

[7] Movebank for animal tracking data.https://www.movebank.org. Abgerufen am 22.04.2021.

[8] ORCID. Connecting research and researchers. https://orcid.org/. Abgerufen am 22.04.2021.

[9] Research Organization Registry. https://ror.org/. Abgerufen am 22.04.2021.

Institutional research data management

Findings from the development and introduction of holistic research data management tools

Luca Leipold[1,*], Janine Straka[2,*] and Kyanoush S.Yahosseini[1,*]

[1]Information and Research Data Management, Robert Koch Institute
[2]Department of Information Sciences, University of Applied Sciences Potsdam
[*]These authors contributed equally to this work

More and more research institutions commit themself to follow data policies when managing their research data. With these policies, they aim to make their research data discoverable, accessible, interoperable and reusable. To support the implementation of these so-called FAIR principles, a suitable technical infrastructure is required. To provide such an infrastructure technical tools that enable easy metadata capture and facilitate research data management are needed. Although tools have already been developed to that end, there is still a lack of integrated platforms that cover the entire lifecycle of research data.

Here we present a prototype of such an integrated platform and describe our learning's from the implementation of this platform at the Robert Koch Institute.

In contrast to previously developed tools, our platform, the DataLinker, captures few data by itself, but mainly connects existing tools and services. More specifically, the DataLinker integrates metadata collected from adjacent systems such as the Research Data Management Organiser (RDMO). This integrated metadata is then linked to all research data associated with a project. This low-threshold approach allows for an easy management of research data at any stage of development. By providing a search interface and contact information, the DataLinker allows research data to be easily located, shared and reused.

We developed the platform to support researchers throughout the research data lifecycle. To do this, we worked closely with researchers during development, requesting and implementing their feedback. However, as we rolled out the platform, we found that requirements to such a platform go beyond the perspective of the researchers' data management needs. More specifically requirements from the perspective of an individual researcher are fundamentally different to the requirements when handling research data in an institutional context. For a platform to actually be used, it has to demonstrate clear added value besides its benefits to research data management. These benefits have to encompass not only the researchers but also the institutional service providers perspective. Providing benefits to the processes of these service providers, such as the IT department, legal department and the data protection office, are significantly relevant to the acceptance of such a platform.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13720 (CC BY 4.0)

Our work shows how project and research data processes are intertwined and cannot be considered independently. Therefore, platforms for simple research data management need to integrate and map both processes. We conclude that researchers need to be supported in both areas to enable them to follow institutional data policies.

1 Introduction

Proper research data management is becoming the focus of research institutions in the field of quantitative research around the world. As a result, more and more research institutions commit themself to follow data policies when managing their research data. With these policies, they aim to make their research data discoverable, accessible, interoperable and reusable. To support the researchers in implementing these so-called FAIR principles [1], a suitable technical infrastructure is required. This infrastructure needs to include technical tools which match the researchers everyday working life. These tools need to allow for easy research data management while providing a direct benefit, for example they should enable the researchers to easily capture, update and keep track of research and metadata. Although various tools have already been developed for this purpose, there is still a lack of integrated platforms that cover the entire lifecycle of research data.

Here we present a prototype of such an integrated platform and explain our experiences from the implementation of this platform at the Robert Koch Institute (RKI). The RKI is the national public health institute Germany [2], consisting of more than 1400 employees. The researchers work in very heterogeneous projects and tasks while handling very diverse research data. Historically (the RKI has been founded in 1891 [2]) additionally hamper the structured and open management of research data. As a result there is no platform or infrastructure to get an overview of all current and past projects and their associated research data so far. This is problematic as making research data findable is the first step towards a FAIR research data management.

Another challenge is the growth of data volume. Exponential growth of data volume brings conventional approaches to a limit, as possibilities for data storage, backup and long-term archiving must keep track of the increasing data column. For example, at the RKI approximately 80-100 terabytes of data are currently generated per year. Without proper research data management keeping track of this enormous quantities of data is next to impossible.

We aim to overcome these problems by introducing a new easy to use software platform for research data management which is developed according to the researchers needs.

In this paper we first introduce our platform while describing the two components of the platform and their interactions in detail. Than we take a more theoretical perspective and describe the challenges we encountered and the lessons learned when introducing the platform to the institute. We suggest that these challenges can not be overcome by mainly technical means but that technical platforms and administrative processes and needs have to be considered side by side.

2 Technical development

We developed a novel integrated platform to guide manage researchers data through the whole research data lifecylce. Our platform consists of two components 1. the open source tool Research Data Management Organiser (RDMO) [4] and 2. a newly developed tool the DataLinker [9]. Both components fulfill different complementary tasks in our platform.

RDMO aims to simplify the process of managing and updating metadata. Hence in our platform RDMO is used to collect project metadata. Through a questionnaire based approach researchers use RDMO to gather various project related metadata, such as the title of their project and their collaborators. Using RDMO as a single-entry point and main hub for project related metadata avoids multiple entries of the same data when fulfilling the administrative demands of different departments in an institution.

The DataLinker automatically reads in data provided in RDMO. By showing a list of all projects which have been created in RDMO it provides a comprehensive overview of all research activity in an institution. Additionally imported projects from RDMO can be further enriched with additional data. Specifically, the DataLinker allows to connect projects with datasets and their physical location. This allows research data to be findable within an institution by searching for its metadata. Finally, the platform provides an easy import and export mechanism to public repositories. The DataLinker is not only limited to RDMO as a data source, but can read, use and digest information from other tools as well using public interfaces (Fig. 1). Hence our platform provides a standardised workflow from data management plan to data publication.

In the following section we introduce the two main components in more detail, while focusing on their technical implementation.

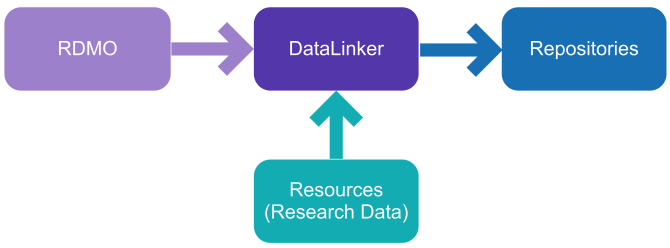


Figure 1: Components of our platform.

2.1 Research Data Management Organiser (RDMO)

The Research Data Management Organiser (RDMO) [4, 5] is a tool to create, update and work together on data management plans (DMPs) and to manage project related tasks. RDMO is an open source web application [6]. It is written in Python utilizing the Django package [7]. It supports a variety of different database systems, such as SQLite, MySQL, and PostgreSQL, as its main data storage. The public facing part of the tool relies on the AngularJS framework. Here we describe the general workflow of the tool followed by the changes we implemented to fit it into our platform.

The general workflow in RDMO focuses on the projects of a user and uses an interview style data entry format (a questionnaire) to collect data about the users' projects (see Fig. 2). A registered and logged in user can create a new project. For each newly created project a questionnaire has to be selected to indicate the type of the project. After selecting a questionnaire the user should answer questions in an interactive questionnaire. Some questions are dynamically generated based on the previous answers of the user. For example indicating that a project is funded by a third-party prompts questions which are related to this kind of projects. Questions which do not apply to a specific project are skipped. Also RDMO handles all answers as work in progress. That is, users can always skip questions or update their answers in a later stage of their progress.

RDMO allows to display questions and answers for each project in a structured way. These views can be used to automatically fill out pre-configured templates allowing for an easy way to inform third-party funders or handle administrative processes. To export answers into different file formats RDMO uses the Pandoc converter, a Python package [8], to export answers into diferent file formats.

The tool also provides an automatically generated list of tasks depending on given answers. These tasks indicate work items for the user. If such a task indicates that someone has to be informed about the state of a project, RDMO allows to generate and send a predefined e-mail with an attached view containing the requested information.

RDMO is a fundamentally collaborative tool. That is, a user can invite other researchers or collaborators to join a project and work on a questionnaire together. RDMO provides role and rights management for this purpose.

2.1 RDMO for institutional research data management

RDMO is extensible and highly configurable to the needs of an institution. To fit the needs of the Robert Koch Institute we adapted RDMO. We changed the look and feel of RDMO to match the look of other platforms of our institution. More importantly we adjusted questionnaires, implemented new tasks and views to mirror the processes and structures at the Robert Koch Institute. All modifications are published in GitHub [10].

We developed new questionnaires in RDMO to fit the needs of our platform. A generic questionnaire with up to 60 questions should be answered for every new project. The

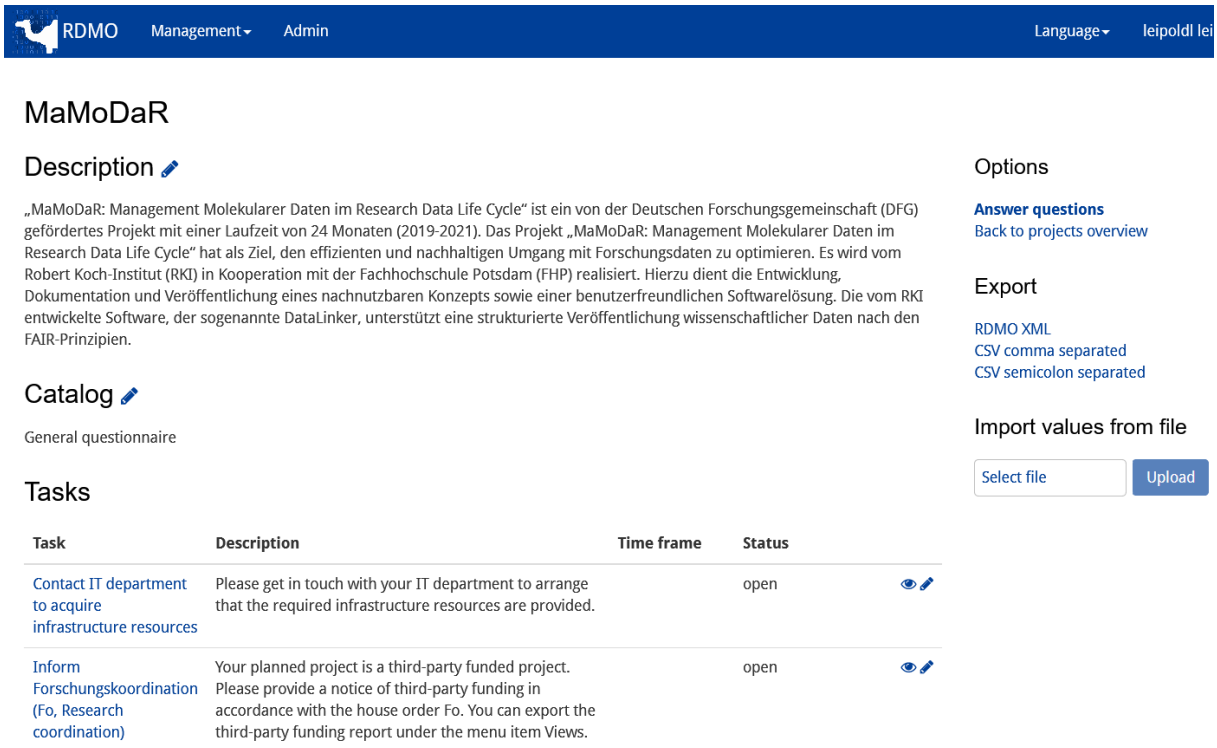


Figure 2: Example overview of a project in RDMO.

given answers in this questionnaire are the main source of metadata in our platform. We also developed three auxiliary questionnaires which are aimed at different project specific areas. These three questionnaires cover the handling of three areas: 1. data management plans, 2. administrative processes regarding the research project and 3. administrative processes regarding research data.

The first area comprises the core competence of the RDMO tool. Users are supported by the tool in creating data management plans allowing them to better manage their research data and to cover requirements for third-party funding applications. In the second area, users are supported in completing administrative applications that derive from their research project. These include cooperation agreements, third-party funding and data publication notifications. These applications can then be directly submitted to cross-sectional departments (such as the legal department, research coordination and the library). The third area includes processes that need to be initiated based on the nature of the research data. For example, in the case of research data which includes personal data, the data protection officer must be involved in coordination processes. Also when publishing sensitive data a dual-use analysis must be carried out. These three areas showcase one benefit of the platform: As many of the processes mentioned above require identical data. RDMO allows data provided in other contexts to be re-used to avoid reentering the same data multiple times.

Another major extension to RDMO was implemented for the so called tasks. Tasks in RDMO indicate work items which the user has to take care of. Tasks are generated

based on the user's answers in the questionnaires. For example, if a user answers the question "Does the dataset contain sensible data?", a task will be created which asks to contact the data security officer. In our platform we reused some of the tasks which were already implemented in RDMO. Additionally, new tasks were created to guide through internal administrative processes. For example based on the question "Is it a third-party funded project?" a task "Inform Forschungskoordination (Fo, Research coordination)" is generated. It includes a detailed description of the task such as "Your planned project is a third-party funded project. Please provide a notice of third-party funding in accordance with the house order. You can export the third-party funding notification under the menu item Views. On a regular basis, Fo should receive the notification at least 8 working days before the planned submission of the application to the funding agency!"

Our goal is to navigate the user of our platform through all mandatory administrative processes. This allows users to more easily comply with those procedures, as the system clearly shows them which applications have to be submitted and which formalities have to be observed. The answers given in the question allows to show only the relevant tasks with their corresponding time line. This makes it easier to meet deadlines, indicate which applications have to be filled and to whom an application should be made available.

2.2 DataLinker

The DataLinker is a web application that supports the capturing of project resources, the searching for project metadata, and the publishing of scientific research data in open repositories in accordance with the FAIR principles. The DataLinker has interfaces to adjacent systems and obtains project metadata primarily from RDMO via a REST API in JSON format. However it is not only limited to RDMO as a source of information. The DataLinker follows a layered architecture, it consists of three layers: a database-based persistence layer for data storage, a business layer implemented in the Java Spring Boot Framework [17] handling the application's logic and a web-based presentation layer implemented in Angular [18]. The business logic and fronted communicate by using a REST API [19]. In order to give the user the feeling of being in a single environment, the design of the DataLinker was based on the design of RDMO. The open source code for DataLinker is available at GitHub [9], a documentation [11] and a manual [13] have been published

The DataLinker provides three core functionalities to expand on the capabilities of RDMO. These three functionalities revolve around 1. search, 2. tracking of research data (so called resources) and 3. publication of research data. To provide these functionalities each project created in RDMO is mirrored into the DataLinker automatically, using the REST API of RDMO. In principle other providers of metadata besides RDMO can be integrated as well.

The DataLinker reconditions the collected metadata and provides a listing of all created projects with their relevant metadata. This metadata includes for example: contact person, title, project description and usage rights. The DataLinker allows resources, the

location and type of data sets, to be attached and tracked to a project. As a result users can easily get an overview of where data is stored, which licences are used and which type of data is been stored.

The three core functionalities of the DataLinker reflect in the tab-based user interface. The tab “Search” allows users to search for other projects and their associated meta- and research data via a keyword search (see Fig. 3). Search results can also be filtered via a faceted search by relevant metadata such as organizational unit, funder, or cooperation partner. This makes allows for different search strategies. For example, researchers interested in influenza can use the search to find already existing research data on their topic. They can also look up who has already worked on influenza and which data is available at which location. Data and also the expertise of others can be reused. Another example refers to department heads who can get an easy overview of all projects which have been carried out in their department, both historically and currently.

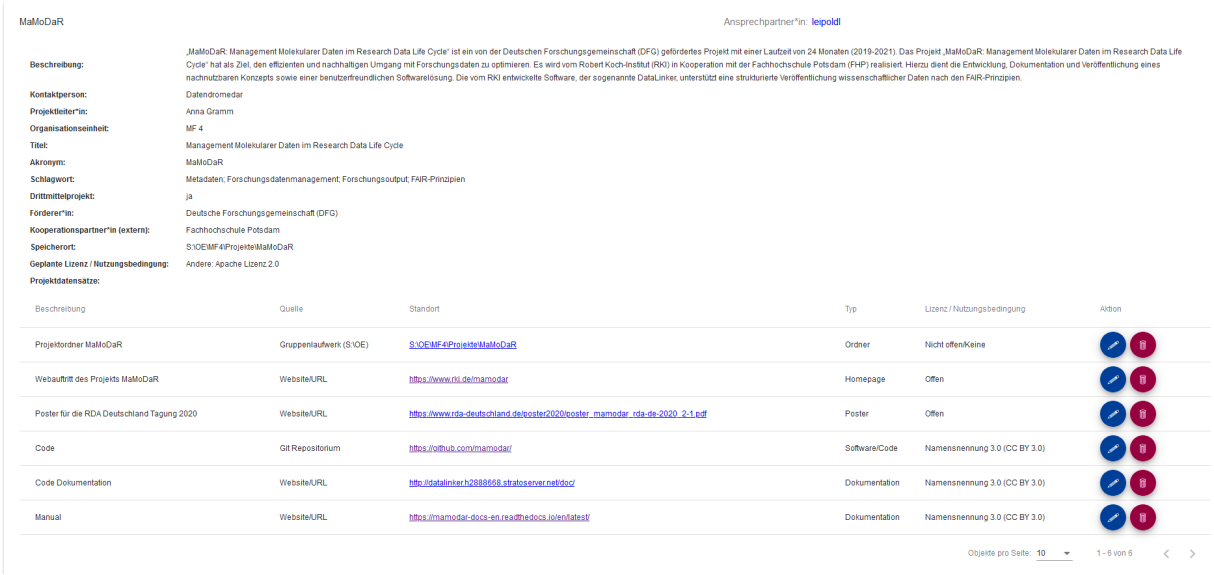


Figure 3: Example projects as shown in the DataLinker with its linked resources.

The tab “My Projects” allows users to attach resources (research data or data sets) to a specific project. This linkage again allows the connection between location of resources and the corresponding projects. The storage of the resources is not altered by the DataLinker, just a pointer to the location of the resources is saved and tracked. This allows the actual research data to remain on a local storage drive, being moved around or to be published to an external source.

The last tab “Publications“ supports the user in publishing data sets to external public repositories such as edoc [15] or Zenodo. Here edoc is the Robert Koch Institute’s own publication server which is based on DSpace [14] which is used in many different institutions. Zenodo is a generic repository which is financed by the European commission. Due to the modular structure of the DataLinker, the data export can also be easily implemented for other repositories. Before publication the user can review their already

collected metadata. After approving the publication metadata and research data is automatically transferred to the selected repository and published. If a DOI [12] is generated during the publication process it is automatically written back into the DataLinker.

3 Platform development

The development of our platform was guided by the needs of the researchers at our institution. To understand their everyday work and to assess their previous knowledge of research data management several approaches were used.

First we launched an institute-wide survey to gain insight into the extent to which researchers are already familiar with research data management and which tools they use for this purpose. We based our survey on the survey conducted at the Potsdam University of Applied Sciences and adapted the questions to our concerns [2]. The survey conducted in February 2020 (51 participants) showed that the majority (86%) of all survey participants (in departments who work with molecular data) deal with research data in their work context. However, experience in research data management is way less common (58% of all participants indicated that they have little to no experience). The survey also indicated that most researchers use digital tools to manage their research data which are not FAIR compliant. More specifically most participants indicated that they use familiar but ill-fitting tools such as the word-processor Microsoft Word and the spreadsheet software Microsoft Excel to handle their research data.

Second, we conducted expert interviews to better understand the users' everyday work and to adapt the tools to their usual work structures. We bundled our results from these interviews with the feedback of the surveys and created a list of requirements. The development of our platform followed this list. Afterwards, usability tests were carried out and adjustments were made to the system in close feedback loops with the researchers. The adaptations made to the platform were discussed in a workshop with internal and external participants. The resulting feedback then again was transferred to the development of the systems.

4 Institutional research data management

Platform requirements for research data management from the perspective of an individual researcher are fundamentally different to the requirements when handling research data in an institutional context. When introducing our platform to a wider audience we found that fulfilling all research data management requirements is not enough for a general acceptance of a system. For the tool to actually be used, it has to demonstrate clear added value besides its benefits to research data management. This is especially important within the framework of institutional research data management. Here it became apparent that the role of cross-sectional departments has to be given a large focus.

Such departments as the IT, the research coordination, the data protection officer and the legal department play a crucial role when introducing new platforms: Their support or rejection of a platform propagates towards the acceptance by the researchers. To generate such acceptance an advertised platform needs to provide a clear added value not only to the researchers but also to these service providers. Hence introducing any new research data management platform is a process which should be jointly initiated.

It is important to realise that research data management is only one module in an institute's landscape. Each department has its own established processes and vocabularies and a justification to use exactly these. Hence, another central finding is that research data management must be integrated into the established administrative processes to achieve wide acceptance. Especially project and research data management processes are closely intertwined, so they need to be considered together. To use RDMO as the single point of entry into research data management proved to be impracticable for our approach. In contrast we suggest to focus on linking the existing processes and their used (technical) tools. In this way, all concerns can be contextualised and applied in the respective expert systems. The necessary data should then be automatically transferred between these expert systems.

5 Conclusions

We developed a research data management platform which consists of two components 1. RDMO as a tool to capture and handle project related metadata and 2. DataLinker as a tool to make this metadata FAIR. We closely cooperated with interested researchers to develop and adapt our platform to their needs. Our experience during the introduction of the platform shows, that focusing on the users and the technical implementation of such a platform is not enough. Instead the introduction of a research data management platform also requires focus on administrative processes. That is why the structures of an institution needs to be addressed in detail and all people involved need to be identified and in-cooperated. To manage this enormous effort we suggest to not replace existing expert systems but to augment them with proper research data management platforms.

In summary, we identified two pillars that have to be given equal attention and are a prerequisite for the success of a project.

The first pillar is the needs of the researchers. Each platform has to be developed according to the users every day work. The goal aim here is to support the user of the platform while avoiding as much additional work as possible. Furthermore, a clear added value must be noticeable to the user, it has to become clear why it is worth to manage research data properly and what benefits the user can have by making research data available to others.

The second pillar comprises the organisational structure. New processes which are inevitably established by the introduction of new technical platforms should be based on existing processes. In particular it is important to include the cross departments in all

developments. These departments, such as the research coordination, the IT department, the data protection officer, the legal department and ethics committee, support scientists in their projects and are hence an integral part of the success of such a platform. Moreover close integration of research data management with project management is essential.

Acknowledgements

Our platform was developed as part of the project "Management Molekularer Daten im Research Data Lifecycle (MaMoDaR)". The project has been funded by the Deutsche Forschungsgemeinschaft (DFG), Germany - Project number 416783714. The Robert Koch-Institute and the University of Applied Sciences Potsdam participated in the realization of the project.

~~The authors thank Linus Grabenhenrich for his valuable advise during the paper creation.~~

Bibliography

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. "The FAIR Guiding Principles for scientific data management and stewardship" Sci Data 3, 160018 (2016): 1-9. `https://doi.org/10.1038/sdata.2016.18.`

[2] Website: `https://www.rki.de/EN/Content/Institute/Mission_Statement/Mission_Statement_node.html` (last visited: 09.03.2021)

[3] Website: `www.rki.de/mamodar` (last visited: 09.03.2021)

[4] Website: `https://rdmorganiser.github.io` (last visited: 09.03.2021)

[5] Website: `https://rdmo.readthedocs.io/en/latest` (last visited: 09.03.2021)

[6] Website: `https://github.com/rdmorganiser` (last visited: 09.03.2021)

[7] Website: `https://www.djangoproject.com` (last visited: 09.03.2021)

[8] Website: `https://pandoc.org` (last visited: 09.03.2021)

[9] Website: `https://github.com/mamodar/datalinker` (last visited: 09.03.2021)

[10] Website: `https://github.com/mamodar/rdmo-rki-catalog` (last visited: 09.03.2021)

[11] Website: `https://github.com/mamodar/datalinker` (last visited: 11.03.2021)

[12] Chandrakar, R. "Digital object identifier system: an overview." The Electronic Library (2006).

[13] Website: `https://mamodar-docs-en.readthedocs.io/en/latest/` (last visited: 06.04.2021)

[14] Website: https://duraspace.org/dspace/ (last visited: 31.03.2021)

[15] Website: https://edoc.rki.de/ (last visited: 06.04.2021)

[16] Arndt, O., Glatz, L., Hummel, B. et al. "Umfrage zum Forschungsdatenmanagement an der FH Potsdam : Projektbericht" Zenodo (2018) https://doi.org/10.5281/zenodo.1161792

[17] Website: https://spring.io/projects/spring-boot (last visited: 06.04.2021)

[18] Website: https://angular.io (last visited: 06.04.2021)

[19] Website: https://restfulapi.net (last visited: 06.04.2021)

Mit AIMS zu einem Metadatenmanagement 4.0: FAIRe Forschungsdaten benötigen interoperable Metadaten

Matthias Grönewald [1], Patrick Mund [2], Matthias Bodenbenner [2], Marc Fuhrmans [1], Benedikt Heinrichs [3], Matthias S. Müller [3], Peter F. Pelz [4], Marius Politze [3], Nils Preuß [4], Robert H. Schmitt [2,5], und Thomas Stäcker [1]

[1]Universitäts- und Landesbibliothek, TU Darmstadt
[2]WZL | RWTH Aachen University
[3]IT Center, RWTH Aachen University
[4]Institut für Fluidsystemtechnik, TU Darmstadt
[5]Fraunhofer-Institut für Produktionstechnologie IPT

Gute wissenschaftliche Praxis erfordert eine präzise und verständliche Dokumentation der Ergebnisse. Dies ist umso wichtiger, wenn Forschende ihre eigenen Forschungsdaten teilen und publizieren oder archivierte Daten Dritter nachnutzen möchten. Forschungsdatenmanagement (engl. research data management) auf Basis weitreichend standardisierter Metadaten ist daher von essenzieller Bedeutung. Standardisierte Metadaten sollen in strukturierter und maschinenlesbarer Form Informationen zu Entstehung, Inhalt und Kontext der beschriebenen Forschungsdaten liefern. Im Zuge der ingenieurwissenschaftlichen Forschung zu Industrie 4.0 werden enorme Mengen heterogener Daten generiert. Ein entsprechendes "Metadatenmanagement 4.0" soll es Forschenden ermöglichen, aus einer Vielzahl bereits vorhandener digitaler Datensätze den richtigen Datensatz für die eigene Forschung zu selektieren. Entsprechend der FAIR-Prinzipien müssen Metadaten für Mensch und Maschine interpretierbar sein.

Im Forschungsprojekt „Applying Interoperable Metadata Standards (AIMS)", gefördert von der Deutschen Forschungsgemeinschaft (DFG), werden die Herausforderungen an ein Metadatenmanagement 4.0 für die ingenieurwissenschaftliche Forschung adressiert. Der Projektfokus liegt auf dem Maschinenbau und verwandten Disziplinen. Das interdisziplinäre Team, verteilt über mehrere Institutionen aus Infrastruktur und Wissenschaft, schafft hier eine Plattform, die es Forschenden ermöglicht, Metadatenschemata zu erstellen und zu teilen. Durch ein Modellierungskonzept, das auf Vererbung und Modularität setzt, kann eine hohe Spezifität bei maximaler Anwendbarkeit und Nachnutzbarkeit der Metadatenschemata erreicht werden. So wird die Akzeptanz der Forschenden erhöht strukturierte Metadaten in ihre Forschungsabläufe zu integrieren, um mit zunehmender Verbreitung den Weg zu gemeinsamen Metadatenstandards zu bereiten. Das Projekt ist von Anfang an als Kooperation zwischen Infrastruktur und Forschung ausgelegt. Dadurch soll das in AIMS realisierte Metadatenkonzept direkt in die Prozesse der beteiligten Forschen-

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13721 (CC BY-SA 4.0)

den integriert werden. Durch die Berücksichtigung von Interoperabilität der entstehenden Metadatenschemata und -schnittstellen wird explizit der generische Transfer der Lösungen auf andere Forschungsfelder verfolgt. In diesem Beitrag werden die Anforderungen aus der Forschung aufzeigt und der Ansatz von AIMS für ein Metadatenmanagement 4.0 vorgestellt. Zudem wird ein Einblick in infrastrukturelle Herausforderungen und mögliche technische Lösungen der aus dem Projekt AIMS hervorgehenden Metadaten-Plattform gegeben.

1 Einleitung

In die Umsetzung von experimentellen Studien (physisch oder virtuell) und die Generierung von großen Mengen an Messdaten wird viel Geld, Zeit und Expert:innenwissen investiert. Neben Publikationen und Berichten sind diese erfassten Forschungsdaten ein wichtiges Produkt wissenschaftlicher Arbeit [1]. Die langfristig nachvollziehbare Dokumentation dieser Daten sowie des Prozesses ihrer Entstehung ist ein Qualitätsmerkmal guter wissenschaftlicher Praxis und Organisationskultur. Sie eröffnet wichtige Anschlussmöglichkeiten für die weitere Forschung und Produktentwicklung. Sie ist außerdem Voraussetzung für einen nachhaltigen Transfer der Forschungsergebnisse in die Industrie bzw. die Weiterverwendung von Daten [2]. Besondere Herausforderungen für die standardisierte Dokumentation von Forschungsdaten in vielen wissenschaftlichen Disziplinen, darunter auch dem Maschinenbau, sind Heterogenität und Komplexität der Daten, sowie ihrer Entstehungsprozesse.

Bei der Durchführung ingenieurwissenschaftlicher Experimente werden i.d.R. sehr individuell konfigurierte Versuchsstände eingesetzt, die zudem meist kontinuierlich weiterentwickelt werden. Infolgedessen stehen keine standardisierten Werkzeuge wie elektronische Laborbücher (ELBs) oder Laborinformationsmanagementsysteme (LIMS), wie sie aus der Chemie und der Biologie bekannt sind, für den Bereich des Maschinenbaus zur Verfügung. Gleiches gilt für kommerzielle Softwarelösungen. Diese setzen meist starre domänenspezifische Standards voraus und versagen bei der Integration von heterogenen und sich verändernden Versuchsaufbauten, Analysemethoden oder Metadaten [3].

Dies führt zu einer hohen Heterogenität der generierten Daten, was sich in Verbindung mit den steigenden Raten der Datengenerierung zunehmend als problematisch hinsichtlich ihrer Nutzbarkeit darstellt [4].
Mitarbeiter:innen sind gezwungen einen beträchtlichen Teil ihrer Zeit aufzuwenden, um Daten aus nicht maschinenlesbaren Quellen (bspw. Grafiken oder unkommentierte Wertetabellen) zu identifizieren, zusammenzuführen und mit relevanten Datensätzen über jeweils spezielle Programmierschnittstellen zu interagieren [5].

Die Umsetzung der FAIR-Prinzipien [6] (engl. findable, accessible. interoperable, re-usable) ist unerlässlich, um Forschungsdaten nachhaltig zu generieren und vorzuhalten, bzw. zu interpretieren und zu nutzen [7, 8]. Dies gilt sowohl für die aktive Nutzung während des Forschungsprozesses als auch für die Nachnutzung nach Abschluss eines Forschungspro-

jektes. Insbesondere beim Austausch und der gemeinsamen Nutzung von Forschungsdaten über Projektgrenzen hinweg werden interoperable Standards für Metadaten und Software-Schnittstellen benötigt, um ihre Nutzbarkeit zu gewährleisten [9, 10]. Es besteht der dringende Bedarf nach der Schaffung übergreifender Plattformen, Etablierung verbindlicher Standards und einheitlicher Dienste, die es erlauben, Forschungsdaten zu sichern, optimal nach zu nutzen und zu vernetzen.

2 Ziel und Ansatz

Das Projekt „Applying Interoperable Metadata Standards" (AIMS) hat die Zielstellung (1) eine interoperable Abbildung komplexer Experimente, Methoden und Datensätze zu erreichen und (2) gleichzeitig die nachträgliche oder manuelle Dokumentation entsprechender Metadaten zu vermeiden.

Zu diesem Zweck werden in AIMS zwei Kernergebnisse erarbeitet: Zum einen wird eine Umgebung geschaffen, die es Forschenden ermöglicht spezifische Metadatenschemata zur Beschreibung ihrer individuellen Forschungsdaten zu erstellen, zu teilen und nach zu nutzen. Sie werden so in die Lage versetzt für ihr direktes Umfeld oder ihre Fachdomäne nach und nach Quasi-Standards zu entwickeln, zu pflegen und zu etablieren. Zum anderen werden Werkzeuge und Abläufe erarbeitet, die sich in den wissenschaftlichen Forschungsalltag integrieren und so, durch eine verbesserte Effizienz der Auswertung, Bearbeitung und Dokumentation von Daten, einen direkten Mehrwert für die Forschung bieten.

Realisiert wird dies auf Basis eines modularen Modellierungsansatzes (siehe Abschnitt 3.1), nutzbar gemacht durch die zur Verfügung gestellte Infrastruktur (siehe Abschnitt 3.2) und umgesetzt durch das Einbinden von mehreren Anwendungsfällen aus der ingenieurwissenschaftlichen Forschung direkt in den Projektablauf (siehe Abschnitt 3.3). Durch diesen integrativen Dialog zwischen Forschenden und Infrastruktur werden Möglichkeiten für ein verbessertes Metadatenmanagement geschaffen, validiert und demonstriert: Spezifische Metadatenbeschreibungen erleichtern nicht nur die Dokumentation, Publikation und Archivierung von Forschungsdaten, sondern auch den Forschungsprozess als solchen. Insbesondere Forschungsprozesse mit großen Datenmengen oder einem hohen Grad an Automatisierung können hier profitieren.

Die Implementierung der Konzepte und Werkzeuge aus AIMS innerhalb der bestehenden Forschungsprozesse ist ein wesentlicher Aspekt auf dem Weg zu Metadatenstandards. In einer Art Konzeptbeweis für ausgewählte Teilbereiche der Ingenieurwissenschaften können hier die Vorteile und Potentiale aufgezeigt werden, die sich für die Forschenden durch ein erfolgreiches Metadatenmanagement bereits während der Forschung ergeben. Dies führt zu einer deutlich gesteigerten Akzeptanz in der Forschungsgemeinschaft als ganzer. Integraler Bestandteil der erfolgreichen Umsetzung des Ansatzes ist deshalb die Verbreitung der positiven Anwendungsfälle aus dem Projekt in der Forschungsgemeinschaft. Durch eine verstärkte Interaktion über bestehende Projekte und Initiativen kann so ein Weg zu gemeinsamen Standards bereitet werden.

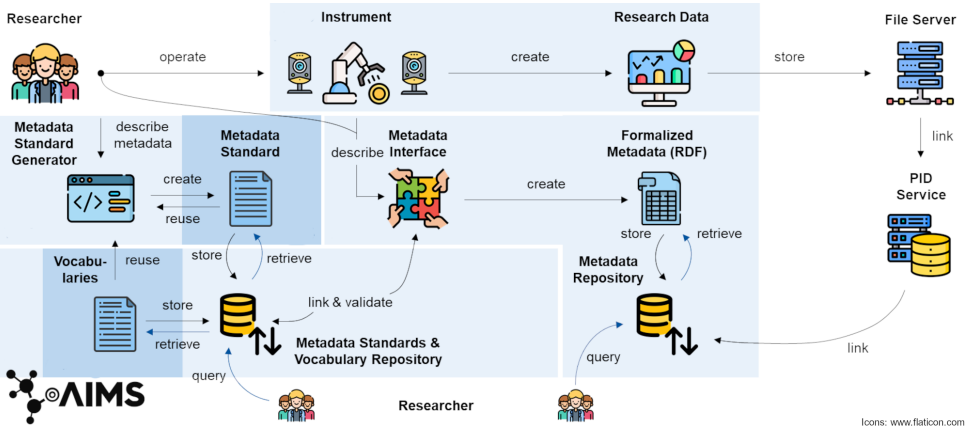


Abbildung 1: Projekt AIMS stellt vernetzte Infrastruktur zum Metadatenmanagement für die Forschenden bereit und bildet damit einen Baustein hin zu einer Linked Open Data Cloud.

3 Applying Interoperable Metadata Standards - AIMS

AIMS widmet sich den beschriebenen Herausforderungen und Lösungsansätzen seit Ende des Jahres 2020 im Rahmen einer Förderung durch die Deutsche Forschungsgemeinschaft (DFG). Von entscheidender Bedeutung ist, dass AIMS keine Initiative rein aus dem Infrastrukturbetrieb ist, sondern maßgeblich von Forschenden initiiert und getragen wird. Konkret arbeiten in einem interdisziplinären Team aus Ingenieur:innen, Informationswissenschaftler:innen und Informatiker:innen der Universitäts- und Landesbibliothek (ULB) und dem Institut für Fluidsystemtechnik (FST) der TU Darmstadt sowie dem IT Center (ITC) und dem Werkzeugmaschinenlabor WZL der RWTH Aachen University Forschende und Infrastrukturbetreibende aus vier unterschiedlichen Institutionen verteilt über zwei Universitäten gemeinsam an einer Plattform zur Erstellung und zum Teilen interoperabler Metadatenschemata und deren Integration in die wissenschaftlichen Abläufe.

AIMS beginnt mit der Integration in einzelnen Anwendungsfällen der beteiligten Institutionen, um von da aus kontinuierlich den Nutzerkreis durch Training und Verbreitung zu erweitern. Ausgehend von Fachgebieten, über Institutionen und Verbundforschung, dient das Projekt als Baustein in einer größeren Initiative zu einem verbesserten Forschungsdatenmanagement in der ingenieurwissenschaftlichen Forschungsgemeinschaft. Hervorzuheben ist dabei zum einen die Integration in die Nationale Forschungsdateninfrastruktur (NFDI) durch intensive Wechselwirkung speziell mit dem Konsortium Nationale Forschungsdateninfrastruktur für die Ingenieurwissenschaften (NFDI4Ing) und die Fokussierung auf offene Schnittstellen, beispielsweise das OAI Protocol for Metadata Harvesting (OAI-PMH, engl.), und der Einsatz standardisierter Abfragesprachen (siehe Abschnitt 3.2).

In AIMS sollen für die zu entwickelte Plattform auch Dienste aus anderen Projekten eingebunden werden, die die Verbreitung und Kuratierung bereits vorhandener Schemata adressieren, beispielsweise CoSciNe [11], oder die über offene Schnittstellen Ontologien

Abbildung 2: Das Konzept von Bausteinen, die von Forschenden als Module zu unterschiedlichen Einheiten zusammengefügt werden können, ist ein zentrales Element hinter AIMS.

verfügbar machen, wie der in NFDI4Ing entwickelte Terminology Service [12]. In seiner Gesamtheit wird so ein weiterer Beitrag zu den Bemühungen einer offenen und vernetzen Forschungslandschaft geleistet.

3.1 Modell

Im Umgang mit der Heterogenität und Komplexität der Untersuchungsgegenstände, Methoden und Werkzeuge in den Ingenieurwissenschaften setzt AIMS auf die flexible Erstellung und Nachnutzung fachspezifischer und anwendungsbezogener Metadatenschemata. Neben der in Abschnitt 4 beschriebenen Plattform (insbesondere dem Generator und dem Repositorium für Metadatenschemata) erfordert dies auch einen Modellierungsansatz für Metadaten, der Spezifität und Flexibilität auf der einen Seite mit einer hohen Nachnutzbarkeit, einer breiten Anwendbarkeit sowie einer maximalen Interoperabilität auf der anderen Seite vereinbart. Dazu wird in AIMS auf ein Baukastenprinzip gesetzt, das auf den folgenden Designkomponenten beruht:

Applikationsprofile

Metadatenschemata werden aus Termen, die aus kontrollierten Terminologien entnommen werden, in Form sogenannter Applikationsprofile [13] zusammengesetzt.

Dies erlaubt die flexible Definition maßgeschneiderter Metadatenschemata, die aufgrund ihres Rückgriffs auf kontrollierte Terme mit anderen Schemata, die dieselben Terme verwenden, interoperabel sind.

Hierarchische Modellierung

Die Applikationsprofile werden darüber hinaus hierarchisch modelliert, indem spezifische Profile nicht losgelöst, sondern als Subklassen generischer Klassendefinitionen definiert werden, die, wie in der objektorientierten Programmierung, die Attribute ihrer Eltern erben und um neu hinzukommende auf die Kinder passende Attribute ergänzen. Durch dieses (mehrstufige) Verfahren sind verwandte Profile stets auf Ebene des letzten gemeinsamen Vorfahrens interoperabel. Ebenso kann die Definition neuer Profile stets an dem passenden Knoten in der Hierarchie erfolgen, die sich auf diesem Weg in die Baumstruktur einfügen.

Modularität

Um die Anwendbarkeit und Nachnutzbarkeit der Applikationsprofile zu erhöhen, werden trennbare Bereiche von Metadaten durch separate Schemata beschrieben. Solche Bereiche sind z. B. Untersuchungsgegenstand, Methode und Werkzeug. Dadurch wird eine unnötige Einschränkung der Anwendbarkeit, die bei Abdeckung mehrerer dieser Bereiche in einem einzelnen Schema aufträte, vermieden. Zur vollständigen Beschreibung von Forschungsdaten werden die einzelnen Profile kombiniert, etwa durch ein Profil für einen "Processing Step", dass Instanzen sämtlicher relevanter Profile bündelt und mit In- und Output des beschriebenen Processing Steps in Verbindung setzt. Darüber hinaus sind weitere direkte Verbindungen zwischen Profilen (sowohl innerhalb eines Bereichs als auch über Bereichsgrenzen hinweg) Teil des Konzepts. So kann etwa ein Werkzeug ein Attribut "enables" besitzen, welches eine durch das Werkzeug implementierte Methode angibt.

Die Realisierung der genannten Designprinzipien erfolgt auf Basis des Resource Description Framework (RDF, siehe Abschnitt 3.2), und fügt sich zu einer Beschreibung von Forschungsdaten in Form eines Knowledge Graph zusammen.

3.2 AIMS Infrastruktur

Zentraler Bestandteil von AIMS ist eine Web-Plattform, die intuitive Werkzeuge zur Erstellung, Veröffentlichung, Weiterentwicklung und Nachnutzung von Metadatenschemata als Applikationsprofile zur Verfügung stellt. Sie basiert auf dem in Abschnitt 3.1 beschriebenen Ansatz und ermöglicht die Dokumentation von Forschungsdaten durch Publikation von Metadatensätzen, die konform zu den mittels der Plattform verwalteten Schemata sind.

In Abbildung 3 ist der geplante Workflow gezeigt, über welchen Nutzende mit der AIMS-Plattform interagieren können. Zur Verdeutlichung des Fokus auf Nachnutzung starten Nutzende in der AIMS Plattform immer bei der Suche für Applikationsprofile (siehe Abb. 3 “search”). Auf Basis der eingegebenen Werte für die Metadaten eines Applikationsprofils (z. B. “Name”, “Beschreibung”, “Fachgebiet” usw.) werden passende Applikationsprofile gefunden. Nutzende haben nun die Möglichkeit eines der gefundenen Applikationsprofile zu verwenden, zu modifizieren oder zu erweitern.

Die Modifikation und Erweiterung finden in einem sogenannten Applikationsprofil-Generator statt. Der Applikationsprofil-Generator erhält Terme von einer “vocabulary database”, die zu einem Applikationsprofil hinzugefügt werden können. Diese Datenbank ist mit den für die AIMS Anwendungsfälle relevanten Terminologien befüllt und mit bestehenden Diensten verknüpft, die existierende Ontologien zur Verfügung stellen, Ein so erstelltes Applikationsprofil wird dann in einer “application profile database” gespeichert und steht für folgende Suchen zur Verfügung. Die Applikationsprofile können anschließend als Grundlage zur Dokumentation der eigenen Forschungsdaten verwendet werden (s. Abschnitt 3.3). Zur Dokumentation von Forschungsdaten anhand der Applikationsprofile steht außerdem ein “metadata store” zur Verfügung. Dieser ist ein Repositorium in dem mittels Applikationsprofilen validierte Metadatensätze abgelegt werden können. Dieses Repositorium für konforme Metadatensätze ist öffentlich zugänglich und durchsuchbar.

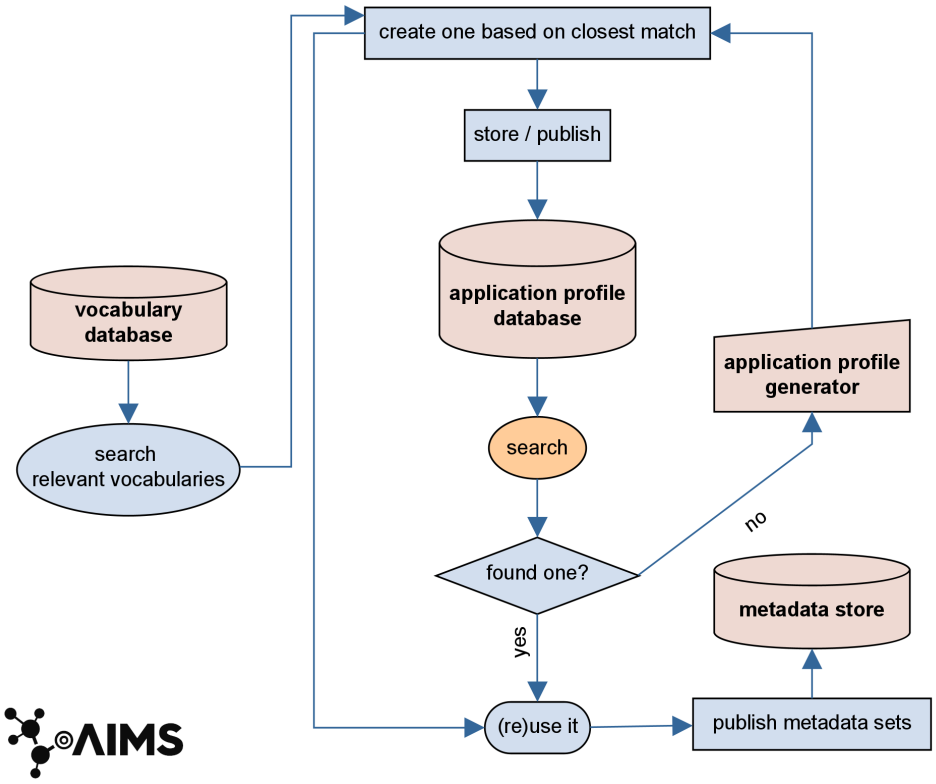


Abbildung 3: Abstrahierter Nutzer-Workflow der AIMS-Plattform.

Repräsentation von Metadaten und Applikationsprofilen

Zur Erstellung von Applikationsprofilen und Metadaten muss ein einheitliches Format gewählt werden, welches interoperabel agieren kann und welches bereits breit unterstützt wird. Auf Basis dieser Kriterien und da es ein bereits genutzter Standard für Ontologien ist, wird das Resource Description Framework (RDF) [14] zur Formulierung von Applikationsprofilen und Metadaten verwendet.

Zur Repräsentation von Applikationsprofilen ist es notwendig, eine einheitliche Beschreibung für diese zu finden. Diese sollte die Möglichkeit bieten, die gewünschte Struktur eines Metadatensatzes unter Rückgriff auf Terme aus kontrollierten Vokabularien nach dem in Abschnitt 3.1 beschriebenen Baukastenprinzip festzulegen. Zudem sollte die Beschreibungssprache bereits verbreitet sein und eine Notation in RDF möglich sein. Auf dieser Basis wird in AIMS die Shapes Constraint Language (SHACL) [15] genutzt, welche einen Standard zur Beschreibung von Applikationsprofilen in RDF darstellt. Zu deren Validierung stehen Implementierungen in den populärsten Programmiersprachen wie C#, JavaScript oder Python zur Verfügung. Mit SHACL ist es möglich für einzelne Attribute spezifische Einschränkungen zu beschreiben, sodass Applikationsprofile auch die Konformität der Metadaten mit bestehenden Definitionen validierbar machen können bspw. einen Wertebereich, ein Zahlenformat oder auch eine Klasse von Applikationsprofilen.

Applikationsprofil-Generator

Das Erstellen von Applikationsprofilen durch Auswahl geeigneter Terme aus kontrollierten Vokabularien wird für den Nutzer intuitiv durch eine grafische Oberfläche, die die Suche nach Termen und deren Zusammenfügen per Drag&Drop ermöglicht, ohne eine Auseinandersetzung mit der exakten Syntax von RDF und SHACL zu erfordern. Dieser Generator stellt eine der Hauptkomponenten von AIMS dar. Je nachdem, ob Forschende ein bereits bestehendes Applikationsprofil erweitern bzw. modifizieren oder ein neues Applikationsprofil erstellen möchten, starten sie mit einem jeweiligen Stand. Bei einer Erweiterung wird sich auf Vererbung gestützt, welche durch das RDF-Schema (RDFS) [16] mit dem Attribut “rdfs:subClassOf” gewährleistet wird.

Im Applikationsprofil-Generator besteht die Möglichkeit für Nutzende nach Terminologien aus Vokabularen zu suchen. Diese Vokabulare stammen aus einer “vocabulary database” und können auf Basis der AIMS Infrastruktur je nach Implementierung und Anwendungsfall verschieden sein. Die Ergebnisse der Suche werden je nach Relevanz, Benutzung und anderer Kategorien sortiert und Nutzenden aufbereitet dargestellt. Sie können nun mithilfe von Drag&Drop die einzelnen gewünschten Terme in ein Applikationsprofil hineinziehen und so zu diesem hinzufügen. Weiterhin können Nutzende ihrem Applikationsprofil und seinen Termen verschiedene Attribute geben, die die Struktur weiter verfeinern und etwa die Kardinalität, mit “sh:minCount” und “sh:maxCount” darstellt, oder den zulässigen Wertebereich festlegen. Über den Wertebereich kann dabei auf weitere Applikationsprofile verwiesen werden, sodass das in Abschnitt 3.1 beschriebene Prinzip der Modularität umgesetzt werden kann.

Sobald ein Applikationsprofil zur Zufriedenheit erstellt wurde, müssen zur Auffindbarkeit und Nachnutzung des Applikationsprofils relevante Metadaten, wie “Name”, “Beschreibung” oder “Fachgebiet”, eingetragen werden. Erst danach kann das Applikationsprofil in der Datenbank veröffentlicht werden.

Applikationsprofil Datenbank

Neben dem Generator stellt die Applikationsprofil-Datenbank ein weiteres zentrales Element der AIMS-Plattform dar. Die dort gespeicherten Applikationsprofile sind über einen kombinierten Ansatz durchsuchbar. Dieser kombinierte Ansatz ergänzt einen aus Titel und Beschreibung gespeisten Suchmaschinenindex zur Textsuche mit Filteroptionen. Die Filterung erfolgt über die zu jedem Applikationsprofil erfassten Metadaten und erlaubt es außerdem Informationen aus den Beziehungen zwischen Applikationsprofilen zur Navigation und weiteren Eingrenzung zu verwenden. Dadurch wird verstärkt die Nachnutzung von Applikationsprofilen angeregt, was Parallelentwicklungen vorbeugt. Gleichzeitig können gewonnene Erkenntnisse aus der Nutzungs- und Suchstatistik für eine kuratierte Terminologieentwicklung im Rahmen von NFDI4Ing genutzt werden.

Um die notwendige Freiheit zur Gestaltung von individuellen, spezifischen Applikationsprofilen zu bieten bleibt die Verantwortung für eine Qualitätssicherung und Pflege der Inhalte der AIMS Infrastruktur letztlich bei den Nutzenden. Nicht zuletzt um dies zu erleichtern stellt die Plattform ein Rechte- und Rollenmodell zur Verfügung.

3.3 Standardisierte Metadaten in der Praxis

Disziplinspezifische Metadatenschemata sind eine Schlüsselkomponente zur Standardisierung von Metadaten und eröffnen neue Möglichkeiten zur Optimierung der Datengenerierung und -analyse während der aktiven Forschung. Allerdings ist hinsichtlich der Akzeptanz bei den Forschenden darauf zu achten, dass die Nutzung von disziplinspezifischen Schemata nicht als einschränkend aufgefasst wird und mit möglichst geringem Mehraufwand verbunden ist.

In AIMS wird daher ein integrativer Forschungs- und Entwicklungsansatz verfolgt, bei dem der in Abschnitt 3.1 beschriebene Modellierungsansatz und die zu entwickelnde Plattform an verschiedenen Prüfständen am FST der TU Darmstadt sowie am WZL der RWTH Aachen praktisch angewendet und iterativ verbessert wird. Die Entwicklung der Plattform wird also aktiv von den Forschenden mitgestaltet und getragen. Dabei ist die aufwandsarme Integration einer strukturierten Dokumentation von Forschungsdaten (in Form von durch Applikationsprofile beschriebener Metadaten) in den Forschungsalltag zentraler Bestandteil des Projekts. Im Vordergrund stehen bei der Entwicklung außerdem, wie durch die entwickelte Plattform und die dort verwalteten Applikationsprofile die Generierung und -analyse von Metadaten vereinfacht und automatisiert werden kann.

Etablierung von Quasistandards für die Generierung von Metadaten

Die Plattform von AIMS unterstützt Forschende bei der Standardisierung der Dokumentation von Prüfständen und Messdaten. In der Praxis müssen die über die Plattform generierten Schemata dafür auf eine Vielzahl unterschiedlicher, disziplinspezifischer Prüfstände anwendbar sein und gleichzeitig eine hohe Benutzerfreundlichkeit aufweisen.

Hier besteht das Risiko, dass sich diese beiden Anforderungen diametral gegenüberstehen. Durch den partizipativen Ansatz, der die aktive Beteiligung der Forschenden fordert und fördert, kann den Herausforderungen, die sich insbesondere aus der Heterogenität des Anwendungsfeldes ergeben, Rechnung getragen werden: Anstatt eine große, aber zwangsweise unvollständige Menge definierter Schemata vorzugeben, ermöglicht AIMS den Forschenden disziplinspezifische, individualisierte Metadatenschemata auf Basis kontrollierter Bausteine und unter Rückgriff auf bereits durch andere Forschende erstellte Schemata eigenständig zu definieren. Durch diese Wiederverwendung und Weiterentwicklung wird in einem evolutionären Prozess die Etablierung von Quasi-Standards unterstützt. D.h. De-Facto-Standards, die sich in der Praxis durchgesetzt haben, aber keinen formalen Prozess zur Erlangung eines Konsenses durchlaufen haben [17].

Ein solcher De-facto-Standard kann durch ein formelles Normungsverfahren in eine weitreichend anerkannte Norm überführt werden (beispielsweise PDF, HTML). Als anerkannte Regeln der Technik können diese so regulierende Bedeutung in Forschung und Produktion erlangen. Durch die in Abschnitt 3.1 beschriebene Modularität und Hierarchie kann dabei ein hohes Maß an Wiederverwendbarkeit bei maximaler Interoperabilität gewährleistet werden.

Zudem ermöglicht ein verfügbares Metadatenschema, Forschungsabläufe so zu gestalten, dass sie die größtenteils automatisierte Generierung von Metadaten unterstützen. Dies gewährleistet die Qualität der Metadaten hinsichtlich des abgebildeten Prüfstandes und der tatsächlichen Messdaten. Benutzereingaben wie Informationen über den Benutzer, das Projekt oder Steuerungsparameter können ebenso adressiert werden. Software, die Forschungsdaten generiert, muss gegebenenfalls beide Wege (manuell und automatisiert) zur Generierung von Metadaten in die Funktionalität jedes ihrer Module integrieren. So ist jedes Modul verantwortlich für die Protokollierung seiner entsprechenden Metadaten. Dadurch wird die Notwendigkeit einer nachträglichen Zuordnung vermieden und Informationsverlust minimiert. Der Prozess wird unterstützt durch grundlegende Kommunikations- (I-O-Link-Kommunikation), Steuerungs- und Online-Verarbeitungsmechanismen.

Automatisierung bei der Datenanalyse zur Ergebnisinterpretation

Über die Dokumentation der Forschungsdaten hinaus, liegt der Nutzen einer standardisierten Dokumentation in der Vereinfachung von Prozessen für Abruf, Kuration und Auswertung von Forschungsdaten. Durch die Verwendung definierter Terminologien sind Forschungsdaten, die durch einheitliche Schemata gemäß der FAIR-Prinzipien dokumentiert sind, leichter auffindbar, verständlich und wiederverwendbar. Die in AIMS verwen-

deten Methoden und Technologien garantieren die Maschinenlesbarkeit der Metadaten, sodass Prozesse zur Verarbeitung und Analyse von Forschungsdaten leicht automatisiert werden können. Durch den Export dieser Metadaten in das in AIMS entwickelte Repositorium können die Indizierungs- und Suchfunktionen des Repositoriums genutzt werden, um selektiv alle Datensätze abzurufen, die gewünschte Kriterien erfüllen. Dies erleichtert die Wiederverwendung vorhandener Daten, auch außerhalb des Projekts in dem sie entstanden sind. Darüber hinaus entsteht durch die einheitliche, mittels Applikationsprofilen standardisierte Dokumentation von Arbeitsabläufen und resultierenden Datenprodukten ein vernetzter Wissensgraph, der das generierte Wissen in einer nachvollziehbaren und verständlicheren Weise bewahrt, wovon andere Projekte in hohem Maße profitieren.

Außerdem können durch die standardisierte Beschreibung der Metadaten nicht nur die (Meta-) Daten selbst, sondern auch Programme, die auf diesen arbeiten, wiederverwendet werden. AIMS ermöglicht es somit, selbst entwickelte Programme nachhaltiger zu gestalten und die Routineschritte in der Datenhandhabung und -analyse wie die Verarbeitung von Rohdaten oder die Quantifizierung der Unsicherheit wiederverwendbar zu automatisieren. Dadurch kann redundanter Code vermieden werden, der üblicherweise bei der Verarbeitung nicht standardisierter Daten entsteht. Die Analysewerkzeuge werden durch breitere Anwendung umfassender und insgesamt weniger fehleranfällig.

4 Zusammenfassung

Es wurde dargelegt, dass die ingenieurwissenschaftliche Forschung mit ihren oft sehr heterogenen Themengebieten, Forschungsfragen und Untersuchungsgegenständen besondere Herausforderungen an die Gestaltung von Metadatenbeschreibungen stellt. Maschinenlesbare Metadatensätze stehen zunehmend im Fokus als Lösungsansatz unter anderem für den Umgang mit zunehmenden Datenmengen, auch im Kontext der Industrie 4.0. Beides erzeugt Anforderungen, für die bisher keine anwendbaren, etablierten Lösungen gefunden werden konnten.

In AIMS wird ein aus der Forschung getriebener Ansatz verfolgt, der diese Anforderungen zielgenau adressiert. Zentrale Elemente sind dabei ein Modellierungsansatz, der flexible und maximal interoperable Metadatenschemata in Form hierarchischer und sich modular referenzierender Applikationsprofile erlaubt, die auf Basis kontrollierter Terminologien im Baukastenprinzip trotzdem hochspezifische Beschreibungen ermöglichen. Dieser Ansatz wird in eine Web-Plattform überführt, die unter einer benutzerorientierten Oberfläche ein Repositorium für fachspezifische Applikationsprofile zur Nachnutzung und einen Generator zu deren Erzeugung und Weiterentwicklung bereitstellt. Diese wird ergänzt um ein Repositorium, in dem den Applikationsprofilen entsprechende validierte Metadatensätze veröffentlicht werden können. Forschenden wird so ein niederschwelliger Zugang zu einem verbesserten Metadatenmanagement ermöglicht.

Ebenso wichtig wie die Erzeugung fachspezifischer Applikationsprofile und Terminologien sind Wege, die es erlauben, sie in die praktische Forschung zu integrieren. Zu diesem Zweck

werden in den beteiligten Forschungsgruppen Verfahren entwickelt, um die Erzeugung und Nutzung standardisierter Metadaten zu optimieren.

Nicht zuletzt ist es ein wichtiges Anliegen von AIMS die Projektergebnisse in der Ingenieurs- und FDM-Community zu vermitteln und einen stetigen Austausch zu den Projektergebnissen und dem entstandenen Wissen zu initiieren, um über vernetzte Initiativen und Projekte die verbundenen Institutionen und letztlich eine größtmögliche Forschungs- und Anwendergemeinschaft im Umfeld der Ingenieurwissenschaften zu erreichen. Zusammen sollen diese Schritte und Maßnahmen ein verbessertes Metadatenmanagement 4.0 gewährleisten und durch AIMS gemeinsam eine Umgebung geschaffen werden, in der Metadaten wachsen können.

Danksagungen

Mit besonderem Dank für die Unterstützung durch Gerald Jagusch (ULB), Ina Heine (WZL), Wolfgang Stille (ULB) sowie durch das eScience Team der RWTH Aachen University vertreten durch Annett Schwarz und Florian Claus. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Projekt-ID 432233186.

ORCID IDs

- Matthias Grönewald https://orcid.org/0000-0002-3480-9102
- Patrick Mund https://orcid.org/0000-0003-0890-5472
- Matthias Bodenbenner https://orcid.org/0000-0002-9413-1874
- Marc Fuhrmans https://orcid.org/0000-0002-9826-018X
- Benedikt Heinrichs https://orcid.org/0000-0003-3309-5985
- Matthias S. Müller https://orcid.org/0000-0003-2545-5258
- Peter F. Pelz https://orcid.org/0000-0002-0195-627X
- Marius Politze https://orcid.org/0000-0003-3175-0659
- Nils Preuß https://orcid.org/0000-0002-6793-8533
- Robert H. Schmitt https://orcid.org/0000-0002-0011-5962
- Thomas Stäcker https://orcid.org/0000-0002-1509-6960

Literaturverzeichnis

[1] J. Ludwig. “Leitfaden zum Forschungsdaten-Management.”, 2013.

[2] Deutsche Forschungsgemeinschaft e. v, ”Leitlinien zum Umgang mit Forschungsdaten.”, 2015.

[3] Laboratory Informatics Institute Inc. “The Complete Guide to LIMS & Laboratory Informatics.”, ed. J. Johnes, Atlanta, 2017.

[4] Z. Chen, D. Wu, J. Lu, and Y. Chen. ”Metadata-based Information Resource Integration for Research Management.”Procedia Computer Science, vol. 17, pp. 54-61, January 01, 2013.

[5] M. Franklin, A. Halevy, and D. Maier. ”From databases to dataspaces: A new abstraction for information management.”, Sigmod Record, vol. 34, no. 4, pp. 27-33, December, 2005.

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

[7] R. Van Noorden. ”Data-sharing: everything on display.”Nature, vol. 500, 2013.

[8] J. Greenberg, S. Swauger, and E. Feinstein. ”Metadata Capital in a Data Repository.Ïnternational Conference on Dublin Core and Metadata Applications; DC-2013– The Lisbon Proceedings, September 2013.

[9] M. Greenwald, T. Fredian, D. Schissel, and J. Stillerman. Ä metadata catalog for organization and systemization of fusion simulation data.ïn English, Fusion Engineering and Design, vol. 87, no. 12, pp. 2205-2208, December 2012.

[10] D. P. Schissel et al. Äutomated metadata, provenance cataloging and navigable interfaces: Ensuring the usefulness of extreme-scale data.”(in English), Fusion Engineering and Design, vol. 89, no. 5, pp. 745-749, May, 2014.

[11] M. Politze, et al. “How to Manage IT Resources in Research Projects? Towards a Collaborative Scientific Integration Environment.” European journal of higher education IT, 1(2020/1), 5. 2020.

[12] R. H. Schmitt, V. Anthofer, et. al., “NFDI4Ing - the National Research Data Infrastructure for Engineering Sciences”, 2020, doi:`https://doi.org/10.5281/zenodo.4015201`.

[13] K. Coyle, T. Baker. “Guidelines for Dublin Core™ Application Profiles” `http://dublincore.org/documents/profile-guidelines/`, 18.05.2009. [Stand 14.04.2021]

[14] O. Lassila. ”Resource Description Framework (RDF) Model and Syntax Specification.” `https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/`, 22.03.1999. [Stand 14.04.2021].

[15] H. Knublauch, Dimitris Kontokostas. “Shapes Constraint Language (SHACL).”https://www.w3.org/TR/shacl/ (20.07.2017) [Stand 14.04.2021].

[16] D. Brickley, R. Guha. “RDF Schema 1.1.” https://www.w3.org/TR/rdf-schema/. (25.02.2014) [Stand 14.04.2021].

[17] T. Carpenter.”9 - Electronic publishing standards.”, Academic and Professional Publishing, pp. 215-241, Chandos Publishing, 2012. https://doi.org/10.1016/B978-1-84334-669-2.50009-3.

Product Life Cycle Oriented Data Management Planning with RDMO at the Example of Research Field Data

Iryna Mozgova[1], Gerald Jagusch[2], Jens Freund[2], Angelina Kraft[3], Tobias Glück[1], Kevin Herrmann[1], Marvin Knöchelmann[1] and Roland Lachmayer[1]

[1]Leibniz University Hannover, Institute of Product Development
[2]Technical University of Darmstadt, University and State Library
[3]TIB - Leibniz Information Centre for Science and Technology University Library, Hannover

A research data management plan (DMP) is an integral part of the organization of working with research data according to the FAIR-Data-Principles. The task area Base Services of the National Research Data Infrastructure for Engineering Sciences (NFDI4Ing) support the open source software Research Data Management Organiser (RDMO) as an important and well established tool for data management planning. The purpose of this contribution is to demonstrate the application of RDMO for the organization of field data throughout a product life cycle at the example of the pilot use case of the archetype Golo. The possibilities of creating DMPs for the outlined use case using RDMO are presented.

1 Introduction

Research data management (RDM) is gradually becoming an integral part of the activities of modern scientific and educational organizations. In engineering disciplines, whose spectrum is itself quite broad and closely related to related scientific fields such as the natural sciences, research data from simulations or experiments with a wide variety of processes and methods are heterogeneous [1]. The heterogeneity of research data ranges from measurement data, observational data, simulation data, surveys and interviews to video and imagery.

Historically, different models, concepts, techniques, methods and data sets used in engineering disciplines have been developed for specific needs within specific areas of research and educational organizations. This can partially explain the fact that at present there is no unified concept of engineering data handling, standardized metadata schemas, generic and domain-specific ontologies and repositories of engineering research data. Typical problems are the exchange of big data, versioning of data, models and the corresponding program code, semantic linking of data, data reusability, reproducibility of results, etc. Cultural and historical aspects of working with research data, the variety of standards and approaches adopted by different research and educational organizations, and, therefore,

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13722 (CC BY-SA 4.0)

taking into account different goals and the need for changes in the culture and processes of organizations, also play an important role [2]. The overall goal of implementing RDM and workflow standards is to systematically make data findable, accessible, interchangeable, and reusable [3]. Germany's National Research Data Infrastructure (NFDI) aims to bring together existing infrastructure components, services and research communities to create a comprehensive, interdisciplinary RDM [4].

1.1 NFDI4Ing - National Research Data Infrastructure for Engineering Sciences

Initiated in 2017 by RWTH Aachen University and the Technical University of Darmstadt, the consortium NFDI4Ing aims to create a research data management system that ensures the preservation of new scientific findings, global aggregation and transparent promotion through standardized data management [5]. For modular and sustainable development, the research profiles of engineers are structured into so-called archetypes, which represent sub-areas of heterogeneous engineering activities and thus enable the systematic and oriented development of research data management solutions. The concept of archetypes has been developed as a result of a systematic analysis of various typical research and engineering activities and has been validated in the engineering community [6].The archetypes are named for easier distinction and reference. In order to bring together the common aspects of the archetype-specific requirements as well as to promote cross-consortium communication, the standardized, modular and central services are implemented via the Base Services.

Archetype Golo deals with typical activities that belong to the research of technical systems, include planning, recording and subsequent analysis of field data for examination of the operating conditions of a technical system and/or subsequent adaptation of a system to the environment and real operating conditions.

1.2 Life cycle of research field data

One example of such a technical system are high-resolution vehicle headlamps (see section 2). Vehicle headlamps can be either considered as a separate independent system, a system within the system of the car, or to study the data obtained in the context of the use and interaction of several technical systems.

The product development process follows a V-model according to VDI guideline 2206 [7]. In the V-model, the process of technical system development starts with the definition of the requirements, followed by a system design considering various influencing factors, a domain-specific design, and the prototype. The prototype goes through one to several development and validation cycles, which in turn can result in the adaptation of the prototype or even its new development. The result of a completed cycle of the V-model is the product, which features a certain degree of maturity of the planned final product. For

the RDM it means the occurrence of different types of collected data like test data and field data. The test data is generated at the beginning of the research process by means of laboratory experiments and various simulation settings (Fig. 1). The field data collected by testing the developed prototype under real environments with new, additional factors is associated to it.

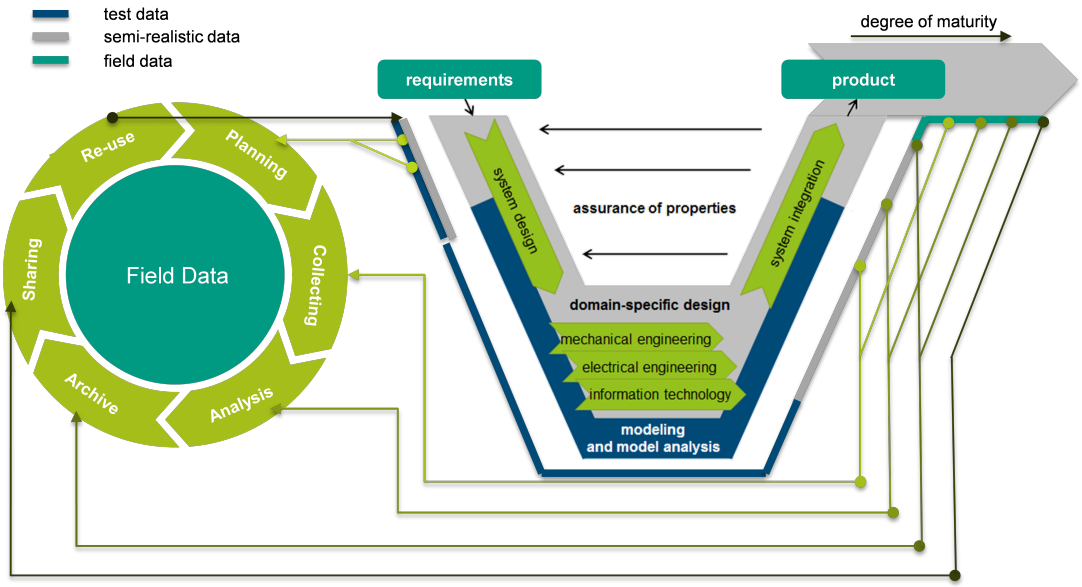


Figure 1: V-model and research data lifecycle.

In the planning phase, both qualitative and quantitative test results must be thoroughly documented. Often this documentation is scattered over several files, is partly done manually, and is rarely standardized. Therefore, it is desirable to use metadata standards and to choose ontologies that allow researchers to select appropriate attributes for the relevant data type and describe them in as standardized a way as possible. In addition, a specification of provenance metadata is required.

Because engineering research projects work with different types of data with different access rights, a data management platform that allows role assignment and rights management is needed to fully collect and store data. This is often accomplished through the services of the research organization and providing access to external project partners. Here, usage of a data management platform that provides functionalities such as standardized data and metadata storage and a possibility for standardized storage of test protocols would be favourable.

During the data analysis phase, depending on the data type, different data volumes in the range of MB to TB are generated; these include raw data as well as extensive simulation data of partly proprietary software components and in-house developments.

Within archiving and providing access to the data the interests of all research participants, e.g., educational and research organizations as well as industrial partners, and the possible mixed project funding should be taken into account. There is a continuous data exchange

between the partners throughout the research data life cycle. This mixed financing of research projects requires graded rights of subsequent use of the resulting research data. The choice of a data management platform for data exchange and storage should be made with the guarantee of long-term archiving of data and metadata and thus ensure their potential reuse.

2 Use Case: Validation of high-resolution vehicle headlamp

The purpose of this example is to demonstrate how the archetype Golo can use DMP for an organization of the RDM process for the development of a high-resolution vehicle headlamp. The fundamental aim of vehicle lighting technology is to optimize visibility for the driver and the visibility of the vehicle for other road users. Two main functions are to be taken into account in the design: lighting the road section and providing information signals to road users, including pedestrians. To realize the functions of lighting, traffic analysis and displaying warning signals a separate block for information processing can be used. An increasing networking, as with the vehicle camera and other sensors, requires an early definition of data and interfaces and requires a frequent exchange of domain-specific field data [8]. Within modeling, implementing and testing all blocks of the system, test sets of quantitative and qualitative data are used.

2.1 Organization of the validation process and data types

The distraction potential for other road users is investigated in a study, when road projections are directed at the driver of the projecting car. For this, a blind study with different test persons driving a constant motorway section are overtaken by a vehicle projecting an image onto the street surface. The distraction potential is examined with an eye-tracking system, which detects the direction of the subjects gaze. Another camera behind the windshield of the test subject's vehicle records the traffic area in front of the vehicle. In addition, the subjects' physiological perception of the headlight projection is recorded with a questionnaire afterwards. Deviations from the intended test procedure are also noted. The sheets with the answered questionnaires and notes are scanned after each trip.

2.2 Project Data Management Plan

For the research data life cycle of this use case, a DMP should be created at the beginning of the product development process, which is adapted to the process steps described in the V-model. The developed DMP template is a Word-document, a checklist in the form of tables with questions. The DMP takes into account the different project phases: initiation, execution, controlling and closing. In the DMP template the project execution set includes subsections: general information, general questions about the experiment, questions about

the raw data of the experiment, questions of the evaluation of the experiments and the subsection other. A cut-out of the section with questions from initiation phase is shown in Fig. 3. The process of filling out the DMP involves quality control of the answers by the project management and the management of the responsible departments, as well as feedback from the project developers who filled out the document, in order to improve the informativeness and quality of the DMP. The first scenario described in the following paragraph explains how to transform the developed DMP template, including the different phases of the project and the corresponding quality control workflows, into a catalog in RDMO.

3 Data Management Planning with RDMO

3.1 What is RDMO?

The Research Data Management Organiser (RDMO) is a software that enables institutions as well as researchers to plan and carry out their management of research data. RDMO can assemble all relevant planning information and data management tasks across the whole life cycle of the research data. RDMO is ready for application all kind of research projects. It is run by several universities and research institutions all over Germany [18]. The current main feature of RDMO is the provisioning of templates for DMPs (so called catalogs), allowing the collaborative creation and maintenance of neatly standardized DMPs. The templates can be customized individually while in same time still being interoperable.

In two project phases funded by DFG, the tool was continuously developed by the project partners Leibniz Institute for Astrophysics (AIP), Fachhochschule/University of Applied Sciences Potsdam (FHP), and the library of Karlsruhe Institute of technolgy (KIT) in close cooperation with users. The tool was extended by enhancing its implementation of roles and interfaces to institutional infrastructure, e.g. repositories, ticketing systems, and the infrastructure for authentication and authorization. The development of RDMO and the organisation of DMP templates is continued by the RDMO Working Group[17].

3.2 How we want to make use of RDMO in NFDI4Ing?

In the NFDI4Ing consortium there are two tasks concerned with RDMO that are included in the Base Service measure ”Quality assurance in RDM processes and metrics for FAIR data”, that is carried out jointly by IPeG Leibniz University Hannover, WZL RWTH Aachen University, University Library RWTH Aachen and University and State Library Darmstadt (ULB Darmstadt). In the one task NFDI4Ing will offer a DMP service to all researchers of engineering through a central multitenant RDMO instance maintained by ULB Darmstadt.

It provides a highly adaptable user interface, authorisation procedures and customizable DMP templates per client, but runs on a single database, thus allowing easy collaborative data management planning in multi-institutional research projects. Local RDMO editors can take care of the content of one client. We will implement new features in RDMO regarding 1) the validation of entries in DMPs in order to connect RDMO with RDM maturity models, 2) Usability improvements, e.g. a recommender for suitable DMP templates, 3) connection to other RDM systems via the RDMO API (e.g., repositories, metadata services, ORCID, CRISs). Furthermore, we will be engaged in the RDMO Working Group regarding the maintenance of the code.

In the second task we will foster the use of specific DMP templates in engineering. We will develop, test and evaluate archetype and sub-discipline specific DMP templates, for different stages of research projects, as well as in close collaboration with all archetypes and community clusters. We will also investigate how to support researchers and local RDM offices by designing a DMP review service based on maturity levels. The establishment of an RDM certification scheme introducing an "NFDI4Ing seal for FAIR data" is our goal in the long run.

3.3 From a local check list to RDMO - a researcher's perspective

3.3.1 An introduction to RDMO's data model

There are two starting situations for researchers who would like to develop their own data management plan questionnaires or *catalogs* in RDMO: Either a data management checklist already exists, e.g. in the form of a text file, which is to be transferred to RDMO, or there is no such preliminary work. In both cases it is highly recommended to build upon existing RDMO catalogs already developed by the RDMO community. On the one hand this approach makes work easier, on the other hand – and even more important – it increases the interoperability and therefore reusability of the new catalogs.

To achieve this, it is essential to have a basic understanding of RDMO's underlying data model, of which Fig. 2 shows an important part.

Catalogs in RDMO consist of *elements*, including *questions*, *attributes*, *options* and *conditions* (the latter not shown in Fig. 2). Within a catalog, questions can be organized in question sets and sections as parent structural elements, options in option sets. These elements are predefined by the creator of the catalog.

To create a data management plan for a specific research project based on a catalog, the researchers first have to create a new project in RDMO. During the following structured interview, RDMO presents them the predefined question texts which are internally stored in the question elements. The answer or *value* given to a specific question by the researchers is stored in RDMO by linking it to an attribute which is also linked to that same question, as shown in Fig. 2. To make the filling of the questionnaire easier, it is possible to present predefined answer *options* to the researchers, from which they can

select the appropriate ones. Options in RDMO are organized in option sets. An example of an option set could be the faculties of a university, with each faculty being one option. Generally, this concept of decoupling the predefined question elements (blue in Fig. 2) from the user specific content (green) through attribute (red) and option elements (orange) makes it possible for researchers to switch the project's catalog while their answers are taken over to the new catalog automatically, provided that the latter elements were chosen in a standardized way.

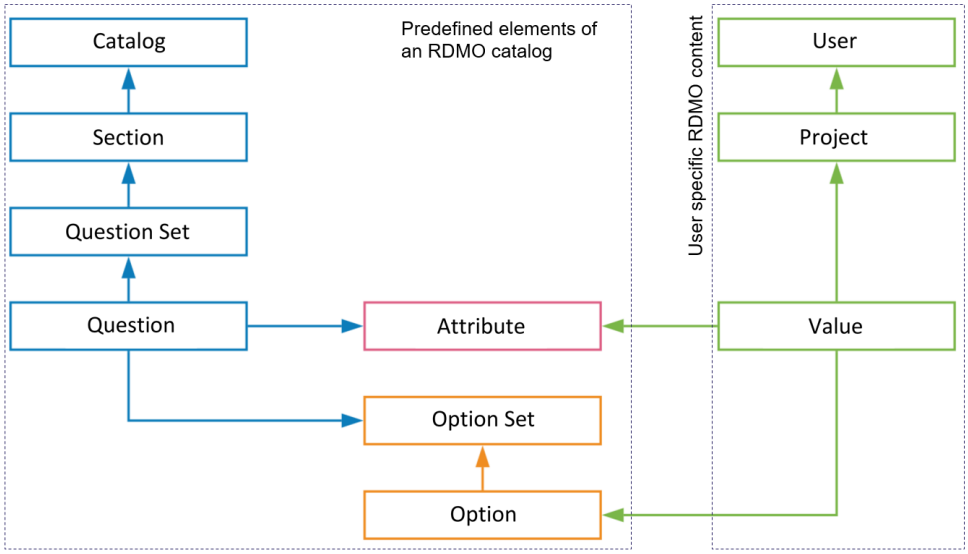


Figure 2: Part of RDMO's data model. Illustration slightly modified taken from [9].

Every element in RDMO is identified by a URI, i.e. a string of characters with the structure of an internet address, but without an actual page behind it. Although, technically, all URIs may be defined completely freely by the developer of a new catalog, for interoperability purposes it is highly desirable to reuse existing attributes with the official RDMO prefix[1] *https://rdmorganiser.github.io*. An RDMO XML file containing all existing attributes with that prefix can be found in the folder *rdmorganiser/domain* of the RDMO-Catalog GitHub repository [19].

By importing this file into an RDMO instance, the attributes become available for catalogs within this instance. It is therefore recommended to import the RDMO Domain XML file immediately after the installation of RDMO.

[1]In RDMO, the URI prefix is defined as the first part of the URI and is important for data exchange between different RDMO instances. [15] Typically, the prefix should indicate the RDMO instance to which the respective element belongs, although in principle it can be freely chosen as long as it adheres to a basic URL scheme (see [15]). In contrast to other elements, such as questions, where the prefix might be changed if they have been imported from an external source to a local RDMO instance, the prefixes of attributes with the `https://rdmorganiser.github.io.` prefix should not be changed because of the interoperability reasons that are described in the text. In this context, an RDMO prefix can be compared to an XML namespace which is also a unique identifier for a collection of (XML element and XML attribute) names.

Each attribute in this file should uniquely represent one concept, e.g. "file size", and the RDMO Content Group, which is part of the RDMO Working Group [17], ensures that only unique attributes are newly added to the file.[2] Suppose, there are two catalogs A and B, each containing its own question about file size, but linked to this same attribute. If researchers decide to create a project based on catalog A, they can switch to catalog B at a later point without losing their given answer since the connection between the question element and the answer value is established via the same attribute. Since each question in RDMO is connected to exactly one attribute and each attribute can represent a concept, conversely, each question should ask for one concept to improve the interoperability of the catalog. If, on the other hand, one question asks for several concepts at the same time, the answer value has to be linked to a more generic attribute and cannot be split up again (at least not in an automatic way) when changing to a catalog with specific questions.

3.3.2 First Scenario: A data management checklist exists

Fig. 3 shows an extract of an existing data management checklist in .docx format used for the use case validation of high-resolution vehicle headlamps (Chapter 2). In order to transfer this checklist to RDMO, at first a new catalog must be created within the RDMO instance. This includes the assignment of a title, a URI prefix and a key. While the URI prefix can be freely chosen in principle, it is a good idea to choose it as your own RDMO's domain name, e.g. `https://tudmo.ulb.tu-darmstadt.de` for the RDMO instance of TU Darmstadt. Next, sections, question sets and questions can be created. Since the technical aspects of creating a catalog are very well described on *forschungsdaten.org*, we refer to this documentation [20] and will focus on the content process of finding and reusing existing attributes to make the new catalog interoperable. In general, if one is working on a new RDMO catalog, one can add information about it to the "Catalogs in Development" table on forschungsdaten.org [16] so that other groups who might be interested in collaborating can get into contact.

Using question 1.5 about the size of the expected research data as an example, the first step in transferring this checklist to the newly created RDMO catalog would be to check whether an appropriate attribute for this question already exists with the official RDMO prefix *https://rdmorganiser.github.io*. Provided that the above mentioned RDMO Domain XML file has been imported to the local RDMO instance, one option would be to go through the attributes in the window of the management interface (see Fig. 4) manually, looking for an appropriate attribute. Although the attributes can also be searched

[2] During the RDMO DFG project phase, Heger (2016) [10] explored the possibility of mappings between RDMO attributes and existing metadata schemas, the most promising being CERIF or SEMCERIF [11] as well as Friend of a Friend (FOAF) [12] and DCMI [13]. However, since these metadata schemas allowed only the mapping of a limited number of attributes, at the moment the RDMO Content Group does not conduct a mapping to existing schemas for new attributes. If, however, persons who have developed new attributes, send them to the Content Group for the review process and decide to provide a mapping to existing metadata standards, this would always be highly appreciated (although it is not mandatory). At the moment, the mapping could be provided in the *dc:comment* elements of the RDMO XML Domain file [14]

for by their URI, it might be difficult to find the right one, in this case *https://rdmorganiser.github.io/terms/domain/project/dataset /size/volume*, since one has to search for the exact term that is used within the URI.

Checklist

Nr.	Questions / Activities	Answers / Releases
1	**Project Initiation**	
1.1	If required: introduction to the research data management by the RDM team.	
1.2	Selection of a person responsible for the research data management of the project.	
1.3	What research activities are involved in the project?	
1.4	What research data is to be expected?	
1.5	What is the expected size of the research data?	
1.6	Who needs to access the data during the project?	

Figure 3: An excerpt from a data management checklist of the IPeG at Leibniz University Hannover. The highlighted question is used in the text as an example of how existing checklist questions can be transferred to RDMO.

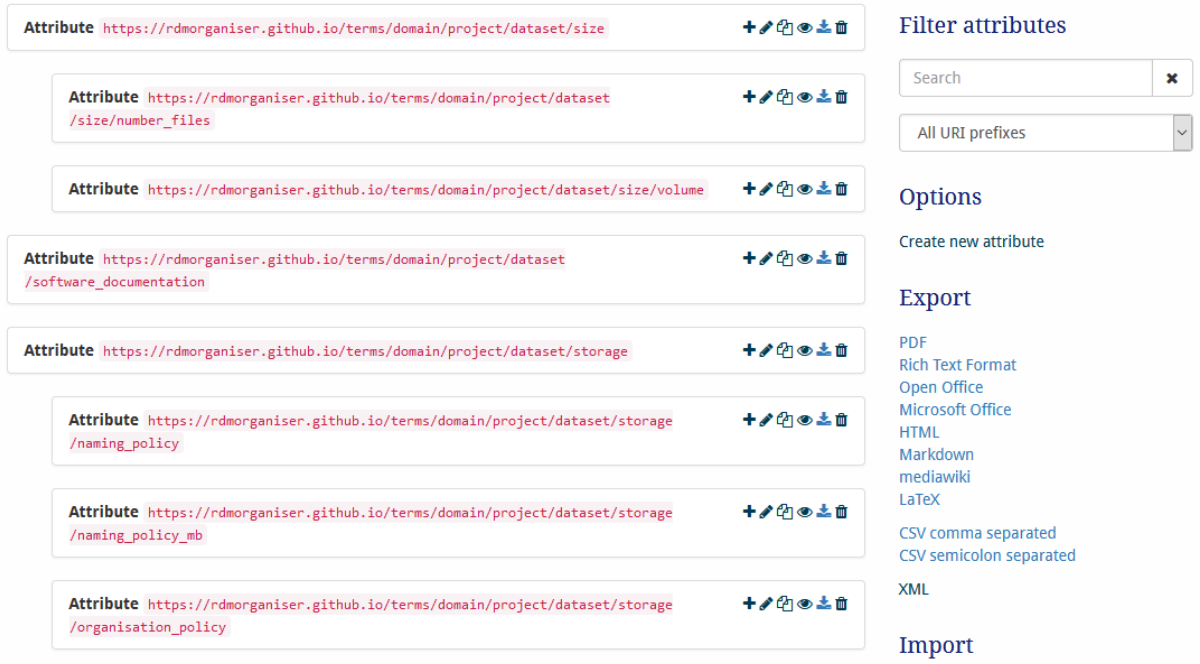


Figure 4: RDMO Domain menu.

A more powerful way to search for attributes is the web tool RDMO-Terms [22] which allows to search all elements of all catalogs available at the central RDMO content repository. By searching for questions regarding file size there is a greater chance of finding the appropriate attribute than by searching for the attribute directly. The reason is that question elements do not only consist of a URI string[3] but also of the question texts,

[3]In fact, attributes can also contain a description of their meaning in form of a comment. However, in practice most attributes do not have such a comment.

mostly at least in English and German, which are also searched. As an example, Fig. 5 shows one of the questions found by a search for the term ”size” from which the appropriate attribute can be identified. It is recommended to check several of the found questions in order to be sure that the most suitable[4] attribute has been identified.

In case an appropriate attribute does not already exist, it should be created, initially using your own organisation's RDMO prefix. When you decide to publish your catalog, it will be checked by the RDMO Content Group, that all newly created attributes are indeed unique, i.e. that they do not have an existing equivalent attributewith the `https://rdmorganiser.github.io` prefix. To facilitate this work, it is beneficial to copy the text of the associated question into the comment field of the attribute. A brief outline of the editorial workflow is given in section 3.3.

While it is possible now to create the question about the file size manually in the new catalog and link it to the found attribute, it can save some time to copy the whole question from the original catalog[5] via RDMO's question copy function (see Fig. 6) into the new catalog. The links to the attribute as well as to a possible option set are taken over at the same time. The prerequisite for this is that the respective catalog's questions and, if applicable, options XML files were previously imported from the RDMO catalog GitHub repository.

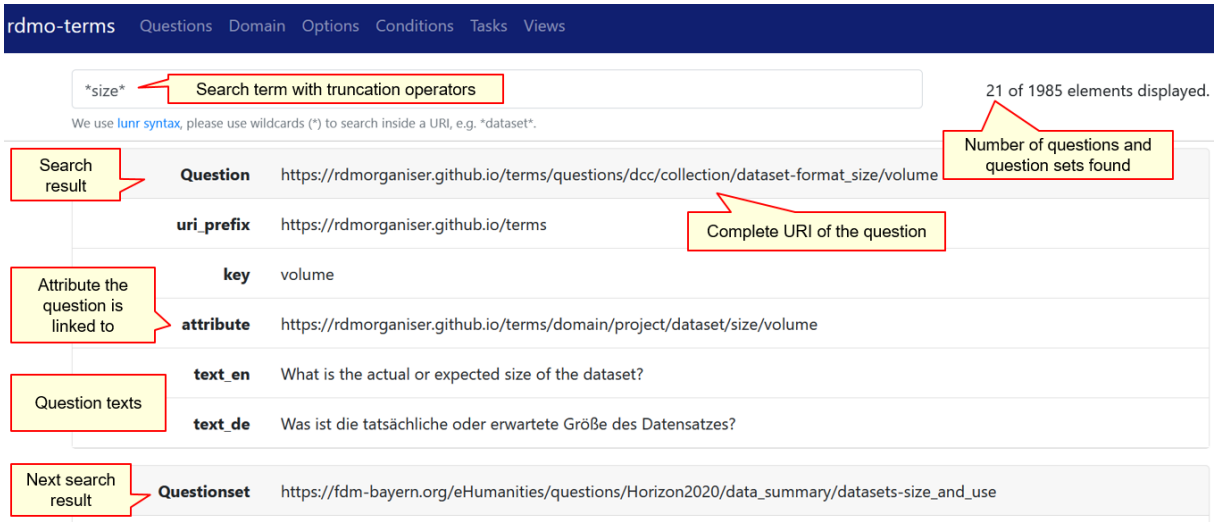


Figure 5: Search for questions containing the term ”size” within the web tool RDMO-Terms.

3.3.3 Second scenario: No data management check list exists

When there is no pre-existing checklist, a good strategy is to first look at the existing RDMO catalogs that have already been developed by the RDMO community and

[4] In the end it is up to the creators of the catalog to decide if the found attribute is appropriate for their question.

[5] In this case the RDMO dcc catalog which can be found under `https://github.com/rdmorganiser/rdmo-catalog/tree/master/rdmorganiser/questions`

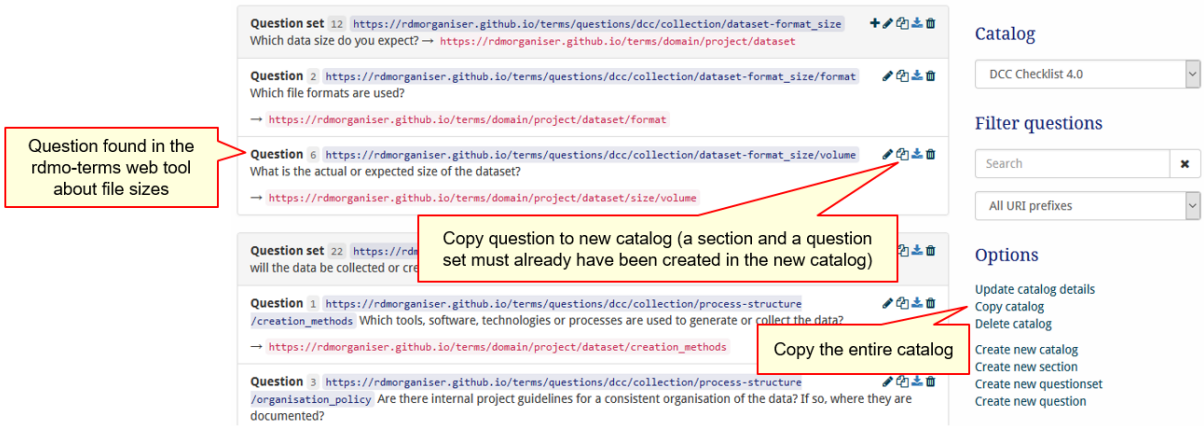


Figure 6: Copy question and copy catalog functions of RDMO.

published in the RDMO-Catalog GitHub repository. There are two folders, one of them ("rdmorganiser") with catalogs developed by the RDMO project and the other ("shared") containing catalogs from different institutions and initiatives. Most of these catalogs come with a documentation, describing the content of the respective catalog, which makes it easy to identify catalogs that might be of interest to ones use case. In addition, the catalogs of the official RDMO project can also be tried out live in the RDMO project's demo instance[6].

Ideally, one of the official or shared catalogs already fulfils the project's requirements. The only task that has to be done then would be to import all the necessary content files of the catalog in the local RDMO instance, which for shared catalogs often also includes files from the official RDMO catalog (especially the domain file with the attributes) since these are often referenced.

Even if none of the catalogs fully meets the requirements, a new catalog can be created reusing content from different existing catalogs. As described in section 3.3, one could first create an empty catalog and fill it with self created questions as well as questions copied from other catalogs. If, on the other hand, one catalog already meets the requirements well and only minor changes are needed, it makes sense to implement these changes either in the imported catalog directly or in a copy, generated via RDMO's copy catalog feature (see Fig. 6). This latter option has the advantage, that one can define a URI prefix which is then automatically applied to all question elements of the copy, making it clear that it is a new catalog of its own.

3.3.4 What happens after this?

When a new catalog has been developed, it helps other RDMO users if it is published in the RDMO-Catalog GitHub repository so that it can be reused. A good way to do this is

[6] `https://rdmo.aip.de/`

to create two pull requests to the dev branch of the repository, one for the domain XML file and the second for the remaining files (including questions, options and conditions).

In preparation for publication, all attributes that have been newly defined for the catalog should be added to the RDMO Domain XML file which is located in the rdmo-catalog repository in the folder rdmorganiser/domain. The new attributes should be inserted into the existing hierarchy of attributes at the positions that seem most appropriate. As prefix, *https://rdmorganiser.github.io* should be used and in the comment section of each attribute the text of the associated question should be inserted. The RDMO Content Group will then review the changes with the discussion taking place in the form of public comments within the pull request. Since the editorial workflow for the inclusion of newly defined options and conditions is still under development, these files can be provided as separate files for the "shared" folder along with the question XML file in the second pull request.

Finally the entire catalog is checked to see if it works technically. If this is the case, both pull requests are merged into the dev and later into the master branch of the repository.

3.4 Where are we now?

Currently, a generic NFDI4Ing client usable for all researchers with DFN-AAI access credentials is being created. NFDI4Ing started a coordination process within the NFDI on DMPs and tools by hosting a workshop on these topics in march 2021 jointly with the RDMO Working Group [23]. A respective working group on the NFDI level will be formed soon. For now, ULB Darmstadt runs RDMO for a network of Hessian universities [21] and for RWTH Aachen.

In 2021 it is planned to conduct a needs assessment for the further RDMO clients within NFDI4Ing and Germany-wide in order to establish further clients. The international perspective (RDA, EOSC) is monitored.

4 Conclusions

The creation of a DMP for a research project using RDMO software has been analyzed. Two scenarios of working with RDMO are considered: the case of an existing DMP and its transformation into a catalog of RDMO questions and the case of creating a catalog of RDMO questions from scratch. The transformation of the existing DMP is illustrated by the example of the Use Case within the archetype Golo of the Consortium NFDI4Ing.

In general, it can be noted that the application of RDMO simplifies the creation of a DMP if one is faced with the necessity to build a plan for the first time. In the case of transforming an existing DMP into an RDMO question catalog template, the advantage of such a transformation is the availability of ready-made plans and the ability to borrow existing ready-made questions. The disadvantages of the transformation process

includes the need for a long search for existing attributes. When searching for appropriate attributes, the general recommendation should be to use the attributes with the official RDMO prefix whenever possible. Since only one attribute can be assigned to a single question during the transformation, complex questions should be avoided and formulated in such a way that they contain, if possible, one key word that can be used as the basis for finding the attribute. Among the strengths of using RDMO is the ability to save catalog of questions in universal formats, which make it possible to link documents semantically and automatically transform catalog of questions and plans into, for example, a wiki-based software such as Semantic MediaWiki.

Further advantages of RDMO sharing in NFDI4Ing, and the perspective in NFDI, are that every researcher can work on a DMP together with every other researcher via the common database. DMP templates can be easily reused, can be adapted by each client for his institution or cooperative project, are jointly developed and can be standardized with regard to completeness, scope, validity, etc. Last but not least, the approach results in very lean IT operations.

Acknowledgements

The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.

Bibliography

[1] Sandfeld, S. et al. (2018) Strategiepapier Digitale Transformation in der Materialwissenschaft und Werkstofftechnik. https://edocs.tib.eu/files/e01fn18/1028913559.pdf (Accessed: 16.05.2021).

[2] Mozgova, I. et al. (2020) Research data management system for a large collaborative project. Proceedings of NordDesign 2020. http://doi.org/10.35199/NORDDESIGN2020.48.

[3] Wilkinson, M.D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3; 160018. https://doi.org/10.1038/sdata.2016.18.

[4] Bierwirth M. et al. Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung (2020) http://doi.org/10.5281/zenodo.3895209.

[5] Schmitt R.H. et al. NFDI4Ing - the National Research Data Infrastructure for Engineering Sciences (2020) http://doi.org/10.5281/zenodo.4015201.

[6] Jagusch, G., Preuß, N. (2019) Umfragedaten zu "NFDI4Ing - Rückmeldung aus den Forschungscommunities". http://doi.org/10.25534/TUDATALIB-104.

[7] Graessler, I., Hentze, J., Bruckmann, T. (2018) V-Models for Interdisciplinary Systems Engineering. In: Proceedings of the DESIGN 2018 15th International Design Conference, S. 747-756. https://doi.org/10.21278/idc.2018.0333.

[8] Knöchelmann, M. et al. (2019) Methodische Entwicklung eines opto-mechatronischen Systems am Beispiel eines hochadaptiven Fahrzeugscheinwerfers. In: Tagungsband der VDI Fachtagung Mechatronik 2019. Paderborn. https://doi.org/10.15488/4683.

[9] Klar, J. (2020) RDMO für Fortgeschrittene. http://doi.org/10.5281/zenodo.3930141.

[10] Heger, M. (2016) Datenmodellierung für Forschungsdatenmanagementpläne. Masterarbeit an der Fachhochschule Potsdam am Fachbereich 5 Informationswissenschaften https://rdmorganiser.github.io/docs/Heger_MA.pdf.

[11] https://www.eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html (Accessed: 15.08.2021).

[12] http://www.foaf-project.org/(Accessed:15.08.2021).

[13] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (Accessed:15.08.2021).)

[14] https://github.com/rdmorganiser/rdmo-catalog/blob/dev/rdmorganiser/domain/rdmo.xml (Accessed: 15.08.2021).

[15] https://rdmo.readthedocs.io/en/latest/management/index.html?highlight=prefix#management(Accessed:15.08.2021).

[16] https://www.forschungsdaten.org/index.php/RDMO(Accessed:15.08.2021).

[17] https://rdmorganiser.github.io/en/rdmo_arge/ (Accessed: 16.05.2021).

[18] https://rdmorganiser.github.io/en/cooperations/ (Accessed: 16.05.2021).

[19] https://github.com/rdmorganiser/rdmo-catalog (Accessed: 16.05.2021).

[20] https://www.forschungsdaten.org/index.php/Katalog_erstellen (Accessed: 16.05.2021).

[21] Hessische Forschungsdaten-Infrastrukturen (HeFDI), https://www.hefdi.de/ (Accessed: 16.05.2021).

[22] https://rdmorganiser.github.io/terms/ (Accessed: 16.05.2021).

[23] https://rdmorganiser.github.io/docs/nfdiws/wsreport/ (in Germany only, Accessed: 16.05.2021).

Die Veröffentlichung von Standardisierten Daten aus der Stadtklimaforschung

Anette Ganske[1], Vivien Voss[2], Amandine Kaiser[3], Angelika Heil[3] und Andrea Lammert[3]

[1]Technische Informationsbibliothek (TIB)
[2]Universität Hamburg, Meteorologisches Institut, CEN
[3]Deutsches Klimarechenzentrum (DKRZ)

Eine Grundvoraussetzung für die Wiederverwendbarkeit von wissenschaftlichen Daten ist die Veröffentlichung der Daten in einer Art und Weise, die garantiert, dass sie auffindbar, zugänglich, untereinander vergleichbar und weiterverwendbar sind. Dies entspricht der Einhaltung der FAIR-Prinzipien.

Innerhalb des Projekts AtMoDat wurde ein Konzept erstellt, wie atmosphärische Modelldaten idealerweise für die Nachnutzbarkeit beschrieben werden sollten. Voraussetzungen hierfür sind die Nutzung des Datenformats NetCDF, die Anwendung des international etablierten Climate and Forecast (CF) Metadatenstandards und die Publikation der Daten mit einem DataCite DOI. Basierend auf dem neu erstellten ATMODAT Standard wurde ein Tool zur Überprüfung der Konformität von Daten und Metadaten, der *atmodat data checker*, entwickelt, welcher über GitHub mit freier Lizenz verfügbar ist.

Neben der klassischen Klimamodellierung gibt es viele Bereiche der atmosphärischen Modellierung, welche noch nicht im CF Metadatenstandard enthalten sind, wie z.B. die mikroskalige Stadtklimamodellierung. Durch die Nutzung des ATMODAT Standards lassen sich nun auch in diesen Bereichen standardisierte Datensätze erstellen. Dies wird exemplarisch anhand des Stadtklimamodells MITRAS vorgestellt.

Nach der Publikation der geprüften Daten können diese mit EASYDAB (Earth System Data Branding) hervorgehoben werden. EASYDAB ist eine neu entwickelte Qualitätskennzeichnung, die potentiellen Datennutzern anzeigt, dass es sich um einen geprüften Datensatz handelt, der einfach wiederzuverwerten ist. Dies wird durch das EASYDAB-Logo auf der Landing Page des Datensatzes angezeigt. Die innerhalb des AtMoDat-Projekts entwickelten Verfahren der Datenstandardisierung und Qualitätskennzeichnung können einfach auf Daten anderen Fachrichtungen übertragen werden und dort zu einer verbesserten Nachnutzbarkeit der Daten beitragen.

1 Einleitung

Klimamodelle sind komplexe Computerprogramme, welche Abschätzungen darüber liefern, wie sich das Klima unter bestimmten Annahmen verändern wird. Der Austausch

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13723 (CC BY-SA 4.0)

von Klimamodelldaten ist weit über die Klimaforschungsgemeinschaft von großer Bedeutung. Klimamodelle liefern sehr große Datenmengen, welche so gespeichert und zugänglich gemacht werden müssen, dass eine Nachnutzung möglichst effizient erfolgen kann. Im Rahmen des Coupled Model Intercomparison Project (CMIP) [1] arbeiten weltweit hunderte Klimaforschungseinrichtungen daran, koordinierte Simulationen mit globalen Erdsystem- bzw. Klimamodellen durchzuführen, die neuesten Modellergebnisse untereinander auszutauschen, zu vergleichen und zu analysieren. Die CMIP-Modellexperimente dienen als wissenschaftliche Basis für die Sachstandsberichte des Weltklimarats IPCC[1] (Intergovernmental Panel on Climate Change). Die IPCC-Berichte tragen den aktuelle Kenntnisstand über den menschengemachten Klimawandel zusammen und bieten Grundlagen für wissensbasierte klimapolitische Entscheidungen.

Um den Datenaustausch und die vergleichende Datenauswertung zu erleichtern, wurden für CMIP genaue Vorgaben für das Dateiformat, die Metadaten und die Datendokumentation gemacht und angewendet [2]. Auch wurde eine einheitliche Namensgebung der Variablen und Dateien vereinbart. Das CMIP Standard-Dateiformat ist NetCDF[2] (Net-work Common Data Form), ein binäres, systemunabhängiges Format, das sich besonders für gegitterte Daten eignet, und das – da es Daten zusammen mit den Metadaten abspeichern kann – selbstbeschreibend ist. Für CMIP wurde festgelegt, dass die Dateien die international vereinbarten *NetCDF Climate and Forecast Metadata Conventions*[3] (im Folgenden CF Conventions genannt) erfüllen sollen. Ein zentrales Element der CF Conventions ist die CF-Standardnamentabelle. Standardnamen sind genau definierte Bezeichner für Größen; sie werden Variablen über das *standard_name* Attribut zugeordnet und machen sie so eindeutig und maschinenlesbar. Die CF-Standardnamentabelle schreibt auch für jeden Bezeichner eine physikalische Einheit vor und sie enthält eine kurze, menschenlesbare Erklärung, wie die einem Standardnamen zugeordnete Größe definiert ist. Derzeit umfasst die Tabelle etwa 4500 Einträge aus verschiedenen geowissenschaftlichen Bereichen. Zusätzlich definieren die CF Conventions wie räumliche und zeitliche Achsen von abgespeicherten Daten zu hinterlegen sind. Weiterhin wurden innerhalb von CMIP Richtlinien erarbeitet, wie die Daten technisch und inhaltlich geprüft werden und in welcher Art und Weise die Ergebnisse der Qualitätsprüfung für den Datennutzer sichtbar gemacht werden.

Im Projekt AtMoDat[4] (Atmosphärische Modelldaten) wurde untersucht, ob die in CMIP angewandten Standards systematisch auf weitere Bereiche der Meteorologie und Klimaforschung übertragen werden können. So gibt es eine Vielzahl kleinerer, CMIP unterstützender Modellvergleichsprojekte (MIPs), wie z.B. die Initiative Aerosols, Clouds, Precipitation and Climate (ACPC)[5]. Eine vollständige Anpassung des CMIP Standards auf solch kleinere MIPs wäre sehr aufwendig und zu kostspielig, so dass die Entwicklung eines reduzierten Standards als sinnvoll erachtet wurde. Weiterhin gibt es den Bereich der Stadtklimaforschung. Hier gibt es bisher keinen etablierten Datenstandard. Daten aus

[1] `https://www.ipcc.ch/`
[2] `https://doi.org/10.5065/D6H70CW6`
[3] `http://cfconventions.org`
[4] `https://www.atmodat.de/`
[5] `http://www.acpcinitiative.org/`

der Stadtklimaforschung sind im Vergleich zu regionalen oder globalen Klimamodelldaten sehr hoch aufgelöst und haben oft eine andere Raumgitterstrukur. Zudem beinhalten sie Variablen, die spezifisch für die Stadtklimaforschung sind, wie z.B. Wärmeflüsse an Gebäudewänden. Um Daten aus der Stadtklimaforschung nachnutzbar veröffentlichen zu können, müssen fachspezifische Standards neu entwickelt werden.

Innerhalb des AtMoDat Projekts wurden Kurationskriterien [3] formuliert, auf deren Basis die Vollständigkeit der Datensätze, die Kohärenz und Konsistenz von Daten- und Metadaten sowie die portalübergreifende Auffind- und Nutzbarkeit der Daten im Zuge des Archivierens und Teilens über Repositorien sichergestellt wird. Die im AtMoDat Projekt entwickelten Standardisierungen und Kurationskriterien sind auf andere Bereiche der Klimaforschung übertragbar; sie können die Reproduzierbarkeit und Überprüfbarkeit deutlich erhöhen und auch die interdisziplinäre Nachnutzung von Klimamodelldaten unterstützen. Diese Kriterien werden im ATMODAT Standard [4] zusammen gefasst. In diesem Standard folgen wir den FAIR-Datenprinzipien[6] [21], welche fordern, dass Daten so veröffentlicht werden sollten, dass sie auffindbar, zugänglich, interoperabel und wiederverwendbar sind. Zudem wurde im Rahmen des AtMoDat Projekts das Earth System Data Branding[7] (EASYDAB) entwickelt, um qualitativ hochwertige Datensätze deutlicher aus der Menge anderer Daten hervorzuheben.

Der ATMODAT Standard wird in Abschnitt 2 näher erläutert. In Abschnitt 3 werden die Resultate der Anwendung des Standards auf die Daten der Stadtklimaforschung beschrieben und in Abschnitt 4 wird EASYDAB detaillierter erklärt.

2 Der ATMODAT Standard

Der ATMODAT Standard ist eine Qualitätsrichtlinie für die FAIRe Veröffentlichung von atmosphärischen Modelldaten mit offenen Lizenzen. Er setzt eine Publikation der Daten mit einem DataCite DOI in einem Repositorium[8] voraus. Zudem muss der DOI, wie von DataCite[9] empfohlen, stets mit einer sogenannten Landing Page verknüpft sein. Der DOI kann für einzelne Datensätze vergeben werden, d.h. für je eine Datei mit wissenschaftlichen Daten und ihre Metadaten. Alternativ kann er auch einer ganzen Datensatz-Kollektion gemeinsam zugewiesen werden, die aus mehreren Datensätzen besteht, deren Daten z.B. mit einer einzigen Modellsimulation berechnet wurden.

Um die FAIRness der veröffentlichten Daten zu ermöglichen, werden im ATMODAT Standard Vorgaben für die Dateien, die Metadaten und die Landing Pages gemacht. Zudem gibt der Standard vor, dass alle veröffentlichten Daten als NetCDF Dateien vorliegen.

[6] FAIR=Findable, Accessible, Interoperable, Reusable. https://doi.org/10.1038/sdata.2016.18
[7] https://www.easydab.de
[8] https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/repositorien/
[9] https://datacite.org/

2.1 Vorgaben für die Landing Page

Wird eine DOI in einem konventionellen Webbrowser aufgerufen, führt diese zu einer menschenlesbaren HTML-Internetseite, die immer auch einen maschinenlesbaren Text, den Seitenquelltext, enthält. Diese Seite wird Landing Page genannt. Die Landing Page wird vom Repositorium, das die Daten archiviert hat, bereitgestellt und enthält grundlegende Metadaten. Die Landing Page kann in Unterseiten gegliedert werden, um sie übersichtlicher zu machen.

Die Vorgaben von DataCite für die Landing Page sind in Best Practices für DOI Landing Pages[10] beschrieben. Auf der Landing Page muss im menschenlesbaren Format immer eine zitierbare vollständige bibliographische Angabe (incl. des DOI) über dem Datensatz selbst stehen, so dass er eindeutig vom Menschen identifiziert werden kann. Zudem soll die DOI im maschinenlesbaren Teil der Landing Page so gekennzeichnet werden, dass Suchmaschinen sie finden können.

Weiterhin muss eine Landing Page immer Informationen darüber enthalten, wie man auf die Daten zugreifen kann. Falls der Datensatz selbst nicht mehr existiert, muss dies auf der Landing Page vermerkt werden (Tombstone Page).

2.2 Vorgaben für die Metadaten

Metadaten müssen die Daten so beschreiben, dass potentielle Datennutzer in der Lage sind zu entscheiden, ob diese Daten für ihre Anwendung von Nutzen sind. Dafür müssen auch der Prozess der Datenerzeugung und die verwendeten Eingangsdaten dokumentiert werden.

Für die Datenveröffentlichung mit einem DataCite DOI werden in den Dateien, beim DOI und auf der Landing Page Metadaten benötigt: Alle notwendigen Informationen können entweder direkt in die Metadaten geschrieben werden oder es kann mit Links auf externe Dokumente verwiesen werden - vorzugsweise über persistente Identifikatoren (PIDs). Ein solches externes Dokument könnte z.B. die Dokumentation des numerischen Modells sein, das zur Berechnung der Daten verwendet wurde. Es wird dringend empfohlen, dass die Metadaten sowohl in maschinenlesbarer als auch maschineninterpretierbarer Form hinterlegt sind. Dies gewährleistet, dass aus den Metadaten automatisierte Listen, z.B. für Auswertungen von Institutionen erstellt werden können. Externe Dokumente selbst, zumindest aber deren Metadaten, sollten ebenfalls maschinenlesbar sein.

Der ATMODAT Standard gibt die folgenden Prinzipien vor, um die Maschinenlesbarkeit und –interpretierbarkeit und damit die FAIRness aller Metadaten zu ermöglichen:

- Alle Angaben zu Personen und Institutionen sollten möglichst immer mit einem PID ergänzt werden, d.h. einem ORCID für Personen oder einem ROR für Institutionen (falls vorhanden). Vorschläge für geeignete PIDs findet man in [6].

[10] `https://support.datacite.org/docs/landing-pages`

- Alle Links auf Dokumente, Homepages usw. sollten persistent angegeben werden, z.B. Dokumente mit einem DOI und Homepages vorzugsweise mit einer URN.
- Alle zeitlichen Informationen müssen in standardisierter Form angegeben werden, z.B. nach ISO 8601 [7] oder ISO 19108 [8].
- Das räumliche Referenzsystem muss immer angegeben werden, z.B. WGS84 [9].
- Schlagwörter, Orte und Arbeitsgebiete sollten aus kontrollierten Vokabularen (CVs) entnommen werden, z.B. geografische Namen aus geonames[11].

2.2.1 Datei-Metadaten

In den CF Conventions werden Spezifikationen für Namen, Einheiten und andere Parameter vorgegeben, um die Dateiinhalte zu vereinheitlichen und sie automatisch verarbeitbar zu machen. Deshalb schreibt der ATMODAT Standard vor, dass die CF Conventions (ab Version 1.4) für die Beschreibung aller Daten verwendet werden. Jedoch werden in den CF Conventions nicht für alle Variablen von atmosphärischen Modellen die sogenannten *standard_names* vorgegeben, wie z.B. für spezielle Variablen von Stadtklimamodellen (siehe Kapitel 3). In diesem Fall kann den Variablen ein eigener *long_name* zuwiesen werden, welcher die Vorschriften der CF Conventions für die Bildung von *long_names* befolgt.

Neben der Einhaltung der CF Conventions fordert der ATMODAT Standard zusätzlich, dass die NetCDF-Dateien eine Beschreibung der Zeit-, Koordinaten- und vertikalen Achsen sowie bestimmte globale Attribute enthalten (siehe [4], Tabelle 11).

2.2.2 DOI-Metadaten

DOI-Metadaten werden für den DOI an DataCite übermittelt und beschreiben die veröffentlichten Daten. Falls ein DOI für mehrere Datensätze (Datensatz-Kollektion) zusammen vergeben wurde, beschreiben die DOI-Metadaten die Datensatz-Kollektion. Die DOI-Metadaten werden in die Metadatenfelder des DataCite Metadatenschemas [10] eingetragen.

Neben den allgemeinen Grundsätzen für alle Metadaten gibt es im ATMODAT Standard zusätzliche Empfehlungen für diese Metadaten:

- Die DataCite-Metadatenfelder *Contributor* und *Creator* sind Pflichtfelder, welche ausgefüllt werden müssen, damit Zusammenfassungen über die Publikation eines einzelnen Forschers, aller Forscher einer Institution oder aller Publikationen innerhalb eines Projekts automatisch zusammengestellt werden können. Aus den gleichen Gründen wird dringend empfohlen, auch die Förderung unter *FundingReference* anzugeben (falls vorhanden).

[11] `https://www.geonames.org`

- Alle Zeitangaben über die Erstellung oder Veröffentlichung des Datensatzes/der Datensatz-Kollektion sind wichtig für den Datennutzer. Sie werden über das Metadaten-Feld *Date* und den zugehörigen Unterfeldern *dateType* erfasst. Eine Ausnahme bildet das Veröffentlichungsjahr, welches in dem Feld *PublicationYear* angegeben wird.

Damit die Daten selbsterklärend beschrieben sind, wird dringend empfohlen, alle für die entsprechende Fachdisziplin sinnvollen Metadatenfelder aus dem DataCite Metadatenschema zu verwenden.

2.2.3 Landing Page-Metadaten

Der ATMODAT-Standard schreibt vor, dass alle Metadatenfelder, die für den DOI angegeben werden, auch auf der Landing Page aufgelistet werden. Dabei müssen die Bezeichnungen der einzelnen Metadatenfelder nicht übernommen werden. Zusätzlich sollen bei Datensatz-Kollektionen oder bei Datensätzen mit mehreren Variablen auch die Beschreibungen der einzelnen Datensätze bzw. Variablen auf der Landing Page stehen. Falls die Datensätze/Datensatz-Kollektionen vor der Veröffentlichung geprüft wurden, sollte das Ergebnis der Prüfung ebenfalls auf der Landing Page vermerkt werden. Diese Prüfungen können ganz unterschiedlich definiert sein: so können z.B. Messdaten auf fehlerhafte Einträge geprüft werden. Es können aber auch Metadaten auf ihre Vollständigkeit oder auf Einhaltung einer Qualitätsrichtlinie geprüft werden.

2.3 Überprüfung der Einhaltung der Vorgaben

Der ATMODAT Standard richtet sich an Repositorien (Bereich Datenkuration), aber auch an die Wissenschaft (Bereich Datenproduktion) und enthält Checklisten, mit denen direkt nach der Datenproduktion und während der Datenkuration die Einhaltung des Standards geprüft werden kann. Zur Erleichterung und vor allem zur Automatisierung dieses Prüfprozesses auch für große Datenmengen wurde der *atmodat data checker* entwickelt. Der Checker ist ein modular aufgebautes Python-Programmpaket, welches mit freien Lizenzen auf Github[12] veröffentlicht wurde. Somit ist der *atmodat data checker* für alle Interessierten nutzbar und hilfreich bei der Produktion und Publikation standardisierter Daten. Ein Beispiel seiner Anwendung wird in Kapitel 3.2 gezeigt. Ein entsprechender Checker für die Prüfung der DOI-Metadaten ist in Arbeit.

3 Anwendung des Standards auf Daten der Stadtklimaforschung

Klima- und Atmosphärenmodelle, die bei CMIP verwendet werden und an die CF Conventions angepasst sind, werden über sehr große Regionen angewendet wie z.B. die ganze

[12] `https://github.com/AtMoDat/atmodat_data_checker`

Erde oder ganz Europa. Die horizontale Auflösung der Modellgitter liegt üblicherweise zwischen 50 und 200 km. Die Modelle rechnen typischerweise in Zeitschritten von Minuten bis Stunden und die Simulationen umfassen Zeiträume von Jahren bis Jahrtausenden. Diese Eigenschaften sind für die Simulation kleinskaliger Prozesse, wie z.B. in der Stadtklimamodellierung, nicht geeignet, denn solch grobe Gitterzellen und Zeitschritte können die städtische Struktur nicht darstellen und kurzlebige Prozesse nicht auflösen. Im Rahmen von AtMoDat Projekts soll eine Erweiterung des CF Standards an die speziellen Anforderungen der mikroskaligen Stadtklimamodellierung angeregt werden.

3.1 Besonderheiten der Stadtklimamodellierung

Die Stadtklimaforschung beinhaltet die Untersuchung von Phänomenen und Prozessen, die im urbanen Raum eine Rolle spielen. Themen in der Stadtklimaforschung sind beispielsweise Effekte des Klimawandels im urbanen Raum [12] oder die Verbesserung der Lebensqualität innerhalb von Städten durch Studien zur Ventilation, Ausbreitung von Luftschadstoffen (z.B. [13]) oder Temperaturverteilungen innerhalb Straßenschluchten [14]. Die Untersuchung solcher Effekte erfolgt zum Einen über die Erhebung und Auswertung von Messdaten (z.B. [12] und [13]), zum Anderen über die Verwendung von Modellen, die solche Prozesse simulieren können.

Modelle, die atmosphärischen Prozesse im städtischen Raum simulieren sollen, müssen Strukturen und Besonderheiten der Stadt darstellen können. Verwendet werden hierbei sogenannte mikroskalige Modelle, wie zum Beispiel ENVI-MET[13], MISKAM[14], MITRAS [15] oder PALM-4U [16]. Die dabei verwendeten Modellgebiete decken Bereiche in der Größenordnung weniger Quadratkilometer ab. Meist umfassen die Gebiete typische Stadtstrukturen, wie zum Beispiel einzelne Stadtteile oder einzelne Straßenzüge. Die Gitterweiten betragen wenige Metern und somit können Hindernisse wie Gebäude oder Bäume dargestellt werden. Die zeitliche Auflösung der Simulationen beträgt wenige Sekunden, so dass kurzlebige Prozesse gut aufgelöst werden können. Insgesamt umfasst eine Simulation meist nur Stunden oder wenige Tage. Beispielsweise lassen sich so Zirkulationen in einzelnen Straßenschluchten oder die Ausbreitung von Schadstoffen innerhalb der Stadt simulieren.

Zur Veranschaulichung der hohen räumlichen Auflösung der mikroskaligen Stadtklimamodelle zeigen wir hier eine Simulation mit MITRAS für die Hamburger Innenstadt (Abbildung 1). Die damit berechneten Windgeschwindigkeiten werden in Abbildung 2 gezeigt. Dabei wurde eine horizontale Gitterweite von 2.5 m verwendet. In der Simulation wurde in rund 1 km Höhe eine südwestliche Anströmung mit einer Windgeschwindigkeit von 3 m/s vorgegeben. Abbildung 2 zeigt die simulierten Windgeschwindigkeiten in vier verschiedenen Höhenschichten, von bodennah (2a) bis zur maximalen Höhe der Gebäude im Gebiet (2d). Es lässt sich deutlich erkennen, wie Gebäude das Windfeld in allen drei Raumrichtungen (horizontale Ebene und in der Vertikalen) beeinflussen. In Bodennähe

[13] `http://www.envi-met.net/documents/papers/geosim2001.pdf`
[14] `https://www.lohmeyer.de/site/assets/files/4559/manual_miskam_63.pdf`

blockiert eine enge Bebauung die Strömung, was sich in niedrigen Windgeschwindigkeiten äußert. Auf Plätzen und breiteren Straßen treten höhere Windgeschwindigkeiten auf. Mit zunehmender Höhe blockieren immer weniger Gebäude den Wind; in 72,5 m Höhe (2d) haben nur noch die Kirchturmspitzen und Hochhäuser direkten Einfluss auf das Windfeld, was sich in der niedrigeren Windgeschwindigkeit im Nachlauf der Strömung widerspiegelt.

Für spezifische Eigenschaften der Stadt, wie die Hindernisse und dazugehörige Variablen, z.B. Wandtemperaturen, Strahlungsflüsse an Gebäudeoberflächen oder Niederschlag auf Dächern, werden in den bisherigen CF Conventions noch keine Vorgaben zur Standardisierung gemacht. Daher beschäftigt sich ein Teil des AtMoDat Projektes damit, typische Variablen aus der Stadtklimaforschung zu identifizieren und ihre Aufnahme in die CF Conventions zu beantragen.

Abbildung 1: Ausschnitt der Hamburger Innenstadt, der für die Simulation mit dem Modells MITRAS verwendet wird. Karte: www.openstreetmap.org

3.2 Anwendung des atmodat data checker

Erste Tests zur Anwendung und Umsetzung des ATMODAT Standards wurden mit Modellausgaben von MITRAS durchgeführt. MITRAS schreibt die Simulationsergebnisse in binäre Dateien, die mit Hilfe eines Post-Processors in NetCDF-Dateien umgewandelt werden. Bisher fehlten dem Post-Processor die Attribute *long_name* und *standard_name*, die die eindeutige Beschreibung und Zuordnung der Variablen definieren. Beide Attribute wurden in den Post-Processor eingebaut. Variablen, die noch nicht in den CF Conventions geführt werden, erhalten zunächst nur das Attribut *long_name*. Weiterhin wurden zusätzliche Metadaten nach Empfehlungen aus dem ATMODAT Standard zur besseren Nachnutzbarkeit hinzugefügt.

Die Datensätze wurden mit Hilfe des *atmodat data checkers* (siehe Abschnitt 2.3) auf die Konformität mit dem ATMODAT-Standard überprüft. Der Checker überprüft neben der

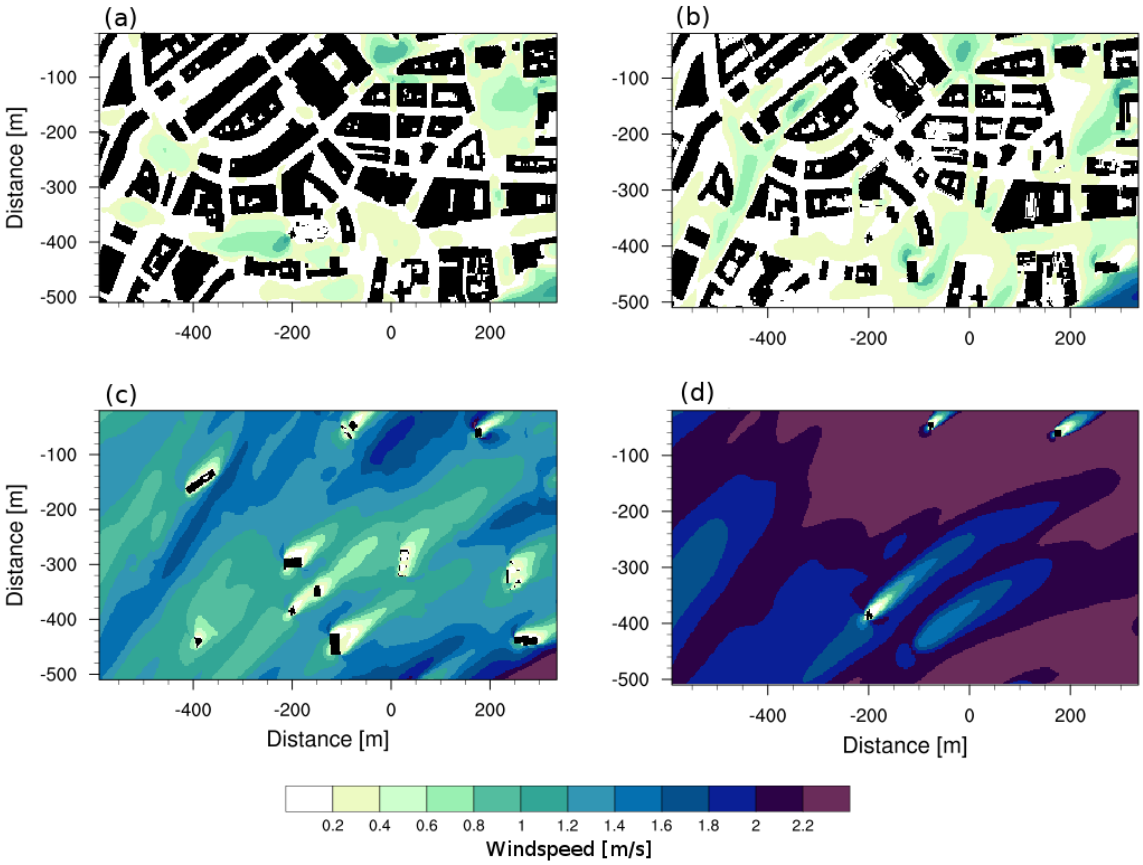


Abbildung 2: Darstellung des Windes in der Hamburger Innenstadt am 21.06.2000 um 4:30 Uhr morgens, simuliert mit dem Modell MITRAS. Farbige Flächen zeigen verschiedene Windgeschwindigkeiten, die schwarzen Flächen beschreiben die Gebäude in dem Modellgebiet. Die Abbildungen zeigen die Windgeschwindigkeiten in (a) 2,5 m, (b) 22,5 m, (c) 48,5 m und (d) 72,5 m Höhe.

Einhaltung der CF Conventions die Dateien auch auf lückenhafte Metadatenangaben und fehlerhafte Dateien. Werden Fehler entdeckt, gibt der Checker eine Meldung aus, in der zwischen Fehlern und Warnungen differenziert wird. Zudem enthält der Output klare Hinweise zur Korrektur der Fehler. Bei den Metadaten unterscheidet der Checker zwischen zwingend erforderlichen (mandatory), empfohlenen (recommended) und sonstigen optionalen Metadatenangaben. Der Output des *atmodat data checkers* ist maschinenlesbar, was z.B. eine automatische Auswertung bei größeren Datenmengen ermöglicht.

Um den *atmodat data checker* anwenden zu können, muss er lokal auf dem Rechner installiert werden. Anweisungen zur Installation und zur Anwendung des Checkers sind auf Github[15] hinterlegt. Die Software lässt sich problemlos auf verschiedenen Betriebssystemen installieren und ist einfach anzuwenden. Im Test mit MITRAS-Ausgabedateien konnte der *atmodat data checker* einwandfrei alle fehlerhaften NetCDF-Dateien, die zum Beispiel keinen *standard_name* oder *long_name* enthielten, identifizieren. Des weiteren

[15] `https://github.com/AtMoDat/atmodat_data_checker`

gibt der *atmodat data checker* hilfreiche Anweisungen, wie die Metadaten der Dateien um im ATMODAT Standard empfohlene Attribute ergänzt werden können, um so die Nachnutzung der Daten zu erleichtern.

4 EASYDAB

Mit dem Earth System Data Branding (EASYDAB) können spezielle Datensätze aus der Erdsystemforschung, die mit einem DataCite DOI veröffentlicht wurden, hervorgehoben werden: Das EASYDAB Logo auf der Landing Page zeigt an, dass die Datensätze eine offene Lizenz haben, den FAIR Data Principles genügen und vom zuständigen Repositorium auf Einhaltung einer Qualitätsrichtlinie geprüft wurden. So können z.B. Datensätze, die die Qualitätsanforderungen des ATMODAT Standards einhalten, mit EASYDAB publiziert werden. Alternativ können Repositorien für eine EASYDAB Veröffentlichung auch auf eigene Qualitätsrichtlinien zurückgreifen, falls mit diesen eine vergleichbar hohe Qualität bei den veröffentlichten Daten erreicht wird, wie das im ATMODAT Standard definiert wird. Die Bedingungen für die Qualitätsrichtlinien findet man unter `https://www.easydab.de/about-easydab/easydab-guideline`.

Das EASYDAB Logo ist geschützt. Um es verwenden zu dürfen, müssen Repositorien einen Vertrag mit der Technischen Informations Bibliothek (TIB)[16] abschließen. In diesem verpflichten sich die jeweiligen Repositorien, dass sie das EASYDAB Logo nur auf den Landing Pages von den Datensätzen/Datensatz-Kollektionen zeigen, bei denen die EASYDAB Richtlinien eingehalten werden.

Mithilfe des EASYDAB Logos können Repositorien anzeigen, dass sie Datensätze unter Berücksichtigung der FAIRness Prinzipien sorgfältig kuratieren und nachnutzbar machen. Forschende können gut beschriebene Datensätze leichter finden, evaluieren und für sie relevante Daten nutzen. In Zusammenarbeit mit dem World Data Center for Climate[17] (WDCC) werden erste Datensätze aus der Stadtklimafoschungvon und von einem kleineren Modellvergleichsprojekt (MIP) zur Veröffentlichung mit EASYDAB vorbereitet.

5 Fazit und Ausblick

Klimamodelldaten sind eine unverzichtbare Grundlage für Klimaforscher und politische Entscheidungsträger. Klimamodelldaten geben zum Beispiel Aufschluss darüber, wie sich atmosphärische Prozesse auf globaler oder regionaler Ebene mit einem Klimawandel verändern; sie geben auch Hinweise, welche Anspassungsmaßnahmen zum Schutz urbaner Gebiete sinnvoll wären.

[16] siehe `www.easydab.de`
[17] `https://www.dkrz.de/up/systems/wdcc`

Die Klima- und Atmosphärenforschung hat umfangreiche Standardisierungen bezüglich der Dateiformate, Metadatengestaltung und Dokumentation etabliert, um die Daten von globalen und regionalen Klimamodellberechnungen zu vereinheitlichen und automatisiert auswertbar zu machen. Im Rahmen des AtMoDat Projekts werden diese Standardisierungen systematisch auf Daten hochauflösender Modelle angepasst. Der hierbei entwickelte ATMODAT Standard wird, wie in diesem Beitrag beschrieben, unter anderem an Daten des Stadtklimamodells MITRAS erprobt.

Nach erfolgreicher Anwendung erfolgt die exemplarische Veröffentlichung erster Datensätze inklusive DataCite DOI im Langzeitarchiv des WDCC. Dies hat besonders im Bereich der Stadtklimaforschung Pionier- und Vorbildcharakter, da im WDCC bisher keine standardisierten Daten aus dieser Disziplin gespeichert werden. Die entwickelten Standardisierungsverfahren lassen sich leicht auf weitere Disziplinen übertragen und, wenn erforderlich, erweitern. Sie stellen somit einen wesentlichen Beitrag dar, die Kompatibilität von Datenstandards sowohl innerhalb einzelner Fachrichtungen als auch disziplinübergreifend zu verbessern und hierdurch den intra- und interdisziplinären Datenaustausch erheblich zu erleichtern.

Um qualitativ hochwertige Datensätze wie solche, die dem ATMODAT Standard entsprechen, in Repositorien deutlich erkennbar hervorheben zu können, wurde EASYDAB, das Earth System Data Branding entwickelt. Das geschützten EASYDAB Logo hilft Nutzern, diese Datensätze schnell zu erkennen und auszuwählen. Der einfache Zugang zu standardisierten, qualitätsgeprüften Forschungsdaten ermöglicht nicht nur eine effizientere Nachnutzung. Er führt auch zu einem erhöhtem Vertrauen der Nutzer in das Repositorium, das diese Daten zur Verfügung stellt, und so wird mit dem EASYDAB Branding auch die Wertigkeit des DataCite DOIs gestärkt. Das längerfristige Ziel ist, dass qualitätsgesicherte Datenpublikationen, wegen ihrer hohen Vorteile, zunehmend von Nutzern sowie Förderorganisationen gewürdigt und eingefordert werden.

Acknowledgements

Das AtMoDat Projekt wurde im Rahmen vom “Forschungsvorhaben zur Entwicklung und Erprobung von Kurationskriterien und Qualitätsstandards von Forschungsdaten” finanziert vom Deutschen Bundesministerium für Bildung und Forschung (BMBF; FKZ: 16QK02A, 16QK02B, 16QK02C, 16QK02D). Wir danken dem DataCite Team, der CF Conventions Community und anderen AtMoDat Partnern für fruchtbare Diskussionen.

Literaturverzeichnis

[1] Eyring, V. , S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer and K. E. Taylor. “Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6)

experimental design and organization." Geoscientific Model Development, Band 9, Heft 5 (2016): 1937-1958. https://doi.org/10.5194/gmd-9-1937-20161937-1958

[2] Juckes, M., K.E. Taylor, P.J. Durack, B. Lawrence, M.S. Mizielinski, A. Pamment, J.-Y. Peterschmitt, M. Rixen and S. Sénési. "The CMIP6 Data Request (DREQ, version 01.00.31)." Geosci. Model Dev. Band 13, Heft 1 (2020): 201–224. https://doi.org/10.5194/gmd-13-201-2020

[3] Ganske, A., D. Heydebreck, H. Höck, A. Kraft, J. Quaas and A. Kaiser. "A short guide to increase FAIRness of atmospheric model data." Meteorologische Zeitschrift Band 29, Heft 6 (2020): 483-491.http://dx.doi.org/10.1127/metz/2020/1042

[4] Ganske, A., A. Kraft, A. Kaiser, D. Heydebreck, A. Lammert, H. Höck, H. Thiemann, V. Voss, D. Grawe, B. Leitl, K. H. Schlünzen, J. Kretzschmar and J. Quaas. "ATMODAT Standard (v3.0)." World Data Center for Climate (WDCC) at DKRZ(2021). https://doi.org/10.35095/WDCC/atmodat_standard_en_v3_0

[5] Wilkinson, M. D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific Data, Band 3 (2016): 160018. https://doi.org/10.1038/sdata.2016.18

[6] Madden, F., R. van Horik, S. van de Sandt, A. Lavasa and H. Cousijn. "Guides to Choosing Persistent Identifiers - Version 2." Zenodo (2020). https://doi.org/10.5281/zenodo.3956569.

[7] ISO 8601-1:2019(en). "Date and time — Representations for information interchange — Part 1: Basic rule."(2019). https://www.iso.org/standard/70907.html

[8] ISO 19108:2002(en). " Geographic information — Temporal schema." (2002). https://www.iso.org/standard/26013.html

[9] Department of Defense. "World geodic system 1984 – its definition and relationships with local geodic systems." Technical Report 3rd Edition, Amendment 1, Geodesy and Geophysics Department, National Imagery and Mapping Agency (NIMA) (2000). https://earth-info.nga.mil/GandG/publications/tr8350.2/tr8350_2.html

[10] DataCite Metadata Working Group. " DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.3." DataCite e.V. (2019). https://doi.org/10.14454/7xq3-zf69

[11] Scherer, D., F. Antretter, S. Bender, J. Cortekar, S. Emeis, U. Fehrenbach, G. Gross, G. Halbig, J. Hasse, B. Maronga, S. Raasch and K. Scherber. "Urban climate under change [UC] 2-A national research programme for developing a building-resolving atmospheric model for entire city regions." Meteorologische Zeitschrift Heft 28, Band 2 (2019): 95-104. https://doi.org/10.15488/10423

[12] Scherer, D., F. Ament, S. Emeis, U. Fehrenbach, B. Leitl, K. Scherber, C. Schneider and U. Vogt "Three-dimensional observation of atmospheric processes in cities." Meteorologische Zeitschrift, Band 28, Heft 2 (2019): 121-138. https://doi.org/10.1127/metz/2019/0911

[13] Martin, D., C.S. Price, I.R. White, G. Nickless, K.F. Petersson, R.E. Britter, A.G. Robins, S.E. Belcher, J.F. Barlow, M. Neophytou, S.J. Arnold, A.S. Tomlin, R.J. Smalley and D.E. Shallcross. "Urban tracer dispersion experiments during the second DAPPLE field campaign in London 2004." Atmospheric Environment, Band 44, Heft 25 (2010): 3043-3052. https://doi.org/10.1016/j.atmosenv.2010.05.007

[14] Bohnenstengel, S., K. H. Schlünzen, and D. Grawe. "Influence of thermal effects on street canyon circulations." Meteorologische Zeitschrift Band 13, Heft 5 (2004): 381-386. https://doi.org/10.1127/0941-2948/2004/0013-0381

[15] Salim, M. H., K. H. Schlünzen, D. Grawe, M. Boettcher, A. M. U. Gierisch and B. H. Fock. "The microscale obstacle-resolving meteorological model MITRAS v2.0: model theory." Geosci. Model Dev., Band 11 (2018): 3427–3445. https://doi.org/10.5194/gmd-11-3427-2018

[16] Maronga, B., G. Gross, S. Raasch, S. Banzhaf, R. Forkel, W. Heldens, F. Kanani-Sühring, A. Matzarakis, M. Mauder, D. Pavlik, J. Pfaffenrott, S. Schubert, G. Seckmeyer, H. Sieker and K. Winderlich. "Development of a new urban climate model based on the model PALM-Project overview, planned work, and first achievements." Meteorologische Zeitschrift, Band 28, Heft 9 (2019): 1-15. https://doi.org/10.1127/metz/2019/0909

DataPLANT – Tools and Services to structure the Data Jungle for fundamental plant researchers

Timo Mühlhaus, Dominik Brillhaus, Marcel Tschöpe, Oliver Maus, Björn Grüning, Christoph Garth, Cristina Martins Rodrigues and Dirk von Suchodoletz

The DataPLANT consortium focuses on the continuous development and improvement of mechanisms and services for collaborative research based on sharing, enrichment and crosslinking of plant-research specific (meta)data. For this purpose, the DataPLANT tool and service chain is intended to facilitate overarching collaboration and research context management, ultimately leading to a more open and cooperative handling of research data through publication. DataPLANT follows a gradual and iterative approach, ensuring the commitment and alignment of expectations of all stakeholders. This particularly emphasizes the interaction between the community and DataPLANT.

The set of tools and microservices developed and advanced in the last couple of months focused on the pre-existing digital landscape of the average plant scientist. The first important step to data management and publication is the assisted annotation of raw data sets through the Swate Workflow Annotation Tool for Excel, which integrates the required external ontologies. The selection of the relevant metadata is simplified by provisioning of metadata templates and the use of non-integrated terms is supported by the Swate OBO Updater. The ArcCommander helps with the creation of the specific folder and file structure following the concept of the Annotated Research Context. In the future, a comprehensive workflow integration and a collaborative platform for data provenance and research sharing will emerge supporting decentralized and centralized digital processes. A central DataPLANT Hub will offer an aggregation of services and knowledge, generating a searchable compendium for research in plant biology.

1 DataPLANT core motivation

In many disciplines, scientists increasingly rely on research data management (RDM) services and infrastructures that facilitate the collection, processing, exchange and archiving of research data sets. A modern, integrated RDM enables reproducible research, the linking of interdisciplinary expertise, the sharing of research for comparison and integration of different analysis results and metadata studies, taking advantage of the immense additional knowledge gained from them. DataPLANT[1] as part of the National Research

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13724 (CC BY-SA 4.0)

Data Initiative (NFDI)[2] aims to generate this added value in the field of fundamental plant research. In this domain, the (molecular) principles of plant life that determine plant growth, crop yield and biomass production are investigated. The methods used for this purpose nowadays often comprise high throughput techniques e.g. *omics and imaging techniques. These generate high-dimensional data which have to be integrated for meaningful interpretation. Successful collaboration and use of data of different modalities – from many sources and experiments, pre-processed or analysed with a variety of algorithms – requires annotation, standardization and contextualization of the data i.e., in a metaphoric sense, a structuring of the data jungle.

The FAIR[9] and Linked Open Data[4] Principles provide an abstract guideline for RDM. Nevertheless, besides these stated best practices it is almost always left to the initiative of individual researchers to implement them, requiring significant time and resources. To address this bottleneck, we opt for a close community-integrative approach mirrored in a three-pillar structure of i) standardization, ii) personal, and iii) technical support for research groups and individual researchers[5]. By combining technical expertise in basic plant research, information and computer sciences and infrastructure specialists, DataPLANT supports plant scientists in all aspects of RDM. It strives to advance a specific community standard for fundamental plant research (meta)data and workflow annotation and provides the necessary tools to facilitate the annotation and handling of data.
Based on the expertise of computer scientists, bioinformaticians, service providers and contribution from the community, development principles were established leading to a first set of tools and workflows has been developed and made available, the elaboration of which is detailed in this paper.

2 Fundamental design principles of the DataPLANT tool chain

Developing applications and tools that support community-driven RDM exceeds beyond writing code. Design principles provide high-level guidelines and a collection of considerations to create successful applications. In DataPLANT, tool development is always motivated by community requirements conveyed by researchers e.g. through data stewards to developers. The objective in DataPLANT is to provide incremental but regular improvements of the digital processes from the very beginning of the project. This will be achieved through iterative but multiple measures allowing a fast start and the possibility of a timely feedback from the practitioners in the field. The main platforms for exchange are the DataPLANT hub for documentation and the public code repository hosted on Github[6]. Ongoing activities are overseen by both a scientific and technical board. Additionally, we continuously survey our community to allow the swift integration of it's needs into the development process. It enables us to integrate our support tools and services in the work processes of the different laboratories.

This means we acknowledge the fact that RDM still represents a considerable additional effort for scientists and is therefore essentially an ad hoc management of experimental

data. Scientists are accustomed to documenting their research in free-text documents or tables loosely organized in file and directory structure. This leads to a preference for a flexible structure to support RDM in practice and reflects the dynamic nature of research. The strong desire to have full control over the research data that originates locally enables decentralized data management and tools that meet the researchers where they are. However, the advantages of cloud-based solutions are obvious and popular when it comes to querying and processing data. We reflect the natural behavior of the researcher in our design principles and avoid creating any type of walled garden or operation lock in. Therefore, we build our tool chains on top of existing tools and standards with an additional layer of comfort for biologists. It should be possible to perform any process in our tool chain without dedicated software. This increases opportunities for others to use and improve upon our work to embrace the open software principles.

3 ARC: a data-centric integration

A major challenge in modern RDM is the scientific integration of different decentralized data- and infra-structures. The evolving nature and needs of various research communities have led to a constant increase in heterogeneity of data standards, software and hardware solutions in the past. Now, given a transformation towards an integrated, multi-provider RDM model, this change in focus has many implications for the development and application of IT systems used in RDM. Regarding interoperability, there are two orthogonal models of thought: an application-centric and a data-centric one.

According to the notion, within an application-centric model, application, software and services are the main focus of integration. This requires a well-defined exchange of all information between the interconnected components. Consequently, it is necessary to agree on the exchange format or APIs to incorporate different functionalities. The main advantage of this approach is the ability to get the most out of legacy software and services due to the large number of systems already existing. However, each application needs to employ its own data model, which depends on specific functions and tasks. Therefore, the complexity increases by the sum of all elements that developers and users need to know in order to master such a system.

The data-centric model[7, 8] is based on an architecture in which data is the primary and permanent asset and applications are interchangeable. In such an architecture, the data model precedes the implementation of any given application. At this point, services and applications are in a state of constant change to meet user requirements and experiences or functionality extensions. In the data-centric world, the data model focuses on semantics. The structure, constraints, and validation that need to be done to the data are only secondarily included. This allows for a local and independent model to support functionality and a separation of concerns regarding system design and interoperability.

In respect to RDM, it seems natural to consider the data as the center and build the DataPLANT tool chain around it. This results in the technical realization and standardized

RDM procedures being process-oriented, meaning that each tool realizes or supports the researcher in a distinct task within the RDM cycle. Consequently, this enables the desired mixed mode of application, in which both human and machine can operate processes simultaneously or asynchronously. In addition, we thereby avoid technological barriers and embrace open software and open science, respectively.

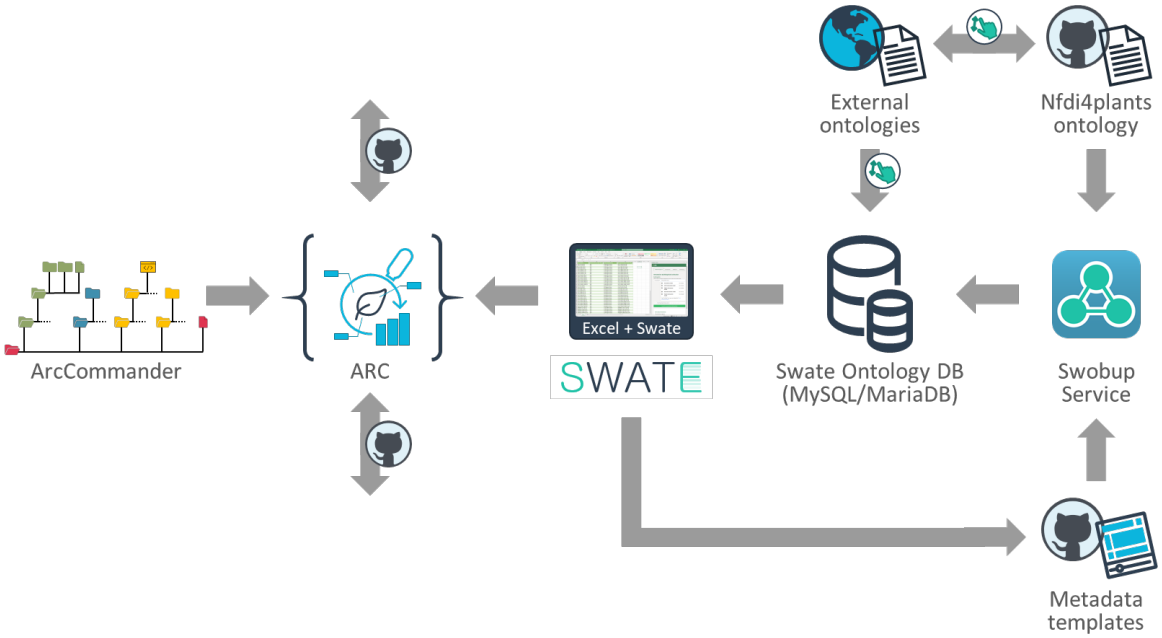


Figure 1: DataPLANT Metadata Toolchain.
The **ArcCommander** helps with the creation of the specific folder and file structure within an **ARC**. In this context, **SWATE** (Swate Workflow Annotation Tool for Excel) supports the metadata annotation process based on the **Swate DB** (Database), which integrates the required external ontologies. The selection of the relevant metadata is simplified by provision of **metadata templates** adapted to public repositories. To enable the user to use non-integrated terms, **Swobup** (Swate OBO Updater) bridges the gap to the **NFDI4plants ontology**, which stores these terms temporarily until incorporation into existing ontologies.

To realize a data centric approach for RDM in DataPLANT, we propose the Annotated Research Context[9], that captures and structures the complete research cycle meeting the FAIR requirements with low friction for the individual researcher. ARCs are self-contained and include assay/measurement data, workflow and computation results accompanied by metadata in one package. Their structure allows full user control over all metadata and facilitates usability, access, publication and sharing of the research. Thereby, ARCs are a practical implementation of existing standards encompassing the benefits of the ISA model, research object crates[10, 11] and the Common Workflow Language[12].

The ARC concept relies on a structure that partitions assays, workflow and results for a granular reuse and development. “Assays” cover biological, experimental and instrumental data including their self-contained description using the ISA model. Similarly,

“workflows” cover all digital steps of a study and contain application code, scripts and/or any other executable description of an analysis ensuring highest flexibility for the scientists. However, to ensure persistence and reproducibility, digital processes or ”workflows” comprise their own containerized running environment. The “result” data is linked to the workflows by a minimal Common Workflow Language file specifying the input and output of the process. The suggested structure for ARCs is a starting point for individual research projects and defines a framework for the organization, sharing, reuse (clone) and evolution (fork/pull request) of research projects in a fashion familiar in open-source software development.

4 Templates for convenient metadata annotation

The tools developed in DataPLANT assist in ARC creation as well as evolution through collaborating, sharing and publishing. However, the most important aspect is to increase the user engagement to actually collect metadata in a human readable and machine usable form. Therefore, a first step is to ease the publication process of research data to public repositories and shift the workload of metadata generation away from the user by template convenience.

Metadata annotation as part of the data submission routine to public endpoint repositories is often bothersome due to a high variability between repository requirements. Differences exist in both the content required (e.g. to comply with underlying metadata standards or minimal reporting guideline like MIAME[13], MINSEQE[14], MIAPPE[15]) as well as the mode format required for submission (FTP, API, webform) and also the presentation (e.g. spreadsheet, web form, check lists or mixes thereof) for meta, raw or processed data, respectively. This can become particularly inconvenient when the same metadata is submitted repeatedly or in large volumes, as in cases of individual researchers submitting the same metadata to multiple unrelated endpoint repositories or data experts requiring different formats (e.g. data champions, core facility heads) that repeatedly submit similar data to the same endpoint repository where duplication of metadata between studies is not supported.

In addition, proper metadata annotation requires use of controlled vocabularies and ontologies, which is often not intuitively supported by repository tools and can be challenging for an untrained user. Post-submission modifications and updates to datasets and metadata can be fragmented and require redundant work, e.g. when metadata on the study level needs to be updated that would affect datasets submitted to different repositories, thus eliminating version control between metadata descriptions. In summary, the wet-lab biologist can easily lose a significant amount of time adapting submission routines. Additionally, lack of flexibility -e.g. rigid requirements for metadata terms- can jeopardize the willingness or even ability to submit, if information is simply not available or the requirement is incompatible with the dataset being described.

To overcome this annotation nuisance, DataPLANT provides a growing collection of templates designed and curated by data stewards that cover the submission routine to selected end-point repositories. The template design process is initiated "backwards", starting from the requirements of end-point repositories and thereby compliance with metadata standards. Data stewards supervise the implementation of ontologies and the use of controlled vocabulary as required by the target repository, and simultaneously contribute to the development of the DataPLANT broker ontology. To provide high quality templates, data stewards cross-validate usability in an independent template reviewing process. This includes periodic mock submissions of "vanilla use-cases" through the pipeline to verify that these are fully compliant and at the same time user friendly and/or where they could be improved. High flexibility is fostered by offering a choice of modes for template distribution, use and customization.

To support this process, DataPLANT introduces SWATE[16] (Swate Workflow Annotation Tool for Excel) as a one-stop-shop (but not one-fits-all) metadata capture approach which leverages on the flexibility of the well-established ISA framework to supplement the ARC research object. As a starting point and user guide for metadata annotation. Once SWATE is installed (on the user-side) or used in the online version, templates can be loaded directly from a database. Meta information supplied by the template authors such as the target repository, study or assay type, enables the selection of suitable templates. Alternatively, SWATE templates can be easily shared and propagated via conventional routes (email, storage cloud or server), also allowing reuse of previous templates. Accordingly, the templates can be filled with or without the help of SWATE. While the latter increases the need for post-annotation curation, the former requires more expertise on the user side, but allows direct linkage to ontology references.
SWATE metadata templates are designed as a non-restrictive starting point and the user is encouraged to expand them with additional attributes. However, to minimize the need for (unsupervised) customization, data stewards interact closely with users and data type experts (champions) to integrate and align their feedback during template design. This can eventually lead to provision of very specific templates e.g. for individual core facilities or research groups to perfectly align with their daily laboratory routines. In addition to leveraging their multiplier role, this supports recording metadata at the place and time of its emergence, mitigating the need for redundant, retrospective annotation or even loss of information.

In this way, DataPLANT rethinks standardization from a purely technical towards a user-friendly, applicable perspective that aligns well with the progress of scientific innovation (e.g. new techniques and data types). With a growing user community and strong data steward interaction, the templates are continuously polished, crystallizing what information is frequently required or lost, and filling the ontology knowledge gaps. Combined with the full suite of the DataPLANT toolbox, SWATE templates lower the users' burden and workload of data publishing in the long run. However, they also allow immediate benefit through repository compliance, harmonized grammar, structure, and use of ontologies, and by guaranteeing usability independent of other DataPLANT mechanisms.

5 Swate for ontology driven metadata annotation

At first glance, the diversity of fundamental research is not ideally mirrored by the rigidity of standardization processes and requirements of metadata management. The balancing act between requirements of the researcher and standardization is especially applicable for the annotation of experimental measurement workflows. The spreadsheet-based version of the ISA standard[17] allows for ontology-driven metadata annotation of technical workflows in a simple and accessible way, compromising between free form and aligning with standardization efforts. However, finding the appropriate ontology term can be extremely tedious and often results in incomplete metadata annotation. In the DataPLANT tool chain we offer Swate to facilitate this via an integrated search function and an ontology guided metadata annotation.

Swate is an add-in that allows the user to easily annotate their data according to ISA standards. It focuses on providing an easy-to-use tool in the widely used and thus familiar environment of the most used spreadsheet editor. In order to directly incorporate collaborative mechanisms, the tool is implemented both as a modern web-based online application and as a desktop application. Fully integrated in Microsoft Excel, users can add and delete columns with specialized headers describing their data in a clear representation. By design, Swate facilitates ontology-driven development of data annotation schemes by the domain experts performing the actual research data generation. Swate features a search function for ontological terms, facilitating an ontology-driven annotation of the data. It can insert ISA-conform protocols and processes that support the DataPLANT template mechanism. By making the trade-off between free form and alignment with a standardization, e.g., ISA syntax, we believe to encourage more researchers to increase their annotation data input. Leveraging standard spreadsheet features, such as color coding, style, and markup as free comment or highlight functionality, ultimately increases the user acceptance and user experience dramatically without polluting the actual metadata information separately stored in the specialized xml dialect named spreadsheetML.

Finally, Swate simplifies mapping between models and their semantic representations in the form of the ISA model, facilitating machine readability of user-annotated data in the result.

6 Swobup and SwateDB, a team for metadata broker to bridge the ontology gap

One of DataPLANT´s core responsibilities is to reduce the effort for users as well as to increase comfort in providing human- and machine-readable metadata. Therefore, the use of controlled vocabularies is indispensable. Controlled vocabularies enable scientists to easily classify, find and reuse data. Reuse is further supported by hierarchical or relational conceptualization in form of an ontology, rendering data and metadata readable to humans and machines. The ambition of DataPLANT lies in addressing all plant research data from

the greenhouse to the lab bench to the endpoint repository. There are excellent ontology portals or Ontology Lookup Services available. These tools and services provide an easy access to all or most available ontologies. However, the problem for our community centric approach is that these services may offer similar terms from different sources leaving the user alone with a decision. This leads to a clatter in ontology references or in the worst case to ambiguity that in itself defeat the purpose to use an ontology in the first place.

In DataPLANT, Data stewards supervise the ontology selection to fine-balance between a limited scope (i.e. pre-selection of ontologies) and the flexibility of the user to provide mandatory and meaningful terminologies for data descriptions. Consequently, ontologies are selected and imported into a database called SwateDB. SwateDB comprises a controlled and practical mix of ontologies developed for specialized plant sciences topics, as well as technical terminologies required for the acquisition (omics) and analysis (bioinformatics) of biological data. Additionally, the SwateDB is the central storage for metadata annotation templates that can be managed and consumed by our tool chain.

However, experience has taught us that missing or unsuitable ontology terms and relations lead to a setback of user motivation and participation. DataPLANT provides a dedicated NFDI4plants ontology (to act as a broker and) bridge between the individual researcher and main ontology provider. This ontology enables the collection of missing vocabulary for immediate use and is also stored in the SwateDB. The automatic process handler Swobup[18] (Swate OBO Updater) simplifies this ingestion process of adding or removing (one or more) terms from an ontology and synchronizes ontology terms in OBO format and publicly hosted metadata templates either by pull request mechanism or a group of authorized users. File versioning and adaptations to the database are outsourced to a shared repository. Any change can be reverted using the repository's built-in features.

Swobup recognizes these changes and reverts to previous versions based on the principles described above. Swobup parses the OBO or template file and incorporates the changes into the SwateDB database. In order to work with Swobup a webhook has to be defined in Github and configured, that it sends a SSL encrypted HTTP Post request to a previously configured URL every time files are changed in the repository. This process allows an immediate update process including version control and history driven by the community. Therefore, anyone is directly or indirectly (via pull request) able to update templates or ontologies without delay and a minimum amount of guidance. This process enables DataPLANT to act as an ontology broker, collecting required terms from the community in the NFDI4plants ontology and forwarding them to the main ontology provider.

7 ARC Commander – Support Tool

For the community of plant researcher, experimental metadata in a structured user-friendly format are most useful to reuse research data and generate new biological knowledge. However, it is advantageous to argument these data with supplementary organizational metadata. Additionally, a solution to organize and manage metadata and research

data practically with low friction for the user needs to be provided. Therefore, DataPLANT introduced the ARC into the research data management landscape. The ARC is an intuitive specification for the primary setup of an experiment and collaboration environment for the storage of research data including context like metadata. Most importantly, the ARC layout follows a specific file and folder structure derived from the RO crate standard and components are registered in a central investigation file that follows the ISA model specification.

Although these requirements are minimal, assisting the researcher in these repetitive tasks and providing structured guidance is beneficial and reduces friction and workload on the side of the user. This is the main aim of the tool ArcCommander[19]. Essentially, this tool provides automation and assistance with processes following the ARC specification. The ArcCommander can be executed to initialize an ARC, creating the basic folder structure, and setting the working environment. Additionally, it can be used to create and modify sub-branches of the ARC, such as assays and workflows. By using the ArcCommander, the researchers are guided during the process and can create or maintain the ARC without needing explicit knowledge of the ARC structure. Besides ARC specifics, general naming recommendations shared across operating systems are adhered to by the ArcCommander. Following the ARC or ISA model respectively, a central registry, called investigation, is stored as a file in which all components of the ARC are registered. Manual registry synchronization upon addition of further content would be time-consuming and error-prone, but can be achieved automatically using the ArcCommander.

In its current state, the ArcCommander is implemented as a command line tool. Often there are experiments that are very similar to each other in some characteristics. For example, proteomics measurements performed in the same laboratory might follow the same protocols. In this case, the resulting ARCs are also likely to have some properties in common. Here, repetition can be easily reduced by concatenating the commands using a script. Commands and parameters are designed to automatically guide the researcher through the process of creating an ARC. This guidance is realized by providing a hierarchically structured and extensively labeled command set, which can be easily and purposefully browsed for the command of interest. Additionally, the user experience is enhanced by a text editor enabling to automatically generate metadata schemata. Instead of specifying all arguments in the command line, a text-based form is created and presented to the user that handles the metadata retrieval.

Enabling successful data sharing, working in teams and information exchange between researchers are the fundamental tasks in RDM. The ArcCommander supports collaborative work by leveraging Git-based version control to keep track of file change history as well as user interactions and contributions. The ArcCommander implements a convenience layer on top of Git to enable synchronization functionality for non-expert users. Besides using standard Git, it can handle large files which are common in research using Git-LFS (Large File Storage). By this, researchers can easily share and control their state of the ARC without additional efforts.

The modularity of the ArcCommander adequately accounts for the dynamic design principles and flexible extensibility required to maintain and extend ARC functionality. A process that requires the most extensibility might be the data publication based on ARCs. After seamless creation of the ARC, a major interest for researchers is the distribution of their research data. An increasing need has been to publish data in a centralized repository that prescribes individual technical metadata information and format requirements. Meeting these requirements imposes a significant additional burden on the researcher. Therefore, the ArcCommander aims to provide automatic export to different central data repositories and to support transformation of the ARC along different formats and requirements. Already included in the ArcCommander is the possibility to easily export the metadata available in "ISAxlsx" format to "ISAJson" and "ISATab" standard formats. This allows for seamless interoperability with available ISA Tools and all central repositories complying with standard ISA model specifications. For the future, a successive extension of the export functionality is planned in order to provide compatibility to additional data repositories and community resources.

8 Workflow and data integration with the Galaxy Gateway

Due to our data-centric approach in DataPLANT, all digital processes are centered around the ARC. This includes or especially applies to data processing and analysis workflows. Galaxy[22] is an integral part of DataPLANT regarding workflow management. Dedicated tools for the plant science community will be integrated during the DataPLANT project and will extend the portfolio already available at Galaxy for Plants[21]. For years, Galaxy has made advanced bioinformatics software accessible to scientists worldwide by providing an intuitive web interface to these applications while fostering reproducibility through the automatic creation of re-runnable protocols of each analysis. The Galaxy community is one of the largest bioinformatics communities world-wide[22]. It provides over 7000 tools and a plenitude of bioinformatics and data processing workflows useful for researchers from the fundamental plant research community. The core framework offers various abstraction layers that offer various extension points and adopt Galaxy to new technologies, while keeping the system maintainable since 16 years.

Tools in Galaxy are independent of each other and contain rich metadata annotations, including all their dependencies. Those dependencies are resolved via different Galaxy plugins for Modules, Conda, Docker, Singularity or others. For truly reproducible research the Galaxy community recommends different approaches depending on the degree of reproducibility and cost. Conda for more flexible and cost efficient tool dependency management and containers for elaborated and isolated environments that are more cost intensive in maintenance. For both scenarios the Galaxy community offers solutions with Bioconda[23] and BioContainers[24].

Another subsystem in Galaxy is the handling of user data. Galaxy supports different kinds of data storages, ranging from hierarchical POSIX storages, to S3 or iRODS. Those can be bound to users, groups or roles and enable flexible quota assignments per object store.

A similar system allows users to browse and import public data deposited on S3, SFTP, Webdav or Dropbox accounts. Galaxy also supports the export of research artefacts, like workflow invocations. Currently, those can be exported as BioComputeObjects and support for ResearchObjects and ARCs are planned.

9 Conclusion and Outlook

Dedicated to structuring the data jungle for fundamental plant researchers, the DataPLANT consortium started its operation by assembling a suite of tools that should greatly facilitate efforts of proper research data management. As a simple structural scaffold, the Annotated Research Context is introduced, which intrinsically follows FAIR requirements.

As a starting point, the ArcCommander simplifies the generation of the ARC folder and file structure. Following the ISA model, this includes a central registry in the form of an automatically updated investigation file. For researchers the annotation of metadata seems to be the most tedious task of the RDM cycle. Swate supports the user during this process. SWATE offers a set of ontologies and metadata templates pre-selected and curated by data stewards and facilitates simple re-use of metadata. At the backend, this is enabled by Swopub, which continuously synchronizes SWATE with the SwateDB according to adaptations to templates and the NFDI4plants broker ontology. As a result of user feedback, it was already possible to create a set of initial templates that converged the user needs with the requirements of corresponding endpoint repositories. This is currently expanded to cover centralized computer-based workflows and integrate already available services such as Galaxy or nf-core[25].

According to DataPLANT's prevailing data-centric view, the listed tools serve to improve user-friendliness based on the current state of the art. Everybody from the open source community is encouraged to take the initiative and adapt existing or own tools to the changing needs or to inquire us directly. Due to the modular structure and ARC centric integrative approach in DataPLANT a continuous improvement of our services comes naturally. DataPLANT envisions a cloud version of the tool chain to be integrated centrally in the DataPLANT Hub in the future. Besides providing a public website that gives information about the project, shares news, and provides links to the project's social media channels and Git repositories, the DataPLANT Hub will create a central environment for the community. A key component will be the integrated search and exploration engine for research data using annotated metadata. Thus, the DataPLANT Hub will be a key component to make research data FAIR.
Thus, an orderly floral bouquet of user-oriented tools can blossom out of the overgrown jungle.

Acknowledgements

We acknowledge support for DataPLANT 442077441 through the German National Research Data Initiative (NFDI 7/1), the TTRR 175 (INF project), and CEPLAS is supported by Deutsche Forschungsgemeinschaft within the Excellence Initiative (EXC 1028) and under Germany's Excellence Strategy – EXC 2048/1 – project 390686111.

ORCID IDs

- Timo Mühlhaus https://orcid.org/0000-0003-3925-6778
- Dominik Brillhaus https://orcid.org/0000-0001-9021-3197
- Marcel Tschöpe https://orcid.org/0000-0002-3731-7664
- H. Lukas Weil https://orcid.org/0000-0003-1945-6342
- Oliver Maus https://orcid.org/0000-0002-8241-5300
- Björn Grüning https://orcid.org/0000-0002-3079-6586
- Christoph Garth https://orcid.org/0000-0003-1669-8549
- Cristina Martins Rodrigues https://orcid.org/0000-0002-4849-1537
- Dirk von Suchodoletz https://orcid.org/0000-0002-4382-5104

Bibliography

[1] DataPLANT. https://www.nfdi4plants.de/ (accessed: 11. May 2021).

[2] NFDI. https://www.nfdi.de/ (accessed: 11. May 2021).

[3] Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. The fair guiding principles for scientific data management and stewardship. Scientific Data 3, Nr. 1. https://doi.org/10.1038/sdata.2016.18

[4] Bizer, C., T. Heath, K. Idehen and T. Berners-Lee. 2008. Linked data on the web (LDOW2008). Proceeding of the 17th international conference on World Wide Web – WWW '08. doi:https://doi.org/10.1145/1367497.1367760

[5] von Suchodoletz, D. , T. Mühlhaus, J. Krüger, B. Usadel, C. Martins Rodrigues, "DataPLANT – Ein NFDI-Konsortium der Pflanzen-Grundlagenforschung", Bausteine Forschungsdatenmanagement, no. 2. German:46-56 (2021). https://doi.org/10.17192/bfdm.2021.2.8335

[6] DataPLANT - Services and infrastructures to support FAIR Data science and good data management practices within the plant basic research community. GitHub. https://github.com/nfdi4plants (accessed: 11. May 2021).

[7] Liu, Z., S. Fan, H. Jiannan Wang and J. L. Zhao. 2017. Enabling effective workflow model reuse: A data-centric approach. Decision Support Systems 93: 11–-25. doi:https://doi.org/10.1016/j.dss.2016.09.002

[8] Wuyts, R., S. Ducasse and O. Nierstrasz. 2005. A data-centric approach to composing embedded, real-time software components. Journal of Systems and Software 74, Nr. 1: 25–-34. doi:https://doi.org/10.1016/j.jss.2003.05.004

[9] ARC. GitHub. https://github.com/nfdi4plants/ARC (accessed: 11. May 2021).

[10] Carragáin, E. Ó., C. Goble, P. Sefton and S. Soiland-Reyes. 2019. A lightweight approach to research object data packaging. Zenodo. doi: https://doi.org/10.5281/zenodo.3250687

[11] eScience Lab at The University of Manchester. Research Object Crate. researchobject.org. https://www.researchobject.org/ (accessed: 11. May 2021).

[12] Amstutz, P., M. R Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer, A. Kartashov, et al. 2017. Common Workflow Language, v1.0. eScholarship, University of California. 1. April. https://escholarship.org/uc/item/25z538jj (accessed: 11. May 2021).

[13] Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, et al. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nature Genetics 29, Nr. 4: 365–-371. doi:https://doi.org/10.1038/ng1201-365

[14] Dimitrova, M., R. Meyer, P. L. Buttigieg, T. Georgiev, G. Zhelezov, S. Demirov, V. Smith and L. Penev. 2020. A Streamlined Workflow for Conversion, Peer-Review and Publication of Omics Metadata as Omics Data Papers. doi:https://doi.org/10.20944/preprints202009.0357.v1

[15] Papoutsoglou, E. A., D. Faria, D. Arend, E. Arnaud, I.s N. Athanasiadis, I. Chaves, F. Coppens, et al. 2020. Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytologist 227, Nr. 1: 260–-273. doi: https://doi.org/10.1111/nph.16544

[16] Swate – Excel Add-In for annotation of experimental data and computational workflows. GitHub.https://github.com/nfdi4plants/Swate (accessed: 11. May 2021).

[17] Sansone, S.-A., P. Rocca-Serra, A. Gonzalez-Beltran, D. Johnson and ISA Community. 2016. ISA Model and Serialization Specifications 1.0. Zenodo. October 28, 2016. doi:https://doi.org/10.5281/zenodo.163640

[18] Swobup - Swate DB update tool. GitHub. https://github.com/nfdi4plants/Swobup (accessed: 11. May 2021).

[19] ArcCommander - Tool to manage your ARCs. GitHub. https://github.com/nfdi4plants/ArcCommander (accessed: 11. May 2021).

[20] Afgan, E., D. Baker, B. Batut, M. van den Beek, D. Bouvier, M. Čech, J. Chilton, et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46, Nr. W1. doi:https://doi.org/10.1093/nar/gky379

[21] Galaxy for Plant Biology. Galaxy. https://plants.usegalaxy.eu/ (accessed: 14. May 2021).

[22] An open source Git extension for versioning large files. Git Large File Storage.https://git-lfs.github.com/ (accessed: 11. May 2021).

[23] Grüning, B., R. Dale, A. Sjödin, B. A. Chapman, J. Rowe, C. H. Tomkins-Tinch, R. Valieris and J. Köster. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature Methods 15, Nr. 7: 475–-476. doi: https://doi.org/10.1038/s41592-018-0046-7

[24] da Veiga Leprevost, F., B. A Grüning, S. Alves Aflitos, H. L Röst, J. Uszkoreit, H. Barsnes, M. Vaudel, et al. 2017. BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics 33, Nr. 16: 2580–-2582. doi: https://doi.org/10.1093/bioinformatics/btx192

[25] Ewels, Phil. nf-core: A community effort to collect a curated set of analysis pipelines built using Nextflow. nf-core. https://nf-co.re/ (accessed: 14. May 2021).

Ein standortübergreifendes Speichersystem für Forschungsdaten

Florian Claus[1], Constanze Curdt[2], Jens Kather[3] und Stephanie Rehwald[3]

[1]RWTH Aachen
[2]Universität zu Köln
[3]Universität Duisburg-Essen

Der digitale Wandel insgesamt und das Forschungsdatenmanagement (FDM) stellen Wissenschaftseinrichtungen vor große Herausforderungen. Eine Antwort ist es, Infrastrukturen zu vernetzen und bereitgestellte Dienste arbeitsteilig zu organisieren. Diese Gedanken waren leitend für das Vorhaben eines Hochschulkonsortiums, um Vorgehensweisen und Infrastrukturen so zu etablieren, dass sie eine Grundlage für die Nachnutzung wissenschaftlicher Informationen schaffen.

Beschafft wurde ein Speichersystem von DELL, das auf einer Kombination von Objekt- und Blockspeicher beruht. Das System wurde georedundant an insgesamt elf Standorten installiert und wird gemeinsam von den antragstellenden Einrichtungen betrieben. Es bietet in der aktuellen Konfiguration ca. 22 PB effektiv nutzbaren Speicherplatz.

Das System hat die Bereitstellung von Speicher für Forschungsdaten grundlegend verändert und ist in die FDM-Konzepte der Einrichtungen eingebunden. Somit trägt es sowohl direkt als auch indirekt als Element der FDM-Konzepte zu einer Verbesserung des FDM bei.

Das Konsortium ist offen für weitere Partner aus NRW, die sich an dem Speichersystem beteiligen möchten. Ebenso kann das System durch Angehörige weiterer NRW-Hochschulen genutzt werden, die nicht Mitglied des Konsortiums sind.

1 Einleitung

Bereits in 2017 hat sich ein Konsortium zusammengefunden, um einen gemeinsamen Antrag für ein Speichersystem für Forschungsdaten zu stellen [1]. Der Antragsgegenstand wurde wenig kreativ, aber sprechend, mit „Forschungsdatenspeicher“ (FDS) bzw. in der englischen Variante „Research Data Storage“ (RDS) betitelt. Das Konsortium besteht aus der RWTH Aachen, der FH Aachen, der Ruhr-Universität Bochum, der Technischen Universität Dortmund, der Universität Duisburg-Essen und der Universität zu Köln. Die RWTH Aachen hat die Konsortialführung übernommen. Beantragt wurde zwar „nur“ ein Speichersystem: Dieses war allerdings in ein Konzept zum Forschungsdatenmanagement

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021.
Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13725 (CC BY-SA 4.0)

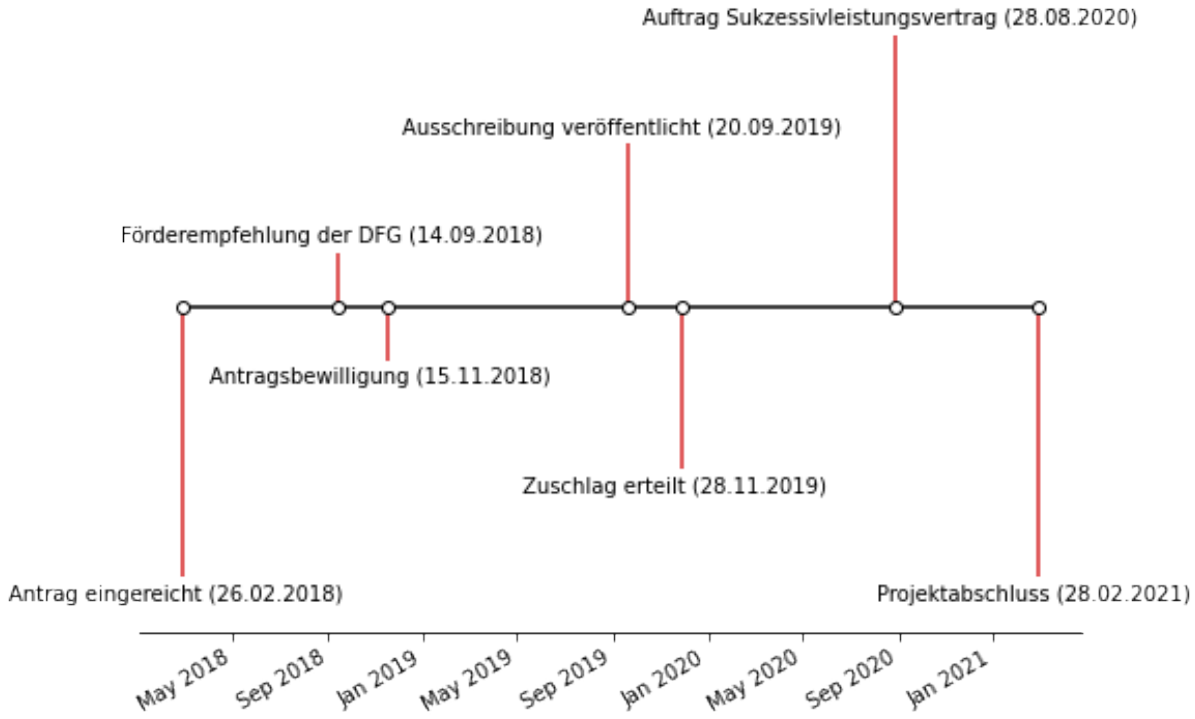


Abbildung 1: Zeitstrahl zum Projektverlauf.

eingebunden, nach dem die neue Infrastruktur nicht nur technisch innovativ sein sollte, sondern auch innovativ auf den Ebenen der Prozesse und Kultur in der Forschung wirken sollte.

Abbildung 1 zeigt den groben zeitlichen Verlauf des Projekts.

2 Ziele

Im Antrag sind vier Zielsetzungen genannt: (1) Zweckbindung des Speichersystems an das Forschungsdatenmanagement (FDM), (2) Möglichkeit der hochschulübergreifenden Speicherung und Nutzung, (3) IdM-basierter Zugang und (4) Vergabe der Ressourcen gemäß wissenschaftsgeleiteter Kriterien.

Auf der Ebene der technischen Infrastruktur sollte ein Speichersystem beschafft und als verteiltes System an den verschiedenen Standorten der antragstellenden Einrichtungen betrieben werden. Durch die Kooperation sollten Skaleneffekte genutzt werden: Das Auftragsvolumen vergrößerte sich durch die gemeinsame Beschaffung, wodurch die Beschaffungskosten gesenkt werden konnten. Durch den verteilten Aufbau konnte eine verbesserte Standortredundanz erreicht werden. Im Betrieb sollte der Aufwand für den Aufbau von Expertise gesenkt und diese Expertise wiederum an allen Standorten gesichert werden.

Auf der Prozessebene war der Ansatz insofern neu, dass explizit ein Speichersystem für Forschungsdaten beantragt wurde. Damit einher ging das Konzept eines wissenschaftsgeleiteten Antragsverfahrens. Speicherplatz sollte nicht mehr an einzelne Hochschuleinrichtungen, sondern, analog zum Ressourcenmanagement im Bereich des Hochleistungsrechnens, an dezidierte Forschungsprojekte vergeben werden. Zudem sollte das Speichersystem in lokale Anwendungsumgebungen integriert werden, die die Anreicherung der Daten mit persistenten Identifiern (PIDs) und Metadaten ermöglichen.

Die Innovationen auf der Prozessebene sollen schließlich zu einem Wandel hin zu einer Forschungskultur beitragen, in der Forschungsdaten als wertvolle Ressourcen wahrgenommen und behandelt werden. Dazu gehört, sie explizit anders als Daten aus dem anderen universitären Kernbereich, der Lehre, oder aus Unterstützungsprozessen (Verwaltung) zu behandeln.

3 Speichersystem

Um die Anforderungen an das zu beschaffende Speichersystem abzuschätzen wurden an den antragstellenden Einrichtungen zur Vorbereitung Bedarfserhebungen vorgenommen. Daraus resultierten Anforderungen an die Größe des Speichersystems, an die Performanz der Zugriffe und die möglichen Zugriffsformen. Das Ergebnis der Anforderungsanalyse war jedoch so wenig überraschend wie hilfreich für eine Priorisierung von Eigenschaften: Das Speichersystem sollte hohes Datenvolumen, hohe Performanz, hohe Datensicherheit und flexible Zugriffsmöglichkeiten mit niedrigen Kosten verbinden. Für die Gestaltung des Speichersystems waren somit eher die konkreten Nutzungsszenarien hilfreich, die in den betrachteten Use Cases konkretisiert worden waren.

Zudem war klargeworden, dass das Speichersystem flexibel erweiterbar sein müsste, da zukünftige Bedarfe nicht sicher abgeschätzt werden können.

Neben den grundlegenden technischen Eigenschaften des Systems ließen sich aus den oben genannten Zielen noch weitere Anforderungen ableiten: Das System sollte verteilt über räumlich weit entfernte Standorte aufgestellt werden. Eine redundante Verteilung der Daten sollte automatisiert und für Nutzende transparent erfolgen. Daten sollten über alle Protokolle eingeliefert und alle Daten auch über alle Protokolle abgerufen werden können.

Den Zuschlag für das Speichersystem erhielt letztlich das Systemhaus Concat als Auftragnehmer, dessen Angebot auf DELL-Systemen beruhte.

Das beschaffte System besteht zum einen aus DELL ECS Servern mit einer Gesamt-Bruttokapazität von 51,56 PB. Diese Systeme können über das S3-Protokoll angesprochen werden. Diese Volumensysteme werden durch hochperformante DELL Isilon Systeme ergänzt, die über die Protokolle SMB/CIFS und NFS erreichbar sind. Diese Systeme verdrängen Daten wiederum in die ECS-Systeme, so dass sie als Cache fungieren und mehr Daten aufnehmen können als ihre Gesamt-Bruttokapazität von 0,59 PB. Auf den Systemen läuft eine proprietäre Managementsoftware von DELL.

Die Systeme sind an insgesamt elf Standorten (Gebäuden) aufgebaut: jeweils drei in Aachen und Köln, zwei in Dortmund und je einer in Bochum, Duisburg und Essen.

Die Systeme können flexibel in Replikationsgruppen organisiert werden, zwischen denen dann automatisiert Daten ausgetauscht werden, um diese gegen den Ausfall einzelner Standorte abzusichern. Gegen irrtümliche Löschung oder Änderung von Daten schützt die Versionierung der Daten. Ein systemexternes Backup ist darüber hinaus nicht vorgesehen und müsste ggfs. an den einzelnen Standorten zusätzlich realisiert werden. Hierfür kann beispielsweise ab 2022 die NRW-weite Infrastruktur des Projektes Datensicherung.NRW verwendet werden. Es wurden zum einen lokale Replikationsgruppen eingerichtet, die nur die Systeme an einer Hochschule einschließen. Somit kann sichergestellt werden, dass Daten die Hochschule nicht verlassen. Dies wird in einzelnen Forschungsprojekten, z. B. bei Industriekooperationen, gefordert.

Daneben gibt es aber auch eine Replikationsgruppe, die alle Standorte einschließt und zur mittel- sowie langfristigen Sicherung von Daten dient. Abbildung 2 zeigt die Aufstellungsstandorte sowie die konfigurierten Replikationsgruppen.

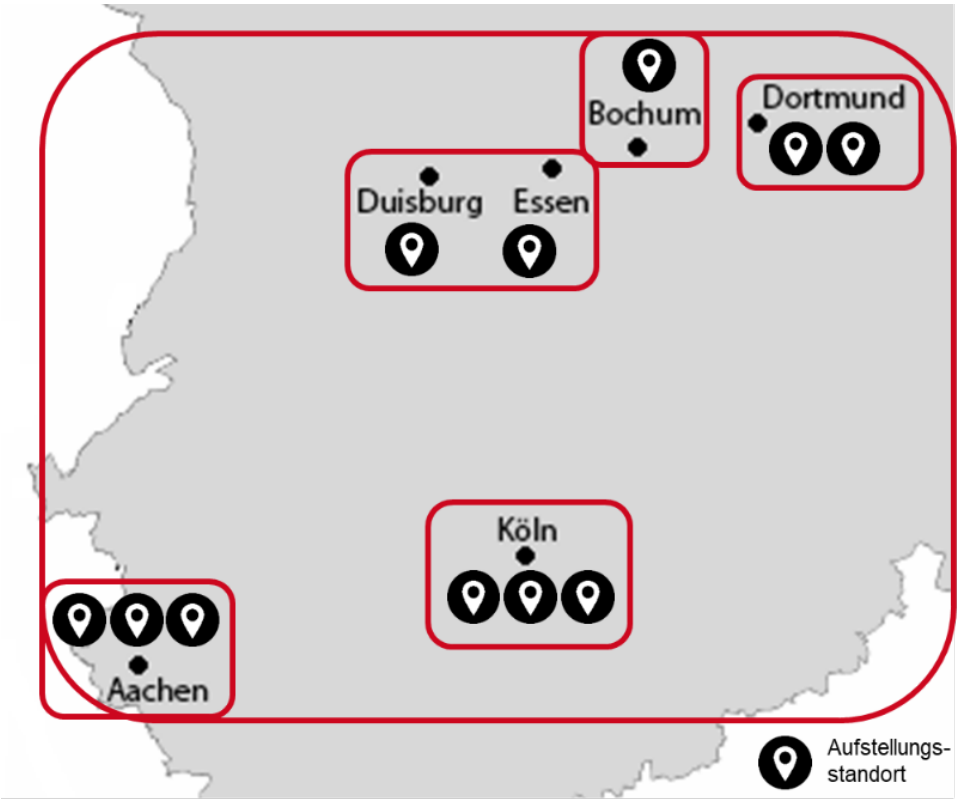


Abbildung 2: Aufstellungsstandorte des Speichersystems.

Durch die Redundanz der Daten verringert sich natürlich der netto nutzbare Speicherplatz. So sind in der aktuellen Konfiguration bei den Isilon-Systemen von den brutto 0,59 PB netto 0,41 PB nutzbar, bei den ECS-Systemen bleiben von brutto 51,56 PB netto 20,81 PB nutzbare Kapazität übrig.

Das System kann flexibel erweitert werden. So können an den aktuellen Standorten weitere Systeme installiert werden, es können aber genauso weitere Standorte eingerichtet und in den Gesamtverbund integriert werden. Somit besteht auch für weitere Hochschulen die Möglichkeit, sich an dem Speicherverbund zu beteiligen. Neu hinzukommende Standorte können in bestehende Replikationsgruppen integriert werden. Ebenso können aber auch neue Replikationsgruppen angelegt werden.

Natürlich ist mit dem beschafften System nicht die Quadratur des Kreises gelungen, es konnten nicht alle Anforderungsdimensionen gleichermaßen erfüllt werden. Zwar lässt sich das Speichersystem durchaus als kosteneffizient beschreiben, bei Verzicht auf einige Anforderungen hätten die Kosten jedoch verringert werden können. Die Verbindung von hohem Datenvolumen, hoher Performanz und hoher Datensicherheit wird durch den hauptsächlichen Einsatz von objektbasiertem Speicher erreicht. Dieser hat gegenüber dateisystembasiertem Speicher einen deutlich reduzierten Metadaten-Overhead, so dass die Synchronisation in einem verteilten System wesentlich effizienter ist. Dieser Speicher kann direkt über das S3-Protokoll genutzt werden. Da aber noch nicht alle Softwarelösungen mit diesem Protokoll arbeiten können, sondern das Vorliegen der Daten in einem lokalen Dateisystem erfordern, ist es ggfs. nötig, Daten zunächst herunterzuladen. Alternativ ist es möglich, den Speicher über die Isilon-Systeme und die „Fileserver“-Protokolle SMB/CIFS und NFS als Netzlaufwerke lokal einzubinden. Allerdings verdrängen diese die Daten in einem proprietären Format in den Objektspeicher. Das bedeutet, Daten, die über die Isilon Systeme eingeliefert werden, können nur über diese auch wieder abgerufen werden, nicht jedoch direkt über das S3-Protokoll.

Die Zusammenarbeit innerhalb des Konsortiums war von Beginn der Antragserstellung, über die Ausschreibung bis hin zur Inbetriebnahme eng. So wurde in allen Phasen gemeinsam Wissen aufgebaut. Dies war wichtig, da keine zusätzlichen Personalressourcen für den Betrieb des Speichersystems beantragt wurden. Durch die Kooperation erreichte das Team jedoch auch ohne zusätzliches Personal eine effektive Größe. Mehrere Personen verfügen über das gleiche Wissen und die gleiche Expertise. So kann im Fall von Personalwechseln Kontinuität gewährleistet werden und es können auch zeitweise Ausfälle kompensiert werden.

Während der Inbetriebnahme und Konfiguration des Systems war die Zusammenarbeit sehr intensiv. Mit dem Abschluss der Projektphase zum 28.02.2021 und der Überführung des Systems in den Regelbetrieb wurde auch das Betriebsgremium neu organisiert und durch eine Geschäftsordnung formalisiert. Lag die zentrale Koordination während des Projekts beim Konsortialführer, gilt nun eine Regelung mit einer halbjährlich zwischen den Konsortiumsmitgliedern wechselnden Leitung.

4 FDM-Konzept

Die Ausgangslage an den Hochschulen des Konsortiums war auf mehreren Ebenen sehr unterschiedlich.

Die zentralen Unterstützungsservices waren unterschiedlich umfangreich und befanden sich teilweise noch im Aufbau. Dies bezieht sich sowohl auf IT-Infrastruktur, die den FAIRen Umgang mit Forschungsdaten unterstützt und von zentralen Rechenzentren bereitgestellt werden, als auch auf Angebote zur Beratung und Weiterbildung.

Die Bedarfe an Speicher variierten stark: Es gibt viele einzelne Projekte, die Bedarf an Speicherplatz für dezidierte Daten und einen dezidierten Zeitraum haben. Es gibt forschende Einrichtungen, die eine einrichtungsweite Lösung suchen, um Daten zu sichern und für die Nutzung in verschiedenen Projekten bereitzustellen. Ebenso gibt es zentrale Großgeräte, die sehr große Mengen an Daten erzeugen und diese wiederum an weitere Forschende zur Analyse verteilen. Und es gibt Verbundforschungsprojekte, die eine einheitliche Plattform für das Management ihrer Daten suchen, wobei die Daten selbst durchaus vielfältig sein können. An den Hochschulen des Konsortiums sind all diese Varianten vorhanden, allerdings mit unterschiedlichen Schwerpunkten. Entsprechend waren auch die Zielsetzungen für den FDS sehr unterschiedlich.

Schließlich unterscheidet sich auch die Struktur der IT-Versorgung. Zentrale Rechenzentren, Fakultäten, Fachgruppen und einzelne Institute spielen bei der Bereitstellung der IT-Infrastruktur eine unterschiedliche und unterschiedlich große Rolle. So bieten einige Rechenzentren stark standardisierte Dienste an, während andere intensiv auf die Entwicklung individueller Lösungen für einzelne Projekte, Fakultäten oder Institute setzen. Auch der Finanzierungsmodus der zentralen Services, zwischen zentral vorfinanziert und abrechnungsbasiert, unterscheidet sich, wobei es natürlich auch hier Mischformen gibt. Selbstverständlich unterscheiden sich auch die bereits eingesetzten Softwareplattformen.

Die Heterogenität der Ausgangslage hatte zur Folge, dass die Einbettung des FDS in ein einheitliches FDM-Konzept nicht trivial war.

Entsprechend wurde bereits frühzeitig parallel zur Abstimmung auf der Betriebsebene eine regelmäßig tagende Runde eingerichtet, die sich um die Prozesse rund um die Bewirtschaftung und Nutzung des FDS kümmerte.

Der Forschungsdatenspeicher hat das Forschungsdatenmanagement auf zwei Ebenen verändert: Er hat zu einem Paradigmenwechsel in der Speicherversorgung geführt und dient als Rückgrat einer Anwendungslandschaft zur Sicherung, Dokumentation und Analyse von Forschungsdaten.

Die Speicherversorgung vor der Einführung des FDS an den beteiligten Hochschulen lässt sich grob in zwei Säulen einteilen: Zum einen gab es zentrale Archivsysteme, auf die Forschungsdaten nach dem Projektende zur Aufbewahrung verschoben wurden. „Lebende“ Forschungsdaten wurden dagegen meist auf Fileservern gespeichert. Diese standen entweder in den zentralen Rechenzentren (selten) oder aber wurden von den einzelnen Einrichtungen in Eigenregie betrieben (häufig).

Die Einführung des FDS stellt demgegenüber einen Paradigmenwechsel auf verschiedenen Ebenen dar. Die dezentral betriebenen Fileserver werden durch ein konsortial betriebenes georedundantes System ersetzt. Damit gehen Effizienzgewinne einher, da die Anzahl der Personen, die insgesamt mit dem Betrieb von Speichersystemen befasst sind, verringert

werden kann. Angesichts der Herausforderungen der Digitalisierung auch im Hochschulbereich und der Knappheit von IT-Fachpersonal bedroht diese Effizienzsteigerung keine Arbeitsplätze.

Fileserver nahmen bisher typischerweise Daten aus allen Bereichen auf (all-purpose). Demgegenüber dient der FDS explizit und ausschließlich dem Management von Forschungsdaten (one-purpose). Dies unterstreicht die Besonderheiten von Forschungsdaten und die damit einhergehenden Anforderungen an den Umgang mit ihnen, wie sie von den FAIR-Prinzipien beschrieben werden.

Risiken wie Datenverlust oder der unbefugte Zugriff auf Daten (data breach) mussten bisher von jeder einzelnen Einrichtung, die einen Fileserver betrieben hat, gemanaged werden. Ebenso musste die langfristige Verfügbarkeit über mehrere Systemlebenszyklen dezentral gesichert werden. Mit dem FDS wird die Verantwortung für das Management dieser Risiken zentral vom Betreiberkonsortium wahrgenommen. Insbesondere die langfristige Verfügbarkeit kann durch das objektorientierte und georedundante System sehr viel besser und einfacher sichergestellt werden.

Bisher gab es für die Hochschulen ein „Dunkelfeld Speicher“. Über die Anzahl der betriebenen Fileserver und die auf ihnen gespeicherten Daten lagen zentral kaum Informationen, geschweige denn ein Überblick vor. Der FDS als zentrales Speichersystem, sofern es lokale Fileserver ersetzen kann, bietet dagegen die Chance, genau diesen Überblick über die gespeicherten Daten zu gewinnen. Da es auch weiter dezentrale Server geben wird, wird dieser Überblick jedoch nie vollständig sein. Dennoch wird der FDS eine wesentlich bessere Grundlage für die Abschätzung zukünftiger Speicherplatzbedarfe bieten.

Zwischen den Fileservern, die der Speicherung von Daten dienen, die aktiv verarbeitet werden, und den Archivsystemen bestand bisher ein Medienbruch. Zur Archivierung war die Migration der Daten auf ein komplett anderes System nötig. Während die bewusste Gestaltung dieses Übergangs durchaus im Sinne des FDM ist, war die Praxis bisher eher unbefriedigend: Zum einen fand der Transfer von Daten ins Archiv häufig vermutlich gar nicht statt. Zumindest lässt der Vergleich der Datenvolumina in den Bereichen Backup und Archiv den Schluss zu, dass nur ein Bruchteil der Daten, die als backup-würdig betrachtet werden, zu irgendeinem Zeitpunkt den Weg ins Archiv finden. Zum anderen fehlte häufig die Anwendungsunterstützung, um sicherzustellen, dass Daten bei der Archivierung mit aussagekräftigen Metadaten versehen sind. Zudem war die Motivation, Metadaten nach Abschluss der eigentlichen Forschungsarbeit noch aufwendig zu erfassen, eher gering.

Der FDS bietet dagegen die technische Grundlage dafür, die Lücke zwischen lebenden Daten und archivierten Daten zu schließen. Er bietet sowohl die Performanz als auch die nötigen Volumina um Daten zu speichern, die neu erzeugt, bearbeitet und analysiert werden, als auch die nötige Persistenz um diese langfristig zu sichern. Die Anwendungsebene muss es ermöglichen, Metadaten so früh wie möglich zu erfassen. Die Archivierung ist dann ein Vorgang, der sich vor allem auf der Ebene der Metadaten und der Policy (read-only) realisiert. Gleichzeitig kann FDS auch als Speicher für veröffentlichte Daten und somit als Backend für Repositorien dienen. Abbildung 3 zeigt das Schließen dieser Lücke schematisch im Domänenmodell des FDM.

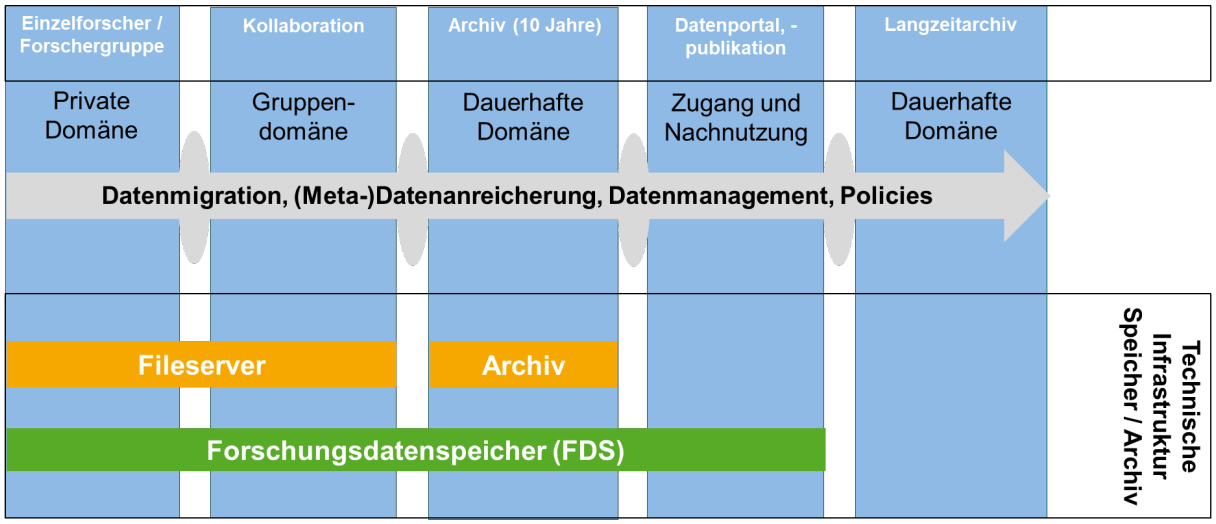


Abbildung 3: Verortung des FDS im Domänenmodell.

Selbstverständlich findet der hier idealtypisch beschriebene Paradigmenwechsel in der Speicherversorgung nicht von heute auf morgen vollständig statt. Vielmehr ist der FDS ein Angebot, dass zunächst bekannt gemacht und dann von den Forschenden angenommen werden muss. Somit vollzieht sich der Wandel nur langsam und schrittweise. Er wird vermutlich auch nie vollständig sein, sondern es werden immer auch Forschungsdaten auf dezentralen Fileservern gespeichert werden. Dennoch verbindet sich mit dem FDS die Hoffnung, über die Infrastruktur den alltäglichen Umgang mit Forschungsdaten und die Kultur in der Forschung im Sinne des FDM beeinflussen zu können.

Die Akzeptanz des FDS hängt in entscheidendem Maße von seiner Einbettung in bestehende oder neu zu entwickelnde Anwendungslandschaften ab.

Bereits im Antrag waren als Ziele festgeschrieben, dass die Vergabe der Ressourcen nach einem wissenschaftsgeleiteten Verfahren erfolgen sollte, dass der Speicher ausschließlich für Forschungsdaten zu nutzen ist und dass das System über definierte Schnittstellen zu vorhandenen Anwendungen verfügen sollte. Ergänzt wurden diese durch die Festlegung auf ein einheitliches Reporting über die Nutzung der Speicherressourcen an allen Hochschulen.

Aus diesen Zielen ergibt sich, dass die Integration mit weiteren Anwendungen auf der Ebene der Beantragung und Provisionierung von Speicherressourcen und auf der Ebene der eigentlichen Nutzung des Speichers erfolgen muss. Das Reporting kombiniert Informationen beider Ebenen: Metainformationen zu den antragstellenden Projekten, wie die beteiligten Hochschulen und die Verortung in der DFG-Fachsystematik, und quantitative Daten zur Speichernutzung, die direkt aus den Administrationsinterfaces des FDS bezogen werden können.

Das Konsortium hat sich darauf geeinigt, dass die Beantragung von Speicherplatz an ein Konzept für das Management der zu speichernden Daten geknüpft ist. Die konkrete Realisierung bleibt dabei den einzelnen Hochschulen überlassen. So ist es möglich, diesen Prozess über die Begutachtung von Datenmanagementplänen zu realisieren, über direkte

Gespräche mit den Forschenden oder über ein automatisiertes Verfahren. Die konkrete Gestaltung variiert je nach den vorhandenen Anwendungen und den hauptsächlich adressierten Use Cases. Während bei der Betreuung von wenigen Großprojekten eine individuelle Betreuung sinnvoll und möglich ist, empfiehlt sich für die Versorgung vieler kleinerer Projekte ein stärker automatisiertes Vorgehen.

In Aachen wird diese Aufgabe beispielsweise von der dort entwickelten Integrationsplattform Coscine übernommen [2]. Sie ermöglicht es zunächst, Projekte zu definieren und die beteiligten Forschenden zur Projektgruppe hinzuzufügen, über die dann auch Berechtigungen verwaltet werden können. Der Speicher wird dann im Rahmen unterschiedlicher Ressourcentypen bereitgestellt. So kann der Speicher direkt über Coscine genutzt werden. Dabei wird sichergestellt, dass zu jedem gespeicherten Objekt Metadaten gemäß einem zuvor von den Forschenden ausgewählten bzw. definierten Schema eingeliefert werden. Die Daten werden automatisch mit PIDs versehen, so dass alle Anforderungen an FAIRe Daten erfüllt werden. Bei der Wahl dieser Nutzungsart erfolgt entsprechend keine weitere Prüfung des FDM-Konzepts. Soll der Speicher direkt über die Protokolle S3, SMB/CIFS bzw. NFS genutzt werden, muss dagegen ein DMP eingereicht werden.

Es besteht auch immer die Möglichkeit einer individuellen Beratung. Insbesondere bei großen und komplexen Projekten wird diese vom FDM-Team initiiert.

Die Möglichkeiten, das Speichersystem zu nutzen, unterscheiden sich zwischen den Standorten. Grundsätzlich ist der direkte Zugriff über die Protokolle S3, SMB/CIFS und NFS möglich. Dieser direkte Zugriff ist insbesondere im Rahmen von Projekten vorgesehen, in denen bereits eigene Softwarelösungen für die Dokumentation und Organisation der Daten vorhanden sind. Das Speichersystem lässt sich so auch in automatisierte Workflows einbinden. Bei dieser Nutzung müssen Forschende allerdings in einem DMP darlegen, wie sichergestellt wird, dass letztlich FAIRe Daten produziert werden.

Viele Forschende können allerdings nicht auf solche bereits etablierten Umgebungen zurückgreifen und verfügen nicht selbst über die notwendigen Programmierkenntnisse. Deswegen wird der Speicher in zentral angebotene Anwendungen integriert, die die Organisation der Daten und die Erfassung von Metadaten unterstützen.

In Aachen übernimmt die bereits erwähnte Plattform Coscine diese Funktion. Daten können hier über die Weboberfläche verwaltet werden. Die Beschreibung der Daten mit Metadaten erfolgt dabei gemäß zuvor definierten Schemata. Dabei kann es sich um bereits etablierte Schemata und Vokabulare, wie EngMeta oder DataCite, handeln, um die Abbildung von Normen oder um selbstdefinierte Schemata. Letztere werden in Zusammenarbeit mit dem FDM-Team erstellt. Dabei findet für Felder, die bereits in einem Standard existieren, ein Mapping auf die existierenden Standards statt. Die Schemata werden RDF-konform in OWL beschrieben und über SHACL in der Anwendung validiert. Daten und Metadaten können auch über eine API eingeliefert und abgerufen werden. Voraussetzung für die Einlieferung von Daten ist aber immer das Vorhandensein der verpflichtenden Metadaten. Auch direkt über S3 eingelieferte Daten können in der Plattform angezeigt und (nachträglich) mit Metadaten beschrieben werden.

An der Universität Duisburg-Essen wird Nextcloud als Software-Plattform in der Kollaborationsphase des FDM-Lebenszyklus eingesetzt, die FDS als Hintergrundspeicher anbindet und die Prozessunterstützung für Nutzende bereitstellt. Die Projektanmeldung und Verwaltung der FDS-Ressourcen soll über eine Schnittstelle zum Coscine-System ermöglicht werden. Alternativ besteht die Möglichkeit zur Nutzung der FDS-Ressourcen nach Durchführung eines Beratungsgespräches mit der Servicestelle RDS und der Vereinbarung eines Datenmanagementplanes.

Der FDS erweist sich so als sehr flexibler Baustein in den FDM-Konzepten der einzelnen Hochschulen. Die Schnittstellen stellen sicher, dass das System interoperabel und die Daten im Objektspeicher flexibel nachnutzbar sind.

5 Offenheit des Konsortiums

Das Konsortium agierte von Anfang an mit einer grundlegenden Offenheit für weitere Nutzende aus Hochschulen, die nicht selbst dem Konsortium angehören.

Angehörige konsortiumsexterner forschender Einrichtungen können natürlich im Rahmen von Kooperationsprojekten Zugang zu dem Speicher bekommen. Darüber hinaus ist aber auch die eigenständige Beantragung von Speicherressourcen durch Forschende an Hochschulen für Angewandte Forschung, Kunst- und Musikhochschulen in NRW, sowie im Rahmen von NFDI-Konsortien, an denen die Konsortialpartner beteiligt sind, vorgesehen. Dies kann ebenfalls über die Plattform Coscine erfolgen, in der ein Login per SSO für DFN-AAI-Angehörige möglich ist. Darüber wird die Zugehörigkeit von Nutzenden zuverlässig mitgeteilt. Für die jeweiligen Hochschulen muss dann in Coscine nur noch eine entsprechende Policy definiert werden, die die Nutzungsmöglichkeiten und Quotarestriktionen für ihre Angehörigen enthält.

Zudem besteht die Möglichkeit für weitere Hochschulen, sich mit eigener Hardware an dem Speichersystem zu beteiligen. Allerdings erfordert dies die Nutzung der gleichen Hardware wie sie bereits verwendet wird, konkret DELL ECS- und Isilon-Systeme. Die lokale Hardware kann dann in bestehende oder neu zu definierende Replikationsgruppen eingebunden werden, so dass die Daten georedundant gesichert werden.

6 Fazit

Das Speichersystem Forschungsdatenspeicher (FDS) stellt einen wichtigen Baustein bei der Entwicklung hin zu einem bewussten und nachhaltigen Umgang mit Forschungsdaten dar. Es wirkt transformierend auf die Art und Weise, wie Speicherplatz für Forschungsdaten zur Verfügung gestellt wird und bietet einen Kristallisationspunkt für die Entwicklung von Anwendungslandschaften für die Forschung.

Objektorientierter Speicher ist eine technologische Grundlage, die für die Ablage von Forschungsdaten hervorragend geeignet ist. Sie reduziert gegenüber klassischem Blockspeicher den Overhead an technischen Metadaten deutlich und ermöglicht so nahezu unbegrenzte Skalierung bei geringen Kosten. Durch die Isilon bietet des FDS aber auch die Möglichkeit, für bestimmte Anwendungsfälle Blockspeicher über die klassischen Fileserver-Protokolle zur Verfügung zu stellen.

Das gesamte Projekt über die Beantragung, Beschaffung, Installation bis hin zum Betrieb des Systems hat den Beteiligten großen Einsatz abverlangt und eine intensive Zusammenarbeit nötig gemacht. Gerade diese intensive Zusammenarbeit und der enge Austausch haben sich jedoch als Wert an sich erwiesen. Das stark zusammengewachsene Team konnte sehr viel technische Expertise aufbauen und durch den intensiven Austausch auch die FDM-Konzepte an den Standorten weiterentwickeln.

Acknowledgements

Research Data Storage (RDS) wurde unter der Fördernummer 124-4.06.05.08-139057 vom Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen finanziert.

Literaturverzeichnis

[1] Eifert, T., Claus, F., and A. Lopez. “Research Data Storage (RDS): Verteilte Speicherinfrastruktur für Forschungsdatenmanagement: Gemeinsamer Antrag (öffentliche Fassung) im DFG-Programm ”Großgeräte der Länder”: RWTH Aachen University (Konsortialführer), Fachhochschule Aachen, Ruhr-Universität Bochum, Technische Universität Dortmund, Universität Duisburg-Essen, Universität zu Köln” Veröffentlicht auf dem Publikationsserver der RWTH Aachen University, (2018): doi: https://doi.org/10.18154/RWTH-2021-04541.

[2] Politze, M., Claus, F., Brenger, B., Yazdi, M. A., Heinrichs, B., and A. Schwarz. “How to Manage IT Resources in Research Projects? Towards a Collaborative Scientific Integration Environment”. European Journal of Higher Education IT 2020-1. Paris, France.

Rechtliche Fragen bei der Nutzung von Abbildungen aus Open-Access-Publikationen

Lucia Sohmen[1] und Fabian Rack[2]

[1]Technische Informationsbibliothek, Hannover
[2]FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur

Die zunehmende Verfügbarkeit von Forschungsdaten eröffnet Forschenden neue Möglichkeiten, mit von Dritten erstellten Forschungsdaten zu arbeiten. Dieser Beitrag befasst sich mit der Frage, welche rechtlichen Rahmenbedingungen gelten, wenn diese nachgenutzten Forschungsdaten öffentlich verfügbar gemacht werden sollen. Im Speziellen geht der Artikel dabei auf Bildersuchmaschinen und das Veröffentlichen von Bildkorpora ein. Dabei wird dargestellt, dass es bei der öffentlichen Zugänglichmachung von unübersichtlichen Bildmengen keine hundertprozentige Sicherheit geben kann. Durch bestimmte Abwägungen und technische Mittel kann sich dieser aber angenähert werden.

1 Einleitung

Forschungsdaten sind ein wichtiger Teil der alltäglichen Forschungsarbeit. Aus diesem Grund gibt es schon seit längerer Zeit Bestreben, Forschungsdaten zugänglicher und auffindbarer zu machen, damit die Allgemeinheit von ihnen profitieren kann. Eines dieser Vorhaben ist die Nationale Forschungsdateninfrastruktur (NFDI). Die Deutsche Forschungsgemeinschaft (DFG) fördert seit 2020 Konsortien aus verschiedenen Fachbereichen, um Qualität und Verfügbarkeit von Forschungsdaten zu erhöhen.

Die Autoren dieses Artikels sind Teil von NFDI4Culture (das Konsortium für Forschungsdaten zu materiellen und immateriellen Kulturgütern). Dort befasst sich die Task Area „Overarching legal and technical infrastructure“ unter anderem mit den rechtlichen Rahmenbedingungen für das Erstellen, Bearbeiten und Veröffentlichen von Forschungsdaten. So hat der Arbeitsbereich einen Help Desk für rechtliche Fragestellungen eingerichtet. Dieser Legal Help Desk bei NFDI4Culture ist eine Anlaufstelle für Rechtsfragen von Forschenden und Institutionen im Umgang mit materiellen und immateriellen Kulturgütern. Er bietet Hilfe zu Fragen u.a. des Urheber-, Persönlichkeits- und Eigentumsrechts an. Der Help Desk soll für Anfragen von Forschenden und Institutionen als „Hilfe zur Selbsthilfe“ dienen – wenn er auch keine individuelle Rechtsberatung leisten kann (und darf).

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13726

Der folgende Artikel soll eine Übersicht über die rechtlichen Anforderungen beim Umgang mit Abbildungen geben und Kriterien herausarbeiten, die für spätere Anwendungsfälle nutzbar gemacht werden können. Im Kulturbereich sind Bilder und audiovisuelle Daten die häufigsten Typen von Forschungsdaten. In diesem Artikel liegt der Fokus auf Abbildungen jeglicher Art.

Das Teilen von Forschungsdaten bringt für die Forschungslandschaft im Allgemeinen große Vorteile mit, da nun auch Forschende teilhaben können, die bisher keinen Zugang zu den entsprechenden Daten hatten und andere sich das aufwändige Erstellen ersparen können, wie z.B. die Durchführung von Gensequenzierungen oder das Digitalisieren von Gemälden. Gleichzeitig werden auch völlig neue Möglichkeiten eröffnet, wenn Forschende nun eine viel größere Datenmenge zur Verfügung haben, weil sie Daten aus verschiedenen Quellen aggregieren können und damit ganz andere Forschungsfragen beantworten können.

Durch die Urheberrechtsreformen der letzten Jahre sind für die Nutzung geschützter Güter für die Wissenschaft einige neue Freiheiten hinzugekommen, unter anderem für das Text- und Datamining. Forschende dürfen nun ihnen zugängliche Daten zur Analyse gemäß den jüngst reformierten und am 7. Juni 2021 in Kraft getretenen §§ 44b, 60d UrhG nutzen. Diese jüngeren gesetzlichen Regelungen erleichtern das Erstellen und Auswerten solcher Datenkorpora. Wer nun die aggregierten Daten weitergeben oder uneingeschränkt öffentlich teilen möchte – also zum Beispiel über eine eigens entwickelte Suchmaschine oder mit der Veröffentlichung des gesamten Datenkorpus –, unterliegt hierbei einigen Restriktionen, die im Folgenden erläutert werden.

2 Problemstellung

Es gibt unterschiedliche Anwendungskonstellationen, in denen Abbildungen gespeichert und der Öffentlichkeit zugänglich gemacht werden können. Im Folgenden werden beispielhaft Anwendungsfälle vorgestellt, bei denen mit nachgenutzten Forschungsdaten gearbeitet wird, also solche Forschungsdaten, die von den Forschenden nicht selbst erstellt wurden, sondern von ihnen gesammelt und gespeichert wurden. Im weiteren Verlauf des Artikels wird besprochen, was beim Umgang mit den verschiedenen Anwendungsfällen beachtet werden muss.

Eine dieser Möglichkeiten sind Suchmaschinen, die Daten – oftmals Abbildungen – aus verschiedenen Quellen aggregieren und für andere durchsuchbar machen. Dazu gehören auch Portale wie die Deutsche Digitale Bibliothek, die Digitalisate von teilnehmenden Bibliotheken, Archiven und Museen präsentieren und sich dabei über Kooperationsverträge absichern.[1] Solche rechtlichen Übereinkünfte als Basis der Zugänglichmachung werden in diesem Artikel nicht behandelt. Stattdessen soll es hier um Suchmaschinen gehen, die automatisiert frei verfügbare Daten aus dem Web finden, analysieren und ihren Nutzern zur Verfügung stellen. Die wohl bekannteste Suchmaschine dieser Art ist Google mit seiner Bildersuche, die sich bekanntlich nicht auf Abbildungen aus der Forschung beschränkt,

[1] `https://pro.deutsche-digitale-bibliothek.de/teilnehmen`

sondern Abbildungen jeglicher Art in ihren Suchindex aufnimmt. Viziometrics[2] und NOA[3] sind Beispiele für Suchmaschinen, die spezifisch wissenschaftliche Abbildungen durchsuchbar machen. Diese Projekte extrahieren Abbildungen aus Artikeln mit einer Open-Access-Lizenz und erschließen sie mit verschiedenen Methoden, um sie einer größeren Anzahl von Nutzenden verfügbar zu machen.

Nachgenutzte Forschungsdaten können nicht nur über Suchmaschinen dargestellt werden. Im Kulturbereich gibt es durch den Hackathon Coding da Vinci inzwischen zahlreiche Projekte[4], die Mediendateien von Bibliotheken, Museen und Archiven verwenden, um daraus neue Anwendungen zu erschaffen. Beispiele für solche Projekte sind spielerische Anwendungen wie Memories, Quizze oder Mandalas, aber es gibt auch Visualisierungen, mit denen die Datensammlungen besser erkundet werden können, wie zum Beispiel VR- und AR-Anwendungen. Auch im naturwissenschaftlichen Bereich gibt es Beispiele für die Visualisierung von Forschungsdaten. Cristallography.net sammelt Daten über Kristallstrukturen, bei denen unter anderem Zeichnungen von Strukturen als Suchbegriff verwendet werden können. Die Webseite sammelt nur textuelle Daten und erstellt daraus automatisiert Abbildungen der Formeln. Andere Beispiele wären Webseiten, die statistische Daten aus verschiedenen Quellen aufbereiten. Besonders bekannt sind hier die Visualisierungen zu Daten über Sars-CoV-2, die unter anderem von ourworldindata.org zur Verfügung gestellt werden. Die Daten werden aus einer großen Anzahl von Quellen gesammelt und zusätzlich zur Visualisierung zum Download bereitgestellt[5].

Zusätzlich zu den dargestellten Anwendungszwecken gibt es viele Forschungsprojekte, die Datenkorpora erstellen, auswerten und die Ergebnisse veröffentlichen, ohne die Forschungsdaten selbst zu präsentieren. In der Regel nimmt die Erstellung eines geeigneten Datensets einen nicht unwesentlichen Teil eines solchen Projekts ein. Es müssen Datenquellen gefunden und evaluiert werden und die Daten selbst müssen heruntergeladen, bereinigt und in ein homogenes Format gebracht werden, um für eine Analyse nützlich zu sein. Für die Gesamtheit der Forschenden ist es also von Vorteil, wenn solche einmalig erstellten Korpora geteilt würden, um anderen die aufwendige Arbeit zu ersparen. Technisch könnte das gelöst werden, indem die Daten über ein entsprechendes Repositorium gebündelt zum Download verfügbar gemacht werden.

In all diesen Szenarien müssen sich die Forschenden fragen, ob das Präsentieren und Veröffentlichen der Forschungsdaten durch die rechtlichen Rahmenbedingungen erlaubt wird. Die Beschränkung auf die Verwendung von Daten aus dem Open-Access-Kontext kann dabei eine gewisse, aber keine absolute Sicherheit schaffen. Zunächst einmal greift man meist auf unübersichtliche Datenmengen zurück, bei denen nicht alle Dateien und ihr Rechtestatus einzeln überprüft werden können. Hier gibt es einige Konstellationen, die dazu führen könnten, dass die Nutzung der Daten rechtlich nicht gestattet ist. Denn die Daten, die aggregiert wurden, könnten auch schon aus unterschiedlichen Quellen kommen und wurden eventuell nicht von der veröffentlichenden Person original erstellt.

[2] `http://viziometrics.org/`
[3] http://noa.wp.hs-hannover.de
[4] `https://codingdavinci.de/de/projekte`
[5] `https://github.com/owid/covid-19-data/tree/master/public/data`

So ist es im Kontext von wissenschaftlichen Artikeln üblich, dass andere Artikel zitiert werden und auch deren Abbildungen wiederverwendet werden. Solange der Zitatzweck diese Verwendung zulässt, ist die Nutzung der zitierten Abbildung (oder allgemein des zitierten Gegenstandes) rechtlich erlaubt; sobald man sie aber aus ihrem Kontext löst, gilt das möglicherweise nicht mehr. Die Abbildungen könnten zudem aus einem Artikel stammen, der keine Open-Access-Lizenz hat. Andere Forschende verwenden auch Abbildungen zur Illustration, die sie an verschiedenen Quellen im Internet gefunden haben. Diese Abbildungen könnten unter einer Open-Access-Lizenz stehen. Es muss aber nicht zwingend die gleiche Lizenz wie die des Artikels sein. In manchen Fällen kommt es auch vor, dass Autoren Abbildungen verwenden, an denen sie keine Rechte hatten und die auch nicht in Open-Access-Artikeln verwendet werden durften, wie zum Beispiel Abbildungen vom Bilderdienst Getty Images. Und selbst wenn die Veröffentlichung dieser Bilder erlaubt war, so ist die Kennzeichnung oft falsch: Häufig finden sich Quellenangaben wie „Google Images“, „Wikimedia Commons“ oder die Angabe fehlt vollkommen.

Forschende sind in der Regel keine Urheberrechtsexperten, weshalb oft Unsicherheiten bezüglich der Anwendung der bestehenden Regelungen bestehen. Wird – auch unwissentlich – gegen das Gesetz verstoßen, so kann dies Konsequenzen für sowohl die Forschenden selbst als auch die Universitäten und Forschungsinstitute, haben. Schadenersatzforderungen und rechtliche Auseinandersetzungen wollen freilich vermieden werden. In manchen Fällen sind die Forderungen der Abmahner nicht einmal gerechtfertigt. Jedoch ist es gerade für kleine Institutionen oft einfacher, einer solchen Forderung nachzukommen, als sich auf einen Rechtsstreit einzulassen [1]. Verständlicherweise möchten Forschende nicht für einen solchen Schaden ihrer Arbeitgeber verantwortlich sein. Zusätzlich könnte aus der Verfolgung von Rechtsverletzungen die Konsequenz gezogen werden, das gesamte Projekt offline zu nehmen. Dadurch drängt sich der Gedanke auf, dass es am sichersten ist, gar keine nicht selbst erstellten Forschungsdaten jedweder Art zu veröffentlichen. Durch die nachfolgende Darstellung sollen solche Ängste vermindert werden, indem die bestehende Rechtslage erläutert wird, um bestehende Freiräume auszuschöpfen.

3 Einschränkungen bei Abbildungen aus Open-Access-Publikationen

Die bereits erwähnte Bildersuchmaschine NOA enthält ausschließlich Abbildungen aus Open-Access-Publikationen. Ist nun wegen des Zuschnitts auf Open-Access-Publikationen eine unbeschränkte Nutzung von Abbildungen aus OA-Publikationen möglich? Um diese Frage zu beantworten, müssen wir zunächst den Begriff Open Access näher beleuchten. Hierzu ziehen wir die Definition von Open Access aus der Berlin Declaration aus dem Jahr 2003 heran [2]. Dort sind wesentliche Bedingungen für OA formuliert – hier ausschnittsweise:

> „The author(s) and right holder(s) of such contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy, use,

> distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship (community standards, will continue to provide the mechanism for enforcement of proper attribution and responsible use of the published work, as they do now)".

Nach dieser Definition bedeutet „Open Access" also nicht nur, dass Publikationen uneingeschränkt im Netz zugänglich und abrufbar sind. Vielmehr muss auch die Nachnutzung der OA-gestellten Publikation zulässig sein, und zwar, wenn auch nicht bedingungslos, zu jedem Zweck.

Teilweise müssen die Bedingungen freier Nachnutzung aktiv hergestellt werden:

- Bei urheberrechtlich geschützten Materialien müssen die Bedingungen von OA mit Hilfe von Lizenzen realisiert werden. Eine Lizenz ist ein Vertrag (in der Regel zwischen zwei Personen), der festlegt, dass und unter welchen Bedingungen die Nutzung eines urheberrechtlich geschützten Werkes erfolgen darf. Die OA-Bedingungen werden in der Wissenschaft häufig mit Standardlizenzen hergestellt. Am bekanntesten ist hier das Lizenzmodell von Creative Commons.[6] Hier erfüllen die Lizenzen CC BY und CC BY-SA die OA-Kriterien.[7]
- Bei urheberrechtlich nicht geschützten (gemeinfreien) Materialien muss – außer dass man den freien Zugang ermöglicht – nichts weiter unternommen werden. Denn es wäre sinnwidrig, durch eine Lizenz eine Nachnutzung zu erlauben, die auch ohne Lizenz ohnehin schon erlaubt ist. Hier ist es aus Nutzendenperspektive gewiss hilfreich, wenn derartiges Material entsprechend gekennzeichnet ist, weil so transparent zu erkennen ist, dass die Materialien nicht urheberrechtlich geschützt sind. Dies wird etwa durch die Public Domain Mark von Creative Commons oder den Rights Statements von rightsstatements.org umgesetzt.

Bei OA-Textpublikationen wie Aufsätzen wird man nun pauschal davon ausgehen können, dass ihr Text auch urheberrechtlich als Werk geschützt ist; der Text einer kunsthistorischen Abhandlung über eine Kinofilmepoche steht also gewiss unter Urheberrechtsschutz. Steht ein solcher Text nun unter keiner freien Lizenz, gelten die Restriktionen des Urheberrechts mit dem Grundsatz „Alle Rechte vorbehalten".[8]

Die Beschränkung auf Publikationen unter Open-Access-Lizenzen allein schafft für die darin enthaltenen Abbildungen allerdings noch keine Sicherheit. Denn eine Publikation kann unter OA-Bedingungen veröffentlicht sein, dabei aber zugleich Fremdabbildungen enthalten, die zum Beleg einer Aussage im Text oder der geistigen Auseinandersetzung mit

[6] Es gibt aber auch weitere Lizenzen wie das Modell der DDPL (Digital Peer Publishing Lizenz), das – anders als die CC-Lizenzen – zwischen der elektronischen und der Offline-Nutzung unterscheidet (d.h. zwischen E-Journalen und gedruckten Journalen).

[7] Die restriktiveren Varianten erlauben die Werknutzung nur für nicht-kommerzielle Zwecke (NC, steht für „non commercial") und/oder unter der Bedingung, dass keine veränderten Versionen des Materials geteilt werden (ND, steht für „no derivatives").

[8] Auch dann gelten gewisse Freiheiten, nämlich insbesondere die urheberrechtlichen Schranken, die wir weiter unten streifen.

einer Abbildung, gedeckt durch § 51 UrhG, *zitiert* werden. Für die Abbildung gilt die freie Lizenz dann nicht (außer, die Abbildung ist ihrerseits frei lizenziert oder gemeinfrei gewesen). Die erwähnte Abhandlung über den Kinofilm kann also als Textpublikation Open Access und damit also frei nachnutzbar sein, während es die enthaltenen Abbildungen nicht sind. Eine solche Publikation mit den zitierten Abbildungen als Ganzes zu speichern und weiterzugeben, ist unbedenklich möglich.[9] Entreißt man allerdings die Abbildung ihrem Kontext, greift die Zitierfreiheit nicht mehr. Die Weitergabe der Abbildungen wäre folglich nur im Kontext des Zitats erlaubt, weil ansonsten der nötige inhaltliche Zusammenhang von zitiertem und zitierenden Inhalt aufgelöst würde. Ein Beispiel: Zitiert eine Abhandlung über Kinofilme geschützte Filmstills aus dem Kino der 1980er-Jahre, sind die extrahierten (!) Filmstills nicht beliebig nachnutzbar.[10]

Lässt sich nun automatisiert erkennen, wenn Abbildungen zitiert werden und damit nicht unter die Lizenz des Artikels fallen? Diese Frage stellt sich, wenn Publikationen auf Abbildungen gecrawlt und diese nicht einzeln entnommen werden. Ansätze hierfür existieren durchaus: Häufig lässt sich beobachten, dass Zitate anhand der Bildunterschrift als solche zu erkennen sind. Ob das der Fall ist, kann stichprobenartig überprüft werden. Dadurch eröffnet sich die Möglichkeit, in den Bildunterschriften automatisiert nach bestimmten Schlüsselwörtern zu suchen, die eine fremde Quelle anzeigen. Beispiele dafür sind Wörter wie „Quelle“, „Foto“ oder die Namen oft verwendeter Bilddatenbanken wie Wikimedia Commons oder Getty Images. So kann erkannt werden, ob es sich um Fremdabbildungen handelt.

Die Organisation Creative Commons bietet seine Standardlizenzen auch in maschinenlesbarem Code an.[11] Damit ist es aus Perspektive der Nutzenden möglich, einzeln unter OA-Bedingungen lizenzierte Abbildungen in einem Artikel maschinenlesbar als frei lizenziert zu erkennen. Von der Einbindung der Codes für die Maschinenlesbarkeit der Lizenzen wird im Kontext von OA-Publikationen nach unserer Beobachtung jedoch leider nur selten Gebrauch gemacht, weshalb es sich derzeit nicht lohnt, danach zu suchen. Die Angaben zur Urheberin oder zum Urheber sind zudem beim Erstellen des maschinenlesbaren Codes optional, obwohl man diese Informationen für eine lizenzgemäße Benutzung zwingend braucht.

Als Zwischenfazit stellen wir fest: Aus Open-Access-Publikationen losgelöste Abbildungen sind nicht stets auch uneingeschränkt nutzbar. Bei extrahierten Abbildungen haben wir es mit heterogenem Material zu tun, dessen Rechtestatus nicht ohne einigen Aufwand ersichtlich oder gar nicht erkennbar ist. Abbildungen aus OA-Publikationen sind nur dann

[9] Ein Restrisiko, dass die Publikation das Zitatrecht missachtet und es nicht zur Anwendung kommt, lässt sich nicht ausschließen. Man wird sich also nicht mit absoluter Gewissheit darauf verlassen können, dass der vorgefundene Inhalt keine Rechte Dritter verletzt. Das ist kein Spezifikum dieser Konstellation: Auch jede individuell erworbene Lizenz bei Drittinhalten kann ins Leere gehen.

[10] Manchmal ist zu Beginn einer Publikation vermerkt, dass zwar der gesamte Beitrag frei lizenziert ist, die Abbildungen aber nicht.

[11] Der License Chooser von Creative Commons (`https://creativecommons.org/choose/?lang=de`) gibt den hierfür nutzbaren HTML-Code aus. Weitere Informationen hierzu unter: `https://wiki.creativecommons.org/wiki/Marking_your_work_with_a_CC_license#Author.2C_License.2C_Machine-readability` (abgerufen am 16. Mai 2021).

Open Content, wenn sie selbst unter die OA-Lizenz fallen, oder wenn sie ihrerseits gemeinfrei sind. Auf die Frage, wann Abbildungen gemeinfrei – also nicht nach den Vorschriften des Urheberrechts geschützt – sind, soll im Folgenden kurz eingegangen werden.

4 Der Schutz von Abbildungen

Bisher sind wir in unserem Beitrag von Abbildungen ausgegangen, die nach den Vorschriften des Urheberrechts geschützt sind und deshalb Restriktionen für die Nachnutzung unterliegen. Dies gilt freilich nicht für alle (Arten von) Abbildungen; hierzu sei auf die einschlägige juristische Fachliteratur verwiesen, z.B. [3][4][5][6]. Im Folgenden werden verschiedene Arten von Abbildungen vorgestellt und erläutert, inwiefern sie unter das Urheberrecht fallen.

Lichtbilder und Lichtbildwerke. Viele Fotografien sind geschützt, und zwar unabhängig von dem Aufwand ihrer Erstellung. Entweder handelt es sich um Lichtbildwerke (§ 2 Abs. 1 Nr. 5 UrhG, v.a. im künstlerischen Bereich), wobei Voraussetzung ist, dass sie einen gewissen künstlerischen Ausdruck haben (Originalität, „persönliche geistige Schöpfung“), oder es handelt sich um Lichtbilder (§ 72 UrhG). In zeitlicher Hinsicht macht dies einen Unterschied, da Lichtbilder 50 Jahre nach ihrer Erstellung bzw. Veröffentlichung, Lichtbildwerke bis 70 Jahre nach dem Todesjahr der Urheber:in geschützt sind. Auch Filmstills sind Lichtbilder oder Lichtbildwerke.

Ähnlich wie Lichtbilder hergestellte Erzeugnisse. Ähnlich wie Lichtbilder hergestellt sind Erzeugnisse, die unter Benutzung strahlender Energie entstanden sind, also bspw. durch Infrarotstrahlen, Kernspin-/CT-Aufnahmen, Röntgenaufnahmen, oder durch spektroskopische Messverfahren. Auch sie sind gemäß § 72 UrhG für 50 Jahre geschützt.

Werkschutz. Weitere Formen des Werkschutzes betreffen den Inhalt von Abbildungen: Sofern eine Reproduktion ein geschütztes Motiv enthält, gilt dieser Schutz. Ein Beispiel wäre der Scan einer geschützten künstlerischen Fotografie (aber auch eines Lichtbildes). Auch geschützt sind originelle Darstellungen wissenschaftlicher oder technischer Art wie Piktogramme, Schaubilder, technische Zeichnungen (§ 2 Abs. 1 Nr. 7 UrhG).

In Abgrenzung zu den genannten Abbildungen ist auch vieles **gemeinfrei**, d.h. frei von Urheberrechten: einfachste Darstellungen ohne Schöpfungshöhe; Scans oder Kopien gemeinfreier Vorlagen sowie – seit 7. Juni 2021 – originalgetreue Reproduktionsfotos bei gemeinfreien Motiven [7]. Auch automatisiert erstellte Abbildungen sind – jedenfalls ohne künstlerischen oder sonst „schöpferischen“ Eingriff durch einen Menschen und wenn die Technik nicht nur Hilfsmittel ist – gemeinfrei.

Ein Algorithmus könnte nun darauf trainiert werden, einige dieser Merkmale über Bilderkennung zu beschreiben. Mit einem passenden Trainingsset könnte wahrscheinlich das Alter der Bilder ungefähr erkannt werden, allerdings sicherlich nicht auf das Jahr genau. Die Erkennung von Scans oder Reproduktionen einzelner Werke könnte auch trainiert werden. Schwieriger würde es, gleichzeitig zu erkennen, ob es sich dabei um gemeinfreie Werke

handelt. Auch auf fehlende Schöpfungshöhe könnte ein Algorithmus trainiert werden. Da deren Vorhandensein auch von menschlichen Entscheidern nicht unbedingt sicher angegeben werden kann, wären die Ergebnisse wahrscheinlich umstritten und nicht rechtssicher. Dazu kommt, dass zum Beispiel in der Kunst die Anforderungen an die Schöpfungshöhe sehr gering sind. Was aus der (zeitgenössischen) Kunst kommt, muss pragmatischerweise auch als unter Werkschutz fallend angesehen werden. Insgesamt könnten durch die beschriebenen Methoden schon einige urheberrechtsfreie Bilder erkannt werden. Bisher ist allerdings kein Algorithmus bekannt, der für diese Aufgaben eingesetzt wird, weshalb alle Überlegungen in der Richtung theoretischer Natur sind.

Gewisse Rechtezuordnungen gibt es schließlich im Hinblick auf den *Inhalt* von Abbildungen. Sofern Personen abgebildet sind, ist deren Recht am eigenen Bild (Persönlichkeitsrecht) berührt. Ob nun bei der Erstellung und Online-Stellung der Inhalte (hier: dem Foto) „alles richtig gemacht wurde“, wird man nur selten bis zur letzten Gewissheit überprüfen können.

Bisher haben wir also dargestellt, dass es bestimmte Mittel der Operationalisierung gibt, um Rechtezuordnungen automatisiert ausmachen zu können. Gleichzeitig stellen diese Mittel keine hundertprozentige Rechtssicherheit her.

5 Phasen der Nutzung

Die oben dargestellten Nutzungskonstellationen (Bildersuchmaschine, Teilen von Korpora) sind urheberrechtlich nur relevant, wenn dabei Nutzungshandlungen im Sinne des Urheberrechts vorgenommen werden. Hier sind allen voran das im Digitalkontext sehr häufige Herstellen von Kopien (die Vervielfältigung, § 16 UrhG) und die öffentliche Zugänglichmachung (§ 19a UrhG) zu nennen, die jeweils grundsätzlich einer Zustimmung der Rechteinhaber bedürfen. Jeder Download ist eine Vervielfältigung; jedes Online-Stellen ist eine Zugänglichmachung.

Die vorliegend dargestellten Konstellationen unterscheiden sich auch von Embedding oder Verlinkungen von Abbildungen, die rechtlich weniger restriktiv gehandhabt werden. Reine Verlinkungen sind in aller Regel unproblematisch; auch Embedding ist in gewissen Grenzen erlaubt. Der Grund hierfür liegt in dem Wertungsgedanken, dass hier die Kontrolle über den Inhalt bei der uploadenden Person liegt. Wir konzentrieren uns vorliegend auf die Konstellation, dass die Inhalte auf den eigenen Server geladen werden und die Kontrolle dann nicht mehr bei der Quelle liegt; dies ist in aller Regel ohne Lizenz nicht zulässig [8].

Bilddaten sammeln und auswerten

Für eigene (nicht kommerzielle) wissenschaftliche Forschung dürfen Abbildungen vervielfältigt werden (§ 60c UrhG). Sie wiederum online weiterzugeben, ist davon nicht gedeckt.

Zu Zwecken der automatisierten Analyse dürfen Abbildungen ebenfalls lokal kopiert und aufbereitet werden (§§ 44b, 60d UrhG, im Folgenden als „TDM-Schranken“ bezeichnet). Mit diesen Schranken erlaubt das Urheberrechtsgesetz nicht nur die Sammlung, sondern auch Aufbereitungshandlungen wie etwa die Normalisierung oder Strukturierung der Daten. Voraussetzung ist dabei, dass der Zweck verfolgt wird, die Daten auch tatsächlich einer Analyse zuzuführen. Mit diesen Erlaubnissen ist zwar sichergestellt, dass zu diesem Zweck die nicht-öffentliche Nutzung erfolgen darf, aber noch nicht, inwieweit die uneingeschränkt öffentliche Weitergabe erlaubt ist.

Der Vollständigkeit halber sei noch erwähnt, dass Gedächtnisinstitutionen zur Bestandserhaltung aus ihren Beständen digitalisieren und Digitalisate an Terminals zeigen dürfen, was wiederum einen Zugang für die Forschung ermöglicht – allerdings *ohne uneingeschränkte öffentliche Weitergabe* (§§ 60e, 60f UrhG).

Bilddaten online publizieren

Die Frage der Nutzungsfreiheiten ohne Zustimmung konzentriert sich in diesem Beitrag auf solche Nutzungshandlungen, an deren Ende die Abbildungen *öffentlich* weitergegeben werden können (primär durch Online-Stellen). Es ist genau die öffentliche, uneingeschränkte Weitergabe, bei der das Urheberrecht abgesehen von der Zitierfreiheit eher strikt ist.

Teilen von Bildkorpora

Die Weitergabe von geschützten und (!) nicht frei lizenzierten Bildkorpora ist durch die TDM-Schranken nicht abgedeckt. Denkbar ist es hier allerdings, durch folgende Schritte zur öffentlichen Weitergabe der Bildkorpora zu kommen:

- Operationalisiertes Erkennen mit möglichst hoher Präzision, ob es sich um OA-Abbildungen handelt;
- Sicherstellen, dass für nicht frei lizenzierte Abbildungen, die zwar lokal einer Analyse zugeführt werden durften, die aber nicht öffentlich geteilt werden dürfen, die Information über den Rechtestatus nicht verloren geht;
- Trennen der freien von unfreien Abbildungen, um nur erstere in den Bildkorpora zu behalten;
- Bei Schutz nach den Vorschriften des Urheberrechts: Entfernen des Schutzes durch Abstraktion, zum Beispiel mit Hilfe von Beschreibungen. Dabei werden nicht die Abbildungen selbst, sondern nur diese abstrakten Beschreibungen veröffentlicht. In der Fachdebatte wird von „Informationsreduktion“ gesprochen, bislang vor allem bei Textbeständen [9]. Entsprechend wäre es hier denkbar, Histogramme, Tonwerte etc. zu entnehmen, durch die allein sich das Bild nicht reproduzieren lässt (diese Information ist dann frei). Der weitere Nutzen dieser Abbildungen ist dadurch aber stark

eingeschränkt. Die Sammlung kommt dann für keine Projekte infrage, bei denen die Bilder einem Publikum zugänglich gemacht werden sollen. Stattdessen kommt für die Nachnutzung nur eine automatische Auswertung der Bilddaten infrage, bei der allerdings nicht die gleichen Informationen wie sonst aus den Bildern gezogen werden können und die Forschenden die Ergebnisse nicht selbst überprüfen können. Einige Forschungsfragen könnten dadurch trotzdem beantwortet werden, zum Beispiel welche Farbtöne in welchen Epochen besonders verwendet wurden oder wie hoch die durchschnittliche Auflösung bei Digitalisaten von bestimmten Werken ist. Da der Anwendungszweck jedoch so eingeschränkt ist, bleibt es fraglich, ob eine Veröffentlichung der reduzierten Information der Mühe wert ist.

Ebenfalls in der Wissenschaft diskutiert werden die Möglichkeiten einer zumindest eingeschränkten öffentlichen Weitergabe von Korpora, die noch geschützte Abbildungen enthalten [10]. Eingeschränkt meint hier meist, dass Korpora geteilt werden, entweder im Rahmen wissenschaftlicher Reviewprozesse, für die gemeinsame wissenschaftliche Forschung oder für die Anschlussforschung. Damit lassen sich folglich nur Teilöffentlichkeiten herstellen.

Bildersuchmaschine

Zum Betrieb von Bildersuchmaschinen hat die Rechtsprechung bereits vor einigen Jahren Kriterien herausgearbeitet. So hatte der Bundesgerichtshof im Jahr 2010 zu entscheiden, ob die Bildersuchmaschine von Google urheberrechtswidrig ist [11]. Hintergrund ist, dass Google für seine Bildersuche die Abbildungen jeweils auf eigenen Servern kopiert und diese Bilder dann in der Bildersuchmaschine angezeigt hatte. Für die Nutzungshandlungen schließt Google freilich keine Lizenzverträge mit den Websitebetreibern ab, um die Abbildungen zu nutzen. Dies wäre praktisch nicht durchführbar.

Für derartige Anwendungen greift keine urheberrechtliche Schranke; insbesondere sind solche Nutzungen nicht vom Zitatrecht gedeckt, weil die Bilder nicht etwa in erläuternde Gedanken eingebettet, sondern ohne inhaltlichen Zusammenhang gezeigt werden. Der BGH hielt diese Nutzungshandlungen dennoch für gerechtfertigt. Um dies zu begründen, behalf er sich in seiner grundlegenden Entscheidung zur Bildersuche im Jahr 2010 einer bemerkenswerten Konstruktion – sinngemäß: Alle, die Abbildungen ins Internet einstellen, müssen davon ausgehen, damit in einer Bildersuche aufgenommen zu werden; wer nicht per Opt-Out erklärt, nicht in der Bildersuche erscheinen zu wollen, willigt in diese Nutzung ein. Einerseits wird damit die Bildersuche als Element des Internets anerkannt, andererseits lässt sich nach der Sinnhaftigkeit eines ausdifferenzierten Schrankensystems fragen, wenn dort, wo keine Schranke existiert, mit einer Einwilligungskonstellation gearbeitet wird.

Was aber, wenn die Bilder ohne Zustimmung ins Netz gekommen sind? Hierfür sagen die Gerichte: Wer bestimmte Sorgfaltspflichten einhält, haftet nicht oder erst ab Kenntnis [12] [13]. Gleichzeitig steht fest, dass diese Rechtsprechung nicht zu einer allgemeinen Privilegierung des Modells Suchmaschinen für jegliche Nutzung urheberrechtlich geschützter Materialien führt [14].

Für wissenschaftliche Anwendungsfälle wie die NOA-Abbildungssuche können wir uns diese Überlegungen zunutze machen: Es ist Grundbedingung von Wissenschaft und Open Access im Internet, nachgenutzt zu werden, auch in Teilen wie hier in Bezug auf Abbildungen. Zumal, wenn dabei folgendes beachtet wird: Die Quelle der Abbildung wird verlinkt und mit angegeben, die jeweilige Lizenz wird – sofern vorhanden – angeheftet.

Dennoch scheint es zu weit gegriffen, dass eine Bildersuche ohne jegliche Restriktion durch die Vorschaubilder-Rechtsprechung legitimiert ist. Denn eine wesentliche argumentative Säule ist der Kontrolleinfluss der Uploadenden: Man geht davon aus, dass sie die Abbildungen selbst ins Netz stellen und technisch Einfluss darauf nehmen können, ob Bildersuchen sie übernehmen dürfen. Hieraus konstruiert die Rechtsprechung die durch schlüssiges Verhalten erteilte („konkludente“) Einwilligung, dass eine dem Netz so immanente Nutzung wie in einer Suchmaschine erfolgen darf. Im Zitierkontext ist dies aber gerade nicht der Fall: Dort hat die Rechteinhaberin der Abbildung gerade nicht Einfluss auf deren Verwendung. Ihre Möglichkeit, ein technisches Opt-Out zu setzen (etwa durch robots.txt), besteht nicht. Ihre Interessen an den Ausschließlichkeitsrechten wiegen *nur im Zitatkontext* weniger schwer als das Allgemeininteresse an der Nutzung. Daher ist es wichtig, den Zitatkontext zu erkennen und Abbildungen gegebenenfalls entsprechend herauszufiltern.

Zugleich hat die Rechtsprechung die Legitimität eines solchen Modells grundsätzlich anerkannt. Daher wäre es nicht berechtigt, der Wissenschaft das Schaffen einer solchen Plattform generell deshalb zu verwehren, weil – und dies ist auch hier eher gemutmaßt – unfreie Abbildungen in OA-Publikationen eine Schlagseite zum Zitat haben. Dies gilt gerade dann, wenn die hier dargestellten Vorsichtsmaßnahmen getroffen werden. In diesem Fall ist die Konstellation letztlich keine funktional andere als bei der Bildersuche. Leider ist dies rechtlich noch nicht abschließend geklärt.

Weitere Sorgfaltsmaßnahmen, beziehungsweise Bedingungen für die Übernahme, sind dann gegebenenfalls die Verkleinerung (Thumbnails), Verlinkung sowie Übernahme jeglicher Lizenzinformation und Identifikatoren.

6 Risiken rechtswidriger Nutzung

Werden Urheberrechte verletzt, weil die Nutzung rechtswidrig erfolgt ist, können Rechteinhaber verlangen, dass die Nutzung der Abbildung beendet („unterlassen“) wird; sie können Lizenznachzahlungen fordern und sich die Kosten für die rechtliche Verfolgung ersetzen lassen. Dabei kommt es nicht zwingend zu einem Gerichtsprozess. Stattdessen ist ein Vorgang vorgeschaltet, um den Streit außergerichtlich beizulegen (Abmahnung).

Das tatsächliche Verfolgungsrisiko dürfte bei öffentlich sichtbaren Nutzungen in der Regel höher sein als bei nicht öffentlichen.

Für eine anwaltliche Aufforderung, eine Nutzung zu unterlassen und eine Abmahnung zu versenden, entstehen Kosten. Es besteht nun ein Unterschied zwischen diesen „Rechtsverfolgungskosten“ und einer Lizenznachzahlung. Letztere basiert rechtlich auf einer Schadensersatzforderung, die stets voraussetzt, dass der Rechtsverletzer schuldhaft, also vorsätzlich oder zumindest fahrlässig, gehandelt hat. Die Höhe von Schadensersatzansprüche hängt vom Einzelfall ab, wird aber gerade in den hier geschilderten Nutzungskonstellationen eher überschaubar sein.

Hier kann man wieder mit den in diesem Beitrag genannten Sorgfaltsmaßnahmen ansetzen, die das Risiko einer Schadensersatzzahlung verringern.

7 Fazit

Für die öffentliche Weitergabe von Abbildungen aus OA-Publikationen gibt es Spielräume, für die sich mit den im Beitrag gezeigten Sorgfaltsmaßnahmen gewisse Freiheiten genießen lassen. Diese lauten: Operationalisierung zum Erkennen des Zustandekommens von Abbildungen (dies beurteilt den Schutzstatus), Übernahme der Lizenzstatus, insbesondere durch Erkennen der maschinenlesbar eingebundenen Lizenzen (sofern vorhanden), Sichern der jeweiligen Lizenzinformation und der Attribution in den Metadaten (sichert Attribution-Pflichten und liegt zudem im Sinne der Guten Wissenschaftlichen Praxis), Erkennung des Kontextes und Hinweisen auf Abbildungszitate. Eine endgültige Rechtssicherheit beim öffentlichen Teilen gibt es nicht, weil – auch für die Wissenschaft – keine Bereichsausnahmen von rechtlichen Anforderungen existieren. Dennoch wollen wir mit dem Appell schließen, Risiken nicht überzubewerten und Spielräume zu nutzen.

Acknowledgements

NFDI4Culture ist gefördert durch die Deutsche Forschungsgemeinschaft (DFG) unter der Fördernummer 441958017.

Literaturverzeichnis

[1] ComputerWeekly.com. „Automated Image Recognition: How Using ‘Free’ Photos on the Internet Can Lead to Lawsuits and Fines“. Zugegriffen 20. Mai 2021. https://www.computerweekly.com/news/252488167/Automated-image-recognition-How-using-free-photos-on-the-internet-can-lead-to-lawsuits-and-fines.

[2] „Berliner Erklärung“. Zugegriffen 20. Mai 2021. https://openaccess.mpg.de/Berliner-Erklaerung.

[3] Ilva Johanna Schiessel, *Reichweite und Rechtfertigung des einfachen Lichtbildschutzes gem. § 72 UrhG* (Nomos Verlagsgesellschaft mbH & Co. KG, 2020), https://doi.org/10.5771/9783748909620.

[4] Dreier/Schulze/Schulze, 6. Aufl. 2018, UrhG § 2 Rn. 222-242.

[5] Dreier/Schulze/Schulze, 6. Aufl. 2018, UrhG § 2 Rn. 189-203.

[6] Schricker/Loewenheim/Vogel, 6. Aufl. 2020, UrhG § 72 Rn. 1-95.

[7] Klimpel, Paul, und Fabian Rack. „Reproduktionen und urheberrechtlicher Schutz Vervielfältigung, Lichtbildschutz und Gemeinfreiheit – was gilt bisher, was wird gelten¿‘ *RuZ - Recht und Zugang* 1, Nr. 2 (2020): 243–57. https://doi.org/10.5771/2699-1284-2020-2-243.

[8] BGH, Entscheidung vom 10.01.2019, Az. I ZR 267/15 – *Cordoba II*.

[9] Raue, Benjamin, und Christof Schöch. „Zugang zu großen Textkorpora des 20. und 21. Jahrhunderts mit Hilfe abgeleiteter Textformate – Versöhnung von Urheberrecht und textbasierter Forschung“. *RuZ - Recht und Zugang* 1, Nr. 2 (2020): 118–27. https://doi.org/10.5771/2699-1284-2020-2-118.

[10] Kleinkopf, Felicitas, Janina Jacke, und Markus Gärtner. „Text-und Data-Mining: urheberrechtliche Grenzen der Nachnutzung wissenschaftlicher Korpora und ihre Bedeutung für die Digital Humanities“, 2021. https://doi.org/10.18419/opus-11445.

[11] BGH, Entscheidung vom 29.04.2010, Az. I ZR 69/08 – *Vorschaubilder*.

[12] BGH, Entscheidung vom 19.10.2011, Az. I ZR 140/10 – *Vorschaubilder II*.

[13] BGH, Entscheidung vom 21.09.2017, Az. I ZR 11/16 – *Vorschaubilder III*.

[14] Ohly, Ansgar. „Zwölf Thesen zur Einwilligung im Internet“. *GRUR*, 2012, 983.

Nicht-lineare Narrative in Netzliteratur: Speicherung und Nachnutzung von Forschungsdaten aus der computergestützten Extraktion von Verweisstrukturen in Hypertexten

Claus-Michael Schlesinger[2], Mona Ulrich[1],Pascal Hein[2], André Blessing[2], Nina Buck[3], Björn Schembera[3], Volodymyr Kushnarenko[3], Andreas Ganzenmüller[3], Lisa Kiss[2], Julia Horvat[2] und Oksana Nedostup[2]

[1]Deutsches Literaturarchiv Marbach
[2]Universität Stuttgart
[3]Höchstleistungsrechenzentrum Universität Stuttgart

Das Forschungs- und Infrastrukturprojekt *Science Data Center for Literature* entwickelt und implementiert ein Repository für born-digital-Bestände am Deutschen Literaturarchiv Marbach (DLA) und ein zugehöriges Portal mit Forschungsumgebung. Wir beschreiben einen exemplarischen Forschungsansatz zur Analyse nicht-linearer Strukturen in Netzliteratur, das Teilkorpus "Literatur im Netz" am Deutschen Literaturarchiv (Entwicklungsgrundlage), das Softwaremodul Warc2graph zur Extraktion von Verweisstrukturen aus den archivierten Objekten, die konzipierte Architektur der Plattform für SDC4Lit als Zielumgebung für das Modul sowie weitere Nachnutzungsmöglichkeiten.

1 Einleitung

Literarische Werke im WWW können von technischen und kulturellen Gegebenheiten des Mediums inspiriert sein. Sie zeichnen sich daher oft durch eine besondere Beziehung zwischen literarischem Text und technischem Medium aus. Neben der Bedeutung grafischer und typografischer Gestaltung zählt dazu insbesondere die Hypertextstruktur und Hypermedialität der Werke[5, 11].
Die Verteilung einer Erzählung auf mehrere miteinander verlinkte Webseiten bedingt eine nicht-lineare Struktur, die oft mit nicht-linearen narrativen Strukturen korrespondiert. Linearität und Nicht-Linearität sind dabei auf den sukzessiven Durchgang durch ein Werk im Zuge der Lektüre oder Interaktion bezogen. Zu unterscheiden sind dabei zwei unterschiedliche Perspektiven auf ein

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13727 (CC BY-SA 4.0)

Hypertextobjekt: Erstens eine Perspektive, die die Hypertextstrukturen ohne den durch Ein- und Ausgabegeräte sowie hypermediale und Interaktionsfunktionen vorgegebenen Verlauf betrachten, und zweitens eine Perspektive, die diese Aspekte mit einbezieht und die Struktur als Gesamtmenge aller möglichen Durchgänge durch ein Werk versteht, d.h. alle möglichen Durchgänge durch ein Werk von allen Einstiegs- zu allen Endpunkten berücksichtigt. Wir sehen eine nicht-lineare Struktur als gegeben, sobald eine Seite mehr als einen Verweis auf Folgeseiten enthält. Ein linearer Durchgang durch den Gesamttext ist dann nicht mehr möglich. Für narrative Texte folgt daraus auch ein zumindest in Teilen nicht-linearer Erzählverlauf. Nicht-linerare Textstrukturen können überwiegend lineare Erzählverläufe mit alternativen Strängen, variablen Enden und zyklischen Elementen oder durch komplexe Verlinkungen verschiedene Erzählverläufe ermöglichen.

Für eine narrative Analyse im engeren Sinn eignen sich nur Werke der Netzliteratur, die entsprechende narrative Eigenschaften aufweisen. Ein klassisches Beispiel sind hier Hyperfiction-Texte wie *Zeit für die Bombe* (1997) von Susanne Berkenheger. [1, 20] In *Zeit für die Bombe* sind Protagonist*innen, Ereignisse und also eine Geschichte im narratologischen Sinn[15] unzweifelhaft vorhanden. Andere Werke wie zum Beispiel Johannes Auers *Kill the Poem* (1997) orientieren sich dagegen eher an Formen der Konkreten Poesie und liefern keine erzählte Handlung, sondern einen Handlungsraum für spielerische ästhetische Ko-Produktion und Interaktion. Auf einer abstrakten Strukturebene ist die Verweisstruktur für alle Objekte im Bereich "Literatur im Netz" ein für die Analyse relevantes Merkmal, in dem sich technische Funktionen mit ästhetischen Eigenschaften verbinden.[5] Im Folgenden geht es dabei weniger um gegenstandsbezogene literaturwissenschaftliche Fragestellungen, sondern um die archivarischen, objektanalytischen und infrastrukturellen Kontexte und Voraussetzungen, in denen solche Fragen gestellt und bearbeitet werden können. Der Fokus liegt auf den Bereichen Archiv, Analyse und Infrastruktur, wie sie sich im Forschungs- und Infrastrukturprojekt *Science Data Center for Literature* (SDC4Lit) darstellen.[1]

In Abschnitt 2 beschreiben wir das Korpus "Literatur im Netz" des Deutschen Literaturarchivs, das die Materialgrundlage bildet für unsere Modellierung von Verweisstrukturen in archivierten Netzobjekten. Zweitens stellen wir mit `warc2graph` ein Softwarepaket vor, das dieses Modell implementiert und Verweisstrukturen aus archivierten Objekten extrahiert.[2]

Insbesondere für die Analyse von Einzelobjekten und kleineren Korpora ist ein hoher Grad an Genauigkeit wünschenswert. Aktuell verfügbare Ansätze zur Extraktion von Verweisstrukturen aus archivierten Netzobjekten und -korpora zielen auf die Analyse von großen Datenmengen und legen den Schwerpunkt daher auf höchstmögliche Effizienz. [13, 4] Warc2graph zielt dagegen auf höchstmögliche Ergebnisqualität (mit den entsprechenden Abstrichen bei der Performanz). Drittens skizzieren wir die Architektur der Forschungs-

[1] Siehe hierzu das Extended Abstract zum Poster von SDC4Lit in diesem Band

[2] Warc2graph ist als Python-Modul über den Python Package Index verfügbar, aktuelle Versionen werden außerdem im Github-Repository von Warc2graph bereitgestellt, `https://github.com/dla-marbach/warc2graph`. Das Softwarepaket in der zum Zeitpunkt der Veröffentlichung dieses Beitrags aktuellen Version 0.1.1 siehe auch [9]

umgebung, in der Warc2graph für die Unterstützung wissenschaftlicher Analysen der archivierten Gegenstände als Modul eingesetzt werden soll. In einem kurzen Ausblick nennen wir abschließend weitere Nachnutzungsmöglichkeiten, die sich aus unserer Sicht für das Modul anbieten.

2 Gegenstand Literatur im Netz

Unsere Forschungsfrage zu nicht-linearen narrativen Strukturen bezieht sich auf literarische Objekte, die im Web veröffentlicht sind und Verweise mittels HTML-Tags oder JavaScript-Funktionen beinhalten. Entwickelt wurde die Forschungsfrage an den archivierten Netzliteraturwerken der Sammlung Literatur im Netz des deutschen Literaturarchivs in Marbach (DLA). Zu der Sammlung gehören neben Netzliteraturwerken auch literarische Blogs und Online-Magazine. Die Archivierung der Quellen erfolgte von 2008 bis 2018. Die archivierten Quellen sind derzeit auf der Plattform *Literatur im Netz* zugänglich.[3] Die Netzliteraturwerke sind zwischen 1995 und 2011 entstanden. Der Großteil der Werke zeichnet sich durch gemeinsame Charakteristiken aus, die aber nicht zwangsläufig die Literaturgattung Netzliteratur an sich beschreiben. Dazu gehört, dass die Webseiten von den Autor*innen meist selbst geschrieben wurden und nicht auf vorgefertigten Templates basieren. Die Objektgrenzen der Werke sind meist klar definierbar und die Objekte beinhalten mehrere HTML-Dokumente, die untereinander verlinkt sind. Durch die Wahl des Mediums haben die Autor*innen die Möglichkeit, ihre Werke beliebig zu strukturieren, wodurch die manifestierten Entscheidungen bezüglich der Struktur bedeutungsvoll sind und daher ebenso wie die textuellen Inhalte bei der Analyse berücksichtigt werden müssen.

Auch für die literarischen Blogs und Onlinemagazine ist eine Analyse der Seitenstrukturen aufschlussreich, unabhängig von der Ausgangsfrage zu nicht-linearen narrativen Strukturen. Die Visualisierung der Seitenstruktur bietet Forscher*innen einen neuen Überblick über das Objekt. Denn anders als bei makrophysikalischen Objekten, zum Beispiel bei einem Buch oder einer Kunstinstallation, sind die Objektgrenzen und der Aufbau eines Netzobjekts nicht sichtbar. Auf der Startseite einer Website ist nicht ersichtlich ob sie zwei oder zweihundert Unterseiten enthält und wie die Seiten sich zueinander verhalten. Forscher*innen müssten sich diese Übersicht aufwendig erarbeiten - für manche Objekte eine fast unlösbare Aufgabe.

Die Strukturinformationen können im hier vorliegenden Anwendungsfall von archivierten Websites, die im WARC-Format[4] gespeichert sind, auch automatisch ausgelesen werden. Das WARC-Format wurde vom International Internet Preservation Consortium (IIPC) aufbauend auf das ARC-Format entwickelt. Das ARC-Format entstand 1996 am Internet Archive, um gecrawlte Webressourcen besser verwalten zu können.[5] In einer WARC-Datei

[3] [14]
[4] IIPC
[5] [16]

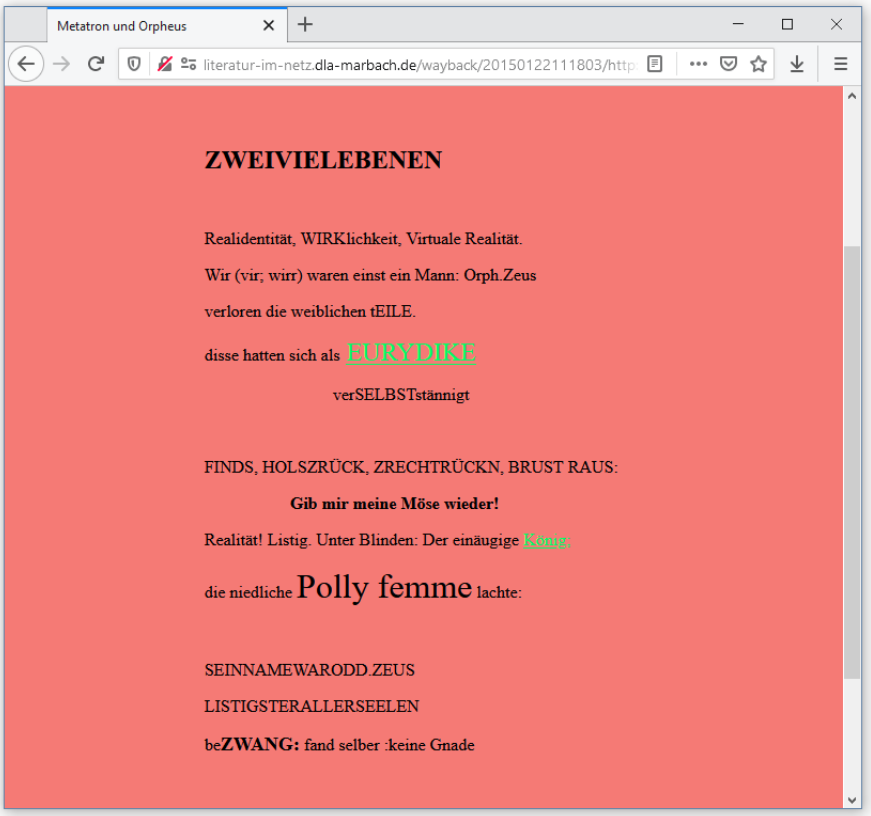


Abbildung 1: Kyon's Metapage, Kyon,1998, `https://metatrons.net/m4.html`, hier Archivversion, Screenshot.

wird die Steuerungskommunikation zwischen Client und Server gespeichert, sowie die vom Client (Webarchivierungstool, Browser) empfangenen Ressourcen.

Die Steuerungskommunikation und WARC-spezifische Metadaten sind als Text und die Ressourcen, sofern sie keine Textdateien sind, in Binärform enthalten. Die Ressourcen werden immer unverändert, das heißt so wie sie empfangen wurden, abgespeichert. Eine WARC-Datei kann mit speziellen Tools, zu nennen sind hier Pywb[6] und OpenWayback[7], wiedergegeben werden. Im Optimalfall wäre der einzige Unterschied zwischen einer originalen Seite im live web und ihrer archivierten Version die URL in der Adresszeile des Browsers.
Die literarischen Objekte am DLA wurden jeweils in eigenen WARC-Dateien gespeichert. Aus einer WARC-Datei können die Strukturinformationen zu einem Objekt mit unterschiedlichen Methoden ausgelesen werden, die teilweise ergänzend eingesetzt werden müssen. Das von uns entwickelte Tool Warc2graph beherrscht mehrere Methoden (einzeln und kombiniert) und ermöglicht dadurch einen Vergleich der Ergebnisse, die durch einzelne oder kombinierte Ansätze gewonnen werden.

[6]Webrecorder Project
[7]IIPC, OpenWayback, `https://github.com/iipc/openwayback`, aufgerufen: 27.04.2021.

3 Warc2graph

Mit Hilfe des Python-Pakets Warc2graph können im WARC-Format archivierte Websites[8] automatisiert ausgelesen und als Netzwerkgraph modelliert werden. Das Paket kann über den Python Package Index bezogen werden. Es stellt sowohl eine in Python importierbare Bibliothek als auch eine einfach zu bedienende Kommandozeilenanwendung bereit. Das Tool öffnet die WARC-Datei mithilfe der Python Bibliothek Warcio[9] und greift auf die Metadaten und die gespeicherten Ressourcen zu, um Verweise zwischen den Ressourcen zu finden, die durch HTML-Tags und deren Attribute bestimmt sind. Für den Netzwerkgraphen werden alle Ressourcen – HTML-, CSS-, Bilddateien und andere Medienformate – als Knoten gespeichert und alle Verweise von und auf die Ressourcen als Kanten. Die Knoten werden über die absolute URL der Ressourcen definiert. Die Kanten sind mit der Information darüber angereichert, welche Art von Verweis sie darstellen. Alle HTML-Tags und ihre Attribute, die URLs enthalten können, werden hier ausgelesen.

3.1 Anwendung

Die Kommandozeilenschnittstelle wird über den Befehl `warc2graph` aufgerufen. Zusätzlich muss der Pfad zu einer WARC-Datei als Parameter mit angegeben werden. Das Tool verarbeitet daraufhin die WARC-Datei und erstellt als Output drei Dateien. Dies ist erstens eine auf XML basierende GEXF-Datei[10], die die Graphdaten enthält. Zweitens werden Visualisierungen erstellt, die einen ersten Überblick zur Struktur der extrahierten Verweise liefern sollen[11]. Dabei werden jeweils drei Netzwerkdiagramme mit drei unterschiedlichen Visualisierungsalgorithmen erstellt, um zu vermeiden, dass ein einziges spezifisches Layout ungerechtfertigte Rückschlüsse auf die Struktur der Website motiviert. Die dritte Outputdatei beinhaltet Metadaten im JSON-Format. Hierbei handelt es sich sowohl um Metadaten, die die ursprüngliche archivierte Website beschreiben, als auch um Metadaten, die den Prozess der Erstellung des Graphen aus der WARC-Datei beschreiben und dabei neben Datum und Uhrzeit auch alle gewählten Parameter beinhalten. Die Metadaten werden automatisiert erstellt, können aber manuell beliebig ergänzt werden.

Alternativ können auch mehrere WARC-Dateien übergeben werden. Hierbei werden die Ergebnisse zusammengefasst und wie ein Archiv einer einzigen Website behandelt. Diese Funktion wird bereitgestellt, weil WARC-Dateien nicht größer als 1GB sein sollen und

[8] Hier und im Folgenden verwenden wir die Begriffe Website und Webpage gemäß der Definitionen des W3C. Eine Webpage ist demnach eine "collection of information, consisting of one or more Web resources, intended to be rendered simultaneously, and identified by a single URI" und eine Website ist definiert als "collection of interlinked Web pages, including a host page, residing at the same network location." (Lavoie und Nielsen 1999)

[9] Webrecorder Project; Webrecorder Project (2021)

[10] [7]

[11] Diese automatisch erstellten Visualisierungen dienen nur zum ersten Eindruck; soll eine Visualisierung erstellt werden, die nicht nur für die Exploration, sondern für die Verwendung als Demonstration und Argumentation gedacht ist, bedarf es einer weitergehenden begründeten Auswahl eines angemessenen Visualisierungsalgorithmus.

Webseiten daher bei der Archivierung auf mehrere WARC-Dateien aufgeteilt werden können. Neben der Analyse von WARC-Dateien können auch Links zu Webpages im live web übergeben und damit nicht-archivierte Websites analysiert werden. Wird das Tool in Form der importierten Python Bibliothek Warc2graph verwendet, eröffnen sich weitere Anwendungsmöglichkeiten. Die Bibliothek stellt die Funktion `create_model` zur Verfügung, die parallel zur Kommandozeilenanwendung einen Pfad zu einer WARC-Datei benötigt. Die Funktion gibt einen gerichteten Graphen zurück, der mithilfe der von NetworkX[12], einer Python-Bibliothek für Graphdaten- und Netzwerkanalyse implementiert ist. Der breite Funktionsumfang von NetworkX kann nun direkt für den erstellten Graphen genutzt werden, um zum Beispiel Zentralitätsmaße zu berechnen oder die Zirkularität und andere spezifische Eigenschaften des Graphen zu prüfen.

3.2 Funktionsweise

Die Modellierung einer Website als Netzwerkgraph mithilfe von Warc2graph läuft in zwei Schritten ab. Die Funktion `warc2graph.extract_links` verarbeitet die WARC -Dateien, liest sie aus und extrahiert daraus alle Verweise zwischen Ressourcen, während die Funktion `warc2graph.create_network` aus den extrahierten Informationen direktionale Netzwerke erstellt. Nutzt man das Python-Modul können beide Schritte unabhängig voneinander durchgeführt und die Daten nach jedem Zwischenschritt überprüft und manuell angepasst werden. Um die Verweise aus der WARC-Datei zu extrahieren wird über alle Einträge der WARC-Datei iteriert. Jede HTML-Datei wird dann mithilfe von drei möglichen Extraktionsmethoden analysiert.

1. Die einfachste Methode liest die in der WARC-Datei gespeicherten Metadaten aus. In den Metadaten können die beim Crawling durch das Webarchivierungstool gefundenen ausgehenden Links (Outlinks) einer Ressource vermerkt sein. Diese Methode ist sehr robust und performant, gleichzeitig aber am wenigsten flexibel. Einzelne Domains können zwar gefilter werden, aber was beim Crawling nicht gefunden wurde kann auch jetzt nicht mehr gefunden werden. Hinzu kommt, dass die WARC-Spezifikationen nicht vorgeben, dass Webarchivierungstools beim Erstellen der WARC-Dateien Metadaten mit Outlinks anlegen müssen. Ob eine WARC-Datei Metadaten mit gefundenen Outlinks einer Ressource enthält, liegt also daran, mit welchem Tool sie erstellt wurde.

2. Bei einer weiteren Methode werden alle HTML-Dateien ausgelesen und mithilfe der Python-Bibliothek BeautifulSoup[13] ausgewertet. Hierbei können alle Tags, die im HTML vorliegen, gefunden werden. Links, die erst durch Javascript generiert werden, können hiermit aber nicht gefunden werden, da das Javascript nicht ausgewertet wird.

[12] [8]
[13] [17]

3. Mit dem Ziel, auch das Javascript auszuwerten, werden die HTML-Dateien mit einem code-gesteuerten Browser (Selenium[14] mit Firefox/Geckodriver) geöffnet und das dabei erstellte Document Object Model (DOM) verwendet, um sowohl die im HTML-Quelltext vorhandenen als auch die dynamisch durch Javascript generierten Links erkennen zu können. Die Steuerung des Browsers und die Auswertung des Javascript macht sich deutlich in der Laufzeit bemerkbar.

Die Informationen werden nach diesem Teilschritt vorerst in einer Liste gespeichert, die aus Tuples mit URLs der Ausgangsressource und URLs der Zielressource bestehen. Im folgenden Schritt wird die erstellte Liste mithilfe des Python Pakets NetworkX in einen Netzwerkgraphen umgewandet, bei dem jede Ressource – identifiziert über ihre URL – als Knoten und jeder Verweis als Kante von der Ausgangsressource zur Zielressource modelliert wird. Informationen über Zeitpunkt der Erstellung des Graphen und über dabei verwendete Parameter werden als Attribut des Graphen gespeichert. Die Daten und Metadaten des gesamten Graphen, aber auch jedes einzelnen Knotens können beliebig erweitert werden.

Der hier beschriebene Prozess von Warc2graph ist modular aufgebaut. Wer nur an einer Extraktion der Verweise, nicht aber an einem Netzwerkgraphen interessiert ist, kann sich die Liste der extrahierten Links ausgeben lassen. Wer Informationen aus anderen Quellen in einen Netzwerkgraphen umwandeln möchte, kann eine Liste aller möglichen Verweise manuell erstellen und diese von Warc2graph in einen Graphen umwandeln lassen, der die selbe Struktur hat, wie die aus einer WARC-Datei erstellten Graphen.

3.3 Reproduzierbarkeit der Ergebnisse

Warc2graph kann auch, per Übergabe einer URL, Websites im live web analysieren. Die Ergebnisse wären allerdings zu einem späteren Zeitpunkt, wenn die Seiten sich verändert haben oder offline sind, nicht reproduzierbar. Eine notwendige Voraussetzung für die vergleichende corpusorientierte Mustererkennung ist aber die Vergleichbarkeit und Reproduzierbarkeit der Einzelanalysen. Weil auch netzliterarische Webseiten sich mit der Zeit ändern können, muss bei wiederholbaren und vergleichenden Analysen mit archivierten Versionen der Seiten gearbeitet werden.

Der hohe Grad an unkontrollierten Veränderungen von Webseiten im live web führt dazu, dass Forschungsergebnisse, die auf extrahierten Daten basieren und nur die extrahierten Daten, nicht aber die Datenbasis verfügbar halten, nach kurzer Zeit nicht mehr reproduziert werden können. Die Extraktion von Verweisstrukturen aus WARC-Dateien ist dagegen dauerhaft reproduzierbar, weil die Datenbasis als Teil der Forschungsdaten mit archiviert werden kann.

Neben der Reproduzierbarkeit von Forschungsergebnissen erlaubt dies auch eine methodisch kontrollierte Verbesserung der eingesetzten Methoden und damit perspektivisch eine Verbesserung der Analyseergebnisse, in diesem Fall der Extraktion von Verweisstrukturen.

[14] Selenium Project

Methodisch kann dieser WARC-Workflow auch auf andere Ansätze übertragen werden, die ihre Daten durch die computergestützte Verarbeitung von von Webseiten gewinnen.

4 SDC4Lit Architektur

4.1 Aufbau und Struktur

Die Aufgabe des Projekts Science Data Center for Literature (SDC4Lit), das den institutionellen Rahmen für die Studie zur narrativen Struktur von Netzliteratur und die Entwicklung von Warc2graph bildet, ist die Realisierung eines nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung zu Born-digital Materialien. Hierzu wird eine Forschungsumgebung aufgebaut, die den Zugang zu den archivierten digitalen Objekten sowie die Nutzung ausgewählter Analysemethoden und-werkzeuge für die Forschung ermöglicht. Die konzipierte SDC4Lit-Architektur besteht aus einem Primärdaten- und einem Forschungsdatenrepositorium mit einer zusätzlichen Analyseschicht und einem übergreifenden Portal.[15]

Im Primärdatenrepositorium werden digitale Objekte aus den Bereichen Literatur im Netz und den Vor- und Nachlässen des Deutschen Literaturarchivs in Marbach (DLA) langfristig gespeichert. Die Aufbereitung, Einarbeitung und Bereitstellung der Daten erfolgt durch Mitarbeiter*innen am DLA. Die geplante Analyseschicht stellt Methoden und Werkzeuge für die computergestützte Arbeit mit den digitalen Archivalien bereit. Zugehörige Forschungsdaten und Forschungsergebnisse von Nutzer*innen werden in einem separaten Forschungsdatenrepositorium gespeichert und können gegebenenfalls für die Nachnutzung zur Verfügung gestellt werden. Um die Bedarfe der Forschungscommunity bei der Entwicklung zu berücksichtigen werden im Rahmen des Projekts wissenschaftliche Fallstudien zu den bereitgestellten Primärmaterialien durchgeführt. Mit Blick auf Textumgangsformen im Bereich Elektronische Literatur und Born-Digitals werden auf diese Weise die Anforderungen an das Portal durch die Abbildung konkreter Arbeitsschritte informiert, wie sie etwa die Extraktion und Analye von Referenznetzwerken für Objekte im Bereich Literatur im Netz darstellt. Das übergreifende SDC4Lit-Portal soll die einzelnen Komponenten miteinander verbinden und den verschiedenen Nutzergruppen (Archiv, Forschung, Lehre, Publikum) einen Zugang zum Archivbestand gewähren. Eine Anbindung externen Repositorien ist geplant.

4.2 Dataverse als Repositoriumssoftware

Die Kernkomponenten des Portals bilden die Repositorien, für deren Aufbau eine passende Repositoriumssoftware ausgewählt sowie ein Datenmodell entwickelt werden muss.

[15] Weitere Informationen und die SDC4Lit Architektur Skizze sind im Extended Poster Abstract „SDC4Lit – Science Data Center for Literature“, der auch im Tagungsband der EST-21 Konferenz veröffentlicht wird, zu finden.

Angesichts der besonderen Form der Primärdaten stellen sich spezifische Anforderungen an das aufzubauende Repositorium. Hier war insbesondere die Herausforderung, dass die Primärdaten in Verzeichnisstrukturen organisiert sind. Da diese für die Forschung auch erhalten werden müssen, ergibt sich daraus die Anforderung, dass das Repositorium auch in der Lage sein muss, mit Verzeichnisstrukturen zumindest umzugehen. Die Anforderungen konnten nicht vollständig von den verfügbaren Softwarelösungen erfüllt werden, sodass hier voraussichtlich Eigenentwicklungen nötig sind. Anhand der ermittelten Anforderungen wurden mehrere Softwarelösungen analysiert und geprüft. Die Entscheidung fiel auf Dataverse[16]. In Dataverse lassen sich Verzeichnisstrukturen über das Metadatenfeld „FilePath“ nachbilden, darüber hinaus verfügt es über eine Große Nutzer*innen-Community, wird stetig fortentwickelt und hat trotzdem Produktcharakter. Darüber hinaus sind in SDC4Lit Erfahrungen aus dem bereits abgeschlossenen DIPL-ING-Projekt vorhanden, mit dem der Aufbau eines institutionellen Forschungsdatenrepositorium für die Universität Stuttgart mit Dataverse geleistet wurde.[19]

Dabei basiert eine Installation von Dataverse aus mehreren logischen Dataverses[17], die eine oberste Strukturierungsschicht darstellen. Im Fall von SDC4Lit werden die drei logischen Dataverses „Literatur im Netz“, „Born digitals“ und „Literarische Computerspiele“ als oberste Strukturierungsebene verwendet, unter die weitere logische Dataverses weitere Strukturierungsebenen einführen können. Am unteren Ende sind die Primärdaten selbst jeweils in Datasets[18] organisiert. Diese bestehen aus den Dateien, zu denen auch jeweils die Verzeichnisinformation als FilePath mit angegeben werden kann, sodass die oben genannte Anforderung erfüllt wird. Auf technologischer Ebene ist die Dataverse-Installation in einer virtuellen Maschine untergebracht, um leichte Migration und Imaging zu erlauben. Die Daten sind auf einer RAID5-Festplattenverbund verortet, um die Datensicherheit zu erhöhen. Darüber hinaus werden derzeit Backup-Routinen entworfen, um die Daten auf Bandspeicher zu sichern und in ein Langzeitarchiv zu überführen.

4.3 Datenmodellierung

Selbstverständlich werden Metadaten benötigt, damit die Forschungsdaten den FAIR-Prinzipien entsprechen [24]. Hier werden bei Dateverse Metadatenblöcke angelegt, die jeweils logisch zusammengehörende Metadaten darstellen und als Felder ausfüllbar sind. Diese Felder müssen von den Administratorinnen und Administratoren als TSV-Dateien eingepflegt werden und richten sich nach einem im Projektrahmen ausgewähltem oder definierten Standard. SDC4Lit entwickelt ein eigenes Datenmodell. Dabei wurden Metadatenstandards, die im Bereich des Bibliotheks- und Archivwesens gängig sind, wie METS [2],MODS [6] und PREMIS [3], evaluiert und ausgewählt. Metadaten, die bereits im Laufe der Sammlung, Archivierung und Erschließung entstehen, werden zusammen mit Primär-

[16]Dataverse Projekthomepage: `https://dataverse.org/`, aufgerufen: 28.04.2021.

[17]Dataverse User Guide: Dataverse Collection Management, `https://guides.dataverse.org/en/latest/user/dataverse-management.html?highlight=dataverse%20management`, aufgerufen: 28.04.2021.

[18]Siehe ebenda (Fußnote 14).

daten in Primärdatenrepositorium gespeichert, während Forschungsdaten und Metadaten, die im Laufe der Forschung entstehen, im parallel aufgebauten Forschungsdatenrepositorium gespeichert werden. Dataverse vergibt den Daten DOIs[19] worüber zusammengehörende Primär- und Forschungsdaten aufeinander referenziert werden können.

4.4 Einbindung von warc2graph in die Infrastruktur

Warc2graph wird als Teil der Analyseschicht in die SDC4Lit Infrastruktur eingebunden. Das Modul wird Input aus dem Primärdatenrepositorium und aus dem Forschungsdatenrepositorium beziehen und Output in das Forschungsdatenrepositorium schreiben können. Die generierten Outputs, die im Forschungsdatenrepositorium liegen, verweisen mittels DOI auf die zugrundeliegenden Primärdaten, und umgekehrt. Für alle Netzliteraturobjekte wird Warc2graph angewendet, um sie mit den Ergebnissen anreichern zu können, und damit einen besseren Zugang zu ermöglichen. Leser*innen können auf einen Blick sehen, ob das Werk wenige oder viele Ressourcen umfasst, mit welchen HTML-Tags die Ressourcen referenziert sind und welche Struktur sich daraus ergibt. In Verbindung mit einem Replay der WARC-Objekte sind diese grundlegenden Sturkturinformationen sowohl für literaturwissenschaftliche als auch für erhaltungsbezogene Forschung relevant sein.

5 Nachnutzung

Die von uns präsentierten Modelle, Workflows und die zugehörige Software wurden anhand eines spezifischen WARC-Korpus entwickelt. Das bedeutet, das bestimmte analytische Verfahrensweisen und die Leistung der Software abhängig ist von den Eigenschaften des Korpus. Gleichzeitig kann durch den hohen Standardisierungsgrad des WARC-Formats eine hohe strukturelle Ähnlichkeit des verwendeten Korpus mit anderen WARC-Korpora vorausgesetzt werden. Die Analyse von Referenznetzwerken in WARC-Korpora ist anschlussfähig für ganz unterschiedliche Fragestellungen, weil die Extraktion der Referenzen und die Transformation der WARC-Daten in ein graphbasiertes Datenformat bei der Korpusanalyse relativ weit am Anfang von analytischen Workflows angesiedelt sind. Wir gehen davon aus, dass die Extraktion von Referenzen generisch für die Verarbeitung von WARC-Korpora eingesetzt werden kann. Aufgrund des spezifischen Entwicklungskorpus ist eine fallbezogene Überprüfung der Extraktionsergebnisse geboten. Eine generische oder zumindest methodisch kontrollierte Nachnutzung ist erst nach Durchführung eines systematischen Qualitäts- und Leistungstests anhand ausgewählter Testkorpora denkbar. Ein solcher Test steht noch aus.

Die Transformation von WARC-Dateien in ein graphbasiertes Datenformat eröffnet über die Weiterverarbeitung der Referenzstrukturen hinaus weitergehende Möglichkeiten der Archivierung und der Analyse, insofern insbesondere hypermediale Objekte zunächst in medienspezifische Elemente aufgeteilt werden. Diese Aufteilung ermöglicht dann medien-

[19] Digital Objekt Identifier System: `https://www.doi.org/`, aufgerufen: 28.04.2021.

bzw. formatspezifische Zusammenstellungen und Analysen, z.B. Textanalysen oder Bildanalysen sowie graphbasierte Analyseansätze. Im hier gewählten Ansatz führt die korpusbasierte Modellierung, Entwicklung und Durchführung der Verweisextraktion zu Daten, die im Forschungsdatenrepositorium strukturiert vorgehalten werden und für die Nachnutzung freigegeben sind. Für den archivarischen Umgang mit den Daten ist entscheidend, dass die Referenzstrukturen und Graphdaten als grundlegende Strukturanalyse in der Regel einer frühen Phase von Forschungsprozessen zugeordnet und daher für weitere fachspezifische Forschungsfragen anschlussfähig sind - etwa für genaue Lektüren einzelner Werke mit Blick auf nicht-lineare Erzählstrukturen.

6 Acknowledgements

Das Science Data Center for Literature wird finanziert vom Ministerium für Wissenschaft und Kultur Baden Württemberg. Am Projekt beteiligt sind das Deutsche Literaturarchiv Marbach, das Höchstleistungsrechenzentrum der Universität Stuttgart sowie das Institut für Maschinelle Sprachverarbeitung und das Institut für Literaturwissenschaft der Universität Stuttgart.

Literaturverzeichnis

[1] Berkenheger, Susanne. 1997. Zeit für die Bombe. Hyperfiction. `http://www.berkenheger.netzliteratur.net/ouargla/wargla/zeit.htm` (zugegriffen: 8. Mai 2021).

[2] Cantara, Linda. 2005. METS: The Metadata Encoding and Transmission Standard. *Cataloging & Classification Quarterly* 40, Nr. 3-4 (September): 237–253. doi: `https://doi.org/10.1300/J104v40n03\_11` (zugegriffen: 10. Mai 2021).

[3] Caplan, Priscilla. 2009. Understanding PREMIS: an overview of the PREMIS Data Dictionary for Preservation Metadata. Library of Congress.

[4] Eldakar, Youssef und Lana Alsabbagh. 2020. LinkGate: Let's build a scalable visualization tool for web archive research. April. `https://netpreserveblog.wordpress.com/2020/04/23/linkgate-update/` (zugegriffen: 8. Mai 2021).

[5] Ensslin, Astrid. 2007. *Canonizing Hypertext : Explorations and Constructions*. London: Continuum International Publishing.

[6] Gartner, Richard. 2003. MODS: Metadata Object DescriptionSchema. JISC Techwatch report TSW.

[7] GEXF Working Group. 2009. GEXF File Format. `https://gephi.org/gexf/format/` (zugegriffen: 29. April 2021).

[8] Hagberg, Aric A., Daniel A. Schult und Pieter J. Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In: *Proceedings of the 7th Python in Science Conference*, hg. von Gaël Varoquaux, Travis Vaught, und Jarrod Millman, 11–15. Pasadena, CA USA.

[9] Hein, Pascal, Mona Ulrich, Claus-Michael Schlesinger und André Blessing. 2021. warc2graph. Zenodo, Mai. `https://zenodo.org/record/4742254` (zugegriffen: 8. Mai 2021).

[10] IIPC. The WARC Format. `https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/` (zugegriffen: 7. Mai 2021).

[11] Landow, George P. 1997. *Hypertext 2.0 : Hypertext - the convergence of contemporary critical theory and technology*. Rev., amplified ed. Baltimore, Md. [u.a.]: Johns Hopkins Univ. Press.

[12] Lavoie, Brian und Henrik Frystyk Nielsen. 1999. Web Characterization Terminology & Definitions Sheet. *Web Characterization Terminology & Definitions Sheet*. `https://www.w3.org/1999/05/WCA-terms/` (zugegriffen: 29. April 2021).

[13] Lin, Jimmy, Ian Milligan, Jeremy Wiebe und Alice Zhou. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *J. Comput. Cult. Herit.* 10, Nr. 4 (Juli): 22:1–22:30. doi:`http://doi.acm.org/10.1145/3097570` (zugegriffen: 19. November 2019).

[14] Marbach, Deutsches Literaturarchiv. 2018. Literatur im Netz. `http://literatur-im-netz.dla-marbach.de/` (zugegriffen: 7. Mai 2021).

[15] Martinez, Matias und Michael Scheffel. 2007. *Einführung in die Erzähltheorie*. München: C.H. Beck.

[16] Mike Burner und Brewster Kahle. 1996. Arc File Format. *Internet Archive: ARC File Format Reference*. September. `https://archive.org/web/researcher/ArcFileFormat.php` (zugegriffen: 10. Mai 2021).

[17] Richardson, Leonard. 2020. Beautiful Soup Documentation. `https://www.crummy.com/software/BeautifulSoup/bs4/doc/` (zugegriffen: 29. April 2021).

[18] Selenium Project. The Selenium Browser Automation Project. `https://www.selenium.dev/documentation/en/` (zugegriffen: 29. April 2021).

[19] Selent, Björn, Björn Schembera, Dorothea Iglezakis und Anett Seeland. 2019. Datenmanagement in Infrastrukturen, Prozessen und Lebenszyklen für die Ingenieurwissenschaften : Abschlussbericht des BMBF-Projektes Dipl-Ing. Universität Stuttgart. doi:10.2314/KXP:1693393980, `https://www.tib.eu/suchen/id/TIBKAT:1693393980/` (zugegriffen: 10. Mai 2021).

[20] Suter, Beat, Michael Böhler und Christian Bachmann, Hrsg. 1999. *Hyperfiction: hyperliterarisches Lesebuch: Internet und Literatur*. Nexus 50. Frankfurt am Main: Stroemfeld.

[21] Webrecorder Project. 2021. webrecorder/warcio. Webrecorder, Mai. https://github.com/webrecorder/warcio (zugegriffen: 5. Mai 2021).

[22] ---. Webrecorder pywb 2.5. *GitHub - webrecorder/pywb: Core Python Web Archiving Toolkit for replay and recording of web archives.* https://github.com/webrecorder/pywb (zugegriffen: 10. Mai 2021a).

[23] ---. WARCIO: WARC (and ARC) Streaming Library. https://github.com/webrecorder/warcio (zugegriffen: 7. Mai 2021b).

[24] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, Nr. 1 (Dezember): 160018. doi:10.1038/sdata.2016.18, http://www.nature.com/articles/sdata201618 (zugegriffen: 10. Mai 2021).

Forschungsdaten in den Naturwissenschaften: Eine urheberrechtliche Bestandsaufnahme mit ihren Implikationen für universitäres FDM

Thomas Hartmann

FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, Forschungsbereich Immaterialgüterrechte in verteilten Informationsinfrastrukturen

Ein Schlüsselfaktor für die Zugänglichkeit, Nachnutzbarkeit und Interoperabilität von Forschungsdaten[1] ist deren urheberrechtlicher Status. In den Geistes- und Sozialwissenschaften unterliegen Forschungsdaten (z. Bsp. Texte in Form von Interviews oder sonstiger Literatur) in den meisten Fällen dem Urheberschutz. Anders ist die Situation in den Naturwissenschaften. Nicht immer, aber häufig bleiben Forschungsdaten aus diesen Fächern urheberrechtsfrei. Dies begründet der Beitrag an einem typischen Beispiel aus dem Forschungsdatenzentrum für Molekulare Materialforschung (SDC MoMaF).[2] Welche rechtlichen Handlungsempfehlungen sich für das Forschungsdatenmanagement (FDM) daraus ergeben, wird am Beitragsende dargestellt.

1 Urheberrechtlicher Rahmen für wissenschaftliche Leistungen (Texte und Forschungsdaten)

Urheberschutz und Patente („geistiges Eigentum“)
Nach immaterialgüterrechtlicher Systematik ist das Urheberrecht ebenso wie etwa der gesetzliche Patent- oder Markenschutz als Schutzrecht eigentumsähnlich mit absoluter, ausschließlicher Wirkung („geistiges Eigentum“ bzw. Intellectual Property, IP) ausgestaltet. Vom Urheberrecht zu unterscheiden sind Patente, die für technische, gewerblich anwendbare Erfindungen gewährt werden und der Anmeldung in einem öffentlichen Register bedürfen. Der gesetzliche Urheberschutz setzt keine Registrierung voraus und beinhaltet Verwertungsrechte und Persönlichkeitsrechte.

[1] Mehr zu den FAIR-Prinzipien des Forschungsdatenmanagements siehe z. Bsp. die GO FAIR Initiative unter `https://www.go-fair.org/`
[2] Mehr zum SDC MoMaF unter `https://momaf.scc.kit.edu/`

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13728 (CC BY-SA 4.0)

Urheberschutz an Texten
Das Urheberrecht gewährt unter bestimmten Voraussetzungen rechtlichen Schutz für geistige Leistungen aus Literatur, Wissenschaft und Kunst. Auf die Wissenschaftlichkeit kommt es dabei nicht an, die gesetzlichen Voraussetzungen insbesondere für Werkschutz sind einheitlich festgelegt für wissenschaftliche wie für nichtwissenschaftliche Werke. Textpublikationen erreichen in der Regel die urhebergesetzlichen Schutzvoraussetzungen unabhängig davon, ob es sich um wissenschaftliche oder etwa um belletristische Texte handelt. Das bedeutet zudem, dass die fachliche Domäne eines wissenschaftlichen Textes nicht für dessen urheberrechtlichen Schutz relevant ist. Ein Artikel oder sonstige Fachveröffentlichung aus der Kunstgeschichte genießt Urheberschutz gleichermaßen wie ein Artikel aus der Chemie, aus der Mathematik oder aus der Materialwissenschaft.

Urheberschutz an Forschungsdaten
Grundlegend anders verhält es sich urheberrechtlich bei Forschungsdaten. Insbesondere die gewählte Forschungsmethode bestimmt die äußere Gestalt der Forschungsdaten, welche für deren urheberrechtlichen Schutzstatus relevant ist. In den Sozialwissenschaften bspw. sind Interviews eine beliebte Forschungsmethode, die als Texte selbstverständlich Urheberschutz erlangen (s.o.). In den Naturwissenschaften hingegen entstehen bei Experimenten, Versuchen und Messungen Forschungsdaten zum Beispiel in Form von Messwerten und in technischen Größen angegebenen Parametern. Auf deren urheberrechtlicher Status wird in diesem Beitrag näher eingegangen.

Urheberschutz in den einzelnen Fachwissenschaften
Vorsicht ist anzuraten bei Verallgemeinerungen für ganze Wissenschaftsdisziplinen: Texte sind zwar in der Regel urheberrechtlich geschützt. Wer in der historischen Forschung jedoch mit hinreichend alten Textmaterialien arbeitet, hat es bei diesen Texten ausnahmsweise mit urheberrechtsfreien Forschungsdaten zu tun. Umgekehrt können in den Naturwissenschaften Aufnahmen und Zeichnungen dem Urheberrecht unterliegen, während andere als Zahlenwerte abgebildete Forschungsdaten urheberrechtsfrei sind. Die Vielfalt an Forschungsmethodik und die daraus resultierende uneinheitliche urheberrechtliche Situation zeigt sich z. Bsp. auch innerhalb der Lebenswissenschaften.

Bedeutung des urheberrechtlichen Status von Forschungsdaten
Ob Forschungsdaten urheberrechtlichen Schutz erlangen,[3] ist eine zentrale Rechtsfrage für das FDM. Antipodisch sind Forschungsdaten entweder urheberrechtlich geschützt oder sie sind urheberrechtsfrei (gemeinfrei). Urheberrechtlich geschützte Forschungsdaten unterliegen vollständig dem gesetzlichen Urheberrecht, urheberrechtsfreie Forschungsdaten haben mit dem Urheberrecht nichts zu tun. Es sollte deshalb im FDM nicht versäumt werden, zunächst den urheberrechtlichen Status der Forschungsdatensätze zu klären.

[3] Näher dazu schon Hartmann, Zur urheberrechtlichen Schutzfähigkeit von Forschungsdaten. In: InTeR – Zeitschrift zum Innovations- und Technikrecht 1, 4 (2013), S. 199 ff. Abrufbar unter `http://hdl.handle.net/11858/00-001M-0000-0014-1208-E` (abgerufen am 27.04.2021).

Wissenschaftliche Leistung unbeachtlich für urheberrechtliche Beurteilung
Da Textpublikationen grundsätzlich urheberrechtlich geschützt sind (s.o.), entfaltet dort die Perspektive wissenschaftlicher Leistung keine weitere Bedeutung.[4] Um Missverständnisse bei Forschenden und in Fachwissenschaften zu vermeiden, sollte jedoch in Bezug auf Forschungsdaten und Urheberrecht klargestellt werden: Urheberschutz wird nicht für bestimmte, fachbezogene wissenschaftliche Leistungen oder Kenntnisse vergeben.[5] Was aus fachwissenschaftlicher Sicht rechtlich schützenswert sein sollte, ist nicht Maßstab der urhebergesetzlichen Beurteilung. Aus Sicht der einzelnen Forschenden und Fachgruppen mag die Zuweisung urheberrechtlichen Schutzes insoweit eher beliebig erscheinen. Die Erklärung dafür sind die allgemein, weitestgehend wissenschaftsunabhängig gehaltenen gesetzlichen Voraussetzungen zur Erlangung von Urheberschutz.[6] Diese urhebergesetzliche Ausgangslage ist bei den juristischen Gestaltungsmöglichkeiten und Handlungsempfehlungen ein maßgeblicher Gesichtspunkt.

2 Ein Forschungsdatensatz im Forschungsdatenzentrum für Molekulare Materialforschung (SDC MoMaF)

Die urheberrechtliche Beurteilung richtet sich auf chemische Forschungsdaten, die im SDC MoMaF mit seinem integrierten Elektronischen Laborbuch (ELN) und seinem Repositorium Chemotion erfasst und veröffentlicht werden. Im Folgenden wird beispielhaft ein typischer Datensatz vollständig abgebildet, für seine urheberrechtliche Beurteilung gegliedert in vier Bestandteile:

a.) Abbildung der Molekülstrukturen,

b.) Tabellarisch dargestellte Eigenschaften von Versuchsmaterialien,

c.) Kurzbeschreibung des Versuchs,

d.) Analyse der Messdaten.

3 Urheberrechtliche Beurteilung

Das Urheberrecht gewährt Schutz für Werke. Im Urheberrechtsgesetz sind als Werkarten insbesondere Texte, Filme, Fotos, technische Zeichnungen angeführt (§ 2 Abs. 1 UrhG). Zudem müssen geistige Leistungen für den Werkschutz insbesondere ein Mindestmaß an

[4] So auch Herrmann/Trottier, Urheberrecht und Werkqualität: Ein Überblick aus der Wissenschaftspraxis. In: Forschung & Lehre 25, 2 (2018). Abrufbar unter `https://www.forschung-und-lehre.de/urheberrecht-und-werkqualitaet-326/` (abgerufen am 27.04.2021).

[5] Vgl. Loewenheim/Leistner, Handbuch des Urheberrechts, 3. Aufl. 2021, § 6 Rz. 8, 21 ff., 34.

[6] Vgl. Herrmann/Trottier, Urheberrecht und Werkqualität: Ein Überblick aus der Wissenschaftspraxis. In: Forschung & Lehre 25, 2 (2018). Abrufbar unter `https://www.forschung-und-lehre.de/urheberrecht-und-werkqualitaet-326/` (abgerufen am 27.04.2021).

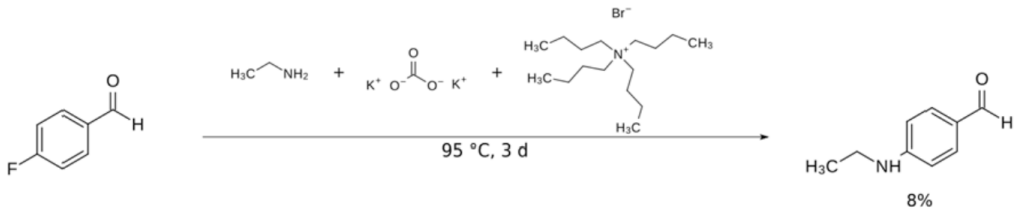


Abbildung 1: Abbildung der Molekülstrukturen (Bestandteil a eines veröffentlichten Datensatzes im SDC MoMaF).

	Formula	Mol mass	Mass [g]	Volume [mL]	Density [g/mL]	mmol	Equiv/yield
4-fluorobenzaldehyde (4-fluorobenzaldehyde)							
S	C7H5FO	124	**2.00**	1.73	1.16	16.1	1.00
ethanamine (ethanamine in Wasser)							
R	C2H7N	45.1	**1.65**	2.04	0.810	24.2	1.50
Potassium carbonate (K2CO3)							
R	CK2O3	138	**2.23**	0.00	0.00	16.1	1.00
tetrabutylazanium;bromide (reactant)							
R	C16H36BrN	322	**1.56**	0.00	0.00	4.83	0.300
4-(ethylamino)benzaldehyde (SG1-V2448-A)							
P	**C9H11NO**	149	**0.181**	0.00	0.00	1.21	8%

Solvent(s):

Abbildung 2: Tabellarisch dargestellte Eigenschaften von Versuchsmaterialien (Bestandteil b eines veröffentlichten Datensatzes im SDC MoMaF).

schöpferischer Qualität erreichen (§ 2 Abs. 2 UrhG).[7] Urheberrechtlich erforderlich ist damit insbesondere eine jeweils äußere Form von Forschungsdaten, die überdurchschnittlich individuell, eigentümlich und originell gehalten sein muss. Auch bei Texten muss für die urheberrechtlich erforderliche Schöpfungshöhe ein Mindestmaß an Individualität erreicht werden.[8] Daran mangelt es z. Bsp. bei systematischen Aufzählungen, nicht kreativen Texten und rein dokumentarischen Aufnahmen.[9]

Für den Wissenschaftsbereich ist urheberrechtlich zudem anerkannt, dass ein besonderes Freihaltebedürfnis besteht. So soll die grundrechtlich geschützte Freiheit des wissenschaftlichen Austausches über Gedanken, Lehren, Ideen, Methoden und Forschungsergebnisse gewährleistet werden.[10] Im Besonderen für Natur-, Lebens- und Technikwissenschaften ist

[7] Näher u.a. zum Werkbegriff in der Wissenschaft Herrmann/Trottier, Urheberrecht und Werkqualität: Ein Überblick aus der Wissenschaftspraxis. In: Forschung & Lehre 25, 2 (2018). Abrufbar unter `https://www.forschung-und-lehre.de/urheberrecht-und-werkqualitaet-326/` (abgerufen am 27.04.2021).

[8] Vgl. zur erforderlichen Gestaltungshöhe bei Texten z. Bsp. Vettermann, Datenschutzrechtliche Informationspflichten zwischen Kreativität und Transparenz – Urheberrechtlicher Schutz von Datenschutzerklärungen. In: ZD – Zeitschrift für Datenschutz 11, 5 (2021), S. 257 ff.

[9] Beispiele von Kuschel, Wem „gehören" Forschungsdaten? In: Forschung & Lehre 25, 9 (2018). Abrufbar unter `https://www.forschung-und-lehre.de/wem-gehoeren-forschungsdaten-1013/` (abgerufen am 27.04.2021).

[10] Vgl. z. Bsp. Loewenheim/Leistner, Handbuch des Urheberrechts, 3. Aufl. 2021, § 7 Rz. 9 f. m.w.N.

Description:

4-Fluorobenzaldehyde (2.00 g, 1.73 mL, 16.1 mmol, 1.00 equiv), ethanamine (66% in water, 1.09 g, 1.35 mL, 24.2 mmol, 1.50 equiv), tetrabutylazanium;bromide (1.56 g, 4.83 mmol, 0.300 equiv) and Potassium carbonate (2.23 g, 16.1 mmol, 1.00 equiv) were stirred at 95 °C over a periode of three days. Water was given to the reaction and the resulting mixture was extracted three times with ethyl acetate. The combined organic phases were washed with brine and dried over Na_2SO_4 and filtered.

Type of Purification: Flash-Chromatography

Dangerous Products:

TLC control: Rf-value: 0.67 (Solvent: cyclohexane/ethyl acetate 1:1)

Additional information for publication and purification details:

As preparation for the column chromatography (dryload), silica gel was added (6 g) and the mixture of combined organic phases + silica gel were evaporated.The obtained crude product was purified via flash-chromatography on silica gel using cyclohexane/ethyl acetate 4:1

Abbildung 3: Kurzbeschreibung des Versuchs (Bestandteil c eines veröffentlichten Datensatzes im SDC MoMaF).

daher klarzustellen: Die fachlichen Gedanken, Inhalte, Fakten und Forschungsergebnisse an sich sind stets urheberrechtsfrei, lediglich deren jeweils konkrete sprachliche Ausformulierung oder etwa Ausgestaltung in einer individuellen Zeichnung können als solche urhebergesetzlichen Schutz erlangen.

a.) **Abbildung der Molekülstrukturen (Bestandteil a des Beispieldatensatzes, Kap. 2)**
Die grafische Darstellung der Molekülstrukturen in den Reaktionen (Bestandteil a des Beispieldatensatzes aus Kap. 2, s.o.) beruht auf durch Software und durch fachwissenschaftliche Gesetzmäßigkeiten vorgegebenen Einstellungen. Die Forschenden geben jeweils die chemischen Strukturen, Reaktionsparameter sowie weitere Bedingungen des Experiments als Zahlenwerte bzw Information ein, daraufhin erstellt die Software die grafischen Abbildungen der jeweiligen Reaktionen bzw. der einzelnen Molekülstrukturen. Es verbleibt damit für die Forschenden in der Regel kein ausreichender Gestaltungsspielraum, um eine individuelle, von der fachwissenschaftlich üblichen Darstellungsweise abweichende, eigentümliche Ausdrucksform zu kreieren. Die Abbildung der Molekülstrukturen ist daher in der Regel urheberrechtsfrei.

b.) **Tabellarisch dargestellte Eigenschaften von Versuchsmaterialien (Bestandteil b des Beispieldatensatzes, Kap. 2)**
Die für eine Reaktion verwendeten Substanzen, deren eingesetzte Menge und ggfs. weitere Details werden präzise beschrieben. Diese Daten (Bestandteil b des Beispieldatensatzes aus Kap. 2, s.o.) sind weder einer urhebergesetzlichen Werkart zuzuordnen noch beruhen sie auf einer schöpferischen Entscheidung der jeweils Forschenden.

Analysis:

C9H11NO (CHMO:0000593 | 1H nuclear magnetic resonance spectroscopy (1H NMR))

^{1}H NMR (400 MHz, $CDCl_3$ [7.27 ppm], ppm) δ = 9.73 (s, 1H), 7.72–7.68 (m, 2H), 6.62–6.60 (m, 2H), 4.37 (br. s., 1H), 3.26 (q, *J* = 7.2 Hz, 2H), 1.30 (t, *J* = 7.2 Hz, 3H).

C9H11NO (CHMO:0000595 | 13C nuclear magnetic resonance spectroscopy (13C NMR))

^{13}C NMR (100 MHz, $CDCl_3$ [77.0 ppm], ppm) δ = 190.2, 153.3, 132.3 (2C), 126.4, 111.7 (2C), 37.9, 14.5.

C9H11NO (CHMO:0000630 | infrared absorption spectroscopy (IR))

IR (ATR, $\tilde{\nu}$) = 3291, 2967, 2923, 2870, 2852, 2809, 2731, 1661, 1586, 1570, 1566, 1562, 1545, 1534, 1508, 1498, 1474, 1451, 1437, 1427, 1395, 1378, 1366, 1342, 1310, 1283, 1231, 1147, 1109, 1062, 1044, 997, 824, 810, 634, 619, 590, 583, 510, 426 cm^{-1}.

C9H11NO (CHMO:0000470 | mass spectrometry (MS))

EI (m/z, 70 eV, 60 °C): 149 (100), 134 (99). HRMS ($C_9H_{11}ON$): Calcd 149.0835, Found 149.0835.

Abbildung 4: Analyse der Messdaten (Bestandteil d eines veröffentlichten Datensatzes im SDC MoMaF).

Eine besondere Originalität und Individualität vor allem der äußeren Darstellung wird – unabhängig vom Einsatz ggfs. hoher fachwissenschaftlicher Expertise und hohem Ressourcenaufwand – nicht erreicht. Auch die Darstellung der Angaben in einer Tabelle führt zu keinem Urheberschutz, denn die Gestaltung der Tabelle ist schlicht und folgt technischen Voreinstellungen des Systems sowie fachwissenschaftlich allgemein üblichen Darstellungen.[11]

Die tabellarisch dargestellten Eigenschaften von Versuchsmaterialien sind daher in der Regel urheberrechtsfrei.

c.) **Kurzbeschreibung des Versuchs (Bestandteil c des Beispieldatensatzes, Kap. 2)**
Der Versuch wird in kurzen, in fachwissenschaftlich üblicher Sprache abgefassten Sätzen dokumentiert. Im Vordergrund der knappen Formulierungen stehen fachwissenschaftliche Konventionen und die Dokumentation technischer Vorgänge.[12]
Urheberrechtlich fragwürdig ist, ob ein solcher Beschreibungstext insgesamt ein Mindestmaß an eigenschöpferischer, individueller Prägung und damit Urheberschutz er-

[11] Vgl. z. Bsp. Loewenheim/G. Schulze, Handbuch des Urheberrechts, 3. Aufl. 2021, § 9 Rz. 268 m.w.N.

[12] Ein urheberrechtlicher Schutz ist ausgeschlossen, wenn Formulierungen, Struktur und Gedankenführung im Wesentlichen durch fachliche Gepflogenheiten vorgegeben sind. So Lauber-Rönsberg/Krahn/Baumann, Gutachten zu den rechtlichen Rahmenbedingungen des Forschungsdatenmanagements (Kurzfassung), 2018, S. 2. Abrufbar unter `https://tu-dresden.de/gsw/phil/irget/jfbimd13/ressourcen/dateien/dateien/DataJus/DataJus_Zusammenfassung_Gutachten_12-07-18.pdf?lang=de` (abgerufen am 27.04.2021).

langen kann. In der juristischen Fachliteratur wird vertreten, dass Forschungsdaten in Form kurzer Versuchsbeschreibungen in der Regel nicht urheberrechtlich schutzfähig sind.[13] Zugleich jedoch wird mit Verweis auf unzureichende Rechtssicherheit im Einzelfall, teils aus allgemeinen Vorsichtigkeitserwägungen empfohlen grundsätzlich einen Urheberrechtsschutz der Forschungsdaten anzunehmen.[14] Solch eine Empfehlung ist mindestens für Forschungsbereiche abzulehnen, welche Forschungsdaten standardisiert nach fachwissenschaftlichen Vorgaben generieren.[15] Wenn Versuchsbeschreibungen einen größeren Umfang[16] erreichen, kann darin grundsätzlich individuelle und originelle Ausdruckskraft erkennbar sein. Wenn zwei Forschende unabhängig voneinander einen Versuch in gewissem (Mindest-)Umfang dokumentieren, sollten sie zwar eine möglichst fachtypisch einheitliche, systematisierte sprachliche Beschreibung verwenden. Im Ergebnis jedoch werden erfahrungsgemäß nur selten identische Formulierungen erzielt. Dies könnte urheberrechtlich ein Anhaltspunkt sein, um doch ein Mindestmaß an eigenschöpferischer Ausdruckskraft und damit Urheberschutz für den Beschreibungstext anzunehmen. Zusammengefasst ist zweifelhaft, ob knappe, sprachlich fachtypisch einheitlich gehaltene Formulierungen von Versuchsbeschreibungen urheberrechtlichen Schutz erlangen. Nur wenn ausformulierte Kurzbeschreibungen eines Versuchs einige Sätze mit Mindestmaß an individueller Gestaltung umfassen, könnten diese – zumindest in ihrer Gesamtheit – urheberrechtlich geschützt sein.

[13] So Kuschel, Wem „gehören“ Forschungsdaten? In: Forschung & Lehre 25, 9 (2018). Abrufbar unter `https://www.forschung-und-lehre.de/wem-gehoeren-forschungsdaten-1013/` (abgerufen am 27.04.2021).

[14] So z. Bsp. Lauber-Rönsberg, Rechtliche Aspekte des Forschungsdatenmanagements, in: Putnings/Neuroth/Neumann (Hrsg.), Praxishandbuch Forschungsdatenmanagement, 2021, S. 91. Abrufbar unter `https://doi.org/10.1515/9783110657807-005` (abgerufen am 27.04.2021). Für „größtmögliche Rechtssicherheit“ empfiehlt auch die von der Universität Konstanz verantwortete Plattform forschungsdaten.info Forschungsdaten „zunächst so zu behandeln, als wären sie regelmäßig schutzwürdig nach dem Urheberrecht, um durch einfache Maßnahmen (z. B. Namensnennung, Einbindung der Urheberin oder des Urhebers in Publikationsentscheidungen) etwaige Probleme zu vermeiden, die bei der Verarbeitung und Organisation vieler solcher Daten anfallen.“ Siehe `https://www.forschungsdaten.info/themen/rechte-und-pflichten/urheberrecht/` (abgerufen am 27.04.2021).

[15] Ähnlich auch Lauber-Rönsberg/Krahn/Baumann, Gutachten zu den rechtlichen Rahmenbedingungen des Forschungsdatenmanagements (Kurzfassung), 2018, S. 3. Abrufbar unter `https://tu-dresden.de/gsw/phil/irget/jfbimd13/ressourcen/dateien/dateien/DataJus/DataJus_Zusammenfassung_Gutachten_12-07-18.pdf?lang=de` (abgerufen am 27.04.2021).

[16] „Metadaten werden die urheberrechtlichen Schutzvoraussetzungen allerdings in der Regel nicht erfüllen, da es sich häufig nur um relativ kurze Beschreibungen handelt.“, so Lauber-Rönsberg, Rechtliche Aspekte des Forschungsdatenmanagements, in: Putnings/Neuroth/Neumann (Hrsg.), Praxishandbuch Forschungsdatenmanagement, 2021, S. 91. Abrufbar unter `https://doi.org/10.1515/9783110657807-005` (abgerufen am 27.04.2021).

d.) **Analyse der Messdaten (Beispiel d des Beispieldatensatzes, Kap. 2)**

Die Charakterisierung und damit der Nachweis der in einer chemischen Reaktion erhaltenen Ergebnisse wird durch die Auswertung von Messdaten erhalten. Diese Messdaten werden durch wissenschaftliche Instrumente generiert, die in der Regel für eine bestimmte Produktklasse jeweils ähnlich gewählt werden. Die Auswertung der erhaltenen Messdaten selbst wird, unterstützt durch wissenschaftliche Software, von den Wissenschaftlern selbst vorgenommen. Die Ergebnisse werden durch Anwenden von chemischem Fachwissen und Interpretation z. Bsp. der erhaltenen Signale in einem Messdatensatz erhalten. Die Angabe der Ergebnisse wird in standardisierter Form, in Anlehnung and etablierte Darstellungsweisen vorgenommen.

Messwerte und ausgewertete Messergebnisse sind urheberrechtlich nicht geschützt. Dies gilt unabhängig vom Einsatz bestimmter Versuchsapparaturen, Messinstrumente und Facheinschätzungen. Urheberrechtlich unbeachtlich sind ebenso weitere fachwissenschaftliche Verfahrensweisen, welche zur Erhebung und Auswertung von Messergebnissen herangezogen werden.

Die Messdaten sind daher in der Regel urheberrechtsfrei.

Fazit

Wesentliche Teile des Datensatzes folgen grafisch und technisch vorgegebenen Darstellungen sowie fachwissenschaftlichen Konventionen und erlangen daher in der Regel keinen Urheberschutz. Ebenso bleiben Mess- und andere Zahlenwerte und einzelne Fachbegriffe in Versuchsbeschreibungen urheberrechtsfrei. In gewisser Länge individuell ausformulierte Beschreibungen können in ihrer äußeren Form urheberrechtlich geschützt sein.

4 Folgen für Lizenzierung: Drohende Scheinrechte

Lizenzierung ist kein juristisch feststehender, im deutschen Gesetzesrecht definierter Begriff.[17] In immaterialgüterrechtlichem, speziell in urheberrechtlichem Zusammenhang gemeint ist damit die Einräumung bestimmter Nutzungsrechte (§§ 31 ff. UrhG). Die Einräumung solcher Nutzungsrechte kann an urheberrechtlich geschützten Materialien erfolgen. Der Anwendungsbereich urheberrechtlicher Lizenzmodelle wie z. Bsp. Creative Commons ist deshalb bezogen auf urheberrechtlich geschützte Materialien.

Werden urheberrechtsfreie Forschungsdaten mit Lizenzen versehen, wirft das grundlegende juristische und wissenschaftspolitische Fragen auf.

Urheberrechtliche Analyse
In der urheberrechtlichen Literatur werden Lizenzierungen nicht urheberrechtlich geschützter Materialien als „Scheinrechte“[18] bezeichnet mit Verweis auf die vom Bundesge-

[17]Eingehend dazu ObergfellHauck (Hrsg.), Lizenzvertragsrecht, 2. Aufl. 2020, S. 1 ff.
[18]Wandtke Urheberrecht, 7. Aufl. 2019, S. 116.

richtshof aufgestellten Grundsätze zu „Leerübertragungen“.[19] Die Verwendung urheberrechtlicher Lizenzen verheißt absolut, ausschließlich wirkenden Urheberschutz (s.o. Kap. 1) der zugrundeliegenden Forschungsdaten, obwohl diese von Gesetzes wegen keinen Urheberschutz haben und Urheberschutz auch nicht vom Lizenzverwender hergestellt werden kann.[20] Die Schieflage ähnlich zum Ausdruck gebracht wird durch „(betrügerische) Rechteanmaßungen“[21] oder „Schutzrechtsrühmung“[22] von Urheberrechten, die an urheberrechtsfreien Forschungsdaten eben gerade nicht bestehen (können). Andere beklagen einen „Copyfraud“ (Urheberrechtsbetrug).[23]

In Betrachtung der für Forschungsdaten häufig konkret empfohlenen Lizenzbedingung Creative Commons Namensnennung 4.0 International (CC BY) gibt die Organisation Creative Commons bei den Lizenzbedingungen diesen Hinweis: „Sie müssen sich nicht an diese Lizenz halten hinsichtlich solcher Teile des Materials, die gemeinfrei sind“.[24] Nach Auffassung von Creative Commons ist damit ausdrücklich eine CC BY-Lizenz an urheberrechtsfreien Forschungsdaten juristisch wirkungslos. Die gerade von Forschenden erwünschte Pflicht zur Namensnennung bzw. Quellenangabe bei Verwendung ihrer Forschungsdaten kann so juristisch nicht herbeigeführt werden.[25]

Zudem ist nicht rechtssicher zu beantworten, welche Rechtswirkung solche Scheinlizenzierungen nach allgemeinen (zivil-)rechtlichen Grundsätzen entfalten können. Denn die übliche Wirkung von Lizenzen bei tatsächlich vorhandenem Urheberschutz, nämlich ein absoluter, eigentumsähnlicher Schutz gegenüber jeder Person, ist ausgeschlossen (s.o. Kap. 1). Es verbleibt eventuell insbesondere eine schuldrechtliche Bindungswirkung wie bei Verträgen. Demnach wird zur Einhaltung der Lizenzbedingungen nur verpflichtet, wer sich dazu vertraglich verpflichtet hat. Eine derartige vertragliche Verpflichtung erstreckt sich

[19] Vgl. näher auch zu den Rechtsfolgen daraus z. Bsp. Wandtke, Urheberrecht, 7. Aufl. 2019, S. 116; Obergfell/Hauck (Hrsg.), Lizenzvertragsrecht, 2. Aufl. 2020, S. 2.

[20] „Ohne Einräumung der Nutzungsrechte ist der Urheberrechtsvertrag ohne Bedeutung“, konstatiert z. Bsp. Wandtke Urheberrecht, 7. Aufl. 2019, S. 115. Er sieht die Einräumung von Nutzungsrechten in Urheberrechtsverträgen als deren „essentialia negotii“ (aaO).

[21] Vgl. Klimpel, Urheberrecht, Praxis und Fiktion, 2013, S. 3 f., 11 f.

[22] Dreier-Schulze (Hrsg.), Urheberrechtsgesetz, 6. Aufl. 2018, § 13 Rz. 37 m.w.N.

[23] Kreutzer/Lahmann, Rechte an Forschungsdaten und Datenbanken. Portal iRights.info, 2019. Abrufbar unter `https://irights.info/artikel/rechte-an-forschungsdaten-und-datenbanken/29587` (abgerufen am 27.04.2021).

[24] Siehe `https://creativecommons.org/licenses/by/4.0/deed.de` (abgerufen am 27.04.2021). Der Lizenzvertrag zu dieser Lizenz trifft dazu folgende Auslegungsregel: „Es sei klargestellt, dass die vorliegende Public License weder besagen noch dahingehend ausgelegt werden soll, dass sie solche Nutzungen des lizenzierten Materials verringert, begrenzt, einschränkt oder mit Bedingungen belegt, die ohne eine Erlaubnis aus dieser Public License zulässig sind.“Siehe `https://creativecommons.org/licenses/by/4.0/legalcode.de` (abgerufen am 27.04.2021).

[25] Dahingehend auch Lauber-Rönsberg, Rechtliche Aspekte des Forschungsdatenmanagements, in: Putnings/Neuroth/Neumann (Hrsg.), Praxishandbuch Forschungsdatenmanagement, 2021, S. 94. Abrufbar unter `https://doi.org/10.1515/9783110657807-005` (abgerufen am 27.04.2021). Über Forschungsdaten hinausgehend für die Nutzung gemeinfreier Materialien ohne Quellenangabe: „Wer gemeinfreie Werke als eigene Schöpfung ausgibt, ist kein Plagiator“, so Loewenheim/Leistner, Handbuch des Urheberrechts, 3. Aufl. 2021, § 8 Rz. 30. Wobei insbesondere wissenschaftsrechtlich (z. Bsp. in Promotionsordnungen) solch ein Verhalten eben doch als Plagiat definiert und mit rechtlichen Konsequenzen verbunden werden kann.

lediglich auf den vertraglich konkret definierten Vertragsgegenstand. Einzelheiten solcher juristischer Konstruktionen sind aktuell Gegenstand rechtswissenschaftlicher Abhandlungen zu Konzeptionen eines Datenrechts[26] und zu Datenlizenzen.[27]

Mit Blick auf den FDM-Kontext unklar ist, wer sich wie (nur) vertraglich wirksam zur Einhaltung der jeweiligen Vertragsbedingungen (ggfs. bezeichnet als „Lizenzbedingungen", „Nutzungsbedingungen" o.Ä.) verpflichtet. Denkbar erscheint eine vertraglich wirksame Einbeziehung z. Bsp., wenn Nutzer/innen bei der Nutzung eines Forschungsdatenzentrums aktiv in Lizenz- bzw. Nutzungsbedingungen einwilligen müssten. Dies bringt einen erhöhten technischen und administrativen Aufwand mit sich und erfordert eine datenschutzfreundliche Ausgestaltung. Vor allem aber können solche vertragsrechtlichen Zugriffsanforderungen zentrale Zielsetzungen digitaler Wissenschaft (Open Access) und des Forschungsdatenmanagements (FAIR-Prinzipien) behindern.
Das eventuelle Anliegen, sich gegen missbräuchliche Nutzungen von Forschungsdaten juristisch zur Wehr setzen zu können, droht ins Leere zu laufen. Denn es ist zu befürchten, dass gerade Personen mit missbräuchlichen oder sonstig unerwünschten Nutzungen rechtlich nicht belangt werden können, weil deren vertragliche Verpflichtung auf bestimmte (Nutzungs-)Bedingungen nicht nachgewiesen werden kann oder die jeweils konkret beanstandete Nutzung nicht von der jeweils zugrundeliegenden vertraglichen Regelung abgedeckt ist. Urheberrechtsfreie Forschungsdaten sollten daher nicht mit üblichen Urheberrechtsvermerken („Urheberschutz", „Alle Rechte vorbehalten", „Alle Rechte beim/ bei der Urheber/in" etc.) und urheberrechtlichen Lizenzen (z. Bsp. Creative Commons-Lizenzen mit Mindestlizenzbedingung CC BY) versehen werden.

Wissenschaftspolitische Analyse
Das Urheberrecht ist ein Verbotsrecht. Urheber/innen und Rechteinhaber/innen sind entsprechend dazu berechtigt, jeden von der Nutzung auszuschließen (§ 31 Abs. 3 Satz 1 UrhG, Prinzip: „alle Rechte sind vorbehalten"). Die Interessen von Nutzern/innen hingegen sind im Urheberrecht nur rudimentär und in engen Grenzen abgebildet.[28]

Diese Ausrichtung ist dem Urheberrecht insgesamt immanent und gilt gleichermaßen auch für Arbeitsbereiche von Forschung, Lehre, Bildung und öffentlichen Informationsinfrastrukturen (s.o. Kap. 1), die besonders auf eine ausgewogene Ausbalancierung von Schutzrechten und Nutzungsinteressen angewiesen sind.[29]
In den letzten 25 Jahren erleben Wissenschaftler/innen und Wissenschaftseinrichtungen das Urheberrecht und insbesondere auch die für sie aufgestellten Schrankenregelungen

[26] Siehe z. Bsp. Hacker, Datenprivatrecht, 2020; Pertot (Hrsg.), Rechte an Daten, 2020; Jöns, Daten als Handelsware 2019; Zech, Information als Schutzgegenstand, 2012.
[27] Siehe z. Bsp. Schur, Die Lizenzierung von Daten, 2020.
[28] Mit daran grundlegender Kritik Kuhlen, Die Transformation der Informationsmärkte in Richtung Nutzungsfreiheit, 2020. Abrufbar unter `https://doi.org/10.1515/9783110693447` (abgerufen am 27.04.2021). Er schlägt ein Konzept der „Nutzungsrechte und Nutzungsfreiheit für Wissen und Information" vor, das an Stelle des individualistischen Urheberrechts treten soll.
[29] In seiner juristischen Abhandlung bezogen auf Open Educational Resources in der Hochschullehre beklagt Horlacher z. Bsp. ein „gesetzgeberisches Unterlassen eines für die Bildung ausgewogenen Urheberrechts". Siehe Horlacher, Die Creative Commons-Lizenzen 4.0, 2021, S. 202.

häufig eher als aus der Zeit gefallenes juristisches Nadelöhr denn als gesetzliche Erlaubnis für eine datengetriebene und innovative, international ausgerichtete Wissenschaft.

Vor diesem Hintergrund sollten Wissenschaftseinrichtungen und Wissenschaftler/innen urheberrechtsfreie Materialien nicht voreilig oder aus eher allgemeinen Vorsichtigkeitserwägungen in den urheberrechtlichen Schutzbereich einsortieren. Neben beträchtlicher juristischer Unstimmigkeiten (s.o.) kann ein solches Vorgehen zentrale Ziele wie insbesondere Open Access in der Wissenschaft und die breitestmögliche Realisierung der FAIR-Prinzipien in der digitalen Forschungsdatenlandschaft gefährden.

Fazit

Urheberrechtliche Lizenzierungen an urheberrechtsfreien Forschungsdaten entfalten nicht die gewünschten und ggfs. von Textpublikationen bekannten Rechtswirkungen und bergen juristische wie wissenschaftspolitische Risiken. Insbesondere im Interesse von Open Access und der FAIR-Prinzipien ist es nicht wünschenswert, wenn urheberrechtsfreie Forschungsdaten wie urheberrechtlich geschützte Materialien behandelt werden. Daher sollten aus urheberrechtlicher und aus wissenschaftspolitischer Betrachtung urheberrechtsfreie Daten nicht mit üblichen Urheberrechtsvermerken („Urheberschutz“) oder üblichen urheberrechtlichen Lizenzen (z. Bsp. Creative Commons BY ggfs. mit weiteren optionalen CC-Lizenzbedingungen) versehen werden.

5 Sonderproblem: Sui Generis-Datenbankherstellerrecht

Vor allem für Forschungsdatenrepositorien, -zentren und andere FDM-Dienste stellt das Sui Generis-Datenbankherstellerrecht (§§ 87a ff. UrhG) ein erhebliches Sonderproblem dar. Dieser spezifische Datenbankschutz schützt nicht die einzelnen Forschungsdaten in einer Datenbank, sondern (nur) die Datenbank vor ihrer (fast) vollständigen Übernahme. Rechteinhaber dieses Datenbankschutzes sind in der Regel nicht die Forschenden, welche die Forschungsdaten jeweils erheben, sondern die Einrichtung, welche die Datenbank bereit stellt.[30] Eine weitere Schwäche dieses spezifischen Datenbankherstellerrechts ist dessen regionale Geltung in der Europäischen Union, nicht aber in anderen Rechtsordnungen.[31]

Wenn Wissenschaftseinrichtungen nicht ausnahmsweise entgegenstehende Schutzinteressen vorbringen können, sollten diese auf ihre Datenbankherstellerrechte z. Bsp. mit der Lizenz CC0 Version 1.0 Universell (CC Zero)[32] verzichten.

[30] Hartmann, Zur urheberrechtlichen Schutzfähigkeit von Forschungsdaten. In: InTeR – Zeitschrift zum Innovations- und Technikrecht, 2013, S. 199 ff. Abrufbar unter `http://hdl.handle.net/11858/00-001M-0000-0014-1208-E` (abgerufen am 27.04.2021).

[31] Weiterführend dazu Duisberg, Datenhoheit und Recht des Datenbankherstellers, Recht am Einzeldatum vs. Rechte an Datensammlungen, in: Stiftung Datenschutz (Hrsg.), Dateneigentum und Datenhandel, 2019, S. 53 ff.

[32] Vgl. dazu Horlacher, Die Creative Commons-Lizenzen 4.0, 2021, S. 190 ff.

Wissenschaftliche Zielsetzungen der Forschenden und der Wissenschaftseinrichtungen können mit anderen rechtlichen Instrumenten z. Bsp. aus den Bereichen des Wissenschaftsrechts sowie des Arbeits-, Dienst-, Förder- und Vertragsrechts wirksamer und vor allem wissenschaftsgeleitet erreicht werden.[33]

6 Handlungsempfehlungen für das FDM

1. Bevor Lizenzstrategien und -empfehlungen entwickelt werden, sollte im fachlichen oder institutionellen FDM juristisch geklärt werden, ob Forschungsdaten überhaupt urheberrechtlich geschützt sind (s.o. Kap. 3).

2. Soweit Forschungsdaten – wie jedenfalls in Kernbereichen des SDC MoMaF – urheberrechtsfrei sind, sollten diese nicht mit urheberrechtlichen Lizenzen versehen werden. So können juristische Unsicherheiten und Missverständnisse sowie Hemmnisse für wissenschaftspolitische Ziele wie Open Access oder FAIR-Prinzipien im FDM vermieden werden (s.o. Kap. 4).

3. An Universitäten, Forschungseinrichtungen und Datenzentren, die in nicht unerheblichen Teilen mit urheberrechtsfreien Forschungsdaten operieren, kann eine urheberrechtliche Klarstellung erreicht werden, indem z. Bsp. die Lizenz CC0 Version 1.0 Universell (CC Zero)[34] bzw. eine „Public Domain Mark“[35] bei den Forschungsdaten vermerkt wird.

4. Die Wissenschaftseinrichtungen sollten ihre unabhängig von den Forschungsdatensätzen bestehenden Datenbankherstellerrechte (§§ 87a ff. UrhG) grundsätzlich mit einem lizenzrechtlichen Verzicht (z. Bsp. CC0) kennzeichnen (s.o. Kap. 5).

5. Regelungsbedarfe[36] an urheberrechtsfreien Forschungsdaten bspw. für abgestufte Zugriffs- und Verfügungsrechte, zur Namensnennung bzw. Quellenangabe, zur Sicherstellung der jeweiligen fachlichen guten wissenschaftlichen Praxis oder bei Datenveröffentlichungen sollten in anderen Rechtsbereichen als dem Urheberrecht erfüllt werden. Während das Urheberrecht eher wissenschaftsfern durch die Bundespolitik vorgegeben wird, haben es die Wissenschafts- und Infrastruktureinrichtungen selbst in der Hand forschungsfreundliche FDM-Bestimmungen etwa im Sinne der

[33] Eine Landkarte mit den vielfältigen für FDM relevanten Rechtsgebieten siehe bei Hartmann, Terra Incognita – digitale Forschungsdaten auf der Suche nach einer rechtlichen Heimat, 2018. Abrufbar unter `https://www.forschungsdaten.org/index.php/Datei:Hartmann_TerraIncognita-Forschungsdaten-RechtlicheHeimat.pdf` (abgerufen am 27.04.2021).

[34] Zu spezifisch dieser Lizenz vgl. Horlacher, Die Creative Commons-Lizenzen 4.0, 2021, S. 190 ff.

[35] Vgl. z. Bsp. Pachali, Creative Commons führt „Public Domain Mark” für gemeinfreie Werke ein. Portal iRights.info, 2010. Abrufbar unter https://irights.info/artikel/creative-commons-fhrt-public-domain-mark-fr-gemeinfreie-werke-ein/6925 (abgerufen am 27.04.2021).

[36] Zu den „beachtlichen Lücken“ der urhebergesetzlichen Bestimmungen für wissenschaftliche Arbeiten vgl. auch Herrmann/Trottier, Urheberrecht und Werkqualität: Ein Überblick aus der Wissenschaftspraxis. In: Forschung & Lehre 25, 2 (2018). Abrufbar unter `https://www.forschung-und-lehre.de/urheberrecht-und-werkqualitaet-326/` (abgerufen am 27.04.2021).

FAIR-Prinzipien mit auch internationaler Ausrichtung in ihrem Wissenschafts- und Hochschulrecht, im Arbeits- und Dienstrecht, bei Förderbedingungen und Kooperationsverträgen zu entwickeln und verbindlich festzulegen.[37] Die FDM-Stellen können so zum rechtlich Motor avancieren und z. Bsp. mit Forschungsdaten-Policies[38] wichtige Leitgedanken einsteuern.[39]

Acknowledgements

Dieser Beitrag entstand im Rahmen des vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg geförderten Forschungsdatenzentrums Science Data Center für Molekulare Materialforschung (SDC MoMaF) am Karlsruher Institut für Technologie (KIT). Besonderer Dank gebührt Dr. Nicole Jung vom Institut für Organische Chemie am KIT für die Erläuterungen der relevanten Forschungs(daten)abläufe bei Chemotion.

[37] Ähnlich auch Arbeitskreis FDM der baden-württembergischen Universitäten, Leitfaden – Verantwortungsvoller Umgang mit Forschungsdaten, 2021, S. 6 f. Abrufbar unter `https://www.forschungsdaten.info/nachrichten/nachricht-anzeige/leitfaden-zum-verantwortungsvollen-umgang-mit-forschungsdaten/` (abgerufen am 27.04.2021).

[38] Zu Policies in der Wissenschaft als Soft Law und Vorstufe für juristisch bindende Regelungen vgl. Hartmann, in Session: Umsetzung und Erfahrung mit Richtlinien und Guidelines (Vortragsaufzeichnung Open-Access-Tage 2014). Abrufbar unter `https://youtu.be/BegYmuqD804` (abgerufen am 27.04.2021).

[39] Für eine aktive Rolle des institutionellen FDM bei Rechtsfragen siehe Hartmann, Rechtsfragen: Institutioneller Rahmen und Handlungsoptionen für universitäres FDM, 2019. Abrufbar unter `https://doi.org/10.5281/zenodo.2654305` (abgerufen am 27.04.2021).

B2FIND – Searching for Research Data across Disciplines

Claudia Martens[1] and Markus Demleitner[2,3]

[1]Deutsches Klimarechenzentrum, Abteilung Datenmanagement
[2]Universität Heidelberg, Zentrum für Astronomie
[3]German Astrophysical Virtual Observatory GAVO

B2FIND is a cross-disciplinary data search engine developed within the pan-European Collaborative Data Infrastructure EUDAT CDI. It has been the central indexing tool for EOSC-hub and it plays a major role in the Data Infrastructure Capacity for EOSC (DICE). Technically, it is a comprehensive joint metadata catalogue that includes metadata records for data stored in various data centres and using different schemas on widely differing granularity. Hence, the major challenge B2FIND faces is bridging the semantic and structural gaps between different disciplines' practices. This contribution discusses how these bridges were built using the example of the Virtual Observatory, a global data infrastructure in Astronomy which itself looks back on 20 years of evolution in metadata management.

1 Introduction

The first letter in the FAIR acronym stands for "Findable", and indeed providing interfaces and APIs that allow for various modes of research data discovery is a very basic requirement for any sort of data infrastructure. Since data structures differ a lot between disciplines, so do the metadata schemas of disciplinary data infrastructures.

On the other hand, data discovery spanning disciplines promises rich rewards in terms of cross-fertilising research. A federation of the various tailored discovery models in the disciplines allows them to retain those while still enabling the cross-disciplinary discovery.

B2FIND has been providing data discovery federated across many disciplines since 2014. This contribution shows how the federation is effected, both centrally at B2FIND (section 2) and on the level of the individual disciplines, which can, to some extent, already accomodate some of the requirements of cross-disciplinary data discovery (section 3). We conclude with a brief look at the bigger picture: What parts of the picture of cross-disciplinary data discovery are still not quite in place?

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13729 (CC BY-SA 4.0)

2 B2FIND – an interdisciplinary discovery portal for research data

2.1 Description

B2FIND is a discovery service for research data distributed within the European Open Science Cloud (EOSC) and beyond. It was built in 2014 as part of the European Union Research and Innovation programme Horizon2020 for the European Data e-Infrastructure Initiative (EUDAT), which is now the pan-European Collaborative Data Infrastructure (EUDAT-CDI), currently consisting of 28 partners, including most major European data centres and research organisations[1]. B2FIND has been the central indexing tool for EOSC-hub[2] and is a basic service of the Data Infrastructure Capacity for the EOSC[3] (DICE).

To fulfill these roles, a comprehensive metadata catalogue was built that includes metadata records for data that is stored in various data centres, using different data formats and metadata schemas on widely divergent granularity levels – where resources are sometimes single files, sometimes complex hierarchies involving millions of files –, representing all kinds of scientific output: from huge netCDF files produced by climate modeling to small audio records of Swahili syllables and phonemes; from the immigrant panel data in the Netherlands to a paleoenvironment reconstruction of the Mozambique Channel and from an image of the "Maison du Chirurgien" in ancient Roman Pompeii to an Excel file giving concentrations of calcium, magnesium, potassium and sodium in throughfall, litterflow and soil in an Oriental beech forest.

In order to enable this interdisciplinary perspective, different metadata formats, schemas and standards are homogenized on the B2FIND metadata schema[4], which is based on the DataCite schema[5] extended with the additional elements `<Discipline>` and `<Instrument>`, allowing users to search and find research data across scientific disciplines and research areas as well as searching for certain measurement tools, e.g., data produced by specific beamlines or measurement stations.
Additionally, the B2FIND schema includes a`<TemporalCoverage>`that allows users to search for a certain time range research data is related to.

Good metadata management is guided by FAIR principles, including the establishment of common standards and guidelines for data providers[6].

In this effort, close cooperation and coordination with scientific communities, research infrastructures and other initiatives dealing with metadata standardisation (OpenAIRE Advance, RDA interest and working groups, and several EOSC related projects) are es-

[1] `https://www.eudat.eu/eudat-cdi`
[2] `https://www.eosc-hub.eu`
[3] `https://www.dice-eosc.eu`
[4] `http://b2find.eudat.eu/guidelines/mapping.html`
[5] `https://schema.datacite.org/`
[6] `http://b2find.eudat.eu/guidelines`

sential in order to establish standards that are both reasonable for community-specific needs and usable for enhanced interoperability.

The main question still is how to find a balance between community-specific metadata that serves their communities' needs on the one side and a metadata schema that is sufficiently generic to capture interdisciplinary research data, but at the same time is specific enough to enable targeted queries yielding results useful to the querier's research. This balance is not a static point, but rather an ongoing process depending on new technical developments as well as on political decisions. Even within the European Open Science Cloud consensus on a "Core Minimum Set" has not been reached yet, and what techniques for metadata exposure and exchange will prevail has yet to be seen.

2.2 Workflow

B2FIND's workflow for metadata ingestion basically consists of three steps: harvesting metadata, mapping them and uploading the final JSON records to a database for indexing and search. These steps are briefly described here in order to provide additional perspective on why the close community involvement described in sect. 3 is helping B2FIND.

2.2.1 Harvesting

Preferably, B2FIND uses the Open Archives Initiative Protocol for Metadata Harvesting OAI-PMH [8] to harvest metadata from data providers. OAI-PMH offers several options that makes it a suitable protocol for harvesting:

(a) a facility to define diverse metadata prefixes (minimal requirement is Dublin Core, further prefixes are optional),

(b) a facility to create subsets for harvesting (useful for large amounts of records or divergent records, e.g., from different projects or sites or measurement stations), and

(c) a facility to configure incremental harvesting.

Nonetheless, B2FIND supports other harvesting methods as well, e.g., the Open Geospatial Consortium Catalogue Service for the Web (OGC-CSW) or various REST APIs. Harvesting triples from SPARQL endpoints is implemented only in a beta version.

2.2.2 Mapping

The mapping process is twofold as it includes a format conversion as well as a semantic mapping based on standardized vocabularies (e.g., the field "Language" is mapped onto

ISO 639 codes, and "Discipline" is mapped on a standardized closed vocabulary[7]). First, entries from XML (or JSON) records are being parsed to assign them to the keys specified in the B2FIND schema. The resulting key-value pairs are stored in JSON dictionaries and checked/validated before being uploaded to the B2FIND repository. Formerly depending on XPATH rules, the current version of the B2FIND ingestion software[8] includes a very flexible mapping procedure: several harvesting endpoints using different metadata standards may be integrated within one "Community". For each endpoint distinct mapping "issues" may be implemented (e.g., specifically defined methods for certain values and attributes, perhaps for representing different sorts of `<identifier>`, for using the header datestamp as `<PublicationYear>` or assigning additional `<keyword>`s)[9].

Currently B2FIND supports generic metadata schemas such as DataCite and Dublin Core. Community-specific metadata schemas are supported as well, e.g., ISO19115/19139 (which is the basis for Inspire) and FGDC (which is a DublinCore crosswalk for ISO 19135) for Environmental Research Communities, or FF for Nordic Archaeologists. In principle, our ingestion software allows us to integrate any metadata schema and we support the implementation of new schemas. The integration of DDI for Social Sciences is currently developed within the frame of a FAIRsFAIR project that aims to improve interdisciplinary research data discovery using DDI-CDI (an "enhanced" version of DDI attempting to make metadata interoperable across disciplines) and DCATv2 (an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web, also a W3C Recommendation).Integration of further RDF vocabularies in our metadata ingestion source code is planned but depends on the option to harvest triples, which again requires resources for software development.

2.2.3 Upload and indexing

B2FIND's search portal and GUI is based on the open source portal software CKAN[10], which comes with an Apache Lucene SOLR servlet. We use it to index the mapped JSON records and offer performant faceted search functionalities. CKAN has a very limited internal metadata schema which has been enhanced for B2FIND by creating additional metadata elements as CKAN "extra" fields.

B2FIND offers a full text search, in addition to which results may be narrowed down using 12 (as of mid-2021) facets, including spatial and temporal search (using a map and an extension for "Timeline Search") and Communities, Keywords, Creator, Publication Year, Discipline, Language, Publisher, Contributor, Resource Type, and Information on licensing and/or access restrictions.

[7] As there is no useful generic classification for scientific disciplines (except perhaps simple taxonomies issued by funding bodies), the B2FIND closed vocabulary is based on re3data's subject schema. A revision is currently ongoing as part of a collaboration project Classification for Research Areas, clara.science.

[8] `https://eudat.eu/news/eudat-unveils-new-improved-b2find-30`

[9] All B2FIND software is openly accessible on Github: `https://github.com/EUDAT-B2FIND`

[10] `https://ckan.org/`

"Community" here is the data provider or research infrastructure that B2FIND harvests from. Using a new OAI-PMH extension for CKAN, all metadata records within our database are also harvested by OpenAIRE, widening the scope of discoverability for research objects. B2FIND in this way acts as a metadata aggregator, exposing metadata that are (in some cases) "hidden" in data repositories which are not crawled by big players like Google[11].

2.3 When is a standard a standard?

The terms metadata "schema" and metadata "standard" are often used interchangeably, and both refer to "the formal specification of the attributes (characteristics) employed for representing information resources" [1]. Thus a metadata schema could be seen as a set of elements with a precise semantic definition and optionally rules how and what values can be assigned to these elements [5]; a metadata standard then is a schema which is developed and maintained by an institution that is a standard-setting one. That means that "a standard is a standard insofar as there is an institutional or organizational standardization unit developing and maintaining a standard - whereas all parties and persons involved agree this institution to be trustworthy and reliable" [9].

Metadata standards evolve in different ways. One way is to enforce their adoption through policies such as EU directive 2007/2/EC[12], which requires institutions and organisations publishing spatial datasets and services to implement Inspire as their metadata schema. Another way is to develop a metadata schema that is useful and thus widely adapted. DataCite may serve as an example here. Nonetheless, as new methods and techniques come into use standards need to continually be developed and adjusted, which is hard work (still) done by humans. As there is no one-and-only standard, interoperability is key.

Reliability is crucial for all aspects of FAIR data principles, escpecially for persistent identification of digital resources. However, while DataCite allows only one type of value for `<identifier>` (its DOI), B2FIND has found it necessary to admit several other types, too (in the present case, the IVOA identifiers discussed below). The internal ranking is: if a DOI is offered it will be displayed; if another PID is offered this will be displayed; if neither DOI nor other PID are offered, B2FIND will display whatever URI is given as `<source>`.

A metadata schema may also be defined as a "logical plan showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata specifically as regards the semantics, the syntax and the optionality" [6], where "syntax" describes the structure of a schema and "semantics" describes the

[11] There are many arguments why a proprietary index like Google dataset search should be viewed with suspicion, many of which have been pointed out elsewhere (e.g., [7]). We mention in passing that Google's revenue from advertisment for less then a second corresponds to the funding B2FIND gets for a year.

[12] `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A32008R1205&from=EN`

meaning of its elements, properties or attributes. We will follow this definition here and try to describe both low-hanging fruit and concrete obstacles in syntax and semantics mapping.

3 Case Study: Mapping VOResource for B2FIND

In practice, communities with discipline-specific metadata schemas are encouraged to help B2FIND by building their half of the bridge to the other communities themselves – in particular because that is usually much simpler for domain experts than for a generic infrastructure. A discipline-specific service will then emit already processed metadata in, for instance, the DataCite schema or possibly even directly in B2FIND's schema; this will in general be a lossy transformation.

```
1  <ri:Resource created="2020-11-17T11:00:00Z"
2      updated="2021-01-29T13:52:42Z"
3      status="active" xsi:type="vs:CatalogResource">
4    <title>Gaia eDR3 source catalogue "light"</title>
5    <identifier>ivo://org.gavo.dc/gaia/q3/edr3lite</identifier>
6    <curation>
7      <publisher>The GAVO DC team</publisher>...</curation>
8    <content>
9      <subject>astrometry</subject>
10     <description>This is a "light" version...</description>
11     <referenceURL>http://dc.g-vo.org/tableinfo/gaia.edr3lite</referenceURL>
12     <relationship>
13       <relationshipType>served-by</relationshipType>
14       <relatedResource ivo-id="ivo://org.gavo.dc/tap">
15         GAVO Data Center TAP service</relatedResource>
16     </relationship>...</content>
17   <capability standardID="ivo://ivoa.net/std/TAP#aux">
18     <interface role="std" version="1.1" xsi:type="vs:ParamHTTP">
19       <accessURL use="base">http://dc.g-vo.org/tap</accessURL>
20     </interface>
21   </capability>
22   <facility>Gaia</facility>
23 <ri:Resource>
```

Figure 1: A sketch of a VOResource metadata record. VOResource includes Dublin Core (e.g., lines 3, 7, 9, 10), and extends it towards items present in DataCite at least in compatible form (e.g., lines 12ff). The core operational metadata (here, the capability in lines 17ff) does not map to anything outside of the VO.

As an example, in this section we discuss the integration of the records in the Virtual Observatory's Registry into B2FIND. As discussed in [2], the native metadata schema in the Virtual Observatory (VO) is built on the discipline-specific standard VOResource

[10] and its extensions. A (severely abridged) sketch of a record is shown in Fig. 1[13]. For illustration, we give the result of the mapping of this information to the oai_datacite-like XML consumed by B2FIND in Fig. 2.

When converting metadata from one schema to the other, there are five major categories of items:

- Elements with exactly matchable semantics
- Elements having simply mappable semantics
- Elements having having similar semantics but requiring more complex mapping procedures
- "Important" elements missing in the source schema
- "Important" elements missing in the target schema

We will discuss examples for each of these in turn.

3.1 Exactly Matching Semantics

When source and target schema have elements with exactly matching semantics, the conversion is a very simple syntactic operation. Even in our example, where both VOResource and DataCite are siblings with respect to their common parent Dublin Core, that is a rare exception. Only the `publisher` matches down to structural details (Fig. 1, line 7 vs. Fig. 2, line 3), and even there some adjustment is necessary, because the element directly sits in the record root in DataCite, where in VOResource it is part of a `<curation>` child element.

Usually, the different requirements in the domain show on the structural level. For instance, in VOResource's all-English world of professional astronomy, a single title is enough (Fig. 1, line 4). The more general DataCite admits multiple titles (Fig. line 2, line 2). As long as the target format admits a superset of the source format's features, this would only turn into a problem if the conversion would ever need to be reversed (e.g., turning DataCite to VOResource); in our case, this does not seem a likely requirement.

Sometimes, a single element in the source schema maps to a pair of element and attribute name in the target schema, as when adorning VOResource's `<description>` (Fig. 1, line 10) with an `descriptionType="abstract"` attribute (Fig. 2, line 16ff). Again, these are operations not trivially invertible.

None of these present technical challenges; the VO's B2FIND interface simply re-uses pre-existing XSLT[14] originally written to facilitate DOI minting from VO resource records.

[13] See `http://dc.zah.uni-heidelberg.de/getRR/gaia/q3/edr3lite` for a full version.

[14] `https://github.com/msdemlei/datalink-xslt`

```
1  <d:resource>
2    <d:titles><d:title>Gaia eDR3 source catalogue "light"...</d:title></d:titles>
3    <d:publisher>The GAVO DC team</d:publisher>
4    <d:publicationYear>2020</d:publicationYear>
5    <d:subjects>
6      <d:subject>observational-astronomy</d:subject></d:subjects>
7    <d:resourceType resourceTypeGeneral="Other"/>
8    <d:alternateIdentifiers>
9      <d:alternateIdentifier alternateIdentifierType="ivoid">
10       ivo://org.gavo.dc/gaia/q3/edr3lite</d:alternateIdentifier>
11     <d:alternateIdentifier alternateIdentifierType="reference URL">
12       http://dc.g-vo.org/tableinfo/gaia.edr3lite</d:alternateIdentifier>
13   </d:alternateIdentifiers>
14   <d:relatedIdentifiers/>
15   <d:formats/>
16   <d:descriptions><d:description descriptionType="Abstract">
17         This is a "light" version...
18   </d:description></d:descriptions>
19 </d:resource>
```

Figure 2: A sketch of the DataCite-like metadata record generated from the VOResource shown in Fig. 1. The various classes of changes between this and the original record are discussed in the text.

3.2 Simply Mappable Semantics

VOResource has the notion of a reference URL (Fig. 1, line 11), which should resolve to a web page with human-readable information on the resource. Also, its identifier (an "ivoid" or IVOA identifier; Fig. 1, line 5) does not directly map to anything in DataCite. For the B2FIND mapping, we decided to map them into alternate identifiers (with types of "reference URL" and "ivoid"; Fig. 2, line 8ff). While for the ivoid, this obviously is the right thing to do – what software there is that knows how to resolve ivoids can easily be taught to pick up the particular alternate identifier type –, the reference URL only passes as an identifer in a very abstract sense; in reality, it would fit a "documentation"-like link a lot better.

However, the B2FIND-internal mapping can easily be taught to pull the reference URL from the alternate identifier (presenting it as the main URI for the result). Again, the pre-existing XSLT mentioned above is sufficient for these transformations.

3.3 Complex Mapping

The case of `<subject>` (Fig. 1, line 9 vs. Fig. 2, line 6) is more complex; in the example, the keyword "astrometry" turns into "observational-astronomy".

Both VOResource and DataCite admit multiple subject elements (albeit in different positions), but expecting a cross-disciplinary search engine to deal sensibly with the keyword

schemes of all disciplines contributing is probably not realistic (although having all the respective vocabularies uniformly in RDF would probably help in such an effort).

Instead, B2FIND has a list of relatively coarse-grained subject keywords. For astronomy, it includes the top-level terms from the Unified Astronomy Thesaurus [4]. Hence, when preparing records for B2FIND, the Virtual Observatory side adds the top-level terms for all subject keywords found where these keywords can be located in the UAT. This is a relatively complex operation using external resources (the UAT vocabulary, database queries) that would hence be very hard indeed to implement in XSLT. Hence, this part is put directly into the code serving the oai_datacite metadata format[15].

3.4 Lacunae in the Source Schema

The most debatable part of the VO-B2FIND interaction is that so far the oai_datacite metadata schema is used with invalid records. That is because most of the resources in the VO do not have DOIs assigned at this point, and even those that have a DOI do not always declare it in their VOResource. However, oai_datacite requires an `<identifier>` element, and it must contain a DOI, so the VO's unique (though not persistent) identifier (Fig. 1, line 5) cannot be used.

The solution for now is to skip the identifier element, thus producing invalid DataCite records. As a bespoke practice between two parties, this works well enough, and it will confuse no generic clients with non-DOI material in identifier elements. As the VO moves to offering the B2FIND metadata schema, this can be cleaned up.

Much less serious is the lack of various optional metadata such as DataCite's `<formats>`. This particular element is not provided because in the VO, almost all resources come as services, which are primarily operated using protocols typically emitting VOTables (an XML-based format for tables with a focus on rich metadata). However, there is often underlying science data, which then presumably is of the primary interest to the researcher. But it is then hard to predict which formats are available, and VOResource records generally do not give the respective metadata. Within the VO, where data formats are relatively standardised (e.g., images will almost always come in FITS), this is not a problem; beyond the VO, it might be.

3.5 Lacunae in the Target Schema

In actual data discovery in the VO, the arguably most important items are the capabilities (Fig. 1, lines 17f), which define what standard protocols can be used to query the data within the resource described, and the tableset (not in the example), which defines the underlying table structure.

[15]The technical details are discussed in a blog post at `https://blog.g-vo.org/semantics-cross-discipline-discovery-and-down-to-earth-code/`.

Both items are not representable in DataCite, and for the capabilities, this is probably not even desirable, as they can, in general, only be used with specialised clients or libraries, which presumably are not available to users from outside astronomy. An exception is when VO resources declare capabilities that offer a web browser interface; in that case, the XSLT produces a URL-typed alternate identifier.

A perhaps less dramatic problem is what makes the relationship declarations in VOResource (Fig. 1, lines 12ff) disappear. What this declares is that the data collection can be queried through a TAP service (cf. [2]). Because DataCite focuses on data rather than services, there is no relationship type "there is a service for the present data" in the controlled vocabulary for relationType. Hence, this particular relationship cannot be declared in the DataCite record, even though the modelling of relationships in general is rather similar in VOResource and DataCite, and DataCite even allows record producers to define the related resources through generic URIs rather than DOIs; this latter property is used by the conversion XSLT to translate several other types of relationships.

Another important piece of metadata is the coverage in space, time, and spectrum, which for the example resource might look like this:

```
<coverage>
  <spatial>0/0-11</spatial>
  <temporal>57174 57174</temporal>
  <spectral>1.986e-19 4.966e-19</spectral>
</coverage>
```

Decoded, this means "data on the whole sky" (`0/0-11` is a representation of that in an astronomy-specific format called MOC [3] that allows operators to define very complex maps on the sky), for midnight 2015-05-31 (`57174` is a "Modified Julian Date" designating that point in time; to be very exact, one should add that this uses a time scale called TDB, and the clock sits in the solar system's center of mass, though for data discovery, this sort of precision probably does not matter), and that the data pertains to the (somewhat augmented) optical band between 400 and 1000 nm (it is the photon energy in Joule).

While temporal and spectral coverage could probably easily be made usable outside of astronomy, for the spatial coverage this seems a lot more problematic. DataCite's GeoLocation could perhaps be extended (somewhat oxymoronically) to sky coverages, but turning complex MOCs into polygons is only approximatively possible and probably unrealistic. Whether MOCs have sufficient utility outside of astronomy to make support for them desirable in a cross-discipline service is, on the other hand, rather questionable.

4 Concluding Remarks

This last question can be put into more general terms: "How do we find a balance between community-specific metadata that serves the needs of specific communities well but is un-understandable to non-experts on the one side, and a metadata schema that is

sufficiently generic to enable meaningful, possibly even blind, interdisciplinary research data discovery?" We do not claim to have an answer, but we tentatively suggest that both the specialised and the generic schemas are needed and that one size will probably never fit all, which would then suggest that the present co-existence of discipline-specific and federating infrastructure is here to stay. Also, developing a generic metadata schema suitable for expressing metadata relevant for cross-disciplinary discovery is an evolving, ongoing process. What is already clear at this point is that creating machine-readable, mappable metadata is hard and requires experts assisting data publishers.

In closing, let us mention that we feel even the most fundamental question, "How do we enable reuse of research data across disciplines?", is not satifactorily answered at this point. Using data usually is quite a bit harder than reading a paper, which very typically already is hard for non-experts in some sub-discipline. Hence, as perhaps already discernible from our emerging collection of (mostly fictitious) user stories[16], cross-discipline data discovery quite likely will more often than not involve expert discovery.

Acknowledgements

Part of the work reported on here was funded by the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Bibliography

[1] G. Alemu and B. Stevens. *An Emergent Theory of Digital Library Metadata - Enrich then Filter*. Amsterdam: Elsevier, 2015.

[2] M. Demleitner. Practical Interoperability in the Virtual Observatory. In *Proceedings of the 3rd Heidelberger e-Science-Tage (this volume)*, 2021.

[3] Pierre Fernique, Thomas Boch, Tom Donaldson, Daniel Durand , Wil O'Mullane, Martin Reinecke, and Mark Taylor. MOC - HEALPix Multi-Order Coverage map Version 1.1. IVOA Recommendation 07 October 2019, October 2019.

[4] Katie Frey and Alberto Accomazzi. The Unified Astronomy Thesaurus: Semantic Metadata for Astronomy and Astrophysics. *Astrophysical Journal Supplement*, 236(1):24, May 2018. arXiv:1801.01021, https://doi.org/10.3847/1538-4365/aab760

[5] B. Haslhofer and W. Klas. A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, 42 (7), 2010.

[6] ISO TC46/SC 11. Managing metadata for records – part 3: Self-assessment method. ISO/TR 23081-3:2011, 2011. URL: https://www.iso.org/standard/57121.html.

[16] https://github.com/msdemlei/cross-discipline-discovery

[7] Peter Kraker. #dontleaveittogoogle - how open infrastructures enable continuous innovation in the research workflow. Presentation held at Open Science Conference, March 19th-20th 2021, Berlin., 2019. URL: https://www.open-science-conference.eu/wp-content/uploads/2019/03/Peter-Kraker.pdf.

[8] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. The open archives initiative protocol for metadata harvesting, version 2.0, 2002. URL: http://www.openarchives.org/OAI/openarchivesprotocol.html.

[9] Erwann Quimbert, Keith G. Jeffery, Claudia Martens, Paul Martin, and Zhiming Zhao. Data cataloguing. In Z. Zhao and M. Hellström, editors, *Towards Interoperable Research Infrastructures for Environmental and Earth Science*. Springer, 2020. https://doi.org/10.1007/978-3-030-52829-4_8

[10] Raymond Plante, Markus Demleitner, Kevin Benson, Matthew Graham, Gretchen Greene, Paul Harrison, Gerard Lemson, Tony Linde, and Guy Rixon. VOResource: an XML Encoding Schema for Resource Metadata Version 1.1. IVOA Recommendation 25 June 2018, June 2018. https://doi.org/10.5479/ADS/bib/2018ivoa.spec.0625P.

Searching Research (Meta-)Data using Semantic Web Technologies

Sarah Bensberg and Marius Politze

IT Center, RWTH Aachen University

Research data can be described with RDF based metadata following discipline specific application profiles. To make the best use of this data stocks according to the FAIR principles, users should be able to search within these metadata sets. Among different approaches it is investigated whether mapping the RDF data into a search index affects the quality of search. The evaluation indicates that approaches are either easy to use but slow or ineffective, or fast and effective but difficult to use. With the transformation of RDF data into a search index, a solution was found that combines benefits of the approaches.

1 Introduction

Digital transformation affects all areas of society: More data is produced and workflows rely on data analysis leading to new challenges in data management. Within the context of research data management, the National Research Data Infrastructure aims to systematize research data stocks and make them accessible. RWTH Aachen University supports this effort with the development of the research data management platform Collaborative Scientific Integration Environment (Coscine): Research data is made accessible independent of the actual storage location and described with metadata that allows a structured search. The goal is to make the data Findable, Accessible, Interoperable and Reusable (FAIR) [1].

1.1 FAIR Metadata

An important aspect is the diversity and heterogeneity of the various scientific disciplines as well as the decentralized nature of data and IT Services [2]. Discipline-specific application profiles are required and should be available to researchers early in the Research Data Life Cycle (RDLC) since implicit knowledge is not passed on and often lost [3]. These application profiles are typically based on existing metadata standards.

The FAIR Digital Object provides a framework for the layout and interleaving of identifiers, metadata and data using a describing record as part of a data identifier [4] in such

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13730 (CC BY-SA 4.0)

a distributed system: Several infrastructure components are available for this purpose: (i) the Persistent Identifier (PID) for registration, permanent and unique referencing, (ii) external "linked" metadata records stored with the PID, which enables the use of discipline-specific systems, (iii) a storage system in which researchers can store the actual bit sequences of their research data, and (iv) the means to publish the results in a suitable repository.

To simplify data exchange and machine processing, the World Wide Web Consortium specifies Semantic Web technologies like Resource Description Framework (RDF) [5] to provide the means to describe digital resources. RDF is suitable because its schema-lessness provides flexibility for changes and different disciplines. In order to follow defined metadata schemas, RDF needs to be restricted by application profiles that allow validation as offered by Shapes Constraint Language (SHACL) [6]. While this makes metadata available as a RDF-based knowledge graph, it presents some general challenges described by Arnaout and Elbassuoni [7]:

- *Data incompleteness*, since RDF triples contain a lot of information, but most knowledge is represented as free text in literal nodes.
- *Inflexible querying*, because triple-pattern queries are highly expressive, but at the same time restrictive, because they have to follow a certain structure.
- *Missing result ranking*, queries may yield results in arbitrary order; a ranking is not provided by SPARQL Protocol And RDF Query Language (SPARQL) and has to be added explicitly.
- *Result diversity*, because it is important that the topmost results give a broad overview of the results of a query and do not all contain similar aspects.

The searchability and thus the sub-goal reusability of the FAIR Guiding Principles is achieved by an associated query language SPARQL [8], requiring considerable technical knowledge. There are two main approaches to overcome this obstacle: the systematic generation of a query or the mapping of RDF (meta-)data into a search index for use in a search engine.

1.2 Hypothesis

Based on the state of research on RDF searches or general Information Retrieval (IR) and taking previous approaches into account, the work is oriented towards the following hypothesis, which is to be tested:

Mapping RDF-based metadata records into a search index for use in a search engine improves the quality of the search and results for the user compared to using generated SPARQL queries.

To validate this hypothesis, different approaches to create an efficient semantic search engine for metadata in RDF-based knowledge graphs are evaluated concerning the resulting quality of the search.

1.3 Assessment of the Quality of Search

In this context, "quality" is regarded in multiple dimensions ranging from user experience when entering their search intention to the results returned and the technical abilities of the system.

Effectiveness According to International Organization for Standardization (ISO) [9]. Referring to search results, a system should find correct data records and satisfy the search intention.

Usability According to ISO [10]. For search, it is therefore primarily a matter of the User Interface (UI) to be appealing and understandable for the user.

Complexity of the creation of a search query is for the user and expected preliminary knowledge.

Efficiency According to ISO [9]. For search, this refers to the effort formulating a query compared to the generated results.

Response Time Nielsen [11] defines two limits: 0.1s for a direct response to an interaction and 1.0s to notice the delay but remain uninterrupted.

Scalability It is expected, that the data stock is growing over time. Hence, the system should be able to handle growing quantities of data.

Additional Effort This requirement assesses how much additional storage or computing resources are necessary to realize the search.

2 Related Work

To generate a better overview, related approaches are divided into different categories, whereby a clear separation is not always possible or ambiguous.

2.1 Retrieval of Entities in the Semantic Web Using Keywords

Swoogle is a crawler-based indexing and retrieval system for RDF documents [12]: an IR system is applied, which uses either character *n-grams* or Uniform Resource Identifiers (URIs) as terms to find documents given the search terms and computes a rank indicating the documents' importance. *Sindice* indexes RDF documents based on resource URIs, Inverse Functional Propertys (IFPs) and keywords [13]. The focus of the approach is retrieval of ontologies and documents in the Semantic Web. Similarly, *SWSE* is a search engine that allows finding and navigating in an object-oriented search space by keyword-based input [14]. *Falcons* provides keyword-based search for concepts and objects on the Semantic Web by indexing all the terms of an entity's virtual document consisting of its names, associated literals, and names of neighboring entities [15]. While in the

context at hand, metadata records are directly persisted in an RDF triplestore and do not need to be indexed, the approaches demonstrate the distinction between object-based and document-based search.

2.2 Generating a SPARQL Query

A very basic approach is to generate a SPARQL query from the keyword-based search of a user where predefined basic graph pattern templates and a mapping of keywords to possible URIs and background information about the existing data structure are used [16]. This shows that background knowledge and structures are helful to generate adequate queries.

SPARQL query builders allow a step-by-step construction of SPARQL queries using cleverly chosen user interfaces. Kuric et al. have compared the best-known query builders regarding their usability for laypeople and divided them into three different categories [17]: (i) *Form-Based Query Builders* use input fields to build the SPARQL query step by step. Classes and objects must either be advised or made available for selection by the user. *ExConQuer* [18], the *Linked Data Query Wizard* [19], *VizQuery* [20], and the *Wikidata Query Service (WQS)* [21] are examples of such approaches. (ii) *Graph-Based Query Builders* make use of a visual and graphical user interface that guides the user in creating valid SPARQL queries. Some examples of these approaches are *iSPARQL* [22], *NITELIGHT* [23], *OptiqueVQS* [24], and *QueryVOWL* [25]. (iii) *Natural Language-Based Query Builders* build the query based on the natural language of the user. *NLP-Reduce* [26], *SPARKLIS* [27], *DEANNA* [28], *AskNow* [29], *Swip* [30] and *ScoQAS* [31] are examples for a Natural Language (NL)-based approaches that also allow generation of SPARQL queries. However, the apporaches focus on enhancing NL approaches and not so much on the quality of search so these are excluded from further consideration.

Faceted Search is used to allow flexible navigation through a large search space and to limit this space by so-called facets [32]. Faceted search can also be applied to RDF knowledge graphs by offering properties (predicates) of entities (subjects) as facets with corresponding values (objects). *OSCAR* [33], the OpenCitations RDF Search Application, and *SemFacet* [34] are examples of such an approach.

The *Assisted SPARQL Editor* leverages the graph summary to support the user in effectively formulating complex SPARQL queries [35]. For this purpose, suggestions for classes, predicates, relationships between variables, and named graphs are provided to the user. *SemSearch* is a search engine that translates user input from a simple syntax into a formal SPARQL query [36]. While loosing much of the expressiveness of SPARQL. *SINA*, a keyword-based search system, uses a hidden markov model to transform input into a SPARQL query [37]. *AutoSPARQL* uses active supervised machine learning to formulate a SPARQL query [38].

2.3 Generating a Search Index to use IR Techniques

Linked swissbib converts bibliographic data into a RDF-based data model and divides it into six different concepts [39]. By using the *JSON-LD* serialization of RDF, they are indexed in the search engine *Elasticsearch (ES)*. *Open Semantic Search* stores the associated labels of the RDF annotation to URIs for indexing data in Solr and allows a faceted search without generting a SPARQL query. *Semplore* also uses a hybrid query option that allows structured queries and faceted searches [40]. Rocha et al. presented a hybrid approach to keyword-based search in the Semantic Web, where the focus is also on finding concepts using an instance graph for each concept, which is searched with traditional IR techniques [41].

3 Approaches for Searching in RDF Metadata

If RDF data is stored in a triple store, the easiest way to access it is the SPARQL Endpoint that allows querying and filtering data. A user must have some knowledge to do this: (i) RDF and SPARQL, (ii) the structure of the stored data, (iii) used vocabularies and terms as well as their corresponding URIs or Internationalized Resource Identifiers (IRIs)[1]. Even if a user would have these information, the use of SPARQL is very expressive and therefore challenging. It can thus be assumed that the interface can only be used by more experienced Data Curators. Unfortunately, many of the presented approaches are outdated, no longer available or there is no code or exact description for their implementation and realization[2], are domain-specific and would have to be widely adapted for use in the domain independent scenario.

3.1 Full-Text Search

SPARQL offers the possibility to filter data using regular expressions. However, a SPARQL query must be generated internally, which applies the user input to all possible occurrences of a literal somewhere in the graph. Virtuoso also offers the option to do a case insensitive full-text search with the `bif:contains` function [42]. Since only indexed literals can be searched with `bif:contains`, it is not possible to define additional rules which compound the literals on the fly in the constructed SPARQL query. Therefore, not all literals in the graph can be searched without onstructed literals and additional triples directly in the triplestore. However, a strategy would have to be considered how to separate the original metadata record, which is why this variant is not further considered.

[1] An IRI is a generalization of an URI. Since both are used very interchangeably in general linguistic usage, only the term URI is used in the following, even if a IRI would be allowed.

[2] Ironically, this is exactly one of the phenomenons, that should be tackeled by an Research Data Management (RDM) system.

3.2 Faceted Search

The approach of a faceted search is especially interesting because it allows for an incremental refinement of data. At the core, this approach generates a SPARQL query that, by selecting facets and values, directly using URIs and data structures of the knowledge graph. Implementations should consider further aspects like the selection of facts and their representation in the UI. In the case of Coscine these could be metadata fields from application profiles, but the question arises how to select the currently visible application profiles. In the scope of this work, however, the focus is on the result which can be achieved with such an approach rather than the generation of an adequate UI.

3.3 Query Builder

The related work shows a lot of different query builders, which all work a bit differently or have a different focus. They will be evaluated in the work because they can provide the most optimal and direct representation in the knowledge graph by gradually assembling a SPARQL query. Various such approaches have already been extensively evaluated [43, 17, 44]. In general, the generated SPARQL queries are to be considered for the evaluation.

3.4 Elastic Search: Index and Search Engine

Creating a search index allows to benefit from functionalities and optimizations for user input and search features like auto-completions, search suggestions, stemming or tolerance towards spelling errors. Positively evaluating this approach would therefore support the core hypothesis.

A conceptual challenge remains how to map the entities from RDF to documents. The mapping offered by Semplore [40] was enhanced to allow custom inference rules, and the possibility to use ES to benefit from advanced search types like value range queries. The resulting model considers resources, the metadata records, as documents in the search index. In principle this defines an entity-object-mapping as basis for the conversion from RDF to the search index.

To define queries, ES provides a Domain Specific Language (DSL) based on JSON. ES offers two different search syntaxes to directly convert user inputs into a corresponding search query. A simpler variant is ignoring syntactically incorrect input while the advanced syntax is offers more search features while being stricter in terms of syntax.

4 Evaluation

In order to evaluate the discussed approaches for searching RDF metadata a two step approach is followed. Firstly they are used generate SPARQL queries that match 13

different search intentions, assuming the user has all necessary knowledge to produce an optimal result. Secondly the generated queries are used for evaluation against a sample dataset.

4.1 Dataset and Evaluation Environment

For the following evaluation of the different approaches, 10.000 exemplary metadata records were generated based on Records from DBPedia. They are based on an application profile based on the EngMeta metadata schema [45].

In the context of the evaluation and for testing the quality of the search results a set of search intentions were considered. The intentions were selected in a way to cover many different search types encountered in the context of RDM e.g.:

- Records with version number 10
- Records about computer science
- Records which are available since 03.07.2020
- Records about political left
- Records which contain a variable with the value 2 meters
- Records about object-oriented software which are published before 2015

Respective search queries are constructed for the approach, that resemble each search intention as close as possible assuming the user not to make any mistakes in this process and achieves a close to optimal result. For the ES approaches, RDF data is first mapped to JSON objects and then inserted into the search index.

4.2 Comparison of the Effectiveness

Based on the sample search queries, confusion matrices were constructed. Table 1 summarizes the results by calculating average precision, recall, and F1 score of the total confusion matrices of the respective approaches.

Table 1: Precision, recall, and F1 score (rounded to two digits)

	Regex	Bif:contains	Query Builder	Faceted Search	ES Simple	ES Advanced
Preciscion	0.6	0.77	0.96	0.88	0.66	0.96
Recall	1	0.67	1	0.9	0.98	1
F1 Score	0.75	0.72	0.98	0.89	0.79	0.98

4.3 Comparison of the Usability

For comparison, the *task execution time* estimated by means of the Keystroke-level Model (KLM), the simplest of the Goals, Operators, Methods, and Selection rules (GOMS) family [46] was considered. The KLM [47] describes various actions with associated physical

operations and specifies an execution time for them. Tasks are converted into operations and the execution time is added up. Table 2 shows average execution times across the approaches.

Table 2: Average task execution time in s based on the KLM

	Regex	Bif:contains	Query Builder	Faceted Search	ES Simple	ES Advanced
⌀	7.79	7.12	34.15	2.86	8.51	17.7

For Faceted Search only the time of the second step is calculated. For the first step, the same time is valid as for the respective search intention in the regex approach.

4.4 Comparison of Complexity

Sepecific preliminary knowledge to express search intentions is quite different: Regex requires egular expression syntax, `Bif:contains` requires Specific Bif:contains operators, Query Builder requires SPARQL Syntax (optional) and knowledge graph structure, Faceted Search requires Regular expressions for preliminary full text search (optional), ES Simple] requires ES simple search syntax and ES Advanced requires ES advanced search syntax.

Individual search queries were considered per approach and classified into categories. The most often selected category was used as an indicator for a final comparison of the complexity as shown in Table 3.

Table 3: Complexity for user to perform search requests ($\surd$ = easy, $-$ = complex)

Regex	Bif:contains	Query Builder	Faceted Search	ES Simple	ES Advanced
$\surd$	$\surd$	$-$	$\surd$	$\surd$	$\surd$

4.5 Compasrison of Efficiency

The categories of effectiveness, usability, and complexity are put in relation to each other. The ranking for the efficiency results from the ratio output/input, where output is the effectiveness and input is the combination of usability and complexity. Table 4 gives a ranking in the categories and overall efficiency.

4.6 Comparison of Response Time

The average wall clock times from sending the request to returning the result set were measured for a small dataset. The standard deviation was calculated to illustrate the inaccuracies of measurements as shown int Table 5.

Table 4: Ranking of effectiveness, usability, complexity, and efficiency whereby the approaches in a higher row are better ranked

Rank	Effectiveness	Usability	Complexity	Efficiency
1	Query Builder ES Advanced	Regex Bif:contains ES Simple	Regex Bif:contains Faceted Search ES Simple ES Advanced	ES Simple ES Advanced
2	Faceted Search	Faceted Search ES Advanced	Query Builder	Faceted Search
3	ES Simple	Query Builder		RegEx
4	Regex			Quiery Builder
5	Bif:contains			Bif:contains

Table 5: Average response time in ms (rounded to a full number) of user requests across all intentions.

	Regex	Bif:contains	Query Builder	Faceted Search	ES Simple	ES Advanced
⌀	401±26	346±24	311±20	327±24	553±30	555±28

For Faceted Search only the time of the second step is calculated. For the first step, the same time is valid as for the respective search intention in the regex approach.

4.7 Comparison of Scalability

To measure scalability of the approaches, the previously presented search queries were executed on the entire 10,000 metadata records the same way described the previous section. The results are shown in Table 6

Table 6: Response time in ms (rounded to a full number) of user requests in large search data record.

	Regex	Bif:contains	Query Builder	Faceted Search	ES Simple	ES Advanced
⌀	2604±97	357±21	335±38	1084±37	552±33	552±27

For Faceted Search only the time of the second step is calculated. For the first step, the same time is valid as for the respective search intention in the regex approach.

4.8 Comparison of Additional Effort

For the SPARQL query builder and the faceted search, no further arrangements need to be made. For the *regex* and *bif:contains* approach, only the literals need to be specified and stored. For ES, in addition to this, the mapping of metadata records to the search index is required additional processing times for different actions on the metadata are shown in Table 7.

Table 7: Times in ms for the additional effort for indexing in ES.

(Re)-Indexing	Add	Update	Delete
1646706±53376	837±143	839±129	735±121

5 Conclusion

Based on the evaluation criteria the approaches were ranked. The order is a direct result of the measured times or classifications without normalizing them or considering further gradations (i.e., to classify similar values equally) in the individual categories. To get a better overview of the results, a preference matrix as shown in Table 8, is used to look at how often one approach is better or worse than another.

Table 8: Preference matrix to compare different approaches where the numbers indicate how often the approach of the top beats the one to be compared

	Regex	Bif:contains	Query Builder	Faceted Search	ES Simple	ES Adv.
Regex	-	2	4	3	3	3
Bif:cont.	1	-	4	3	3	3
Query B.	3	3	-	3	4	4
Faceted S.	4	3	3	-	4	4
ES Simp.	2	3	3	2	-	2
ES Adv.	2	2	2	1	2	-
Win/Loss	11/15	13/14	16/17	12/17	16/11	16/10
Diff	-4	-1	-1	-5	+5	+6

The extended ES approach performs best. Next comes the simple variant of ES, and after that the SPARQL query builder together with the *bif:contains* approach. This largely confirms the hypothesis: The mapping of RDF-based metadata records into a search index for use in a search engine improves the quality of the search for the user, compared to the use of generated SPARQL queries. Essentially, the same results can be achieved with less effort for the user.

In conclusion, the main finding of this work is that the use of search engines can also be suitable for searching in RDF-based knowledge graphs if an adequate entity-object-mapping is applied. In the case at hand a mapping was derived from and enhanced.

The consecutive step is the full integration of the search functionality in the research data management platform Coscine. In addition, an appropriate user interface for displaying the found metadata records must be considered. Exemplary ES specific search interfaces also include ES features like auto-complete or search suggestions [48], which reduce the error rate and help the user to enter the most optimal search query.

Additionally, some tests or considerations for the optimal configuration of ES could still be done. Here, for example, the indexing and search analyzers, tokenizers, synonyms, and ranking functions could be considered to enhance results and rankings.

Supplementary Information

The datasets and applications used for the evaluation are available:

- "Search Engine Evaluation for Research Data in an RDF-based Knowledge Graph", DOI: https://doi.org/10.18154/RWTH-2020-09885.
- "Sample Dataset for Search Engine Evaluation for Research Data in an RDF-based Knowledge Graph",https://doi.org/10.18154/RWTH-2020-09886.
- "RDF-based Knowledge Graph Mapping for Elastic Search", https://doi.org/10.18154/RWTH-2020-09884.

A more extensive report on the background, especially the entity-object-mapping is available in the master thesis by Sarah Bensberg: "An Efficient Semantic Search Engine for Research Data in an RDF-based Knowledge Graph". 2020. DOI: https://doi.org/10.18154/RWTH-2020-09883.

Acknowledgments

The work was partially funded with resources granted by NFDI4Ing and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 432233186 – AIMS.

The authors would like to thank Prof. Stefan Decker and Prof. Matthias S. Müller for supervising the master thesis that served as a basis for this report.

Bibliography

[1] Marius Politze, Florian Claus, Bela Brenger, M. Amin Yazdi, Benedikt Heinrichs, and Annett Schwarz. How to manage it resources in research projects? towards a collaborative scientific integration environment. In Yves Epelboin, Michele Mennielli, Anna Pacholak, Pekka Kähkipuro, Gill Ferell, Carmen Diaz, Ligia M. Riberio, Johan Bergström, Thierry Koscielnieak, Elsa Cardoso, Raimund Vogl, Bas Cordewener, Noel Wilson, Carla Vilarrasa, Nicole Harris, and Outi Tasala, editors, *European Journal of Higher Education IT 2020-2*. 2020.

[2] Dominik Schmitz and Marius Politze. Forschungsdaten managen – bausteine für eine dezentrale, forschungsnahe unterstützung. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB*, 5(3):76–91, 2018. https://doi.org/10.5282/o-bib/2018H3S76-91

[3] Marius Politze, Sarah Bensberg, and Matthias Müller. Managing discipline-specific metadata within an integrated research data management system. In Joaquim Filipe, editor, *Proceedings of the 21st International Conference on Enterprise Information Systems ICEIS 2019, Heraklion, Crete - Greece, May 3 - 5, 2019*, ICEIS (Setúbal). SciTePress, 2019. https://doi.org/10.5220/0007725002530260

[4] Koenraad de Smedt, Dimitris Koureas, and Peter Wittenburg. Fair digital objects for science: From data pieces to actionable knowledge units. *Publications*, 8(2):21, 2020. https://doi.org/10.3390/publications8020021

[5] Richard Cyganiak, David Wood, and Markus Lanthaler, eds. RDF 1.1 Concepts and Ab- stract Syntax. 2014. (Visited on 02/20/2020).

[6] Shapes constraint language (shacl). URL:https://www.w3.org/TR/shacl/.

[7] Hiba Arnaout and Shady Elbassuoni. Effective searching of rdf knowledge graphs. *SSRN Electronic Journal*, 2018. https://doi.org/10.2139/ssrn.3199315

[8] Sparql 1.1 overview. URL:https://www.w3.org/TR/sparql11-overview/.

[9] En iso 9000:2015 d/e quality management systems— fundamentals and vocabulary.

[10] Iso 9241-11:2018(en) ergonomics of human-system interaction — part 11: Usability: Definitions and concepts.

[11] Jakob Nielsen. *Usability engineering*. Academic Press, Boston, 1993.

[12] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs. Swoogle. In David Grossman, Luis Gravano, ChengXiang Zhai, Otthein Herzog, and David A. Evans, editors, *Proceedings of the Thirteenth ACM conference on Information and knowledge management - CIKM '04*, New York, New York, USA, 2004. ACM Press. https://doi.org/10.1145/1031171.1031289

[13] Sindice.com: a document-oriented lookup index for open linked data. *International Journal of Metadata Semantics and Ontologies*, 3(1), 2008. https://doi.org/10.1504/IJMSO.2008.021204

[14] Andreas Harth, Aidan Hogan, Jurgen Umbrich, and Stefan Decker. *SWSE: Objects before documents*. 2012.

[15] Gong Cheng, Weiyi Ge, and Yuzhong Qu. Falcons: Searching and browsing entities on the semantic web. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, New York, NY, USA, 2008. Association for Computing Machinery. https://doi.org/10.1145/1367497.1367676

[16] Saeedeh Shekarpour, Soren Auer, A.-C. Ngonga Ngomo, Daniel Gerber, and Claus Stadler. Keyword-driven sparql query generation leveraging background knowledge. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2011. https://doi.org/10.1109/WI-IAT.2011.70

[17] Emil Kuric, Javier D. Fernández, and Olha Drozd. Knowledge graph exploration: A usability evaluation of query builders for laypeople. In Maribel Acosta, Philippe Cudré-Mauroux, and Maria Maleshkova, editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, Information Systems and Applications, incl. Internet/Web, and HCI, Cham, 2019. Springer International Publishing. https://doi.org/10.1007/978-3-030-33220-4_24

[18] Judie Attard, Fabrizio Orlandi, and Sören Auer. Exconquer: Lowering barriers to rdf and linked data re-use. *Semantic Web*, 9(2), 2018. https://doi.org/10.3233/SW-170260

[19] Patrick Hoefler, Michael Granitzer, Eduardo Veas, and Christin Seifert. Linked data query wizard: A novel interface for accessing sparql endpoints. *7th Workshop on Linked Data on the Web 2014*, 2014.

[20] Vizquery - hay's tools. URL: https://hay.toolforge.org/vizquery/.

[21] Wikidata query service. URL: https://query.wikidata.org/.

[22] Openlink isparql. URL: http://dbpedia.org/isparql/.

[23] Paul R Smart, Alistair Russell, Dave Braines, Yannis Kalfoglou, and Nigel R. Shadbolt. A visual approach to semantic query design using a web-based graphical query designer. 2008. https://doi.org/10.1007/978-3-540-87696-0_25

[24] Ahmet Soylu, Evgeny Kharlamov, Dmitriy Zheleznyakov, Ernesto Jimenez-Ruiz, Martin Giese, Martin G. Skjæveland, Dag Hovland, Rudolf Schlatte, Sebastian Brandt, Hallstein Lie, and Ian Horrocks. Optiquevqs: A visual query system over ontologies for industry. *Semantic Web*, 9(5), 2018. https://doi.org/10.3233/SW-180293

[25] Florian Haag, Steffen Lohmann, Stephan Siek, and Thomas Ertl. Queryvowl: Visual composition of sparql queries. In Fabien Gandon, Christophe Guéret, Serena Villata, John Breslin, Catherine Faron-Zucker, and Antoine Zimmermann, editors, *The semantic web: ESWC 2015 satellite events*, Lecture notes in computer science Computer communication networks and telecommunications, Cham and Heidelberg and New York, 2015. Springer. https://doi.org/10.1007/978-3-319-25639-9_12

[26] Esther Kaufmann, Abraham Bernstein, and Lorenz Fischer. Nlp-reduce: A naive but domainindependent natural language interface for querying ontologies. In *4th European Semantic Web Conference ESWC*, 2007.

[27] Sébastien Ferré. Sparklis: An expressive query builder for sparql endpoints with guidance in natural language. *Semantic Web*, 8(3), 2016. https://doi.org/10.3233/SW-150208

[28] Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, Maya Ramanath, Volker Tresp, and Gerhard Weikum. *Natural Language Questions for the Web of Data*. EMNLP-CoNLL '12. Association for Computational Linguistics, USA, 2012.

[29] Mohnish Dubey, Sourish Dasgupta, Ankit Sharma, Konrad Höffner, and Jens Lehmann. Asknow: A framework for natural language query formalization in sparql. In Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Simone Paolo Ponzetto, Christoph Lange, and Chiara Ghidini, editors, *The semantic web*, Lecture notes in computer science Theoretical computer science and general issues, Cham and Heidelberg, 2016. Springer. https://doi.org/10.1007/978-3-319-34129-3_19

[30] Camille Pradel, O. Haemmerlé, and N. Hernandez. Demo: Swip, a semantic web interface using patterns. 2013.

[31] Majid Latifi, Horacio rodriguez, and Miquel Sànchez-Marrè. Scoqas: A semantic-based closed and open domain question answering system. *Procesamiento de Lenguaje Natural*, 59, 2017.

[32] Hearst Marti A. Uis for faceted navigation: Recent advances and remaining open problems. 2008.

[33] Ivan Heibi, Silvio Peroni, and David Shotton. Enabling text search on sparql endpoints through oscar. *Data Science*, 2(1-2), 2019. https://doi.org/10.3233/DS-190016

[34] Marcelo Arenas, Bernardo Cuenca Grau, Evgeny Kharlamov, Šarūnas Marciuška, and Dmitriy Zheleznyakov. Faceted search over rdf-based knowledge graphs. *Journal of Web Semantics*, 37-38, 2016. https://doi.org/10.1016/j.websem.2015.12.002

[35] Stéphane Campinas. *Graph summarisation of web data: data-driven generation of structured representations*. PhD thesis, NUI Galway, 2016.

[36] Yuangui Lei, Victoria Uren, and Enrico Motta. Semsearch: A search engine for the semantic web. In Steffen Staab and Vojtěch Svátek, editors, *Managing knowledge in a world of networks*, volume 4248 of *Lecture notes in computer science Lecture notes in artificial intelligence*. Springer, Berlin, 2006. https://doi.org/10.1007/11891451_22

[37] Saeedeh Shekarpour, Edgard Marx, Axel-Cyrille Ngonga Ngomo, and Sören Auer. *SINA: Semantic Interpretation of User Queries for Question Answering on Interlinked Data*, volume 30. 2015. Semantic Search. https://doi.org/10.1016/j.websem.2014.06.002

[38] Jens Lehmann and Lorenz Bühmann. Autosparql: Let users query your knowledge base. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter de Leenheer, and Jeff Pan, editors, *The semantic web: research and applications*, Lecture Notes in Computer Science, Berlin, 2011. Springer. https://doi.org/10.1007/978-3-642-21034-1_5

[39] Markus Mandalka. Open semantic search: Your own search engine for documents, images, tables, files, intranet & news. URL: https://www.opensemanticsearch.org//.

[40] Lei Zhang, Qiaoling Liu, Jie Zhang, Haofen Wang, Yue Pan, and Yong Yu. Semplore: An ir approach to scalable hybrid query of semantic web data. In Karl Aberer, editor, *The semantic web*, Lecture Notes in Computer Science, Berlin, 2007. Springer. https://doi.org/10.1007/978-3-540-76298-0_47

[41] Cristiano Rocha, Daniel Schwabe, and Marcus Poggi Aragao. A hybrid approach for searching in the semantic web. In Stuart Feldman, Mike Uretsky, Marc Najork, and

Craig Wills, editors, *13th International Conference on World Wide Web Conference*, New York, N.Y., 2005. ACM Press. https://doi.org/10.1145/988672.988723

[42] 16.3.1.using full text search in sparql. URL: http://docs.openlinksw.com/virtuoso/rdfsparqlrulefulltext/.

[43] Pavel Grafkin, Mikhail Mironov, Michael Fellmann, Birger Lantow, Kurt Sandkuhl, and Alexander V. Smirnov. SPARQL query builders: Overview and comparison. In Björn Johansson and Filip Vencovský, editors, *Joint Proceedings of the BIR 2016 Workshops and Doctoral Consortium co-located with 15th International Conference on Perspectives in Business Informatics Research (BIR 2016), Prague, Czech Republic, September 14 - 16, 2016*, volume 1684 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[44] Adam Styperek, Michal Ciesielczyk, Andrzej Szwabe, and Pawel Misiorek. Evaluation of sparql-compliant semantic search user interfaces. *Vietnam Journal of Computer Science*, 2(3), 2015. https://doi.org/10.1007/s40595-015-0044-y

[45] Engmeta - beschreibung von forschungsdaten | informations- und kommunikationszentrum | universität stuttgart. URL: https://www.izus.uni-stuttgart.de/fokus/engmeta/.

[46] Shiroq Al-Megren, Joharah Khabti, and Hend S. Al-Khalifa. A systematic review of modifications and validation methods for the extension of the keystroke-level model. *Advances in Human-Computer Interaction*, 2018, 2018. URL: https://www.hindawi.com/journals/ahci/2018/7528278/, https://doi.org/10.1155/2018/7528278

[47] David Kieras. *Using the Keystroke-Level Model to Estimate Execution Times*. 2003.

[48] Suggesters | elasticsearch reference [7.9] | elastic. URL: https://www.elastic.co/guide/en/elasticsearch/reference/current/search-suggesters.html.

bwHPC-S5: Scientific Simulation and Storage Support Services

Robert Barthel[1] und Jürgen Salk[2]

[1]Steinbuch Centre for Computing, Karlsruher Institut für Technologie
[2]Kommunikations- und Informationszentrum, Universität Ulm

Das Projekt bwHPC-S5 ist das aktuelle Begleitprojekt zum Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das Hochleistungsrechnen (HPC), Data Intensive Computing (DIC) und Large Scale Scientific Data Management (LS^2DM) und dient als Bindeglied zwischen Wissenschaft und den Infrastrukturen für HPC, DIC und LS^2DM. Es beinhaltet eine landesweit aufgestellte Nutzerbetreuung und unterstützt Infrastrukturbetreiber und deren Nutzenden mit IT-Services. Phase 1 des Projekts läuft von Juli 2018 bis März 2021 und wird vom Ministerium für Forschung, Wissenschaft und Kunst Baden-Württemberg (MWK) finanziert. Ab April 2021 wird das Projekt mit der bereits bewilligten Phase 2 fortgesetzt.

1 Einleitung

Seit vielen Jahren sind die Landeskonzepte für das Hochleistungsrechnen und datenintensive Dienste in Baden-Württemberg bekannt und etabliert. Das Konzept baut auf den Grundsätzen der kooperativen Bereitstellung von Ressourcen und Servicestrukturen auf. Besonders der Aspekt der Kooperation schafft Synergien und bündelt Expertisen zur Verbreiterung der Nutzerbasis, bedarfsgerechten Bereitstellung von Diensten und zielgerichteten Unterstützung von Forschenden und Studierenden bei der Nutzung dieser Dienste.

bwGRiD (2008) als Startpunkt einer standortübergreifenden HPC-Infrastruktur und bwGRiD ergänzende Maßnahmen (2011) als Beginn eines föderativen Nutzerunterstützungsprojekts schufen eine bis heute andauernde enge Zusammenarbeit in Baden-Württemberg, eine Profilbildung bei der Softwarebereitstellung und für Nutzende eine standortunabhängige und flexible Ressourcennutzung auf Basis einer vereinheitlichten Arbeitsumgebung mit homogener Rechnerinfrastruktur. Im Kontrast dazu standen die Zugangshürden durch Grid-Zertifikate, die dezentrale Wissenschaftsunterstützung und lokale clusternahe Datenhaltungskonzepte.

Die nutzerseitigen Anforderungen nach Vereinfachung des Zugangs und wissenschaftsangepasster Infrastruktur wurden HPC-seitig mit dem Konzept bwHPC [1, 2] von 2013 bis 2018 umfassend adressiert. Die nachfolgend genannten Kernpunkte des bwHPC–Konzeptes stellten bundesweit ein Alleinstellungsmerkmal dar und dienten dem Wissenschaftsrat als

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13731

Vorbild bei dessen Empfehlungen zur Finanzierung des nationalen Hoch-und-Höchstleistungsrechnens in Deutschland [3]:

- Niederschwelliger Zugang zu den Ressourcen auf Basis der landesweit föderierten Identitätsinfrastruktur bwIDM,
- Ausdifferenzierung der HPC-Infrastruktur in ein landesweit nutzbares Rechensystem bwUniCluster („Baden-Württemberg-Universalcluster“) zur Abdeckung der Grundversorgung an Rechenleistung und in vier verschiedene, landesweit nutzbare bwForCluster („Baden-Württemberg-Forschungscluster“) für verschiedene Wissenschaftsgemeinden.
- Verzahnung der Angebote durch „vertikale“ Durchlässigkeit der Tier 3 hin zur Tier 2 bis 0
- „Horizontale“ Verzahnung durch die Anknüpfung an fachliche Kompetenzzentren zwischen Rechnerbetrieb und Anwendungsforschung

Die konkrete Umsetzung des Konzeptes wurde durch die Begleitprojekte bwHPC-C5 Phasen 1 und 2 (2013 – 2018) kanalisiert [4, 5]; mit den erfolgreichen Ergebnissen:

- Aufbau einer landesweiten Supportstrukturen zur Nutzerunterstützung der bwForCluster und des bwUniCluster sowie sowie die Integration der Hochschulen für angewandte Wissenschaften
- Begleitung der Vorbereitungsarbeiten zur Inbetriebnahme der Clustersysteme (Technologieevaluation, bwIDM) und Aufbau einer bedarfsgerechten Softwareinfrastruktur
- Etablierung zentraler Dienste, u.a zur Aussteuerung der Ressourcen (ZAS) gemäß dem bwHPC-Konzept, landesweites Ticketsystem und Dokumentenplattform (bwHPC-Wiki)
- Etablierung eines landesweiten Schulungsprogramms und zentral koordinierten Öffentlichkeitsarbeit

Parallel zur HPC-Infrastruktur und den bwHPC-C5-Projekten wurde als Teil des bwDATA-Konzeptes [6, 7] eine umfangreiche Speicherinfrastruktur mit der Large Scale Data Facility (LSDF), der bwCloud und dem bwDataArchive aufgebaut. Die Landesprojekte bwFDM-Info I und II etablierten Plattformen zur Bedarfserfassung, Weiterbildung und zum Dialog im Bereich des Forschungsdatenmanagements. Dazu reihen sich Spezialsysteme zur Visualisierung, bwVisu, und zur Datenanalyse (z.B. das Smart Data Innovation Lab).

Aus Sicht der Nutzenden ist das Management und die Verarbeitung von Forschungsdaten in getrennten Infrastrukturen wenig praktikabel. Entsprechenden Empfehlungen folgend (z.B. [8]) werden mit dem seit 2018 beschlossenen Umsetzungskonzept für HPC, DIC und LS^{2}DM des Landes Baden-Württemberg [9] (*kurz: „bwHPC 2.0“*) die bisher eigenständigen, erfolgreichen bwHPC- und bwDATA-Konzepte durch nachfolgende Maßnahmen in eine gemeinsame Sicht überführt:

- Erweiterung der Aufgabenbereiche der Kompetenzzentren auf Datenhaltung und Einrichtung weiterer Kompetenzzentren

- Bedarfsgerechter Ausbau und Erneuerung der HPC-Systeme aller drei Ebenen und der Datenmanagementsysteme

- Entwicklung einer landesweiten Datenföderation

- Fortführung des Begleit- und Nutzerunterstützungsprojektes mit der Integration dienenden Schwerpunkten

Als Synonym für diese Integration stehen die fünf „S“ – *Scientific Simulation and Storage Support Services* – des Begleitprojekts bwHPC-S5.

Zur Umsetzung der gemeinsamen Sicht auf HPC, DIC und LS^{2}DM werden im bwHPC-S5-Projekt alle erbrachten Dienste für die wissenschaftlichen Nutzer so organisiert, dass die begleitende Supportleistung, unabhängig ob es dabei um das Thema HPC, Daten oder einer Kombination aus beiden geht, über die gleichen Schnittstellen angeboten wird. Das Konzept „one face to the customer“ gilt dabei nicht nur für die Schulungen, Best-Practices-Dokumentationen, Dienstschnittstellen oder die allgemeine Nutzerunterstützung sondern insbesondere auch für die gemeinsam zwischen Betreibern und Nutzern durchgeführten Tigerteam-Unterstützungsprojekte.

Das Projekt bwHPC-S5 ist aufgrund der langen Laufzeit des neuen Umsetzungskonzepts in mehrere Phasen unterteilt, um auf geänderte Bedarfslagen und Anforderungen flexibel reagieren zu können. Die Unterteilung in Phasen und deren Evaluierung ermöglicht u.a. die gezielte Nachsteuerung der Umsetzungsmaßnahmen. Der 1. Phase von Juli 2018 bis März 2021 waren sechs Ziele auferlegt:

1. Ausbau der föderativen Wissenschaftsunterstützung

2. Fortschreibung der fachlichen Ausprägung im Bereich HPC

3. Umsetzung einer landesweiten Datenföderation

4. bedarfsgetriebene gemeinsame Technologieevaluierung

5. weitere Professionalisierung der Öffentlichkeitsarbeit sowie

6. Fortschreibung und Weiterentwicklung der gemeinsamen Software-Versorgung der HPC-Systeme

Die 2. Phase von April 2021 bis Juni 2023 zielt insbesondere auf die Fortführung der erfolgreich begonnenen Anstrengungen zur einheitlichen Sicht auf die Dienste zur Festigung des Erstanlaufpunkts zu HPC, DIC und LS^{2}DM für Einsteigende und Bestandsnutzende, aber auch auf die konzeptionelle und technische Ausgestaltung der BaWü-Datenförderation.

2 Projektpartner und Governance

Das Konsortium des Projekts bwHPC-S5 Phase 1 besteht aus den Universitäten Freiburg, Heidelberg, Hohenheim, Konstanz, Mannheim, Stuttgart, Tübingen und Ulm, dem Karlsruher Institut für Technologie sowie der Hochschule Esslingen und der Hochschule für Technik Stuttgart.

Um die Realisierung und Überwachung der Projektziele möglichst effektiv zu gestalten, wird eine bewährte hierarchische Struktur für die Koordination umgesetzt, wie sie für kollaborative Forschungsprojekte dieser Größe und Komplexität üblich ist [10].

Strategische Entscheidungen über die Fortentwicklung der Tier-3 Landes-Cluster und der landesweiten Dateninfrastruktur werden vom Arbeitskreis der Leiterinnen und Leiter der Wissenschaftlichen Rechenzentren und Informationszentren des Landes Baden-Württemberg
(ALWR-BW), eingerichtet durch die Landesrektorenkonferenz (LRK), getroffen. Gegenüber dem ALWR-BW wird alle drei Monate zum Status und zu den Risiken des Projektverlaufs sowie zur Auslastung der bwHPC-Infrastruktur berichtet. Anforderungen und Vorgaben anderer Steuergremien werden über den ALWR-BW und die Projektverantwortlichen koordiniert.

Für die wissenschaftliche Begleitung des Betriebs und der Weiterentwicklung definierter digitaler Forschungsinfrastrukturen (u.a. Vorhaben der HPC/DIC-Landesstrategie) ist der Landesnutzerausschuss Baden-Württemberg (LNA-BW) zuständig. Der LNA-BW stellt die systematische Erhebung der wissenschaftlichen Nutzungsanforderungen sicher, bewertet die Wirksamkeit der Mechanismen zur Aussteuerung der Nutzung und Auslastung der Infrastrukturen, schlägt Verbesserungen der Aussteuerungs- bzw. Auslastungsmechanismen vor und verfasst Empfehlungen und Stellungnahmen zur Gewährleistung der optimalen wissenschaftlichen Ausrichtung der Vorhaben und Rahmenkonzepte der HPC/DIC-Landesstrategie. Gegenüber dem LNA-BW wird zu jeder Sitzung zum Status des Projekts, der bwHPC-Infrastruktur und der Umsetzung der mit dem ALWR-BW abgestimmten Maßnahmen berichtet.

Der Steuerkreis digitale Forschungsinfrastruktur Baden-Württemberg (SK DigiForInfraBW) repräsentiert die übergreifende Instanz der Betreiber und Nutzenden der digitalen Forschungsinfrastrukturen innerhalb der HPC/DIC-Landesstrategie[1] sowie des bwDATA-Rahmenkonzeptes. Gegenüber dem SK DigiForInfraBW berichtet der ALWR-BW einmal in Jahr u.a. zum Fortschritt des bwHPC–S5 Projektes.

Das Projekt bwHPC-S5 wird aufgrund seiner Komplexität und Bedeutung für das Umsetzungskonzept HPC, DIC und LS^{2}DM von zwei Vertretern des ALWR-BW begleitet und verantwortet. Für die operative Leitung des Projekts ist dagegen das Projektbüro (PMO) verantwortlich. Fortschritt und Risiken auf Projektebene bzw. beim Zusammenwirken der Arbeitspakete werden im Kernteam thematisiert. Das Kernteam setzt sich aus

[1] Ausgenommen HPC-Systeme der Leistungsklasse 1 (Universität Stuttgart) und 2 (Karlsruher Institut für Technologie)

allen Leitenden der Arbeitspakete, dem Projektbüro und einer Vertretung pro Projektpartner ohne AP-Leitung zusammen. Aufgaben der Arbeitspaketleitenden umfassen die technische Leitung, Planung und Überwachung ihrer Arbeitspakete aber auch das Berichtswesen an die übergeordneten Gremien. Aus den bwHPC-Kompetenzzentren bilden zudem je ein stimmberechtigter Vertreter ein Team für die Zuweisung der Rechenvorhaben an eine Clusterressource: das sogenannte Clusterauswahlteam (CAT).

Für die Abstimmung in betrieblichen Belangen der HPC- und Datenföderation, u.a. bezüglich Entwicklung und Fortschreibung der Betriebsmodelle sowie der betrieblichen Aspekte in der Produktion, wurde ein Technical Advisory Board (TAB) etabliert, in welches jede Universität ein Mitglied entsendet. Die Hochschulen entsenden ebenfalls ein Mitglied in das TAB.

Die oben genannten Gremien treffen sich regelmäßig, um innerhalb des Projektes die notwendige Abstimmungen zu vollziehen. Neben Präsenztreffen stimmen sich Kernteam bzw. Technical Advisory Board in mindestens zweiwöchentlich stattfindenden Video- und Telefonkonferenzen ab.

3 Projektstruktur der 2. Phase

Aus der beschriebenen Ausgangslage ergeben sich für die weitere Begleitung des Umsetzungs-konzepts – insbesondere für die Fortschreibung der Verzahnung der landesweiten Rechen- und Dateninfrastruktur und für die kontinuierliche Fortentwicklung zugehöriger Dienste sowie effizienter und effektiver Nutzerunterstützung – die Projektziele **P.1** bis **P.4**:

- P.1 - Weiterentwicklung der föderativen Wissenschaftsunterstützung
- P.2 - Weiterentwicklung der landesweiten HPC-Infrastruktur
- P.3 - Weiterentwicklung der landesweiten Datenföderation
- P.4 - Erschließung neuer Technologien

Diese übergeordneten Projektziele bestimmen gleichzeitig die Zielstellungen und Aufgaben der nachfolgend beschriebenen Arbeitspakete und werden bei ihrer Fortschrittsbewertung durch definierte Kennzahlen ergänzt.

Das Projekt bwHPC-S5 Phase 2 verortet dabei die Arbeitspakete (AP), je nach deren Schwerpunktaufgaben und Zielsetzungen, in jeweils einem der drei nachfolgenden Aktivitätsebenen (AE):

1. Nutzerbezogene Aktivitäten und Öffentlichkeitsarbeit
2. Föderativer Betrieb und systembezogene Aktivitäten
3. Innovations- und Evaluationsaktivitäten

Anhand der thematischen Strukturierung ergeben sich dabei drei APs für die erste Aktivitätsebene; vier APs für die zweite und ein AP für die dritte. Zusammen mit dem

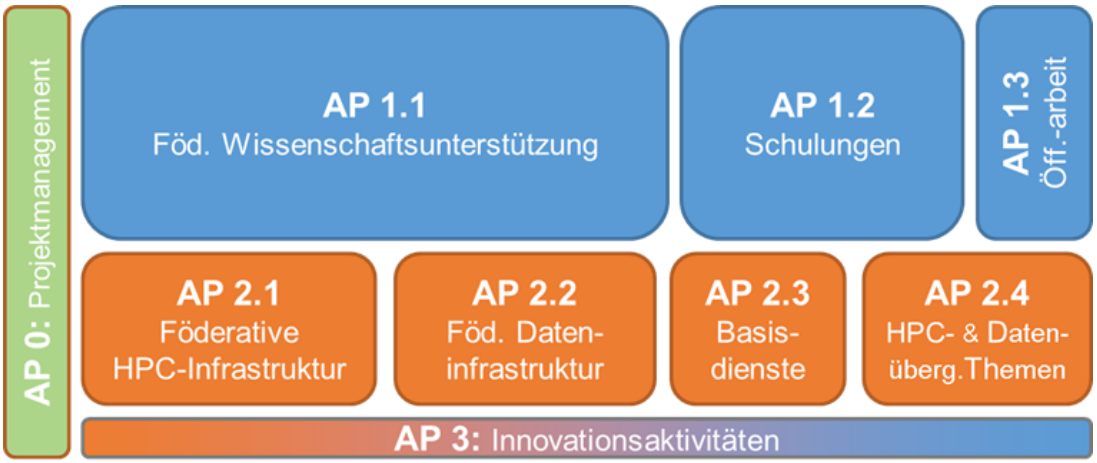


Abbildung 1: Arbeitspaketstruktur von bwHPC-S5 Phase 2.

Projektmanagement strukturiert sich das Projekt bwHPC-S5 Phase 2 in neun Arbeitspakete, deren Personalmittelaufwand in Abbildung 1 flächenmäßig ins Verhältnis gesetzt wurden. Der Schwerpunkt des Projekts liegt dabei auf der ersten Aktivitätsebene mit rund 50% der personellen Ressourcen.

Aktivitätsebene 1

Aktivitätsebene 1 mit den Arbeitspaketen 1.1, 1.2 und 1.3 stellt die direkte Schnittstelle zu Nutzern und anderen (auch externen) Interessengruppen sowie der Öffentlichkeit dar. Jedes einzelne dieser drei Arbeitspakete beinhaltet sowohl HPC-bezogene Anteile als auch Themenbereiche mit Bezug zum datenintensiven Rechnen (Data Intensive Computing, DIC) und zum Umgang mit umfangreichen wissenschaftlichen Datenmengen (Large Scale Scientific Data Management, LS^{2}DM).

Im AP1.1 „Föderative Wissenschaftsunterstützung“ steht die zielgerichtete und fachlich fundierte Unterstützung von Nutzern der landesweiten bwHPC- und Speichersysteme durch fachspezifisch ausgerichtete bwHPC-Kompetenzzentren im Vordergrund. Unter einem bwHPC-Kompetenzzentrum ist eine Organisationsstruktur zu verstehen, in der Fachkompetenzen zur Anwenderunterstützung in verschiedenen Wissenschaftsbereichen gebündelt werden. Die personelle Zusammensetzung der Kompetenzzentren erfolgt standortübergreifend zur optimalen Ausnutzung des landesweit vorhandenen Expertenwissens. Mit dem landesweiten Ticketsystem werden alle üblichen Anfragen zentral erfasst und unabhängig vom Standort des Nutzers oder des Dienstes vom fachlich am besten geeigneten Expertenteam bearbeitet. Natürlich wird dies bei betriebsnahen Themen oft vom jeweiligen Betreiberstandort erbracht, aber bei spezifischen und fachlich tiefergehenden Anfragen erfolgt dies standortübergreifend. Komplexe und tiefergehende Fragestellungen, die nicht mit den üblichen Unterstützungsstrukturen schnell und effizient erbracht werden können, werden durch Bildung standortübergreifender Tigerteams adressiert. Ein Tigerteam ist Teil eines Kompetenzzentrums und wird zeitlich befristet, meist in enger

Kooperation mit einer wissenschaftlichen Arbeitsgruppe, zur Umsetzung von konkreten Unterstützungs- und Optimierungsmaßnahmen aufgestellt. Die personelle Zusammensetzung eines Tigerteams kann standortübergreifend erfolgen, um das an unterschiedlichen Standorten vorhandene Expertenwissen optimal zu nutzen.

Ergänzend zu AP1.1 werden in AP1.2 „Schulungen“ durch bedarfsorientierte und landesweit abgestimmte Schulungen die Fachkompetenzen an eine breite Nutzerschaft vermittelt. Dazu wird in Grundlagen- und Aufbaukursen sowohl Basiswissen zur effizienten Nutzung der bwHPC-Systeme für rechen- und datenintensive Aufgaben adressiert als auch spezielle Anwendungsbereiche zu HPC und Datenmanagement (z.B. zur parallelen Programmierung oder zum Einsatz spezieller Bibliotheken und Werkzeuge zur Performancesteigerung) sowie komplexer Softwaresysteme. Neben Präsenzkursen sind auch digitale Lehrformen, wie E-Learning und Web-Seminaren, Teil des Projekt-Portfolios, um möglichst viele Nutzer erreichen zu können. Wesentliche Kernaufgaben in bwHPC-S5 Phase 2 bestehen im weiteren Ausbau des Schulungsangebotes zu Speichersystemen unter Berücksichtigung des Themas FDM sowie in der Erweiterung der Teilnehmerkreise, hier insbesondere um Angehörige der HAWen und Nutzer der landesweiten Speicherdienste.

Das AP1.3 „Öffentlichkeitsarbeit“ unterstützt und koordiniert alle Aktivitäten zur Kommunikation der Ziele und Erfolge des Projektes bwHPC-S5 im Kontext des landesweiten Umsetzungskonzeptes bwHPC, datenintensiven Rechnen (DIC) und Large Scale Scientific Data Management (LS^2DM) gegenüber verschiedenen Zielgruppen. Dazu gehören neben bestehenden Nutzern auch potentiellen Nutzer, die bisher auf lokalen Systemen arbeiten und über die besonderen Möglichkeiten der zentralen Rechner- und Speicherinfrastruktur informiert werden sollen, sowie Entscheidungsträger an den Universitäten, die HPC- und Daten-Community und die interessierte Öffentlichkeit. Ziele sind hier die Sichtbarkeit von bwHPC weiter zu erhöhen, das Diensteangebot zu vermitteln und insgesamt die Anzahl der Nutzer in Baden-Württemberg weiter zu steigern.

Aktivitätsebene 2

Aktivitätsebene 2 fasst alle Aktivitäten zusammen, welche den landesweiten, föderierten und koordinierten Betrieb der Rechen- und Dateninfrastruktur sicherstellen. Mit den Arbeitspaketen 2.1 bis 2.4 werden dazu interne Dienste und Dienstleistungen zur Verfügung gestellt, die nicht unmittelbar nach außen sichtbar sind, sondern auf technischer und administrativer Ebene auf die produktive Umsetzung des vorliegenden Landeskonzeptes ausgerichtet sind. Weiterhin stellen diese Arbeitspakete ein internes Unterstützungsangebot für die darüber liegende nutzerbezogene Schicht, insbesondere für AP1.1, dar.

Die technisch fokussierten Aktivitäten in den Bereichen HPC (AP2.1 „Föderative HPC-Infrastruktur“) und Daten (AP2.2 „Föderative Dateninfrastruktur“) sind im Bereich der Föderation noch in unterschiedlichen Entwicklungsstadien. Während im HPC-Bereich die Föderierung der bestehenden Systeme bereits sehr erfolgreich umgesetzt wurde und daher mehr eine Verstetigung mit Integration neuer Rechnersysteme, Optimierung von Betriebskonzepten und Vereinfachung der Nutzung im Vordergrund stehen, zielt das AP2.2

basierend auf den Ergebnissen von bwHPC-S5 Phase 1 auf den produktiven Aufbau eines Verbundes von heterogenen und verteilten Speicher und Datenmanagementsystemen zur BaWü-Datenföderation, inklusive der Bereitstellung von geeigneten Plattformen, Schnittstellen und Werkzeugen zum Datentransfer, sowie der Einbindung neuer Speichersysteme an unterschiedlichen Standorten in Baden-Württemberg. Dazu wird im Bereich Daten noch mehr Entwicklungs- und Konfigurationsaufwand notwendig sein. Die bisher sehr erfolgreich etablierten Basisdienste werden im AP2.3 „Basisdienste“ fortgeführt und weiterentwickelt werden. Dazu gehört insbesondere die Umsetzung von notwendigen Erweiterungen von Schnittstellen zu Rechen- und Speicherdiensten mit Integration neuer Konzepte zur Nutzerauthentifizierung. Neben der Vereinheitlichung der Basisdienste gibt es eine Reihe von Technologiebereichen, die nicht klar HPC- oder Datenthemen zugeordnet werden können und daher in AP2.4 „HPC- und datenübergreifende Themen“ eigenständig und übergreifend umgesetzt und angeboten werden.

Aktivitätsebene 3

Die Aufgaben der Aktivitätsebenen 1 und 2 stellen aufgrund der notwendigen Ausrichtung über Standortgrenzen hinweg bei allen Aufbau- und Integrationsaufgaben in den Bereichen HPC, DIC und LS^{2}DM eine hohe Anforderung an darauf abgestimmte neue und innovative Lösungen, müssen aber gleichzeitig die Erwartungshaltung der Anwender erfüllen, die einen verlässlichen und möglichst störungsfreien Betrieb fordern. Die Dynamik im Bereich der verfügbaren Technologien und zugehörigen Software- und Betriebsmodellen erfordert jedoch für zukünftige Systeme ggf. noch nicht für den Betrieb geeignete Technologien bereits im Vorfeld zu untersuchen. In AP3 „Innovationsaktivitäten“ werden daher Technologietrends beobachtet und im Rahmen von Technologie-Sprints hinsichtlich ihres Potentials zur Verbesserung bestehender Dienste und des Aufbaus neuer Dienste evaluiert. Im Rahmen der Arbeiten zur Datenföderation in bwHPC-S5 Phase 1 hat sich insbesondere herausgestellt, dass es auch für die in Entstehen befindliche BaWü-Datenföderation verstärkter Evaluationsaktivitäten bedarf. In Phase 2 werden daher in AP3 die vormals stark HPC-bezogenen Technologie-Sprints thematisch um Innovationsaktivitäten im Bereich des datenzentrischen Rechnens und des föderativen Forschungsdatenmanagements erweitert. Dieses Arbeitspaket fungiert somit als Vorstufe für die Überführung von innovativen Konzepten in den produktiven Betrieb und stellt damit auch eine Schnittstelle zwischen der auf den Produktionsbetrieb ausgelegten zweiten Aktivitätsebene und eher prototypischen Neuentwicklungen dar.

Querschnittsebene

Die koordinierende und leitende Rolle bei der thematischen und strukturellen Zusammenführung von HPC, DIC und LS^{2}DM im Projekt und die damit notwendige Abstimmung zwischen diesem Projekt, anderen Landesprojekten und den Steuergremien erfolgt über AP0 „Projektmanagement“.

4 Fazit

Die Etablierung des Projekts bwHPC-S5 stellt einen der notwendigen Schritte zur Verknüpfung von HPC- und Forschungsdateninfrastrukturen im Land Baden-Württemberg dar. Die einzelnen Arbeitspakete tragen dabei wechselseitig zur Erreichung der Projektziele bei, zur deren Fortschrittsbewertung eine Reihe komplementärer Erfolgsindikatoren herangezogen werden, die einer regelmäßigen Bewertungsanalyse unterliegen. Basis für die Bewertungsanalyse sind definierte Erwartungswerte bzw. kritische Handlungsschwellen.

So belegen die Kennziffern zur Systemnutzung und Nutzeraktivität, dass durch das Zusammenwirken aller Arbeitspakete eine sehr hohe Auslastung der vorhandenen Ressourcen durch einen breiten wissenschaftlichen Nutzerkreis erzielt werden konnte. Insbesondere sind die Nutzungsanteile der verschiedenen Universitäten auf eine Verteilung eingeschwungen, wie sie für landesweit verfügbare Rechnersysteme erwartet und erwünscht ist. Der fortwährende Trend zur Zunahme der Nutzeranzahl kann dabei als Indikator für die hohe Akzeptanz der bwHPC-Systeme in Tier 3 bei den Anwendern gewertet werden.

Auch die während der bisherigen Betriebszeiten erzielten mehr als 1500 Veröffentlichungen in internationalen wissenschaftlichen Fachzeitschriften, Büchern und Konferenzbänden, zu denen die Nutzung der Systeme im bwHPC-Verbund nachweislich beigetragen haben, belegen sehr eindrucksvoll die herausragende Bedeutung der fünf bwHPC-Clustersysteme und der kontinuierlichen Unterstützungsmaßnahmen des Begleitprojektes bwHPC-S5 für den Wissenschaftsstandort Baden-Württemberg.

Bestätigt wird das auch im Rahmen von regelmäßig stattfindenden landesweiten Nutzerbefragungen, in denen die Qualität der Supportleistungen des Projektes sowie der angebotenen Schulungen durchgängig als sehr hoch bewertet wird. Insbesondere der Erfolg der stattgefundenen Tigerteam-Maßnahmen wird dabei von den betreffenden Nutzergruppen als sehr gut eingestuft.

Auch der hohe Nutzungsgrad der bereitgestellten Softwaremodule, einschließlich zugehöriger Online-Hilfen, technischer Dokumentationen und Beispielskripten für die Batchsysteme der Cluster, belegt die hohe Relevanz des landesweit abgestimmten und fachlich ausdifferenzierten Softwareportfolios auf den Clustersystemen für die wissenschaftliche Nutzerschaft der Systeme.

Mit bwHPC-S5 Phase 1 wurden bereits entscheidende Schritte zum Aufbau der BaWü-Dat-en-föderation und der Erweiterung des Unterstützungsangebots auf datenbezogene Themen vollzogen und bewertet. Unter Beibehaltung der etablierten Projektstrukturen sowie der Dienst- und Unterstützungsangebote wird in Phase 2 insbesondere die Verzahnung mit den Forschungsdateninfrastrukturen und deren Verknüpfung mit den existierenden HPC-Infrastrukturen noch stärker intensiviert und konkretisiert werden, um die Speicher- und Datenmanagementsysteme als produktiv nutzbare Dienste mit entsprechenden Unterstützungsangeboten für die Wissenschaftler zu realisieren.

Danksagungen

Die Autoren danken dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg für die Finanzierung des Projekts bwHPC-S5 sowie allen nicht namentlich erwähnten Projektmitarbeitenden für ihren unermüdlichen Einsatz und für ihre Unterstützung beim Erstellen des Projektsantrags bwHPC-S5 Phase 2.

Literaturverzeichnis

[1] H. Hartenstein, T. Walter, and P. Castellaz. Aktuelle Umsetzungskonzepte der Universitäten des Landes Baden-Württemberg für Hochleistungsrechnen und datenintensive Dienste. *Praxis der Informationsverarbeitung und Kommunikation*, 36(2):99–108, 2013.

[2] G. Schneider, M. Hebgen, Horstmann K., H. Hartenstein, M. Waldvogel, P. Leinen, M. Resch, T. Walter, H. Großmann, and P. Castellaz. Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das Hochleistungsrechnen, 2012. http://dx.doi.org/10.15496/publikation-21185.

[3] Wissenschaftsrat. Empfehlungen zur Finanzierung des Nationalen Hoch- und Höchstleistungsrechnens in Deutschland (Drs. 4488–15). Stuttgart, Apr 2015.url:https://www.wissenschaftsrat.de/download/archiv/4488-15.html.

[4] H. Hartenstein, S. Wesner, R. Barthel, T König, and T. Nau. bwHPC-C5: Coordinated Compute Cluster Competence Centers (2013-2018): Ein Begleitprojekt zum Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das Hochleistungsrechnen (bwHPC), Jun 2013.

[5] B. Neumair, S. Wesner, R. Barthel, T. Nau, J. Salk, O. Schneider, and A. Streit. bwHPC-C5: Coordinated Compute Cluster Competence Centers (2013-2018): Phase 2 des Begleitprojekts zum Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das Hochleistungsrechnen (bwHPC), Oct 2015.

[6] G. Schneider, M. Hebgen, Horstmann K., H. Hartenstein, M. Waldvogel, P. Leinen, M. Resch, T. Walter, H. Großmann, and P. Castellaz. Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für datenintensive Dienste – bwDATA Phase I (2013-2014), 2012. url:http://dx.doi.org/10.15496/publikation-21188.

[7] G. Schneider, V. Heuveline, Horstmann K., B. Neumair, M. Waldvogel, P. Leinen, M. Resch, T. Walter, S. Wesner, P. Castellaz, H. Hartenstein, A. Streit, and M. Bestenlehner. Rahmenkonzept der Hochschulen des Landes Baden-Württemberg für datenintensive Dienste – bwDATA (2015-2019), 2015. url:http://dx.doi.org/10.15496/publikation-21187.

[8] Rat für Informationsinfrastrukturen. Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutsch-

land. Göttingen, 2016. url:http://www.nbn-resolving.de/urn:nbn:de:101:1-201606229098.

[9] G. Schneider, V. Heuveline, K. Horstmann, B. Neumair, P. Hätscher, J. Kolbitsch, S. Rehm, M. Resch, T. Walter, S. Wesner, and P. 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^{2}DM), 2018. url:http://dx.doi.org/10.15496/publikation-27872.

[10] S. Wesner, T. Walter, B. Wiebelt, and D. von Suchodoletz. *Strukturen und Gremien einer bwHPC-Governance – Momentaufnahmen und Perspektiven*, pages 315–330. De Gruyter Oldenbourg, 2016. https://doi.org/10.1515/9783110459753-027

Daten teilen – aber wie? Angebote der Informationsplattform forschungsdaten.info

Elisabeth Böker

bw2FDM, Team Open Science, Universität Konstanz

Wie gelingt das Teilen von Daten? In diesem Beitrag werden die Informationsangebote der zentralen deutschsprachigen Informationsplattform zu Forschungsdatenmanagement, forschungsdaten.info, vorgestellt. Die Bandbreite des Angebots reicht von inhaltlichen Beiträgen und fachspezifischen Informationen über ein FAQ bis hin zu Tutorials und Tools. Die Informationen werden von 30 FDM-Expertinnen und Experten aus Deutschland, Österreich und der Schweiz erstellt und unterliegen einer redaktionellen Prüfung. Abschließend wird darauf eingegangen, wie die Bekanntheit der Informationsplattform sichergestellt wird.

1 Einführung

Share Your Research Data! Dieser Aufruf sollte in sämtlichen Disziplinen eine Selbstverständlichkeit sein, denn im Forschungsdatenmanagement wird davon ausgegangen, dass die Forschung durch zugängliche Forschungsdaten profitiert. Allerdings stellt sich die Frage, wie die Daten geteilt werden sollen und was dabei zu beachten ist. Da dies nicht ohne Weiteres zu beantworten ist – nicht nur unterscheidet es sich von Disziplin zu Disziplin, sondern es ist auch von der Art der Daten abhängig – bedarf es profunder Informationen. Hier setzt die zentrale deutschsprachige Informationsplattform forschungsdaten.info an.

2 Über forschungsdaten.info

Die 2018 an den Start gegangene Plattform forschungsdaten.info [1] hat das Ziel, mit praxisnahen Artikeln in das Forschungsdatenmanagement (FDM) einzuführen. Dabei werden neben den Phasen im FDM im Bereich „Themen“ [2] auch disziplinspezifische Informationen unter „Wissenschaftsbereiche“ [3] aufbereitet. Im Bereich „FDM im deutschsprachigen Raum“ [4] wird ein Überblick gegeben, welche Projekte und Initiativen existieren. Das Angebot reicht von Landesinitiativen bis zur NFDI. Abgerundet wird die Informationszusammenstellung durch die Rubrik „Praxis kompakt“ [5], die mit einem FAQ, Glossar, einer Toolliste, Empfehlungen für Tutorials und einem Quiz aufwartet (Abbildung 1).

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13732 (CC BY 4.0)

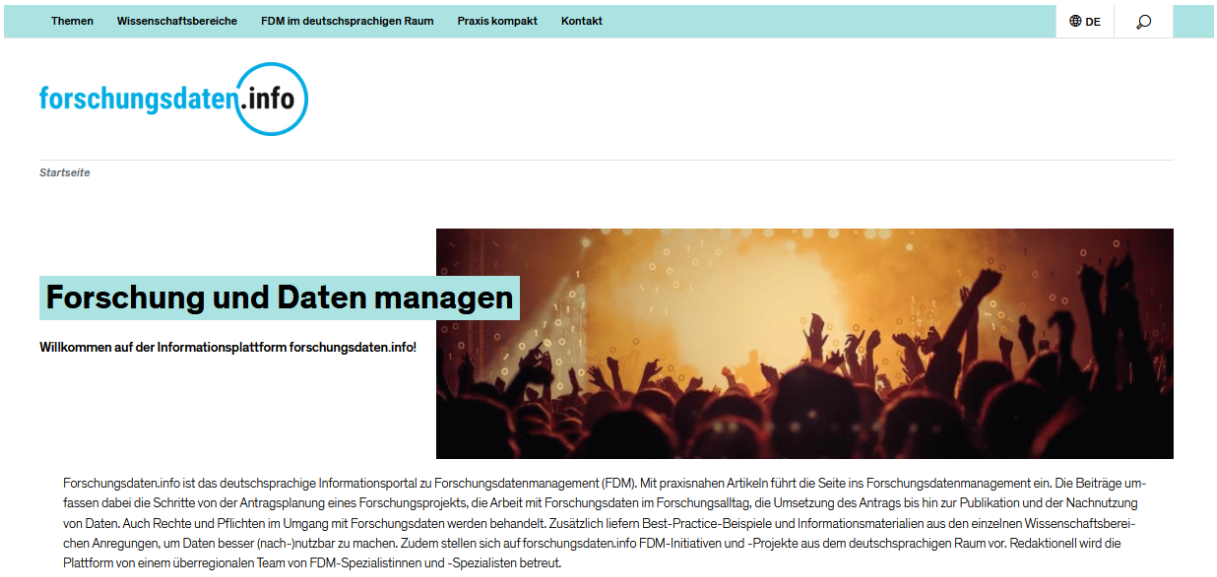


Abbildung 1: Startseite von forschungsdaten.info.

Die große Bandbreite der Artikel und die fundierten Inhalte sind insbesondere auf das große Redaktionsteam von forschungsdaten.info zurückzuführen. Rund 30 Expertinnen und Experten arbeiten an der Seite mit. Die Artikel sind aufgrund der Kennzeichnung mit CC 0 einfach nachnutzbar: Forschende finden gebündelt Informationen, FDM-Kontaktpersonen haben eine verlässliche Quelle und Einrichtungen können diese individuell für eigene Angebote – von Webseiten über Schulungen bis Prospekte – nachnutzen [6].

3 Angebote der Informationsplattform speziell zum Thema „Daten teilen“

Forschende sowie andere Interessierte finden in diesem breiten Spektrum an Themen auch etliche spezifische Angebote zum Thema „Daten teilen“. An erster Stelle ist die Rubrik „Veröffentlichen und Archivieren“ aus dem Bereich „Themen“ zu nennen [7]. Aspekte des Datenteilens werden von der Auswahl und Bewertung der Daten über die Datenpublikation bis hin zu Hinweisen zu Repositorien und Datenjournalen behandelt. Die Rubrik hat das Ziel, Forschenden zu vermitteln, was beim Teilen von Daten bei der Publikation zu beachten ist und wie sie eigene Daten teilen können. Ein wesentlicher Aspekt beim Teilen von Daten sind rechtliche Fragestellungen. In der Rubrik „Rechte und Pflichten“ [8] im Artikel „Recht und Forschungsdaten - Ein Überblick“ wird beispielsweise auf Regelungen zur Erhebung und Nutzung von Forschungsdaten eingegangen. Noch konkretere Informationen sind im Beitrag „Forschungsdaten veröffentlichen“ aufgearbeitet. Hier wird u. a. mit übersichtlichen Schaubildern gezeigt, welche Rechtsfragen zu klären sind und welche rechtlichen Hindernisse bestehen könnten. Dabei liegt das Augenmerk auf einer verständlichen und gleichzeitig korrekten Aufbereitung. Außerdem behandelt der Artikel, welche

Lizenzen sich für eine Publikation eignen. Disziplinspezifische Informationen zum Datenteilen sind vorrangig im „Wissenschaftsbereich“ zu finden [9]. Das Angebot reicht von einer Übersicht über Repositorien und Datenjournale, über Tools und Services bis zur Vorstellung von Projekten und Initiativen, wobei – und sogar in einer eigenen Rubrik – auch die NFDI-Konsortien aufgeführt sind. Wie alle Angebote der Seite werden diese von fachkundigen FDM-Kolleginnen und -Kollegen ausgewählt und bewertet, sodass eine Qualitätsprüfung gegeben ist. Für einen ersten Überblick zu einem Thema steht der Bereich „Praxis kompakt“ bei forschungsdaten.info [10]: Für das Thema „Daten teilen“ finden sich Antworten auf häufige Fragen im „FAQ“ [11], FDM-Begriffe werden im „Glossar“ [12] erklärt, und eine Übersicht an „Tools“ [13] sind auf der gleichnamigen Seite aufgeführt. Querverweise zwischen den einzelnen Beiträgen verknüpfen die Inhalte der Seite zu einem Wissensnetz, in dem sich die Nutzerinnen und Nutzer unkompliziert zurechtfinden.

4 Bekanntheit der Plattform

Doch wie werden Forschende auf diese Angebote aufmerksam, um von diesen Informationen zu profitieren? Ein gutes Ranking der Seite in Suchmaschinen ist die erste Hürde – soweit die Kriterien transparent sind, wird darauf geachtet. Darüber hinaus ist es hilfreich, dass die FDM-Community über das Angebot Bescheid weiß. Sichergestellt wird dies über Vorträge auf einschlägigen Konferenzen und die Verbreitung von neuen Informationen über die zentralen FDM-Mailinglisten. So können die FDM-Kolleginnen und -Kollegen an den einzelnen Standorten auf das Angebot etwa in Schulungen oder Beratungen verweisen. Zudem binden mehrere Einrichtungen ausgewählte Seiten von forschungsdaten.info ein, verweisen darauf bzw. nutzen diese für die eigenen Angebote [14]. Mit Werbematerialien wie Postkarten und Plakaten wird zudem auf forschungsdaten.info hingewiesen, teils wurde der Forschungsdatenpapagei auf Screens in den Bibliotheken etc. gezeigt [15] (Abbildung 2).

Auch wenn das Auslegen von Postkarten durch die Pandemie derzeit entfällt, so war dies zuvor, gerade kombiniert mit Vorträgen oder Informationsständen, eine Maßnahme, die die Sichtbarkeit für das Angebot erhöhte. Vorträge bei Fachtagungen oder in Rahmen von Workshops von NFDI-Konsortien sind in Planung und werden insbesondere die Kenntnisse über das Angebot bei den Forschenden selbst erhöhen. Auch auf Twitter ist die Redaktion aktiv. Anfang Juli 2021 folgten mehr als 730 Personen/Institutionen @ForschDatenInfo [16]. Für die weitere Bekanntmachung und für die Intensivierung des Austauschs mit Forschenden und der FDM-Community hat die Redaktion zudem zwei neue Angebote geschaffen: forschungsdaten.info live [17] bietet Vorträge der Radaktionsmitglieder zu aktuellen FDM-Themen und lädt die Teilnehmenden zur gemeinsamen

Diskussion ein. Der Newsletter forschungsdaten.info aktuell [18] informiert alle zwei Monate über Neuigkeiten von der Seite, eingeschlossen der FDM-News und Termine.

Die Nutzungszahlen von forschungsdaten.info sprechen dafür, dass das Angebot nicht nur wahrgenommen wird, sondern sich auch intensiver Nutzung erfreut (Abbildung 3).

Abbildung 2: Werbematerialien von forschungsdaten.info, Couleur/pixabay - Bearbeitung E. Böker/forschungsdaten.info CC BY 4.0.

Von 2018 bis 2020 stiegen die Besuche (unique visits, nicht zu verwechseln mit einzelnen Besuchern - unique visitors) um jeweils 50 %. Die Verweildauer auf der Seite stieg von 2 Minuten und 43 Sekunden im Jahr 2019 auf 3 Minuten 28 Sekunden im Jahr 20202. Auch für 2021 erwartet die Redaktion einen nochmaligen deutlichen Anstieg, der sowohl auf die wachsende Bedeutung des Themas als auch die Bekanntheit – verstärkt durch die Marketingmaßnahmen – zurückzuführen sind [19].

Danksagung

Die Informationsplattform wird im Rahmen des Landesprojektes bw2FDM durch das Ministerium für Wissenschaft Forschung und Kunst Baden-Württemberg gefördert. Seit Mai 2018 wird sie auf der Basis eines Memorandums of Understanding von der Universität Konstanz, dem KIT Karlsruhe und der Universität Hohenheim sowie zahlreichen Partnerinstitutionen und Einzelpersonen mittlerweile aus ganz Deutschland, Österreich und der Schweiz gepflegt und weiterentwickelt. Eine enge Zusammenarbeit mit dem Wiki forschungsdaten.org wurde 2019 vereinbart. Die Autorin dankt allen, die forschungsda-

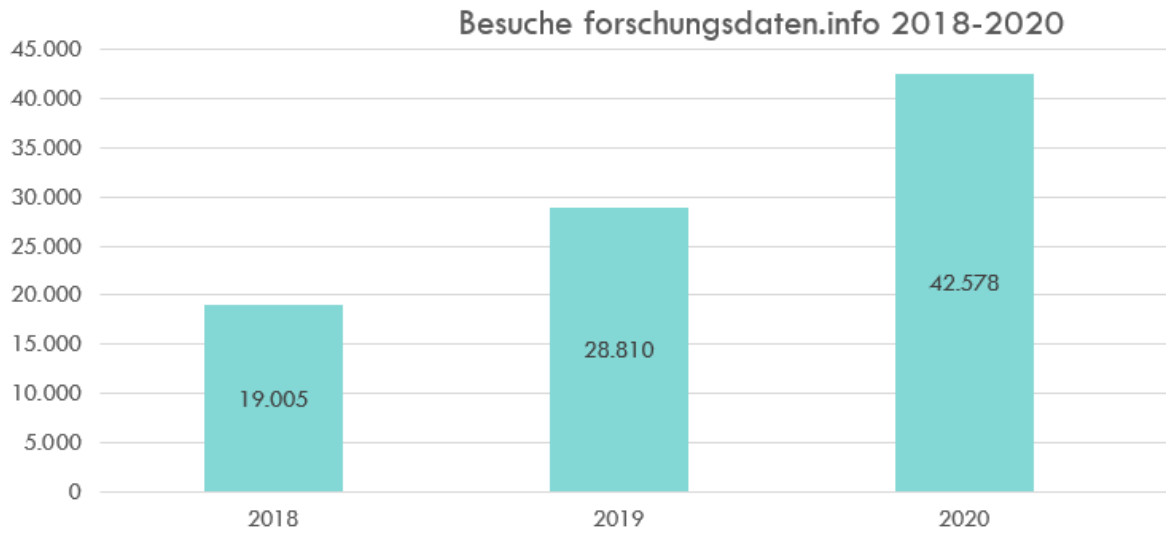


Abbildung 3: Besuche (unique visits) forschungsdaten.info 2018-2020.

ten.info ermöglichen, insbesondere den Kolleginnen und Kollegen aus der Redaktion, die das umfangreiche Angebot an Informationen mit Artikeln wie auch Ideen in den Redaktionsbesprechungen sehr bereichern.

ORCID ID

- Elisabeth Böker https://orcid.org/0000-0002-6025-3144

Literaturverzeichnis

[1] Informationsplattform forschungsdaten.info. `https://www.forschungsdaten.info` [26.4.2021].

[2] Informationsplattform forschungsdaten.info.Themen: `https://www.forschungsdaten.info/themen` [27.4.2021].

[3] Informationsplattform forschungsdaten.info.Wissenschaftsbereiche: `https://www.forschungsdaten.info/wissenschaftsbereiche` [27.4.2021].

[4] Informationsplattform forschungsdaten.info.FDM im deutschsprachigen Raum: `https://www.forschungsdaten.info/fdm-im-deutschsprachigen-raum` [27.4.2021].

[5] Informationsplattform forschungsdaten.info. Praxis kompakt: `https://www.forschungsdaten.info/praxis-kompakt` [27.4.2021].

[6] Böker, E., Brettschneider, P., Axtmann, A., Mohammadianbisheh, N. (2020). Kooperation im Forschungsdatenmanagement: Dimensionen der Vernetzung im Forschungsdatenmanagement am Beispiel der baden-württembergischen Landesinitiative bw2FDM. O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB, 7(4), S. 3-5. https://doi.org/10.5282/o-bib/5636; Kröger, Jan, und Kerstin Wedlich-Zachodin. 2020. „Das Beteiligungsmodell Von forschungsdaten.Info: Ein Kleines ABC Der Nachhaltigkeit“. Bausteine Forschungsdatenmanagement, Nr. 1 (April). German:86-95. https://doi.org/10.17192/bfdm.2020.1.8160.

[7] Informationsplattform forschungsdaten.info.Veröffentlichen und Archivieren: https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren [27.4.2021].

[8] Informationsplattform forschungsdaten.info.Rechte und Pflichten: https://www.forschungsdaten.info/themen/rechte-und-pflichten [27.4.2021].

[9] Informationsplattform forschungsdaten.info.Wissenschaftsbereiche: https://www.forschungsdaten.info/wissenschaftsbereiche [27.4.2021].

[10] Informationsplattform forschungsdaten.Praxis kompakt: https://www.forschungsdaten.info/praxis-kompakt [27.4.2021].

[11] Informationsplattform forschungsdaten.info.FAQ: https://www.forschungsdaten.info/praxis-kompakt/faqs-frequently-asked-questions [6.5.2021].

[12] Informationsplattform forschungsdaten.info.Glossar: https://www.forschungsdaten.info/praxis-kompakt/glossar [27.4.2021].

[13] Informationsplattform forschungsdaten.info.Tools: https://www.forschungsdaten.info/praxis-kompakt/tools [27.4.2021].

[14] Zum Beispiel: FID Crossasia: Forschungsdaten in den asienbezogenen Wissenschaften, crossasia.org, 2020, https://crossasia.org/service/forschungsdaten [22.11.2020]. Integriert werden u. a. die News, Veranstaltungshinweise und Themenbereiche auch bei den Projektpartnern selbst in Konstanz und am KIT. Vgl. KIM Konstanz: Forschungsdatenmanagement, kim.uni-konstanz.de, 2020, https://www.kim.uni-konstanz.de/openscience/forschungsdatenmanagement [22.11.2020], KIM Konstanz: Open Science, kim.uni-konstanz.de, 2020, https://www.kim.uni-konstanz.de/openscience/, Stand 22.11.2020. Auch Inhalte der FDM-Seiten der Universität Aachen basieren auf den Angeboten von forschungsdaten.info, etwa das Glossar. Vgl. RWTH Aachen University: Forschungsdatenmanagement von A bis Z, rwth-aachen.de, https://www.rwth-aachen.de/cms/root/Forschung/Forschungsdatenmanagement/~svkj/A-bis-Z/?page=1 [22.11.2020].

[15] Informationsplattform forschungsdaten.info. Materialien zu forschungsdaten.info: https://www.forschungsdaten.info/kontakt/materialien-zu-forschungsdateninfo/ [27.4.2021];

Böker, E., Brettschneider, P., Axtmann, A., Mohammadianbisheh, N. (2020). Kooperation im Forschungsdatenmanagement: Dimensionen der Vernetzung im Forschungsdatenmanagement am Beispiel der baden-württembergischen Landesinitiative bw2FDM. O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB, 7(4), S. 3-5. https://doi.org/10.5282/o-bib/5636.

[16] Twitteraccount der Informationsplattform forschungsdaten.info: https://twitter.com/ForschDatenInfo [7.6.2021].

[17] Informationsplattform forschungsdaten.info., forschungsdaten.info live: https://www.forschungsdaten.info/kontakt/forschungsdateninfo-live [27.4.2021].

[18] Informationsplattform forschungsdaten.info., forschungsdaten.info aktuell: https://www.forschungsdaten.info/kontakt/forschungsdateninfo-aktuell [27.4.2021].

[19] Die Nutzungszahlen beruhen auf einer Auswertung der Redaktion basierend auf dem Webanalytik-Bericht von Matomo Analytics. Nicht ausgerechnet wurden interne Nutzende, da dies aufgrund der diversen IP-Adressen (Büro und Homeoffice etc.) ein zu hoher Verwaltungsaufwand darstellen würde. Zentral ist hier die Tendenz der Entwicklung.

Extending a SKOS-based taxonomy catalog with collaborative features and an interface to provide terminologies to describe research data with interdisciplinary, semantic concepts

André Langer, Bach Tran and Martin Gaedke

Professorship for Distributed and Self-organizing Systems,
Chemnitz University of Technology, Germany

Publishing research data in the World Wide Web is typically done by uploading scientific files into a research data repository. Additional meta information can be provided, which is then used to improve the discoverability of this research dataset. However, search operations and filters are mainly keyword-based and commonly result in additional irrelevant or even missing search results, especially in an interdisciplinary research data sharing context. A semantic, concept-based approach can address this issue by relying on well-established taxonomies and linking similar concepts together. Taxonomy services already exist in different knowledge domains and provide concepts with identifiers in a controlled, quite static and isolated way. Essential features, such as collaboration, linking and integration, are often limited or missing, which are success factors for Web 2.0 applications and services. We therefore envision an interdisciplinary taxonomy service both accessible for humans and applications that can provide research concepts from different domains together with unambiguous identifiers and a flexible API to retrieve and manage available terms.

1 Introduction

In the context of OpenScience, researchers are encouraged to publish their research datasets in common data repositories so that others can find and reuse it. Following the FAIR Principles for scientific data management and stewardship [1], „(Meta)data (shall be provided with) assigned globally unique and persistent identifiers“ (F1). Establishment already took place for persistent identifiers that are relevant for mainly administrative and citation metadata, encompassing approaches such as DOI for unambiguously identifying publications, ORCID for identifying authors, ROR for organizations, and further more [2]. Beside that, standardized vocabularies are increasingly used to provide metadata for

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13733

research publications in a structured way, such as based on OpenAIRE Guidelines for Data Archives [8] (DataCite, DCAT-AP, Dublin Core).

Nevertheless, finding existing relevant research datasets in practice for reuse, replay or repurpose is still a challenge, as search queries have to be formulated carefully [4] and search results have to be reviewed individually, if they are actually related to the current research focus and required research data characteristics, or not, as shown in figure 1.

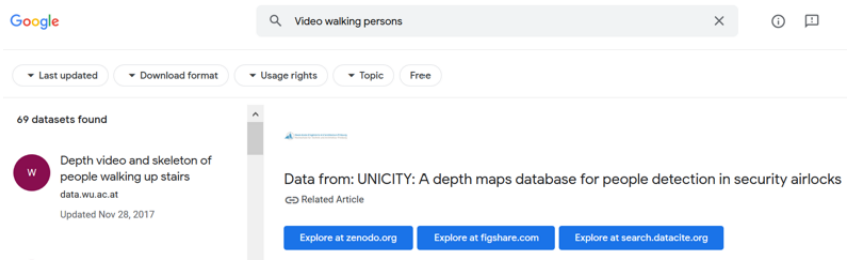


Figure 1: Example research data search query using Google Dataset search.

Reasons for that are, but are not limited to:

1. Filter operations are restricted to general characteristics, such as the download format, usage rights or topic area, as structured metadata descriptions commonly focus on general aspects as the least common denominator among knowledge areas
2. Search operations are commonly based on a keyword-based search within the available metadata title, list of keywords, and natural-language/floating text-based abstract
3. Providing highly structured metadata descriptions with mappings to similar concepts from other disciplines is demanding for users as not many appropriate research data publishing tools exist that consider structured concepts for descriptive purposes.

Especially the last aspect is astounding, as the FAIR principles suggest the user to provide rich metadata (F1) in standardized vocabularies (I2) with accurate and relevant attributes (R1). By using established terminologies with unique concept identifiers, researchers can make sure, that the provided meta information is structured, unambiguous and appropriate for retrieval activities in the future. Several web-based taxonomy catalogs and services already exist providing services for accessing a collection of existing terminologies and concepts. But provided terms might be incomplete, outdated or even contain wrong information. Instead, it would be a benefit if such a knowledge base of research-related concepts can be extended by a broader community, including the creation of missing terms, the categorization of existing concept groups and the interlinking between related concepts.

As part of the PIROL PhD project on Publishing Interdisciplinary Research Over Linked data [4], we currently implement and assess the extension of traditional terminology services with collaborative features. The results will directly contribute to the activities

in the ongoing collaborative research center SFB1410 Hybrid Societies. In this highly interdisciplinary research projects, more than 100 scientists from all main research areas closely work together to investigate the future interaction between people and smart devices, which will also require new methods to publish and discover relevant research data in a structured way.

The rest of the paper is structured in the following way: In section 2, we first elaborate the problem scenario and formulate objectives for improving the interdisciplinary research publishing and discovery process. Existing services are characterized in section Related Work. After that, we illustrate a conceptual architecture for an interdisciplinary, collaborative taxonomy service in section BIRD Concept. Section Conclusion will summarize our idea and describe our next steps.

2 Problem Analysis

When publishing research data, a metadata description should be provided that contains relevant indexable content for any potential stakeholder that might search for this dataset. The scientist carrying out the data description and publishing activity therefore has to carefully describe relevant characteristics from the user's point of view by providing appropriate words as keywords or textual content. In the following, we will focus on an uni-language scenario.

The main challenge from our perspective is, that even a very careful description can be insufficient for the discovery of this research data, because users in the future might use

1. **Generic or specialized terms**: broader class types or specialized words for a certain characteristic
2. **Homonyms**: equal words in a totally different context
3. **Synonyms**: different words than the publisher to express the same characteristic
4. **Weasel Words**: equal words that have a somehow different meaning in their discipline
5. **Unaware terms**: aspects that are applicable to the current data publication but not considered in the description by the publishing user

Aspect 1 can be addressed by relying on taxonomic (hierarchical) relationships and inference possibilities. Aspect 2 is related to the ambiguity of certain word labels and requires a classification or typification of terms with the same name but a different meaning. Aspect 3 and 4 can be taken into consideration by thesauri, that link similar words together. Aspect 5 cannot easily be addressed and is out-of-scope of our investigation.

Semantic technologies already exist that provide basic solutions for these aspects by introducing a uniform representation of knowledge-domain specific terminologies together with persistent unambiguous identifiers that represent a particular concept of the real world ("from strings to things") [6]. The feasibility and strength of such an approach was already shown in industrial application scenarios, such as manufacturing process chains, e-commerce or general-purpose search engines. In a scientific context, taxonomies for a certain knowledge area also already exist, but commonly in a decentralized, independent

or even unstructured availability. Thus, it is not trivial to access existing established identifiers for research-specific concepts, expose a list of known species for a particular concept type or map terms with a similar meaning together [7].

Our proposition therefore emphasizes an approach on how to make scientific concepts in existing research terminologies accessible in an interdisciplinary context by focusing on the following objectives:

OBJ1 Provision of an interdisciplinary semantic taxonomy platform

OBJ2 Ability to import existing taxonomies

OBJ3 Ability to filter for particular concepts, types and identifiers

OBJ4 Ability to collaboratively retrieve, add or update concepts

OBJ5 Ability to tag concepts or group of concepts

OBJ6 Ability to link concepts among different terminologies

The described semantic taxonomy service shall be a collaborative point for research concept information access and interdisciplinary reuse, and primarily designed as a knowledge base for application-based access by research data publishing tools. It can assist a user to describe research data submissions in a highly structured, unobtrusive, transparent way.

3 Related Work

The curation and standardization of relevant terms for a particular knowledge discipline is increasingly demanded for the digitization and exchange in scientific communication. One prominent representative is the initiative National Research Data Infrastructure (NFDI)[1] in Germany, which is currently being established through multiple consortia to offer platforms, services and counseling among all major knowledge disciplines in order to also create a link to international efforts, such as the European Open Science Cloud (EOSC) [8].

Dedicated services to access controlled terminologies are already offered [9] in various representations: *Terminology catalogs* (such as NCBO BioPortal, AgroPortal, gfbio [10]) and ontology collections (such as LOV, AberOWL, ORR, OLS, Ontobee) offer directories to search for existing controlled vocabularies, concepts and relationships within a particular research area. *Authority services* (such as LCSH, MeSH, FAST, EuroVoc) provide general classification concepts. And *knowledge bases* (such as DBpedia, Wikidata and Yago) offer structured data about encyclopedic information for general concepts from the real world [11].

The technical basis to realize these platforms are either dedicated implementations or Open source software projects (such as Skosmos, iQvoc). Especially the last group already makes use of a semantic (RDF) data model with standardized vocabularies (SKOS), however, access is commonly limited to read-only operations. Collaborative approaches

[1] `https://www.nfdi.de/`

are rare and realized in Wiki-based environments (such as based on Semantic MediaWiki), but naturally focus on general-purpose entity collections without a scientific research concept focus and limited internal data organization, import and homogeneous mapping possibilities.

4 Concept

To facilitate the interdisciplinary description of research data on a meta level, essential building blocks are currently missing that can provide structured information about sets of scientific concepts, such as investigated objects, applied methods, used devices, described characteristics or research objectives, and links to similar concepts among research disciplines.

In order to combine the strengths of a dedicated taxonomy registry with collaborative features, previously only known from general-purpose Wikis, we envision the **extension of a semantic base platform** that can set concepts from existing scientific terminologies into a relationship. These terminologies will remain decentralized and independent in their actual location and namespace, but the interdisciplinary taxonomy service will enhance access and mapping, and can collect improvement suggestions. A conceptual architecture is shown in figure 2.

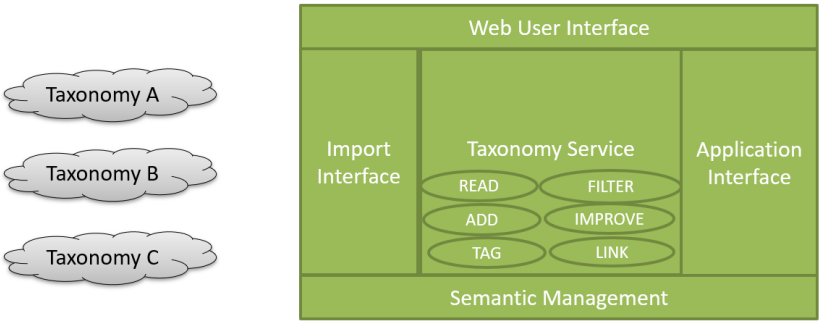


Figure 2: Interdisciplinary research taxonomy service with collaborative extensions.

An **Import Interface** offers a frontend possibility to import an already existing scientific taxonomy into the data corpus of the terminology service. In existing registries, this was either not possible at all, or only manually for administrators in the backend. This is especially relevant for the curation of proposed controlled vocabularies among different disciplines, e.g., across NFDI consortia. The taxonomy has to be represented in a common structured serialization format and is converted into a SKOS representation. Newly added taxonomies can of course go through a review process before being publicly incorporated.

A **Web User Interface** and **Application Interface** is offered to retrieve structured information (also including unambiguous persistent identifiers) about a particular taxonomy or concept. Extended filter possibilities will also allow querying for lists of concepts of a certain type or for synonyms among different terminologies.

An **Add and Improve functionality** allows suggestions for additions and corrections of the existing taxonomical content, as this might not be complete due to the open nature. These operations do not have to be carried out manually via the frontend interface, but can be announced in the background through any data annotation or publishing tool. The suggestions will be set into relationship to a specific taxonomy, but not incorporated in the original, controlled namespace.

A **Tagging component** realizes grouping operations of similar concepts among taxonomies from different disciplines to improve filtering operations for specific types.

Furthermore, a **Linking extension** sets the basis for further link discovery algorithmic operations and inference possibilities along similar research concepts.

5 Conclusion

In this paper, we described our vision for an interdisciplinary taxonomy service providing research concepts in a semantic way, both accessible for humans and machines. It adresses the challenge, that existing general encyclopedic services contain research-related concepts only in an incomplete way with limited categorization and filter possibilities. Controlled taxonomies that standardize scientific concepts and characteristics already exist, but in a decentralized, independent and inhomogeneous way. We have the hypothesis, that bridging research disciplines for research data publishing and discovery is a crowd-based effort to facilitate interdisciplinary research data publishing. Therefore, dedicated taxonomy platforms with additional collaborative features to improve, tag and link similar groups of research concepts have to be realized in order to make them accessible both for humans and research data management applications.

We currently work on establishing such an interdisciplinary taxonomy service originating in the Collaborative Research Center SFB1410 Hybrid Societies. It is built upon Skosmos and allows the collaborative collection, enhancement and integration of terminologies related to human-computer-interaction, overspanning all major research areas. Our next step is to populate it with existing taoxnomies from this knowledge domain and demonstrate its user input assistance in the submission form of research data publishing tools.

Acknowledgements

This work is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) Project ID 416228727 SFB 1410.

ORCID IDs

- André Langer https://orcid.org/0000-0001-7073-5377
- Martin Gaedke https://orcid.org/0000-0002-6729-2912

Bibliography

[1] Mark D. Wilkinson, Michel Dumontier, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. https://doi.org/10.1038/sdata.2016.18

[2] Angela Dappert, Adam Farquhar, Rachael Kotarski, and Kirstie Hewlett. Connecting the persistent identifier ecosystem: Building the technical and human infrastructure for open research. *Data Science Journal*, 16(0):1–16, jun 2017. https://doi.org/10.5334/dsj-2017-028

[3] Pedro Príncipe, Najla Rettberg, Eloy Rodrigues, Mikael K. Elbæk, Jochen Schirrwagen, Nikos Houssos, Lars Holm Nielsen, and Brigitte Jörg. Openaire guidelines: Supporting interoperability for literature repositories, data archives and cris. *Procedia Computer Science*, 33:92–94, 2014. 12th International Conference on Current Research Information Systems, CRIS 2014. URL: https://www.sciencedirect.com/science/article/pii/S1877050914008059, https://doi.org/https://doi.org/10.1016/j.procs.2014.06.015

[4] Hamlet Batista. 7 Reasons Why Search Engines Don't Return Relevant Results 100% of the Time. 2007. https://moz.com/blog/7-reasons-why-search-engines-dont-return-relevant-results-100-of-the-time, Accessed: 2021-05-01.

[5] André Langer. PIROL : Cross-domain Research Data Publishing with Linked Data technologies. In Marcello La Rosa, Pierluigi Plebani, and Manfred Reichert, editors, *Proceedings of the Doctoral Consortium Papers Presented at the 31st CAiSE 2019*, pages 43–51, Rome, 2019. CEUR.

[6] Christopher Erdmann, Natasha Simons, et al. Top 10 FAIR Data & Software Things, 2019. https://librarycarpentry.org/Top-10-FAIR/2019/09/05/linked-open-data/, Accessed: 2021-05-01. https://doi.org/10.5281/zenodo.2555498

[7] Ceri Binding and Douglas Tudhope. Improving interoperability using vocabulary linked data. *International Journal on Digital Libraries*, 17(1):5–21, 2016. https://doi.org/10.1007/s00799-015-0166-y

[8] Javad Chamanara, Angelina Kraft, Sören Auer, and Oliver Koepler. Towards Semantic Integration of Federated Research Data. *Datenbank-Spektrum*, 19(2):87–94, jul 2019. https://doi.org/10.1007/s13222-019-00315-w

[9] Andreas Ledl. Demonstration of the BAsel Register of Thesauri, Ontologies & Classifications (BARTOC). In *NKOS Workshop 2015*, 2015.

[10] Naouel Karam, Claudia Müller-Birn, Maren Gleisberg, David Fichtmüller, Robert Tolksdorf, and Anton Güntsch. A Terminology Service Supporting Semantic Annotation, Integration, Discovery and Analysis of Interdisciplinary Research Data. *Datenbank-Spektrum*, 16(3):195–205, nov 2016. URL: `www.naturkundemuseum.berlin`, `https://doi.org/10.1007/s13222-016-0231-8`

[11] André Langer, Christoph Göpfert, and Martin Gaedke. Querying the Semantic Web for Concept Identifiers to Annotate Research Datasets. *Fourteenth International Conference on Advances in Semantic Processing SEMAPRO 2020*, (c):49–55, 2020.

Forschungsdatenmanagement für ein interdisziplinäres Verbundprojekt

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

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

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

1 Einleitung

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

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

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

2 Ziel und Rahmenbedingungen

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

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

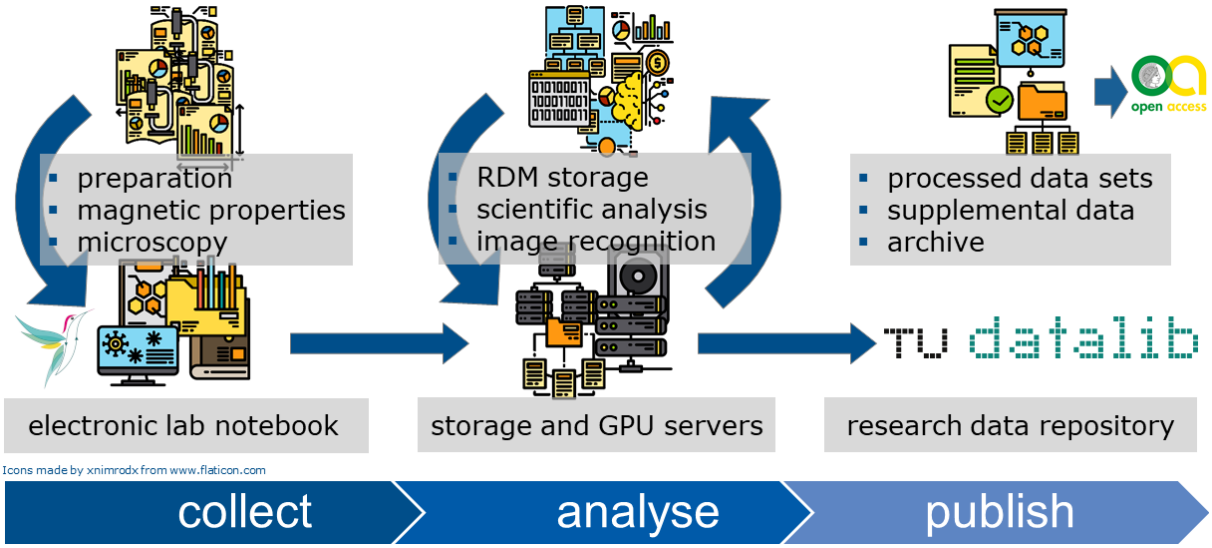


Abbildung 1: Die drei Hauptelementen der SFB/TRR 270 FDM-Infrastruktur: gemeinsam genutztes ELB, dezentraler „micro-cluster" aus Datenservern und Systemen optimiert für Maschinelles Lernen, institutionelle Forschungsdatenrepositorien zu Langzeitarchivierung.

3 Umsetzung

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

Das elektronische Laborbuch

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

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

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

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

Integration des SFB/TRR 270 FDM mit der lokalen Infrastruktur

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

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

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

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

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

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

Interaktion der Nutzer und die betrachteten Nutzerszenarien

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

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

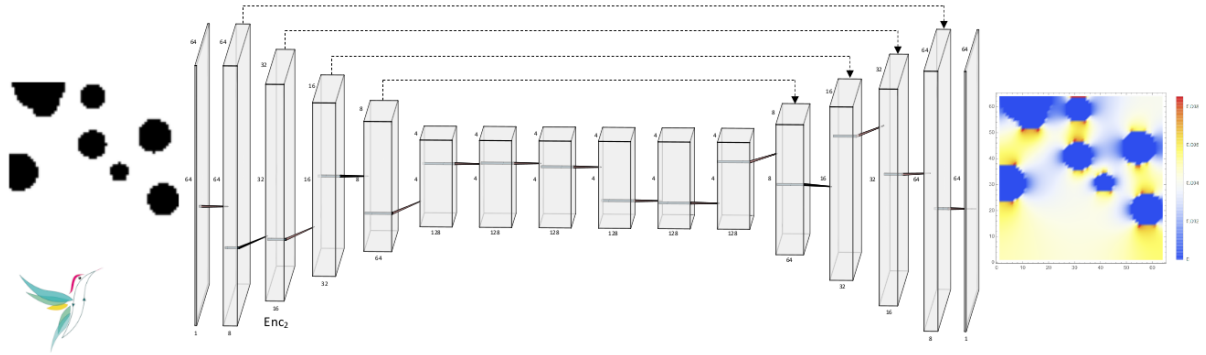


Abbildung 2: Im Prozesse des Maschinellen Lernens werden die von den erhobenen Forschungsdaten von den „Datenanalyst:innen" strukturiert und auf Deskriptoren reduziert und damit dann mehrstufig ein künstliches neuronales Netzwerk trainiert, um damit Strukturen mit interessanten Eigenschaften zu vorherzusagen.

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

4 Zusammenfassung

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

Danksagungen

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

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

ORCID IDs

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

Literaturverzeichnis

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

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

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

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

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

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

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

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

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

Forschungsdaten und Textpublikationen verknüpfen – Potenziale, Umsetzung und Herausforderungen

Julian Naujoks und Patrick J. Droß

Wissenschaftszentrum Berlin für Sozialforschung (WZB)

Als Extended Abstract zum Lightning-Talk „Gut verknüpft – besser auffindbar?" auf den E-Science-Tagen 2021 thematisiert dieser Beitrag die Verknüpfung von Text- und Datenpublikationen vor dem Hintergrund unserer Erfahrungen aus der Kuratierungs- und Veröffentlichungspraxis von sozialwissenschaftlichen Forschungsdaten am Wissenschaftszentrum Berlin für Sozialforschung (WZB). Im Sinne eines Werkstattberichts wird dabei auf Potenziale und Mehrwert, Umsetzung und Herausforderungen eingegangen. Unserer Erfahrung nach spielen bei der Verknüpfung vor allem das richtige Timing, die Heterogenität der Datenformate und die Prozesshaftigkeit eine zentrale Rolle. Dabei geht es nicht zuletzt um das komplexe und teils organisatorisch anspruchsvolle Vorhaben netzwerkartiger Verknüpfungen zwischen traditionellen und neuen Publikationsformaten im Sinne einer qualitätsgesicherten, offenen Wissenschaft.

1 Einleitung

Im Zentrum der Bemühungen um mehr offene Forschungsdaten stehen nach wie vor die Kernforderungen einer grundsätzlichen Zugänglichkeit und eindeutigen Refererenzierbarkeit von Forschungsdaten als eigenständige Forschungsergebnisse. Gleichwohl lohnt es sich, mit der zunehmenden Konsolidierung im Open-Data-Bereich auch einige bisher eher randständige Aspekte in den Blick zu nehmen. Hierzu zählt bspw. die Verknüpfung von Text- und Datenpublikationen. Mit dem folgenden Beitrag wollen wir uns diesem wichtigen Teilbereich im Gesamtprozess der Kuratierung und Veröffentlichung von Forschungsdaten widmen. Wir folgen dabei der Annahme, dass die Verbindung von nachhaltig veröffentlichten Daten mit (im besten Falle sogar frei zugänglichen Open-Access-) Textpublikationen erst die Transparenz schafft, die den Grundgedanken von Open Science, aber auch die Idee einer guten wissenschaftlichen Praxis auszeichnet. Im Rahmen dieses Extended Abstracts werden wir zum einen schlaglichtartig auf die Potenziale und praktischen Herausforderungen der Verbindung von Text- und Datenpublikationen hinweisen. Zum anderen werden wir unser eigenes Vorgehen am WZB vorstellen und wollen mit diesem Einblick zur aktuellen Diskussion beitragen.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13735

2 Potenziale

Die Verbindung von frei zugänglichen Daten mit (möglichst) ebenso frei verfügbaren Textpublikationen stellt hinsichtlich der Umsetzung der FAIR-Prinzipien in puncto Offenheit und Transparenz wissenschaftlicher Ergebnisse einen Idealzustand dar. Dies betrifft insbesondere den Start- und Zielpunkt von FAIRness (Auffindbarkeit und Nachnutzbarkeit). Zwar wird die Findability von Forschungsdaten durch neue Initiativen, Datenbanken und Suchmaschinen Schritt für Schritt erhöht, gleichwohl weisen aktuelle Studien auf „eine gewisse Divergenz zwischen Angeboten zur Datensuche und dem tatsächlichen Suchverhalten“ hin [1]. Demnach stoßen viele Forscher*innen im Alltag auch weiterhin vor allem über Textpublikationen bzw. allgemeine Forschungsliteratur auf die für sie relevanten Forschungsdaten (neben sozialen Kontakten/Netzwerken sowie Webseiten). Der direkte Verweis auf die einer Textpublikation zugrunde liegenden Primärdaten und deren Veröffentlichungsort würde die Auffindbarkeit von Daten entsprechend stark vereinfachen und weitere, ggf. aufwendige Rechercheschritte erübrigen.

Gleichzeitig fördern Textpublikationen, die aus einer Datenveröffentlichung direkt angesteuert werden können, das inhaltliche und methodische Verständnis der Daten (Stichwort Reusability). Sie erhöhen damit nicht nur das Potenzial ihrer direkten Nachnutzbarkeit, sondern leisten durch wichtige Kontextinformationen und -interpretationen insgesamt einen Beitrag zu einer verbesserten Datenqualität. Dies ist insofern von Bedeutung, da die Qualität der Datenveröffentlichungen letztlich mit darüber entscheidet, ob sich das grundlegende Ziel, nämlich die Nachnutzung der Daten, künftig in der erhofften Breite verwirklichen lässt [2].

3 Umsetzung am WZB

Das WZB war im Rahmen des Projektes SowiDataNet gemeinsam mit dem GESIS-Leibniz-Institut für Sozialwissenschaften, dem Deutsches Institut für Wirtschaftsforschung (DIW) und der ZBW-Leibniz-Informationszentrum Wirtschaft am Aufbau eines sozial- und wirtschaftswissenschaftlichen Fachrepositorium für Forschungsdaten beteiligt. Seit 2018 werden Daten von WZB-Forscher*innen in diesem Repositorium veröffentlicht. Parallel dazu wurde durch das Forschungsdatenmanagement am WZB ein umfangreiches Supportangebot etabliert. Dieses unterstützt die WZB-Forscher*innen von der finalen Aufbereitung der Daten, über notwendige Schritte der Anonymisierung bis hin zur Metadatenbeschreibung und finalen Veröffentlichung von Forschungsdaten im Repositorium. Flankiert wurde dies durch allgemeine Informationsangebote und Schulungen zu diesem Serviceangebot und einer grundsätzlichen Bewusstseinsförderung für die Idee offener Forschungsdaten. Wiederholt zeigte sich dabei, wie wichtig es ist, möglichst frühzeitig mit den Forscher*innen in Kontakt zu kommen. Um so eher die Datenpublikation eingeplant wird, um so besser lassen sich Daten vorbereiten, ggf. auch Bedenken der Forscher*innen aufgreifen und besprechen und im besten Falle auch Daten- und Textpublikationen koordinieren.

Im Sinne eines offenen Zugangs zu Forschungsoutput können die nachhaltig verfügbaren Forschungsdaten häufig auch noch durch Open-Access-Textpublikationen ergänzt werden.

Aus der Perspektive einer wissenschaftlichen Infrastrukturabteilung haben diese neuen Serviceangebote mit ihren vielseitigen Potenzialen natürlich auch diverse Fragen aufgeworfen. Die Verknüpfung von Forschungsdaten und Textpublikationen steht exemplarisch für die Anforderung, neue Services mit bestehenden Nachweissystemen und Publikationsprozessen zusammenzubringen. So waren wir am WZB – wie viele andere Einrichtungen vermutlich auch – damit konfrontiert, neben Bibliothekskatalog, Forschungsinformationssystem und Open-Access-Repositorium mit dem Open-Data-Repositorium ein weiteres System mit neuen Anforderungen in die Arbeitsabläufe zu integrieren.

Aufgrund fehlender Schnittstellen, abweichender (Metadaten-)Standards und nicht zuletzt unterschiedlicher Zuständigkeiten innerhalb unserer eigenen Abteilung war eine vollständig integrierte Lösung aller Informationssysteme vorerst nicht realisierbar. Neben der allgemeinen Bewusstseinsförderung galt es daher, sich um das konkrete organisatorische Zusammenspiel der Infrastrukturangebote zu kümmern, vor allem auch in Bezug auf die Verweise zwischen Text- und Datenveröffentlichungen. Aktuell bedeutet dies zumeist eine systematische manuelle Pflege der Verknüpfungen über die einzelnen Systeme hinweg: Am WZB bemühen wir uns im Zuge von Datenveröffentlichungen um die Aufnahme der Datenzitation inkl. Digital Object Identifier (DOI) direkt in den Zeitschriftenaufsatz oder in andere Textpublikationen (z.B. Monografie, Discussion Paper).
Zentrale Textpublikationen der Primärforscher*innen werden im Gegenzug im Datenrepositorium aufgenommen. Im Bibliothekskatalog erfolgt der Nachweis der Datenpublikation bei den entsprechenden Texten. Gleiches gilt – falls möglich – auch für die Metadaten im Open-Access-Repositorium. Im Idealzustand erhält man so eine wechselseitige Referenzierung, die – unabhängig vom Einstiegspunkt der Suche – Daten und dazugehörige Publikationen oder umgekehrt Publikation und dazugehörige Daten leicht auffindbar macht. Letztlich entsteht so ein komplexes Netzwerk wechselseitiger Verweise zwischen alten und neuen Publikationsformaten. Ein Trend, der sich generell abzeichnet und der Ausdruck einer digitalisierten Wissenschaft ist: „Je mehr sich die Kultur des Data-Sharing verbreitet, desto häufiger wird es damit auch zu netzwerkartigen Verzweigungen zwischen dem Output dieser ehemals prä-publikatorischen Phasen und den traditionellen Formaten der Ergebnispublikation kommen“ [3].

4 Herausforderungen

Bei der Verknüpfung von Text- und Datenpublikationen stoßen wir in unserer Kuratierungspraxis jedoch immer wieder auch auf Herausforderungen. Drei zentrale Themenbereiche sollen nachfolgend geschildert werden.

Timing der Verknüpfung

Zunächst ist die Verknüpfung von Text und Daten im Wesentlichen eine Frage des Timings. Um die Datenzitation bspw. in einen Journal-Artikel aufnehmen zu können, muss die Datenveröffentlichung naturgemäß bereits erfolgt sein, bevor die finale Version für den Druck an das Journal geht. Aufgrund zum Teil langer Review-Verfahren stehen die Chancen hier grundsätzlich nicht schlecht, aber es bedarf stets einer gewissen Planung. Oft verläuft dies aber auch nicht idealtypisch und das Zeitfenster ist deutlich kleiner, um die Verknüpfung herzustellen. Dann kann es dazu kommen, dass die Daten noch nicht fertig kuratiert und veröffentlicht sind, aber der Artikel bereits final eingereicht werden muss. Diese Fragen stellen sich zudem in der Praxis oft am Ende des Projekts. Dies ist häufig der Zeitpunkt, bei dem die Forscher*innen sich gedanklich oder gar physisch schon an einer ganz neuen Etappe des klassischen Forschungszyklus befinden (neuer Projektantrag, neue Einrichtung, usw.). Wir versuchen dies in zweifacherweise zu entzerren, indem wir erstens seitens des Forschungsdatenmanagements am WZB beraten und Strategien für ein geschicktes Handling der Publikationen vorschlagen. Zweitens nutzen wir die hilfreiche Möglichkeit, den DOI vorab im Repositorium zu reservieren. So kann eine vollständige Datenzitation bereits in den Artikel aufgenommen werden, auch wenn der Datensatz noch nicht publiziert ist. Bei diesem Vorgehen müssen jedoch einige zentrale Punkte berücksichtigt werden: So sollte definitiv klar sein, dass die Daten veröffentlicht werden können und dass dies zeitnah geschehen kann, damit ab Erscheinen des Journals auch der DOI auflöst. Wesentliche Prüfschritte der Kuratierung, wie z.B. die Eignung der Daten, Datenschutz und urheberrechtliche Fragen, müssen folglich geklärt sein.

Heterogenität der Datenformate und Veröffentlichungsorte

Best Practice im Open-Data-Bereich ist nach wie vor zweifellos die Veröffentlichung in einem nachhaltigen Forschungsdatenrepositorium. Hierzu zählt je nach Disziplin die Veröffentlichung eines Rohdatensatzes bzw. aufbereiteten Datensatzes, der mit standardisierten und fachlichen Metadaten beschrieben ist, einen persistenten Identifikator hat und ggf. mit zusätzlichen Begleitdokumenten versehen ist (in den Sozial- und Wirtschaftswissenschaften sind dies häufig Fragebogen und Codebook). In der Praxis bemerken wir aber gerade im Bereich der Verknüpfung von Text und Daten eine zunehmende Ausdifferenzierung der Datenformate und Publikationsorte. Ein Beispiel dafür ist die von wissenschaftlichen Verlagen verstärkt geförderte und teils geforderte Veröffentlichung von Forschungsdaten zu Replikationszwecken auf den Verlagsseiten/-angeboten. Prinzipiell sind verschiedene Wege der freien Verfügbarmachung von Forschungsdaten aus einer Open-Data-Perspektive zu begrüßen, doch leider sind diese in der Praxis nicht immer nachhaltig gestaltet. Dies ist ein Umstand, auf den auch der Rat für Informationsinfrastrukturen (RfII) kürzlich im Zuge der Diskussionen um die sogenannten Enhanced Publications aufmerksam machte [4]. Wir beobachten, dass in vielen dieser Fälle die Daten unvollständig aufbereitet und kaum beschrieben sind. Als Supplementary Material werden sie teils sogar als ZIP- oder Word-Datei auf der Journal-Webseite abgelegt. Häufig handelt es sich um Tabellen

und Auswertungen, aber nicht um die Daten selbst. Bei diesen im Rahmen des Einreichungsprozesses des Artikels eher beiläufig stattfindenden Datenveröffentlichungen fehlt es zumeist an Kuratierung, Standardisierung und Qualitätssicherung durch unterstützende Infrastrukturangebote. Im ungünstigsten Fall verschwinden Daten sogar hinter einer Paywall, so dass sich nur noch schwer von offenen Forschungsdaten sprechen lässt. Sollte dieser Weg gewählt werden, versuchen wir in der Praxis angesichts der Heterogenität von Datenformaten gerade bei den Replikationsdaten darauf hinzuwirken, dass hier ebenso ein Mindestmaß an Qualitätsstandards berücksichtigt wird. Falls möglich, empfehlen wir jedoch bevorzugt, die Daten über ein nachhaltiges Datenrepositorium verfügbar zu machen.

Prozesshaftigkeit der Forschung

Der stark prozesshafte Charakter der Forschung ist für die Erstellung einer netzwerkartigen Referenzierung zwischen den Publikationsformen eine besondere Herausforderung. So stellt sich bspw. die Frage, auf welche Textpublikationen im Datenrepositorium verwiesen werden soll und ob bzw. wie diese Verknüpfungen aktuell zu halten sind. Dies ist vor allem bei langfristig angelegten Forschungsprojekten der Fall, in denen ein zentraler Projektdatensatz veröffentlicht wird und in Folge noch eine Vielzahl von Textpublikationen anschließen kann. Wie können die Verknüpfungen aktuell gehalten werden, vor allem wenn zwischen der Publikation der Forschungsdaten und den daraus entstehenden Texten längere Zeiträume liegen? Am WZB versuchen wir, auf verschiedensten Ebenen die Verknüpfungen so gut wie möglich zu erfassen und herzustellen. Zunächst entscheiden die Forscher*innen – als Expert*innen für ihr jeweiliges Forschungsprojekt – natürlich selbst, welche relevante Textpublikationen initial im Datenrepositorium aufgenommen werden sollen. Wir weisen zudem darauf hin, dass jederzeit weitere Textpublikationen nachgemeldet werden können. Gleichzeitig versuchen wir aber auch in etwas systematischerer Form, die Angaben im Forschungsinformationssystem zu nutzen, in dem Forscher*innen sowohl ihre Text- als auch ihre Datenpublikationen eintragen und dabei auch wechselseitige Verbindungen angeben können. Schließlich wird durch uns beim Upload von Textpublikationen in das Open-Access-Repositorium auf Verknüpfungen geachtet, um auch dort Verweise auf Daten herzustellen.

5 Fazit

Die Verknüpfung von Text- und Datenpublikationen ist ein wichtiger Teilbereich des Kuratierungs- und Veröffentlichungsprozesses von Forschungsdaten. Auch wenn das technologische Potenzial einer automatisierten Verknüpfung riesig ist, zeigt unsere Erfahrung am WZB, dass wir im aktuellen Stadium noch vielfach mit manuellen Mitteln aktiv sein müssen. Gleichwohl können auch damit schon wichtige Ergebnisse erzielt werden: Die Auffindbarkeit der Daten steigt und die gegenseitige Referenzierung leistet nicht nur der

Transparenz und Reproduzierbarkeit der Forschungsergebnisse Vorschub. Sie steht gleichzeitig auch exemplarisch für Verknüpfungen zwischen alten und neuen Publikationsformaten einer digitalisierten Wissenschaft. Auch wenn einige wichtige Entwicklungen noch abzuwarten bleiben und mit den Fragen des Zeitpunkts, der Ausdifferenzierung der Datenformate sowie der Prozesshaftigkeit von Forschung wichtige Herausforderungen skizziert wurden, zeigt sich: Bei der Thematik geht es nicht nur um die bloße Verknüpfung von Texten und Daten. Vielmehr stehen hier sowohl Fragen der Datenqualität als auch grundsätzliche Aspekte der Openness von Forschung im Fokus.

Literaturverzeichnis

[1] Friedrich, Tanja; Recker, Jonas (2021): Auffindbarkeit und Nutzbarkeit von Daten. In: Markus Putnings, Heike Neuroth und Janna Neumann (Hg.): Praxishandbuch Forschungsdatenmanagement. Berlin: De Gruyter Saur, S. 405-426 (hier S. 423). DOI: http://doi.org/10.1515/9783110657807-023

[2] Blasetti, Alessandro; Droß, Patrick J., Fräßdorf, Mathis; Naujoks, Julian (2017): „Digital ist teilbar. Potenziale und Erfolgsbedingungen von Open Access und Open Data". In: WZB-Mitteilungen, H. 155, S. 34-37. DOI: https://doi.org/10.5281/zenodo.4733943

[3] Breuer, Constanze, Trilcke, Peer (2021): Die Ausweitung der Wissenschaftspraxis des Publizierens unter den Bedingungen des digitalen Wandels, Herausgegeben von der Arbeitsgruppe „Wissenschaftspraxis" im Rahmen der Schwerpunktinitiative „Digitale Information" der Allianz der deutschen Wissenschaftsorganisationen, S. 6. DOI: https://doi.org/10.48440/allianzoa.041

[4] Rat für Informationsinfrastrukturen (RfII) (2019): Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, S. 89. Online verfügbar unter: https://rfii.de/?p=4043

FDM-Landesinitiativen und NFDI

Magdalene Cyra und Matthias Fingerhuth

Landesinitiative für Forschungsdatenmanagement – fdm.nrw

In zahlreichen Bundesländern gibt es sogenannte FDM-Landesinitiativen (FDM-LI). Diese sind in ihren Strukturen und Aufgaben sehr unterschiedlich, sie bergen jedoch ein großes Potential, um die NFDI mit der Gesamtheit der Forschungseinrichtungen in Deutschland zu verknüpfen. Die Möglichkeiten der systematischen Zusammenarbeit sollten in einem Dialog entwickelt werden.

1 Einleitung

FDM-Landesinitiativen (FDM-LI) besitzen das Potential, wesentliche Vermittler und Multiplikatoren der Nationalen Forschungsdateninfrastruktur zu sein. Im März 2021 haben Vertreter zahlreicher Landesinitiativen ein gemeinsames Positionspapier veröffentlicht, das sich für eine Integration der FDM-LI in die NFDI ausspricht [1]. Dieses Papier greift einige dort formulierte Aspekte auf und geht auf sie in weiterem Detail ein. Dafür wird zunächst der Hintergrund der FDM-LI dargestellt, der gegenwärtig noch sehr uneinheitlich ist. Daraufhin wird die Bedeutung lokaler FDM-Infrastrukturen und die Notwendigkeit ihrer Einbindung in die NFDI diskutiert, um schließlich auf die mögliche Rolle der FDM-LI einzugehen, gleichzeitig aber auch auf bestehende Hürden für ihren Einsatz einzugehen.

2 FDM-Landesinitiativen – Begriffsbestimmung und Netzwerk

Für den Begriff der FDM-Landesinitiative (FDM-LI) gibt es bislang keine allgemeingültige Definition. Grundsätzlich können unter dem Namen solche Einrichtungen gefasst werden, deren zentraler Schwerpunkt auf der Förderung von Forschungsdatenmanagement liegt, wobei sie ihre Aktivitäten auf einzelne Bundesländer fokussieren (ohne dass ihre Aktivitäten notwendig an der Landesgrenze halt machen). Diese Begriffsbestimmung gibt einen recht weiten Rahmen dafür, was als FDM-LI verstanden werden kann. Davon ausgehend ist es jedoch wichtig, auf die Unterschiede zwischen den Einrichtungen einzugehen. FDM-Landesinitiativen sind

- Vom Land einberufen (z.B. fdm.nrw) oder auf Eigeninitiative beruhend (z.B. RLP-FDM)
- Vom Land finanziell gefördert (z.B. fdm.nrw) oder nicht (z.B. SaxFDM)

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13736 (CC BY-SA 4.0)

- Zentralisiert (z.B. fdm.nrw) oder verteilt (z.B. HeFDI)
- In weitere Landesaktivitäten eingebettet (z.B. Hamburg Open Science) oder als Einzelinitiative existent (z.B. FDM-RLP)
- Als Zielgruppe auf Forschende (z.B. TKFDM) oder Infrastrukturmitarbeitende (z.B. fdm.nrw) fokussiert
- Fachspezifisch (z.B. FDM-Bayern) oder generisch (z.B. fdm.nrw)
- In ihrem Fokus auf Hochschulen beschränkt (z.B. fdm.nrw) oder auf die Breite der Forschungsinstitutionen ausgerichtet (z.B. SaxFDM)

Darüber hinaus treten sie teils als Bereitsteller von Infrastrukturen oder aber lediglich in einer vermittelnden Rolle zwischen Einrichtungen auf. Ebenfalls wichtig ist es zu erwähnen, dass sie sich hinsichtlich ihrer personellen Stärke, ihren Gründungszeitpunkten und ihren Laufzeiten als Projekte – sowohl hinsichtlich ihres Beginns als auch ihres Projektendes – deutlich voneinander unterscheiden. Auch hinsichtlich der konkreten Arbeitspakete, mit denen sie ihre Ziele verfolgen, sind sie überaus divers, wobei sie auf sich verändernde Umstände reagieren.[1] Unter diesen Aufgaben treten Vernetzung und Informationsvermittlung sowie Aus- und Weiterbildung wiederkehrend hervor. Neben diesen offiziellen oder inoffiziellen FDM-Landesinitiativen gibt es jedoch auch in weiteren Bundesländern Personen, die innerhalb des Bundeslandes im Bereich FDM eine herausgehobene oder gut vernetzte Position einnehmen.

Im Kontext der RDA-Deutschland Tagung 2020 hat sich ein Netzwerk der FDM-LI und Kontaktpersonen in den Bundesländern gegründet, das sich über die Gemeinsamkeit der Arbeit auf Landesebene definiert. Das Netzwerk trifft sich regelmäßig für einen allgemeinen Austausch, darüber hinaus gibt es einen Dialog zu spezifischen Themen. Das gemeinsame Positionspapier markiert einen Auftakt für diese Zusammenarbeit. Es soll hier nicht wiedergegeben werden, im Folgenden werden jedoch einige wesentliche Punkte des Positionspapiers nochmals hervorgehoben und erläutert. Dies betrifft zunächst die Bedeutung des Einbezugs lokaler Strukturen und Dienste in die NFDI. Darauf aufbauend werden die Bedingungen und Möglichkeiten einer systematischen Zusammenarbeit von FDM-LI und NFDI diskutiert, die im Positionspapier angeregt wurde.

3 Die Unumgänglichkeit lokaler Strukturen und Dienste

FDM-Infrastruktur – sowohl technisch als auch personell – gewinnt an Bedeutung für sämtliche Forschungsstandorte. Es ist davon auszugehen, dass die Arbeit der NFDI (verstanden als Gesamtheit der Konsortien und des Direktorats) die Landschaft der FDM-Infrastrukturen nachdrücklich dynamisieren und wesentliche Dienste in sie erbringen wird,

[1] Die wohl umfassendste Zusammenstellung von Daten zu den FDM-LI ist jüngst in einer Umfeldanalyse der brandenburgischen LI veröffentlicht worden [3]. Sie zeigt jedoch auch die schnelle Entwicklung in der Projektlandschaft, da einige der Projektlaufzeiten seit Erhebung der Daten bereits verlängert worden sind.

doch dies lässt einzelne Standorte nicht aus ihrer Verantwortung, auch selbst ein gewisses Spektrum an Diensten anzubieten. Auch wenn die NFDI in den nächsten Jahren ihre Ziele erfolgreich umsetzt, wird sie auf absehbare Zeit nicht zu einer vollständigen Virtualisierung der Forschungsdateninfrastruktur und der angegliederten FDM-Services – also des gesamten Prozesses, der Daten von der Planung über die Dokumentation bis zur Archivierung und einer mögliche Nachnutzung begleitet – in Deutschland führen. Um die gegenwärtig vorhandenen und entstehenden Bedarfe der Forschenden bedienen zu können, braucht es deshalb lokale Infrastrukturen, die natürlich in die Gesamtheit der NFDI integriert werden müssen.[2]

Doch die Bedeutung des lokalen erschöpft sich nicht in einer Grundversorgung, auf die Spezialbedarfe aufbauen. Vielmehr kann man lokale Policies – etwa zur Verknüpfung von Praktiken an die Nutzung von Ressourcen – als das Rückgrat guter FDM-Praxis sehen, die dem Umgang mit Forschungsdaten an einzelnen Standorten einen Rahmen geben. Lokale Praktiken wiederum sind Grundlage für den vielfach angemahnten notwendigen kulturellen Wandel in Bezug auf Forschungsdaten. Dies ist auch mit Blick auf die Vermittlung dieses Umgangs mit Daten in der wissenschaftlichen Ausbildung an den Hochschulen von großer Bedeutung, denn dieser lässt sich nicht in Weiterbildungen vermitteln – vielmehr muss er von Grund auf in die Ausbildung integriert werden.

Die Bedeutung lokaler Angebote wird somit angesichts der bereits bestehenden und in Zukunft absehbar wachsenden Anforderungen an FDM kaum abnehmen. Ihre Integration in die NFDI ist eine Aufgabe, die es fortwährend umzusetzen gilt.

4 Möglichkeiten der Zusammenarbeit von FDM-Landesinitiativen zur NFDI

Es liegt nahe, Strukturen zu schaffen, die Einbeziehung der Basisdienstleister in die NFDI sicherstellen. Natürlich gibt es vielfältige Organisationen, die sich in diesem Zusammenhang einbringen können, etwa Verbände von Bibliotheken und Rechenzentren. Hier soll jedoch lediglich über die Rolle von FDM-LI eingegangen werden, denn durch die Verbindung von thematischem und räumlichen Fokus können FDM-LI einen spezifischen Beitrag zu diesen Strukturen leisten.

Dies ist keineswegs so zu verstehen, dass sie diese Strukturen bereits umfassend bieten. Zwar arbeiten FDM-LI und Konsortien der NFDI bereits gegenwärtig punktuell zusammen oder sondieren entsprechende Möglichkeiten. So positiv diese Berührungspunkte sind,

[2] Auch der Rat für Informationsstrukturen (RfII) hat ähnliches im Papier „Leistung aus Vielfalt"formuliert: „In einer nationalen, föderierten Infrastruktur wird weiterhin auch eine Grundversorgung mit Diensten durch kleinere Akteure zu finanzieren sein. Hier sind beispielsweise Bibliotheken, Archive, kleinere Rechenzentren oder IT-Infrastrukturen an wissenschaftlichen Instituten zu nennen. Diese bilden eine eigene, ebenfalls auf Kooperation angelegte Ebene in der NFDI und sind durch geeignete Organisationsstrukturen in funktionaler Weise in das System einzubinden-[2], S. 44.

so können sie mit Blick auf die Notwendigkeiten der Vernetzung vielleicht nicht die letzte Ausbaustufe darstellen. Mit alleine rund 400 Hochschulen in 16 Bundesländern und bis zu 30 NFDI-Konsortien können individuelle Kooperationen zwischen FDM-LI und Konsortien nur bedingt eine Breitenwirkung entfalten.

Gleichwohl ist festzuhalten, dass FDM-LI keine Partner sind, die in ihrer Gesamtheit aus dem Stehgreif in die NFDI integriert werden können. Ihre heterogenen Aufgabenzuschnitte und Strukturen, und nicht zuletzt der Umstand, dass sie bislang noch nicht umfassend existieren, machen sie kaum zu einem Universalwerkzeug für die Breitenwirkung der NFDI. Die Zusammenarbeit zwischen NFDI und FDM-LI muss deshalb gemeinschaftlich ausgearbeitet und entwickelt werden. Schon jetzt lassen sich aber Punkte ausmachen, die für eine Einbindung der Landesinitiativen sprechen.

Bewusstsein für die Herausforderungen, die im Zuge der NFDI auf die Forschungsinstitutionen zukommen, ist eine zentrale Voraussetzung dafür, dass sie bewältigt werden können. Wie vielfach beschrieben geht es um nicht weniger als einen Wandel in der Wissenschaftskultur, der neben der Wissenschaft selber auch die Einrichtungen betrifft, die diese tragen. Hier können FDM-LI einen wesentlichen Beitrag dazu leisten, Einrichtungen abseits der fachlichen Ebene, auf denen sich die Konsortien der NFDI bewegen, in den Transformationsprozessen zu begleiten. Dies betrifft darüber hinaus aber auch die Weiterbildung im Bereich FDM. Eine Abstimmung der Aktivitäten ermöglicht nicht nur ein effizienteres Erreichen der Zielgruppen. Sie erlaubt es auch, durch die Identifizierung zentraler Inhalte und durch die Formalisierung von Bildungsinhalten die Professionalisierung des Arbeitsfeldes FDM voranzutreiben.

Auch mit Blick auf die Gliederung von Aktivitäten auf politischer Ebene ist eine Einbindung der FDM-LI in größere Kontexte sinnvoll. Wo der NFDI e.V. eine bundesweite Organisation ist und Deutschland in der EOSC mandatiert vertritt, bieten FDM-LI das föderale Gegenstück zu dieser zentralen Einrichtung. Sie können angepasst an die Strukturen in den Ländern eine Brücke zur Breite der wissenschaftlichen Einrichtungen schlagen. Dabei können sie auch einen Schwerpunkt auf einen Einbezug auf einer infrastrukturellen Ebene legen, der die fachliche Ausrichtung, die im Fokus des NFDI e.V. steht, ergänzt. Diese Breitenvernetzung der FDM-LI ist auch vor dem Hintergrund von Kooperationen außerhalb des NFDI e.V. von Relevanz. Kooperationen im Bereich der Informationsinfrastrukturen bieten großes Potential für die Steigerung sowohl hinsichtlich der Qualität der angebotenen Dienste als auch ihrer Effizienz. Dies wäre nicht zuletzt im Sinne der Länder als wesentliche Träger von Forschungseinrichtungen. Überhaupt stellen FDM-LI hier eine mögliche Schnittstelle zu den Ministerien im Land dar, über die Desiderate und Entwicklungen gezielt gespiegelt werden können.

Im gemeinsamen Positionspapier nehmen die FDM-LI die Position ein, dass ihre Einbindung in die NFDI über eine Sektion erfolgen sollte. Eine solche förmliche Integration der FDM-LI ist natürlich keine Voraussetzung für eine Zusammenarbeit und Abstimmung mit den Akteuren der NFDI. Dennoch ist denkbar, dass sie klare Vorteile sowohl für NFDI als auch für FDM-LI bietet. Grundsätzlich bietet die Integration den FDM-LI größere Sichtbarkeit und dürfte ihren Stand in der Zusammenarbeit mit der NFDI verbessern. Für die

NFDI wiederum böten sie zunächst zentrale Ansprechpartner, über die sie Einrichtungen in den Bundesländern erreichen könnten. Weiter könnte sich aus einer Integration auch die Abstimmung eines Kerns von Aufgaben ergeben, die die FDM-LI innerhalb der NFDI einnehmen. Es versteht sich, dass dafür eine flächendeckende Existenz von LI notwendig wäre, die, wie bereits erwähnt, gegenwärtig nicht gegeben ist. Dies verweist auf die Rolle der Entscheidungstragenden in den Ländern. Eine starke Zusammenarbeit von NFDI und FDM-LI, die auch eine Perspektive auf die langfristige Übernahme von Aufgaben in der Struktur der NFDI entwickelt, böte sicherlich nicht nur Gründe für eine allgemeine Einrichtung von FDM-LI, sondern auch Gründe, sie aus ihrem Projektstatus zu heben. Angesichts des für die NFDI angelegten Zeitraums von mindestens 10 Jahren und der Hoffnung auf die Entwicklung fester Strukturen wären parallele Entwicklung für die Landesinitiativen letztlich im Sinne aller Beteiligten.

Literaturverzeichnis

[1] Axtmann, Alexandra; Böker, Elisabeth; Brand, Ortrun; Cyra, Magdalene; Dworschak, Nina; Fingerhuth, Matthias et al. (2021): "Wir bringen die breite Basis mit"– Gemeinsames Plädoyer für eine enge Einbindung der Landesinitiativen für Forschungsdatenmanagement in die Nationale Forschungsdateninfrastruktur. DOI: https://doi.org/10.5281/zenodo.4524655

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

[3] Wuttke, Ulrike; Neuroth, Heike; Rothfritz, Laura; Straka, Janine; Zeunert, Miriam; Schneemann, Carsten et al. (2021): Umfeldanalyse zum Aufbau einer neuen Datenkultur in Brandenburg. DOI:https://doi.org/10.25932/publishup-48090

The ReSUS Project - Infrastructure for Sharing Research Software

Markus Hirsch[1], Dorothea Iglezakis[1], Frank Leymann[2] and Michael Zimmermann[2]

[1]University Library, University of Stuttgart, Germany
[2]Institute of Architecture of Application Systems, University of Stuttgart, Germany

The goal of the ReSUS project is to develop an overarching concept and an operable solution to make research software more findable, accessible, interoperable, and reusable. It will help to archive and provision research software and data, and ensure that it will be executable and usable in the future regardless of the execution environment. It will create a concept to store the software and associated research data in one place, combine it with metadata, license information, a unique identifier, and other relevant information.

1 The Motivation for the ReSUS Project

There has been a lot of progress in the last years concerning FAIR data [1] in the scientific community. Several institutions, initiatives and working groups have exchanged ideas, developed principles, standards, and created initiatives to spread those principles and ideas [2, 3]. It has also become clear that not only research data should become findable, accessible, interoperable, and reusable, but also the research software that was used in the process of generating, analyzing, and interpreting this data [4, 5].

Currently, the awareness for the importance of publishing the code of self-created scientific software is slowly rising [6]. Publishing the plain code on platforms like GitHub[1] is a first step in the direction of more FAIR research software. However, investigations show that lack of documentation leads to difficulties in executing such code [7]. Even if a description on how to run the code is available, it might be too complicated for non-tech scientists to execute. Descriptions may also be incomplete or outdated as execution environments and required components may change over time. Legal uncertainties in the licensing of software can also reduce the likelihood of publications.

Therefore, the goal of ReSUS is on the one hand to make it as convenient as possible for the creators to publish their code by supporting them in selecting a suitable license, in describing the software with appropriate metadata and in modelling dependencies. On the other hand, ReSUS will make it easier to find and use existing research software by

[1] `https://github.com/`, all links last accessed on 2021-04-29.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13737 (CC BY-SA 4.0)

offering a search index and by automatically providing the software with all its dependencies. Encapsulating the code and all necessary artifacts in a self-describing open source container format will ensure that software can be executed independently from any particular executing environment. This will increase usability regardless of future technological developments and also benefit technically less experienced users.

2 Providing Help in the License Selection Process

Software licenses are very important for the reuse of research software. They determine under which conditions and restrictions a software can be used or integrated into own code. In the ReSUS project, we will focus on free/libre open source licenses (FLOSS)[3] for software.

For the creators of scientific software, who in the most cases are not trained software developers, it is difficult and time consuming to get a general overview about license types and regulations, and select a suitable license for their own code. When their own code is based on a multitude of other code parts or libraries under different licenses, even computer scientists and software developers may struggle with the complexity of license regulations and their implications [8]. With rising complexity, the danger of license violations rises, as certain licenses, even from the same license type, might be mutually incompatible.

There is a variety of accessible information and tools that help with this. The SPDX license list[4] is widely used as a reference. It gives a comprehensive presentation of FLOSS licenses and their versions. For each of them, it provides a permanent URL and a unique identifier.

Choosealicense[5] presents a good overview and categorization of the main features of a selection of the most widely used FLOSS licenses. It also provides a brief description for each license and a very concise guide to select a license according to one's preferences. This is very useful for newly created software, but does not take into account restrictions of licenses, if software is built upon existing code.

The Joinup Licensing Assistant[6] allows to search and compare FLOSS licenses according to a large variety of criteria. It also provides a compatibility checker[7] to verify, if a certain license may be assigned if the own code is built upon code with another license. However, only two licenses can be compared at the same time (one inbound, one outbound).

[3] https://dwheeler.com/essays/floss-license-slide.html
[4] https://spdx.org/licenses/
[5] https://choosealicense.com
[6] https://joinup.ec.europa.eu/collection/eupl/solution/joinup-licensing-assistant/jla-find-and-compare-software-licenses
[7] https://joinup.ec.europa.eu/collection/eupl/solution/joinup-licensing-assistant/jla-compatibility-checker

So the case with multiple inbound licenses can not be covered easily, plus a user without knowledge about licenses may not know which licenses to compare, as no recommendation is made.

As the information and functionality of these tools is very useful and may be used in other projects, the license checker in the ReSUS project will build upon this existing knowledge. The goal is to create an automated compatibility checker integrated in the publishing process that guides the creator to a suitable license.

To do so, our knowledge about FLOSS licenses will be represented in a license ontology, which is accessible via a SPARQL endpoint. It will be based upon existing information as mentioned above. It will contain the properties of licenses, such as permissions, limitations, and the condition of reuse. The relation and mutual compatibility of the licenses will be described via these properties.

We use Fossology [9] to look for existing licenses of used components or libraries in the source code. The license checker will then get a list of used licenses as an input, access the ontology, and return only licenses that do not contradict those existing ones. This will prevent unintended license violations. In a second step, creators will be able to select their preferences guided by questions ("Are there prerequisites from your institution concerning licenses?", "Do you want to have your software used by as many people as possible?"). The questions and the compatibility check will prevent the creator from unintentionally choosing a license that might inhibit reuse. More in depth information will be supplied for those who want more explanation, while keeping the process as simple as possible.

3 Adding Software Metadata

Crucial for the citability, the findability, and the reusability of research software are structered metadata. As stated in the Software Citation Principles in [10], most important for the citation of software are a persistent identifier (PID), information about the name and version of the software, the authors, the release date, and the location or repository of the software. With the Citation File Format[7] there is an uncomplicated way to add these information in form of a structured YAML-file to software code.

For the reusability of software, there has to be some more information in the metadata. According to the redefinition of the FAIR principles for software [4], software should be described with metadata "so that it can be replicated, combined, reinterpreted, reimplemented, and/ or used in different settings". This includes (quite vaguely) "a plurality of accurate and relevant attributes", a detailed provenance and qualified references to other software. Most important for the reusability of code is the statement of all dependencies that are necessary to get this code run. CodeMeta [11] addresses these challenges with a multitude of attributes not only to citation metadata but also - among others - about the status of development, required or optional dependencies, links to help, documentation

[7] `https://citation-file-format.github.io/`

and issue trackers, linking to a funding and embargos. CodeMeta bases on schema.org[8] the ontology for structured information on the web and is defined as a JSON-LD scheme. CodeMeta files are therefore somewhat more complicated to create, but much more powerful for describing research software.

But what about the plurality of accurate and relevant attributes that metadata of FAIR software should include? So far, both the Citation File Format and CodeMeta include general bibliographic information for citing or technical information about the software part of research software, but not about the research part. What kind of research can be done with this software? Which models or procedures are implemented? What methods can be used? These are information important for searching and important for linking software to other research outputs.

Within the ReSUS platform, the component to describe software with metadata is our data (and code) repository DaRUS basing on the repository software Dataverse[9]. In Dataverse, metadata schemes are defined in the form of metadata blocks that can then be activated for a dataset collection (so called dataverse) if required to describe the contained datasets. Dataverse is maintained and developed from an international community led by the Institute of Quantitative Social Science (IQSS) of Harvard. Within this community, a working group with our participation is discussing and working on the handling of research software within Dataverse including agreeing upon a metadata block for software basing on CodeMeta. Having this metadata block for the technical description of research software, the next step will be a guide to use descriptive metadata to describe the research aspects of the software.

4 Prototypical Implementation

In this section, we present our prototypical implementation planned for the ReSUS platform. Therefore, we firstly present an overview of the overall architecture of the platform. Furthermore, we describe the individual components of the platform in more detail and show where we are building on existing software components and standards.

Figure 1 shows the five main components of our prototype: (i) the *Library System*, (ii) the *Storage Backend*, (iii) the *ReSUS Frontend*, (iv) the *ReSUS Backend*, and (v) the *ReSUS Modeling Tool*. Furthermore, on the bottom of the figure, used external services are depicted, for example, Fossology[10] for checking the licenses or a DOI Provider for creating persistent IDs. On the right side of the figure, the ReSUS Modeling Tool for creating Research Object Archives (ROARs) [12] is depicted. In summary, ROARs enable the packaging, publishing, and installation of research software using a self-contained and portable packaging format.

[8] `https://schema.org/`

[9] `https://dataverse.org/`. For an overview of all components of ReSUS see section 4

[10] `https://www.fossology.org`

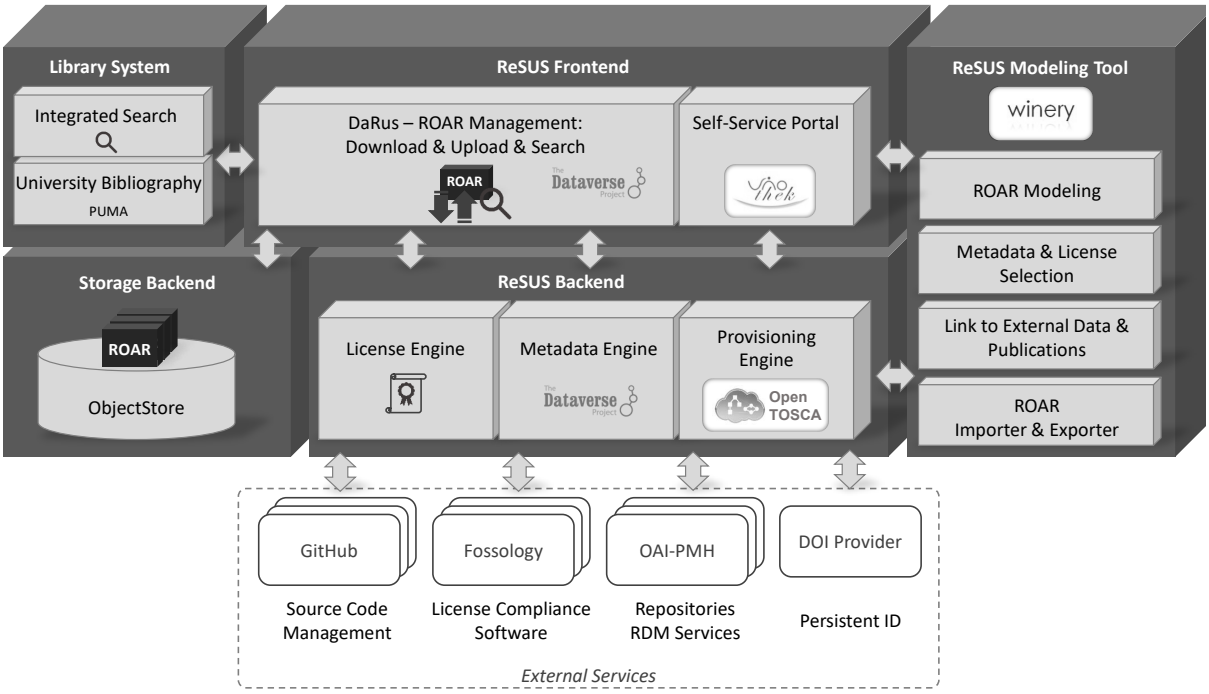


Figure 1: System architecture of the ReSUS platform.

In addition, all important information of research software, especially its technical dependencies, related research data, descriptive metadata, licenses, and references to corresponding publications can be bundled within a ROAR.

The ReSUS Modeling Tool is based on the OpenTOSCA modeling tool Winery [13] and extends it with additional required functionalities regarding the management and creation of ROARs. Here, the topology of the application with all required dependencies can be modeled and managed. Moreover, metadata can be added to describe the ROAR and a license can be selected. Required data can either be packaged directly within the ROAR, or referenced to a remote location. Likewise, related publications can be referenced. The ROARs can be exported from Winery as well as imported. They are self-describing and contain all artifacts and information necessary for the automated provisioning [14].

For modeling the topology of the application, the Topology and Orchestration Specification for Cloud Applications (TOSCA) [15, 16] is used. The OASIS standard TOSCA allows to define the deployment of applications by topology models as well as management plans. A topology model consists of the components of the application as well as their relations to each other. For example, it can be defined that a web application shall be hosted on an application server and also shall be connected to a database, where it reads data from. Such modeled applications can be executed automatically by a corresponding TOSCA runtime environment, like for example the TOSCA-based provisioning engine OpenTOSCA Container [17].

An extended version of the OpenTOSCA provisioning engine is part of the ReSUS Backend and is able to consume the ROARs exported from Winery and interpret the topology model describing the application. Furthermore, it executes all required steps in order to provision an instance of the modeled application [18].

Beside the modeling tool Winery and the provisioning engine Container, the self-service portal Vinothek [19] is also part of the OpenTOSCA Ecosystem.

In our ReSUS platform, the Vinothek is part of the ReSUS Frontend, and allows the end-user to manage and initiate the provisioning of the modeled application contained in a selected ROAR. If required, user specific as well as use-case specific variables, such as credentials to a cloud service or the location of data to be processed, are requested to be entered here by the end-user.

In addition to the OpenTOSCA ecosystem, with Dataverse and DaRUS we also build on further existing software. DaRUS[11] is part of the ReSUS Frontend for managing the ROARs. It is based on Dataverse and allows to upload, download, and search for ROARs. In the ReSUS Backend, our Metadata Engine is also based on Dataverse (see section 3). It supplies all components with required metadata and provides standardized interfaces for retrieving metadata and registering persistent IDs. The persistent ID ensures the citability and findability of the research software. The License Engine, which is also part of the ReSUS Backend, checks and manages the licenses of the research software. To do this, it uses external license checking software, such as Fossology or ScanCode[12]. Moreover, the License Engine is intended to assist the researcher in the selection of a license by suggesting a compatible license (see section 2).

On the left side of figure 1, the Library System is shown. The goal is to connect the ReSUS platform with the existing Library System in order to be able to search for ROARs and integrate them into the already available bibliography. Moreover, the depicted Storage Backend is used in order to store the ROARs in a sustainable way.

To sum up, we want to implement our concepts of the ReSUS platform based on standards and already existing software, such as TOSCA, the OpenTOSCA Ecosystem, and Dataverse. In the ReSUS project, we plan on extending and adapting the mentioned components by adding additionally functionality regarding the handling of ROARs, especially the linking of data, referencing of publications, creation of IDs, adding of metadata, and selection of licenses. The presented software components are open-source and can be obtained from GitHub[13,14].

[11] `https://darus.uni-stuttgart.de`
[12] `https://github.com/nexb/scancode-toolkit`
[13] `https://github.com/OpenTOSCA`
[14] `https://github.com/IQSS/dataverse`

5 Conclusion

In this paper, we outlined our goals and concepts for an operable solution to make research software more discoverable, accessible, interoperable, and reusable. Therefore, in this work, we first illustrated the current problems in the area of research software, for example, regarding the selection of an appropriate license in order to be able to publish it. Furthermore, we highlighted the importance of metadata for research software in order to fulfill the FAIR principles. Moreover, we presented our ideas and concepts in order to tackle the illustrated issues. Finally, we depicted an architecture overview and presented the single components of our planned ReSUS platform, which is based on existing software components and standards, such as the OpenTOSCA ecosystem and the TOSCA standard. In the future work, we focus on refining and implementing the presented concepts.

Acknowledgements

This work was partially funded by the German Research Foundation (DFG) project "ReSUS (Reusable Software University of Stuttgart)" (GEPRIS 425911815).

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 Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018–, March 2016. URL: `http://dx.doi.org/10.1038/sdata.2016.18`.

[2] Mark D Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. A design framework and exemplar metrics for FAIRness. *bioRxiv*, 2017. URL: `https://www.biorxiv.org/content/early/2017/12/01/225490`, `arXiv:https://www.`

biorxiv.org/content/early/2017/12/01/225490.full.pdf, https://doi.org/10.1101/225490

[3] Christopher Erdmann, Natasha Simons, Reid Otsuji, Stephanie Labou, Ryan Johnson, Guilherme Castelao, Bia Villas Boas, Anna-Lena Lamprecht, Carlos Martinez Ortiz, Leyla Garcia, Mateusz Kuzak, Paula Andrea Martinez, Liz Stokes, Tom Honeyman, Sharyn Wise, Josh Quan, Scott Peterson, Amy Neeser, Lena Karvovskaya, Otto Lange, Iza Witkowska, Jacques Flores, Fiona Bradley, Kristina Hettne, Peter Verhaar, Ben Companjen, Laurents Sesink, Fieke Schoots, Erik Schultes, Rajaram Kaliyaperumal, Erzsébet Tóth-Czifra, Ricardo de Miranda Azevedo, Sanne Muurling, John Brown, Janice Chan, Niamh Quigley, Lisa Federer, Douglas Joubert, Allissa Dillman, Kenneth Wilkins, Ishwar Chandramouliswaran, Vivek Navale, Susan Wright, Silvia Di Giorgio, Mandela Fasemore, Konrad Förstner, Till Sauerwein, Eva Seidlmayer, Ilja Zeitlin, Susannah Bacon, Katie Hannan, Richard Ferrers, Keith Russell, Deidre Whitmore, and Tim Dennis. Top 10 FAIR Data & Software Things, February 2019. https://doi.org/10.5281/zenodo.2555498

[4] Daniel S. Katz, Morane Gruenpeter, and Tom Honeyman. Taking a fresh look at FAIR for research software. *Patterns*, 2(3), March 2021. https://doi.org/10.1016/j.patter.2021.100222

[5] Anna-Lena Lamprecht, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, Stephanie van de Sandt, Jon Ison, Paula Andrea Martinez, Peter McQuilton, Alfonso Valencia, Jennifer Harrow, Fotis Psomopoulos, Josep Ll. Gelpi, Neil Chue Hong, Carole Goble, and Salvador Capella-Gutierrez. Towards FAIR principles for research software. *Data Science*, pages 1–23, November 2019. URL: https://doi.org/10.3233%2Fds-190026, https://doi.org/10.3233/ds-190026

[6] Hartwig Anzt, Felix Bach, Stephan Druskat, Frank Löffler, Axel Loewe, Bernhard Y. Renard, Gunnar Seemann, Alexander Struck, Elke Achhammer, Piush Aggarwal, Franziska Appel, Michael Bader, Lutz Brusch, Christian Busse, Gerasimos Chourdakis, Piotr Wojciech Dabrowski, Peter Ebert, Bernd Flemisch, Sven Friedl, Bernadette Fritzsch, Maximilian D. Funk, Volker Gast, Florian Goth, Jean-Noël Grad, Jan Hegewald, Sibylle Hermann, Florian Hohmann, Stephan Janosch, Dominik Kutra, Jan Linxweiler, Thilo Muth, Wolfgang Peters-Kottig, Fabian Rack, Fabian H.C. Raters, Stephan Rave, Guido Reina, Malte Reißig, Timo Ropinski, Joerg Schaarschmidt, Heidi Seibold, Jan P. Thiele, Benjamin Uekermann, Stefan Unger, and Rudolf Weeber. An environment for sustainable research software in germany and beyond: current state, open challenges, and call for action. *F1000Research*, 9:295, January 2021. URL: https://doi.org/10.12688%2Ff1000research.23224.2, https://doi.org/10.12688/f1000research.23224.2

[7] Christian Collberg, Todd Proebsting, and Alex M. Warren. Repeatability and benefaction in computer systems research. studie, 2015. URL: http://reproducibility.cs.arizona.edu/v2/RepeatabilityTR.pdf.

[8] Daniel A. Almeida, Gail C. Murphy, Greg Wilson, and Mike Hoye. Do software developers understand open source licenses? In *Proceedings of the 25th International Conference on Program Comprehension*, ICPC '17, pages 1–11. IEEE Press, 2017. https://doi.org/10.1109/ICPC.2017.7

[9] Michael C. Jaeger, Oliver Fendt, Robert Gobeille, Maximilian Huber, Johannes Najjar, Kate Stewart, Steffen Weber, and Andreas Würl. The FOSSology Project: 10 Years Of License Scanning. *Journal of Open Law, Technology & Society*, 9(1):9–18, 2018. URL: https://www.jolts.world/index.php/jolts/article/view/123.

[10] Arfon M. Smith, Daniel S. Katz, Kyle E. Niemeyer, and FORCE11 Software Citation Working Group. Software citation principles. *PeerJ Computer Science*, 2(e86), 2016. URL: https://www.force11.org/software-citation-principles, https://doi.org/10.7717/peerj-cs.86

[11] Matthew B. Jones, Carl Boettiger, Abby Cabunoc Mayes, Arfon Smith, Peter Slaughter, Kyle Niemeyer, Yolanda Gil Gil, Martin Fenner, Krzysztof Nowak, Mark Hahnel, Luke Coy, Alice Allen, Mercè Crosas, Ashley Sands, Neil Chue Hong, Patricia Cruse, Dan Katz, and Carole Goble. Codemeta: an exchange schema for software metadata. version 2.0. 2017. https://doi.org/10.5063/schema/codemeta-2.0

[12] Michael Zimmermann, Uwe Breitenbücher, Jasmin Guth, Sibylle Hermann, Frank Leymann, and Karoline Saatkamp. Towards Deployable Research Object Archives Based on TOSCA. In *Papers from the 12th Advanced Summer School on Service-Oriented Computing (SummerSoC 2018)*, pages 31–42. IBM Research Division, October 2018.

[13] Oliver Kopp, Tobias Binz, Uwe Breitenbücher, and Frank Leymann. Winery - A Modeling Tool for TOSCA-based Cloud Applications. In *Proceedings of 11th International Conference on Service-Oriented Computing (ICSOC'13)*, volume 8274 of *LNCS*, pages 700–704. Springer Berlin Heidelberg, December 2013. https://doi.org/10.1007/978-3-642-45005-1_64

[14] Michael Zimmermann, Uwe Breitenbücher, Lukas Harzenetter, Frank Leymann, and Vladimir Yussupov. Self-Contained Service Deployment Packages. In *Proceedings of the 10th International Conference on Cloud Computing and Services Science (CLOSER 2020)*, pages 371–381. SciTePress, May 2020.

[15] OASIS. *Topology and Orchestration Specification for Cloud Applications (TOSCA) Primer Version 1.0.* Organization for the Advancement of Structured Information Standards (OASIS), 2013.

[16] OASIS. *Topology and Orchestration Specification for Cloud Applications (TOSCA) Version 1.0.* Organization for the Advancement of Structured Information Standards (OASIS), 2013.

[17] Uwe Breitenbücher, Christian Endres, Kálmán Képes, Oliver Kopp, Frank Leymann, Sebastian Wagner, Johannes Wettinger, and Michael Zimmermann. The OpenTOSCA Ecosystem - Concepts & Tools. *European Space project on Smart*

Systems, Big Data, Future Internet - Towards Serving the Grand Societal Challenges - Volume 1: EPS Rome 2016, pages 112–130, December 2016. https://doi.org/10.5220/0007903201120130

[18] Michael Zimmermann, Uwe Breitenbücher, Michael Falkenthal, Frank Leymann, and Karoline Saatkamp. Standards-based Function Shipping - How to use TOSCA for Shipping and Executing Data Analytics Software in Remote Manufacturing Environments. In *Proceedings of the 21st IEEE International Enterprise Distributed Object Computing Conference (EDOC 2017)*, pages 50–60. IEEE, October 2017.

[19] Uwe Breitenbücher, Tobias Binz, Oliver Kopp, and Frank Leymann. Vinothek - A Self-Service Portal for TOSCA. In *Proceedings of the 6th Central-European Workshop on Services and their Composition (ZEUS 2014)*, pages 69–72. CEUR-WS.org, February 2014.

NFDI4Cat: Local and overarching data infrastructures

Sonja Schimmler[1], Thomas Bönisch[2], Martin Thomas Horsch[2], Taras Petrenko[2], Björn Schembera[2], Volodymyr Kushnarenko[2], Bianca Wentzel[1], Fabian Kirstein[1], Harald Viemann[3], Martin Holeňa[3] and David Linke[3]

[1]Fraunhofer Institute for Open Communication Systems, FOKUS
[2]High Performance Computing Center Stuttgart, HLRS
[3]Leibniz Institute for Catalysis, LIKAT

The NFDI is a German national initiative that aims to develop repositories, tools, standards, and best practices for research data management across all scientific disciplines. Until 2022, approximately 30 consortia will be formed under the umbrella of the NFDI e.V. association. NFDI for Catalysis-Related Sciences (NFDI4Cat) is one of these consortia, which targets research data management for catalysis-related sciences, a field that is of strategic importance for the economy and the society as a whole. In this paper, we give a brief overview of the consortium and introduce its planned local and overarching data infrastructures. We further describe our approach for requirements analysis, and provide some first insights on our findings.

1 Introduction

Catalysis is one of the key technologies for tackling challenges related to climate change. This research field is investigating the acceleration of chemical transformation by using a catalyst to increase the reactions efficiency and minimize unwanted side products at the same time. Each advancement in this catalytic research is an essential foundation for addressing problems like attaining CO_2 neutrality, finding a sustainable way to feed the worlds population or improving the valorization of plastic waste. The field of catalytic research is highly interdisciplinary covering bio-, electro-, photo-, heterogeneous and homogeneous catalysis as well as disciplines like reactor design and process engineering.

Catalysis-related sciences are currently facing some problems resulting in a slowdown of research advancement. There are many different companies and institutes working on catalysis research but most of the simulations and experiments take place in isolation, resulting in the repetition of experiments and simulations. There is a lack of standardization regarding the documentation of experiments, simulations and its data and metadata. There is also a lack of exchange. These problems can be mitigated by standardization and by setting up local and overarching data infrastructures that are specifically designed for catalysis-related sciences.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13738 (CC BY-SA 4.0)

As part of the NFDI (National Research Data Infrastructure), the consortium NFDI4Cat (NFDI for Catalysis-Related Sciences) was formed to tackle these challenges in catalysis-related sciences. The core objective of NFDI4Cat is to facilitate a fundamentally improved understanding of catalysis by building a bridge between theory, simulation, and experimental studies by addressing all aspects in the catalysis value chain from the catalyst design over characterization and kinetics to engineering aspects [13].

The consortium aims for a more standardized way of handling data throughout the research data lifecycle. It will develop ontologies for catalysis-related sciences to fully describe data and processes and build local and overarching data infrastructures for the community to enable storage and exchange of semantic rich data. The local and overarching research data infrastructures will support the whole research data lifecycle and will serve as an e-science solution for the field.

One challenge is to identify and serve the real needs of the NFDI4Cat community. Therefore, we will involve different stakeholders in the whole process, including a user-centered requirements analysis and setting up a pilot system to get early feedback from the users. Another challenge is to avoid fragmentation and data silos. Therefore, we will proceed with a coordinated approach.

The remainder of this paper is structured as follows: In Section 2, we will give a brief overview of the consortium. In Section 3, we will introduce the local and overarching data infrastructures that are planned within the consortium. In Section 4, we will describe our approach for requirements analysis, and provide some first insights on our findings. In Section 5, we will give an outlook and conclude the paper.

2 Consortial structure

The NFDI4Cat consortium assembles experts from homogeneous, heterogeneous, photo-, bio-, and electrocatalysis on the one side, and from process engineering and data technology on the other side. It gathers a total of 16 dedicated partners, experts from process engineering and data technology (High Performance Computing Center Stuttgart (HLRS), Fraunhofer Institute for Open Communication Systems (FOKUS), Max Planck Institute for Dynamics of Complex Technical Systems (MPI-DCTS), Karlsruher Institute for Technology (KIT)) and from catalysis and data driven catalysis research (Leibniz Institute for Catalysis (LIKAT), Max Planck Institute for Chemical Energy Conversion (MPI-CEC), RWTH Aachen, TU Berlin, TU Braunschweig, TU Darmstadt, TU Dortmund, Friedrich-Alexander-University Erlangen-Nürnberg, University of Greifswald, University of Leipzig, TU München, University of Rostock) coordinated by the DECHEMA Gesellschaft für Chemische Technik und Biotechnologie e.V. The project is supported by an advisory board of industrial partners, including BASF SE, Clariant Produkte GmbH, Covestro Deutschland AG, Evonik Industries AG, hte GmbH, Linde AG and Thyssenkrupp Industrial Solutions AG.

To achieve the project goals, the working programme consists of eight task areas. Task area one is responsible for metadata standardization and ontology development. Task area two focuses on data standards, data collection and interfaces, and task area three on data analysis, quality management and data reusability. Task area four handles the development of linked extensible infrastructures and data access management. Task area five takes care of the dissemination, outreach and training of catalysis researchers, task area six of the networking with other NFDIs, SFBs (Collaborative Research Centres) and international initiatives. Task area seven focuses on intellectual property, licences and reward models, and task area eight on the overall management of the consortium.

3 Planned national data infrastructures

One main goal of NFDI4Cat is to set up and establish local and overarching data infrastructures, as shown in Figure 1. This includes a distributed repository infrastructure and other local and overarching services that are needed by the NFDI4Cat community, in order to put forward a national environment for catalysis-related research data. To ensure future viability, the infrastructure will be built on existing standards and principles, *e.g.*, by using established vocabularies such as schema.org or W3C DCAT, and synchronized with other consortia and other communities. Open source solutions will be favored, relying on modern technologies, and using Semantic Web technologies.

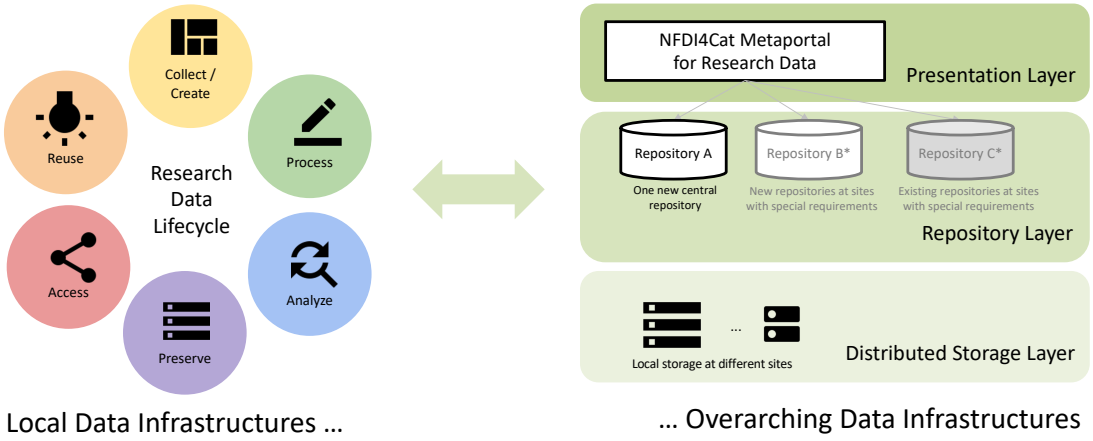


Figure 1: Local and Overarching Data Infrastructures.

3.1 Overarching data infrastructures

A distributed repository infrastructure will be set up, which will serve as an overarching data infrastructure. A layered architecture is planned, which includes a distributed storage layer, a repository layer, and a presentation layer. The distributed storage layer will enable the local storage at different sites. The repository layer will provide one new

central repository at HLRS and, where required, existing and new local repositories. The central repository as well as the new local repositories will be based on well suited existing solutions, such as Dataverse, DSpace or Invenio. The presentation layer will provide a general access point to the metadata and data that is openly available in the different repositories and will offer other overarching services that are identified of being useful for the NFDI4Cat community. This includes a graphical user interface as well as a SPARQL endpoint. The presentation layer will also provide an interface to interact with related infrastructures.

3.2 Local data infrastructures

Besides the overarching data infrastructure, local data infrastructures will be put forward, which support the whole research data lifecycle – collect/create, process, analyze, preserve, access and reuse. For this purpose, several labs working in different catalysis disciplines will setup pilots. Other groups will benefit from these pilots, either by reusing some of the local services established, or by learning from the pilots. Long-term goal of this effort is to include these services in a general toolbox.

4 Requirements analysis

Within the consortium, we follow an agile approach, meaning that the development is performed incrementally. Focusing on the overarching data infrastructure, we started with an initial requirements analysis, which will be refined later on. As a next step, a pilot system will be set up to get early feedback from the users.

4.1 User-centered approach for requirements analysis

Potential benefits from novel development work usually cannot be fully appreciated *in advance* by the majority of its future users; therefore, user-centered requirements analysis is limited and cannot be used as a substitute for good design. Within these limitations, however, it can play a role as an element of an encompassing strategy, supporting developers at anticipating community concerns and at ranking multiple design objectives. Thereby, it can assist at making and justifying conceptual design choices.

Accordingly, it is common practice in major coordinated software development efforts to conduct a user-centered requirements analysis. Previous experience has corroborated that this can be beneficial in developing research data infrastructures. In related fields, *e.g.*, Chen and Wu [3] chose a similar proceeding. Within the NFDI, *e.g.*, the sister project NFDI4Chem conducted an extensive community survey, identifying requirements for data generation, processing, annotation, sharing, publication, and reuse [1]. A similar questionnaire-based approach was followed for the NFDI4Ing requirements analysis [2].

The NFDI4Cat consortium relies on extensive in-person discussions with prospective users for its initial requirements analysis. At the time of writing, NFDI4Cat is in the process of collecting and evaluating a set of typical perspectives of prospective users of the research data infrastructures. So far, 17 individual users affiliated with member institutions have been interviewed, collecting pre-existing data management practices as well as requirements for the system architecture, for data documentation and annotation, and, hence, for metadata standardization.

In accordance with agile practices [4], we are currently in the process of extracting *personas*, *epics* and *user stories* from these perspectives. The expectation is that the requirements and design objectives obtained by analysing these perspectives are a sufficient first approximation to the needs of scientific research and development in catalysis at large, and that any further refinements can be carried out at a subsequent stage on the basis of first concrete experiences in working with a pilot system. So far, 4 personas (scientist, administrator/developer, data officer, external) were identified, and approx. 50 epics and 250 user stories were gathered.

4.2 Requirements for system architecture

The majority of the interview partners have expressed a strong desire or mandatory requirement for the bulk of their data to be hosted locally, with reliable mechanisms in place to ensure that intellectual property is protected and non-disclosure agreements with industrial and other research partners are honoured.

In some exceptional cases, especially where extensive previous work on digitalization of research data can be reused, bespoke local repositories will be developed or maintained; in other exceptional cases, where a local repository solution is required, a generic local repository (installed locally, but developed centrally) is preferable, since the benefit from reusing an established architecture will outweigh any potential benefit from developing a dedicated architecture for a bespoke local repository at the respective institution; in most other cases, however, a central repository will fulfil the requirement by enabling local storage at different sites. This also includes cases, where extremely large data sets (*e.g.*, from synchrotron beamline experiments) need to be processed.

For publishable data as well as training materials, and similar research and development assets, obversely, the interview partners expect NFDI4Cat to support a wide dissemination by enhancing their findability and accessibility, which would be optimally handled by a central component of the infrastructure, in particular, through a single point of entry.

4.3 Requirements for interoperability

Intra-platform interoperability requirements are deduced from the need for local and overarching components as well as multiple tiers of the repository architecture to communicate with each other in a well-defined way. This also concerns the exchange of data

and metadata with electronic laboratory notebooks (ELNs) used by the consortial partners and other researchers in catalysis-related sciences. The user interviews indicate that solutions for interoperating should be explored for various open source ELNs like Chemotion [7], Kadi4Mat ELN component [6] or LARAsuite [9] but also for commercial ELNs like FURTHRmind [8].

Inter-platform interoperability requirements, obversely, focus on the communication with external digital infrastructures that are expected to interact closely with NFDI4Cat in the future. A cross-disciplinary integration of services with other NFDI consortia is of interest (in particular, concerning data ingest, retrieval, and extraction), where major synergies are expected from interactions with NFDI4Chem[1] (for chemistry), NFDI4Ing[2] (for engineering sciences), and FAIRmat[3] (for materials science). A coordination with similar domain-specific consortia such as the UK Catalysis Hub, working towards inter-platform interoperability, may be advisable as well. Furthermore, it is highly desirable to attain a status of affiliation with the European Open Science Cloud (EOSC), and to develop the required cross-platform standards.

5 Conclusion and outlook

In this paper, we have given an overview of NFDI4Cat and its planned local and overarching data infrastructures. We have further described our approach for requirements analysis, and provided some first insights on our findings.

The requirements analysis is accompanied by documenting research workflows within the different labs. Research workflows discussed in the user interviews are documented and annotated, yielding preliminary semantic artefacts, *e.g.*, lists of typical steps in catalysis research workflows, measurement methods and tools, observed properties and data formats, and key performance indicators associated with catalyst performance assessment.

The requirements analysis is further accompanied by collecting competency questions from the users [10, 11]. This way, input for data documentation and annotation, and, hence, metadata standardization and semantic interoperability are retrieved. Competency questions are representative queries formulated by prospective users in informal language (*e.g.*, "what experimental data on catalyst material class X for reaction Y were published in year Z?"), which are expected to become formally expressible as SPARQL queries by ontology-based metadata standardization [11, 12].

[1] https://nfdi4chem.de
[2] https://nfdi4ing.de
[3] https://www.fair-di.eu/fairmat

Acknowledgments

This work was funded by the German Research Foundation (DFG) through the National Research Data Infrastructure for Catalysis-Related Sciences (NFDI4Cat), DFG project no. 441926934, within the National Research Data Infrastructure (NFDI) programme of the Joint Science Conference (GWK).

Bibliography

[1] S. Herres-Pawlis, J. C. Liermann, O. Koepler, "Research data in chemistry: Results of the first NFDI4Chem community survey," *Zeitschrift für allgemeine und anorganische Chemie* 646(21), 1748–1757, https://doi.org/10.1002/zaac.202000339,2020.

[2] G. W. Jagusch, N. Preuß, "NFDI4Ing: Rückmeldung aus den Forschungscommunities," Umfragedaten, NFDI4Ing consortium, https://doi.org/10.25534/tudatalib-104,2019.

[3] X. Chen, M. Wu, "Survey on the needs for chemistry research data management and sharing," *Journal of Academic Librarianship* 43(4), 346–353, https://doi.org/10.1016/j.acalib.2017.06.006,2017.

[4] M. Cohn, *User Stories Applied for Agile Software Development*, Boston: Pearson Education, ISBN 978-0-321-20568-1, 2004.

[5] S. Herres-Pawlis, O. Koepler, C. Steinbeck, "NFDI4Chem: Shaping a digital and cultural change in chemistry," *Angewandte Chemie International Edition* 58(32), 10766–10768, https://doi.org/10.1002/anie.201907260,2019.

[6] N. Brandt *et al.*, "Kadi4Mat: A research data infrastructure for materials science," *Data Science Journal* 20(1), 8, https://doi.org/10.5334/dsj-2021-008,2021.

[7] P. Tremouilhac *et al.*, "Chemotion ELN: An open source electronic lab notebook for chemists in academia," *Journal of Cheminformatics* 9, 54, https://doi.org/10.1186/s13321-017-0240-0,2017.

[8] H. Roth, D. Menne, J. Kamp, S. Emonds, H. Wollf, M. Wessling, "Schnell zu neuen Materialien: Effizientes Forschungsdatenmanagement an der Aachener Verfahrenstechnik," *Chemie Ingenieur Technik* 92(9), 1254–1255, https://doi.org/10.1002/cite.202055486,2020.

[9] M. Dörr, U. T. Bornscheuer, "Program-guided design of high-throughput enzyme screening experiments and automated data analysis/evaluation," pp. 269–282 in U. T. Bornscheuer *et al.* (eds.), *Protein Engineering: Methods and Protocols*, New York: Humana, ISBN 978-1-4939-7364-4, 2018.

[10] P. C. Barbosa Fernandes, R. S. S. Guizzardi, G. Guizzardi, "Using goal modeling to capture competency questions in ontology-based systems," *Journal of Information and Data Management* 2(3), 527–540, 2011.

[11] A. Fernández Izquierdo, R. García Castro, "Requirements behaviour analysis for ontology testing," pp. 114–130 in C. Faron Zucker, C. Ghidini, A. Napoli, Y. Toussaint (eds.), *Proceedings of EKAW 2018*, Cham: Springer, LNCS 11313, ISBN 978-3-030-03666-9, 2018.

[12] C. Bezerra, F. Santana, F. Freitas, "CQChecker: A tool to check ontologies in OWL-DL using competency questions written in controlled natural language," *Learning and Nonlinear Models* 12(2), 115–129, https://doi.org/10.21528/LNLM-vol12-no2-art4,2014.

[13] C. Wulf, M. Beller, T. Boenisch, O. Deutschmann, S. Hanf, N. Kockmann, R. Kraehnert, M. Oezaslan, S. Palkovits, S. Schimmler, S. A. Schunk, K. Wagemann, D. Linke, "A Unified Research Data Infrastructure for Catalysis Research - Challenges and Concepts," *ChemCatChem* 12(2), 115–129, https://doi.org/10.1002/cctc.202001974R2,2021.

Nationale Forschungsdateninfrastruktur für die Ingenieurwissenschaften (NFDI4Ing)

Britta Nestler[1c], Peter F. Pelz[2c], Robert H. Schmitt[3c], Marco Berger[4a], Hauke Dierend[5a], Benjamin Farnbacher[6a], Bernd Flemisch[7c], Dennis Gläser[7a], Ina Heine[3c], Nils Hoppe[6a], Gerald Jagusch[2a], Roland Lachmayer[5c], Jan Lemmer[2a], Jan Linxweiler[8a], Amelie I. Metzmacher[3a], Iryna Mozgova[5c], Nils Preuß[2a], Manuela Richter[2a], Stefanie Roski[9a], Hartmut Schlenz[10a], Michael Selzer[1a] und Christian Stemmer[6c]

[a]Autor:innen
[c]Kontribuierende
[1]Karlsruher Institut für Technologie
[2]Technische Universität Darmstadt
[3]Rheinisch-Westfälische Technische Hochschule Aachen
[4]Technische Universität Dresden
[5]Leibniz Universität Hannover
[6]Technische Universität München
[7]Universität Stuttgart
[8]Technische Universität Braunschweig
[9]Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden
[10]Forschungszentrum Jülich

NFDI4Ing ist ein 2017 gegründetes Konsortium mit dem Ziel, Wissenschaftler:innen aller Disziplinen zu ermöglichen, ingenieurwissenschaftliche Forschungsprozesse in ihrer Gesamtheit nachvollziehen oder reproduzieren zu können. Die Besonderheit an NFDI4Ing ist der Aufbau, welcher sich in drei Bereiche aufteilt. Die Archetypen, die an den methodischen Bedarfen ausgerichtet sind, die Community Cluster und die Base Services. NFDI4Ing erarbeitet technologische Methoden und Lösungen, bietet Aus- und Weiterbildungsprogramme und trägt zur Verbreitung des Forschungsdatenmanagements (FDM) in den Ingenieurwissenschaften bei.

1 Einleitung

NFDI4Ing hat das Ziel, Forschungsdaten in den Ingenieurwissenschaften FAIR zu machen – auffindbar, zugänglich, interoperabel und nachnutzbar [1]. Dabei stellt sich das Konsortium drei zentralen Herausforderungen [2]: (i) Bildung zur Sicherstellung der Datenkompetenz von Anfang an, (ii) Validierung von technologischen Lösungen und Methoden für Forschungsdaten sowie (iii) Erprobung von Konzepten, die Förderung von Data Governance und der Datenkuration. Die meisten TU9-Universitäten sowie das Deutsche Zentrum für Luft- und Raumfahrt (DLR), das Forschungszentrum Jülich, als auch

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13739 (CC BY-SA 4.0)

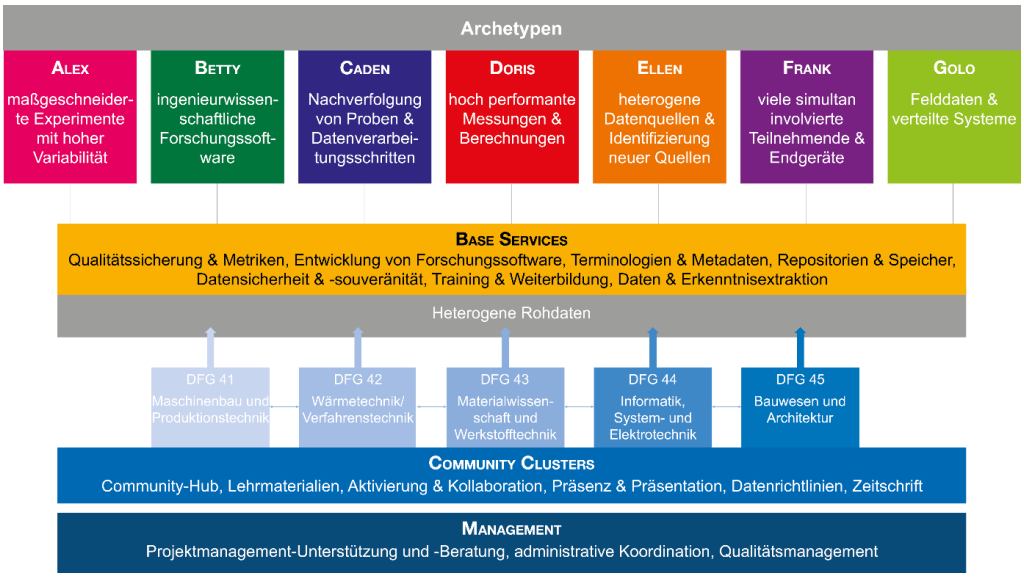


Abbildung 1: Aufbau von NFDI4Ing.

das Leibniz-Informationszentrum Technik und Naturwissenschaften Hannover (TIB) sind Gründungsmitglieder des Konsortiums.

Der Aufbau von NFDI4Ing orientiert sich an dem Aufbau eines Unternehmens mit Entwicklung, Produktion, Vertrieb und der Unternehmensleitung, s. Abb. 1. Die Entwicklungsabteilung wird durch die einzelnen Archetypen repräsentiert. In diesem modularen und methodenorientierten Ansatz bearbeitet jeder Archetyp spezifische Aufgaben und Ziele. Für den Vertrieb sind die Community Cluster (CC) zuständig. Sie bilden die Schnittstellen zu den jeweiligen Fachgebieten in den Ingenieurwissenschaften (DFG Fachgebiete 41-45). Die Base Services stehen für die Produktion, indem sie zentrale Dienste für die Archetypen und die Community Cluster zur Verfügung stellen. Weiterhin werden Dienste gepflegt und entwickelt, die von den Community Clustern als relevant erachtet werden. Das Management koordiniert und verwaltet die anderen Organisationsstrukturen.

2 Organisationsstrukturen

Archetypen

Die große Vielfalt ingenieurwissenschaftlicher Problemstellungen führt zu hochspezialisierten, individualisierten Forschungsansätzen und -methoden. Gleichzeitig lassen sich interdisziplinäre Gemeinsamkeiten auf der Ebene der verwendeten Forschungsmethoden und -prozesse erkennen [3]. Die Archetypen der NFDI4Ing, von Alex (maßgeschneiderte Experimente) bis Golo (Felddaten), harmonisieren diese Gemeinsamkeiten und repräsentieren die Vielfalt der FDM-Anforderungen im Ingenieurwesen.

1. Alex
Der Archetyp Alex repräsentiert Forschende, die physico-chemische Effekte technischer Systeme mit Hilfe maßgeschneiderter Experimente untersuchen. Diese können sowohl real als auch software-gestützt sein. In ihrer täglichen Arbeit konkurrieren Kompatibilität und Flexibilität, so werden häufig einzigartige Teillösungen für den jeweiligen Versuch entwickelt. Daher sind modulare, selbstdokumentierende und wiederverwendbare Softwarebausteine notwendig. Ziel ist, Konzepte und Best-Practice Beispiele für modulare Teillösungen zu erarbeiten. Dadurch werden die Nachvollziehbarkeit der Systemfunktionen, des Datenflusses und der Konfiguration verbessert. Weitere Schwerpunkte sind die Selbstdokumentation der Software und die automatische Generierung von maschinenlesbaren Metadaten. Dies soll zukünftig sowohl die Veröffentlichung und Wiederverwendung von Forschungsdaten fördern als auch die Entkopplung von Code und Daten erleichtern.

2. Betty
Der Archetyp Betty repräsentiert Entwickelnde von Forschungssoftware, bspw. für die Simulation physikalischer Systeme mit eigens entwickelten Methoden. Um die Transparenz der Forschung und die Reproduzierbarkeit der Ergebnisse zu gewährleisten, muss die Forschungssoftware nachnutzbar publiziert werden. Ziel der Task Area (TA) ist es deshalb, Forschende mit Werkzeugen und Wissen für die Entwicklung validierter und qualitätsgesicherter Ingenieurs-Forschungssoftware auszustatten. Dies beinhaltet Guidelines und Standards für die Entwicklung und Veröffentlichung von Forschungssoftware, sowie die Definition und/oder Anwendung von Metadatenstandards zur Beschreibung der implementierten Methoden und der Abhängigkeiten von anderen Softwarepaketen. Die in der TA entwickelten Ansätze werden zunächst in drei Pilotanwendungen aus den Ingenieurwissenschaften exemplarisch umgesetzt und durch Hinzunahme neuer Anwendungen systematisch erweitert und verallgemeinert.

3. Caden
Ein zentrales Interesse des Archetyps Caden ist das sog. Provenance-Tracking von Proben und Daten. Die zentrale Anforderung besteht darin, Datenentitäten (d. h. sowohl Daten als auch Metadaten) zu speichern und Parameter von Aktivitäten (z. B. Temperaturen, Drücke, Simulationsparameter) strukturiert und nachverfolgbar zu erfassen. Darüber hinaus müssen die Verknüpfungen der Entitäten zur Beschreibung einer Graphentopologie erstellt werden. Der Graph kann bei einer Vielzahl von Prozessschritten sehr komplex und nicht-linear sein (d. h. Verzweigungen und Abzweigungen enthalten). Eine weitere Herausforderung in diesem Archetyp ist die Kooperation unterschiedlicher Einrichtungen. Institutionen speichern Metadaten bisher isoliert voneinander und eine maschinenverarbeitbare Verknüpfung zwischen den Fragmenten des Graphen ist bisher nicht etabliert. Eine Möglichkeit dies zu lösen, ist der Einsatz einer einheitlichen Forschungsdateninfrastruktur, wie z. B. *Kadi4Mat* [4] oder *eLabFTW* [5].

4. Doris
Im Archetyp Doris werden Forschende repräsentiert, die hochgenaue und hochauflösende Messungen durchführen, und/oder Simulationen mit Hilfe von Höchstleistungsrechnern (HPMC) ausführen. Durch die Größe der erzeugten Daten (hunderte TB bis PB) sind diese nicht mobil, sondern nur über lokale und personalisierte Accounts bei den Re-

chenzentren abrufbar und nicht in Datenbanken oder Repositorien aufzufinden. Durch die Entwicklung einer interoperablen HPMC-Ontologie sowie passgenauer Softwarelösungen für die Auffindbarkeit und den Zugang zu HPMC-Daten und -Metadaten, soll der Zugriff durch Drittnutzende ermöglicht werden. Um diese Werkzeuge zu etablieren und deren Nutzung zu fördern, werden Forschende mit Workshops und Best-Practice-Guides unterstützt. Durch breiten Zugang zu interoperablen HPMC-Daten werden neue Forschungsansätze ermöglicht (Forschung an Daten), dazu zählen die Bereiche maschinelles Lernen, künstliche Intelligenz und neuronale Netzwerke.

5. Ellen

Der Fokus des Archetypen Ellen liegt auf komplexen Systemen, welche durch eine große Anzahl von multidisziplinären Abhängigkeiten charakterisiert werden. Diese sind dem Bereich der Datenwissenschaften zugeordnet, bspw. modellbasierte Simulationen und Optimierungsrechnungen. Die Berechnungen sind typischerweise sehr datenintensiv und erfordern Informationen aus vielen verschiedenen Disziplinen. Das Ziel der TA ist es, Ingenieur:innen bei dem Such- und Implementierungsprozess zu unterstützen, die Anzahl potenzieller Datenquellen zu erhöhen, deren Integrationsgrad zu steigern und den Zeitaufwand für den Suchprozess zu reduzieren. Dazu sollen methodische Konzepte und deren Softwareimplementierungen zur Verfügung gestellt werden, welche ebenfalls in der Lage sind, fehlende Daten zu generieren.**6.** Frank

Der Archetyp Frank repräsentiert Forschende, die sich interdisziplinärer Forschungsmethoden, z. B. aus der Produktionstechnik, der Ergonomie oder dem Wirtschaftsingenieurwesen, bedienen. Diese resultieren in einer Vielzahl verschiedener und heterogener, z. T. zu verknüpfender, Datenquellen, z. B. Fertigungsdaten in Echtzeit, Simulations- oder Probandendaten. Eine große Herausforderung bezieht sich somit auf die Synchronisation und das Zugriffsmanagement dieser heterogenen Daten in interdisziplinären Forschungsteams. Die Aufgaben der TA bestehen in der Etablierung gemeinsamer Ontologien und der Neugestaltung eines entscheidungsunterstützenden FDM-Frameworks, das u. a. die Datenrückverfolgbarkeit optimiert. Diese tragen zu der übergeordneten Zielsetzung bei, die Akzeptanz von FDM durch verbesserte Anwendbarkeit zu erhöhen und somit interdisziplinäre Forschungskollaborationen zu fördern.

7. Golo

Der Archetyp Golo repräsentiert Forschende, die sich mit der Planung, Erfassung, Klassifizierung und Analyse von Felddaten technischer Systeme beschäftigen. Felddaten dienen dazu, Modelle eines technischen Systems zu optimieren sowie detaillierte Analysen unter realen Betriebsbedingungen entlang des Lebenszyklus eines technischen Systems durchzuführen. Im Fokus steht die Entwicklung von Methoden und Werkzeugen zur Erfassung und Wiederverwendung der Felddaten. Anhand des Konzeptes eines Digitalen Zwillings erfolgt die Datenrepräsentation und -strukturierung nach den FAIR-Data-Prinzipien [6]. Zudem werden Checklisten und Best-Practice-Ansätze für die Arbeit mit Felddaten definiert. Durch eine Standardisierung von Schnittstellen und Abläufen soll die Wiederverwendung von Felddaten gefördert und erleichtert werden.

Base Services

Es sind sieben Base Services identifiziert worden, die für alle Archetypen und damit für die übergreifenden Leitziele von NFDI4Ing relevant sind und zentral entwickelt und erbracht werden: (i) Bereitstellung von Datenqualitätssicherungsprozessen, den entsprechenden Werkzeugen und Datenqualitätsmetriken, um Daten FAIR [6] zu machen; (ii) Unterstützung bei der Entwicklung von Forschungssoftware; (iii) Bereitstellung einfach zu bedienender und verständlicher Metadatenwerkzeuge und die Etablierung detaillierter ing.-wiss. Terminologien; (iv) sichere Speicherung und Langzeitarchivierung von Daten sowie die Möglichkeit, diese zu teilen oder zu publizieren; (v) Authentifizierungs-, Autorisierungs- und Rollenmanagementinfrastruktur (notwendig für Daten aus industrienahen Forschungsprojekten); (vi) Konzepte und Materialien für die Ausbildung; (vii) Techniken der Datenextraktion und Wissensentdeckung in der technischen Literatur.

Community Cluster

Der Aufgabenbereich Community Cluster ist das Vertriebsorgan von NFDI4Ing und verfolgt zwei Aufgaben. Erstens wird in sechs Arbeitsbereichen an Maßnahmen zur Verbreitung der entwickelten Dienstleistungen und Best-Practices gearbeitet. Diese sind: (i) Kommunikation, (ii) Lehre, (iii) Zusammenarbeit, (iv) Partizipation, (v) Standardisierung und (vi) Journal. Zweitens dient er als Schnittstelle zu den fünf DFG-Forschungsgebieten 41-45 und deren Bedürfnissen, wobei jedes Gebiet durch Co-Spokespersons repräsentiert wird. Die Forschungsgebiete werden im Community Cluster gezielt eingebunden, um die entwickelten Lösungen umzusetzen:

41 Maschinenbau und Produktionstechnik
Die Community 41 repräsentiert Fachdisziplinen, z. B. Mechanik, Textiltechnik, Produktionsorganisation und Betriebswissenschaften, die über die Verbindung von Grundlagen- und angewandter Forschung industrielle Herausforderungen adressieren. Die Community zeichnet sich durch zahlreiche Querschnittsthemen aus, die z. B. in Verbänden und Vereinen wie der Wiss. Gesellschaft für Produktionstechnik (WGP) gebündelt werden. Diese Vereinigungen werden im Community Cluster gezielt eingebunden, um einerseits die Bedürfnisse der Community zu ermitteln und andererseits die in NFDI4Ing entwickelten Lösungen in die Anwendung zu bringen.

42 Thermofluidmechanik und Verfahrenstechnik
Die Einbindung der Community erfolgt über Projekte, Vorträge und Workshops. Neben der bestehenden Anbindung an Verbundforschung werden Transferprojekte mit der Industrie durchgeführt. Die Zusammenführung und der Austausch mit den Teilcommunities und deren fachliche Perspektiven findet z. B. auf der PAAT 2020 [6], der 2. jährlichen NFDI4Ing Konferenz [7] mit dem Schwerpunkt FDM in der Thermofluidik, statt. In 2021 wird ein Community Board eingerichtet, welches hilft die Lösungen von NFDI4Ing in die Community 42 hinein zu tragen.

43 Materialwissenschaft und Werkstofftechnik
In den Materialwissenschaften ist die Umsetzung digitaler und automatisierter Arbeitsflüsse für eine beschleunigte und fortschrittliche Entwicklung neuer Materialien von großer Bedeutung. Repräsentative Arbeitsflüsse verknüpfen etablierte Algorithmen der Datenvorbereitung und -analyse. Die datengetriebenen Erkenntnisse und Informationen können als Graphen veranschaulicht werden und ermöglichen eine gezielte, anwendungsorientierte, maßgeschneiderte Auslegung neuer Werkstoffe.

44 Informatik, System- und Elektrotechnik
Die Community umfasst Disziplinen aus den Fachrichtungen Informatik, System- und Elektrotechnik. Fachübergreifende Querschnittsfragen ergeben sich dabei bspw. bei Anforderungen zur Handhabung großer Datenmengen, dauerhafter Datenaufbewahrung und -bereitstellung sowie disziplinübergreifender Modellentwicklung. Addressiert werden diese und weitere Fragestellungen durch Vorträge, Workshops und weitere partizipative Veranstaltungsformate.

45 Bauwesen und Architektur
Die Aktivierung der Community erfolgt u. a. durch Einbeziehen des Fakultätentags für Bauingenieurwesen, Geodäsie und Umweltingenieurwesen (FTBGU) sowie der Dekane- und Abteilungsleiterkonferenz für Architektur, Raumplanung und Landschaftsarchitektur (DARL). Dabei liegt ein besonderer Fokus auf dem Austausch der Communities mit den Archetypen, der bspw. durch regelmäßige Umfragen und das Einrichten von Community Boards realisiert wird.

Special Interest Groups (SIGs)

SIGs behandeln NFDI4Ing übergreifenden Themen bzw. Interessen und stellen die Querkommunikation sicher. Diese erfolgt sowohl zwischen den Organisationsstrukturen innerhalb von NFDI4Ing (Archetypen, Base Services, Community Cluster) als auch über die Grenzen von NFDI4Ing hinaus, bspw. durch die Einbindung externer Expert:innen. Bisher wurden zwei SIGs gegründet: (i) Metadaten und Ontologien und (ii) Basic FDM Training & Weiterbildung.

3 Ausblick

Wichtige Ergebnisse der Arbeiten von NFDI4Ing werden in einem neu geschaffenen Journal für Forschungsdatenmanagement im Ingenieurwesen sowie auf der Webseite des Konsortiums (nfdi4ing.de) veröffentlicht und auf der jährlichen NFDI4Ing-Konferenz vorgestellt.

Danksagung

Die Autor:innen und die Kontribuierenden möchten sich bei Bund, Ländern und bei der Gemeinsamen Wissenschaftskonferenz (GWK) für die Förderung und Unterstützung im Rahmen des Konsortiums NFDI4Ing bedanken. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 442146713.

Literaturverzeichnis

[1] Mark D. Wilkinson u. a. The FAIR Guiding Principles for scientific data management and stewardship. 3:160018, 2016. Journal Article. doi:https://doi.org/10.1038/sdata.2016.18.eprint:26978244.

[2] Robert H. Schmitt u. a. 209 .YHAu6j9CSUk. NFDI4Ing - the National Research Data Infrastructure for Engineering Sciences, 2020. https://zenodo.org/record/4015201#.YHAu6j9CSUk.

[3] Gerald Wolfgang Jagusch and Nils Preuß. Umfragedaten zu "NFDI4Ing - Rückmeldung aus den Forschungscommunities", 2019. doi:https://doi.org/10.25534/TUDATALIB-104.

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

[5] Nicolas CARPi, Alexander Minges, and Matthieu Piel. eLabFTW: An open source laboratory notebook for research labs. 2(12):146, 2017. doi:https://doi.org/10.21105/joss.00146

[6] DECHEMA e.V. Jahrestreffen der ProcessNet-Fachgemeinschaft Prozess-, Apparate- und Anlagentechnik (PAAT), 2020. https://dechema.de/en/PAAT2020_Programm/_/AK_Progr_2020.pdf(besucht am 09. 04. 2021).

[7] TU Darmstadt. Gepflegt, geteilt, digital lesbar, 2020-12-08.https://www.tu-darmstadt.de/universitaet/aktuelles_meldungen/einzelansicht_287808.de.jsp (besucht am 09. 04. 2021).

heiARCHIVE, a long-term preservation service at Heidelberg University

Martin Baumann[1,2], Florian Heß[1,3], Leonhard Maylein[1,3], Tatjana Mechler[1,2], Benjamin Scherbaum[1,2] and Eric Volkmann[2]

[1]Competence Centre for Research Data, Heidelberg University
[2]University Computing Centre, Heidelberg University
[3]University Library, Heidelberg University

heiARCHIVE is a new institutional service for long-term data preservation at Heidelberg University. It offers researchers an easy-to-use end-user platform for archival of their research data as well as the possibility to do a OAIS compatible long-term preservation containing features like format recognition, validation and conversion of files of appropriate file formats. heiARCHIVE is developed and will be operated by the Competence Center Research Data - a joint service facility of the University Computing Center and Heidelberg University Library. This work outlines the concept of the service and its current status.

1 Introduction

heiARCHIVE[1] is an upcoming institutional service for long-term data preservation offering researchers an easy-to-use end-user platform for archival of their research data. It is a dark archive that is based on the OAIS reference model, cf. [1]. heiARCHIVE is based on an in-house software development that offers features like format recognition/validation and extraction of metadata from files. A storage abstraction is realized based on the open source data management software iRODS[2] to manage data copies and geo-replication and the BagIt file packaging format (RFC 8493[3]) is used for structuring and naming directories and files. A dedicated right and role concept including billing management is available. Through service-local identity management, also alumni can use the service and users will prospectively also be able to do authentification using their ORCID[4].

heiARCHIVE is a service developed and maintained by the Competence Centre for Research Data[5] (KFD), a joint facility of the university's Computing Centre and the Uni-

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029723 veröffentlicht.

[1]https://heiarchive.uni-heidelberg.de/
[2]https://irods.org
[3]https://tools.ietf.org/html/rfc8493
[4]https://orcid.org
[5]https://data.uni-heidelberg.de

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13740 (CC BY-SA 4.0)

versity Library. In accordance with Heidelberg University's Research Data Policy it is the mission of KFD to provide the best possible support for the comprehensive and coherent management of research data for the university and its researchers. Amongst others, KFD offers also an institutional repository for open research data heiDATA[6] that is based on the open source web application Dataverse[7]. This service is used for data publication which is preferred over pure archiving wherever allowed and useful.

The software behind the heiARCHIVE service was started to be developed in 2017 since no available software could be found that would fulfill all specific requirements. On the one hand, an archiving solution for researchers to preserve their research data was needed, i.e. the software must be usable for end-users and allow for potentially very large capacity. On the other hand, elaborate long-term preservation processes for data of cultural heritage must be feasible. The long-term preservation of data must be included in existing processes of digitalization, data presentation or publication which requires certain levels or automatization and also of continuous adjustment between the primary data location and the archive. Additionally, a billing system must be integrated. Based on these requirements and expecting furthermore to come in the future, we decided to start a software development project to be flexible and put the focus on the features that we have the strongest demands.

2 Modular design and implementation

The modular design of heiARCHIVE follows the OAIS concept in implementing the data flow in terms of SIP, AIP and DIP, see Fig. 1. The data flow is managed by process steps that run sequentially and correspond to one module each: the inbox (prepare the data), ingest (package the data), storage (securely store the data) and access (to access the data). These modules are conceptually separated, can run on different (virtual) servers and interact only through a dedicated API. The heiARCHIVE admin module controls the overall process and also the state of all archive packages. Scaling compute resources and network bandwidth is possible by adding additional (virtual) servers, and can be helpful e.g. for intensive checksum operations. Today, data transfers from/to heiARCHIVE are realized via SFTP, but further protocols are intended. There is a graphical web interface for user interaction (GUI), both for end users and for maintainers. The full software stack of heiARCHIVE is based on Python 3. For the GUI, the scheduling and API endpoints are realized using the high-level Python Web framework Django[8].

Within the heiARCHIVE admin module, an indexer based on Solr[9] is integrated. Its main task is to make certain element contents of the metadata of all archived data retrievable. Today, this feature can be used by heiARCHIVE admins only, but might be activated for end-users in the future.

[6] https://heidata.uni-heidelberg.de/
[7] https://dataverse.org/
[8] https://www.djangoproject.com/
[9] https://solr.apache.org/

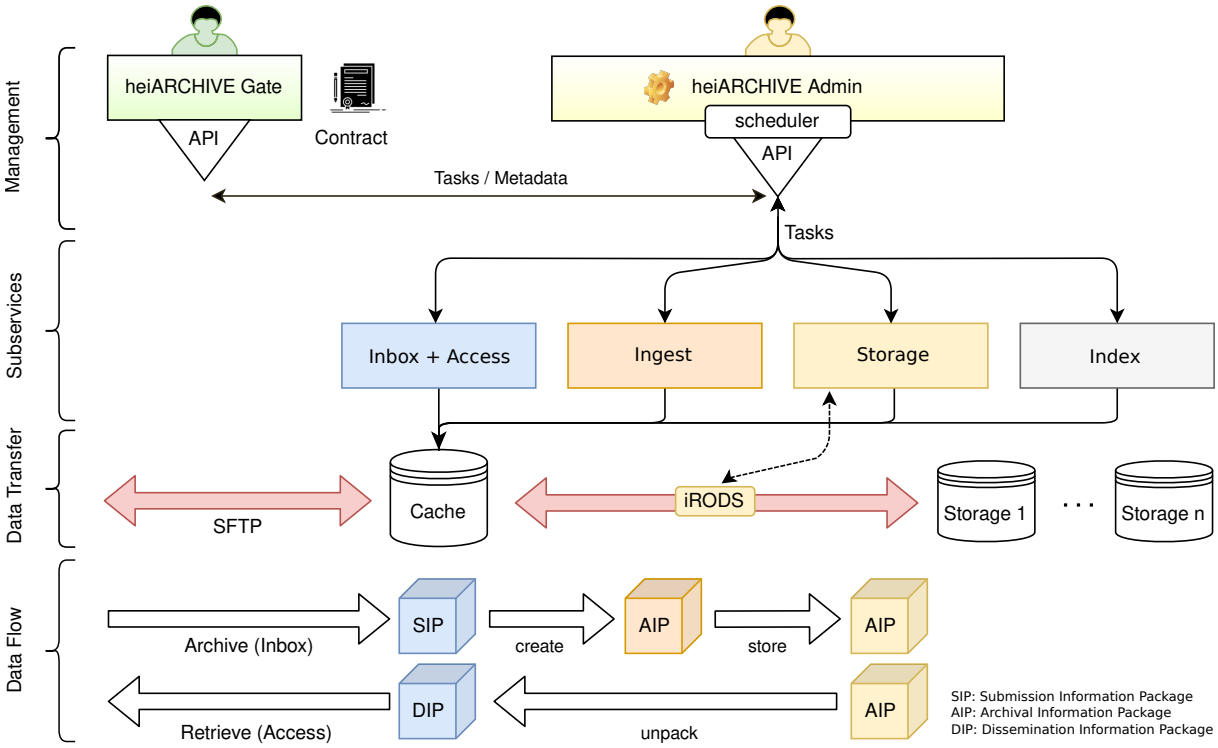


Figure 1: Illustration of the data flow, data management and modules that correspond to subservices of heiARCHIVE.

The indexer reads and validates the metadata files generated during the ingest process. In order to feed Solr with items in the expected format, XSLT transformations are used. Additionally, a selection of admin database items is indexed for faster retrieval compared to SQL-based queries.

3 Archive- and role management

heiARCHIVE has a hierarchical management structure and a related right and role concept for the users. The main navigation of the web interface reflects this management structure, see Fig. 2. The highest-level structure is denoted by "project" that can only be created by entitled persons (e.g. professors). A project establishes the link to a cost center ("Kostenstelle") and can also define a financial quota.

A project can contain one or multiple "archives" each of which is related to a data responsible person (e.g. a PhD student). The archive involves a set of archival parameters, e.g. the archive mode that is set to be either "long-term archiving" to include format validation or to be "bitstream-preservation". Descriptive metadata such as title, short description, project context, etc. is set at this level as well.

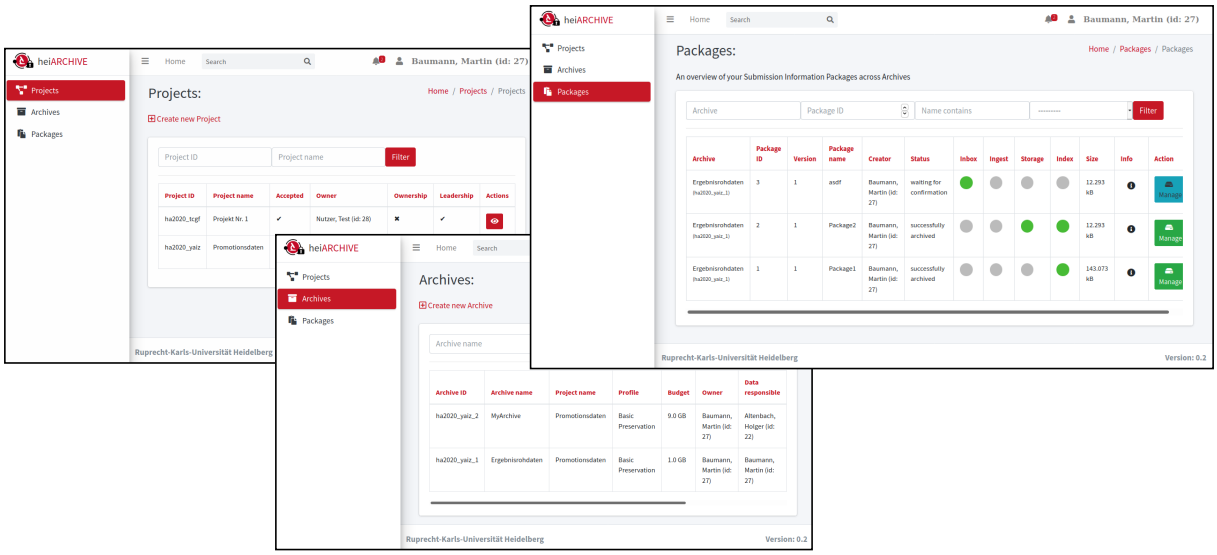


Figure 2: Exemplary views of heiARCHIVE's GUI: Users have an overview of their projects, archives and archive packages including the states of the archive packages within the archiving pipeline.

The archive is the container for one or multiple "archive packages" – the structure which finally contains the data to be archived. The data responsible person has the permission to upload data into an archive package, add metadata and start the packaging and storage process.

4 Metadata

For subsequent use of the preserved data and also for a description of the preservation process, a minimal set of mandatory metadata is stored in the heiARCHIVE database and the index, but also within the AIPs. Some metadata is demanded from the user, e.g. the creator of the data and additional descriptive information. Other metadata can be determined from the user's data to be archived (denoted by *payload*) itself. More extensive descriptive metadata may be stored in a pre-defined location within the data package that might be considered in the future by the indexer.

The metadata procedure during the ingest process is sketched in Fig. 3. Metadata is collected from the user via the GUI or via the API and is stored in the file"heiarchive-metadata.xml" complying a custom schema. Using a custom schema for this intermediate format has practical reasons, since it is easy to use and we may change at our sole discretion to meet new institutional conditions. During the ingest workflow, the data of this file is read into an SQLite database which is also the reporting target of several tools that analyze the structure and content of the payload. After modifications to certain items in the database, e.g. placeholder replacements and reference resolutions, all items are processed and put into place to forge piece by piece the final XML file which can be

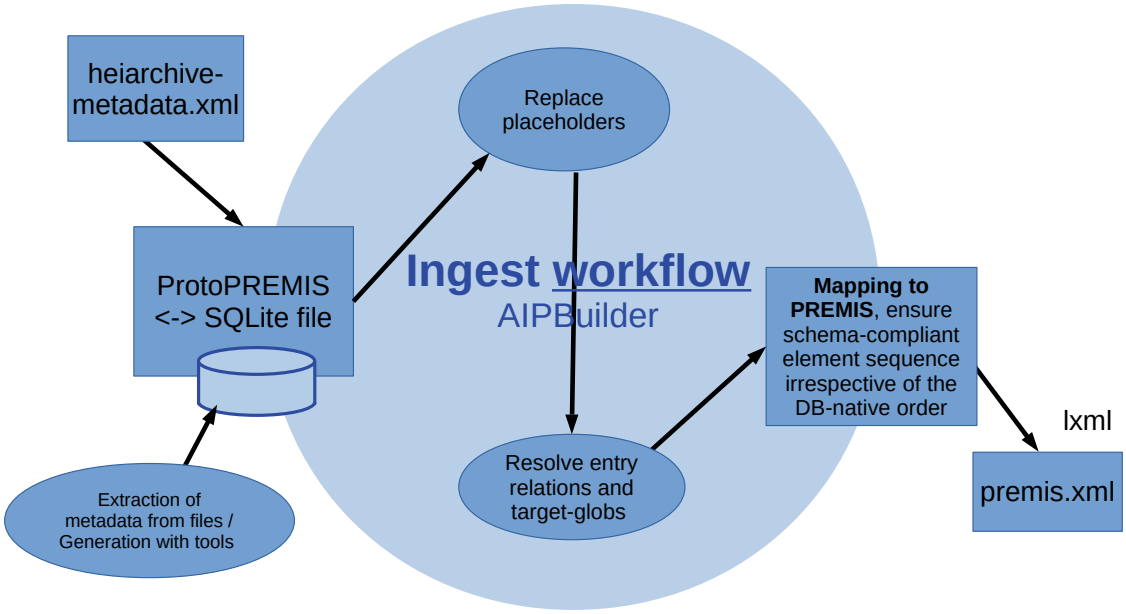


Figure 3: Process of metadata collection, its metadata processing towards a standardized schema and finally the creation of a standard conforming metadata file.

written in one single pass using the Python library lxml[10]. The SQLite database is used as a cache and collection facility for easy enrichment of the data without need to read in and write out large portions of XML by separate tools in rather resource-inefficient ways. The database file is stored within the AIP besides the"heiarchive-metadata.xml" to be on the safe side when detecting and resolving issues.

The XML file resulting at the end of the process complies different standards. The METS standard[11] defines a container for descriptive, administrative, and structural metadata. The PREMIS standard[12] defines the metadata for the preservation of the data objects and their long-term usability. And, most likely, DataCite[13] will be used to represent the descriptive metadata in the future (currently in investigation). DataCite is considered to be a suitable schema for descriptive metadata and - due to its wide distribution - facilitates inter-operation with other archive or repository services, also at University Library.

5 Status and next steps

The main features of the software behind heiARCHIVE are implemented, the submission process is running and the dissemination is technically prepared. Extensive testing has been done to ensure the GUI and backend functions are working reliably. Currently, the

[10] https://lxml.de/
[11] https://www.loc.gov/standards/mets/
[12] https://www.loc.gov/standards/premis/
[13] https://schema.datacite.org/

metadata model and the author's contract are not finally defined and the access of the available tape library via iRODS is under investigation. Geo-replication has not been realized yet but is in preparation. Although heiARCHIVE is a dark archive, a publicly accessible registry is planned that contains an excerpt of the metadata of the archived data together with a persistent identifier.

In the next few months, these tasks will be tackled and the service will afterwards be started step by step towards productive operation: First, selected researchers will be invited to do an archiving for research data that has no challenging requirements (e.g. not very large capacity). Then, the service will be opened for general use for Heidelberg University researchers. At the same time, a connection between services at University Library to heiARCHIVE will be realized via the API for an automated long-term preservation of data of these services.

Acknowledgements

The development of heiARCHIVE has been funded in parts by the ministry of science, research and arts of the state of Baden-Württemberg, Germany.

Bibliography

[1] CCSDS - Consultative Committee for Space Data Systems. "Reference Model for an Open Archival Information System (OAIS), Recommended Practice." Recommendation for Space Data System Practices, CCSDS 650.0-M-2, Magenta Book, June 2012.

Storage for Science – Aktueller Stand und anstehende Entwicklungen eines verteilten FDM-Systems

Dirk von Suchodoletz [*], Ulrich Hahn, Jonathan Bauer, Kolja Glogowski und Mark Seifert

[*] RZ Universität Freiburg

Der Speicherbedarf über die verschiedenen wissenschaftlichen Fach-Communities hinweg ist in den letzten Dekaden erheblich gestiegen. Neben der reinen Verarbeitung und Speicherung kommen zunehmend Anforderungen an Such-, Auffind- und Verfügbarkeit der Daten hinzu, die sich aus der „Guten wissenschaftlichen Praxis“ oder den FAIR-Prinzipien ergeben [1, 9]. Da sich große Speichersysteme mit Laufzeiten von mindestens zehn Jahren nicht mehr durch kleinere Arbeitsgruppen realisieren lassen, wurde für das Speichersystem bwSFS (Storage-for-Science) ein föderativer Ansatz gewählt und über einen gemeinsamen Antrag der Universitäten Freiburg und Tübingen umgesetzt. Die fachliche Zuordnung der wissenschaftlichen Arbeitsgruppen orientiert sich an den bwHPC-Communities und folgt den Konzepten aus bwDATA Phase III. Das Speichersystem bwSFS realisiert eine georedundant verteilte technische Plattform für Basis-Speicherdienste mit darauf aufsetzendem Forschungsdatenmanagement und erlaubt das Teilen wissenschaftlicher Daten über deren gesamten Datenlebenszyklus hinweg. Es ist damit sowohl ein zentraler Baustein für das Data Intensive Computing der BinAC- und NEMO-HPC-Communities [3, 4] und stellt gleichzeitig Kapazitäten und Dienste für Forschende ohne HPC-Bezug der beteiligten Universitäten, des Science Data Centers BioDATEN und des NFDI-Konsortiums DataPLANT bereit.

1 bwSFS – Hardwaregrundlage und Basisdienste

Die Zielgruppen und die zu integrierenden Forschungsinfrastrukturen umfassen primär die Arbeitsgruppen aus den Anträgen, sowie zusätzliche AGs aus den abgedeckten Fachbereichen und lokale AGs zur Grundversorgung mit Langzeitspeicherdiensten [5]. Dabei strebt bwSFS eine effiziente und langfristig gesicherte Ablage von Forschungsdaten an, die ergänzend und integrierend zu bereits bestehenden Repositorien der Fachwissenschaften operiert. Mit bwSFS werden die infrastrukturellen Ressourcen einzelner Fachwissenschaften für das Forschungsdatenmanagement gebündelt, um eine bessere Unterstützung in

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029643 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13741 (CC BY-SA 4.0)

der Umsetzung spezifischer FDM-Anforderungen zu erreichen. Die notwendige umfassende Beratung der Forschenden sollte über parallel laufende Aktivitäten im FDM der jeweiligen Einrichtung oder durch FDM-orientierte Projekte gewährleistet werden. Zur Umsetzung der FAIR-Anforderungen und OpenAccess-Prinzipien wird initial auf Datenmanagementpläne gesetzt, die durch Vorgaben der jeweiligen Fach-Communities mit Richtlinien für Metadatenmanagement, Archivierung und Lizenzmodelle unterstützt werden. Organisatorische Fragen auf Seiten der verschiedenen institutionellen Ebenen werden durch den von den Science Data Centern und dem Arbeitskreis FDM erstellten „Leitfragen zum verantwortungsvollen Umgang mit Forschungsdaten“ vorangetrieben. Im Folgenden wird der aktuelle Stand der aus [6], [7] und [8] weiterentwickelten Überlegungen sowie die noch anstehende Entwicklung eines verteilten FDM-Systems vorgestellt.

Die zentralen Speicherinstallationen befinden sich an den Standorten Tübingen und Freiburg, zusätzlich kommen Cache-Systeme an den Universitäten Konstanz und Stuttgart zum Einsatz. bwSFS stellt insgesamt knapp 20 Petabyte nutzbare Speicherkapazität auf Basis von NetApp-Komponenten der FAS- und StorageGrid Produktlinien bereit. Das System arbeitet mit Speicherplatzoptimierung durch Kompression und Deduplizierung, so dass insbesondere für noch in Verwendung befindliche, ungepackte Daten virtuell zusätzliche Kapazität zur Verfügung steht. Die Basis-Speicherdienste sind als Netzwerkdateisysteme NFS und SMB sowie als Objektspeicher (S3) ausgeführt. Die Filesysteme werden primär lokal an den Hauptstandorten oder via Caching-Komponente zusätzlich für Arbeitsgruppen in Stuttgart und Konstanz transparent lokal verfügbar gemacht. Ein Teil des Objektspeichers wird weltweit verfügbar sein, um insbesondere in verteilten Workflows und Kooperationen eingesetzt werden zu können. Eine Anbindung an die bwHPC-Systeme erfolgt mittels SFTP und S3. Zudem stehen verschiedene Überlegungen zum effizienten Datenaustausch innerhalb der Baden-Württembergischen Datenföderation im Rahmen von bwHPC-S5 an. Ein weiteres Ziel besteht im Angebot automatisierter Workflows für die Speicherverwaltung und Anbindung an Dienste von Fachwissenschaftens. Das System verfügt über eine solide Hardwarebasis mit moderner Überwachung und verschiedenen, teilweise über die Standortgrenzen hinweg reichenden Redundanzen in Form einer kompletten Spiegelung des Filesystem-Bereichs und Erasure Coding für den Objektspeicher. Die Installation ist auf Erweiterbarkeit an beiden Standorten angelegt; zusätzliche Komponenten wurden bereits für die de.NBI-Cloud am Standort Freiburg und das QBIC in Tübingen hinzugefügt.

2 Dienste für das Forschungsdatenmanagement

bwSFS wird eine Reihe von FDM-Diensten für einzelne Fachwissenschaften der Antragsteller, der beteiligten Projekte sowie Universitäten offerieren, die im Backend auf die Basisdienste aufsetzen. Zur Unterstützung von Datenpublikationen wird innerhalb von bwSFS InvenioRDM verwendet, welches ein komfortables Userinterface und die OAI-PMH-Schnittstelle bereits beinhaltet. In diese Entscheidung wurden frühzeitig alle am FDM-Prozess beteiligten zentralen Einrichtungen und Projekte einbezogen. In Tübingen

sind das die Universitätsbibliothek, die Core-Facility eScience-Center und das SDC BioDATEN. In Freiburg erfolgt die Koordination mit zentralen Einrichtungen und die Communities durch die Research Data Management Group. Für die DOI-Vergabe in Invenio wird auf etablierte Dienste der beteiligten Universitätsbibliotheken zurückgegriffen. Als Speicher-Backend wird von InvenioRDM das S3-Protokoll nativ unterstützt und erlaubt so die Anbindung an die Object-Storage-Infrastruktur von bwSFS. In der Implementierung macht sich Invenio die vollen Vorteile von S3 zu nutze und dient dabei als Broker. Das System händigt bei Datenübertragungen pre-signed URLs aus, um eine direkte Verbindung zwischen Clients und Objektspeicher zu erlauben. Auf diese Weise wird die hohe Verfügbarkeit und Performanz der zugrundeliegende bwSFS-Infrastruktur direkt genutzt. Die umfangreiche REST-API von InvenioRDM bietet weiterhin viele Möglichkeiten zur Integration in Drittsysteme. Einige Workflows konnten bereits evaluiert werden, etwa das automatische Erstellen von Publikationsentwürfen aus unterschiedlichen Systemen heraus. Dazu zählen Code-Versionierungsplattformen wie GitLab oder GitHub bei neuen Code-Releases, andere Compute-Umgebungen wie HPC oder Galaxy nach Beendigung von Jobs oder auch in SDC-Gateways über Veröffentlichungstemplates. Nutzer könnten diese Entwürfe dann auf der InvenioRDM-Webseite fertigstellen und, falls erwünscht, veröffentlichen. InvenioRDM etabliert keinen Ersatz bestehender Daten-Repositorien, sondern komplementiert fehlende Angebote und bietet Ergänzungen zu existierenden Systemen, um beispielsweise eine gesicherte, lokale Zweitkopie eines Datensatzes zu hinterlegen.

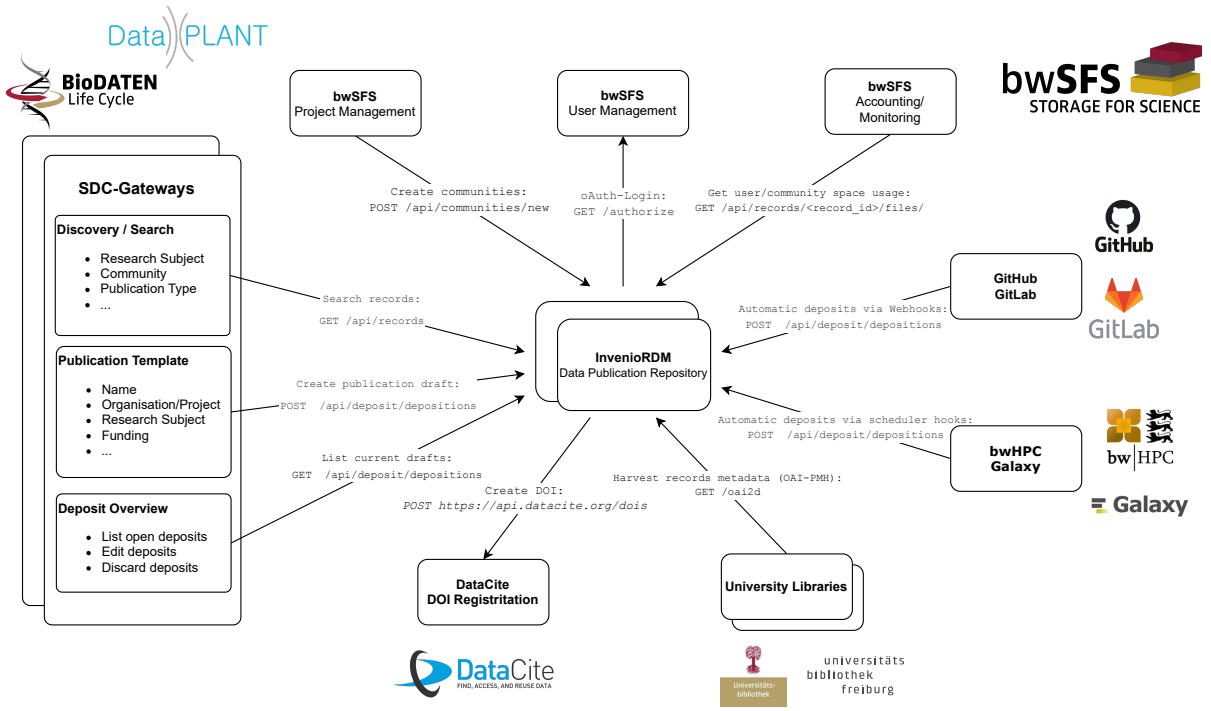


Abbildung 1: InvenioRDM als zentraler Bausteinfür das Publizieren und Teilen von Forschungsdaten.

In Freiburg wird eine GitLab-Instanz für Versionierung, Kollaboration und zum Teilen von Daten und Code laufender Projekte zum Einsatz kommen, ein Konzept welches bereits seit längerem durch die Research Data Alliance für viele Fachwissenschaften vorgeschlagen wird. Um dabei effizient mit großen Datenmengen umzugehen und diese möglichst direkt im Objektspeicher von bwSFS ablegen zu können, werden aktuell im Rahmen von BioDATEN und DataPLANT verschiedene Ansätze untersucht. Das Bilddatenmanagement-System OMERO bietet nutzerfreundliche Schnittstellen für den Zugriff, die Darstellung und die Arbeit mit Bilddaten aus der Mikroskopie. Je nach Bedarf können OMERO Instanzen als Gruppen-eigenes Repositorium z.B. auch für sensible Daten, als Kollaborationsplattform oder zur Bereitstellung öffentlicher Daten für Websites und Publikationen eingesetzt werden. OMERO legt einen starken Fokus auf den Erhalt und die Erweiterung der Metadaten. Mit OMERO erhalten Forschende der beteiligten Universitäten die Möglichkeit die stetig wachsende Menge an Mikroskopie-Bilddaten aufsetzend auf der bwSFS-Infrastruktur zu verwalten. Dazu wäre eine technische Basisinfrastruktur aus sicherer Speicherung im Filesystem und gehosteter OMERO-Instanz denkbar, die über automatisierte Deployments beispielsweise mit Kubernetes erzeugt werden kann. Das erlaubt mehrere parallel laufende Instanzen mit unterschiedlicher Konfiguration, die durch die jeweilige Arbeitsgruppe an ihre Bedürfnisse angepasst und durch diese selbst verwaltet wird. Dieses entlastet die Forschenden ebenso wie im Fall InvenioRDM und GitLab vom Betrieb eigener Basisinfrastrukturen und erlaubt ihnen den Fokus auf ihre fachlichen Belange zu richten.

Da nicht für alle Bedarfe der Antragstellenden auf bereits existierende Software zurückgreifen kann, sollen zudem eigene Dienste[1] der Fachwissenschaften gehostet werden können. Weiterhin ist eine sichere Langzeitspeicherung, abseits etablierter Repositorien, für eine DFG-konforme Ablage nicht veröffentlichter Daten mit S3-Backend vorgesehen. Die technische Grundlage der FDM-Dienste basiert auf einer "Hyper Converged Infrastructure"(HCI), die mit Kubernetes und Rancher orchestriert wird und die verschiedenen Micro-Services des Gesamtsystems betreibt. Für die Bereitstellung umfangreicherer Dienste wird je nach fachwissenschaftlicher Zuordnung auf Ressourcen der bwCloud oder der de.NBI-Cloud zurückgegriffen.

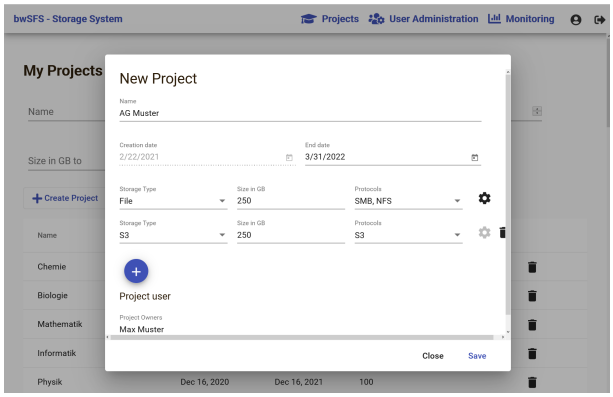
3 Nutzer- und Projektmanagement

Schon in der Implementierungsphase der Software und Dienste, die die Fachwissenschaften einbezieht, zeichnet sich ab, dass die klassischen Methoden des Identitätsmanagements nicht genügen. Im Vergleich zu HPC-Diensten erfordern Speicherdienste wegen ihrer vielgestaltigen Nutzerschaft und langen Haltefristen von Forschungsdaten eine wesentlich tiefere Integration bestehender Infrastrukturen und ein flexibleres Nutzermanagement. Um die vorgesehene Nutzerbasis des Systems von verschiedenen Standorten und aus den unterstützten Fachwissenschaften verwalten zu können und zukünftig eine nahtlose Inte-

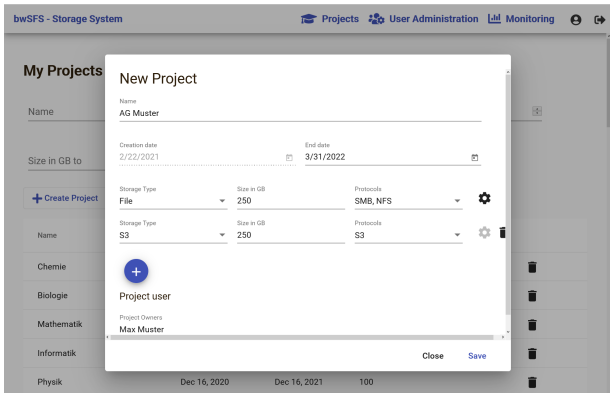
[1] Zum Beispiel die Oberflächendatenbank, https://contact.engineering/, gefördert im Rahmen des ERC StG-757343, `https://cordis.europa.eu/project/id/757343 undlivMatS`

gration in die Datenföderation zu erreichen, ist ein föderiertes Management der Projekt-, User- und Gruppendaten notwendig. Hierbei wird einerseits auf etablierte Strukturen aus dem HPC-Umfeld wie bwIDM aufgesetzt, in denen ein Teil der Nutzerbasis bereits beheimatet ist. Andererseits sind weitere Quellen für Benutzer-Identitäten beziehungsweise -informationen, wie Elixir-AAI oder zusätzliche Dienste wie ORCID vorgesehen. Hierzu erfolgt eine enge Kooperation mit dem geplanten bwIDM2-Projekt, welches die Weiterentwicklung der badenwürttembergischen Identitäts Föderation um Aspekte des FDM, wie beispielsweise Fragen zur langfristigen Nutzeridentifizierung, berücksichtigt.

Unter Einbeziehung der Anforderungen der Speichersysteme und FDM-High-Level-Dienste soll ein einrichtungsübergreifender persistenter Identifikator zum bwIDM Datensatz hinzugefügt werden. Hier folgt bwSFS der Empfehlung des AK FDM in Baden-Württemberg und setzt auf die langzeitstabile ORCID ID [9], welche eine Infrastruktur für die Wiedererkennung derselben Person unabhängig von ihrer aktuellen Heimateinrichtung erlaubt.



(a) UI 1



(b) UI 2

Abbildung 2: Web-GUI der eigens entwickeltes Nutzerinterface zur Projektverwaltung.

Zur Anlage, Verwaltung und Konfiguration von Speicherprojekten wird parallel zur Inbetriebnahme der Hardware eine Projekt-Management-Software für bwSFS entwickelt. Nach der Beantragung und Bewilligung eines Speicherprojekts kann hier ein berechtigter Administrator das Projekt anlegen, projektverantwortliche Personen und technische Projektadminstratoren festlegen und die dazugehörigen Speicherressourcen erstellen. Die

Software umfasst drei wesentliche Komponenten, die in einem Kubernetes-Cluster bereitgestellt werden. Als Schnittstelle zum NetApp-ONTAP-System werden Ansible-Playbooks zur Provisionierung der Speicherressourcen, derzeit NFS-/CIFS-Shares, später ebenfalls S3-Tenants/Buckets, geschrieben und durch die Integration in einer AWX-Instanz über REST-API angesteuert. Die Projekt- und Nutzerverwaltung werden als eigenständige Micro-Services mit dem Spring-Framework für die Java-Plattform realisiert; die anfallenden Daten werden in einer PostgreSQL Datenbank abgelegt. Das in Angular2 entwickelte Web-Frontend dient schließlich als graphische Schnittstelle zu den Micro-Services und ist momentan nur für berechtigte Administratoren des Systems zugänglich. Um die Basis-Funktionalität des Web-Frontends erweitern zu können, wird ein Plugin-Mechanismus entwickelt. Durch die Entwicklung eigener Plugins können weitere Speicherressourcen, wie beispielsweise der Zugriff auf FDM-Dienste, in der Web-Oberfläche angeboten werden und dabei auf die Schnittstellen der Projekt- und Nutzerverwaltung zurückgreifen. Darüber hinaus werden Monitoring und Accounting Informationen über eine im Kubernetes-Cluster betriebenen Grafana-Instanz bereit gestellt, um unter anderem eine graphische Übersicht über die Belegung aller Speicherressourcen innerhalb eines Speicherprojektes zu erhalten.

4 Fazit und Ausblick

Die Entwicklung der Konzepte für bwSFS, die Definition des Funktionsumfangs und die Ausschreibung des Systems werden nun mit der Inbetriebnahme der ersten Dienste auf einen produktiven Stand gebracht. Die Etablierung dieses Gesamtsystems über zwei Hauptstandorte und zwei Cache-Standorte und für verschiedene Fachwissenschaften stellt eine große Herausforderung dar. Dies gilt insbesondere weil es keine vorgefertigten kommerziellen Lösungen gibt, sondern eine Kombination aus einzelnen Komponenten auf Hard- und Softwareebene geschaffen werden musste. Alle im Rahmen von bwSFS entwickelten Softwarekomponenten werden unter eine Open Source Lizenz gestellt, um sowohl die Nachnutzung, als auch Erweiterung und Anpassung offen und transparent zu gestalten. Während die Basisdienste bereits verfügbar sind und von ersten Antragstellern genutzt werden, sind viele High-Level-Dienste noch in Entwicklung oder Erprobung. Auch die breite Unterstützung des Forschungsdatenmanagement stellt ein erhebliches Unterfangen dar. Die Kooperation aller beteiligter Einrichtungen und Forschenden und der übergreifende Austausch von Daten und Erfahrungen müssen angestrebt werden, da der Umfang der zu berücksichtigenden Aspekte von den einzelnen Einrichtung dauerhaft nicht zu stemmen ist. Im Erfolgsfall wird eine solche Infrastruktur die Basis, oder zumindest eine Blaupause, für größere und umfassendere Aktivitäten, wie die NFDI bilden können. Jedoch erschöpfen sich die Kosten des FDMs nicht in der Hardware [10]. Deshalb ist ein Begleitprogramm auf organisatorischer und technischer Ebene wie beispielsweise durch BioDATEN und DataPLANT notwendig, um die personellen Ressourcen für die notwendige dauerhafte Betreuung der Forschenden vorzuhalten.

Danksagungen

Wir danken der Deutsche Forschungsgemeinschaft DFG für die Unterstützung der Projekte bwSFS und DataPLANT. bwSFS wird durch die Deutsche Forschungsgemeinschaft DFG gefördert: GZ: INST 37/1046-1 FUGG, GZ: INST 37/1047-1 LAGG, GZ: INST 39/1099-1 FUGG, GZ: INST 39/1098-1 LAGG DataPLANT wird durch die Deutsche Forschungsgemeinschaft DFG gefördert: GZ: 670407 (NFDI 7/1) auf Basis der Bund-Länder-Vereinbarung zum Aufbau einer nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018 finanziert. Wir Danken dem Land Baden-Württemberg für die Unterstützung des Science Data Centers BioDATEN und bwSFS-Infrastruktur.

ORCID IDs

- Dirk von Suchodoletz https://orcid.org/0000-0002-4382-5104
- Ulrich Hahn https://orcid.org/0000-0003-4471-9263
- Jonathan Bauer https://orcid.org/0000-0002-5624-2055
- Kolja Glogowski https://orcid.org/0000-0002-1361-5712
- Mark Seifert https://orcid.org/0000-0002-1042-6107

Literaturverzeichnis

[1] Deutsche Forschungsgemeinschaft. *Sicherung guter wissenschaftlicher Praxis/Safeguarding Good Scientific Practice. Denkschrift/Memorandum.* Wiley Online Library, 2013. URL: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527679188`, `https://doi.org/10.1002/9783527679188`.

[2] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[3] Jens Krüger, Volker Lutz, Felix Bartusch, Werner Dilling, Anna Gorska, Christoph Schäfer, and Thomas Walter. Bioinformatics and Astrophysics Cluster (BinAC). In *Proceedings of the 3rd bwHPC-Symposium*, pages 91–95, 2017. URL: `https://books.ub.uni-heidelberg.de/heibooks/reader/download/308/308-4-79226-1-10-20171002.pdf`, `https://doi.org/10.11588/heibooks,308.418`.

[4] 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. URL: http://hdl.handle.net/10900/87655, https://doi.org/10.15496/publikation-29041.

[5] Rahmenkonzept der Hochschulen des Landes Baden-Württemberg für datenintensive Dienste – bwDATA Phase III (2020-2024), 2021. URL: https://ub01.uni-tuebingen.de/xmlui/bitstream/handle/10900/114548/bwDATA_III__1_001.pdf, https://doi.org/10.15496/publikation-55923.

[6] Dennis Wehrle, Bernd Wiebelt, and Dirk von Suchodoletz. Design eines FDM-fähigen Speichersystems. In *10. DFN-Forum Kommunikationstechnologien, 30.-31. Mai 2017, Berlin, Gesellschaft für Informatik eV (GI)*, pages 145–154, 2017. URL: https://dl.gi.de/bitstream/handle/20.500.12116/470/paper10.pdf.

[7] 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. URL: http://hdl.handle.net/10900/87672, https://doi.org/10.15496/publikation-29058.

[8] Felix Bartusch, Kolja Glogowski, Ulrich Hahn, Michael Janczyk, Steve Kaminski, Jens Krüger, Volker Lutz, Gerhard Schneider, Mark Seifert, Dirk von Suchodoletz, Thomas Walter, and Bernd Wiebelt. Defining the future scientific data flow for multi-disciplinary research data. In *E-Science-Tage 2019: Data to Knowledge*, pages 110–127. heiBOOKS, March 2020. URL: https://books.ub.uni-heidelberg.de/heibooks/reader/download/598/598-4-88224-1-10-20200325.pdf, https://doi.org/10.11588/heibooks.598.

[9] Dirk von Suchodoletz, Elisabeth Böker, Peter Brettschneider, and Franziska Rapp. Entwicklung in Baden-Württemberg: ORCID und ROR IDs als Standard für langfristige Personen-und Institutionen-Identifier. *Bausteine Forschungsdatenmanagement*, (2):80–88, 2020.

[10] Jan Leendertse and Dirk von Suchodoletz. Kosten und Aufwände von Forschungsdatenmanagement. *Bausteine Forschungsdatenmanagement*, (1):1–7, April 2020. URL: https://bausteine-fdm.de/article/view/8246, https://doi.org/10.17192/bfdm.2020.1.8246 .

iVA: Ein interaktiver Virtueller Assistent von BERD@BW zur Aufbereitung von Rechtsfragen im Bereich Open Science

Markus Herklotz[1] und Lars Oberländer[2]

[1]Universität Mannheim, Lehrstuhl für Statistik und Sozialwissenschaftliche Methodenlehre
[2]Universitätsbibliothek Mannheim

Da die Kenntnis des Datenschutzrechts häufig nicht zur Kernkompetenz von Forscherinnen und Forschern gehört und die Unsicherheit bezüglich des Themenfeldes den Open-Science-Bestrebungen entgegenstehen kann, hat es sich das Projekt BERD@BW (Business and Economic Research Data Center) zur Aufgabe gemacht, den mit Daten Arbeitenden einen Einstieg in die Thematik zu bieten. Im Rahmen von BERD@BW entwickeln wir einen interaktiven Virtuellen Assistenten (iVA), der dabei helfen soll, einen Zugang zu den relevanten rechtlichen Regelungen zu finden. In diesem Beitrag soll neben dem Entwicklungsprozess und der Funktionsweise auch im Mittelpunkt stehen, wie ein solcher Assistent durch seine individualisierte Informationsvermittlung einen Beitrag zur Förderung von Open Science und dem Austausch von (Forschungs-) Daten leisten kann.

1 Wozu einen Assistenten zur Aufbereitung von Rechtsfragen?

Der Austausch und die Weiternutzung von Forschungsdaten schließt ein wichtiges Themenfeld mit ein, das häufig nicht zu den Kernkompetenzen der Forschenden gehört: Datenschutz und Urheberrecht. Nicht erst seit der Etablierung der Datenschutzgrundverordnung (DSGVO) und der damit verstärkten Diskussion um die rechtlichen Aspekte zur Nutzung von Forschungsdaten müssen sich Wissenschaftler/innen damit beschäftigen, welche Daten sie unter welchen juristischen Voraussetzungen verarbeiten und bereitstellen können. Durch die rasanten Entwicklungen beispielsweise im Bereich Big Data werden darüber hinaus auch neue Fragestellungen des Datenschutzes aufgeworfen (vgl. [6], [5]). Unsicherheiten hinsichtlich der rechtlichen Grundlagen können auch dazu führen, dass Forschende von einer Veröffentlichung der Daten „sicherheitshalber“ absehen.

Dieses Problemfeld kann somit die Durchsetzung von FAIR-Prinzipien erschweren und den Open-Science-Bestrebungen entgegenstehen.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029644 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021.
Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13742 (CC BY-SA 4.0)

Die Nachfrage nach Informationen zum Datenschutz ist ungebrochen hoch, was beispielsweise vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) und dessen aktualisierten Handreichung bestätigt wird.[1] Dabei handelt es sich um eine hilfreiche Aufarbeitung der essentiellen datenschutzrechtlichen Themen insbesondere für Forschende aus den Sozial-, Verhaltens- und Wirtschaftswissenschaften (vgl. [1]). Es zeigt allerdings auch die Komplexität des Themenfeldes und verdeutlicht die Hürde, die es darstellen kann, sich als Nicht-Jurist/in in dieses umfangreiche Material einzuarbeiten.

Da das Projekt *Business and Economic Research Data Center (BERD@BW)* das Ziel verfolgt, ein Kompetenzzentrum für Datenverfügbarkeit, Datenaustausch und Datenanalyse in den Wirtschaftswissenschaften aufzubauen, nimmt die Bereitstellung eines Informationsangebots u.a. zu rechtlichen Fragestellungen darin einen hohen Stellenwert ein.[2] Um den mit den Daten Arbeitenden und Forschenden einen niedrigschwelligen Zugang zu den grundlegenden rechtlichen Regelungen zu ermöglichen, wird in BERD@BW ein *interaktiver virtueller Assistent (iVA)* entwickelt. Dieser soll den Forschenden als erster Einstieg dienen, wodurch eine Gesprächsgrundlage mit den zuständigen Datenschutzbeauftragten geschaffen und die tiefergehende Auseinandersetzung mit weiteren Materialien gefördert wird.

2 Entwicklung

Die Bereitstellung spezifischer Informationen bedarf zunächst eines Ausgangspunktes. Hierfür wurde die juristische Problematik auf eine anwendungsorientierte Fragestellung aus der Perspektive der Forschenden kondensiert: „Darf ich die Daten für mein Forschungsprojekt verarbeiten?“. Die Praxisnähe der Frage erleichtert den Forschenden die Auseinandersetzung mit dem juristischen Material und dient als Leitlinie bei der Auswahl der Informationen. Der hierauf aufbauende Entwicklungsprozess soll im Folgenden in den zwei Teilbereichen der juristischen Aufarbeitung sowie der didaktisch-technischen Implementierung der Information dargestellt werden.

2.1 Juristische Grundlagen

Relevant zur Beantwortung der Ausgangsfrage sind insbesondere die Normen aus den Bereichen Datenschutz und geistiges Eigentum, da diese grundsätzlich die Datenverarbeitungen zu Forschungszwecken beschränken können. Diese beiden Rechtskomplexe wirken weitgehend unabhängig voneinander und konnten somit in der Entwicklung von iVA problemlos getrennt werden. Als Reaktion auf einen hohen Bedarf an datenschutzrechtlichem Informationsmaterial aufgrund relativ neuer restriktiver Regelungen und der erheblichen praktischen Bedeutung steht im Rahmen des Projekts BERD@BW zunächst die datenschutzrechtliche Komponente im Fokus.

[1] https://www.konsortswd.de/aktuelles/pressemitteilungen/15072020/ [Letzer Zugriff: 28.04.2021]
[2] Für mehr Informationen zu BERD@BW s. [3]

Eine mit dem Datenschutzrecht verbundene Herausforderung ist dessen weitgehende Zersplitterung über zahlreiche Rechtsquellen unterschiedlicher Gesetzgeber. So finden sich neben expliziten Datenschutzgesetzen sowohl in der Gesetzgebung der Europäischen Union (DSGVO), des Bundes (BDSG) als auch der Länder (LDSG BW) einzelne Vorschriften in fachspezifischen Gesetzen (PolG BW, BStatG, TKG) und Leitlinien beratender Gremien (Artikel-29-Datenschutzgruppe).[3]

Um dieser Problematik zu begegnen, wurden dem Anwendungsvorrang des Unionsrechts entsprechend zunächst die Tatbestände der DSGVO als höchstrangiges Fachrecht hinsichtlich ihrer Anwendbarkeit auf die hier zugrundegelegte Fragestellung („Darf ich die Daten für mein Forschungsprojekt verarbeiten?“) sondiert (vgl. [7, vor Art. 1, Rn. 20 ff.]). Allerdings zeichnet sich die DSGVO durch die Verbindung ihrer direkten Geltung in den Mitgliedstaaten mit zahlreichen Öffnungsklauseln, welche den nationalen Gesetzgebern Spielraum für konkretisierende Regelungen einräumen, als ”hinkende Verordnungäus (vgl. [2, vor Art. 1, Rn. 88]). Dementsprechend wurden anhand der einschlägigen Öffnungsklauseln die anwendbaren Bundes- und Landesvorschriften ermittelt, welche ausgehend von ihrer praktischen Relevanz in das weitere Vorgehen miteinbezogen wurden. Die Einbeziehung von Fachgesetzen fand hierbei bewusst nicht statt, um eine Überfrachtung iVAs mit für die meisten Anwender/innen irrelevanten Informationen zu vermeiden.

Die anwendbaren Vorschriften wurden anhand ihrer Rechtsfolgen eingeteilt und entlang der Gesetzessystematik gegliedert. Aus dieser Gliederung ließ sich ein schematischer Prüfungsaufbau zur Beantwortung der Ausgangsfragestellung in den drei Hauptkomplexen Anwendungsbereich, Rechtsgrundlage und Verarbeitungsmodalitäten entwickeln. Diese so erreichte Struktur wurde nun mit den forschungsrelevanten Detailinformationen aus Gesetzestexten, untergesetzlichen Leitlinien und einschlägiger Literatur auf den unterschiedlichen Prüfungsebenen angereichert, um eine praxisorientierte Informationssammlung zu schaffen.

2.2 Didaktische & Technische Implementierung

Das so entstandene abstrakte Prüfungsgutachten stellt allerdings lediglich einen Zwischenschritt des Informationsangebotes dar. Durch seinen erheblichen Umfang und seine juristisches Prägung fehlt es dem Prüfgutachten an einem laienfreundlichen Zugang und in weiten Teilen einer Verständlichkeit für fachfremde Leser/innen. Entsprechend bedurfte es einer Darstellung der gesammelten Informationen in einer Form, die leicht zugänglich ist und anhand einer Vorauswahl zielgerichtete Informationen nach dem konkreten praktischen Bedarf ermöglicht.

[3] DSGVO=Datenschutzgrundverordnung, BDSG=Bundesdatenschutzgesetz, LDSG BW=Landesdatenschutzgesetz Baden-Württemberg, PolG BW=Polizeigesetz Baden-Württemberg, BStatG=Bundestatistikgesetz, TKG=Telekommunikationsgesetz

Um nicht mit den Vorschriften des Rechtsdienstleistungsgesetzes in Konflikt zu geraten und die Informationen einem möglichst großem Publikum zugänglich zu machen, gilt es allerdings, die Grenze zum Angebot einer einzelfallbezogenen Rechtsberatung einzuhalten.[4]

Aus dieser Gemengelage resultiert die Notwendigkeit einer zugänglicheren Form für die Darstellung der Prüfungssystematik und rechtlichen Informationen. Dieser Notwendigkeit wurde durch die Entwicklung eines Informationskonzepts begegnet, welches sich durch Interaktivität, praxisorientiert zielgerichtete Informationsbereitstellung und Anwender/innenbezogenheit auszeichnet. So wurde die Prüfungssystematik zunächst durch die Aufspaltung in einzelne Fragen mit binärer Antwortmöglichkeit in die Struktur eines Entscheidungsbaumes überführt (siehe Abb. 1).

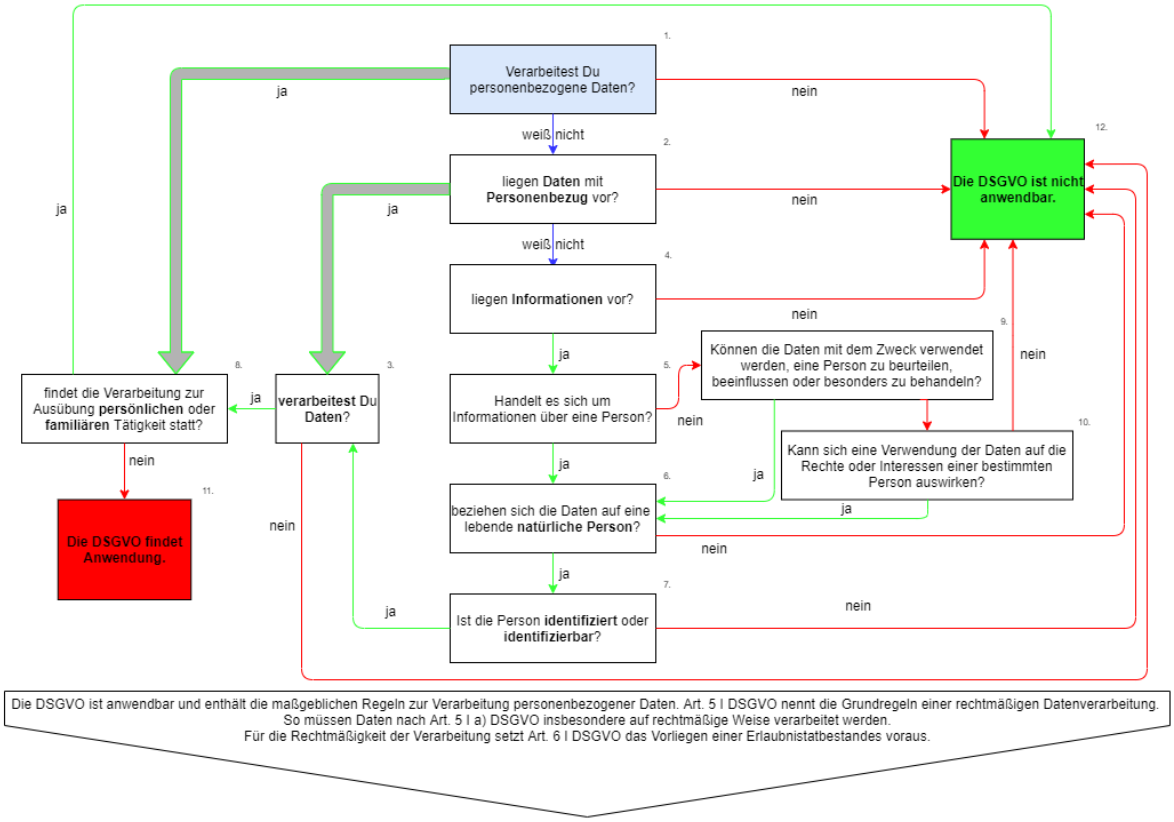


Abbildung 1: Ausschnitt des Entscheidungsbaums zum Anwendungsbereich der DSGVO (Stand: Januar 2021).

Im Anschluss an die Entwicklung und grafische Darstellung des Entscheidungsbaums wurden verschiedene Methoden in Betracht gezogen, um diesen online umzusetzen. Dabei stand neben einer intuitiven Bedienung im Vordergrund, die Online-Umsetzung interaktiv zu gestalten, um die Informationen bedarfspezifisch darzustellen zu können und die

[4] Um diese Abgrenzung auch den Anwender/innen zu verdeutlichen, weisen wir hierauf in der Anwendung explizit hin.

Nutzer/innen aktiv in die Rezeption einzubinden. Das Angebot soll die Anwender/innen ermutigen, ihre konkreten Datenprojekte zu reflektieren und damit den Lernerfolg erhöhen (vgl. *self-reference-effect*[5]). Dafür wurde zudem angestrebt, den Nutzer/innen nicht nur ein Ergebnis anzuzeigen (bspw. ob die DSGVO Anwendung findet oder nicht), sondern darüber hinaus abzubilden, wie dieses Ergebnis erreicht wurde. Von diesen Eckpunkten wurde sich erhofft, dass die Hürde bei der Auseinandersetzung mit dem Themenfeld des Datenschutzes genommen wird und zusätzlich zum ergebnisorientierten Aufbau auch der Lerneffekt verstärkt werden kann. Darüber hinaus soll das Objekt möglichst offen zugänglich und auf verschiedenen Plattformen implementierbar sein, um das Angebot möglichst vielen Interessierten verfügbar zu machen.

Diese didaktischen und technischen Anforderungen an das Lernobjekt konnten letztendlich mit dem Open-Source Content-Authoring-Tool Xerte[6] erfüllt werden. Xerte bietet u.a. einen Baukasten für Entscheidungsbäume, durch welchen sich die Inhalte technisch reibungslos und mit einer gleichzeitig großen Palette an Einstellungsmöglichkeiten umsetzen ließen. Die so entwickelte, erste Fassung des interactive Virtual Assistant (iVA) zum Anwendungsbereich der Datenschutzgrundverordnung konnte nach Fertigstellung per iFrame auf der BERD@BW-Website zur freien Verfügbarkeit eingebunden werden.[7]

3 Funktionsweise

iVA leitet die Nutzer/innen schrittweise mit konkreten Fragen durch den Entscheidungsbaum, wodurch die Anwender/innen ihre eigenen Datenprojekte outcome-orientiert reflektieren können. Durch den daraus entstehenden Workshop-Charakter können die Nutzer/innen herausfinden, ob ihr Projekt datenschutzrechtlich problematisch sein könnte. Die Funktionsweise erinnert an Fragebogen-Formate, womit die meisten Nutzer/innen vertraut sein dürften. Die einzelnen Seiten von iVA folgen einem konsistenten Aufbau (siehe Abb. 2) um den Anwender/innen die Navigation und die Rezeption der Informationen zu erleichtern.

Jeder Frage ist ein Einleitungstext vorangestellt, der auf die Frage mit Informationen vorbereitet. Daraufhin ist die eigentliche Frage mit den darunter befindlichen Antwortmöglichkeiten zu sehen. Zusätzlich werden den Nutzer/innen bei vielen Fragen unter der Schaltfläche "Mehr Informationen"Beispiele präsentiert oder weiterführende Hinweise gegeben. Parallel wird auf der linken Seite des Screens konstant eine Gliederung eingeblendet, wobei der auf der jeweiligen Seite geprüfte Punkt fett hervorgehoben wird. Dies soll den Anwender/innen die Orientierung im Prüfungsaufbau erleichtern und einen Gesamtüberblick ermöglichen.

Diese Fassung von iVA mündet in einen von zwei Ergebnisscreens: In diesem Fall ob auf Grundlage der Antworten die DSGVO Anwendung findet oder nicht. Dass konkret zu

[5] s. [4], [8], [9]
[6] www.xerte.org.uk [Letzter Zugriff: 14.05.2021]
[7] www.berd-bw.de/iva [Letzter Zugriff: 14.05.2021]

Abbildung 2: iVA Samplescreen (Stand: März 2021).

einem Ergebnis geführt wird, ist ein wichtiger Aspekt des avisierten praxisorientierten Workshop-Charakters. Gemeinsam mit dem Ergebnisscreen wird den Nutzer/innen ein Überblick über die gestellten Fragen und gegebenen Antworten angezeigt (siehe Abb. 3). Dies beinhaltet auch eine Version des Ergebnisscreens, die zum kopieren optimiert ist. Dadurch soll gefördert werden, dass die Anwender/innen sich tiefergehend mit dem Ergebnis auseinandersetzen und mit ihren zuständigen Datenschutzbeauftragten besprechen können.

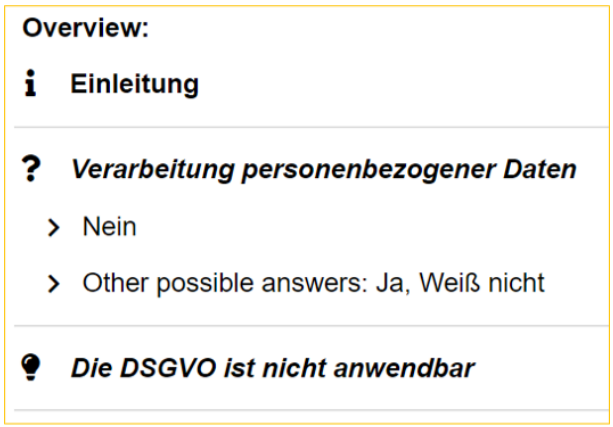


Abbildung 3: iVA Ergebnis-Überblick (Stand: März 2021).

4 Resümee & Ausblick

iVA wurde im Rahmen des Projekts BERD@BW an der Schnittstelle zwischen der Aufarbeitung juristischer Fragestellungen und dem Angebot von Aus- & Weiterbildung entwickelt und ist auf der Website des Kompetenzzentrums abrufbar. Durch seine einfache Verfügbarkeit und leichte Zugänglichkeit kann es auch problemlos auf anderen Websites eingebunden und so einer möglichst großen Zielgruppe verfügbar gemacht werden. Die flexibile Einbindung ermöglicht zudem, das Angebot je nach Bedarf mit weiteren Materialien zu ergänzen. iVA wird dabei als lebendiges Objekt gedacht, das durch Nutzer/innen-Erfahrung und neue Erkenntnisse erweitert und weiterentwickelt wird. Um dies sicherzustellen, wurde von dem aus Juristen und Bildungsforschern bestehenden iVA-Entwicklungsteam bereits im Entstehungsprozess projekt-extern Feedback von Anwender/innen und Datenschützer/innen eingeholt.

Das Prinzip iVAs, das in Hinsicht auf die identifizierten Anforderungen zur niedrigschwelligen Vermittlung von Rechtsinformationen entwickelt wurde, ist auch auf weitere Themenbereiche anwendbar. Das Repertoire soll im Projekt auf weitere Felder innerhalb und außerhalb des Datenschutzes erweitert werden. Dies bietet sich vor allem für die im Bereich Open Science relevanten Themen an, die häufig relativ disziplinunabhängige Kompetenzen im Forschungsdatenmanagement erfordern. Durch Online-Angebote wie iVA werden niedrigschwellige Einstiegsmöglichkeiten in diesen Themenkranz geschaffen, die flexibel innerhalb des geringen Zeitbudgets der Anwender/innen abgerufen werden können. Auf diese Weise kann iVA als Open Educational Resource einen Beitrag zum Aufbau von Rechtskompetenzen bei mit Daten Forschenden leisten, die für die Leitlinien von Open Science unabdingbar sind.

Anmerkungen

Das Projekt Business and Economic Research Data Center (BERD@BW) wird finanziert vom Ministerium für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg.

Literaturverzeichnis

[1] Matthias Bäcker and Sebastian Golla. *Handreichung Datenschutz*. Rat für Sozial- und Wirtschaftsdaten (RatSWD), 2nd edition, 2020.

[2] Eugen Ehmann and Martin Selmayr. *DS-GVO: Datenschutz-Grundverordnung: Kommentar*. Beck'sche Kurz-Kommentare. München: CHBeck : LexisNexis, München, Wien, 2nd edition, 2018. URL: `https://beck-online.beck.de/?vpath=bibdata/komm/EhmannSelmayrKoDSGVO_2/cont/EhmannSelmayrKoDSGVO.htm`.

[3] Sabine Gehrlein, Irene Schumm, and Renat Shigapov. BERD@BW – Ein Science Data Center für unstrukturierte Daten in den Wirtschafts- und Sozialwissenschaften. In *Tagungsband E-Science-Tage 2021: Share your Research Data*. Im Druck, 2021.

[4] Cynthia S. Symons and Blair T. Johnson. The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, 121(3):371–394, 1997. Publisher: American Psychological Association. https://doi.org/10.1037/0033-2909.121.3.371.

[5] Gerrit Hornung and Constantin Herfurth. Datenschutz bei Big Data Rechtliche und politische Implikationen. In Christian König, Jette Schröder, and Erich Wiegand, editors, *Big Data*, pages 149–183. Springer Fachmedien Wiesbaden, Wiesbaden, 2018. URL: http://link.springer.com/10.1007/978-3-658-20083-1_11, https://doi.org/10.1007/978-3-658-20083-1_11.

[6] J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, editors. *Privacy, Big Data, and the Public Good: Frameworks For Engagement*. Cambridge University Press, Cambridge, 1st edition, 2014.

[7] Boris P. Paal and Daniel A. Pauly. *Datenschutz-Grundverordnung: Bundesdatenschutzgesetz*. Beck'sche Kompakt-Kommentare. CHBeck, München, 3rd edition, 2021. URL: https://beck-online.beck.de/?vpath=bibdata/komm/PaalPaulyKoDSGVO_3/cont/PaalPaulyKoDSGVO.htm.

[8] Karen L. Hartlep and G. Alfred Forsyth. The Effect of Self-Reference on Learning and Retention. *Teaching of Psychology*, 27(4):269–271, October 2000. Publisher: SAGE Publications Inc. https://doi.org/10.1207/S15328023TOP2704_05.

[9] Ronald R. Schmeck and Scott T. Meier. Self-reference as a learning strategy and a learning style. *Human Learning: Journal of Practical Research & Applications*, 3(1):9–17, 1984. Place: US Publisher: John Wiley & Sons.

BERD@BW – A Science Data Center to foster Open Science in Business, Economics and Social Sciences

Sabine Gehrlein, Irene Schumm and Renat Shigapov

Mannheim University Library, University of Mannheim

The Center for Business, Economic and Related Data in Baden-Württemberg(BERD@BW) is one of the four science data centers funded by the Ministry of Science, Research and Arts of Baden-Württemberg within the digitization strategy "digital@bw". BERD@BW is aimed to improve sharing, finding and reusing unstructured and semi-structured research data in the social sciences in accordance with the FAIR principles (findable, accessible, interoperable and reusable). BERD@BW is built by the University of Mannheim and the Leibniz Center for European Economic Research (ZEW). Both institutions are experienced in infrastructure projects and in the empirical social sciences, including business and economics. BERD@BW is based on four pillars: 1) building up methodological knowledge, 2) developing tools and services dealing with unstructured and semi-structured data, 3) training and consulting with respect to legal and technical issues in research data management, and 4) engaging in national and international networking. The services and materials developed within BERD@BW are available as openly as possible on the project homepage: `https://www.berd-bw.de`.

1 Introduction

The social sciences, including business studies and economics, have a long tradition of handling structured research data in standardized mainly-tabular forms with proper metadata. These datasets often origin from public administration processes or surveys. In many cases they are sensitive and restricted in use. Infrastructure institutions (e.g., research data centers[1] and libraries) and commercial providers make the data available for academic purposes under clearly specified licenses and guaranteed authenticity. Data access and research methods with structured data are well established in both teaching and research.

Unstructured and semi-structured data, on the other hand, pose new challenges for data management and research. The licenses are harder to specify. The data authenticity

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: `https://doi.org/10.11588/heidok.00029645` veröffentlicht.

[1] organized by KonsortSWD, `https://www.konsortswd.de/en/ratswd/`, retrieved May 14, 2021.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13743 (CC BY-SA 4.0)

is trickier to guarantee. Unstructured data come in a variety of forms (pictures, text, video, and audio) from many non-standard sources (social media, web pages, and mobile phones) and in very large volumes ("big data"). To make the datasets findable, accessible, interoperable and reusable (FAIR),[2] new technologies have to be adapted and skilled staff is needed to implement suitable solutions.

To face these challenges, the Center for Business, Economic and Related Data in Baden Württemberg (BERD@BW in the following) was established by partners from infrastructure (Mannheim University Library, University-IT Mannheim, and research data center of ZEW) and research (Prof. Stahl and Prof. Kreuter from the Mannheim Center for Data Science, MCDS, and Dr. Licht from ZEW). BERD@BW is based on four pillars: 1) collecting methodological competence for modern data analysis, 2) developing the tools and services for collecting, analyzing and archiving unstructured and semi-structured data, 3) training and consulting on legal and technical aspects of research data management, and 4) networking. In the following, we describe these pillars and draw conclusions.

2 The Pillars of BERD@BW

Methodological competence. According to the publications in the top business and economic journals, unstructured data has become more and more important within the last ten years [1]. However, among the social scientists the data science and coding competencies needed to process and analyze unstructured and semi-structured data are not yet broadly spread. To improve those data science competencies among the target group of BERD@BW, knowledge about the methods and algorithms was collected and provided in an easily approachable form[3]. Machine learning, artificial intelligence and natural language processing are clearly explained without the use of too much technical jargon. A benchmarking system for algorithms dealing with unstructured data is currently under development.

Tools and services. Various tools and services are developed in BERD@BW to collect, process, archive and link semi-structured and unstructured data and to make the data available in compliance with the FAIR principles. A web-scraping engine "ARGUS"[4] is developed in order to collect unstructured data from non-standard sources. A corpus of websites of German companies is built using ARGUS [2-4] by scraping the web pages twice a year. The dynamically growing corpus is used as a prototype for archiving and providing unstructured data.[5] Social scientists are used to deal with sensitive data provided by research data centers. Access to sensitive data may be restricted due to privacy regulations and confidentiality obligations. A typical way to access the data is using a guest workstation on the premises of the respective research data center. Since the start of

[2] `https://www.go-fair.org/fair-principles/`, retrieved May 14, 2021.
[3] `https://digitaleconomy.org`, retrieved May 14, 2021.
[4] Available on GitHub
[5] The corpus „Mannheimer Webpanel" is provided by the research data center of the ZEW, see `https://kooperationen.zew.de/zew-fdz/datenangebot/mannheimer-webpanel.html`.

the Corona pandemic, physical access to many research data centers has been very limited or not possible at all. Therefore, the "BERD desktop" was developed for a secure remote access to sensitive data. The BERD desktop is a remote desktop server using a docker container and a virtual machine. To protect the sensitive data against unauthorized use, the configuration of the BERD desktop prohibits changes in system configuration, internet access, and uploads as well as downloads by the user. Data transfer into and out of the BERD desktop works only through the data administrator of the research data center. To enable users to run their analyses, the most popular tools for data analysis in the social sciences (i.e., R, Python and Stata) are provided. The first prototype of the BERD desktop in the bwCloud[6] is implemented. First tests of external users with data from the research data center of the ZEW are currently made.

Many datasets in the social sciences are still not findable, accessible, interoperable, and reusable. To improve this at least partially, a knowledge graph-based research data management infrastructure for German company datasets is developed. The Wikibase software was chosen as a backend for the infrastructural services. A test frontend in Python and JavaScript was created. Valuable historic datasets were digitized, OCR-ed (optical character recognition) and structured [5]. To speed up data integration and knowledge graph construction with Wikibase, the open source tool "RaiseWikibase" is implemented [6]. For automatic annotation services the open source semantic annotator "bbw" was designed and coded in Python [7]. The annotator performs named entity linking, property matching and type linking for tabular data without metadata using a Wikibase knowledge graph.

Training and consulting. Based on the collected knowledge and using the tools and services created, several training and consulting measures (both synchronous and asynchronous) have been initiated. The online micro workshop series "Data Literacy Snacks", which takes place during lunchtime for free, is offered as a part of synchronous training.[7] The events provide introductions into several topics of research data management tailored to the target group of data scientists and empirical researchers in business, economics and the social sciences. The first series comprises sessions about reproducible research, privacy regulations for research data and knowledge graphs in research data management. The number of registrations and participants exceeded expectations and personal feedback was very positive so far. As a part of asynchronous training and consulting the "interactive Virtual Assistant" (iVA) was developed [8] which guides users through different topics in research data management. For example, the first iVA implementation provides an interactive introduction into privacy law [8].

The iVA concept is currently under evaluation and will be rolled out to other topics relevant for the BERD community. The iVA implementations will be available as Open Educational Resources. Furthermore, a concept for an (a)synchronous online course "Good Practices for Managing Data" is developed in BERD@BW. While many discipline-specific courses in the area of data science and empirical research focus on methodology, this course

[6] `https://www.bw-cloud.org/`, retrieved May 14, 2021
[7] `https://www.berd-bw.de/snacks/`, retrieved May 14, 2021

specifically targets governance questions in research data management. The content of the course is mainly based on the Train-the-Trainer concept in research data management [9] and the concepts published by the UK Data Service [10].

Networking. A major aim of the funding program for the science data centers was fostering community-based networking and involvement into the National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur, NFDI).[8] The NFDI is a nationwide initiative of all German States and the Federal Government for building up and connecting discipline-specific research data networks in order to improve research data management within those disciplines, especially long-term storage, backup and accessibility.[9] The discipline-specific networks shall work together on cross-cutting topics and engage in international research data management initiatives.

One consortium covering the social sciences is KonsortSWD, which mainly focuses on sensitive structured data managed by the research data centers accredited by the RatSWD. Based on BERD@BW, the complementing consortium BERD@NFDI[10] was formed and organized with the main focus on unstructured data. More highly-recognized infrastructure institutions joined BERD@NFDI: ZBW (Leibniz Information Centre for Economics) and GESIS (Leibniz Institute for the Social Sciences). They are the most relevant providers of research data and information infrastructure in business, economics and the social sciences in Germany.[11] IT resources (storage and computing power) are provided by the LRZ (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities), which is world-famous for its high-performance computing resources. In addition, more well-known research institutions dealing with unstructured data in business, economics and the social sciences joined the consortium. Among them are the Universität Hamburg, the University of Cologne, and the Ludwig-Maximilians-Universität München.

It is worth to mention that the co-spokesperson Prof. Bernd Bischl is a co-founder of OpenML[12], which could become a groundwork for a FAIR machine learning and data analysis platform within BERD@NFDI. The partners of BERD@NFDI are important players in international and subject-specific research data networks such as GoFAIR, the European Open Science Cloud, and the Coleridge Initiative. Therefore, the connections to these initiatives are established as well.

[8] `https://mwk.baden-wuerttemberg.de/de/service/presse-und-oeffentlichkeitsarbeit/pressemitteilung/pid/vier-science-data-centers-in-baden-wuerttemberg/`, retrieved May 14, 2021.

[9] `https://www.dfg.de/en/research_funding/programmes/nfdi/index.html`, retrieved May 14, 2021.

[10] `https://www.berd-nfdi.de/`, retrieved May 14, 2021.

[11] ZBW and GESIS run the DOI registration agency for Social Sciences research data da|ra, for example, `https://www.da-ra.de/`, retrieved May 14, 2021.

[12] `https://www.openml.org/`, retrieved May 14, 2021.

3 Conclusions

In this paper, we sketched the foundations of the science data center BERD@BW. The contributions to FAIR and open research data management in business, economics and related domains are described. Apart from collecting and processing the subject-specific data science knowledge, the tools and services, which improve collecting, storing, reusing, and providing unstructured and semi-structured data, are presented. The training and consulting concepts, developed to improve research data handling, are introduced. The sustainable development of BERD@BW is guaranteed through embedding it into the new national consortium BERD@NFDI (https://www.berd-nfdi.de). An extended international networking is expected within BERD@NFDI.

Acknowledgements

The Business and Economic Research Data Center (BERD@BW) is funded within the digitization strategy "digital@bw" by the Ministry of Science, Research and Arts of Baden-Württemberg, Germany.

Bibliography

[1] Bayerl, A., Kluge, S., Beichert, M., Stahl, F. Methods and applications of Data Science in the context of Business & Economics. Available at DigitalEconomy.org (2020), https://tinyurl.com/3vjmx9j5.

[2] Kinne, J. and D. Lenz. Predicting Innovative Firms Using Web Mining and Deep Learning, ZEW Discussion Paper No. 19-001 (2019), Mannheim, https://doi.org/10.1371/journal.pone.0249071.

[3] Krüger, M., Kinne, J., Lenz, D., Resch, B.:The Digital Layer: How Innovative Firms Relate on the Web (2020). ZEW - Centre for European Economic Research Discussion Paper No. 20-003, Available at SSRN: https://ssrn.com/abstract=3530807.

[4] Kinne, J., and Axenbeck, J. Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics 125, 2011–2041 (2020). https://doi.org/10.1007/s11192-020-03726-9.

[5] Gehrlein, S., Kamlah, J., Pintsch, M., Schumm, I., Weil, S.: Vom Papier zur Datenanalyse. "Neue" historische Forschungsdaten für die Wirtschaftswissenschaften. In: Heuveline, V. (ed.) E-Science-Tage 2019 : Data to Knowledge. vol. 598, pp. 140-152. heiBOOKS (2020), https://doi.org/10.11588/heibooks.598.c8423.

[6] Shigapov, R., Mechnich, J., Schumm, I. RaiseWikibase: Fast inserts into the BERD instance. ESWC2021 Poster and Demo Track (2021).

[7] Shigapov, R., Zumstein, P., Kamlah, J., Oberländer, L., Mechnich, J., Schumm, I. bbw: Matching CSV to Wikidata via Meta-lookup. In: SemTab at ISWC 2020. vol. 2775, pp. 17-26 (2020), http://ceur-ws.org/Vol-2775/paper2.pdf.

[8] Herklotz, M., and Oberländer, L. iVA: Ein interaktiver Virtueller Assistent von BERD@BW zur Aufbereitung von Rechtsfragen im Bereich Open Science. E-Science-Tage 2021: Share your research data (2021).

[9] Biernacka, K., Bierwirth, M., Buchholz, P., Dolzycka, D., Helbig, K., Neumann, J., Odebrecht, C., Wiljes, C., Wuttke, U. Train-the-Trainer Concept on Research Data Management (Version 3.0). Zenodo (2020), http://doi.org/10.5281/zenodo.4071471.

[10] Corti, L., Van den Eynden, V., Bishop, L. and M. Woollard: Managing and Sharing Research Data – A Guide to Good Practice. SAGE (2020), http://hdl.handle.net/11329/297.

Automatisiertes Deployment von Elektronischen Laborbüchern mit Ansible

Henning Timm[1], Anne Wittkamp[2] und Stefan Beyer[3]

[1]Research Data Services, Universitätsbibliothek, Universität Duisburg-Essen, Deutschland
[2]Stabsstelle eScience, Zentrum für Informations- und Mediendienste, Universität Duisburg-Essen, Deutschland
[3]Geschäftsbereich IT-Infrastruktur, Zentrum für Informations- und Mediendienste, Universität Duisburg-Essen, Deutschland

Die spezifischen Anforderungen verschiedener Disziplinen an Elektronische Laborbücher (ELB) lassen sich nicht mit einer einzelnen Software erfüllen. Auch ist die Verwaltung einer zentralen, institutionellen ELB-Instanz für viele Anwendungsfälle zu unflexibel. Die Installation und der Betrieb vieler paralleler Instanzen erzeugen jedoch einen erheblichen Mehraufwand für die betreibenden Institutionen. Dieser Mehraufwand kann durch die Verwendung von automatisiertem Deployment mit Hilfe einer Virtualisierungsumgebung und einer Konfigurations- und Automatisierungssoftware, wie z. B. Ansible reduziert werden.

1 Betriebskonzepte für Elektronische Laborbücher

Die Dokumentation von Forschungsprozessen in der Laborforschung erfolgt bislang zum überwiegenden Teil analog in papierbasierten Laborkladden oder in einer Mischform von papierbasierten Laborkladden („Laborjournal“) und digitaler Datenerfassung. Elektronische Laborbücher bilden diese papierbasierten Laborkladden in Form von Webanwendungen ab. Neben der schlichten Möglichkeit die Dokumentation der Versuche direkt in digitaler Form zu erzeugen, bieten diese ELB viele weitere Features. Im Sinne eines digitalen Forschungsdatenmanagements erleichtert die Nutzung eines ELB die direkte Verknüpfung von Rohdaten mit zusammenhängenden Metadaten, Workflows und prozessierten und analysierten Daten und macht Forschungsprozesse sowie deren Ergebnisse besser nachvollziehbar und nachnutzbar. Die Gefahr eines Informationsverlustes oder fehlerhafter Ergebnisse, zum Beispiel wenn durch Zwischenschritte eines Workflows auf nicht aktuelle Daten zurückgegriffen wird, kann so vermieden werden. Ebenso können Daten zwischen Kooperationspartner:innen meist in nachhaltiger Weise ausgetauscht und transferiert werden.

Letzteres ist durch die Implementierung als Webanwendung möglich, wodurch die Inhalte des ELB direkt zentral abgelegt werden und über das Setzen von Berechtigungen mit anderen Nutzern geteilt werden können.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13744 (CC BY-SA 4.0)

Im Moment ist der Markt verfügbarer ELB-Lösungen sehr umfangreich [1]. Die speziellen Anforderungen verschiedener Disziplinen an ELB – etwa in Bezug auf die Anbindung von Datenbanken, technischen Geräten aus den Laboren oder besondere Darstellungen z. B. von Molekülen – gehen weit auseinander. Dadurch reichen die ELB-Lösungen von generischen hin zu fachspezifischen Programmen, die wiederum kommerziell oder als Open Source-Lösung zur Verfügung stehen. Bei der Einführung eines ELB am Standort müssen daher die Bedarfe der Forschenden an die Software gegen die technische und personelle Umsetzbarkeit der betreibenden Einrichtungen abgewogen werden. Die Erfahrung der Autor:innen zeigt, dass eine zentrale ELB-Instanz meist nicht ausreichend ist, um die fachspezifischen Bedarfe hinreichend abzudecken. Resultierend werden derzeit (auch an ein und demselben Standort) verschiedene ELB parallel eingeführt.

Der nachhaltige, zentrale Betrieb von mehreren heterogenen ELB-Instanzen stellt lokale FDM-Infrastruktur vor große Herausforderungen, nicht nur bei der Frage nach Speichermöglichkeiten und dem Identity Management.

Jede neue ELB-Instanz erzeugt Aufwände zur initialen Inbetriebnahme, wie die für Installation und Konfiguration benötigte Zeit, und bindet Personal für regelmäßige Wartung und Updates. Die Erfahrung der Autor:innen aus der Inbetriebnahme von vier ELBs zeigt, dass die Installation eines ELB auf einer Virtuellen Maschine (VM) ca. einen Personentag benötigt. Einige Arbeitsschritte, wie die Vergabe eines SSL-Zertifikates, hängen dabei von externen Stellen ab. Außerdem steigt durch ein breites, passgenaues Angebot an ELBs der Wartungsaufwand erheblich. Jede Instanz muss unabhängig Sicherheitsupdates erhalten und neue Versionen der ELB-Software müssen installiert werden.

Unabhängig von der konkret verwendeten Software weisen die meisten ELBs die Struktur einer typischen Webanwendung auf: Ein für das ELB spezifisches Frontend wird über einen Webserver zur Verfügung gestellt und interagiert mit einer Datenbank. Durch diese Struktur ähneln sich bestimmte Installationsschritte von ELBs, was bei der Automatisierung der Installation ausgenutzt werden kann. Weiterhin werden ELBs häufig mit Hilfe einer Virtualisierungsumgebung auf VMs installiert, die auf Hardwareanforderungen des konkreten ELBs angepasst werden können.

2 Automatisierung mit Ansible

Durch ein automatisches Deployment von ELB-Instanzen mittels einer Konfigurationssoftware wie Ansible[2] kann der Arbeitsaufwand erheblich verringert werden [3].[1] Automatisiertes Deployment bezeichnet hierbei die Installation und Konfiguration der benötigten ELB-Software auf einer bestehenden VM. Mit Hilfe von Ansible werden die dafür benötigten Schritte durch sogenannte Ansible-Playbooks zentral gesteuert und abstrahiert für ein sogenanntes Inventory von VMs durchgeführt: Ausgehend von einem Ansible-Server, auf dem die Playbooks hinterlegt sind, greift ein spezieller Ansible-Nutzer automatisiert

[1] Es existieren weitere Softwarelösungen, wie z. B. Saltstack oder Puppet, die ebenfalls für diese Aufgabe genutzt werden könnten.

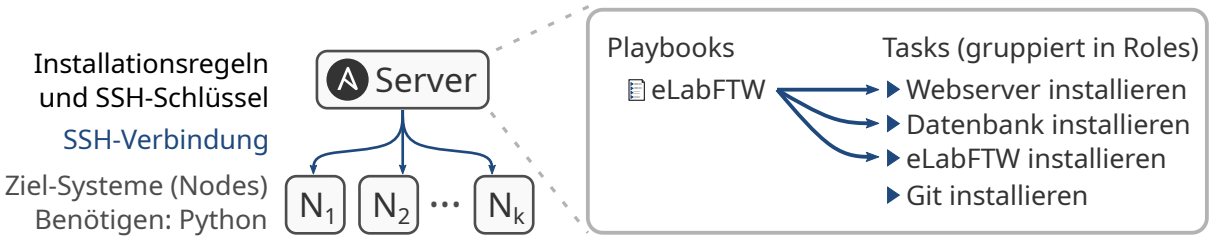


Abbildung 1: Beispielhafte Struktur eines Ansible-Servers mit k angeschlossenen VMs $N_1, .., N_k$ und einem Playbook zur Installation von eLabFTW. Die Roles (z. B. Webserver Installieren) beinhalten jeweils mehrere Tasks (Installation der Software, Konfiguration, ...).

per SSH auf die angeschlossenen Systeme zu und führt dort die für die jeweilige VM vorgesehenen Aufgaben durch. Die Zielsysteme werden vom Ansible-Server in einem oder mehreren Inventories verwaltet.

Playbooks werden in der weit verbreiteten Markup-Sprache YAML definiert, was eine leichte Nachnutzung ermöglicht. Sie enthalten Tasks, einzelne Installations- und Konfigurationsschritte wie z. B. die Installation einer Software mit Hilfe der Paketverwaltung, die in Roles für bestimmte Aufgaben zusammengefasst sind. Abbildung 1 zeigt beispielhaft die Struktur eines Playbook für die Installation des ELBs eLabFTW [4].

Installationen und sichere Konfigurationen (z. B. gehärtete Webserverkonfiguration, Firewalleinstellungen) müssen so nur einmal an zentraler Stelle erstellt, geprüft und aktualisiert werden. Soll eine neue Instanz eines bereits verwendeten ELBs in Betrieb genommen werden kann das bestehenden Playbook verwendet werden und es müssen nur Anpassungen, wie z. B. Passwörter, Namen und IP-Adressen, für die neue VM in das Inventory eingepflegt werden.

Im Gegensatz zu anderen Deploymentlösungen, wie z. B. dem Klonen eines Golden Image bei Erstellung einer VM, ermöglicht Ansible auch eine automatisieren Wartung der Systeme. So können Backups, System- und Softwareupdates ebenfalls zentral für alle angeschlossenen VMs ausgelöst werden. Da die angeschlossenen VMs mit dem selben Playbook installiert wurden und daher eine weitestgehend identische Konfiguration aufweisen reduziert sich der Wartungsaufwand für das betreibende Personal. Durch die zentrale Steuerung von Updates über den Ansible-Server werden ebenfalls die Einarbeitungszeiten in die jeweilige Software reduziert, was den Prozess sowohl beschleunigt als auch gegen Flüchtigkeitsfehler absichert. Zudem fungieren die gesammelten Playbooks als Dokumentation des Installationsprozesses und als zentraler Ort, an dem Besonderheiten der verwendeten Software notiert werden können.

Die gewonnene Zeitersparnis wird allerdings erkauft durch einen höheren Vorbereitungsaufwand, namentlich die Entwicklung (und Wartung) der Playbooks. Da die Entwicklung eines Playbooks nur einmalig anfällt und dessen Wartung für alle assoziierten Instanzen nur an einer Stelle (eben im Playbook) durchgeführt werden muss, amortisieren sich dieser Aufwand. Dieser Effekt wird noch verstärkt, wenn neu entwickelte Playbooks auf

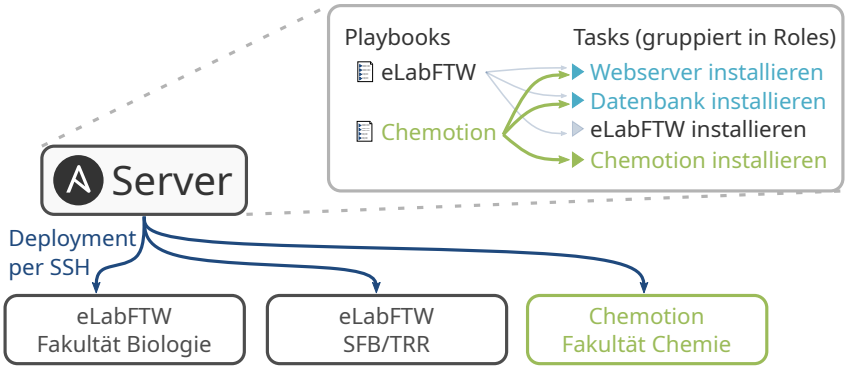


Abbildung 2: Fallbeispiel für die Entwicklung eines neuen Playbooks für die Software Chemotion, aufbauend auf einem existierenden Playbook für eLabFTW. In diesem Fall soll Chemotion mit der selben Webserver- und Datenbankkonfiguration wie eLabFTW betrieben werden, so dass die entsprechenden Roles nachgenutzt werden können.

bereits bestehende Roles und Tasks zurückgreifen können. Initial hat die Entwicklung des eLabFTW-Playbooks ca. 20 Stunden in Anspruch genommen und momentan werden neun Instanzen von eLabFTW mit diesem System betrieben. Die größte Zeitersparnis entsteht jedoch in Betrieb und Wartung durch die Möglichkeit Updates zentral ausgelöst, parallel und automatisiert für alle Instanzen durchzuführen. Der Wartungsaufwand, welcher vorher ca. 1 Stunde pro Monat pro Instanz betrug, konnte so auf 1 bis 2 Stunden pro Monat für alle neun Instanzen reduziert werden.

Bestehende Playbooks können aber auch als Grundlage für weitere Applikationen dienen. Bei ELBs die der oben beschriebenen Struktur von Frontend, Webserver und Datenbank folgen, aber auch bei verwandter Software können entwickelte Regeln für neue Playbooks nachgenutzt werden. Die Entwicklung neuer Playbooks (s. u.) benötigt aktuell nur noch 3 bis 6 Stunden, da große Teile existierender Playbooks nachgenutzt werden können.

In dem in Abbildung 2 dargestellten Fallbeispiel soll zu zwei bestehenden Instanzen von eLabFTW ein mit der Software Chemotion[5] realisiertes ELB hinzugefügt werden. Ein für Chemotion neu entwickeltes Playbook kann, abhängig von der angestrebten Konfiguration, z. B. die für eLabFTW entwickelten Roles für Webserver und Datenbank nachnutzen. Damit reduziert sich die Entwicklung des Playbooks auf die für Chemotion spezifischen Schritte.

3 Inter-Institutionelle Kollaboration

Eine Stärke der Open Source Software Ansible, im Gegensatz zu vergleichbarer Softwarealternativen, liegt in der geringen Einstiegshürde für den Betrieb des Servers, sowie für Entwicklung und Austausch von Playbooks. Playbooks können unabhängig von Inventories (die sicherheitskritische und interne Informationen beinhalten) leicht über

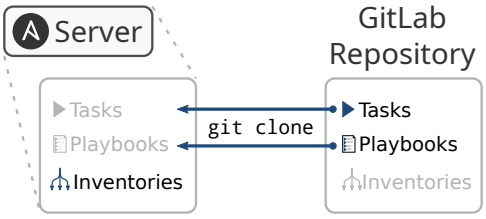


Abbildung 3: Beispiel für die Arbeit mit kollaborativ entwickelten Playbooks. Öffentliche Informationen, wie die verwendeten Playbooks, können über eine Plattform wie GitLab geteilt werden. Interne und sicherheitskritische Anpassungen, wie z. B. Inventories oder Anpassungen für bestimmte Systeme, können von der Institution intern verwaltet werden.

Software-Plattformen wie z. B. eine institutionelle GitLab-Instanz verwaltet und ausgetauscht werden. Auf dem Ansible-Server werden diese dann via Git synchronisiert und mit den Institutionsspezifischen Inventories verknüpft. Dabei muss besonders beachtet werden, dass das Playbook keine internen Informationen, wie z. B. Zugangsdaten, preis gibt. Im Fall der Autor:innen wurde das Playbook in Workshop-Form (z. T. mit externen Gästen) auf einem Testsystem entwickelt, so dass diese Abgrenzung von Beginn an eingehalten wurde. Das entwickelte eLabFTW-Playbook ist auf GitHub veröffentlicht (`https://github.com/RDS-UDE/eLabFTW-ansible-playbook`).

Dieses in Abbildung 3 illustrierte Vorgehen erlaubt die kollaborative Entwicklung im Austausch mit anderen Institutionen, ermöglicht aber ebenfalls die Verwaltung interner Komponenten. Soll z. B. eine neue FDM-Software getestet oder eingesetzt werden, für die bereits ein Playbook an einer anderen Institution entwickelt wurde, so kann dies den Aufwand der Inbetriebnahme deutlich reduzieren. In diesem Fall können dieses Playbook und die assoziierten Roles überprüft und (z. B. via Git) in die bestehende Ansible-Architektur eingebunden werden, so dass nur das Inventory angepasst werden muss. Durch eine gemeinsame Entwicklung mit Kooperationspartner:innen anderer Institutionen können so zum einen einheitliche Konfigurationen entwickelt werden, zum anderen erhöht die mit einer Nachnutzung einhergehende Überprüfung die Qualität des entwickelten Codes.

4 Fazit

Mit Hilfe von Ansible kann der erhöhte Installations- und Wartungsaufwand, der mit dem Betrieb individueller ELBs einher geht, reduziert werden. Durch ihre Struktur als Webanwendung sind ELBs besonders geeignet für dieses Vorgehen, da ein großer Teil der benötigten Installationsschritte Potential zur Nachnutzung in anderen Playbooks aufweist. Zusätzlich wird durch die Erstellung und Wartung von Playbooks eine lebendige Dokumentation der betriebenen Systeme geschaffen. Ein späterer Umstieg auf Automatisierungstechnologien mit anderem Fokus (z. B. Docker und Kubernetes) bleibt durch die Struktur von Ansible-Projekten ebenfalls möglich.

Die Verwendung einer Open Source Software mit niedriger Einstiegshürde ermöglicht die Nachnutzung von Tasks und Playbooks, sowohl intern – zur Entwicklung neuer Playbooks – als auch in Kooperation und Kollaboration mit externen Institutionen (z. B. organisiert über GitLab).

Acknowledgements

Die Rechte an Logos gehören den jeweiligen Rechteinhabers: Das verwendete Ansible-Logo ist abgeleitet von der Datei Ansible_logo.svg, erstellt durch den Wikimedia Commons Nutzer Vulphere, welche gemeinfrei zur Verfügung steht.

Literaturverzeichnis

[1] ZB MED (Hrsg.). 2020. *Elektronische Laborbücher im Kontext von Forschungsdatenmanagement und guter wissenschaftlicher Praxis – ein Wegweiser für die Lebenswissenschaften*, 2. aktualisierte und erweiterte Fassung, Köln, doi: `https://doi.org/10.4126/FRL01-006415715`.

[2] Red Hat, Inc. 2021. „Ansible is Simple IT Automation“. Abgerufen 14. Mai 2021 von: `https://www.ansible.com/`.

[3] Masek, Pavel, Martin Stusek, Jan Krejci, Krystof Zeman, Jiri Pokorny und Marek Kudlacek. 2018. „Unleashing Full Potential of Ansible Framework: University Labs Administration“. 22nd Conference of Open Innovations Association (FRUCT), 144-150, doi: `https://doi.org/10.23919/FRUCT.2018.8468270`.

[4] CARPi, Nicolas, Alexander Minges und Matthieu Piel. 2017. „eLabFTW: An open source laboratory notebook for research labs“. Journal of Open Source Software, 2(12), 146, doi: `https://doi.org/10.21105/joss.00146`.

[5] Tremouilhac, Pierre, An Nguyen, Yu-Chieh Huang, Serhii Kotov, Dominic Sebastian Lütjohann, Florian Hübsch, Nicole Jung und Stefan Bräse. 2017. „Chemotion ELN: an Open Source electronic lab note-book for chemists in academia“. J Cheminform 9, 54, doi: `https://doi.org/10.1186/s13321-017-0240-0`.

BIRD: Using Conversational User Interfaces to Provide Relevant Metadata for Interdisciplinary Research Data Publishing

André Langer, Lukas Schmolke and Martin Gaedke

Professorship for Distributed and Self-organizing Systems,
Chemnitz University of Technology, Germany

By the digitization of science, publishing research data to the scientific community is increasingly demanded in order to allow the replay, reuse or repurpose of existing research results. It is often necessary to describe the original research data files to increase its findability according the the FAIR principles for good scientific practice. So far, this is typically done by providing an additional descriptive floating text or by stating specific information in a submission form of a research data repository. Alternatively, structured machine-readable metadata description files can directly be provided. However, the meta description result quality depends on the experience of the user, commonly focuses on general aspects, and tool assistance is often limited. To address these issues, we applied as an alternative a conversational user interface paradigm to the description of research data and present the BIRD prototype (Bot-based Interface for Research Descriptions). It realizes a chat dialog which will assist scientists in providing an appropriate, structured metadata description based on the OpenAIRE Guidelines for Data Archives, independent of a particular technical repository platform. The tool is offered as a demonstrator for public access.

1 Introduction

When publishing scientific artifacts, such as recorded files from an experiment, generated files from a software, or developed application components, researchers are encouraged to provide additional structured meta information about certain characteristics of these scientific datasets, as these are normally not self-descriptive. For that purpose, several proposed metadata standards and schemas already exist [1]. Such a metadata description nowadays commonly comprises a title, some information about the author and institution, some other administrative or citation metadata, some simple and maybe ambiguous keywords and an unstructured free-text description of the main content. However, especially for early-career researchers, it is an obstacle to start with research data publishing

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029704 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021.
Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13745 (CC BY-SA 4.0)

(RDP) because they might not be aware of relevant existing standards, are bored to fill out extensive, static, text input-oriented submission forms in well-established research data repository applications, or see it as a time-consuming activity without support or interaction.

In 2020, we conducted a survey among early 24 career researchers of varying discipline, experience level and demographic characteristics at Chennitz University of Technology, which focused on their research data publishing and metadata description behavior [2]. As a result, it is shown in Fig. 1, that a vast amount of the participants has never published research data with an appropriate description so far and that knowledge about existing standards on schemas and vocabularies is limited.

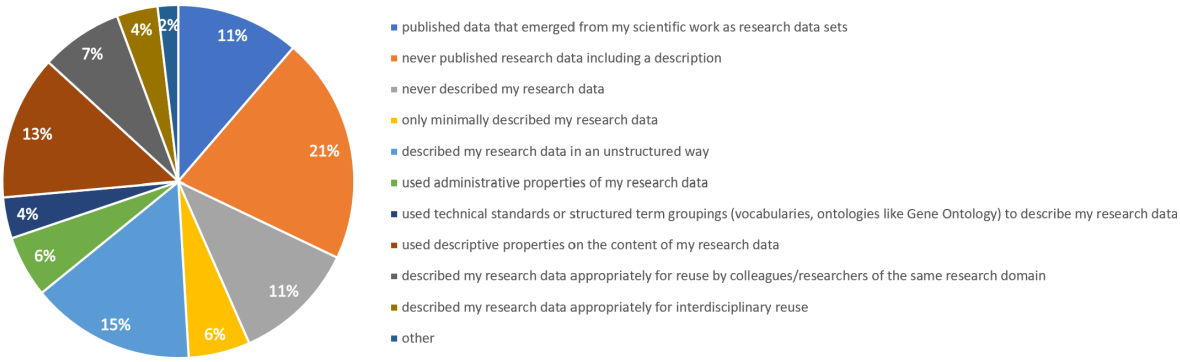


Figure 1: Preliminary survey among 24 early-career researchers on research data publishing.

Furthermore, the findability and filter-ability for particular research data requires highly structured, unambiguous, fine-grained metadata, which is often not provided in floating text data descriptions or free text form fields, especially for interdisciplinary purposes.

To address this issue, Chatbot-like user interfaces are a promising approach that were already successfully applied in other knowledge domains to request structured information from a user and guide the user through a set of relevant questions in an adaptive fashion [3]. In the particular domain of scientific metadata management, the number of existing approaches is still limited.

We investigate the feasibility and challenges of such a conversational UI-based approach to build the prototype of a dialog system for research data descriptions based on the Rasa framework and the OpenAIRE Guidelines for Data Archives which will generate a semantically enriched XML file result. This export file can then be used as a structured data source in a consecutive application or tool chain, or it can simply be published as microdata together with the corresponding dataset on web platforms, in order to improve the structured description and discoverability of the shared research data according to the FAIR principles. This research is part of the PIROL PhD project on Publishing Interdisciplinary Research Over Linked data [4] and carried out in the context of the highly interdisciplinary collaborative research center SFB1410 Hybrid Societies.

The rest of the paper is structured in the following way: In section Related Work, we discuss existing metadata standards for research data meta descriptions, user interface

approaches on how to provide them and related work on the application of conversational user interfaces to that scenario. Section BIRD Concept states objectives and the general concept to realize such a description tool. The implementation of a prototypical demonstrator is discussed in section Realization. Section Conclusion summarizes our results and provides an outlook to future activities.

2 Related Work

Providing an additional metadata description for published research data can be done in various ways, as stated in [5], encompassing also traditional approaches based on simple, rarely structured readme plain text files. Encouraged by the FAIR Principles from [6], vocabularies can alternatively be used to provide such a metadata description in a structured way, such as DCAT-AP [7], the OpenAIRE Guidelines [8] for Data Archives based on DataCite [9] or schema.org/Dataset [10]. Beside these general-purpose controlled vocabularies, more than 1.500 other ontologies are listed in 2021 in [1].

Tools already exist that support users in the provision of required meta information based on these vocabularies, which are typically form-based submission interfaces in research data management applications, wizard-based approaches, such as [11], or markup generators [12, 13].

Chatbots are an alternative, already well-understood paradigm to interact with the user. They were already successfully applied in several knowledge domains [14] and studies also show benefits in the application of a dialog-based system in the acquisition of scientific data [15, 3]. Intelligent Conversational User Interfaces as an advancement are expected to be one of the most rapid growing markets within the next years [16]. To the best of our knowledge, no contribution exists so far that demonstrated the feasibility of an applied chatbot interface for generating platform-independent research metadata descriptions. Thus, we will do that in the following.

3 BIRD Concept

Based on the hypothesis, that especially researcher groups with limited previous experience in publishing research data can benefit from a chatbot-based, ontology-considering dialog system, we formulate the following five objectives:

OBJ1 **Dialog - based user interaction**

OBJ2 **Collection of descriptive research metadata**

OBJ3 **Adaptive conversation progress**

OBJ4 **Possibility for additions and changes**

OBJ5 **Structured, platform-independent result export**

To address OBJ1, we focus in a straight-forward fashion on a rule and story-based chatbot interface to facilitate the creation of FAIR research data meta descriptions with an emphasis on descriptive metadata decoupled from a particular application.

We exemplarily base it for OBJ2 on the DataCite class and property definitions and OpenAIRE guidelines to realize the main interaction path ("Happy Path").

A basic dynamic interaction (OBJ3) with the user is realized with alternative select options for particular questions in the dialog and with the possibility to rewind and correct the last action (OBJ4), or by simply requesting additional explanation.

A structured, schema-based XML metadata export functionality is provided for OBJ5.

4 Realization

We implemented the concept as a Proof-of-Concept (PoC) and published the BIRD prototype for public access[1]

The implementation of an intent-based solution for general metadata was straight-forwardly realized with slots. In the backend, we have chosen the OpenSource Python-based Rasa[2] Natural Language Processing (NLP) and Understanding (NLU) framework for realizing the logic of the chatbot based on a dedicated action server. The frontend user interface was realized based on React/SocketIO/NodeJS and a rasa-webchat component[3] as depicted in figure 2.

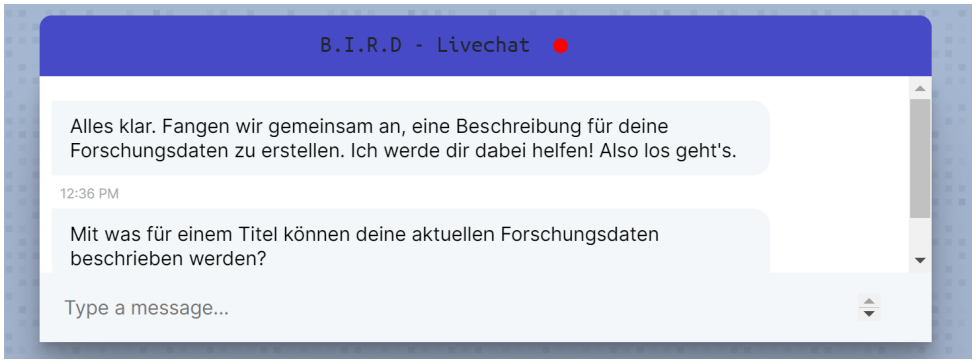


Figure 2: Basic BIRD UI as React Rasa-Webchat component (with German content).

Using intents in this scenario turned out to be challenging, as the chatbot asked the question in an active fashion, and the provided user answer had to be carefully interpreted and filtered. NLP processing performance was OS-dependent and the framework documentation had pitfalls. Handling robustness and a posteriori corrections was challenging and only partially practicable.

[1] https://www.pirol-data.de/bird.
[2] https://rasa.com/
[3] https://github.com/botfront/rasa-webchat

5 Conclusion

To foster the interdisciplinary publication and discovery of research data, tools with a better user interface experience have to be developed that allow a natural and effective provision of structured, relevant meta information with limited prior knowledge.

The BIRD prototype shifts away from traditional forms and allows a dialog-based textural or even lingual metadata collection.

After deployment, it is going to be assessed in a real-world usage scenario. The main focus is then on incorporating taxonomical persistent semantic concept identifiers for common research characteristics and to improve the adaptive dialog behavior. Alternatively, it is worth to investigate hybrid approaches that combine a conversational user interface with a web form.

Acknowledgements

This work is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) Project ID 416228727 SFB 1410.

ORCID IDs

- André Langer https://orcid.org/0000-0001-7073-5377
- Martin Gaedke https://orcid.org/0000-0002-6729-2912

Bibliography

[1] FAIRsharing Standards Overview page. `https://fairsharing.org/standards/`, Accessed: 2021-04-22.

[2] Matthias Tietz, André Langer, and Martin Gaedke. Survey results on the interdisciplinary description of research data. `https://purl.org/net/vsr/storch/survey`, Accessed: 2021-04-22.

[3] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. `https://doi.org/10.1145/3290605.3300316`

[4] André Langer. PIROL : Cross-domain Research Data Publishing with Linked Data technologies. In Marcello La Rosa, Pierluigi Plebani, and Manfred Reichert, editors, *Proceedings of the Doctoral Consortium Papers Presented at the 31st CAiSE 2019*, pages 43–51, Rome, 2019. CEUR.

[5] Dong Joon Lee and Besiki Stvilia. Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLOS ONE*, 12(3):1–44, 2017. https://doi.org/10.1371/journal.pone.0173987.

[6] Mark D. Wilkinson, Michel Dumontier, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. https://doi.org/10.1038/sdata.2016.18.

[7] Bert Van Nuffelen. DCAT Application Profile for data portals in Europe Version 2.0.1. 2020. URL: https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe

[8] Pedro Príncipe, Najla Rettberg, Eloy Rodrigues, Mikael K. Elbæk, Jochen Schirrwagen, Nikos Houssos, Lars Holm Nielsen, and Brigitte Jörg. Openaire guidelines: Supporting interoperability for literature repositories, data archives and cris. *Procedia Computer Science*, 33:92–94, 2014. 12th International Conference on Current Research Information Systems, CRIS 2014. URL: https://www.sciencedirect.com/science/article/pii/S1877050914008059, https://doi.org/https://doi.org/10.1016/j.procs.2014.06.015.

[9] Noémie Ammann, Lars Holm Nielsen, Sebastian Peters, and Madeleine de Smaele. Datacite metadata schema for the publication and citation of research data. 2011.

[10] Data and Datasets - schema.org. https://schema.org/docs/data-and-datasets.html, Accessed: 2021-04-22.

[11] KNB. Morpho, data management for earth, environmental and ecological scientists. https://knb.ecoinformatics.org/tools/morpho, Accessed: 2021-04-23.

[12] DataCite Metadata Generator - Kernel 4.3. https://dhvlab.gwi.uni-muenchen.de/datacite-generator/, Accessed: 2021-04-22.

[13] NustartSolutions. Awesome Step-by-Step JSON-LD Schema Generator Tool (2019). https://nustart.solutions/tools/json-ld-schema-generator-tool/, Accessed: 2021-04-23.

[14] Stefano Valtolina, Serena Di Gaetano, and Pietro Diliberto. Chatbots and Conversational Interfaces: Three Domains of Use. Technical report, 2018. URL: http://ceur-ws.org.

[15] Irene Celino and Gloria Re Calegari. Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. *International Journal of Human-Computer Studies*, 139:102410, 2020. URL: https:

//www.sciencedirect.com/science/article/pii/S107158192030015X, https://doi.org/https://doi.org/10.1016/j.ijhcs.2020.102410.

[16] Mordor Intelligence. Chatbot Market - Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026). https://www.researchandmarkets.com/reports/4622740/chatbot-market-growth-trends-covid-19-impact, Accessed: 2021-04-22.

Mit welchem Aufwand bekommen wir Skripte FAIR(er)?

Denis Arnold und Christian Lang

Leibniz-Institut für Deutsche Sprache

In diesem Beitrag widmen wir uns der Frage, welche Schritte unternommen werden müssen, um Skripte, die bei der Aufbereitung und/oder Auswertung von Forschungsdaten Anwendung finden, so FAIR wie möglich zu gestalten. Dabei nehmen wir sowohl Reproduzierbarkeit, also den Weg von den (Roh)daten zu den Ergebnissen einer Studie, als auch Wiederverwertbarkeit, also die Möglichkeit, die Methoden einer Studie mittels des Skripts auf andere Daten anzuwenden, in den Fokus und beleuchten dabei die folgenden Aspekte: Arbeitsumgebung, Datenvalidierung, Modularisierung, Dokumentation und Lizenz.

1 Einleitung

Die Offenheit von Forschungsdaten ist spätestens seit der Postulierung der TOP-Guidelines im wegweisenden Papier von Nosek et al. [1] ein oft wiederholtes und in weiten Teilen noch unerfülltes Desideratum, das beispielsweise auch die Deutsche Forschungsgemeinschaft im Kodex "Leitlinien zur Sicherung guter wissenschaftlicher Praxis" [2] und im Impulspapier "Digitaler Wandel in den Wissenschaften" [3] adressiert.

Ähnlich gelagert, jedoch weniger als die TOP-Guidelines auf völlige Offenheit fokussiert, sind die von Wilkinson et al. [4] 2016 formulierten FAIR-Prinzipien. Demnach sollen Forschungsdaten auffindbar (**F**indable), zugänglich (**A**ccessible), interoperabel (**I**nteroperable) und nachnutzbar (**R**eusable) sein. Ziel der FAIR-Prinzipien als Leitlinien guten Datenmanagements sind die Vereinfachung des Auffindens, der Evaluation und der Weiterverwendung von entsprechend publizierten Daten [4, S.1].

Gutes Datenmanagement als Herausforderung für Wissenschaftlerinnen und Wissenschaftler datenintensiver Disziplinen[1] berührt verschiedene Bereiche des wissenschaftlichen Prozesses. Im Zentrum einer Vielzahl an Diskussionen steht dabei oftmals die Schaffung von Repositoriumsinfrastrukturen, die es – entsprechende Metadaten vorausgesetzt – ermöglichen, möglichst flexibel Forschungsdaten der (wissenschaftlichen) Öffentlichkeit zur Verfügung zu stellen. In unserem Beitrag wollen wir dagegen den Fokus auf einen Aspekt datengetriebener Forschungsarbeit legen, der in der Diskussion um die Offenheit von

[1] Durch die digitale und empirische Wende lassen sich zunehmend auch Teile der klassischen Geisteswissenschaften als datenintensive Disziplinen einordnen.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13746

Forschungsdaten oftmals nur am Rande oder gar nicht erwähnt wird. Genauer gesagt, betrachten wir Skripte, die Forscherinnen und Forscher zur Datenaufbereitung und Datenauswertung schreiben. Unserer Erfahrung nach werden diese vielfach als Mittel zum Zweck auf dem Weg zum Ergebnis und nicht als integraler Bestandteil der Forschungsdaten und einer offenen Publikationsstrategie betrachtet. Dabei schließen Wilkinson et al. in ihre FAIR-Prinzipien explizit auch sogenannte "non-data assets" ein und nennen in diesem Kontext Algorithmen, Tools und Workflows, die zu den Forschungsdaten geführt haben und auf die ebenso die FAIR-Prinzipien anzuwenden seien [4, S. 4f.]. Auch im Impulspapier "Digitaler Wandel in den Wissenschaften" [3, S. 10] wird "[...] der Zugang zu Daten und Software für die Wissenschaften [...] nach dem FAIR-Prinzip [...]" als Herausforderung benannt und auch im Kodex "Leitlinien zur Sicherung guter wissenschaftlicher Praxis" [2] wird auf die Bedeutung von "(Forschungs)software" verwiesen. Skripte als eigene Gattung werden hier oder auch an anderen Stellen nicht explizit ausgezeichnet, aber es ist davon auszugehen, dass auch sie unter diese Gattung fallen, auch wenn sie häufig nicht als vollwertige Software angesehen werden.

Auch Nosek et al. [1] fordern neben der Offenheit der Daten(grundlage) und der Ergebnisse unter dem Stichwort "Analytic methods (code) transparency" eine entsprechende Offenheit von Methoden und Code, die im idealen Fall (also der höchsten Transparenzstufe) darin besteht, dass der Code in einem entsprechenden Repositorium abgelegt und die Methoden vor der Publikation unabhängig reproduziert werden [1, S. 1424].

Was kann also getan werden, um auch Skripte FAIR(er) zu gestalten? Im Zentrum unseres Interesses steht dabei weniger die Frage, wie Skripte als Forschungsdaten veröffentlicht werden können[2] und welche Anforderungen an entsprechende Repositorien zu stellen sind. Vielmehr konzentrieren wir uns auf die Beschaffenheit der Skripte an sich und fragen uns, was bei deren Erstellung und bezüglich ihrem Aufbau getan werden kann, damit diese für den Fall, dass sie in einem Repositorium gemeinsam mit den Forschungsdaten abgelegt werden, von einem möglichst breiten wissenschaftlichen Publikum nachvollzogen und wiederverwendet werden können. Dabei dienen uns die von Wilkinson et al. formulierten FAIR-Prinzipien als grundlegende Orientierung mit einem besonderen Fokus auf die Interoprabilität (**I**nteroperable) und Wiederverwertbarkeit (**R**eusable) von Skripten.

2 Aspekte

Im Folgenden gehen wir auf verschiedene Aspekte ein, die bei der Erstellung von Aufbereitungs- und Analyseskripten Anhaltspunkte für eine FAIRe(re) Gestaltung im Allgemeinen und eine Erhöhung der Wiederverwendbarkeit im Besonderen bieten. Hierbei versuchen wir eine von konkreten Sprachen unabhängige Perspektive einzunehmen.

[2] Hier bieten Zenodo (https://zenodo.org), Open Science Framework (OSF) (https://osf.io), aber auch Fachrepositorien die Möglichkeit, Skripte auffindbar zugänglich zu machen (zur Kritik an einem vorhandenen/entstehenden Repositorien-Pluralimus siehe [4, S. 2]).

Infolgedessen gehen wir nicht auf spezifische Weisen der Code-Optimierung ein, sondern fokussieren uns auf übergeordnete, allgemeine und (weitestgehend) sprachenunabhängige Aspekte von Skripten.

2.1 Arbeitsumgebung

Skripte werden für bestimmte Sprachen oder Programme geschrieben und sind nicht ohne Weiteres in anderen Versionen lauffähig. Skriptsprachen sind von Interpretern abhängig, die es häufig ermöglichen, den gleichen Code auf verschiedenen Betriebssystemen auszuführen. Trotzdem unterscheidet sich der standardmäßige Gebrauch z. B. von Encodings (etwa Windows-1252 vs. UTF-8) oder auch der Umgang mit Systembibliotheken zwischen verschieden Interpretern und Betriebssystem zum Teil stärker, als dass man sie vernachlässigen könnte. Nahezu alle Skriptsprachen werden durch umfangreiche Bibliotheken besonders attraktiv, die ihrerseits fortlaufender Entwicklung unterliegen. Ein wichtiger Schritt stellt somit eine Zusammenstellung der beteiligten Programme und Bibliotheken und ihren Versionen dar. Viele Skripte benutzen in ihrem Workflow weitere Skripte, Programme (z. B. Datenbanken) oder auch Onlinedienste. Neben der Benennung der Werkzeuge sind außerdem Konfigurationen und Interaktionen mit den Schnittstellen wichtige Bestandteile der Arbeitsumgebung.

2.2 Datenvalidierung

2.2.1 Prüfung auf Datengleichheit

Für eine Reproduktion muss die Datengrundlage verifizierbar sein. Dies ist umso wichtiger, wenn die Daten nicht frei im Sinne von Open, sondern geschützt im Sinne von FAIR sind. Hierzu eigenen sich beispielsweise Listen mit Prüfsummen. Pfade zu Dateien und anderen Programmen sollten in jedem Fall relativ angegeben sein und es sollte in den Schnittstellen die Möglichkeit gegeben sein, Pfade anzupassen.

2.2.2 Testsuiten

In der professionellen Softwareentwicklung sind Implementierung von Tests und Testdaten, also kleinen Datensätzen, die eine Überprüfung der Funktionalität ermöglichen, üblich. Testsuiten ermöglichen eine automatisierte Evaluierung des Codes und unterstützen nicht nur stark in der Entwicklung, sondern dienen auch zur abschließenden Überprüfung. Diesen Testsuiten kommt insbesondere dann eine wichtige Bedeutung zu, wenn die originalen Datensätze zu groß sind oder rechtlichen Beschränkungen unterliegen. Außerdem könnten sie Respositorien automatisch überprüfbare Kriterien für die Übernahme an die Hand geben. Hierzu bräuchte es dann entsprechende Vorgaben durch das Repositorium.

2.2.3 Test auf verarbeitbare Datenstruktur

Wenn Skripte im Sinne der Nachnutzung auf neue Datensätze angewendet werden sollen, müssen Skripte überprüfen, dass Daten auch die vorausgesetzten Strukturen und Eigenschaften besitzen. Mindestens sollte eine Überprüfung in den Funktionen erfolgen und mit sprechenden Fehlermeldungen versehen werden.

2.3 Modularisierung

Für eine Wiederverwendbarkeit ist es besonders günstig, den Workflow möglichst weit zu modularisieren und die Schnittstellen umfassend zu beschreiben. Übliche Schritte eines Workflows sind das Laden von Daten, Vorverarbeitung, Anreicherung, Analysen und Visualisierung. Ein erster Schritt wäre hier, die einzelnen Bestandteile des Workflows zu identifizieren und modular anzulegen. Wenn nun bei der Nachnutzung andere Formate eingelesen werden sollen, kann ein neues Modul die gewünschten Funktionen hinzufügen, während die weiteren Module weiterverwendet werden können. Dies gilt in gleicher Weise, wenn in den übrigen Schritten andere Methoden angewendet werden sollen.

2.4 Dokumentation

Die Dokumentation sollte die Arbeitsumgebung und Nutzung des Skripts vollständig beschreiben. Hierbei sollte möglichst die konkrete Benutzung der einzelnen Skripte auch beispielhaft aufgezeigt werden. Hierfür eignet sich am besten das Testset. Besondere Wichtigkeit kommt den Interaktionen mit anderen Werkzeugen zu. Hier gehört dann die Konfiguration der Komponenten in die Dokumentation. Auch im Skript selbst sollte Dokumentation durch sprechende Fehlermeldungen und Nachrichten, sprechende Benennungen von Funktionen, Variablen und schließlich Kommentare im Quellcode gegeben sein.

2.5 Lizenz

In den FAIR Prinzpien werden Lizenzen in R1.1 adressiert: “(meta)data are released with a clear and accessible data usage license.” Die Frage, welche konkrete Lizenz man für sein Skript, aber auch Daten verwenden sollte, ist nicht leicht, insbesondere wenn die eigenen Arbeiten auf dem Werk Dritter aufbauen. Einen guten Einstig in die Thematik bietet [5]. Der dort beschriebene Public License Selector ist in verschiedenen Diensten integriert worden und lässt sich auch hier finden: `http://ufal.github.io/public-license-selector/`. Licentia [6] ist ein weiteres Werkzeug, das Lizenzen vergleicht und die Übersetzung von Lizenzen in RDF unterstützt und findet sich hier: `http://licentia.inria.fr/`

3 Diskussion und Ausblick

Viele der in Abschnitt 2 genannten Aspekte sind letztlich generelle Hinweise zur Qualitätssicherung von Skripten und Code im Allgemeinen und werden als solche zum Teil auch durch die Leitlinien des *clean codings*, best practices zur Erstellung guten Codes, in der Softwareentwicklung abgedeckt [7]. Einige Gesichtspunkte sind ohne weiterführende Programmierkenntnisse mit weniger (z. B. sprechende Variablen-/Funktionsbenennungen, Kommentare im Skript) oder mehr (z. B. Erstellung einer umfassenden und durch unterschiedliche Nutzergruppen anwendbaren Dokumentation) Aufwand umsetzbar.[3] Andere erfordern fortgeschritteneres (und für die eigentliche Aufgabe nicht unmittelbar und zwangsweise notwendiges) Wissen (z. B. die Implementierung von Datenstrukturtests). Hinsichtlich einer Kontrolle der Arbeitsumgebung bietet sich für die Bereitstellung von lokalen Abhängigkeiten, also dem Interpreter, system- und sprachspezifischen Bibliotheken und weiteren lokalen Softwarekomponenten, grundsätzlich eine Virtualisierung in virtuellen Maschinen und Containerlösungen an. Hier entsteht allerdings Mehraufwand, der wiederum zusätzliche Kenntnisse sowie mehr Zeit bei der Implementierung und Testung benötigt und schlussendlich mehr Speicherplatz belegt, da nicht nur das Skript und die Dokumentation gespeichert werden müssen, sondern der ganze Container beziehungsweise die ganze virtuelle Maschine. Für verschiedene Aufgaben bietet sich diese Herangehensweise an und es gibt einige Projekte, die hierzu Infrastruktur anbieten. Die Langfristigkeit solcher Lösungen ist hierbei auch vom Bestand der Formate oder gegebenenfalls Emulatoren abhängig.

Von Seiten der Forschungsinfrastrukturen stellt sich für die Übernahme von Skripten in Repositorien die Frage nach geeigneten Metadatenschemata und danach, wie man den Nutzen aufwandsarm validieren kann. Für Ersteres könnten bestehende Standards ein guter Ausgangspunkt sein, für Zweiteres könnten Testsuiten einen ersten Schritt darstellen.

Eine FAIR(ere) Aufbereitung von Skripten setzt – wie die vorangegangene Diskussion zeigt – über die Untersuchung der eigentlichen Forschungsfrage hinausgehenden Aufwand und Kompetenzen voraus (und dies von Fachwissenschaftlerinnen und Fachwissenschaftlern, die in der Regel keine ausgebildeten Programmierer, sondern häufig Autodidakten sind, die bedarfsgeleitet Programmiersprachen einsetzen). Dennoch sind Skripte – verstanden als Implementierung analytischer Workflows – als "non-data assets" nach Wilkinson et al. eine zentrale Komponente des wissenschaftlichen Prozesses und deshalb lohnt es sich, den Aufwand zu investieren, diese so FAIR wie möglich zu gestalten: "Analytical workflows, for example, are a critical component of the scholarly ecosystem, and their formal publication is necessary to achieve both transparency and scientific reproducibility." [4, S. 5] Über die Nachnutzbarkeit hinaus nutzt eine solche Vorgehensweise letztlich auch dazu, den eigenen Forschungsprozess besser zu handhaben.

[3] Für viele verschiedene Programmiersprachen existieren Pakete, die Schnittstellen zu *pandoc* [8] bieten und somit eine Dokumentationserstellung direkt aus dem Code in unterschiedliche Ausgabeformate ermöglichen.

4 Danksagung

Wir bedanken uns bei zwei anonymen Gutachtern und Bernhard Fisseni und Thorsten Trippel für hilfreiche Kommentare und Diskussionen.

Literaturverzeichnis

[1] Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

[2] Deutsche Forschungsgemeinschaft "Leitlinien zur Sicherung guter wissenschaftlicher Praxis", 2019. `https://doi.org/10.5281/zenodo.3923602`

[3] Deutsche Forschungsgemeinschaft "Digitaler Wandel in den Wissenschaften", 2020. `https://doi.org/10.5281/zenodo.4191345`

[4] Wilkinson, M. D., M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[5] Kamocki, Pawel, P. Straňák, M. Sedlák. "The Public License Selector: Making open licensing easier." In *Proceedings of the Tenth International Conference of Language Resources and Evaluation (LREC 2016)*, edited by N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard et al., 2533-2538. Paris: European Language Resources Association (ELRA), 2016.

[6] Cardellino, C., S. Villata, F. Gandon, G. Governatori, H. Lam, A. Rotolo. "Licentia: a Tool for Supporting Users in Data Licensing on the Web of Data." In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014)*, edited by M. Horridge, M. Rospocher, J. van Ossenbruggen, 277-280, CEUR-WS.org, 2014. http://ceur-ws.org/Vol-1272

[7] Martin, R. C. *Clean Code. A Handbook of Agile Software Craftsmanship*. Upper Saddle Rive, NJ et al.: Prentice Hall, 2009.

[8] McFarlane, J. "pandoc: A universal document converter [Software]" (2016)

NFDI4BIOIMAGE – An Initiative for a National Research Data Infrastructure for Microscopy Data

Christian Schmidt[1] and Elisa Ferrando-May[1,2]

[1]Bioimaging Center, University of Konstanz, Konstanz, Germany
[2]German BioImaging – Gesellschaft für Mikroskopie und Bildanalyse e.V.

Bioimaging and biophotonics are key enabling research technologies in the natural and biomedical sciences. They both foot to a large extent on (light) microscopy, which has transformed from a mainly qualitative observational method to a big-data quantitative approach, as exemplified by automated high-content imaging. Advancements in microscopy instrumentation are achieved at an unprecedented pace driving the production of vast amounts of bioimage data. Novel AI-based bioimage informatics tools are emerging and facilitate knowledge extraction from these highly complex data with high information density. Therefore, image processing and analysis have become an intrinsic and essential component of bioimage-based research. At present, bioimaging data is mainly stored locally and is often not systematically annotated. Proprietary software and heterogeneous file formats impede comparability. Unfolding the full potential of bioimaging and biophotonics requires a culture of image data sharing and re-use that could advance research in multiple scientific disciplines. FAIRification of bioimage data management demands the development and adoption of common standards, harmonizing data handling practices, and extensive training and user education. Leveraging our experience within German BioImaging – Society for Microscopy and Image Analysis (GerBI-GMB) in bringing together microscopy users, IT infrastructure providers, image analysts, and application specialists at core facilities in Germany, we aim to tackle these challenges. We intend to submit our proposal for a consortium within the national research data infrastructure (NFDI) in 2021 to foster a state-of-the-art, high-quality bioimage data management ecosystem in Germany's research data management landscape.

1 Introduction

The national research data infrastructure (NFDI) is currently being established in Germany as a network of closely collaborating consortia. Aims are to manage the scientific

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029489 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13747 (CC BY-SA 4.0)

data systematically, develop standards and solutions for data handling, and provide concepts for long-term storage according to the FAIR principles, i.e., data must be findable, accessible, interoperable, and re-usable [1, 2]. Individual NFDI consortia represent a community of data users and data providers based on a scientific discipline or scientific method. Up to 30 consortia will be incorporated into the NFDI framework in a science-driven process orchestrated by the German Research Foundation, DFG, in three consecutive application calls between 2019 and 2021 [2]. By close collaboration across disciplines, all consortia contribute to research data management standards to the mutual benefit. Basic services relevant for all consortia and common aspects with shared solutions among a subgroup of the consortia will be addressed as cross-cutting topics [3, 4].

Bioimaging methods enable in-depth insight into biological and biomedical samples or materials at high spatial and temporal resolution. Method development and research in bioimage informatics facilitate the extraction of knowledge from images while promoting the advancement of both instrumentation and software. At present, data acquired in bioimaging experiments are often stored locally, and are not systematically annotated and archived according to common standards. The lack of shared research data management standards for bioimaging is also reflected in the often incomplete description of imaging methods in published research articles [5].

Recently established archives for published image data like the BioImage Archive (BIA) hosted at the European Bioinformatics Institute or the Image Data Resource (IDR) showcase the benefit of highly curated and cross-referenced image datasets to the scientific community [6, 7]. However, a large number of often proprietary file formats, insufficient metadata standards, and the need to establish new formats for cloud-based computing and storage impose considerable challenges on the FAIRification efforts for bioimage data [8, 9]. The integration of bioimage (meta)data with multimodal and multidisciplinary data sources, e.g., from established "-omics" research areas like genomics, proteomics, metabolomics, and more, needs to be addressed. Increased data accessibility and interoperability will enable the re-use of bioimage data to extract novel information from existing data of diverse sources. These developments will advance the field of bioimage informatics, including machine-learning algorithms for bioimage analysis and computational modeling.

2 Resources

One important pillar of the NFDI4BIOIMAGE initiative is German BioImaging, a well-organized network of >50 imaging core facilities spread across research institutions.
Through this network, the initiative has established contacts with a large number of microscopy users [10]. A significant part of bioimaging data in Germany is acquired at or in close collaboration with core facilities, which puts NFDI4BIOIMAGE in an excellent position to survey community needs and distribute novel standards and solutions for bioimage data management. The consortium has both a methodological and disciplinary scope, the latter comprising the field of bioimage informatics. Task Areas will be devoted to, e.g., (meta)data standards and formats, including the development of next-generation file for-

mats to overcome current limitations in data handling and transformation [8]. The team includes among others leading experts in software engineering specialized in bioimage data formats and data management tools. Furthermore, NFDI4BIOIMAGE leverages the work of the Research Data Management for Microscopy (RDM4Mic) group, which is part of German BioImaging and has a long-standing collaboration with the Open Microscopy Environment (OME) consortium. OME is home to OMERO (OME Remote Objects), the most widely used management platform for microscopy data. Thus, NFDI4BIOIMAGE will promote OMERO as one possible solution for FAIR bioimage data handling. In particular, common metadata recommendations and best practices for running OMERO instances at local institutions will be addressed [11]. Members of yet another consortium, the Quality Assessment and Reproducibility for Instruments and Images in Light Microscopy (QUAREP-LiMi) group [12], will contribute to NFDI4BIOIMAGE with work on quality criteria for bioimage acquisition and their representation in metadata. This work includes collaboration with commercial vendors. With members in Euro-BioImaging and Global BioImaging, the NFDI4BIOIMAGE initiative has further well-established international connections.

Task Areas on technical infrastructures and on data linking & multimodal data integration will guide the choice and configuration of local and decentralized hardware. They will work towards increased interoperability of microscopy data with other data types and platforms. Integration of these solutions with laboratory information management systems and electronic lab notebooks is another topic dealt with by the consortium. We also build on established connections to the IDR and BIA and will promote user-friendly workflows and guidelines to connect these repositories to national storage and archiving solutions. A Task Area on bioimage informatics aims to establish user-friendly tools and interfaces for FAIR bioimage analysis, enabling researchers from all disciplines to use state-of-the-art tools and software.

The resources mentioned above enable the initiative to focus on all aspects of the bioimaging data life cycle from experiment planning to data acquisition, annotation and storage, image analysis, data processing, and finally to publication and archiving. Ultimately, the re-use and mining of existing data should be routinely considered when planning to conduct a bioimaging experiment and should belong to the standard skills of researchers in the life sciences. NFDI4BIOMAGE aims to become a reliable resource for bioimage data users, data generators, and data stewards, and will do so by making services and solutions sustainable and openly available to the whole community. Notably, the consortium will generate benefits for the NFDI as a whole by collaborating with discipline-specific NFDI consortia and contributing to cross-cutting topics.

3 Support

German BioImaging has been funded by DFG from 2012 to 2017 as a scientific network. In 2017, it was transformed into a scientific society with the legal form of a non-profit association. A project for bioimage data management within the DFG funding line In-

formation Infrastructures for Research Data initiated by members of German BioImaging has been recently granted funding. It represents a first step for developing FAIR bioimage data management practices based on the platform OMERO. The project, termed I3D:bio (Information Infrastructure for BioImage Data) will start at the beginning of 2022. Fostering good practices of research data management and open science is a shared vision of the participating institutions.

4 Conclusions

The National Research Data Infrastructure paves the way for a community-driven, collaborative effort to facilitate research data management in Germany and network data internationally. We aim to make a substantial contribution within the framework of the NFDI with our focus on microscopy, biophotonics, and bioimage analysis as a consortial initiative applying in the third call for proposals in 2021. We welcome interested researchers, IT professionals, and, in general, people from the community and NFDI consortia to get in contact and collaborate on the aims of NFDI4BIOIMAGE, provide input and feedback. Please visit https://nfdi4biomage.de for further information.

Acknowledgements

The authors present the envisioned NFDI4BIOIMAGE consortium on behalf of the initiative's members, participants, and supporters – a growing community of committed people. We would like to thank Stefanie Weidtkamp-Peters (Heinrich Heine University of Düsseldorf, spokesperson of the initiative), Pavol Bauer (LIN Magdeburg), Markus Becker (INP Greifswald), Thomas Bocklitz (IPHT Jena), Timo Dickscheid (FZ Jülich), Marc Thilo Figge (HKI Jena), Susanne Kunis and Karen Bernhardt (University of Osnabrück), Matthias Landwehr (University of Konstanz), Josh Moore (University of Dundee, OME-Team), Roland Nitschke (University of Freiburg), and Stephanie Rehwald (University of Duisburg-Essen) for their contributions in the early phases of shaping the NFDI4BIOIMAGE initiative. We are thankful for the support by the mentioned initiatives and the growing list of contributors and participants. We also thank the University of Konstanz for supporting the preparation phase of the NFDI4BIOIMAGE proposal.

Bibliography

[1] Wilkinson, M.D., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data, 2016. 3: p. 160018. https://doi.org/10.1038/sdata.2016.18

[2] Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018, in BAnz AT 21.12.2018 B10. 2018: Germany.

[3] The development of the National Research Data Infrastructure (NFDI). Second statement of the NFDI Expert Committee. 2020. URL: https://www.dfg.de/download/pdf/foerderung/programme/nfdi/stellungnahme_nfdi_201112_en.pdf (last visit: May 10th, 2021)

[4] Bierwirth, M., et al.: Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung. Zenodo, 2020. http://doi.org/10.5281/zenodo.3895209

[5] Marques, G., T. Pengo, and M.A. Sanders, Imaging methods are vastly underreported in biomedical research. Elife, 2020. 9.e55133 https://doi.org/10.7554/eLife.55133

[6] Ellenberg, J., et al.: A call for public archives for biological image data. Nat Methods, 2018. 15(11): p. 849-854. https://doi.org/10.1038/s41592-018-0195-8

[7] Williams, E., et al.: The Image Data Resource: A Bioimage Data Integration and Publication Platform. Nat Methods, 2017. 14(8): p. 775-781. https://doi.org/10.1038/nmeth.4326

[8] Moore, J., et al.: OME-NGFF: scalable format strategies for interoperable bioimaging data. bioRxiv, 2021. https://doi.org/10.1101/2021.03.31.437929

[9] Swedlow, J.R., et al.,: A Global View of Standards for Open Image Data Formats and Repositories. Nature Methods, 2021, https://doi.org/10.1038/s41592-021-01113-7

[10] Ferrando-May, E., et al.: Advanced light microscopy core facilities: Balancing service, science and career. Microsc Res Tech, 2016. 79(6): p. 463-79. https://doi.org/10.1002/jemt.22648

[11] RDM4Mic. 2021; URL: https://german-bioimaging.github.io/RDM4mic.github.io/ (last visit: Aug 9th, 2021)

[12] Nelson, G., et al.: QUAREP-LiMi: A community-driven initiative to establish guidelines for quality assessment and reproducibility for instruments and images in light microscopy, J Microsc. 2021 Jul 2. doi: 10.1111/jmi.13041. Online ahead of print

SDC4Lit – Science Data Center for Literature. Aufbau eines nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung

Jan Hess[1], Alexander Holz[1], Nina Buck[2], Andreas Ganzenmüller[2], Volodymyr Kushnarenko[2], Björn Schembera[2], André Blessing[4], Pascal Hein[3], Kerstin Jung[4], Heinz Werner Kramski[1], Claus-Michael Schlesinger[3], Mona Ulrich[1], Thomas Bönisch[2], Andreas Kaminski[2], Roland S. Kamzelak[1], Jonas Kuhn[4] and Gabriel Viehhauser[3]

[1]Deutsches Literaturarchiv Marbach
[2]Höchstleistungsrechenzentrum Stuttgart
[3]Institut für Literaturwissenschaft/Digital Humanities, Universität Stuttgart
[4]Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Das Science Data Center for Literature (SDC4Lit) hat sich das Ziel gesetzt, die Anforderungen, die (Digitale) Literatur an ihre Archivierung, Erforschung und Vermittlung stellt, systematisch zu reflektieren und entsprechende Lösungen für einen nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung langfristig umzusetzen. Im Zentrum stehen dabei der Aufbau langzeitverfügbarer Repositories für (Digitale) Literatur und die Entwicklung einer Forschungsplattform.

1 Projektziel

Die Digitalisierung verändert die Bedingungen für die Produktion, Distribution, Rezeption und damit auch für die Erforschung von Literatur. Die veränderten medialen Bedingungen führen nicht nur zur Übersetzung von gedruckten Texten in digitale Objekte, sondern bringen selbst produktiv neue Literaturformen und -gattungen hervor. Hierzu zählen etwa literarische Hypertexte, Blog-Formate, literarische Tweets und Twitter-Bots, aber auch Texte und Textgeneratoren, die auf computerlinguistische Methoden setzen. Zum einem scheinen sich diese Texte zur Anwendung computergestützter Analysemethoden besonders anzubieten, da sie genuin in elektronischer Form vorliegen. Zum anderen bringt diese Form für ihre Archivierung und Bereitstellung eine Reihe von besonderen Anforderungen mit sich.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00030001 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13748 (CC BY-SA 4.0)

So führen die hochfrequenten Erneuerungszyklen digitaler Technik dazu, dass die ursprünglichen Darbietungsformen historischer elektronischer Texte teils aufwendig rekonstruiert werden müssen, da die entsprechende Hard- oder Software schnell veraltet ist.

Das Science Data Center for Literature (SDC4Lit)[1] hat sich das Ziel gesetzt, die Anforderungen, die Digitale Literatur an ihre Archivierung, Erforschung und Vermittlung stellt, systematisch zu reflektieren und entsprechende Lösungen für einen nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung langfristig umzusetzen. Im Zentrum stehen dabei der Aufbau eines langzeitverfügbaren Repositoriums für (Digitale) Literatur und die Entwicklung einer Forschungsplattform, die die Möglichkeit zum computergestützten Arbeiten mit den Beständen der Repositorien bietet. Da eine solche Repositoriumsstruktur, die Sammeln, Archivieren und Analysieren miteinander verzahnt, nur in der interdisziplinären Zusammenarbeit zu bewerkstelligen ist, sind mit dem Deutschen Literaturarchiv Marbach (DLA), dem Höchstleistungsrechenzentrum Stuttgart (HLRS), dem Institut für Maschinelle Sprachverarbeitung (IMS) sowie dem Institut für Literaturwissenschaft/Digital Humanities (ILW) an der Universität Stuttgart im Projekt Partner mit Expertisen in den verschiedenen Bereichen miteinander vereint.

2 Datenmaterial

Die Daten, die im Repositorium gesammelt, archiviert und – sofern rechtlich möglich – zur Verfügung gestellt werden, stammen aus der Sammlung des Deutschen Literaturarchivs Marbach und werden stetig durch neue Bestände und Objekte ergänzt. Die Datenmenge lässt sich in drei Themenbereiche gliedern.

Den ersten Teil bildet der Bereich Literatur im Netz. Hierbei handelt es sich um archivierte Webseiten, literarische Blogs oder Online-Magazine mit Bezug zur Neueren deutschen Literatur. Dieses Korpus wurde am DLA von 2008 bis 2018 zusammengestellt und kuratiert. In dieser Zeit wurden etwa 500 Internet-Quellen einmalig oder wiederholt gespeichert und so insgesamt ca. 3.840 Speicherungen vollzogen. Zur Archivierung der Webseiten wurden über die Jahre unterschiedliche Capturing-Tools und -Techniken eingesetzt (u. a. HTTrack[2] und Heritrix[3]), sodass aktuell ca. 75% der Speicherungen im Zielformat WARC[4] und die restlichen als offene Ressourcen in Verzeichnissen vorliegen. Beide Speicherungsformen werden ins Repositorium übernommen. Der zweite Teil besteht aus genuin digitalen Vor- und Nachlässen. Hierbei handelt es sich um Born-digitals[5], die dem DLA von Autorinnen und Autoren oder Institutionen zur Archivierung überlassen wurden. Diese

[1] SDC4Lit Homepage, `https://www.sdc4lit.de/`, letzter Abruf: 10.05.2021.

[2] Roche, Xavier et al., "HTTrack Website Copier 3.49-2",`https://www.httrack.com/`, letzter Abruf: 10.05.2021.

[3] "Heritrix3", Internet Archive, `https://github.com/internetarchive/heritrix3/wiki`, letzter Abruf: 10.05.2021.

[4] "The WARC Format 1.1", IIPC,`https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/`, letzter Abruf: 10.05.2021.

[5] Kramski, Heinz Werner, "Stichwort Born-digitals", `https://edlex.de/index.php?title=Born-digitals`, letzter Abruf: 10.05.2021.

Vor- und Nachlässe können alle technisch möglichen Datenträger, Dateiformate und Datenmengen enthalten und müssen mehrfach bearbeitet werden, um sie strukturiert und nutzbar im Repositorium ablegen zu können. Aktuell werden Born-digital Vor- und Nachlässe von etwa 75 Bestandsbildnern am DLA archiviert. Diese Vor- und Nachlässe befinden sich in unterschiedlichen Bearbeitungsstadien und stellen eine Gesamt-Datenmenge von ca. 2,8 TB, verteilt auf etwa 2000 Datenträger, dar.

Den dritten und bisher kleinsten Teil bilden die Computerspiele mit Bezug zur (deutschsprachigen) Literatur. Die Sammlung reicht hier von einfachen, textbasierten Spielen bis hin zu aufwändigen 3D-Adventures und weist somit auch hier eine weite Spanne von unterschiedlichen technischen Anforderungen, Datenformaten und Zugangsmöglichkeiten auf. Diese Breite an Sammlungsobjekten aus unterschiedlichen Zeiten, Quellen und Systemen stellt in mehrfacher Hinsicht eine große Herausforderung dar. Bereits die Vorbereitung dieser Daten für das Repositorium gestaltet sich aufgrund der Obsoleszenz von Datenträgern und -formaten, möglicher unlesbarer bzw. defekter Dateien, der Formatmigration etc. anspruchsvoll. Neben den technischen Herausforderungen erschwert die heterogene Datenlage auch die Auswahl einer geeigneten Repositoriumssoftware bzw. die Auswahl und Anwendung verschiedener Metadaten-Standards. Ein Ziel von SDC4Lit besteht darin, die beschriebenen Daten nutzbar und in ihrer Ästhetik authentisch für die Forschung und Vermittlung zu Verfügung zu stellen.

3 Architektur-Entwurf

Zu den Kernaufgaben des Projekts gehört der Aufbau einer Plattform, die archivierte digitale Objekte nicht nur passiv zur Verfügung stellt, sondern auch eine Interaktion mit ebendiesen Objekten erlaubt, um weitergehende Forschung und Analyse zu ermöglichen (Abb. 1). Im Gegensatz zu bisherigen generischen Lösungen wie [1], [8] wird in SDC4Lit ein disziplinspezifischer Ansatz mit starker Integration in die Forschungsinfrastruktur verfolgt.

Eine der Kernkomponenten der SDC4Lit-Architektur ist das Primärdaten-Repositorium, in dem alle Objekte der digitalen Literatur langfristig gespeichert werden. Dazu werden die Archivmaterialien zunächst entsprechend vorbereitet und in das Repositorium eingefügt. Kuratiert werden die Bestände des Primärdaten-Repositoriums ausschließlich von DLA-Mitarbeiterinnen und -mitarbeitern. Nutzerinnen und Nutzer des Repositoriums sollen gezielt nach Objekten suchen und mit gefundenen Objekten weiterarbeiten können.

Für die weitere Arbeit mit den über die Plattform zur Verfügung gestellten Archivalien ist der Aufbau einer Forschungs- und Analyseumgebung geplant, die niedrigschwellig zu bedienende digitale Analysemethoden und -werkzeuge zusammenführt, dokumentiert und in Form modularer Pipelines für die Forschung bereitstellt. Die Ergebnisse der Analysen können von den Nutzerinnen und Nutzern als Forschungsdaten in einem separaten Repositorium gespeichert werden, um sie für die weitere Nachnutzung zur Verfügung zu stellen.

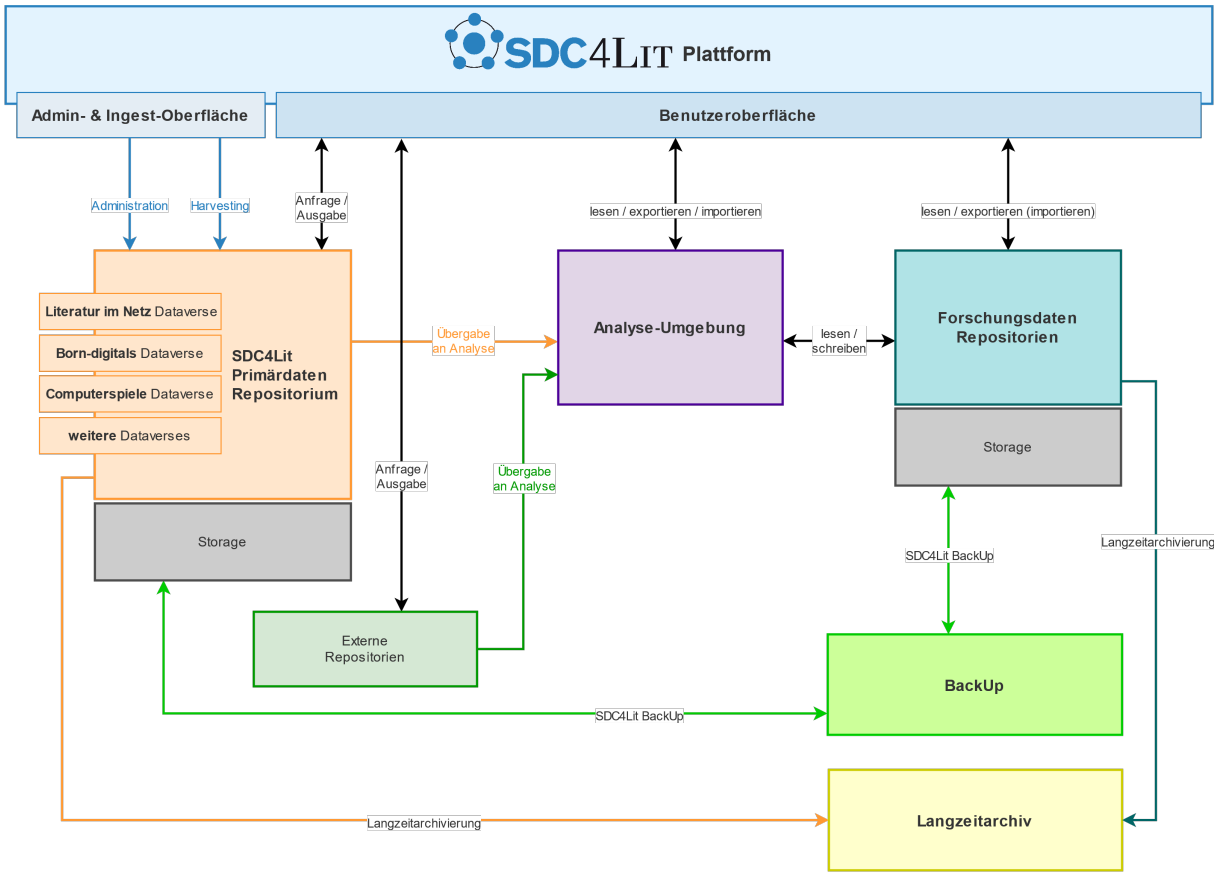


Abbildung 1: Entwurf der SDC4Lit-Architektur.

Um die Nachhaltigkeit und Sicherheit der Daten gewährleisten zu können, ist ein Tape-Backup für beide Repositorien sowie die Anbindung an ein Langzeitarchiv vorgesehen. Da sich die Frage der Nachhaltigkeit nicht allein in der Datenspeicherung erschöpft, sondern auch die Vermittlung von Methodenkompetenzen sowie die Rückkopplung mit den Bedürfnissen der Fachcommunity betrifft, verfolgt das SDC4Lit einen forschungs- und vermittlungsorientierten Ansatz, der auf Fallstudien zur (digitalen) Literatur basiert und in dessen Kontext bereits vorhandene literaturwissenschaftliche Methoden und Werkzeuge erprobt oder (weiter-)entwickelt werden. Neben verschiedenen anderen Fallstudien und methodischen Ansätzen wurden und werden beispielsweise verschiedene Entitätenerkenner hinsichtlich ihrer Verwendbarkeit für literarische Blogs und Zeitschriften getestet. [7] Zur Unterstützung der Erschließung und der textgenetischen Erforschung digitaler Vor- und Nachlässe werden Textähnlichkeitsmaße und Text-Reuse-Werkzeuge erprobt und weiterentwickelt. In einer Beta-Version zur Verfügung steht bereits ein in Zusammenarbeit mit Studierenden der Universität Stuttgart im Rahmen eines Projektseminars entwickeltes Software-Modul zur Erforschung nicht-linearer narrativer Strukturen von Netzlite-

ratur.[6] Dieses warc2graph[7] benannte Python-Modul, mit dem sich WARC- und Webobjekte graphbasiert modellieren lassen, wurde im Rahmen der E-Science-Tage 2021 in einem Tandem-Talk vorgestellt. Zusammen mit anderen bereitgestellten Methoden und Werkzeugen wird warc2graph in der späteren Projektphase in die Forschungsumgebung eingebunden und über die SDC4Lit-Plattform zur Verfügung stehen, um die im SDC4Lit-Primärdaten-Repositorium bereitgestellten, aber auch eigene Korpora analysieren zu können.

Die beschriebenen Infrastrukturkomponenten sollen über die gemeinsame SDC4Lit-Plattform bedient werden können, auf der alle Arbeitsschritte für die Nutzerinnen und Nutzer transparent und selbsterklärend aufbereitet werden. Geplant ist, zukünftig auch andere externe Primärdaten-Repositorien anzubinden und so verschiedene Communitys auf einer Plattform zusammenzubringen.

4 Herausforderungen

Die Obsoleszenz der Dateiformate und Datenträger, der große Umfang an Daten der digitalen Vor- und Nachlässe sowie die Diversität der Formate aller Primärmaterialien sind große Herausforderungen bei der Archivierung und Analyse der Daten. Eine daraus resultierende Kernaufgabe war die Auswahl einer geeigneten Repositoriumssoftware. Viele vorhandene Softwarelösungen zum Aufbau von Repositorien bringen zum Teil einen sehr unterschiedlichen Funktionsumfang mit. Um ein geeignetes Softwarepaket auswählen zu können, wurden anhand von User Storys ein Katalog von „Anforderungen an Repositorien“ formuliert und die einzelnen Anforderungen jeweils priorisiert und gewichtet. Anhand dessen wurden unterschiedliche Softwarelösungen, namentlich DSpace, Dataverse, MyCoRe, Fedora bzw. Islandora, Invenio und AtoM, in mehreren Schritten analysiert und evaluiert. Da keine der vorhandenen Open-Source-Produkte sämtliche unserer Anforderungen [2] vollumfänglich erfüllen konnte, müssen für bestimmte Funktionalitäten Eigenentwicklungen geleistet werden. Die Entscheidung fiel letztendlich auf Dataverse[8], da hier die meisten Kriterien erfüllt wurden und auf eigene Erfahrungen, beispielsweise aus dem Projekt DIPL-ING [7], zurückgegriffen werden kann.

Eine weitere Herausforderung, die derzeit im Fokus steht, ist die (Meta-)Datenmodellierung. Da bisher keine Standards existieren, die die SDC4Lit-Primärdatenbestände annähernd vollständig abdecken, werden in SDC4Lit eigene Datenmodelle entwickelt, die auf Vorarbeiten und Erfahrungen der Projektpartner basieren und den FAIR-Prinzipien

[6] Weitere Informationen sind dem Full Paper zum Vortrag zu entnehmen, das ebenfalls im Tagungsband der EST-21-Konferenz unter dem Titel ”Nicht-lineare narrative Strukturen in Netzliteratur: Speicherung und Nachnutzung von Forschungsdaten aus der computergestützten Extraktion von Verweisstrukturen in Hypertexten”veröffentlicht wird.

[7] Der Programm-Code von warc2graph wird auf `https://github.com/dla-marbach/warc2graph` veröffentlicht. Das Paket kann über den Python Package Index installiert werden: `https://pypi.org/project/warc2graph/`. Ein aktueller Snapshot (Version 0.1.1) ist über Zenodo verfügbar: DOI:10.5281/zenodo.4742254 (Hein et al. 2021 [6]).

[8] “The Dataverse Project”, Dataverse, `https://dataverse.org/`, letzter Abruf: 10.05.2021.

entsprechen [10]. Um Interoperabilität und Nachnutzbarkeit der Daten zu gewährleisten, orientiert sich SDC4Lit an Metadatenstandards wie METS [3], MODS [5] und PREMIS [4], die im Bereich des Bibliotheks- und Archivwesens etabliert sind.

Zurzeit[9] wird ein Repositoriumsprototyp auf Basis von Dataverse aufgesetzt und demnächst evaluiert. Im nächsten Schritt werden die ersten Daten in das Repositorium eingepflegt und für die weitere Bearbeitung bereitgestellt. Die Analysemethoden und -werkzeuge werden ausgewählt oder (weiter-)entwickelt und in die Forschungsumgebung integriert, sodass das Hauptziel von SDC4Lit – der Aufbau eines nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung – langfristig umgesetzt werden kann.

Danksagung

SDC4Lit wird gefördert vom Ministerium für Wissenschaft, Forschung und Kunst in Baden-Württemberg.

Literaturverzeichnis

[1] Felix Bach, Björn Schembera, and Jos Van Wezel, “Design and Implementation of the first Generic Archive Storage Service for Research Data in Germany”, International Journal of Digital Curation 15.1 (2020).

[2] Felix Bach et al., “Kriterien für die Auswahl einer Softwarelösung für den Betrieb eines Repositoriums für Forschungsdaten”. Bausteine Forschungsdatenmanagement (`https://bausteine-fdm.de`) (eingereicht; Publikation vorauss. 2021).

[3] Linda Cantara, “METS: The metadata encoding and transmission standard”, Cataloging & classification quarterly 40.3-4 (2005): 237-253.

[4] Priscilla Caplan, “Understanding Premis”, Washington DC, USA: Library of Congress, 2009.

[5] Richard Gartner, “MODS: Metadata object description schema”, JISC Techwatch report TSW (2003): 3-6.

[6] Pascal Hein et al., „Warc2graph,“ Zenodo, 2021. DOI: `https://doi.org/10.5281/zenodo.4742254`, `https://zenodo.org/record/4742254`.

[7] Kerstin Jung et al., Workshop “Ensemble-Methoden aus menschlichen und maschinellen Bewertungen - Ein Entitäten Abstimmungsexperiment für 3 Tools und N Forschende”, `https://vdhd2021.hypotheses.org/197`, letzter Abruf: 10.05.2021.

[8] Angelina Kraft et al., “The RADAR Project – A Service for Research Data Archival and Publication”, ISPRS International Journal of Geo-Information 5.3 (2016): 28.

[9] Stand: April 2021.

[9] Björn Schembera et al., “Datenmanagement in Infrastrukturen, Prozessen und Lebenszyklen für die Ingenieurwissenschaften: Abschlussbericht des BMBF-Projektes Dipl-Ing”, (2019). https://doi.org/10.2314/KXP:1693393980.

[10] Mark D.Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship”, Scientific data 3.1 (2016): 1-9.

HUBzero als open-source Science Gateway im Rahmen des Science Data Centers BioDATEN

Holger Gauza, Fabian Wannenmacher, Johannes Werner, Thomas Zajac und Jens Krüger

Eberhard Karls Universität Tübingen, High Performance und Cloud Computing Gruppe, Zentrum für Datenverarbeitung, Wächterstraße 76, 72074 Tübingen, Germany

Der offene Austausch von Forschungsdaten ist essentiell für die Kollaboration von Forscherinnen und Forschern und steigert den Erkenntnisgewinn. Im Unterschied zu Plattformen für Datenpublikation oder kollaboratives Arbeiten ermöglichen Science Gateways eine umfangreiche Integration von Storage- und Compute-Infrastrukturen sowie die Anbindung von Repositorien für fachspezifische Communities. Darüber hinaus bieten Science Gateways Module zur Interaktion mit anderen Forscherinnen und Forschern und zur Dokumentation von Prozessen, Know-how und Metadaten. Der Zugang zu Analysewerkzeugen, Storage und Modulen zur Erleichterung von Projekt- und Wissensmanagement erfolgt webbasiert. Durch die breite Integration von Infrastrukturen und Modulen unter anderem zur Analyse, Speicherung und Veröffentlichung von Forschungsdaten decken sie den gesamten Lebenszyklus von Forschungsdaten ab. Zusätzlich können Science Gateways dazu dienen, die öffentliche Sichtbarkeit von wissenschaftlichen Communities und Forschungsgebieten zu verbessern. Das Science Data Center BioDATEN baut ein solches Science

1 Einleitung

Science Gateways dienen einer wissenschaftlichen Community als zentraler und webbasierter Einstiegspunkt zu Ressourcen wie Schulungsmaterialien, Werkzeugen für die Datenanalyse, Datensätzen und erlauben den einfachen Zugang zu verteilten IT-Infrastrukturen für Storage und Computation [1]. Das Science Data Center *Bioinformatics DATa ENvironment* (SDC BioDATEN) baut ein solches Science Gateway auf Basis von HUBzero auf [2]. Der Anspruch eines Science Gateways ist es, Forscherinnen und Forscher über den gesamten Lebenszyklus von Forschungsdaten hinweg zu unterstützen. Darüber hinaus bieten Science Gateways Werkzeuge für die Außendarstellung einer wissenschaftlichen Community wie Newsletter, Blogartikel und Ankündigungen. Ein gutes Beispiel für die

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029705 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13749 (CC BY-SA 4.0)

Umsetzung eines Science Gateways liefert das nanoHUB [3]. Der positive Einfluss von Science Gateways auf die Nutzung verteilter IT-Infrastrukturen konnte bereits nachgewiesen werden. So übertraf die Nutzung verteilter IT-Infrastrukturen über Science Gateways im Jahr 2014 die deren Nutzung via Kommandozeile [4]. Die internationalen Bemühungen um die Verbreitung von Science Gateways werden von dem *Science Gateway Community Institute* (SGCI) und dem *International Workshop on Science Gateways* (IWSG) vorangetrieben [5, 6, 7]. Der Katalog des SGCI listete im April 2021 über 620 Science Gateways aus unterschiedlichen Forschungsfeldern auf [8]. Neben den dort gelisteten Gateways existieren weitere Projekte und Initiativen die sich als Science Gateway beschreiben lassen, wie beispielsweise das CAMPOS Data Cockpit [9], MoSGrid [10] und NFDI4Chem [11]. Im Folgenden werden der gegenwärtige Stand und die weiteren Zielsetzungen für den Aufbau eines Science Gateways für BioDATEN dargestellt.

2 Ein Science Gateway für BioDATEN

BioDATEN setzt für den Aufbau eines Science Gateways für die Bioinformatik-Community in Baden-Württemberg auf HUBzero. HUBzero ist ein Open-Source-Framework und erlaubt durch das integrierte Joomla Content-Management-System (CMS) den Aufbau einer Website für die Außendarstellung des Projekts. Der Vorteil von HUBzero, im Vergleich zur Nutzung eines einfachen CMS, liegt in seinem weitreichenden Funktionsumfang für die wissenschaftliche Community. Durch die Integration von Komponenten, Modulen und Plugins können Funktionen erstellt und eingebunden werden, welche die Nutzerinnen und Nutzer bedarfsgerecht in ihrer Arbeit unterstützen. Darüber hinaus umfasst HUBzero ein vorkonfiguriertes Rechte- und Rollenmanagement, mit dem Berechtigungen für die Administration des Gateways und der Veröffentlichung von Inhalten gesteuert werden können. Ein weiteres, vorkonfiguriertes Modul erlaubt die Erstellung von News-Artikeln und steuert bei Bedarf Zeitpunkt und Dauer der Veröffentlichung. Durch die Kombination von Maßnahmen zur Außendarstellung mit der Integrierbarkeit weiterer Funktionen in einem Science Gateway wird die Community sowohl hinsichtlich ihrer Sichtbarkeit als auch ihrer wissenschaftlichen Arbeit unterstützt. Bei letzterem liegt der Fokus auf frühzeitige Berücksichtigung und Umsetzung des Lebenszyklus von Forschungsdaten.

2.1 Gegenwärtiger Stand

Die öffentliche Webseite des Projekts BioDATEN (*https://portal.biodaten.info*) basiert bereits auf HUBzero. Daneben steht den registrierten Nutzerinnen und Nutzern ein Mitgliederbereich zur Verfügung. Zum Login wird die etablierte *ELIXIR Authentication and Authorization Infrastructure* (AAI) in Kombination mit Keycloak genutzt [12, 13]. Die Anbindung an die ELIXIR AAI hat für die Nutzerinnen und Nutzer den Vorteil, dass diese sich mit den bereits vorhandenen Zugangsdaten der Heimatorganisation am Gateway anmelden können. Der Einsatz von Keycloak erlaubt die Anbindung weiterer Dienste mit gleicher Nutzerbasis. Durch die ELIXIR AAI wird der Login als Single Sign-on

zur Nutzung aller Dienste des Gateways umgesetzt (Abbildung 1). Nach der Anmeldung ermöglicht ein zentrales und konfigurierbares Dashboard eine Übersicht über Gruppen, Projekte und offene Tickets. Zusätzlich können Nutzerinnen und Nutzer über ihre Forschungsarbeit in Blogs berichten, an Kursen teilnehmen und sich in Gruppen austauschen. Besondere Bedeutung kommt der Verwaltung von Projekten zu: Nutzerinnen und Nutzer können eigene Projekte anlegen, Mitarbeiterinnen und Mitarbeiter einladen, Dateien austauschen, To-do-Listen pflegen und Aufgaben zuweisen. Dem jeweiligen Projekt steht eine Speicherplatz-Quota zur Verfügung, die durch die Betreiber angepasst werden kann. Das Gateway ermöglicht auf diese Weise eine zentrale Datenablage und den einfachen Datenaustausch unabhängig von kommerziellen Diensten, E-Mails oder USB-Sticks. In Kombination mit dem Gruppenmodul steht Projekten ein Wiki-System zur Verfügung, um das Wissen im Projekt für alle verfügbar und auf aktuellem Stand zu halten.

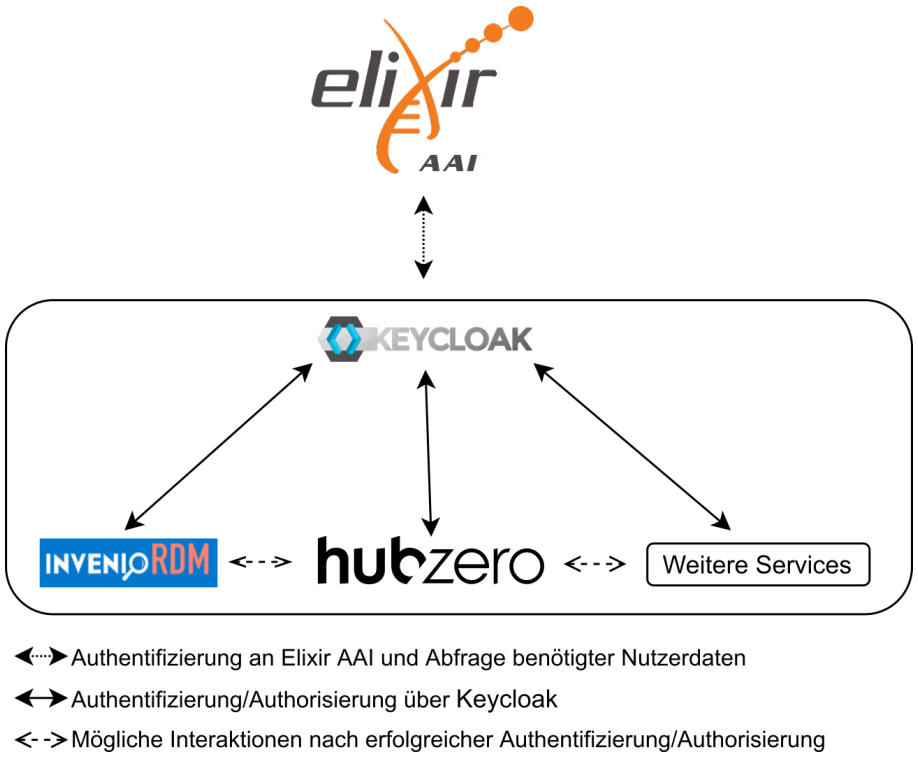


Abbildung 1: Anbindung verschiedener Services via Keycloak und ELIXIR AAI.

2.2 Zukünftige Entwicklung

Die Zielsetzung von BioDATEN liegt in der Unterstützung der Wissenschaftlerinnen und Wissenschaftlern über den gesamten Lebenszyklus der erhobenen Forschungsdaten, von der Datengenerierung über die Annotation mit Metadaten bis zur Datenpublikation. Das zentrale Science Gateway auf Basis von HUBzero dient als erste Anlaufstelle für die Community und bündelt bereits etablierte und neu zu implementierende Ressourcen und In-

frastrukturen. Die Integration etablierter Systeme und bestehender Infrastrukturen trägt zur Nachhaltigkeit des Science Data Centers BioDATEN bei.

Anbindung an die bwHPC-Infrastruktur BioDATEN entwickelt einen Workflow, um direkt bei Lifecycle-Beginn Daten und Metadaten aus der bwHPC-Infrastruktur zu übernehmen [14]. Hierbei dienen Job-Feedback-Skripte des Bioinformatics and Astrophysics Cluster (BinAC) als Ausgangspunkt für die automatische Generierung eines Rumpf-Metadatensatzes, der primär prozessuale Metadaten über die erzeugten Dateien und die verwendeten Ressourcen enthält. Dieser Metadatensatz wird anschließend von den Forscherinnen und Forschern über das Gateway mit Angaben zur Person und wissenschaftlichen Angaben erweitert. Soweit möglich wird dabei automatisch auf bereits vorhandene Metadaten zurückgegriffen, um den Aufwand für die Nutzerinnen und Nutzer so gering wie möglich zu halten. Die notwendigen Angaben richten sich nach DataCite-Schema und wissenschaftlich einschlägigen Schemata, wie beispielsweise dem *Minimum Information about any (x) Sequence* Schema [15, 16]. Durch die Integration des BinAC und vorhandener Metadatenschemata wird ein Forschungsdatensatz früh auf eine mögliche Publikation mit allen benötigten Metadaten vorbereitet.

Datenpublikation Zur Publikation von Forschungsdaten setzt das SDC BioDATEN auf InvenioRDM [17]. Die Publikation soll direkt aus dem Gateway ermöglicht werden, indem die Daten und Metadaten über eine Schnittstelle an InvenioRDM übertragen und dort veröffentlicht werden. Durch den Rückgriff auf das DataCite-Schema liegen alle notwendigen Angaben für die Veröffentlichung und Registrierung eines DOIs vor, so dass Nutzerinnen und Nutzer von der frühen Integration standardisierter Schemata in den Forschungsalltag profitieren können.

Suche Nutzerinnen und Nutzer können die Inhalte des Gateways mithilfe einer entsprechenden Funktion durchsuchen. Zur Realisierung der Suche wird Apache Solr in Kombination mit VuFind eingesetzt [18, 19]. Die Umsetzung der facettierten Suche basiert auf einem eigens erstellten Metadatenschema.

Storage Die Langzeit-Speicherung großer Datenmengen ist eine Herausforderung in datenintensiven Forschungsfeldernwie der Bioinformatik oder Astrophysik. Beim Umgang mit solchen (größtenteils heterogenen) Daten spielt die Wahl eines geeigneten Speichermodells, das den verschiedenen Anforderungen während der Laufzeit eines Projekts genügt, eine wichtige Rolle. In letzter Zeit gewinnt S3-Objektspeicher immer mehr an Bedeutung. Dabei werden Dateien als Objekte abgebildet, die in so genannten Buckets hinterlegt und ähnlich einer Ordnerstruktur mithilfe von Präfixen gruppiert werden können. Im Gegensatz zu traditionellen Dateisystemen müssen sich der Nutzerinn und Nutzer dabei keine Informationen merken, die seine Daten nicht direkt betreffen, wie Laufwerksbuchstabe und den vorausgehenden Pfad, was den Zugriff erleichtert. S3-Objektspeicher erlaubt zudem die Zugriffsbeschränkung auf hinterlegte Objekte und Freigabe dieser unter einer URL.

So können bestimmte Forschungsdaten vor der Publikation auf eine einfache Weise einem wachsenden Kreis an Interessenten freigegeben werden. Das baden-württembergische Landesprojekt *Storage for Science* (bwSFS) stellt S3-Objektspeicher ebenso bereit wie die Cloud des *Deutschen Netzwerks für Bioinformatik-Infrastruktur* (de.NBI) [20, 21]. Den Nutzerinnen und Nutzern wird mittels Gateway Objektspeicher über die de.NBI Cloud angeboten, dessen Nutzung durch die Hinterlegung von S3-Credentials im Gateway erfolgt. Durch diese Integration wird ein wichtiger Beitrag zur Nachhaltigkeit des SDC BioDATEN geleistet und der Nutzen für die Forscherinnen und Forscher erhöht.

Integration von Tools Die enge Anbindung an die bwHPC/BinAC-Infrastruktur, de.NBI-Cloud und die Integration von Galaxy Workflows [22] ermöglicht die Verwendung und Integration vorhandener Workflows und Tools zur Datenanalyse. Die web- und UI-basierte Bereitstellung dieser Tools und Workflows entlang des Lebenszyklus von Forschungsdaten über das Gateway ist ein weiteres Ziel des SDC BioDATEN.

3 Zusammenfassung

Science Gateways ermöglichen den Aufbau eines zentralen Einstiegspunkts für die wissenschaftliche Community zu Diensten, Daten und Materialien. Darüber hinaus erlauben sie den Aufbau einer Website für die Verbreitung von News, Blogs und Öffentlichkeitsarbeit. Ein Beispiel für die Integration von Diensten und verbesserte Sichtbarkeit einer Community ist das nanoHUB [3]. Im Rahmen des Science Data Centers BioDATEN wird ein solches Science Gateway auf Basis von HUBzero aufgebaut, das Wissenschaftlerinnen und Wissenschaftler über den gesamten Lebenszyklus von Forschungsdaten hinweg unterstützen soll. Die Integration in den wissenschaftlichen Alltag erfolgt durch den Rückgriff auf etablierte und verbreitete Infrastrukturen wie die ELIXIR AAI für Login und Accountgenerierung sowie auf de.NBI Cloud und bwHPC/BinAC für die Bereitstellung von Storage und Computation. Die Anbindung an den BinAC ermöglicht die automatische Generierung eines Rumpf-Metadatensatzes, welcher im Laufe des weiteren Lebenszykluses vervollständigt wird. Durch die frühzeitige Erhebung notwendiger Metadaten wird die Veröffentlichung in den späteren Phasen des Lebenszyklus vorbereitet. Die geplante Integration von Analysewerkzeugen und Workflows wird einen weiteren Beitrag zur Integration des Science Gateways in den wissenschaftlichen Alltag leisten und die Unterstützung und Umsetzung des Lebenszyklus von Forschungsdaten weiter ausbauen.

Förderung

Das Science Data Center BioDATEN wird vom Ministerium für Wissenschaft, Forschung und Kunst Baden -Württemberg aus Mitteln der Landesdigitalisierungsstrategie digital@bw gefördert.

Literaturverzeichnis

[1] Nancy Wilkins-Diehr, Michael Zentner, Marlon Pierce, Maytal Dahan, Katherine Lawrence, Linda Hayden, and Nayiri Mullinix. The science gateways community institute at two years. In *Proceedings of the Practice and Experience on Advanced Research Computing*, PEARC '18, New York, NY, USA, 2018. Association for Computing Machinery. https://doi.org/10.1145/3219104.3219142.

[2] Michael McLennan and Rick Kennell. HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering. *Computing in Science Engineering*, 12(2):48–53, 2010. https://doi.org/10.1109/MCSE.2010.41.

[3] nanoHUB. https://nanohub.org/. Zuletzt abgerufen am 29.07.2021.

[4] Katherine A Lawrence, Nancy Wilkins-Diehr, Julie A Wernert, Marlon Pierce, Michael Zentner, and Suresh Marru. Who cares about science gateways? a large-scale survey of community use and needs. In *2014 9th Gateway Computing Environments Workshop*, pages 1–4. IEEE, 2014.

[5] IWSG - International Workshop on Science Gateways. https://sites.google.com/site/iwsglife. Zuletzt abgerufen am 27.04.2021.

[6] Katherine A. Lawrence, Nayiri Mullinix, Maytal Dahan, Linda Hayden, Marlon Pierce, Nancy Wilkins-Diehr, and Michael Zentner. How the science gateways community institute supports those who are creating websites to access shared resources. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)*, PEARC '19, New York, NY, USA, 2019. Association for Computing Machinery. https://doi.org/10.1145/3332186.3333256.

[7] SGCI - Science Gateways Community Institute. https://sciencegateways.org/. Zuletzt abgerufen am 27.04.2021.

[8] Science Gateways Catalog. https://catalog.sciencegateways.org/#/home. Zuletzt abgerufen am 27.04.2021.

[9] M. Finkel, A. Baur, T. K. D. Weber, K. Osenbrück, H. Rügner, C. Leven, M. Schwientek, J. Schlögl, U. Hahn, T. Streck, O. A. Cirpka, T. Walter, and P. Grathwohl. Managing collaborative research data for integrated, interdisciplinary environmental research. *Earth Science Informatics*, 13(3):641–654, Sep 2020. https://doi.org/10.1007/s12145-020-00441-0.

[10] Jens Krüger, Richard Grunzke, Sandra Gesing, Sebastian Breuers, André Brinkmann, Luis de la Garza, Oliver Kohlbacher, Martin Kruse, Wolfgang E. Nagel, Lars Packschies, Ralph Müller-Pfefferkorn, Patrick Schäfer, Charlotta Schärfe, Thomas Steinke, Tobias Schlemmer, Klaus Dieter Warzecha, Andreas Zink, and Sonja Herres-Pawlis. The mosgrid science gateway – a complete solution for molecular simulations. *Journal of Chemical Theory and Computation*, 10(6):2232–2245, 2014. PMID: 26580747. https://doi.org/10.1021/ct500159h.

[11] Nicole Jung, Steffen Neumann, Oliver Koepler, Felix Bach, Christian Popp, Sonja Herres-Pawlis, Johannes Liermann, Matthias Razum, and Christoph Steinbeck. NFDI4Chem – Infrastruktur für den digitalen Wandel in der Chemischen Forschung. *Bunsen-Magazin*, 2021. URL: https://bunsen.de/bmo/nfdi4chem, https://doi.org/10.26125/r978-6f93.

[12] Mikael Linden, Michal Prochazka, Ilkka Lappalainen, Dominik Bucik, Pavel Vyskocil, Martin Kuba, Sami Silén, Peter Belmann, Alexander Sczyrba, Steven Newhouse, et al. Common elixir service for researcher authentication and authorisation. *F1000Research*, 7, 2018.

[13] Keycloak - Open Source Identity and Access Management. https://www.keycloak.org/. Zuletzt abgerufen am 27.04.2021.

[14] bwHPC. https://www.bwhpc.de/. Zuletzt abgerufen am 29.07.2021.

[15] DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4. 2021. URL: https://schema.datacite.org/meta/kernel-4.4/, https://doi.org/10.14454/3W3Z-SA82.

[16] Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature biotechnology*, 29(5):415–420, 2011.

[17] InvenioRDM. https://inveniosoftware.org/products/rdm/. Zuletzt abgerufen am 27.04.2021.

[18] Apache Solr. https://solr.apache.org/. Zuletzt abgerufen am 27.04.2021.

[19] VuFind. https://vufind.org/vufind/. Zuletzt abgerufen am 27.04.2021.

[20] bwSFS - Storage for Science. https://www.alwr-bw.de/bwsfs/. Zuletzt abgerufen am 27.04.2021.

[21] Peter Belmann, Björn Fischer, Jan Krüger, Michal Procházka, Helena Rasche, Manuel Prinz, Maximilian Hanussek, Martin Lang, Felix Bartusch, Benjamin Gläßle, et al. de. NBI Cloud federation through ELIXIR AAI. *F1000Research*, 8, 2019.

[22] E. Afgan, D. Baker, B. Batut, M. van den Beek, D. Bouvier, M. Cech, J. Chilton, D. Clements, N. Coraor, B. A. Grüning, A. Guerler, J. Hillman-Jackson, S. Hiltemann, V. Jalili, H. Rasche, N. Soranzo, J. Goecks, J. Taylor, A. Nekrutenko, and D. Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res*, 46(W1):W537–W544, 07 2018.

Data Stewards as ambassadors between the NFDI and the community

Dirk von Suchodoletz, Timo Mühlhaus, Dominik Brilhaus, Hajira Jabeen, Björn Usadel, Jens Krüger, Holger Gauza and Cristina Martins Rodrigues

The NFDI consortium DataPLANT focusing on fundamental plant research, provides data stewards as a core element of its strategy for dissemination of common standards, concepts of research data management, and workflow services. Data stewards play a special hinge role between service providers, individual researchers, groups, and the wider community. They help to bridge the gap between the scientists working in the lab and the technical solutions and services. Project groups and individual researchers will profit from direct support in their daily tasks ranging from data organization to the selection and continuous development of the proper tools, workflows and standards. This leads to a community-wide dissemination and development of data management strategies especially suited to support plant research. In particular, the convergence of researcher and repository requirements is of great importance, and crucial for the success of RDM in general. Additionally, the data steward service concept of DataPLANT is designed for effective capacity building and training to ensure sustainability in the research landscape.

1 Motivation – What is a data steward?

The slow adoption and dissemination of common standards, the concepts of research data management, and workflow services is still a hindrance to collaboration, data sharing-and-reuse, as well as open science in many scientific communities [1, 2]. The responsible and informed handling of research data is part of good scientific practice [3, 4]. The central goals of DataPLANT [5, 6] are, to provide appropriate infrastructure and workflows, and to train researchers of varying experience towards data stewardship and research data management (RDM). In the long run, such qualification measures should be included in the relevant curricula. The task for the support and community domain of the project is to prepare tailored content for the various data management mechanisms over the entire lifecycle. Hence, data stewards are experienced individuals with strong communication skills, expertise in plant biology, bioinformatics tool development and familiar with heterogeneous infrastructure. Data stewards operate at the core of DataPLANT and fulfill

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029712 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13750 (CC BY-SA 4.0)

a special hinge role between the various stakeholders and the wider community to bridge the gap between researchers and technical infrastructure (see Figure 1). DataPLANT introduces a community-integrative approach of data stewardship that is supported by internally governed and associated data stewards with aligned functions. Internally governed data stewards are funded and orchestrated by the NFDI consortium itself. With a focus on DataPLANT's core mission, they support multiple consortia and individual research groups. This allows the DataPLANT consortium to provide on-site support for the individual project partners and participants either in person or remotely. Associated data stewards are funded by and seated at DataPLANT project partners such as collaborative research centres, typically familiar with local scientific workflows and RDM practices. The common goal of data stewards is to integrate institutional and community RDM concepts as well as aligning the standards in the domain and infrastructural support environments both on a practical and operational level [7]. This bidirectional communication fosters to interlink RDM activities within the community.

2 Contribution to the community

Data stewards target the community on different levels and provide specifically tailored data management strategies that enable the community to use existing standards and facilitate the use of technology and infrastructure for data management [8]. Through the community-integrative model, they interact directly with core facilities, research groups and individual researchers. As the major (*omics) data providers, core facilities play a special role in the development and dissemination of DataPLANT. They are experts in measurement technologies that are central to the community and know most about method-specific metadata and infrastructure requirements. Due to their community network and diverse client base, they take a multiplier role, allowing an indirect reach out to participants, plus possible links to other scientific communities and NFDI consortia. Data stewardship of core facilities thus has a manifold effect by finding an RDM solution that suits the facility and improving user-friendliness for clients who use the same DataPLANT mechanisms established in other facilities. Research groups profit from data stewards in multiple ways. Data stewards advise on data management and standards related questions of a grant application or during the setup phase of a research project. Project managers and principal investigators can request information on the ongoing activities in standards development. In addition, data stewards offer proven and well established procedures to handle research data aiming at the improvement of digital lab organisation according to the FAIR data principles [9].

2.1 Dissemination and development of data management strategies

A holistic planning phase including a data management plan (DMP) is a prerequisite of a successful grant application and project start. Together with the participants, data stewards develop a plan fitting their project requirements. The DMP of the proposed project

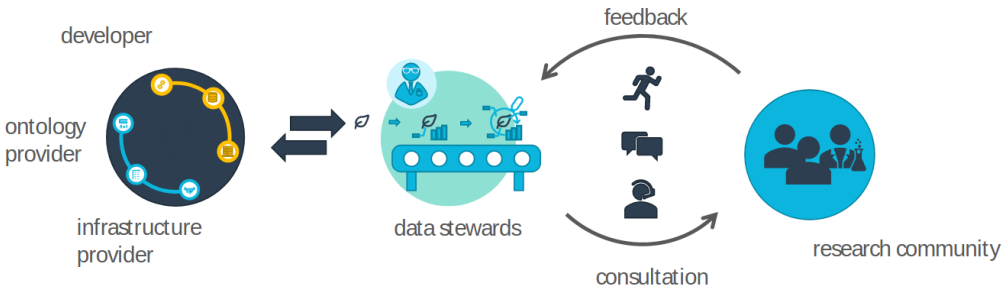


Figure 1: The hinge role of DataStewards between the community and infrastructure.

estimates the required funds and compute resources as well as the amount for data to be stored and published in the long run. DataPLANT employs a data-centric approach towards FAIRness of plant biological data. At the heart of this approach lies the ARC (annotated research context) [10] as the data packaging format for research objects, which expands the widely established metadata grammar of ISA [11] to enrich the ARC with content and provides further context e.g. on the workflows and tools used. Its flexible and open nature guarantees long-term accessibility and sustainability. The central DataPLANT mechanisms of data stewardship and data management planning evolve around the ARC environment, accessible directly or through the DataPLANT Hub [5]. Data stewards help developing the ARC environment to offer a common suite of suitable data formats, standards, and repositories for an increasing range of data types and integrate associated tools and workflows for data processing and publication. These developments are elaborated in the DMP and enable the community to use the DataPLANT technologies and infrastructures and facilitate data publication in community-specific repositories.

2.2 Converging researcher and repository requirements

As the sustainability of DataPLANT depends on the convergence between its data-centric approach and the current state of the individual plant science communities, data stewards participate in implementing suitable operating procedures into the participant groups. Proper metadata description is the basis for data findability and accessibility. Data stewards support a structured collection of metadata for common experimental and computational workflows by drafting metadata templates and guiding participants on creating templates or adapting existing ones to their needs. They foster compliance with the submission requirements of end-point repositories and associated metadata standards and minimal reporting guidelines. This ensures that metadata is (readily) usable independent of DataPLANT services. To facilitate the collection of metadata at its point of emergence, data stewards support the FAIRification of the whole scientific process – from

experiment planning to data acquisition and processing. The light-weight standardization convention of the ARC environment can easily be adapted to or implemented into daily laboratory routines. Data stewards help the participants to develop suitable solutions for data storage and sharing, for the lab organisation or to adapt local software packages. Through the development of digital workflows such as Galaxy [12] and Nextflow [13], they enable access to necessary infrastructures and harness remote resources. Data FAIRness and preparation of high quality ARCs for sharing and publication is assured by active participation of data stewards during the iterative cycles of metadata annotation and data handling. The development of ARCs is a bidirectional, iterative effort. Data stewards continuously monitor and evaluate participant feedback on tools and services. This process of incorporating case-by-case specific requirements into a widely adoptable consensus, shapes ARC's flexibility and the further route of development of tools and services. Retracing participant input and adaptations will propel the development of the ARC environment and facilitates to address frequently missed information in metadata templates, fragmentary ontologies, and existing standards. Furthermore, the direct and timely interaction with the active research community enables the flexible integration of future developments, including new techniques and data types.

3 Capacity building

Significant dissemination to the community is achieved through a comprehensive training program that introduces DataPLANT services and tools as well as general data literacy and analysis capabilities to the researcher [14]. Individual consultation of participants will be complemented with on-site workshops for research groups adapted to the needs of the community and the stage of association with DataPLANT. During the onboarding phase, the activities cover general data management practices and familiarization with DataPLANT tools and services. In-depth expertise on specific topics is elaborated with respective stakeholders in the participating groups. For a continuous exchange between the data stewards and the research groups, DataPLANT encourages the appointment of data management representatives (DMRs), who – similar to core facilities – act as relevant multipliers. They take a bidirectional role by (i) spreading knowledge on data management, standards and services in their groups and (ii) reporting back common hurdles and requirements. Both DMRs and core facility heads will specifically be addressed and qualified by DataPLANT data stewards. In addition to workshops, a continuously updated knowledge base provides teaching materials, tutorials for tools, services and best practices that reflect the development of DataPLANT. The ultimate goal of DataPLANT is to enable the researcher to produce ARCs without or only minimal support by the data steward. Training is not exclusive to participants, but likewise enables the continuous qualification of data stewards ("train the trainer"). Data stewards attend training and workshops to keep track of all relevant developments in the field as well as international activities and achievements. In regular meetings and through a central data stewardship knowledge base, data stewards exchange on best-practices, qualify on new standards, learn on legal issues, updates on extended modified ontologies and metadata schemas as well

as on potential new workflow and software options. FAIRification use cases at the participants' sites are shaped into general best practices and common data stewardship tasks. This rich support resource will particularly be useful to freshly onboarding data stewards, but may also be transferred into the plant science or NFDI community to set new standards for data stewardship in general. Besides disseminating DataPLANT mechanisms, the data stewards consulting and qualification capacities need to be extended over time. This challenge to personnel development is shared with other consortia in the NFDI as well and addressed through cross-cutting activities [15].

4 Data steward dispatch model

Substantial data stewardship time is allocated to consulting services and capacity building, in addition to self-qualification and dissemination. Data steward support can be requested in any stage of the research process. The group of data stewards maintains connections with the community as they accompany scientists and research groups in the various stages of the research data life cycle. Until the data stewardship is institutionalized, we follow a distribution model to optimize leveraging effects in the community. Therefore, efficient scheduling of resources suggests focusing the support on data generating hubs within the community. However, in order to follow the consortium's objectives of transparent communication and broad user involvement, a balanced mechanism that ensures fair allocation of resources is envisioned with the following dispatch model:

1. First time request is (automatically) granted but goes with conditions (e.g commitment to the NFDI objectives, provisioning of the data to the NFDI).
2. FairShare: Available data stewards hours are divided by the number of requests. Additionally, 30% are reserved for future requests.
3. Later, the allocation could take input parameters like the size of a research group, the provision of additional resources (e.g grant money, material costs of their accepted grant) and bonus points.
4. The bonus points are allocated to groups or individuals after quality assessment of the provided data, and these points can be translated into additional hours or resource allocation using an evaluation system.
5. In the future extra points may be awarded for exemplary data sets published and referenced.
6. During phases of higher loads, the multiple incoming requests can be ordered by waiting time. Groups which interacted more recently with a data steward will wait comparably longer than researchers who used their services a longer time ago. A weighted queue can be maintained for high load, less resource time-period.

The preliminary strategy combines factors of fair distribution of resources with incentive schemes to improve the metadata quality and FAIRness of data sets. Given that it is

challenging to know the demand in advance, it is anticipated that this set of rules will be further polished and adjusted according to the existing resources and data management demands from the community. Special requests, conflicts which are not solvable on that layer will be passed on to the Senior Management Board to decide. Additionally, this body takes steering responsibility to adapt the distribution if necessary, after a ramp-up period followed by an evaluation of the process. We assume a rising demand from the wider community.

5 Sustainability and outlook

To foster a broader adaptation of DataPLANT within the community and to grow with the demand for new participants, data stewardship should be complemented by co-funding or own personnel of new members. If a broad range of future individual project proposals or large-scale projects like collaborative research centres plan for personnel and infrastructure services directly by contributing to the NFDI, a sustainable financing and reimbursement model can be created benefiting the broader community. Small projects can then receive qualified support from a range of experts according to their contribution. Data stewards in large projects get integrated into a broadly qualified team working on cutting-edge research and workflows. The consortium's and NFDI's governance structures ensure the orientation of the data stewards' support on the actual demands of the community.

Acknowledgements

CEPLAS has been supported by Deutsche Forschungsgemeinschaft within the Excellence Initiative (EXC 1028) and under Germany's Excellence Strategy – EXC 2048/1 – project 390686111. We acknowledge support for DataPLANT 442077441 through the German National Research Data Initiative (NFDI 7/1) and the Science Data Center BioDATEN which is supported by the Ministry of Science, Research and Art Baden-Württemberg.

ORCID IDs

- Dirkvon Suchodoletz https://orcid.org/0000-0002-4382-5104
- Timo Mühlhaus https://orcid.org/0000-0003-3925-6778
- Dominik Brilhaus https://orcid.org/0000-0001-9021-3197
- Hajira Jabeen https://orcid.org/0000-0003-1476-2121
- Björn Usadel https://orcid.org/0000-0003-0921-8041
- Jens Krüger https://orcid.org/0000-0002-2636-3163

- Holger Gauza https://orcid.org/0000-0003-0191-3680
- Cristina Martins Rodrigues https://orcid.org/0000-0002-4849-1537

Bibliography

[1] Sara Rosenbaum. Data governance and stewardship: designing data stewardship entities and advancing data access. *Health services research*, 45(5p2):1442–1455, 2010. https://doi.org/10.1111/j.1475-6773.2010.01140.x.

[2] Ge Peng. The state of assessing data stewardship maturity–an overview. *Data science journal*, 17, 2018. https://doi.org/10.5334/dsj-2018-007.

[3] Deutsche Forschungsgemeinschaft. DFG guidelines on the handling of research data. https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_research_data.pdf, 2015. [Online; accessed 28-April-2021].

[4] Guidelines for Safeguarding Good Research Practice. Code of Conduct, September 2019. Available in German and in English. https://doi.org/10.5281/zenodo.3923602.

[5] DataPLANT NFDI webpage. https://nfdi4plants.de/. [Online; accessed 16-April-2021].

[6] Dirk von Suchodoletz, Timo Mühlhaus, Jens Krüger, Björn Usadel, and Cristina Martins Rodrigues. Dataplant – ein nfdi-konsortium der pflanzen-grundlagenforschung. *Bausteine Forschungsdatenmanagement*, (2):46–56, 2021. https://doi.org/10.17192/bfdm.2021.2.8335.

[7] Dorothea Iglezakis and Sibylle Hermann. 4.4 disziplinspezifische und – konvergente fdm-projekte. In *Praxishandbuch Forschungsdatenmanagement*, pages 381–398. De Gruyter Saur, 2021. https://doi.org/10.1515/9783110657807.

[8] Daniela Hausen, Jessica Rosenberg, Ute Trautwein-Bruns, and Annett Schwarz. Data stewards an der rwth aachen university–aufbau eines flexiblen netzwerks. *Bausteine Forschungsdatenmanagement*, (2):20–28, 2020. https://doi.org/10.17192/bfdm.2020.2.8278.

[9] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[10] C. Garth, J. Lukasczyk, T. Mühlhaus, B. Venn, , K. Glogowski, C. M. Rodgrigues, and D. von Suchodoletz. Immutable yet evolving: ARCs for permanent sharing in the research data-time continuum. 2021.

[11] Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor, Oliver Hofmann, Hong Fang, Steffen Neumann, Weida Tong, Linda Amaral-Zettler, et al. Toward interoperable bioscience data. *Nature genetics*, 44(2):121–126, 2012.

[12] Jorrit Boekel, John M Chilton, Ira R Cooke, Peter L Horvatovich, Pratik D Jagtap, Lukas Käll, Janne Lehtiö, Pieter Lukasse, Perry D Moerland, and Timothy J Griffin. Multi-omic data analysis using galaxy. *Nature biotechnology*, 33(2):137–139, 2015.

[13] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017. `https://doi.org/10.1038/nbt.3820`.

[14] Sarah Jones, Robert Pergl, Rob Hooft, Tomasz Miksa, Robert Samors, Judit Ungvari, Rowena I Davis, and Tina Lee. Data management planning: How requirements and solutions are beginning to converge. *Data Intelligence*, 2(1-2):208–219, 2020. `https://doi.org/10.1162/dint_a_00043`.

[15] Frank Oliver Glöckner, Annette Pollex-Krüger, Kirsten Toralf, Juliane Fluck, Birgitta König-Ries, Chris Eberl, Torsten Schrade, Anton Güntsch, Birgit Gemeinholzer, Thomas Schörner-Sadenius, et al. Berlin declaration on nfdi cross-cutting topics. Technical report, Jülich Supercomputing Center, 2019.

Immutable yet evolving: ARCs for permanent sharing in the research data-time continuum

Christoph Garth, Jonas Lukasczyk, Timo Mühlhaus, Benedikt Venn, Jens Krüger, Kolja Glogowski, Cristina Martins Rodrigues and Dirk von Suchodoletz

Scientific research data is often viewed as monolithic and immutable – once created and processed, it is published in archives for transparency and reproducibility. In this paper, we argue that research data sets should by default be viewed as dynamic and evolving, and should be created, managed, and curated by processes that mirror the development of software rather than via a publication-focused approach. We propose *Annotated Research Contexts* (ARCs), a lightweight basis for such processes, and illustrate how ARCs will assist research data management in plant biology, within a framework developed by the NFDI consortium DataPLANT.

1 Introduction

Research data management has received a growing amount of attention as many scientific disciplines are increasingly data-driven. For example, the *FAIR* principles [1] state that data should be made available in a findable and accessible manner, i.e., in open, publicly archives, and be interoperable and reusable, i.e. published in non-proprietary formats and annotated with metadata that describes contents and provenance. Mounting adoption of FAIR requirements by funding agencies — in particular, for publicly funded research — has greatly benefited overall quality, reuse, and sharing of research data. As a consequence, a plethora of open formats for FAIR data publication were developed in the past decade, for example the *schema.org*-based *Research Object* format [14] specification that is generic and domain agnostic, among many others.

While the FAIR principles define desiderata for *published* research data, there is an implicit presumption that data is published exactly once, and immutable henceforth. In practice, major incentives to publish or share research data are typically coupled to research cycles [2], e.g. sending a research manuscript for review to a publisher who requires

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029769 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13751 (CC BY-SA 4.0)

data co-publication, or publishing all collected data at the end of a project. Focusing, or limiting, research data management processes to these two outcomes forgoes a host of opportunities that arise from a more continuous, dynamic approach. We will, in the following, argue that typical processes found in software development align well to the goals of research data management (RDM), and thus could be adapted to and adopted in an RDM context. We outline the current discussion on and status of the efforts of the DataPLANT consortium [3], within the framework of the National Research Data Infrastructure, to provide the technical basis — in the form of *Annotated Research Contexts* (ARCs) — for implementing corresponding processes for the plant research community. Our approach focuses not only raw data gathered during experiments, but also considers analysis routines and derived data products within the context of research data management.

2 RDM and Software Development

Software development is a well-studied aspect of Computer Science, and exhibits a comprehensive set of best practices to structure and manage the process of developing software systems at all scales, from small-scale personal software to largest-scale, multi-entity projects. These guidelines have emerged over half a century of practice. Others have written at length and much more formally about this [4]; here, it is not our intention to provide a detailed description. However, we single out a few broad insights and considerations that map well to the requirements of RDM, and from this derive requirements for a potential technical basis for RDM processes.

Holistic life cycle. Software development follows the goal completing a software product: a problem is analyzed, a design is derived, and an implementation is developed and released. This process is often cyclic, i.e. following a first release, requirements are reanalyzed, new features are designed, etc. However, the ultimate goal of each iteration of the cycle, i.e., producing a release, is anticipated throughout all steps. This has led to the development of techniques such as *unit testing* and *continuous integration*, which ensure that software is syntactically and semantically correct throughout the development phase. This also ensures that an implementation can be released at any point in time, e.g., if additional testing is required, or on a regular schedule.

RDM in most cases also follows a cycle: from a specific research question, experiments are performed to gather data, which is subsequently analyzed; results are published, and new or modified research questions are identified, restarting the cycle [2]. While publication of the data typically occurs at least at the end of each cycle, it appears beneficial to anticipate the need for publication already in earlier stages of the process, namely during data gathering and analysis. To give an example, the FAIR principles require that data is correctly annotated with metadata, such that others may interpret and reuse it. While it is absolutely feasible to perform this annotation just in time for publication, it often is a

substantial burden of work, and in practice not performed with ideal diligence, resulting in metadata of insufficient quality [5].

Borrowing from the software development playbook, we advocate *continuous annotation*, i.e. spreading the data annotation throughout the RDM life cycle by ensuring that data is annotated as it is gathered, and annotations are carried through analysis. More generally, we argue that a data set should be annotated completely at every stage of its life cycle. The same applies to *continuous reproducibility*; instead of focusing a large effort on making analysis reproducible just before publication, analysis on a data set should always be reproducible.

At first glance, treating a large burden for a sequence of smaller burdens appears as a zero-sum game. However, in collaborative settings, the overall effort is quickly reduced as the number of collaborators grows. Moreover, ensuring that data is always well-annotated and reproducible affords opportunities for unplanned, ad hoc collaboration without incurring additional overhead.

Rapid and Scalable Collaboration A key aspect of software development is scalability: similar processes are used whether the development team is small or large. Effective collaboration is achieved using *distributed version control systems* such as e.g. Git, Mercurial, DARCS, etc., that store an evolving version of source code and arbitrate changes made by many developers.

It appears useful to use the collaboration opportunities afforded by version control also in the context of RDM [6], where collaboration also occurs on a variety of scales: from the individual researcher performing an experiment in isolation to a large multi-national inter-group collaborative effort, all forms of collaboration should be possible with minimal friction, while simultaneously ensuring that data is always in a consistent, well-annotated and reproducible state. Version control as a documentation and arbitration mechanism is ideally suited to this purpose, as it ensures an atomic and unambiguous history, without requiring diligence of participating researchers. Therefore, our goal is to amend ad hoc collaboration mechanisms (file sharing, email, etc.) by a systematic approach rooted in decentralized version control [7].

Automation A further ingredient in software development is that all aspects of development that can be automated, are automated. Two prominent examples are quality control (via continuous integration testing, see above) and adherence to common conventions (e.g., code formatting). RDM can borrow from this regarding a multitude of aspects. For example, the quality of metadata and reproducibility can be automatically assessed to a certain degree [8], allowing to quickly bring deficiencies to the attention of researchers and ensure continuous annotation and continuous reproducibility. Using these mechanisms, reviewers can rely on automation to ensure that a data set is in an advertised state. Finally, automatic annotation, where possible (e.g., for data products from computational analysis) can free researchers from manual annotation.

Provenance Often, software development requires a detailed understanding of the provenance of a particular aspect of an implementation. The history afforded by version control makes this straightforward. In the context of RDM, the detailed recording of changes through version control history, identifying the researcher responsible for each change, allows an in-depth understanding of who contributed in which manner to a data set, and in what form. While this information is crucial to collect in the first place, e.g. to be able to assign credit to all contributors to a data set, version control stands to effectively automate this previously tedious and manual process.

Cathedral vs. Bazaar In his seminal and highly influential essay, Raymonds [9] discusses two models for releasing source code during development, which are similarly applicable to the sharing of research data. In the *Cathedral* model, corresponding to a top-down approach, data would be made available at discrete releases, for example upon publication of a paper that relies on it. In contrast, in the *Bazaar* model, in a bottom-up manner, research data is constantly shared and evolved in full view of the public. His central thesis in contrasting these two models, adapted to an RDM context, is that data that is permanently shared during its evolution in a bazaar-type approach can benefit strongly from increased opportunities for collaboration and rapid identification of problems, such as lacking reproducibility or metadata annotation. Again, a version control approach to RDM may improve permanent sharing of research data and act as a catalyst toward increased data quality.

Open Standards It has been long recognized that an adherence to open standards, tools, and community-specific conventions is highly beneficial towards encouraging wide and opportunistic collaboration. A core consideration to support RDM technically is hence to rely on open and established tools with a wide user base, to the largest possible extent. For the case of version control, for example, this suggests to use the widely-used Git. Relying on overly specific or restricted solutions stands to exclude (opportunistic) contributions from others ("walled garden").

3 Annotated Research Context

A core objective in DataPLANT is to provide a technical basis that implements software development-inspired processes and integrates these into RDM workflows. While many technical solutions exist, the barrier of entry is significant due to their substantial complexity (e.g., version control systems); furthermore, the processes themselves must be negotiated, understood, and applied with stringency. This places a substantial burden on researchers and is a prohibitive time investment. Therefore, an additional objective of DataPLANT is to define and standardize easy-to-use RDM procedures and their technical realization, specifically targeted at the needs of the fundamental plant research community. In the following, we describe the *Annotated Research Context* (ARC) that DataPLANT is developing toward these goals.

ARC structure An ARC is a collection of files and folders, laid out in a specific schema, following the ISA model [10], which distinguishes *investigation*, *study*, and *assay* and is ubiquitous in the plant research community. ARCs are intended to capture all research data pertaining to a single study within a larger investigation; the scope of a single ARC is intended to scale from encompassing the research data of smaller research projects such as a single publication or a thesis, to larger lab-wide or even multi-lab investigations. We reserve a more detailed description of the ARC file structure, metadata schema, and workflow description for future work, once an initial specification is finalized.

ARCs are Version Control Repositories While based on a specific file structure, ARCs are version controlled repositories [6] and capture their evolution in the form of a version history. Version control system operations are not exposed directly to users, as their complexity is overwhelming to most non-experts and full flexibility is not needed. Rather, we define an **update** operation that records the current state of all files in the history. In addition to allowing an understanding of the provenance of an ARC, update operations are canonical points for automation of quality control and other tasks.

Git is a suitable initial choice for the underlying version control system, due to several factors. It is very lightweight and decentralized, allowing researchers to operate individual ARCs without overhead or requiring a centralized resource. Furthermore, Git is technically mature, widely adopted, and very well documented. Others have employed Git for RDM with same reasoning (cf. Section 4).

Collaboration with ARCs ARCs can be easily used collaboratively shared through existing Git repository hosting mechanisms, such as e.g. Github or a Gitlab server instance. To keep complexity manageable, low-level Git operations *push* (propagating local changes to a remote repository) and *pull* (merging remote changes with the local history) are replaced by a **sync** operation which combines these two. If conflicts occur, i.e., when remote changes would overwrite local changes or vice versa, researchers can opt to overwrite either local changes or remote changes, or address the conflict manually.

Multiple researchers can collaborate on a single ARC by selecting a hosting facility (e.g. public or lab-specific) and placing a central copy of their ARC there. They can concurrently modify and **update** their respective ARC copies, and **sync** to propagate their own changes to the central ARC and obtain the other researchers' changes. Moreover, ARCs will support an **import** operation to selectively reuse parts of one ARC (e.g. a data set or workflow) in another, while retaining the capability of sending changes back to the original ARC.

Within DataPLANT, the **DataPLANT** hub will provide a central hosting facility for ARC Git repositories, open to all researchers participating in the consortium and their collaborators, further reducing the burden of initiating collaboration with others.

ARC Automation Towards ensuring continuous annotation and continuous reproducibility, the version control inherent in ARCs could serve as a vehicle to easily implement automation for each new version. Currently, automation is considered in checking for adherence to file schema, metadata completeness and quality, reproducibility, among other aspects.

ARC Sharing and Publication Publication of ARCs is fairly straightforward [11], especially using services that already allow formal publication of Git repositories, such as e.g. Zenodo.org or the Open Science Foundation (osf.io). The DataPLANT hub will also offer publication facilities. ARCs, with full modification history, will be archived long-term and accessible via DOI.

To avoid lock-in into a specific ecosystem of tools and ensure that ARCs play well in the diverse ecosystem of research data publication facilities, it is planned to offer automatic conversion of ARCs into other ubiquitous formats for research data archival, such as e.g., *RO Crates* [12] or formats compatible with OpenAIRE / H2020 Data Management Plans. Unless explicitly hosted on private repository servers, ARCs are shared by default, in real time — no additional steps are needed to share an ARC and its evolution with others.

Implementation. ARCs and the RDM processes they support are conceptually straightforward and rely on open standards and tools (e.g., Git). To support the RDM procedures outlined above without having to resort to low-level operations (e.g. *git push*) dedicated support software is currently under development, initially as a command line tool, and later as a GUI client and web-based interface. Several alternatives are currently under consideration.

4 Discussion & Conclusion

Naturally, others have pursued a similar approach to the one we propose here, and unsurprisingly come up with similar designs to support RDM in daily use [11]. We will here briefly consider two representative concepts (resp. tools) to illustrate similarities and differences to our approach.

The notion of fundamentally conducting RDM on the basis of version control has been investigated in a broad manner [6]. For example, the powerful and domain-agnostic *Datalad* tool [15] is based on a Git variant (git-annex) to manage and describe the evolution of research data and enable low-friction collaboration, and supports a variety of generic operations. However, this generality incurs complexity: clear processes are not defined. Defining and negotiating these between collaborators can be a substantial burden towards effective collaboration. In contrast, the ARC process trades generality for convenience, focusing on few important operations relevant to plant research workflows. For example, while ARCs are always self-contained, Datalad provides a more general form of linking

to other data sets, which however makes it difficult to check data set completeness (e.g., for archiving) or provenance. The *DVC* tool [13] takes a more holistic view and adds tailored processed for Machine Learning to versioned data management. It is related to our approach, however, but its specificity prohibits an application in plant research.

ARCs are designed as a support vehicle for RDM processes in plant research, and borrow in design from software development best practices. The ARC concept at heart embodies the idea of permanent sharing of research data. Fundamentally, this implies a strong reliance on open standards and tools, and version control as a candidate supporting technology. In this manner, ARCs are a key driver towards fulfilling DataPLANT's mission: to enable plant researchers to participate as first-class citizens in a large, vibrant, and growing ecosystem of data-driven research.

Acknowledgements

We acknowledge support for DataPLANT 442077441 through the German National Research Data Initiative (NFDI 7/1).

ORCID IDs

- Christoph Garth
- Jonas Lukasczyk
- Timo Mühlhaus
- Benedikt Venn
- Jens Krüger
- Kolja Glogowski
- Cristina Martins Rodrigues
- Dirk von Suchodoletz

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 Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[2] Philippa C Griffin, Jyoti Khadake, Kate S LeMay, Suzanna E Lewis, Sandra Orchard, Andrew Pask, Bernard Pope, Ute Roessner, Keith Russell, Torsten Seemann, et al. Best practice data life cycle approaches for the life sciences. *F1000Research*, 6, 2017.

[3] Dirk von Suchodoletz, Timo Mühlhaus, Jens Krüger, Björn Usadel, and Cristina Martins Rodrigues. Dataplant – ein nfdi-konsortium der pflanzen-grundlagenforschung. *Bausteine Forschungsdatenmanagement*, (2):46–56, 2021.

[4] David King and David Katch. *Current Practices in Software Development: A Guide to Successful Systems*. Yourdon Press Englewood Cliffs, New Jersey, 1984.

[5] Rafael S Gonçalves and Mark A Musen. The variable quality of metadata about biological samples used in biomedical experiments. *Scientific data*, 6(1):1–15, 2019.

[6] Christian T Jacobs and Alexandros Avdis. Git-rdm: A research data management plugin for the git version control system. *Journal of Open Source Software*, 1(2):29, 2016.

[7] Eric C Kansa, Sarah Whitcher Kansa, and Benjamin Arbuckle. Publishing and pushing: mixing models for communicating research data in archaeology. *International Journal of Digital Curation*, 9:57—70, 2014.

[8] Vidya Ayer, Christian Pietsch, Johanna Vompras, Jochen Schirrwagen, Cord Wiljes, Najko Jahn, and Philipp Cimiano. Conquaire: Towards an architecture supporting continuous quality control to ensure reproducibility of research. *D-Lib Magazine*, 23(1/2), 2017.

[9] Eric Raymond. The cathedral and the bazaar. *Knowledge, Technology & Policy*, 12(3):23–49, 1999.

[10] ISA TAB format. `https://isa-specs.readthedocs.io/en/latest/isatab.html`, accessed 2021-04-26.

[11] Jean-Baptiste Poline. From data sharing to data publishing [version 1; referees. 2018.

[12] Eoghan Ó Carragáin, Carole Goble, Peter Sefton, and Stian Soiland-Reyes. RO-Crate, a lightweight approach to Research Object data packaging, July 2019.

[13] DVC – open-source version control system for machine learning projects. `https://dvc.org`, accessed 2021-04-26.

[14] Eoghan Ó Carragáin, Carole Goble, Peter Sefton, and Stian Soiland-Reyes. A lightweight approach to research object data packaging. In *Bioinformatics Open Source Conference (BOSC) 2019*, 2019.

[15] Yaroslav O Halchenko, Benjamin Poldrack, and Michael Hanke. Datalad–decentralized data distribution for consumption and sharing of scientific datasets. In *Organization of Human Brain Mapping Poster. Organization of Human Brain Mapping Annual Meeting, Geneva, Switzerland*, 2016.

Nationale Forschungsdateninfrastruktur (NFDI)

Sophie Kraft[1], Hendrik Seitz-Moskaliuk[1], York Sure-Vetter[1], Elena Wössner[1], Nils Bohmer[2], Jan Eufinger[3], Juliane Fluck[4], Oliver Koepler[5], Jan Korbel[6], Bernhard Miller[7], Sarah Pittroff[8], Cristina Martins Rodrigues[9], Thorsten Schwetje[10], Dirk von Suchodoletz[9] und Judith Sophie Weber[11]

[1]Nationale Forschungsdateninfrastruktur (NFDI) e.V., Karlsruhe
[2]DECHEMA e.V., Frankfurt am Main (NFDI4Cat)
[3] Deutsches Krebsforschungszentrum Heidelberg (DKFZ) (GHGA)
[4]ZB MED - Informationszentrum Lebenswissenschaften, Köln (NFDI4Health)
[5]TIB - Leibniz Informationszentrum Technik und Naturwissenschaften, Hannover (NFDI4Chem)
[6]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg (GHGA)
[7]GESIS - Leibniz-Institut für Sozialwissenschaften. Mannheim und Köln (KonsortSWD)
[8]Akademie der Wissenschaften und der Literatur | Mainz (NFDI4Culture)
[9]Albert-Ludwigs-Universität Freiburg (DataPLANT)
[10]Technische Universität Darmstadt (NFDI4Ing)
[11]MARUM - Center for Marine Environmental Sciences University of Bremen (NFDI4BioDiversity)

Im Oktober 2020 gingen die ersten 9 Konsortien sowie der Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V. an den Start. Ziel von NFDI ist die Schaffung übergreifender Strukturen für das Forschungsdatenmanagement in Deutschland in enger Anbindung an die Fachcommunities. Dieser Artikel fasst die Kurzvorstellungen der 9 Konsortien und des NFDI-Direktorats aus dem gemeinsamen Workshop der E-Science-Tage 2021 zusammen.

1 Einleitung

In der Nationalen Forschungsdateninfrastruktur (NFDI) werden die wertvollen Datenbestände von Wissenschaft und Forschung für das gesamte deutsche Wissenschaftssystem systematisch erschlossen, vernetzt und nachhaltig sowie qualitativ nutzbar gemacht. Bislang sind sie oftmals projektbezogen oder auf Zeit verfügbar und meist nicht miteinander verknüpft.

Bund und alle 16 Länder fördern NFDI gemeinsam auf Grundlage des §91b Grundgesetz. Der NFDI-Verein hat als wesentlichen Zweck die Förderung von Wissenschaft und Forschung durch eine Nationale Forschungsdateninfrastruktur, die ein übergreifendes Forschungsdatenmanagement in Deutschland etabliert und fortentwickelt und die Effizienz

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13752 (CC BY-SA 4.0)

des gesamten deutschen Wissenschaftssystems steigert. Eine ausführliche Beschreibung des Vereins findet sich in [1].

Zur Erfüllung des Vereinszwecks werden in einem wissenschaftsgeleiteten Verfahren, durchgeführt von der DFG, zwischen 2020 und 2022 bis zu 30 Konsortien ausgewählt, die Service-Portfolios für einen definierten Nutzerkreis entwickeln. Konsortien bilden sich entlang von Fachcommunities oder gemeinsam genutzter Datentypen. Stand 2021 werden bereits neun Konsortien gefördert.

Ein wesentliches Merkmal eines jeden Konsortiums ist die enge Zusammenarbeit von Wissenschaft und Wissenschaftsinfrastruktur. Die Konsortien arbeiten gemäß der in ihren Communities ermittelten Bedarfe und beteiligen sich darüber hinaus am Aufbau von NFDI insgesamt. Ziele dieser Partnerschaften sind die nachhaltige, qualitative und systematische Sicherung, Erschließung und Nutzbarmachung von Forschungsdaten über regionale und vernetzte Wissensspeicher, die Etablierung eines Forschungsdatenmanagements nach den FAIR-Prinzipien sowie die Anbindung und Vernetzung zu internationalen Initiativen wie der European Open Science Cloud (EOSC).

2 NFDI Konsortien

Im Folgenden stellen sich die neun Konsortien der ersten Auswahlrunde kurz vor.

DataPLANT

Das Hauptziel von DataPLANT besteht in der Unterstützung von Pflanzen-Grundlagenforschenden in Fragen des Forschungsdatenmanagements (FDM) [2]. DataPLANT als ein eher kompaktes Konsortium adressiert die wichtigsten Kernpunkte eines fachorientierten FDMs in drei zentralen Task Areas zu Standardisierung, zu technischen Diensten und Infrastruktur sowie zur persönlichen Unterstützung vor Ort. Letzteres geschieht über die sogenannten Data Stewards als Kernelement der DataPLANT FDM-Strategie [3]. Durch das in DataPLANT vereinbarte Präsenz-Modell der Data Stewards in den Forschungsgruppen profitieren diese von direkter und bedarfsgerechter Unterstützung. Data Stewards bilden die entscheidende Brücke zwischen den technischen Lösungen, der Infrastruktur und den Forschenden und fördern durch Unterstützung und Beratung im Daten- und Workflow-Management den Prozess der Standardisierung von Metadaten, Nutzung von Ontologien und Provenienz in der Datenverarbeitung. Bereits in der Antragsphase haben sich 32 Forschungsgruppen aus dem Bereich der Pflanzenforschung für eine Zusammenarbeit entschieden. Um sukzessive die gesamte Forschungslandschaft abzudecken, strebt DataPLANT an, zunächst große Verbünde, dann kleinere Forschungsgruppen und anschließend einzelne Personen einzugliedern.

DataPLANT arbeitet datenzentriert und baut auf bestehenden Strukturen auf. Ein zentrales Moment zur Zielerreichung ist der **A**nnotated **R**esearch **C**ontext (ARC), welcher

als Einstiegspunkt fungiert und zukünftig die Struktur einer Datenpublikation im Fachgebiet definiert. Der ARC wird den gesamten Forschungszyklus abdecken, vom Experiment über die rechnerischen Aspekte bis hin zu den eigentlichen Daten und Metadaten sowie die daraus resultierenden Publikationen. Es basiert auf bestehenden Formaten, Terminologien und Richtlinien, sodass die Integration bestehender Daten in bereits existierende Repositorien möglichst reibungslos erfolgen kann. Mit einem initialen Fokus auf Proteomik- und Transkriptomik-Datensätzen stehen bereits eine Reihe von Tools und Services [4], um von Beginn an einen schnellen Einstieg in ein verbessertes Datenmanagement zu gewährleisten. Gleichzeitig wird die technische Infrastruktur für die gemeinsame Nutzung und Versionieren von ARCs aufgebaut [5]. Dementsprechend stellt DataPLANT die zentrale Anlaufstelle für Forschende im Bereich der Pflanzen-Grundlagenforschung dar, um ein entsprechendes Forschungsdatenmanagement einzurichten.

GHGA

Ziel des NFDI-Konsortiums GHGA (German Human Genome-Phenome Archive) ist der Aufbau eines nationalen Genomarchivs für die sichere Speicherung, den Zugriff und die Analyse menschlicher Omics-Daten (z.B. Genome, Transkriptome) in einem einheitlichen ethisch-rechtlichen Rahmen. Solche Genomdaten und andere verwandte sogenannte Omics-Daten, die mithilfe moderner Sequenzierverfahren gewonnen werden, sind integraler Bestandteil der biomedizinischen Forschung. In Zukunft werden diese Daten auch die klinische Versorgung immer stärker prägen. Im Rahmen der Datennutzung für die Forschung muss das Bedürfnis, Daten offen und FAIR zu teilen immer mit dem Schutz der Privatsphäre der Patient:innen ausbalanciert und gegeneinander abgewogen werden. Zugriff kann dabei nur unter Einhaltung der notwendigen technischen und organisatorischen Schutzmaßnahmen und für legitime Forschungszwecke gewährt werden. Auf europäischer Ebene gibt es für diesen Zweck bereits das Europäische Genom-Phänom-Archiv (EGA). Da die zentrale EGA Infrastruktur die spezifischen nationalen Regelungen zum Datenschutz nur ungenügend abbilden kann, bereitet das EGA aktuell eine Umwandlung in eine föderierte Infrastruktur aus nationalen Knoten (“föderiertes EGA”) vor. GHGA wird hierzu den deutschen Knoten innerhalb des föderierten EGA aufbauen [6].

Bei der Entwicklung von GHGA werden auch die Wünsche der Forschungsgemeinschaften nach effizienten, benutzerfreundlichen Analysen im großen Maßstab und zur Replikation von Ergebnissen auf anderen Kohorten berücksichtigt. GHGA setzt dabei auf existierenden, nationalen Omics-Datenlieferanten und deren IT-Infrastrukturen auf, um eine harmonisierte, interoperable Infrastruktur zu schaffen. Ziel ist es, Forschende in Deutschland in die Lage zu versetzen, humane Genomdaten rechtssicher entsprechend der FAIR-Richtlinien austauschen zu können und dabei internationale Standards zum Datenaustausch stärker mitzugestalten. GHGA ist dabei eingebunden in flankierende internationale Forschungsnetzwerke wie etwa die europäische *Beyond One Million Genomes Initiative*. Eine ausführliche Beschreibung von GHGA ist hier zu finden [7].

KonsortSWD

Mit dem Konsortium für die Sozial-, Bildungs-, Verhaltens- und Wirtschaftswissenschaften (KonsortSWD) wird innerhalb der NFDI ein bereits erfolgreich etabliertes Kooperationsnetzwerk zu einer integrierten Dateninfrastruktur weiterentwickelt [8]. Das Netzwerk hat das geteilte Verständnis, dass viele Daten nur dann überhaupt für die (Sekundär-)Nutzung bereitgestellt werden können, wenn sie bei den Datenanbieter:innen verbleiben. Dass diese Dezentralität gleichzeitig auch Effizienz bedeutet, hat mit der Vielfalt von Datentypen, vor allem aber mit der Vielfalt an rechtlichen und forschungsethischen Einschränkungen in den Disziplinen von KonsortSWD zu tun. KonsortSWD gründet wesentlich auf der Arbeit des Rates für Sozial- und Wirtschaftsdaten (RatSWD), der seit 2004 dem forschungsfreundlichen Zugang zu administrativen und später auch anderen Daten im Rahmen von Forschungsdatenzentren (FDZ) zum Erfolg verholfen hat. Heute stellen 39 FDZ Daten deutlich über die namensgebenden Sozial- und Wirtschaftswissenschaften hinaus zur Verfügung.

KonsortSWD setzt sich für möglichst leicht nutzbare, qualitativ hochwertige Daten für die Gesellschaftsforschung ein. Unser Forschungsdatenmanagement (FDM) wird daher so angelegt sein, dass es Forschende und FDZ technisch und inhaltlich bei Sicherung und Nachnutzung (neuer) sensibler und nicht sensibler Daten unterstützt. Eine Grundlage dabei sind die Arbeiten der Data Documentation Initiative (DDI) aufbauen, einem international verwendeten Metadatenstandard, der maßgeblich von der KonsortSWD Community entwickelt wurde und wird. Insbesondere baut KonsortSWD die Unterstützung für die qualitative Sozialforschung aus: Forschende sollen u.a. durch Anonymisierungstools, Handreichungen zu Informed Consent über den gesamten Forschungsprozess beim FDM unterstützt werden. Auch werden die Zugangspunkte zu den Daten stärker vernetzt, um den Aufwand für die Nutzenden, z.B. beim Zugang zu sensitiven Daten in den FDZ, zu reduzieren.

NFDI4BioDiversity

So mannigfaltig die Fragestellungen und Zusammenhänge innerhalb der Biodiversität sind, so komplex und heterogen sind auch die Daten, die erhoben und analysiert werden, um diese zu durchdringen. Das Konsortium NFDI4BioDiversity arbeitet innerhalb der NFDI daran, diese Daten besser verfügbar zu machen, diese zu harmonisieren, zu integrieren und eine breitere Basis für Analysen zu schaffen. Und nicht nur die Daten selber, sondern auch die Akteure der Biodiversitätslandschaft sind vielfältig und reichen von wissenschaftlichen Institutionen zu Behörden, Bürgerwissenschaften, Museen, Sammlungen und Fachgesellschaften. Vertreter dieser Akteure bündelt NFDI4BioDiversity mit seinen 49 Partnern.

NFDI4BioDiversity baut dabei auf die Vorarbeiten der German Federation for Biological Data (GFBio)[1] auf, die bereits erprobte Services und Beratungsangebote zur Unterstützung von Forschenden in allen Phasen des Datenlebenszyklus aufgebaut hat. Das Angebot wird jetzt auf den NFDI4BioDiversity Nutzerkreis erweitert und um weitere Services und Tools ausgebaut. Nutzereinbindung und Praxistauglichkeit der Angebote werden durch insgesamt 23 Use Cases gewährleistet, die von der Anbindung einzelner institutioneller Datenressourcen über die Verbindung internationaler Netzwerke bis zum Ausrollen neuer Software-Tools für das Management und die Darstellung von Daten reichen.

Auf der technischen Ebene ist das zentrale Arbeitsziel die Entwicklung einer cloud-basierten Infrastruktur, der NFDI Research Data Commons. Dort sollen Daten aus verschiedenen Quellen zusammengebracht, integriert und harmonisiert werden. Dies beinhaltet auch standardisierte Workflows der jeweiligen Datenprovider, die ausgehandelt und weiterentwickelt werden müssen, sowie die Verwendung gängiger (Meta)datenstandards und Ontologien. Diese Schritte bilden die Grundlage für z.B. die Anwendung von Analysetools, das Bereitstellen von Trainingsdatensätzen, die Visualisierung der Daten in Dashboards oder die Darstellung von Datenpaketen für bestimmte Zielgruppen. Eine ausführliche Darstellung von NFDI4BioDiversity findet sich in [9].

NFDI4Cat

Um das volle Potenzial der in der Katalyseforschung generierten Daten auszunutzen, ist ein grundlegender, digitaler Wandel in diesem Wissenschaftsbereich sowie in der Prozess- und Verfahrenstechnik erforderlich. Aus dieser Notwendigkeit heraus wurde im Rahmen der Initiative zur Nationalen Forschungsdateninfrastruktur (NFDI) das NFDI4Cat-Konsortium gebildet, das seit Oktober 2020 von der DFG gefördert wird. Das NFDI4Cat-Konsortium zielt darauf ab, eine Transformation vom derzeitigen Status zur “digitalen Katalyseforschung” zu ermöglichen und zu beschleunigen. Die entscheidenden Elemente, die für eine ”digitale Katalyseforschung”benötigt werden, sind die Vereinheitlichung von Konzepten, Vokabularien und Datenformaten sowie die Schaffung von vernetzten Informationsarchitekturen, die die Speicherung und den Austausch von semantisch reichen Daten ermöglichen. Diese Dateninfrastrukturen sollen den Einsatz moderner Analyseansätze ermöglichen, insbesondere mit Werkzeugen, die auf künstlicher Intelligenz basieren.

Das NFDI4Cat-Konsortium besteht aus 16 erfahrenen Partnern aus dem Bereich der homogenen, heterogenen, Photo-, Bio- und Elektrokatalyse und wird von der DECHEMA Gesellschaft für Chemische Technik und Biotechnologie e.V. koordiniert. Das Konsortium besteht außerdem aus Datenproduzent:innen und -nutzer:innen aus dem akademischen Bereich und aus anderen Kooperationsverbünden innerhalb der NFDI. Die Besonderheit innerhalb dieses Konsortiums ist die Beteiligung von industriellen Fachkolleg:innen mit beratender Funktion.

[1] `http://www.gfbio.org`

NFDI4Chem

NFDI4Chem ist das Fachkonsortium Chemie innerhalb der NFDI, mit der Vision alle Prozesse im Umgang mit Forschungsdaten aus der chemischen Forschung zu digitalisieren und zu vernetzen [10]. Beginnend bei der Datenerzeugung, über deren Verarbeitung und Analyse bis hin zur Publikation entwickelt NFDI4Chem eine modular aufgebaute, vernetzte Infrastruktur. Zum Beispiel stellt sie Software-Tools wie das elektronische Laborjournal und Daten-Repositorium „Chemotion“ Forschenden zur Unterstützung im Laboralltag zur Verfügung. NFDI4Chem fördert das Forschungsdatenmanagement und Open Science gemäß der FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable). Der Digitalisierungsprozess wird weiterhin durch die Entwicklung von Ontologien zur semantischen Beschreibung von Daten, Standards für Daten- und Metadatenformate sowie Minimalinformationen unterstützt. NFDI4Chem möchte gemeinsam mit der wissenschaftlichen Community einen kulturellen Wandel zur Etablierung und Akzeptanz eines FAIRen Datenumgangs gestalten.

Die Gründung des Konsortiums begann im Jahr 2018 als Graswurzelbewegung durch den Zusammenschluss von Vertreter:innen aus universitärer und außeruniversitärer Forschung, Infrastruktureinrichtungen, Rechenzentren und nationalen Fachgesellschaften, wie der Gesellschaft Deutscher Chemiker (GDCh), der Deutschen Bunsen-Gesellschaft (DBG) und der Deutschen Pharmazeutischen Gesellschaft (DPhG). Aus dieser Initiative bildete sich das Fachkonsortium Chemie NFDI4Chem unter der Leitung von Christoph Steinbeck (Friedrich-Schiller-Universität, Jena) und Oliver Koepler (TIB - Leibniz Informationszentrum Technik und Naturwissenschaften, Hannover), welches damit mehr als 40.000 Mitglieder der chemischen Gemeinschaft repräsentiert.

NFDI4Culture

Ziel von NFDI4Culture ist der Aufbau einer bedarfsorientierten, forschungsgetriebenen Infrastruktur für Forschungsdaten des materiellen und immateriellen Kulturerbes innerhalb der NFDI. Mit seinem Fokus auf Kulturgüter deckt das Konsortium ein breites Spektrum an akademischen Disziplinen und ihrer Gegenstände ab: von der Musikwissenschaft, Kunstgeschichte und Architektur bis hin zur Theater-, Film- und Medienwissenschaft. Bislang existiert auf nationaler Ebene keine Struktur, die sich um die fächerübergreifende Auffindbarkeit und Zugänglichkeit sowie um die langfristige Sicherung und kontinuierliche Pflege von Forschungsdaten des kulturellen Erbes bemüht. NFDI4Culture will diese Lücke schließen [11]. Zu den Gegenständen des Konsortiums gehören digitale Repräsentationen (2D-Digitalisate von Gemälden, Fotografien und Zeichnungen, digitale 3D-Modelle kulturhistorisch bedeutender Gebäude, Denkmäler, audiovisuelle Daten von Musik-, Film und Bühnenaufführungen u.a. mehr), genuine Forschungsdaten sind ebenso ihre Metadaten, Annotationen oder andere durch Forschung am Objekt gewonnene Daten. Bei der Sicherung, Standardisierung und Bereitstellung von nachhaltiger Infrastruktur und an den Nutzer:innen orientierten Diensten des Forschungsdatenmanagements ist das Konsortium entlang der FAIR- und CARE-Prinzipien konzipiert.

Die von NFDI4Culture adressierte Forschungslandschaft ist durch eine hohe institutionelle Diversität gekennzeichnet. Die Forschungseinheiten in der Interessengemeinschaft des Konsortiums reichen von individuellen Forscher:innen bis zu Universitätsinstituten, Kunsthochschulen, Akademien sowie Institutionen des kulturellen Erbes wie Galerien, Bibliotheken, Archiven und Museen (GLAM). Konzept und Struktur des Konsortiums wurden über zwei Jahre in enger Zusammenarbeit zwischen elf Fachgesellschaften, neun Trägerinstitutionen und 52 Partnern entwickelt. Zu den Trägerinstitutionen gehören vier Universitäten (Köln, Heidelberg, Marburg, Paderborn), drei Infrastruktureinrichtungen (FIZ Karlsruhe, TIB Hannover, SLUB Dresden) und die Stiftung Preußischer Kulturbesitz. Sprecherinstitution ist die Akademie der Wissenschaften und der Literatur | Mainz.

NFDI4Health

Das Ziel von NFDI4Health ist es, in Deutschland eine Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten aufzubauen. Die Integration wichtiger deutscher Forschungsinstitute mit Erfahrung als Datenhalter, -analyst und Methodenentwickler macht NFDI4Health zu einem interdisziplinären Konsortium, das auf etablierten Strukturen, Kompetenzen und Know-How sowie einer zunehmenden Unterstützung und Teilnahme der Forschungsgemeinschaft aufbaut.

Deutschland verfügt über eine Fülle gesundheitsbezogener Daten aus gut strukturierten Langzeitstudien und Datenerhebungen bei gesunden Personen (epidemiologische / Public Health Studien) sowie aus klinischen Studien mit Patient:innen in Krankenhäusern, die eine umfassende Beschreibung der Studienteilnehmenden anhand von Fragebögen, medizinischen Untersuchungen und molekularen oder genetischen Profilen aufweisen. Durch ihren Längsschnittcharakter und ihre hohe Qualität sind diese Daten eine wertvolle Ressource für die Entwicklung von präventiven und therapeutischen Maßnahmen auf Individual- und Populationsebene.

NFDI4Health möchte ein umfassendes Inventar deutscher epidemiologischer, Public Health- und klinischer Studiendaten aufbauen. Diese Daten sollen nach den FAIR-Prinzipien (inter-)national zugänglich gemacht werden.

Die in die NFDI eingebetteten Aufgaben von NFDI4Health sind:

1. Auffindbarkeit von und Zugang zu strukturierten Gesundheitsdaten ermöglichen.
2. Föderalen Rahmen für Datenhaltungsorganisationen erhalten.
3. Austausch und Verknüpfung von personenbezogenen Daten unter Wahrung des Datenschutzes ermöglichen.
4. Automatisierte Dienste (z.B. Suche, Analysetools) etablieren.
5. Interoperabilität und Wiederverwendbarkeit der Daten etablieren und verbessern.
6. Anwendungsfallorientierte Zusammenarbeit zwischen Forschungsgemeinschaften fördern.

Eine ausführliche Beschreibung des Konsortiums findet sich in [12].

NFDI4Ing

NFDI4Ing repräsentiert die deutsche ingenieurwissenschaftliche Forschungslandschaft.

Das Konsortium vereint Institutionen und Personen aus allen Bereichen ingenieurwissenschaftlicher Tätigkeiten, die ihre jeweils spezifische Expertise, Erfahrung und Netzwerke in das Arbeitsprogramm des Konsortiums einbringen. Damit ermöglicht NFDI4Ing die enge Vernetzung der Perspektiven von Ingenieurwissenschaftler:innen, Anwender:innen und Diensteanbieter:innen, um auf das gemeinsame Ziel hinzuarbeiten, die FAIR-Prinzipien wissenschaftlichen Datenmanagements für die Ingenieurwissenschaften Realität werden zu lassen.

Eine zentrale Aufgabe von NFDI4Ing ist es, die große Vielfalt ingenieurwissenschaftlicher Forschungsansätze und -methoden in einer begrenzten Anzahl gemeinsamer Standards und Anforderungen an Forschungsdatenmanagement (FDM) zu konsolidieren. Dies ist ein fortlaufender Prozess. Seit 2017 konnten systematisch und methodengestützt Bedürfnisse, Erwartungen und Arbeitsabläufe in der ingenieurwissenschaftlichen Forschung identifiziert und in sieben typische ingenieurwissenschaftliche Forschungsprofile klassifiziert werden. Wir nennen diese Forschungsprofile unsere *Archetypen* und verwenden sie in NFDI4Ing zur Strukturierung und stetigen Weiterentwicklung unseres Arbeitsprogramms. In gleichnamigen Arbeitsgruppen werden typische Methoden und Verfahrensweisen harmonisiert und neue Anwendungsfelder eröffnet. Flankiert werden die Archetypen-Arbeitsgruppen durch Teilprojekte, die den Austausch mit den Fachcommunities gewährleisten (*Community Clusters*), und spezialisierte Dienste und Dienstleistungen für Ingenieur:innen entwickeln, standardisieren und bereitstellen (*Base Services*).

Angeleitet werden wir dabei von unseren acht *Kernzielen* [13], die z.B. von der lückenlosen Reproduzierbarkeit aller Teilschritte des Forschungsprozesses über die Automatisierung der Erzeugung und Verwaltung von Metadaten bis zur Etablierung von data literacy in allen Phasen ingenieurwissenschaftlicher Aus- und Weiterbildung reichen, und im Ergebnis ingenieurwissenschaftliches Forschungsdatenmanagement FAIR machen sollen.

3 Fazit

So unterschiedlich die fachlichen Hintergründe der 9 Konsortien sind, so ähnlich sind die Fragestellungen, die sie im Rahmen von NFDI bearbeiten: Wie werden die Bedarfe der Nutzenden von Dateninfrastrukturen ausreichend im Konsortium abgebildet und umgesetzt? Wie können Daten dauerhaft nach den FAIR-Prinzipien gespeichert und zugänglich gemacht werden? Welche rechtlich-ethischen Aspekte sind dabei zu beachten? Was sind geeignete Technologien und Metadatenformate? Wie können Forschende im Datenmanagement geschult werden? Und noch vieles mehr. Innerhalb der Konsortien liegt der Fokus

auf communityspezifischen Antworten. Gleichzeitig bietet der NFDI-Verein den Rahmen, konsortienübergreifende Lösungen zu finden und dabei von den unterschiedlichen Expertisen der Mitglieder zu profitieren. Der Verein nimmt seit Januar 2021 Mitglieder auf und die verschiedenen Vereinsorgane konstituieren sich im Lauf des ersten Halbjahres 2021. Gleichzeitig nimmt der konsortienübergreifende Austausch Fahrt auf, um gemeinsam am Forschungsdatenmanagement der Zukunft zu arbeiten.

Literaturverzeichnis

[1] S. Kraft, A. Schmalen, H. Seitz-Moskaliuk, Y. Sure-Vetter et al., "Nationale Forschungsdateninfrastruktur (NFDI) e.V.: Aufbau und Ziele", Bausteine Forschungsdatenmanagement, zur Veröffentlichung akzeptiert (2021).

[2] D. von Suchodoletz, T. Mühlhaus, J. Krüger, B. Usadel, C. Martins Rodrigues, "DataPLANT – Ein NFDI-Konsortium der Pflanzen-Grundlagenforschung", Bausteine Forschungsdatenmanagement, in Revision (2021).

[3] D. von Suchodoletz, T. Mühlhaus, D. Brillhaus, H. Jasbeen, B. Usadel, J. Krüger, H. Gauza, C. Martins Rodrigues, "DataStewards as ambassadors between the NFDI and the community", E-Science-Tage 2021, zur Veröffentlichung akzeptiert (2021).

[4] C. Martins Rodrigues, J. Krüger, T. Mühlhaus, B. Usadel, M. Tschöpe, D. von Suchodoletz, "DataPLANT – Tools and Services to structure the Data Jungle of fundamental plant researchers", E-Science-Tage 2021, zur Veröffentlichung akzeptiert (2021).

[5] B. Venn, T. Mühlhaus, D. von Suchodoletz, J. Krüger, B. Usadel, C. Garth, "Immutable yet evolving: ARCs for permanent sharing in the research data-time continuum", E-Science-Tage 2021, zur Veröffentlichung akzeptiert (2021).

[6] G. Saunders, M.l. Baudis, R. Becker, S. Beltran, C. Béroud, E. Birney, C. Brooksbank, et al. "Leveraging European Infrastructures to Access 1 Million Human Genomes by 2022." Nature Reviews. Genetics 20, no. 11, 693–701 (2019). `https://doi.org/10.1038/s41576-019-0156-9`.

[7] J. Eufinger, J. Korbel, E. Winkler, O. Kohlbacher, O. Stegle, "Genomdaten FAIR und sicher teilen: Das Deutsche Humangenom-Phänom Archiv (GHGA) als Baustein der Nationalen Forschungsdateninfrastruktur" Bausteine Forschungsdatenmanagement, zur Veröffentlichung akzeptiert (2021).

[8] B. Hollstein, B. Miller, P. Siegers und C. Wolf, "KonsortSWD: Vom Netzwerk zur integrierten Dateninfrastruktur der Gesellschaftsforschung", Bausteine Forschungsdatenmanagement, zur Veröffentlichung akzeptiert (2021).

[9] J. S. Weber, B. Ebert, M. Diepenbroek, I. Kostadinov, F. O. Glöckner, "NFDI4BioDiversity - NFDI-Konsortium für Biodiversitäts-, Ökologische und Umweltdaten" Bausteine Forschungsdatenmanagement, in Revision (2021).

[10] J. Ortmeyer, F. Schön, S. Herres-Pawlis, N. Jung, F. Bach, J. Liermann, S. Neumann, C. Popp, M. Razum, O. Koepler, C. Steinbeck, NFDI4Chem - Fachkonsortium für die Chemie, Bausteine Forschungsdatenmanagement, in Revision (2021).

[11] R. Altenhöner, T. Schrade et al., NFDI4Culture - Consortium for research data on material and immaterial cultural heritage (2020). https://doi.org/10.3897/rio.6.e57036

[12] J. Fluck und I. Pigeot et al., "NFDI4Health - Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten", Bausteine Forschungsdatenmanagement, zur Veröffentlichung akzeptiert (2021).

[13] R. H. Schmitt, V. Anthofer et. al. "NFDI4Ing - the National Research Data Infrastructure for Engineering Sciences" (2020). https://doi.org/10.5281/zenodo.4015201

Entwurf einer Infrastruktur für den Datenaustausch großer Forschungsdatenmengen mittels WebDAV, FTS3 und OIDC

Martin Baumann[1], Frauke Bösert[2], Sven Siebler[1], Paul Skopnik[2] und Jan Erik Sundermann[2]

[1]Universitätsrechenzentrum, Universität Heidelberg
[2]Steinbuch Centre for Computing, Karlsruher Institut für Technologie

Standortübergreifende Zusammenarbeit innerhalb wissenschaftlicher Communities im Umfeld föderierter Speicher- und Compute-Umgebungen erfordert häufig den Transfer großer Datenmengen zwischen Zentren und Speichersystemen. Vorgestellt wird eine prototypische Implementierung einer Infrastruktur, die auf einem mit Plugins erweiterten Apache-Webserver für den Datenaustausch mittels WebDAV basiert. Die Server ermöglichen dabei den Zugriff auf Speichersysteme und ergänzen diese um ein weiteres Zugriffsprotokoll, welches die Anbindung an FTS3 unterstützt.

Die erforderlichen Erweiterungen und erste Ergebnisse zum Transfer von Daten mittels WebDAV werden vorgestellt.

1 Einleitung

Die Zusammenarbeit in wissenschaftlichen Communities erfordert oft Transfers großer Datenmengen zwischen Zentren und Speichersystemen sowie authentifizierte Datenzugriffe. Die einfache Einbindung existierender Speichersysteme von Hochleistungsrechnern, Archivierungsdiensten oder Repositorien ist wünschenswert wenn nicht sogar notwendig, da auf diesen Systemen die großen Datenmengen erzeugt, analysiert oder später archiviert bzw. veröffentlicht werden. Der im folgenden vorgestellte Ansatz unterstützt bei der Lösung dieser Anforderungen und greift existierende bzw. sich aktuell in intensiver Erprobung befindliche Konzepte zum automatisierten und effizienten Datentransfer auf Basis von FTS3 und Standard-Protokollen [1, 2] aus dem Umfeld des LHC-Computings auf und ergänzt diese um technische Lösungen, die es ermöglichen sollen, bereits existierende und heterogene Speichersysteme in die föderierte Infrastruktur zu integrieren.

Die Durchführung entsprechender Transfers setzt ein geeignetes Netzwerkprotokoll, eine flexibel nutzbare Verwaltungsschicht für die Steuerung und Überwachung der Übertragungen und eine passende Authentifizierung voraus. Für den Transfer großer Datenmengen

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13753 (CC BY-SA 4.0)

hat sich das HTTP-basierte Netwerkprotokoll WebDAV als vielversprechende Technologie in verschiedenen Anwendungsfeldern bereits bewährt. Zur Steuerung und Überwachung von Transfers großer Datenmengen kann die Open-Source Software FTS3 genannt werden, die zur Übertragung der Daten des Teilchenbeschleunigers am CERN seit längerem produktiv eingesetzt wird. FTS3 fungiert dabei als zentrale koordinierende Stelle oder "dritte Partei", die Datentransfers zwischen zwei Speichersystemen initiieren kann. Der Datenfluss erfolgt dabei immer direkt zwischen den beiden beteiligten Speichersystemen. Mit OpenID Connect ist eine auf OAUTH2 aufbauende Authentifizierungsschicht gegeben, auf Basis derer eine token-basierte Authentifizierung ermöglicht wird. Mittels dieser Authentifizierungstokens können Berechtigungen für Datentransfers vorübergehend, beispielsweise an den FTS3-Transferdienst, delegiert werden.

2 Involvierte Technologien

FTS3: Die Open-Source Software FTS3 (`https://fts.web.cern.ch/fts/`) [3] wurde entwickelt, um Datenmengen im Petabyte-Bereich vom Teilchenbeschleuniger Large Hadron Collider am Europäischen Kernforschungszentrum CERN global zu verteilen und ist seit längerem produktiv im Einsatz. Sie ermöglicht es, Datenübertragungen effizient zu planen und die vorhandenen Netzwerkressourcen optimal einzusetzen. Zur Steuerung bietet FTS3 ein Kommandozeileninterface (CLI), eine REST API und mit WebFTS (`https://webfts.cern.ch`) ein web-basiertes Nutzerinterface. Die Authentifizierung erfolgt mittels x.509 Zertifikaten oder per OpenID Connect (OIDC/OAUTH2). Als Datenübertragungsprotokolle werden HTTPS/WebDAV, XrootD und GridFTP unterstützt. Mit FTS3 kann ein Dienst zur Orchestrierung von Third-Party-Datentransfers (TPC) zwischen verschiedenen Speichersystemen (URIs) betrieben werden.

WebDAV: Das im RFC 4918 spezifizierte Netzwerkprotokoll WebDAV (`http://webdav.org/`) basiert auf HTTP/HTTPS und erweitert dieses Protokoll um einen Satz neuer Befehle. Diese Befehle erlauben es beispielsweise, Eigenschaften von Dateien, Verzeichnissen oder Verzeichnisstrukturen abzurufen, Verzeichnisse zu erstellen, sowie Dateien oder Verzeichnisse zu kopieren, zu verschieben oder zu löschen. Als HTTP-basiertes Protokoll ist es standardisiert und darüber hinaus weit verbreitet. Entsprechende Implementierungen stehen für die üblichen Betriebssysteme (Win/Lin/Mac auch Android/iOS) wie auch in den gängigen Webservern zur Verfügung. Das WebDAV-Protokoll ermöglicht die Übertragung von einzelnen Dateien wie auch von ganzen Verzeichnissen. Die Verwendung von WebDAV ist gerade für standortübergreifende Verbindungen interessant, da der benötigte Port 443 als Standardport für HTTPS-Verbindungen in der Regel bereits freigegeben ist und somit meist keine speziellen Firewallanpassungen erforderlich werden. WebDAV-Verbindungen lassen sich abhängig von der gewählten Serverimplementierung auf verschiedene Weisen authentifizieren und es ist im Allgemeinen möglich, eine token-basierte Authentifizierung zu realisieren, wie sie bei dem hier beschriebenen Ansatz zur Integration mit FTS3 favorisiert wird.

OpenID Connect (OIDC): OIDC (`https://openid.net/connect/`) ist eine auf OAuth 2.0 basierte, inzwischen sehr weit verbreitete Authentifizierungsschicht. OIDC ermöglicht die Authentifizierung von Nutzern bei Diensten, ohne dass diese Passwörter austauschen oder verwalten müssen. Durch den Austausch von Authentifizierungstokens können Berechtigungen für Datentransfers vorübergehend beispielsweise an den FTS3-Transferdienst delegiert werden. Bei Verwendung geeigneter Werkzeuge [4, 5] können diese ähnlich wie ssh-Keys in übliche Workflows beim Datenzugriff integriert werden.

3 Aufbau des Prototyps

Es wurde begonnen, einen Prototyp der geplanten Infrastruktur aufzubauen. Mit diesem Prototyp sollte zunächst die Durchführbarkeit von TPC und die Funktionalität im Zusammenspiel der involvierten Technologien demonstriert werden und er sollte anschließend weiteren Untersuchungen dienen. Der Prototyp besteht aus den bereits genannten Technologien WebDAV, FTS3 und OIDC, die geeignet erweitert bzw. modifiziert wurden, sowie einzelnen Speicherendpunkten.

Zur Bereitstellung der Protokolle für den Datenzugriff wurden WebDAV-Server auf Basis des Apache-HTTP-Servers mit erweiterten und neuen Modulen verwendet. Nutzerdaten lagen in lokalen Dateisystemen vor. Der Ansatz lässt sich auf beliebige POSIX-Dateisysteme generalisieren, wie sie beispielsweise in den zentralen Speichersystemen an HPC-Clustern eingesetzt werden. Um den Apache WebDAV-Server auf beliebigen POSIX-Dateisystemen einsetzen zu können, ohne die existierenden Eigentümer und Berechtigungsstrukturen ändern zu müssen, wurde dieser so konfiguriert, dass WebDAV-Befehle mit der UID des authentifizierten Nutzers ausgeführt werden. Dazu wird das Apache-Modul mpm-itk [6] eingesetzt, welches es ermöglicht jede HTTP/HTTPS-Anfrage einzeln unter möglicherweise verschiedenen POSIX-Benutzern zu verarbeiten. Zu diesem Zweck verändert mpm-itk den Anfragen-Verarbeitungs-Mechanismus von Apache: Statt Anfragen von langlebigen Prozessen mit einer statisch konfigurierten UID verarbeiten zu lassen, wird jede Anfrage in einem eigenen Prozess verarbeitet, wobei dessen UID anfragespezifisch abgeleitet wird. Der zusätzliche Aufwand, einen neuen Prozess pro Anfrage zu starten, ist unbedeutend im Vergleich zum Datentransfer von großen Dateien. Für den WebDAV-Anwendungsfall wurde das Modul weiter angepasst, so dass die Entscheidung über die UID erst nach erfolgreicher Authentifizierung des Nutzers passiert. Mit diesem Setup ist es möglich, authentifizierten Nutzern Zugriff auf existierende POSIX-Speichersysteme zu geben und dabei alle durch das Dateisystem unterstützten Berechtigungsstrukturen inkl. erweiterter ACLs zu berücksichtigen.

Die Nutzerauthentifizierung erfolgt mittels OAUTH2 Bearer-Tokens und OpenID Connect [7, 8]. Die so authentifizierten Nutzer werden mit Hilfe eines dedizierten Moduls lokalen POSIX-Benutzern zugeordnet. Unter deren UID wird die Anfrage dann vearbeitet. Die eingesetzten Speichersysteme verwenden LDAP-Facaden und basieren damit vollständig auf bwIDM, dem föderiertes Identitätsmanagement in Baden-Württemberg [7].

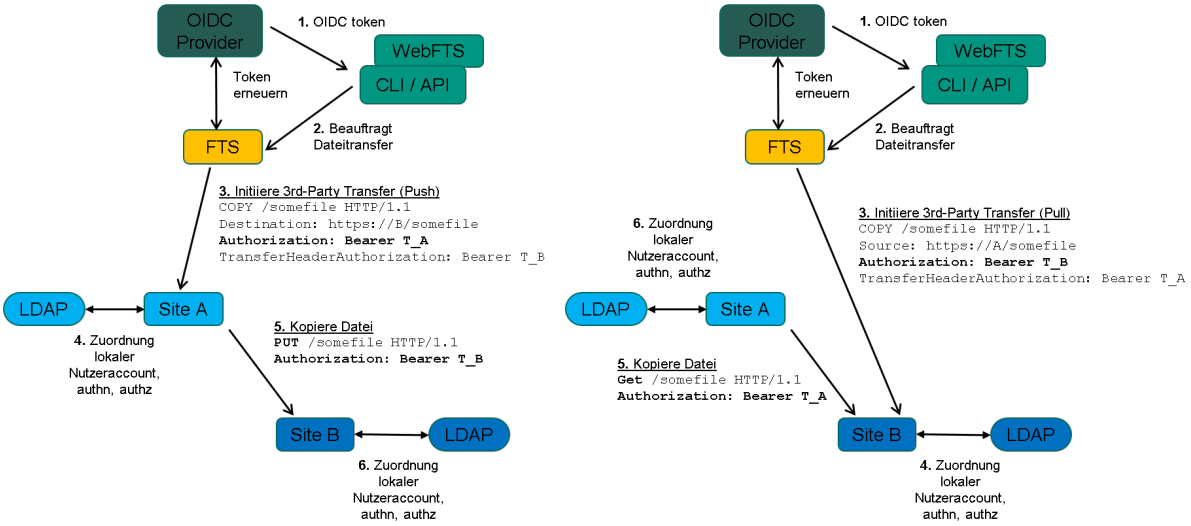


Abbildung 1: Third-Party-Dateitransfers mit WebDAV / FTS3 / OIDC im Push- (links) und Pull-Modus (rechts).

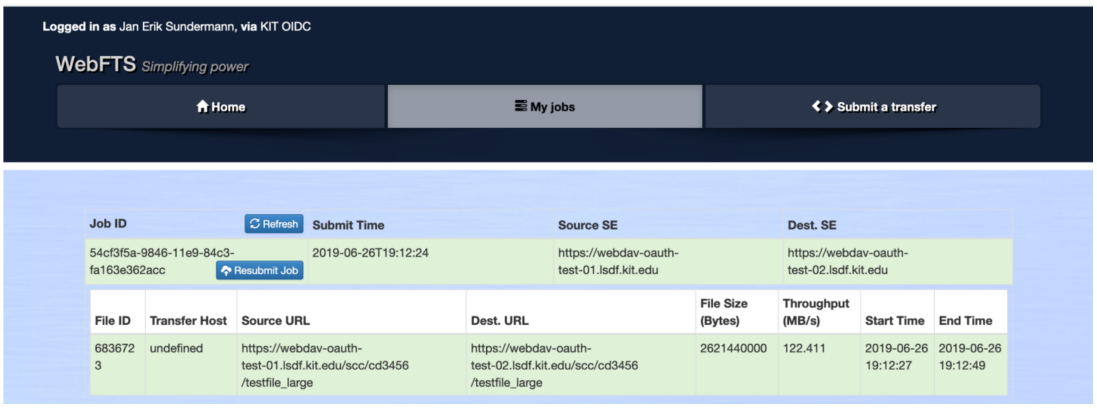


Abbildung 2: WebFTS-Instanz mit Beispiel eines erfolgreichen Transfers zwischen zwei Endpunkten.

Für die Integration eines vorhandenen Speichersystems an FTS3 muss der entsprechende Speicherendpunkt einen erweiterten Befehlssatz von WebDAV verstehen, der es ihm ermöglicht, TPC durchzuführen (siehe [10]). Zur Integration in den beschriebenen Prototypen wurde das existierende Apache WebDAV-Modul um den TPC-Befehl COPY sowie um Performance-Marker erweitert. Die Implementierung verwendet die GFAL2-Bibliothek [11]. Für diese Evaluation wurde ein existierender FTS3-Server am CERN verwendet, sowie eine neue WebFTS-Instanz am KIT aufgesetzt. Abbildung 1 illustriert den Ablauf eines von FTS initiierten TPC im Push- und im Pull-Modus. Nur einer der involvierten Speicherendpunkte muss einen WebDAV-Zugang bereit stellen, der den erweiterten Befehlssatz für TPC unterstützt.

4 Ergebnisse

Auf Basis des zuvor beschriebenen Prototypen wurde zunächst die grundsätzliche Funktionalität erprobt. Hierbei wurden mit Hilfe von WebFTS standortübergreifende TPC durchgeführt (siehe Abb. 2). Weiterhin wurden erste vergleichende Performancemessungen des WebDAV- und SFTP-Protokolls zwischen Heidelberg und Karlsruhe durchgeführt. Hierfür wurde das Tool rclone (`https://rclone.org/`) mit einem synthetischen Datensatz verwendet, bestehend aus je 147 GB Daten verschiedener Dateigröße (16 MB - 10 GB; 9362 - 14 Dateien). Die Benchmarkergebnisse für die Transfers mit WebDAV und SFTP sind in Abb. 3 dargestellt.

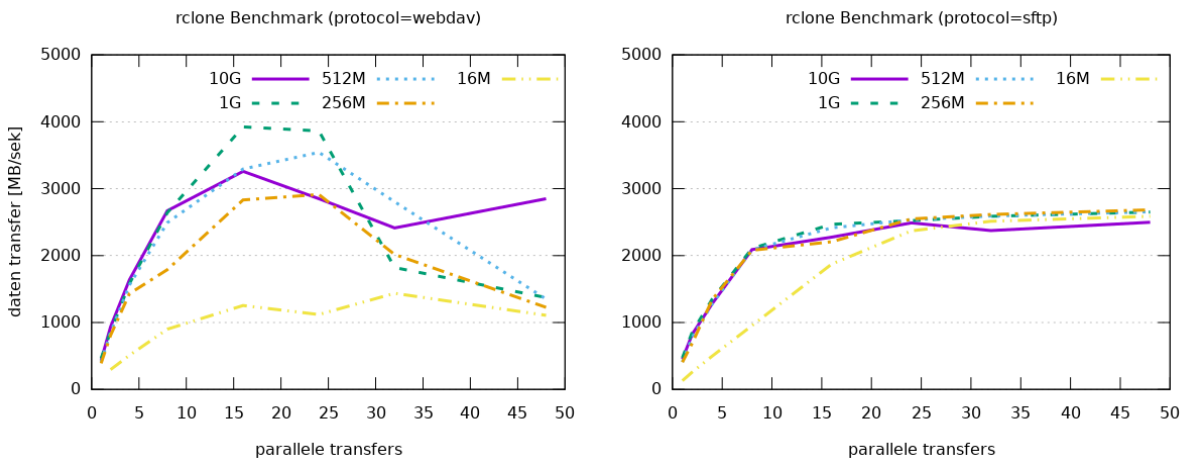


Abbildung 3: rclone Benchmark auf einem synthetischen Datensatz (je 147 GB einer Dateigröße).

In beiden Fällen kann ein deutlicher Performancegewinn durch Verwendung paralleler Transfers erreicht werden. Bei WebDAV fällt dieser Effekt zum einen stärker aus, zum anderen wird (bei bis zu 25 Transfers) mit 3-4 GB/s auch eine höhere Maximalgeschwindigkeit erreicht. Der Performanceeinbruch ab ca. 25 Transfers ist vermutlich auf serverseitige Parameter zurückzuführen, so dass durch weitere Optimierung auch eine höhere Skalierung zu erwarten ist. Bei Verwendung des SFTP-Protokolls erkennt man im Vergleich, dass das mögliche Maximum von 2.2-2.5 GB/s bereits durch 8-10 Transfers erreicht wird. Lediglich bei kleineren Dateigrößen skaliert SFTP deutlich besser als WebDAV, was wahrscheinlich auf den Verbindungsoverhead durch die unterschiedliche Dateianzahl zurückzuführen ist.

WebDAV ist für den skalierenden Transfer von großen Datenmengen, auch standortübergreifend, geeignet. Insbesondere die Möglichkeit des automatisierten Datentransfers durch TPC bietet hierbei einen deutlichen Mehrwert für die Zusammenarbeit beim föderierten Datenaustausch.

5 Ausblick

Für den vorbereitenden Einsatz des Prototypen in Produktivumgebungen und die Verbesserung der Performanceskalierung sind im Weiteren folgende Aktivitäten vorgesehen: Optimierung verwendeter Apache-Webserver-Parameter, Verbesserung der TPC-Implementierung in WebDAV, Anbindung weiterer Speicherendpunkte und OIDC-Provider. Zusätzlich ist die Erprobung des WebDAV-Protokolls mit gängigen Anwendungen und Nutzerworkflows geplant, z.B. für die Datenanalyse mit Scikit-Learn und Jupyter Notebooks. Darüber hinaus soll evaluiert werden, wie sich die so aufgesetzte Infrastruktur zum automatisierten und regelbasierten Datentransfer z.B. mit Datenmanagementwerkzeugen wie Rucio [12] einsetzen lässt.

Danksagung

Die vorgestellten Ergebnisse wurden im Rahmen des Projekts bwHPC-S5 erarbeitet, das durch das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK) gefördert wird.

Die Benchmarkmessungen wurden auf den Speicherdiensten “LSDF Online Storage” (Karlsruhe) und “SDS@hd” (Heidelberg) durchgeführt, die vom MWK und der Deutschen Forschungsgemeinschaft (DFG) gefördert sind (INST 35/1314-1 FUGG, INST 35/1503-1 FUGG).

Literaturverzeichnis

[1] B. Bockelman, A. Ceccanti, F. Furano1, P. Millar, D. Litvintsev and A. Forti. “Third-party transfers in WLCG using HTTP”, EPJ Web Conf., Volume 245 (2020), 24th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2019)

[2] B. Bockelman, A. Hanushevsky, O. Keeble, M. Lassnig, P. Millar, D. Weitzel, W. Yang. “Bootstrapping a New LHC Data Transfer Ecosystem”, EPJ Web Conf., Volume 214 (2019), 23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2018)

[3] Kiryanov, A., T. A. Ayllon, and O. Keeble. “FTS3 / WebFTS – A Powerful File Transfer Service for Scientific Communities”, Procedia Computer Science, volume 66 (2015): 670-678.

[4] OIDC-Agent Projektseite, `https://indigo-dc.gitbook.io/oidc-agent/` [abgerufen 27. Juli 2021].

[5] G. Zachmann. “OIDC-Agent: Managing OpenID Connect Tokens on the Command Line”, In: Becker, M. (Hrsg.), SKILL 2018 - Studierendenkonferenz Informatik. Bonn: Gesellschaft für Informatik e.V. (S. 11-21).

[6] The Apache 2 ITK MPM, http://mpm-itk.sesse.net/,[abgerufen 27. Juli 2021].

[7] Apache-Modul mod_oauth2, https://github.com/zmartzone/mod_oauth2 [abgerufen 27. Juli 2021].

[8] Apache-Modul mod_auth_openidc, https://github.com/zmartzone/mod_auth_openidc,[abgerufen 27. Juli 2021].

[9] J. Köhler, S. Labitzke, M. Simon, T. Dussa, M. Nussbaumer, H. Hartenstein. “bwIDM – Federated Access to IT-Based Services at the Universities of the State of Baden-Württemberg”, De Gruyter, Online erschienen: 25. Januar 2014, https://doi.org/10.1515/pik-2013-0025.

[10] HTTP/WebDAV Third-Party-Copy Technical Details, CERN Wiki(25. März 2020): https://twiki.cern.ch/twiki/bin/view/LCG/HttpTpcTechnical [abgerufen 27. Juli 2021].

[11] Grid File Access Library (GFAL2), https://dmc-docs.web.cern.ch/dmc-docs/gfal2/gfal2.html,[abgerufen 27. Juli 2021].

[12] M. Barisits, T. Beermann, F. Berghaus et al. “Rucio: Scientific Data Management”, Computing and Software for Big Science volume 3, Article number: 11 (2019), https://doi.org/10.1007/s41781-019-0026-3.

PsyCuraDat: The development of a user-friendly curation standard for psychological research data

Katarina Blask, Marie-Luise Müller, Marc Latz, Valentin Arnold and Stephanie Kraffert

Leibniz-Institute for Psychology

The project *PsyCuraDat* aims to develop a user-friendly curation standard for psychological research data based on method-specific curation criteria, considering the needs of researchers in their roles as data providers and data users. This standard will bring enhanced effectiveness and efficiency, as well as increased satisfaction and quality to the curation and reuse of data in the field of psychology.

1 Introduction

The Open Science movement has brought useful knowledge and important changes to the scientific communities. Also within psychological science, the guidelines of open science practices are increasingly being implemented, in order to make research data openly accessible, and to enable sustainable (re-)use of data. However, standards specifically aiming at psychologists from all sub-disciplines, and allowing them to optimally prepare their data for reuse, hardly exist. To counteract this problem, we have started the project *PsyCuraDat*, to develop a user-friendly curation standard for psychological research data, meeting all necessary requirements to assure the data's long-term interpretability and reusability.

2 Approach & Development

In order to develop a curation standard, guaranteeing the sustainable use of psychological research data while meeting the needs of researchers in their role as data users and data providers, we began our project with a detailed examination of existing curation criteria and quality standards in psychology. We assessed whether existing standards are suitable for the documentation of all central phases of the research process (i.e., conceptualization, data collection, data analysis, publication and archiving). First results showed that none

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029714 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13754 (CC BY-SA 4.0)

of the existing standards could adequately provide the documentation of the whole psychological research process. Also, existing standards are barely known, and thus barely used. Our research suggests that this is mostly due to too complicated handling [2]. Subsequently, we explored which data, and metadata, are needed for which types of reuse. We found that researchers primarily depend on information about the research design, hypotheses, a codebook and a study protocol [2] . Proceeding from those results, we elaborated method-specific curation criteria for a documentation standard, and developed a first prototype. Those curation criteria have led to a contextual specification, divided into three documentation levels based on the psychological research process: The first level represents the research objective on a *conceptual level*, and consists of the hypotheses, and the research design, as well as information on the sample. On the second level, the research design is presented on an *operational level*, on which researchers describe the methods used to operationalize the variables (e.g., through an extended codebook). The third level offers a detailed *description* of the research process, including a procedure graphic, as well as a presentation of the data preparation and analysis steps.

In order to test the usability of the prototype, two user studies were conducted with samples composed of psychological researchers. Following researchers' direct interaction with datasets, prepared in accordance with the standard, we used cognitive interviews to investigate the standard's usability. We primarily focused on evaluating the formal specification, as well as exploring which metadata must be provided, and in what form it should be presented. Our findings in user study 1 suggested that participating researchers perceived the contextual specification and the three documentation levels as rather technical, and not interlinked enough. Thus, we refined the standard towards a more comprehensible structure, which was then tested in user study 2. Besides improving the standard's contextual structure, additional data preparation and analysis scripts, as well as a short manual, describing the different documentation parts and their functional properties, were added [3]. Furthermore, results of both user studies showed that researchers most strongly relied on the conceptual description of the research design, the codebook, the analysis script and the graphical description of the procedure. Moreover, they requested a more hierarchical documentation structure of the procedure graphic (i.e., the graphic presented as a flow chart). Lastly, all participating researchers stated that they could imagine integrating the standard into their research process, and that they perceived the benefits of using this documentation standard to outweigh the costs and efforts [4].

3 Conclusions & Outlook

This empirically developed documentation standard has been based on curation criteria created in collaboration with psychological researchers from various sub-disciplines. Its content specification reflects the psychological research process and enables a detailed documentation of all central phases by providing a three-level-structure. This structure allows for a detailed description of the data on a study level (i.e., the conceptual de-

scription of the research design) as well as on a data level (i.e., the codebook) which, according to our empirical results, is both essential in order to fully understand and reuse a dataset. Furthermore, information provided on these two levels have to be linked to the data, and to each other, through a comprehensive description of the data collection and analysis process, including a summarizing procedure graphic, as well as all materials used during the data collection and analysis process. Besides developing a user manual and guidelines about how to implement the standard in everyday research routine, we are currently working out the standard's information architecture, aiming at a high level of comprehensibility and usability. That is to say, an "easy-to-use" and "easy-to-learn" data documentation structure, enabling effective and efficient use and reuse of psychological research data. When completed, our standard aims to offer a user-friendly tool to allow psychologists from all sub-disciplines to optimally prepare their data for reuse. It offers a contribution to Open Science and to the sustainable use of research data.

Acknowledgements

The project *PsyCuraDat* has been funded by the German Federal Ministry of Education and Research (funding number: 16QK08).

Bibliography

[1] Blask, K., Gerhards, L., & Jalynskij, M. (2020). Metadata in Psychology 2.0: What researchers really need - Study description of the data referring to the online survey conducted in the BMBF-funded project PsyCuraDat. ZPID (Leibniz Institute for Psychology). https://doi.org/10.23668/PSYCHARCHIVES.2757

[2] Blask, K., Jalynskij, M., & Gerhards, L. (2020). Metadata in Psychology 1.0: What researchers really need - Study description of the data referring to the expert interviews conducted in the BMBF-funded project PsyCuraDat. ZPID (Leibniz Institute for Psychology). https://doi.org/10.23668/PSYCHARCHIVES.2750

[3] Blask, K., Müller, M.-L., Arnold, V., & Kraffert, S. (2020). Evaluation of the PsyCuraDat- Specification 1.0 - Study description of the data referring to the first user study conducted in the BMBF-funded project PsyCuraDat. ZPID (Leibniz Institute for Psychology). https://doi.org/10.23668/PSYCHARCHIVES.4318

[4] Blask, K., Müller, M.-L., Arnold, V., & Kraffert, S. (2021). Evaluation of the PsyCuraDat- Specification 2.0: Study description of the data referring to the second user study conducted in the BMBF-funded project PsyCuraDat. ZPID (Leibniz Institute for Psychology). https://doi.org/10.23668/PSYCHARCHIVES.4459

V-FOR-WaTer - a virtual research environment for environmental research

Marcus Strobl[1], Elnaz Azmi[1], Sibylle K. Hassler[2], Mirko Mälicke[2], Jörg Meyer[1], Achim Streit[1], and Erwin Zehe[2]

[1]Steinbuch Centre for Computing, Karlsruhe Institute of Technology
[2]Institute of Water and River Basin Management, Chair of Hydrology, Karlsruhe Institute of Technology

Extent and diversity of environmental data are continuously increasing due to more sensor networks with higher spatial and temporal resolution. To find appropriate data for analyses and especially for large scale models and simulations in this data explosion can take up to several months. The preprocessing of these heterogeneous datasets from different research disciplines to acquire a coherent dataset, can be done with a wide range of algorithms and tools. The outcome is a base dataset that is not reproducible and in consequence, neither are the resulting analyses [3, 9]. The datasets therefore do not obey the FAIR principles [13]. The V-FOR-WaTer web portal [11] aims to improve this situation by collecting data and metadata from a wide variety of sources and by offering preprocessed data.

1 Objectives

Huge effort is made to improve the availability of data by establishing data portals and data repositories. They range from project-specific data portals, that provide access to project datasets, via data repositories for already published datasets (e.g. GFZ data services, PANGAEA), through to data portals provided from federal state offices (e.g. LUBW, USGS, NASA). However, most of these portals do not allow for proper preprocessing of heterogeneous data to prepare scientific analyses.

The V-FOR-WaTer web portal represents an enhancement for data repositories to facilitate a standardized generation of reproducible base datasets. We want to give direct access to data and metadata to some of the most important and promising data sets for environmental research, that can be scaled and preprocessed with data uploaded from the scientists. These preprocessed datasets can be downloaded to the user's local PC or processed online using tools within the portal, provided either by V-FOR-WaTer or from

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029717 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13755 (CC BY-SA 4.0)

the scientific community. For efficiently working in and with the portal, we offer well documented open source tools and a simple workflow for the scientists, which can be saved and restored to perform reproducible preprocessing and analyses. All these components constitute the virtual research environment of V-FOR-WaTer.

Scientists can also upload their own data in combination with the corresponding metadata. The metadata is used for filtering, for tools to restrict which and how datasets can be combined and can be utilized to simplify data publication in established data repositories. To secure new and yet unpublished data sets, we have a fine-grained access management that enables scientists to upload data with an embargo period.

2 The V-FOR-WaTer Web Portal

The web portal is being developed with JavaScript and the Python web framework Django and is published open source [10]. These technologies are widely used and actively maintained by a large community and promise fast bugfixes and a long service life. The focus of the front end is on simple and intuitive usability. The design follows classical geographic information systems (GIS) with a map as main element on the start page (Fig. 1). Filtering of data takes place on the map and with a filter menu in the sidebar. To use data with restricted access, users can send requests to the data owner through the portal. For the identity management we use the external tool B2ACCESS [1], that is provided by EUDAT. With this identity provider researchers can reuse existing federated accounts from their home universities to log in.

The original database was designed with focus on the data sets of the Catchments as Organized Systems (CAOS) project [14]. These data sets are of special interest due to the large amount of data and their heterogeneity that represent a wide range of data used by the hydrological community. By now the metadata model is in its second version and it contains the necessary flexibility to hold all data types necessary for water and terrestrial environmental research and is continuously adapted to the needs of new data. Besides the CAOS data we already integrated data from the hydrology group of the Institute of Water and River Basin Management of KIT and from the Landesanstalt für Umwelt Baden-Württemberg (LUBW) [6]. The integration of data from more projects is work in progress. The data can be downloaded from the web portal in formats that are commonly used in environmental science, and the accompanying metadata follows the international standards of INSPIRE [4] and ISO19115 [5].

Instead of downloading, the data can also be processed within the web portal. For more flexibility to access the tools we use separate packages that are connected through a web processing server (WPS) [12], and not implemented as static part of the V-FOR-WaTer code. This way the selection of tools can easily be extended, and access is possible through the web portal or through an API from a local PC as well. The toolbox is constantly growing and already comprises among others the geostatistical toolbox SciKit GStat [7, 8]. Tools for scaling and uncertainty quantification for evapotranspiration (ET)

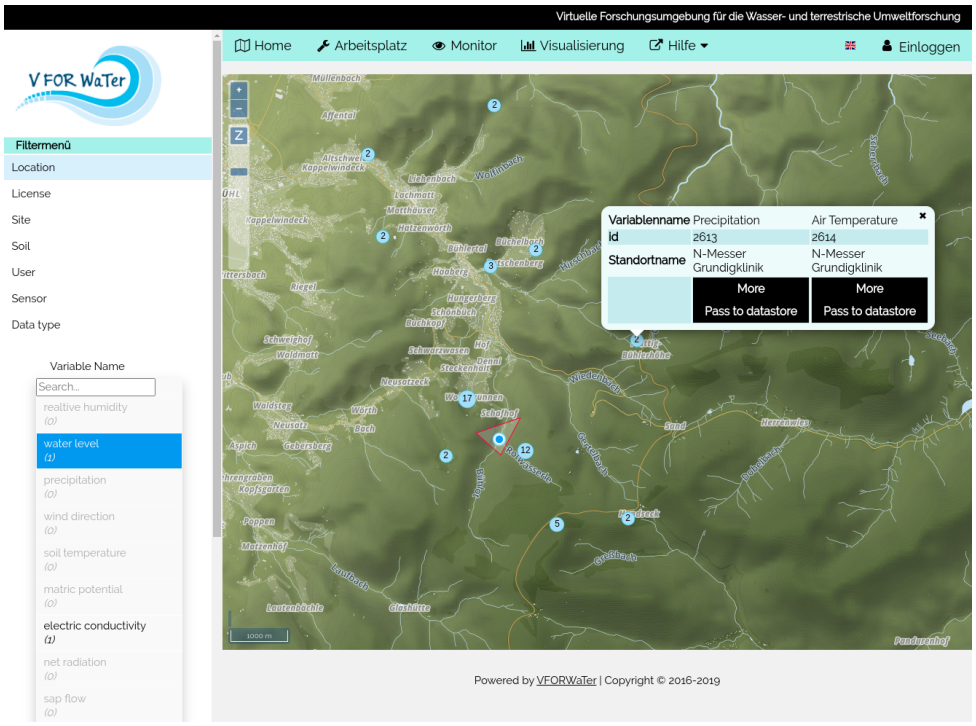


Figure 1: Screenshot of map and filter menu.

data are in development and will also be integrated in the V-FOR-Water portal. For complex workflows connecting several tools a graphical workflow editor, based on the JavaScript application framework Draw2D [2], is being integrated in the portal.

3 Conclusion

The V-FOR-WaTer web portal offers scientists centralized access to relevant data and tools, hence strongly supports them with their search, preparation, analysis and publication of data. The use of the web portal accelerates science and helps to make science reproducible.

Acknowledgments

V-FOR-WaTer has been funded by the ministry of science, research and arts of the state of Baden-Wuerttemberg, Germany.

Bibliography

[1] B2ACCESS Service, accessed April 27, 2021, https://eudat.eu/services/b2access.

[2] Draw2D, accessed May 21, 2021, https://www.draw2d.org/draw2d.

[3] Hutton, Christopher, Thorsten Wagener, Jim Freer, Dawei Han, Chris Duffy, and Berit Arheimer. "Most computational hydrology is not reproducible, so is it really science?" Water Resources Research 52.10 (2016), 7548–7555. DOI: https://doi.org/10.1002/2016WR019285.

[4] Infrastructure for spatial information in Europe (INSPIRE), accessed July 13, 2021, https://inspire.ec.europa.eu/inspire-directive/2

[5] ISO 19115-1:2014 Geographic information — Metadata, accessed July 13, 2021, https://www.iso.org/standard/53798.html

[6] Landesanstalt für Umwelt Baden-Württemberg (LUBW), accessed May 21, 2021, https://www.lubw.baden-wuerttemberg.de/startseite.

[7] Mälicke, Mirko, Helge D. Schneider, Sebastian Müeller, and Egil Möller. "mmaelicke/scikit-gstat: A scipy flavoured geostatistical variogram analysis toolbox (Version v0.5.0)" Zenodo (April 20, 2021). DOI: https://doi.org/10.5281/zenodo.1345584

[8] Müller, Sebastian and Lennart Schüler. "GeoStat-Framework/GSTools: v1.3.0 'Pure Pink" Zenodo (April 14, 2021). DOI: https://doi.org/10.5281/zenodo.4687075

[9] Stagge, James H., David E. Rosenberg, Adel M. Abdallah, Hadia Akbar, Nour A. Attallah, and Ryan James. "Assessing data availability and research reproducibility in hydrology and water resources" scientific data 6 (2019), 190030. DOI: https://doi.org/10.1038/sdata.2019.30.

[10] V-FOR-WaTer project on github, accessed April 27, 2021, https://github.com/VForWaTer/vforwater-portal.

[11] V-FOR-WaTer web portal demo instance, accessed April 27, 2021, https://portal.vforwater.de.

[12] Web Processing Service, accessed April 27, 2021, https://www.opengeospatial.org/standards/wps.

[13] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. "The FAIR Guiding Principles for scientific data management and stewardship." scientific Data 3 (2016), 160018. DOI: https://doi.org/10.1038/sdata.2016.18

[14] Zehe, Erwin, Uwe Ehret, L. Pfister, T. Blume, B. Schröder, M. Westhoff, C. Jackisch et al. “HESS Opinions: From response units to functional units: a thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments.” Hydrology and Earth System Sciences 18, 11 (2014), 4635–4655. DOI: https://doi.org/10.5194/hess-18-4635-2014

Concepts and services for the homogenization and management of file structures in collaborative neuroscientific projects

Thorsten Arendt[1], Achilleas Koutsou[2], Deepti Mittal[3], Keisuke Sehara[4], Rike-Benjamin Schuppner[4], Matthew Larkum[4], Thomas Wachtler[2] and Julien Colomb[4]

[1]Philipps-Universität Marburg
[2]Ludwig-Maximilians-Universität München
[3]Ruprecht-Karls-Universität Heidelberg
[4]Humboldt-Universität zu Berlin

With the GIN-Tonic tool, we provide researchers with a default file organization and file sharing system for research projects in order to facilitate research collaboration and lab management. In contrast to software developers, researchers mostly do not organize files according to a common standard. While data managers propose to design and follow such an organization, they fail at providing clear recommendations or examples to researchers; and there is no time specifically assigned to this task in the researcher's work. We believe that providing researchers with a commonly accepted folder tree structure template could make a huge difference in promoting data management and facilitating research collaboration. This paper presents the results of an initial survey run in three neuroscientific collaborative research centres in Germany (CRC 1315, CRC 1158, CRC/TRR 135), including a presentation of a new folder structure and its technical implementation in the GIN-Tonic application.

1 Introduction

Every day, researchers spend time doing file management on their computers (creating, downloading, naming, moving, saving, copying, reviewing, navigating, searching for, sharing, and deleting files and folders). While many different initiatives and tools have tried to improve file management (using tags, databases and search algorithms), the use of a folder tree structure appeared to be unavoidable and necessary [1]. In addition, both proponents of reproducible research and data management experts recommend researchers the use of an appropriate folder organizational structure [2, 3].

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029718 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13756 (CC BY-SA 4.0)

However, only few actually provide examples or attempt to bring uniformity in such structures (see [4] and [5] for exceptions).

As data managers of different institutions working with neurobiologists, we teamed up with the NFDI Neuroscience community to develop a new strategy to support researchers in data management. We hypothesize that implementing a homogeneous directory structure using a template could help researchers collaborate on their projects, and manage data and files better.

In a first step, we collected the feedback from 51 neuroscientists presented with two initial template drafts (see figure in the upper-left part of the corresponding poster), analysed their responses, and built an updated template (see figure in the right part of the corresponding poster and `https://doi.org/10.5281/zenodo.4410128` to download the template(s)). The new template takes three levels of projects organization into account (experiment, project, and laboratory), while staying fairly simple and flexible. This extended abstract ends by outlining the GIN-Tonic application, that brings some technical solution (based on the git sub module technology) to add flexibility and ease of use to the template.

2 The Survey

Two template structures

In order to obtain a practical flair in comments and feedback received, we provided two templates, both having a similar number of folders, but organized differently. The templates were obtained by analysing current research work flows discussed during a number of interviews with researchers of the involved consortia. The *5_top* template represented a more hierarchical structure, while the *9_top* template represented a more flat structure (see figure in the upper-left part of the corresponding poster). Then, we asked researchers to browse the folder tree while asking them to place or find specific files, hoping researchers will make themselves familiar with the template before giving us their feedback.

No clear preference for one or the other template

We ran the survey on the involved CRCs during autumn 2020 (using the lime survey application) and finally got 51 responses. In the meantime, we prepared the analysis code using synthetic data. The analysis was run on the final data to produce a reproducible report. In general, researchers reacted very positively to our project. Surprisingly, about half of the participants preferred one template, while the other half preferred the other one (see figure in the lower-left part of the corresponding poster; participants had to choose one or the other template). This preference was highly correlated with the similarity of the template to the structure they currently use (Pearson Chi-square, p-value = 2.23e-06, no

correction for multiple tests was performed). However, we could not identify any notable effect of career stage or research domain on this preference, or any effect of this preference on willingness to use such a template. In particular, computer scientists did not seem to differ from wet-lab researchers in these aspects.

We asked three questions about where they would save or search for specific documents in the different structures. Researchers were indicating different folders, and few chose the folder we designed for that data, showing that a too detailed folder structure seems to be rather inconvenient. On the other hand, researchers would navigate the repository to find specific files using similar strategies, suggesting that having a structure can be helpful, and may reduce the time to browse for specific information.

Issues

Most participants do see the advantages of having a standardized structure for their files, but were critical about the **cost to benefit ratio** of the process, especially for ongoing projects. Many mentioned that it would only reach full impact if the whole lab would be using it, emphasizing the advantage of such a system for collaborative work. In addition, they mentioned the time saved by not having to create a template for themselves, but only use an existing standard. While they were quite unanimous about using a template for new projects, they were sceptical about making a transition for ongoing projects, as the cost of transition might be too high. More generally, they were worried that learning a new work flow to use the template could be time consuming. In many cases, people mentioned that the **files are organized via experiments**, not at the project level. In particular, people tend to pool data and code in a single experiment folder. This is reminiscent of the two example structures given in the library carpentry course [2]. On the other hand, some **files are organized outside of project folders**, that is in particular places irrespective of the project they belong to. Many researchers reported having a folder for all conference reports or all manuscripts, for instance.

3 The revised Template

Template overview

Data management principles recommend (1) to keep all files related to a project in a single folder (this facilitates sharing of these files with the whole team working on the project, for instance), and (2) to manage data and code differently (this allows different version control systems, as well as independent sharing and reusing of data and code). We finally designed a template that follows these principles, but added some recommendations and technical solutions in order to permit users to have laboratory and experiment-level organization of their files, nevertheless.

The figure in top-right part of the poster shows the folder structure developed after analysing the answers of the survey. The template works mostly on the project level (one unique folder for all files related to one project). The experiment level is taken care by specifying several experiment sub folders when new experiments are started. By sharing specific sub folder independently in a cloud solution, one can reorganize information in cross-project directories that host sub folders coming from different projects. Note that both the creation of experiment sub folders and the creation of laboratory level organization is automated in GIN-Tonic.

Table 1: Our definition of the organisation levels.

Organisation level	Definition
Experiment	The unit of research involving a statistically dependent datasets that are analyzed together. It mostly produces one figure. Different experiments typically involve different methods, or different samples, and could be ending in different publications.
Research Project	The unit of research that address a specific research question. It can mostly be delimited by the team involved and typically produce a unique research paper.
Laboratory	Any organization that involves files from several projects. It can also be for a unique researcher, or for a consortium of laboratories.

The experiment level

We propose to keep data and code in different first level folders, and to create several new folders (new data, analysis, and figure folders) for each new experiment. In addition, some of these new folders may also follow their own templates. For instance, some researchers could use a specific BIDS standard template [6] for some imaging experiments.

The project and laboratory level

We propose to share some sub folders independently (for shared figures, report and conferences, and manuscripts) in order to be able to have them in the project folder or in a different folder structure merging information coming from different projects. One could for instance create a folder containing all manuscripts prepared in the lab (see figure in top-right part of the poster).

4 Sharing and automation

Automation possibilities

The creation of different sub folders for one experiment could easily be automated in your computer language of choice. An automation would make sure that the folder names are kept consistent for each experiment. Working on the laboratory level is more complex. If one wants to have cross-folder organization locally, one can work with *alias* folders, where the user can create short-cuts to specific folders using a different organization. The data exists only once and there is no issue of synchronization.

The expected use case is different though, and we expect some users to have certain files organized by projects and other users having the same files organised by file type. This requires to share sub folders independently and set the different instance of these sub folders to be linked together. This is pretty complex to set up using common cloud technology (DropBox-like) that cannot be easily automated. As explained below, the open source GIN-Tonic tool allows to set it up automatically, using the git sub-module technology.

GIN-Tonic implementation

GIN is the G-Node infrastructure. It is based on gogs, git and git annex technologies and brings non only most of the project management and coordination tools that made the success of open source software development, but also large file support and data publication. It is compatible with the git sub-module technology, where sub folders can be synchronized, shared and published independently of the other sub folders, while looking completely normal on one's computer. GIN brings therefore the possibility to publish sub-modules independently of each others, which will ease the opening of research data. It might also make the use of markdown and LaTeX to write manuscripts a straightforward choice, as these technologies can use git as a version control system. We are building an extension that will facilitate some administrative tasks and automate some complex work flows linked to the use of the template. We could not resist calling it *Tonic*, in reference to the vigor added to the GIN tool. The Tonic tool is still under development, and we are also working on the implementation of the Tonic concept for GitLab-based platforms (that is, using git and git-LFS), called LAB-Tonic.

The Tonic application creates a new project repository, clones the research folder structure (with some folders being created as sub-modules), and adds a script that will synchronize the repository and its sub-modules on a double click. It also adds sub-modules (*shared_figures*, *manuscripts*, and *report_conf*) to laboratory-wide repositories, automatically allowing laboratory level organization of some files. This means that for example a manuscript draft can be available in the project folder on the student computer, while the same data will be available in the *lab_manuscript* folder on the PI computer, both versions being synchronized with a unique version on the GIN server. Furthermore, Tonic

will also be able to add sub-modules and folders to the parent repository. For each experiment performed, a data sub-module will be created, so that the data can be curated and published independently of the other experiments, while other normal folders will be created (see figure in the upper-right part of the corresponding poster). A synchronization of the computer version will then bring these changes to the local version, showing the new folders ready to be filled with data, code, or figures.

5 Conclusions

With this project, we hope to provide the research community with a useful project folder structure template. A follow-up survey will tell us how wide it could be applied, and whether domain specific templates may be needed. The template will get its full power when used inside the GIN-Tonic application. Tonic will indeed automate several administrative tasks, like the production of sub folders upon new experiments, and come with a predefined rule for sharing one's files in the lab.

Acknowledgements

This project has been partially funded by Deutsche Forschungsgemeinschaft (DFG), project numbers 222641018 – CRC/TRR 135 TP INF, 255156212 – CRC 1158 TP Z01, and 327654276 – CRC 1315 TP Z.

Bibliography

[1] Jesse David Dinneen and Charles-Antoine Julien. "The Ubiquitous Digital File: A Review of File Management Research" Journal of the Association for Information Science and Technology 71, no. 1 (January 2020), https://doi.org/10/ghssbm.

[2] Library Carpentry. September 2019.https://librarycarpentry.org/lc-fair-research.

[3] Arnold, Becky, Louise Bowler, Sarah Gibson, Patricia Herterich, Rosie Higman, Anna Krystalli, Alexander Morley, Martin O'Reilly, Kirstie Whitaker, and The Turing Way Community. "The Turing Way: A Handbook for Reproducible Data Science" 2019. https://doi.org/10.5281/zenodo.3233986.

[4] Vuorre, Matti, and Matthew J. C. Crump. "Sharing and Organizing Research Products as R Packages" Behavior Research Methods, September 2020. https://doi.org/10/gg9w4c.

[5] Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. “Good Enough Practices in Scientific Computing“ PLOS Computational Biology 13, no. 6 (June 2017): e1005510. https://doi.org/10/gbkbwp.

[6] Gorgolewski, K., Auer, T., Calhoun, V. et al. “The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments“ Sci Data 3, 160044 (2016). https://doi.org/10.1038/sdata.2016.44.

Transparently Safeguarding Good Research Data Management with the Lean Process Assessment Model

Hendrik Geßner

Institute of Computer Science, University of Potsdam

In the last years, research data management moved into the spotlight of the scientific community. Organizations like the DFG and projects like FDMentor updated their guidelines to include current research software and data developments, while concepts like FAIR publishing gained traction interdisciplinarily. However, research guidelines often either take an abstract policy-driven perspective or solely focus on practices that, by omitting the underlying principles, become obsolete as the state-of-the-art advances. When looking at quality and evaluation methods in the industry, especially in systems and software development, models like CMMI, SPICE, or Six Sigma take a holistic approach by combining a process or life cycle perspective, clear goals, and target-oriented practices. These models were created with industrial processes in mind and applying them to research projects directly is counterintuitive.

We developed a Lean Process Assessment Model (LPAM) for research software and data that adheres to the CMMI framework. CMMI allows individual practices to be replaced by equivalent ones if they are suitable for achieving the overall objective. This framework allows LPAM to stay up-to-date, even when the state-of-the-art advances.

Together with interviews and discussions, existing guidelines and practices were analyzed and grouped into processes and goals. LPAM was developed with continuous researcher feedback. This procedure resulted in a discipline-agnostic model to manage and assess research projects, chairs, or organizations.

The different processes were assigned to CMMI's Maturity Levels, which rank each process's priority and give a clear improvement path. The model helps researchers in balancing goals and practices in their work. For assessing the state of a research project, we propose a peer-review based procedure that is intuitive and well-established for researchers.

We are convinced that LPAM narrows the gap between goals, principles, and practices and is a suitable tool to safeguard good research data management transparently.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: `https://doi.org/10.11588/heidok.00029719` veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13757 (CC BY 4.0)

1 Introduction

In August 2019, the DFG updated its *Guidelines for Safeguarding Good Research Practice*, a Code to "create a deeply rooted culture of research integrity at higher education institutions"[1] [5]. The precursor white paper *Safeguarding Good Scientific Practice* was published in 1998 with updates in 2013 [4]. For about 21 years, one of the most important scientific codes in Germany never even mentioned *research data*. Until 2019, the only requirement was to store primary data for 10 years, the one rule that every German scientist seems to be aware of when it comes to handling research data.

The DFG relied on separate subject-specific recommendations on the handling of research data, which are more in-depth[2]. These subject-specific recommendations vary in detail, and some subject areas, such as mathematics, are completely absent. Other projects like FDMentor [2] or groups like the Software Carpentry [22, 23] try to create a common understanding with regards to data management practices.

Yet, we find that there is a gap in practice when dealing with research data. While research guidelines often take a policy-driven perspective in order to remain current in the long term, workshops and handouts for researchers focus on concrete practices that quickly become outdated without reference to the underlying principles.

When looking at quality and evaluation methods in the industry, especially in systems and software development, models like CMMI, SPICE, or Six Sigma take a holistic approach by combining a process or life cycle perspective, clear goals, and target-oriented practices. Nevertheless, these models were created with industrial processes in mind and applying them to research projects directly proved to be counterintuitive.

We combined both the scientific and the industry perspective by using the industry model CMMI and re-implementing it with content from existing research guidelines and practices. Our goal was not to reinvent research data management practices, but to bring existing practices into a shape that enables transparent and reliable assessment of applied research data management in research projects and groups.

We developed the Lean Process Assessment Model (LPAM) to assess, support, and improve the quality of research processes within the CRC 1294 "Data Assimilation". It provides a common foundation of good scientific practices for research activities. LPAM is meant to be research field-agnostic so that all research projects are able to apply its base practices.

The presented model focuses on the certification of research projects. However, it is also a checklist, a self-assessment tool, and a learning resource. LPAM allows to certify good scientific practice, to uncover improvement capabilities and to reach defined quality levels. It is intended for peer review between research projects.

[1] `https://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice/`, retrieved on 16.05.2021

[2] `https://www.dfg.de/foerderung/antrag_gutachter_gremien/antragstellende/nachnutzung_forschungsdaten/`, retrieved on 16.05.2021

This document is kept short. For a visualization of the topics, please refer to the poster of the same title [8].

2 Existing Literature

There are previous publications with similar approaches. The closest one is the *RDM CMM* or *CMM4RDM*, which used the precursor of CMMI, CMM [17]. Our model differs primarily in that *LPAM* explicitly considers research software and research data while *RDM CMM* focuses solely on research data.

The German Aerospace Center developed software engineering guidelines which present an extensive overview of required practice while differentiating between four project sizes, but does not provide guidance on priorities [18].

Similar to the DFG's current *Guidelines for Safeguarding Good Research Practice* [5], CMMI and *LPAM* use a multi-level abstraction structure. While the DFG differentiates between *guidelines*, *explanations* and *detailed, subject-specific information*, CMMI separates into *process areas*, *processes*, *specific goals*, *specific practices*, and *detailed hints*. The DFG's subject-specific information are maintained as a list of links to further external resources on a web page, which makes them unsuitable for assessments. In LPAM, practices are part of the model. In addition, all processes are assigned to maturity levels, which provide a clear improvement path for research projects and groups. The DFG code misses this prioritization.

3 Development and Sources

The goal behind LPAM was not to reinvent research data management practices, but to bring existing practices into a shape that enables transparent and reliable assessment of applied research data management in research projects and groups. Therefore, the model is based on existing guidelines by funding agencies and RDM handbooks [2, 4, 5], practices from existing literature and knowledge bases [1, 6, 7, 9, 10, 13, 15, 16, 17, 18, 21, 22, 23], interviews with research projects and RDM experts, discussions with DFG reviewers for infrastructure projects, and impressions from a conference on research software engineering.

To get an understanding of the current software and data practices within the CRC 1294 "Data Assimilation", we conducted interviews with representatives from all CRC 1294 research projects and used that knowledge to find and assess suitable processes from the parent model. Comparing the results to known best practices in research made it clear that the focus of the parent model processes was mostly inadequate and far too abstract for our small-scale research projects. Therefore, we distilled new process areas from scientific literature and support materials, examining research life cycles while keeping the model framework.

We originally planned to base LPAM on the Software Process Improvement and Capability Determination (SPICE) [11, 12], an industry process assessment model mainly used within the automotive industry as Automotive SPICE 2.5 [20]. When it became apparent that our lack of understanding of SPICE and a general public knowledge resource gap on SPICE hindered our work, we decided to switch to the far more widespread Capability Maturity Model Integration (CMMI) [3]. This step proved to be more straightforward than expected as CMMI and SPICE share compatible roots. Today, LPAM follows the CMMI framework.

In all phases of LPAM creation, we gathered feedback from members of three CRC research projects. The feedback loop led to a less abstract, more hands-on model with detailed descriptions that guide researchers in their attempt to adhere to the presented practices.

4 Process Areas and Processes

LPAM contains three main process areas, each divided into several processes that bundle the specific goals and practices into coherent groups. The model also contains an assessment section that explains the ideas, methods and limitations behind the peer-review based procedure with regards to CMMI. The generic goals, which are the final part of LPAM, are still a work in progress as their implementation in scientific research projects is still unclear.

Following the CMMI framework, LPAM is divided into three process areas: *data*, *software*, and *project management / support*, with the main emphasis being on the data and software process areas. The process layout of the data process area follows the research data life cycle as presented by the UK Data Archive/UK Data Service [19], including *plan*, *collect*, *process/analyse*, *publish/share* and *archive*. *Reuse* was skipped because of its unique status in the research data life cycle, but its ideas were incorporated into the *collect* process. The specific goals, practices and hints in the data process area are based on the research data management field which has made huge steps forward in the last few years.

In contrast to research data management, good research software management practices are far less developed. We incorporated practices from multiple sources and combined them into the software process area. The structure of *planning*, *implementation*, *verification*, *automation and tools*, *publishing*, and *archiving* loosely follows the *DLR Software Engineering Guidelines* [18]. There are efforts to apply the FAIR principles to software [7, 9], as well as documented good practices based on experience from training researchers [22, 23].

The project management process area mainly consists of training and infrastructure. Both topics result from requirements formulated in other parts of LPAM. The training process lists skills that are prerequisites for practices in the data, software or manuscript process areas, such as FAIR data publishing or DRY programming techniques.

The infrastructure process collects necessary services that other practices build upon, such as the provision of an archive. All in all, the project management process area is focused on services that the research community should provide and implement.

5 Maturity Levels

The goal of LPAM is to improve on and safeguard a set of good scientific practices. CMMI is built on the idea that planned improvement can only be achieved by structurally identifying and eliminating weaknesses. Projects would start from *initial* (level 1), where processes are unpredictable and success is random, and develop into more predictable, less random entities [14]. Both CMMI and LPAM focus on stabilising the processes in the first levels, while optimizing results and processes at higher levels. Conducting a CMMI appraisal results in a rating.

CMMI provides a staged representation with five maturity levels. These are called *initial* (level 1), *managed* (level 2), *defined* (level 3), *quantitatively managed* (level 4), and *optimizing* (level 5). Each maturity level has defined characteristics and scopes. With LPAM, these scopes translate to *individual researchers* (for managed), *chairs and projects* (for defined), *quantitative metrics* (for quantitatively managed), and *process improvement* (for optimizing). LPAM also establishes an additional level *legal and DFG minimum* to pool minimal practices that always have to be adhered to first.

Each maturity level dictates a specific set of processes which a project has to tackle successfully. Together with the requirement that maturity levels have to be reached one after another, this results in a clear improvement path. It provides a precise understanding of which process to tackle first, and which process to tackle later on at a higher level.

We decided to only support the staged representation in LPAM. We are still working on maturity levels 4 and 5 as the associated generic practices of CMMI are very ambitious and we are yet unsure whether it is realistic for small research teams to reach them.

6 Conclusions

The LPAM presented here bridges a gap between policies and practices in research data management. It combines research guidelines and practices with the industry model CMMI. This approach makes good scientific practice clearly improvable and transparently certifiable. The feedback loop led to a less abstract, more hands-on model with detailed descriptions that guide researchers in their attempt to adhere to the presented practices.

Because LPAM is guided by the current state of research, several questions remain unanswered. In contrast to research data management, good research software management practices are far less developed. For instance, there is no consensus on the application of FAIR data principles to software.

The generic goals of LPAM are still a work in progress as their implementation in scientific research projects is still unclear. Interviews with the pilot research projects could not conclusively clarify how an implementation of the generic goals would have to look like. Similar difficulties arose for the two highest maturity levels, *quantitatively managed* and *optimizing*.

An application of the presented model to research projects outside the pilot projects is still pending. Nevertheless, we are convinced that LPAM narrows the gap between goals, principles, and practices and is a suitable tool to safeguard good research data management transparently.

Acknowledgements

The research of Hendrik Geßner has been partially funded by the Deutsche Forschungsgemeinschaft (DFG) - Project-ID 318763901 - SFB1294.

ORCID ID

- Hendrik Geßner https://orcid.org/0000-0002-7786-2587

Bibliography

[1] Biernacka, K. *Wie publiziere ich Forschungsdaten?* Zenodo, 2018. https://doi.org/10.5281/zenodo.1440956.

[2] Biernacka, K., P. Buchholz, S. A. Danker, D. Dolzycka, C. Engelhardt, K. Helbig, J. Jacob, J. Neumann, C. Odebrecht, C. Wiljes, and U. Wuttke. *Train-the-Trainer Konzept zum Thema Forschungsdatenmanagement.* Zenodo, 2020. https://doi.org/10.5281/zenodo.4322849.

[3] CMMI Product Team. *CMMI®for Development, Version 1.3.* Technical report, Software Engineering Institute, 2010.

[4] Deutsche Forschungsgemeinschaft. *Sicherung guter wissenschaftlicher Praxis*, pp. 1–109. John Wiley & Sons, Ltd, 2013. https://doi.org/10.1002/9783527679188.oth1.

[5] Deutsche Forschungsgemeinschaft. *Guidelines for Safeguarding Good Research Practice. Code of Conduct.* Zenodo, 2019. http://doi.org/10.5281/zenodo.3923602.

[6] Dietrich, C. and D. Lohmann. "The dataref versuchung." *ACM SIGOPS Operating Systems Review* 49, no. 1 (2015): 51–60. https://doi.org/10.1145/2723872.2723880.

[7] Erdmann, C., N. Simons, R. Otsuji, S. Labou, R. Johnson, G. Castelao, B. V. Boas, A.-L. Lamprecht, C. M. Ortiz, L. Garcia, M. Kuzak, P. A. Martinez, L. Stokes, T. Honeyman, S. Wise, J. Quan, S. Peterson, A. Neeser, L. Karvovskaya, O. Lange, I. Witkowska, J. Flores, F. Bradley, K. Hettne, P. Verhaar, B. Companjen, L. Sesink, F. Schoots, E. Schultes, R. Kaliyaperumal, E. Tóth-Czifra, R. de Miranda Azevedo, S. Muurling, J. Brown, J. Chan, N. Quigley, L. Federer, D. Joubert, A. Dillman, K. Wilkins, I. Chandramouliswaran, V. Navale, S. Wright, S. Di Giorgio, M. Fasemore, K. Förstner, T. Sauerwein, E. Seidlmayer, I. Zeitlin, S. Bacon, K. Hannan, R. Ferrers, K. Russell, D. Whitmore, and T. Dennis. *Top 10 FAIR Data & Software Things*. Zenodo, 2019.

[8] Geßner, H. "Transparently Safeguarding Good Research Data Management with the Lean Process Assessment Model." In *E-Science-Tage 2021: Share Your Research Data*. Heidelberg, 2021. https://doi.org/10.11588/heidok.00029719.

[9] Hong, N. C. and D. S. Katz. *FAIR enough? Can we (already) benefit from applying the FAIR data principles to software?*. Figshare, 2018. https://doi.org/10.6084/m9.figshare.7449239.

[10] Hrynaszkiewicz, I. and M. J. Cockerill. "Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals." In *BMC research notes* 5, no. 494 (2012), editorial. https://doi.org/10.1186/1756-0500-5-494.

[11] ISO/IEC JTC 1/SC 7 Software and systems engineering. *Information technology – Process assessment – Part 1: Concepts and vocabulary*. Technical report 15504-1:2004, ISO/IEC, 2004.

[12] ISO/IEC JTC 1/SC 7 Software and systems engineering. *Information technology – Process assessment – Concepts and terminology*. Technical Report 33001:2015, ISO/IEC, 2015.

[13] Klimpel, P. *Folgen, Risiken und Nebenwirkungen der Bedingung "nicht-kommerziell - NC"*. Wikimedia Deutschland, iRights.info, CC DE, 2012.

[14] Kneuper, R. *CMMI: Verbesserung von Softwareprozessen mit Capability Maturity Model Integration* (1st ed.). Heidelberg: dpunkt.verlag, 2003.

[15] Lauber-Rönsberg, A., P. Krahn, and P. Baumann. *Gutachten zu den rechtlichen Rahmenbedingungen des Forschungsdatenmanagements im Rahmen des DataJus–Projektes*. 2018.

[16] Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. "The preregistration revolution." *Proceedings of the National Academy of Sciences of the United States of America* 115, no. 11 (2018): 2600–2606. https://doi.org/10.1073/pnas.1708274114.

[17] Qin, J., K. Crowston, and A. Kirkland. "Pursuing Best Performance in Research Data Management by Using the Capability Maturity Model and Rubrics." *Journal of eScience Librarianship* 6, no. 2 (2017): e1113. https://doi.org/10.7191/jeslib.2017.1113.

[18] Schlauch, T., M. Meinel, and C. Haupt. *DLR Software Engineering Guidelines: Version: 1.0.0.* Zenodo, 2018. https://doi.org/10.5281/zenodo.1344612.

[19] UK Data Service. *Research data lifecycle.* UK Data Service, 2019.

[20] VDA QMC Working Group 13 / Automotive SIG *Automotive SPICE Process Assessment / Reference Model*, version 3.1. VDA QMC, 2017.

[21] Wilkinson, M. D., M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3, no. 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

[22] Wilson, G., D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson. "Best practices for scientific computing." *PLoS biology* 12, no. 1 (2014): e1001745. https://doi.org/10.1371/journal.pbio.1001745.

[23] Wilson, G., J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal. "Good enough practices in scientific computing." *PLoS computational biology* 13, no. 6 (2017): e1005510. https://doi.org/10.1371/journal.pcbi.1005510.

Der Zertifikatskurs Forschungsdatenmanagement als adaptierbares Aus- und Weiterbildungsangebot

Mirjam Blümm[1,2], Konrad U. Förstner[1,3], Marvin Lanczek[4], Birte Lindstädt[3], Rabea Müller[3], Ulrike Nickenig[5], Stephanie Rehwald[5], Benjamin Slowig[5] und Jessica Stegemann[5]

[1]Institut für Informationswissenschaft, Technische Hochschule Köln
[2]Advanced Media Institute, Technische Hochschule Köln
[3]ZB MED - Informationszentrum Lebenswissenschaften
[4]ZBIW - Zentrum für Bibliotheks- und Informationswissenschaftliche Weiterbildung, Technische Hochschule Köln
[5]Landesinitiative für Forschungsdatenmanagement in NRW – fdm.nrw

Mit dem ersten, berufsbegleitenden Zertifikatskurs „Forschungsdatenmanagement“ (FDM) wird ab August 2021 dem stetig wachsenden Bedarf an qualifiziertem Personal im Kontext FDM begegnet. Im Rahmen verschiedener Module erhält die heterogene Zielgruppe von Beschäftigten aus Bibliotheken, Rechenzentren, Forschung und der Forschungsförderung eine fundierte Grundausbildung; zudem ist eine individuelle Spezialisierung für ein FDM-Themengebiet möglich. Einen Überblick bietet das Poster, welches bei den E-Science-Tagen 2021 vorgestellt wurde [1].

1 Einleitung: Ausbildungsmöglichkeiten zum FDM in Deutschland

Der Bedarf an qualifiziertem Personal im Bereich Datenmanagement wurde nicht zuletzt vom Rat für Informationsinfrastrukturen wiederholt festgestellt und die Schaffung entsprechender Qualifizierungsmöglichkeiten mit wachsender Dringlichkeit angemahnt [2]. Die Fachhochschule Potsdam und die Humboldt-Universität zu Berlin bieten derzeit mit ihrem kooperativen Masterstudiengang „Digitales Datenmanagement“ (seit Sommersemester 2020) deutschlandweit das erste Format an, welches dezidiert auf die im digitalen Forschungsprozess nötigen Kompetenzen von Data Stewards abgestimmt ist [3].

Das Poster zu diesem Beitrag ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00030639 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13758 (CC BY-SA 4.0)

Zugleich eröffnet er eine Weiterbildungsperspektive: Module in den Bereichen Grundlagen, Technologen und Methoden des Datenmanagements können alternativ auch mit einem Zertifikat abgeschlossen werden [4].

In Nordrhein-Westfalen (NRW) hat die Technische Hochschule Köln (TH Köln) den Bachelorstudiengang „Data and Information Science“ (seit Wintersemester 2018/19) und den konsekutiven Masterstudiengang „Digital Sciences“ (ab Sommersemester 2022 geplant) eingeführt, der einen Fokus auf Methoden und Techniken im Umgang mit digitalen Daten, z. B. Textmining-Verfahren, bei der Generierung, Analyse und Visualisierung von Daten legt. Eine individuelle Spezialisierung als „Data Analyst“ oder „Data Librarian“ ist möglich [5].

Ein dezidierter Zertifikatskurs, der berufsbegleitend absolviert werden kann, erscheint besonders für Personengruppen, die bereits im Berufsleben stehen, eine interessante Option zu sein. Hiermit wird eine gezielte Weiterqualifikation ermöglicht, ohne ein komplettes Hochschulstudium absolvieren zu müssen. Der Umfang geht dabei deutlich über bisher etablierte Schulungsangebote wie den Train-the-Trainer-Workshop zum Forschungsdatenmanagement [6] oder die auf bibliothekarische Bedürfnisse zugeschnittenen „Library Carpentry“-Workshops [7] hinaus, erlaubt also eine tiefere Beschäftigung mit den Inhalten und eine individuelle Schwerpunktsetzung. Das ZBIW hat hierzu den Zertifikatskurs „Data Librarian“ [8] entwickelt und nach der ersten Durchführung sehr positives Feedback von Teilnehmenden erhalten [9]. Durch die Vermittlung von breit gefächertem Überblickswissen zu verschiedenen Themenfeldern des Kurses wie Suchmaschinentechnologie, Programmierung, Maschinelles Lernen, Information Retrieval, Statistik, Datenvisualisierung und Open Access, können sich Beschäftigte wissenschaftlicher Bibliotheken für neue Arbeitsfelder qualifizieren und dieses Wissen direkt in ihren beruflichen Alltag integrieren.

Der Zertifikatskurs „Forschungsdatenmanagement“, der ab August 2021 von der ZBIW durchgeführt wird, ist ein neuer Baustein berufsbegleitender Weiterqualifizierung, der im vielschichtigen Thema FDM eine Spezialisierung für verschiedene Zielgruppen erlaubt.

2 FDM als Kernthema für die forschungsnahen Infrastruktureinrichtungen und die Forschungsbereiche

Immer mehr Hochschulen und Forschungseinrichtungen sind in den vergangenen Jahren dazu übergegangen, Serviceangebote im Bereich FDM für verschiedene Interessengruppen, vor allem für Forschende (Schwerpunkt bei den Promovierenden) einzurichten. Hierzu zählen beispielsweise Beratungen zur Erstellung von Datenmanagementplänen, zur Veröffentlichung von Forschungsdaten und Publikationsformen oder zur Vergabe von Metadaten. Darüber hinaus werden auch technische Infrastrukturen bzw. Tools bereitgestellt, mit denen aktives FDM betrieben werden kann (z. B. Elektronische Laborbücher) oder Forschungsdaten – gemäß guter wissenschaftlicher Praxis [10] – gespeichert werden können. Akteure sind hierbei insbesondere die Bibliotheken, Rechenzentren und die Forschungsförderung.

Aus den Aktivitäten dieser Infrastruktureinrichtungen heraus etablieren sich in der Regel nach und nach spezialisierte FDM-Kontakt- und Beratungsstellen an Hochschulen und Forschungseinrichtungen.

Um diesem neuen Aufgabengebiet kompetent begegnen zu können, ist es wichtig, die Beschäftigten dieser Einrichtungen im komplexen Feld des Forschungsdatenmanagements weiterzubilden. Um ein an den Bedarfen der Forschenden orientiertes Angebot bereitstellen zu können, ist ein breites generisches wie zunehmend auch fachspezifisches Wissen zu digitalen Forschungstätigkeiten, FDM in allen Phasen des Lebenszyklus von Daten sowie zu geeigneten Vermittlungsmethoden notwendig. Mit Blick auf NRW und ausgehend von der Landesinitiative für Forschungsdatenmanagement – fdm.nrw [11] wurde Ende 2019 eine Umfrage zu spezifischen Weiterbildungsbedarfen der Beschäftigten aus forschungsnahen Infrastruktureinrichtungen aufgesetzt und durchgeführt. Anhand von leitfadengestützten Gesprächen wurden bis Anfang 2021 an insgesamt 44 Einrichtungen in NRW der FDM-Sachstand und die Wünsche im Bereich Weiterbildung von 85 Personen eingeholt.

Die Befragten arbeiten in Bibliotheken, Rechenzentren und der Forschungsförderung; ferner zählen auch Fachwissenschaftler:innen zu den potentiellen Teilnehmenden von Weiterbildungsangeboten. Diese Personen(-gruppen) haben je nach hochschulinterner Einrichtung spezifische Qualifikationen und unterschiedliche Berührungspunkte mit dem Bereich FDM im Rahmen ihrer täglichen Arbeit. Zudem sind sie in unterschiedlicher Form und Ausführung an der Etablierung bzw. Erweiterung von FDM-Diensten an ihren Standorten beteiligt. So sind einige Personen in etablierten bzw. im Aufbau befindlichen FDM-Kontaktstellen (mit stark variierenden Stellenanteilen) tätig, während andere nur vereinzelte Aspekte als „FDM-Beauftragte“ ihrer Einrichtung bearbeiten – dies häufig nur begleitend und mit einem sehr geringem Stellenanteil neben ihren regulären Aufgaben. Insgesamt weisen die im FDM-Kontext beschäftigten Personen sehr heterogene Bildungshintergründe auf: eine abgeschlossene Berufsausbildungen (z. B. Fachangestellte für Medien- und Informationsdienste oder Fachinformatiker:innen), ein abgeschlossenes Studium (Bachelor, Master etc.) und/oder einen erworbenen Doktorgrad. Diese berufliche wie wissenschaftliche Varianz spiegelt sich in den ermittelten Weiterbildungsbedarfen wieder:

- Überblick zu rechtlichen Aspekten
- Überblick über FDM-Tools und -Angebote
- inhaltliche Verknüpfungen verschiedener FDM-Aspekte,
- Strategien zu Vernetzung und Austausch zwischen Forschenden und FDM-Servicestellen
- didaktische Konzepte und Vermittlungsformate

Neben den Bedarfen aus den Infrastruktureinrichtungen, die jeweils auf die Anliegen der Forschenden am Standort rekurrieren, bestehen daneben eine Reihe von übergreifenden Aufgaben, die in größeren Verbünden geleistet werden können. Die Nationale Forschungsdateninfrastruktur (NFDI) [12] soll für die gesamte deutsche Forschungsgemeinschaft Grundlagen für die effiziente Speicherung, Analyse und Nachnutzung von Forschungs-

daten bieten. Da die NFDI-Konsortien fachlich orientiert sind, sind deren Bedarfe oftmals darauf ausgerichtet, sogenannte „(embedded) Data Stewards“ bzw. „Data Scientists“ einzustellen, also Fachwissenschaftler:innen mit vorhandenen Kompetenzen im Bereich Forschungsdatenmanagement oder Datenanalyse. Die Kompetenzen im Bereich Forschungsdatenmanagement umfassen dabei drei Kernpunkte: 1) Leitlinien und Standards – u. a. Sichtung und Formulierung von Publikationsleitlinien und fachbezogenen Metadaten- und semantischen Standards für die verschiedenen Fachbereiche, dies wird in Modul 3 für sechs unterschiedliche Fachdisziplinen vermittelt 2) Forschung – mit dem Fokus auf der Unterstützung der Forschenden in der praktischen Arbeit, z. B. bei der Umsetzung von Policies sowie bei der Nutzung von Diensten und Softwarewerkzeugen sowie 3) Infrastruktur – mit dem Akzent auf der Kommunikation von Feedback der Nutzenden mit Softwareentwickler:innen und Serviceprovidern, sodass die Services den Bedürfnissen der Forschenden angepasst werden können. Neben diesem Profil gibt es einen Bedarf an Softwareentwickler:innen (z. B. Research Software Engineers) mit Erfahrungen in bestimmten Programmiersprachen bzw. mit bestimmten Tools.

Sowohl die aus den Konsortien formulierten Interessen an fachspezifisch ausgerichtetem geschultem Personal als auch die aus den Interviews mit Beschäftigen der Infrastruktureinrichtungen in NRW destillierten Zielgruppen- und Bedarfsanalysen lieferten die Grundlage für die strukturelle und inhaltliche Ausgestaltung des Kurses.

3 Der Zertifikatskurs als strukturiertes und berufsbegleitendes Weiterbildungsangebot

Um dem komplexen Handlungsfeld FDM zu begegnen und den angesprochenen Zielgruppen eine strukturierte wie zertifizierte Vermittlung von FDM-bezogenen Informationen, Tools und Erfahrungswerten anzubieten, haben die TH Köln, ZBIW, ZB MED und die Landesinitiative fdm.nrw 2019 begonnen, gemeinsam ein modulbasiertes Konzept für den Zertifikatskurs „Forschungsdatenmanagement“ [13] zu entwickeln.

Der Zertifikatskurs besteht aus 9 Modulen, die sich in 3 Basismodule, 5 Aufbaumodule und 1 Projektmodul aufteilen. Insgesamt hat der Zertifikatskurs einen Umfang von 8 ECTS (240 Std.), die in rund 10 Monaten absolviert werden. Neben dem Organisationsteam des Zertifikatskurses wurden weitere Referent:innen aus ganz Deutschland akquiriert, die einzelne Module (analog und digital) inhaltlich gestalten und auch die umfassenden Selbstlernphasen moderieren. Es können 15 Personen aus Einrichtungen innerhalb von NRW an dem Kurs teilnehmen. Zur Förderung der Breitenwirkung des Zertifikatskurses werden über die Digitalisierungsoffensive des Landes NRW Stipendien zur Verfügung gestellt, die einen Großteil der Teilnahmegebühr kompensieren.

Inhaltlich werden im Rahmen der Basismodule drei grundlegende Themengebiete behandelt: Im Rahmen des Moduls 1 „Grundlagen des FDM“ wird das erfolgreiche FDMentor-Konzept des „Train-the-Trainer-Workshops zum FDM“ [6] integriert. Modul 2 „Open Science & rechtliche Aspekte“ nimmt verschiedene ethische wie auch rechtliche Themen

in den Fokus und liefert Orientierungen u. a. zu Open Science, Urheberrecht, Lizenzierung oder Datenschutz. Modul 3 „Forschung, Forschungsdaten & Forschungsdatenmanagement in verschiedenen Fachgebieten“ gibt einen praxisbezogenen Einblick in fachspezifische Forschungsprozesse, die damit zusammenhängenden Maßnahmen im Forschungsdatenmanagement und die Entwicklungen einzelner NFDI-Konsortien.

Den Basismodulen folgen fünf Aufbaumodule, von denen die Teilnehmenden (basierend auf ihren Interessen) vier belegen müssen. Modul 4 „Hacken & Experimentieren mit Daten“ beinhaltet einen 2-tägigen Library-Carpentry-Workshop, in dessen Rahmen Werkzeuge wie Unix Shell, Python oder GitHub vermittelt werden. Im Modul 5 „(Meta-)Daten verwalten & teilen“ liegt der Fokus auf den Forschungsdaten selbst und den damit verbundenen Aspekten, die für eine weitere Nutzung bzw. Verbreitung elementar sind (u. a. Datenformate, Metadaten oder Persistent Identifiers). Im Rahmen des Moduls 6 „Technische Infrastruktur“ erfahren die Teilnehmenden, welche technischen Lösungen zur Verfügung stehen, um Forschungsdaten zu speichern, sie langfristig zu archivieren, auffindbar und verfügbar zu machen (u. a. Speicher-Systeme, Langzeit-Archive bzw. -verfügbarkeit und Repositorien).

Eine weitere Perspektive verfolgt Modul 7 „Daten- & Projektmanagement in der Forschung“, in dem ein Forschungsprojekt von der Planung bis zur Umsetzung betrachtet wird. Elementar sind hierbei Aspekte wie Datenmanagementpläne (DMP), Vorgaben zu Fördermitteln oder auch Tools zum Projektmanagement. Modul 8 „FDM-Beratung & Schulung“ widmet sich mit dem Blick auf die Interaktion mit verschiedenen Zielgruppen (u. a. Forschende oder Promovierende) einem zentralen Aufgabenfeld für FDM-Akteure. So werden u. a. Techniken der Gesprächsführung, Theorien und Strategien oder auch didaktische Konzepte behandelt, mithilfe derer unterschiedliche Beratungssettings oder Schulungsformate gestaltet werden können. Im Rahmen des Modul 9 „Projektmodul“ haben die Teilnehmenden die Möglichkeit, ein FDM-bezogenes Vorhaben praktisch durchzuführen und zu dokumentiert.

Mit Abschluss des Zertifikatskurses sollen die Teilnehmenden

- wichtige Aspekte des FDM identifizieren und diese kompetent sowie didaktisch einfallsreich vermitteln können;
- in der Lage sein, rechtliche Aspekte einzuordnen;
- Kenntnisse über Funktion und Handhabung forschungsunterstützender Software erlangt haben;
- elementare Aspekte der Datendokumentation als Grundlage der Speicherung und Nachnutzung von Daten kennen;
- Grundzüge des Projektmanagements nachvollziehen und
- das erworbene Wissen in der Praxis generisch und fachspezifisch anwenden können.

4 Ausblick

Die erste Gruppe startet im August 2021. Ein erstes Zwischenfazit ist für Anfang 2022 angedacht, an welches sich eine umfassende Evaluation nach Ende des ersten Durchlaufs anschließen wird. Geplant ist, mindestens zwei weitere Kurse 2022-2023 anzubieten. Hierfür werden auf Grundlage der Evaluation die Module ggf. angepasst, aktualisiert und erweitert.

Die Deckung des Bedarfes an ausgebildetem Personal in Infrastruktureinrichtungen am Standort und in den Konsortien der NFDI ist eine nationale Aufgabe, der mit dem Zertifikatskurs „Forschungsdatenmanagement“ und seiner berufsbegleitenden Ausrichtung begegnet wird. Als regional konzipiertes Angebot aus NRW werden mit dem Pionierprojekt Erfahrungen gesammelt, die auf überregionale und nationale Strategien für die FDM-Ausbildung übertragen werden können.

Literaturverzeichnis

[1] Blümm, Mirjam et al.: Zertifikatskurs Forschungsdatenmanagement - ein adaptierbares Aus- und Weiterbildungsangebot, Konferenzbeitrag, E-Science-Tage Heidelberg (2021)urn:nbn:de:bsz:16-heidok-306934

[2] Rat für Informationsinfrastrukturen. “Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland.” (2016) urn:nbn:de:101:1-201606229098; Rat für Informationsinfrastrukturen: “Digitale Kompetenzen - dringend gesucht! Empfehlungen zu Berufs- und Ausbildungsperspektiven für den Arbeitsmarkt Wissenschaft.” (2019) urn:nbn:de:101:1-201908071103224970621 8

[3] https://www.ddm-master.de/, zuletzt abgerufen am 10.08.2021

[4] https://www.ddm-master.de/ddm-als-weiterbildung/, zuletzt abgerufen am 10.08.2021

[5] https://www.th-koeln.de/studium/data-and-information-science-bachelor_52793.php, zuletzt abgerufen am 10.08.2021

[6] Biernacka, K., P. Buchholz, D. Dolzycka, K. Helbig, J. Neumann, C. Odebrecht, C. Wiljes, und U. Wuttke. “Train-the-Trainer Konzept zum Thema Forschungsdatenmanagement.” Version 3.0. (2020), https://doi.org/10.5281/zenodo.3938533

[7] Martin-Konle, C. “Library Carpentry - neues Werkstatt-Weiterbildungsformat für Bibliotheken: VDB-Fortbildung in Marburg: Werkzeuge und Konzepte zum praktischen Umgang mit Daten im Berufsalltag.” O-Bib. Das Offene Bibliotheksjournal, 5(3) (2018): 162–165. https://doi.org/10.5282/o-bib/2018H3S162-165

[8] https://www.th-koeln.de/weiterbildung/zertifikatskurs-data-librarian_63393.php, zuletzt abgerufen am 10.08.2021

[9] ZBIW. Zentrum für Bibliotheks- und Informationswissenschaftliche Weiterbildung. „Jahresbericht 2020“, S. 50–55, https://www.th-koeln.de/mam/downloads/deutsch/weiterbildung/zbiw/allgemein/zbiw_jahresbericht_2020.pdf, zuletzt abgerufen am 10.08.2021

[10] Deutsche Forschungsgemeinschaft. „Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Kodex.“ (2019). http://doi.org/10.5281/zenodo.3923602

[11] https://www.fdm.nrw/, zuletzt abgerufen am 10.08.2021

[12] https://www.nfdi.de/, zuletzt abgerufen am 10.08.2021

[13] https://www.th-koeln.de/weiterbildung/zertifikatskurs-forschungsdatenmanagement_82048.php, zuletzt abgerufen am 10.08.2021

Managing large research data with SDS@hd

Sabine Richling, Sven Siebler, Alexander Balz, Robert Kühl and Martin Baumann

University Computing Centre, Heidelberg University

The scientific data storage 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 for individuals and groups. The service is operated by the Heidelberg University Computing Centre and running in production since 2017 with a continuously growing number of users and storage projects. Access management can be done via a predefined set of roles and also based on access control lists on the filesystem level enabling researchers to share data in a collaborative fashion.

1 Introduction

In many fields of research, the capacity of generated scientific data is enormous and continuously growing. This is a consequence of technical progress in data generating devices (e.g. high-throughput microscopes, telescopes and genome sequencers) and increasing performance capability of computer systems (e.g. high-performance compute clusters or cloud systems). Research projects have additional requirements including group and access management for co-operational setups, data protection for projects dealing with sensitive data, the demand in flexibly sharing data with others, data publication and long-term preservation.

The scientific data storage service SDS@hd [1, 2] builds on top of the second generation hardware of the "Large Scale Data Facility" (LSDF2) offering a fast and large-capacity storage backend for such needs. The LSDF2 is part of the state of Baden-Württemberg's concept for data-intensive services bwDATA [3].

SDS@hd has been developed to improve the value of available data storage resources in the context of research projects. It offers processes for user registration, role management, access management and collection of information for reporting. The service is open to all scientists at Baden-Württemberg's universities in the sense of a "Landesdienst".

SDS@hd is tailored primarily to those phases in the research data life cycle in which frequent and fast data accesses are necessary. In this phase, users can profit from the fast direct connection to the local high-performance compute cluster "bwForCluster MLS&WISO" [4] and can share their data with other registered users of SDS@hd.

For data publication or long-term preservation of research data, appropriate platforms must be used in addition to SDS@hd, see Fig. 1. Several community-specific platforms

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13759 (CC BY-SA 4.0)

Figure 1: For data publication or archiving, data has to be transferred from SDS@hd to dedicated services. The institutional services heiDATA and heiARCHIVE of Heidelberg University facilitate open-access data publication and long-term preservation and are connected to SDS@hd.

exist for many disciplines and are often preferred over alternatives (e.g. institutional platforms). This is the case since appropriate features for the respective community are available including support for specific metadata standards which simplifies the locating and reuse of data. There are needs where community specific platforms cannot or should not be used. For such situations, many universities and other research institutions are operating institutional repositories for their members. In this context, Heidelberg University offers the institutional open-access data publication service heiDATA (`https://heidata.uni-heidelberg.de`) and the upcoming institutional long-time preservation service heiARCHIVE [5]. Both services are connected to SDS@hd and allow for metadata management and allocation of persistent identifyers. Whenever data is transmitted to and registered in such platforms, more or less requirements related to the data structure, file formats, and standardized metadata have to be fulfilled which most often requires some manual steps during the ingest process. For the transfer of the data from or to such services, different access protocols are available in SDS@hd and additional technologies will be added as required, cf. [6].

Subsequently in this publication, we outline the access management followed by some statistics of users and projects of SDS@hd. We briefly outline how the service is used in different research projects based on a selection of use-cases. We present a statistic of publications by users of SDS@hd and, finally, describe some planned developments.

2 Access management

For the management of the storage resources and accesses, storage projects are introduced as the organizational units of SDS@hd, denoted “Speichervorhaben” or short “SV”. One such storage project consists of one storage share with a defined quota and corresponds to a dedicated group of users. Members of the same storage project are able to share data easily while the access of non-members by default is not possible. There is a responsible person for each storage project who takes responsibility for the stored data and also for the

reporting (e.g. for evaluations by the DFG). The responsible person can manage members and roles for his or her storage project via a web management tool. A fine-grained access management within one SV on the level of user roles is available. For example, a guest role allows only restricted access to the project, which can be used to exchange data with external collaborators in a dedicated download/upload area of the project. Additionally, it is possible to use ACLs on filesystem level to customize access permissions for members of a storage project.

The use of SDS@hd is possible in general for all bwIDM member organizations and involves a registration procedure on user level. bwIDM [7] is the federated identity management of Baden-Württemberg's universities which is realized as sub-federation of the DFN-AAI [8]. This technology allows researchers to use the ID of their home institution when using SDS@hd. The access is controlled by the entitlements "sds-hd-user" and "sds-hd-sv" which are granted by each organization for their own staff and students. A concept for the creation and management of guest accounts for research partners outside of bwIDM is in development.

3 Usage statistics

In the last 5 years the number of storage projects and users of SDS@hd was continuously growing. In 2017, the service started after a migration from a previous storage service with little over 50 projects and 150 users. Meanwhile, there are 850 registered users which are organized in 235 storage projects (see Figs. 2,3).

Looking at these numbers, it has to be noted that the number of persons who profit from (and de facto use) SDS@hd is higher. This is due to the fact, that several groups and projects implement a "proxy concept" where some (potentially high) number of end-users access the data via a proxy service, e.g. an analysis platform. The connection between the proxy service and SDS@hd is realized on behalf of only one registered user and is therefore only counted as one user in the statistics. The actual number of such end-users is not known but can be substantial. Such proxy services are implemented e.g. by core-facilities providing storage space to their customers for microscopy or sequencing tasks, or web-services like Omero (`https://www.openmicroscopy.org/omero/`).

As a central storage service for research projects, SDS@hd is used by researchers from different scientific fields. The user communities of SDS@hd and their shares in storage projects are shown in Fig. 4. Life sciences, medical sciences, and digital humanities have the largest shares. Together they have a share of 81%. The remaining share covers scientific computing, astrophysics and further disciplines.

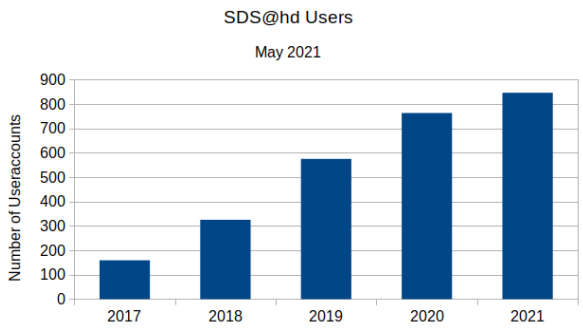


Figure 2: Number of SDS@hd users.

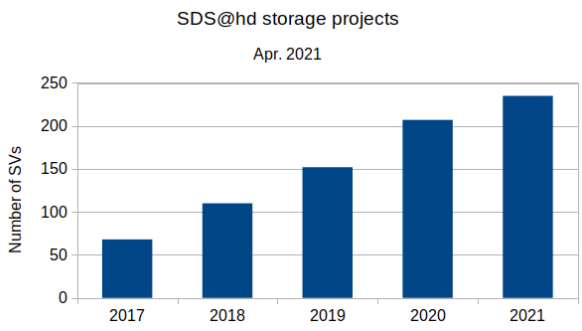


Figure 3: Number of SDS@hd storage projects.

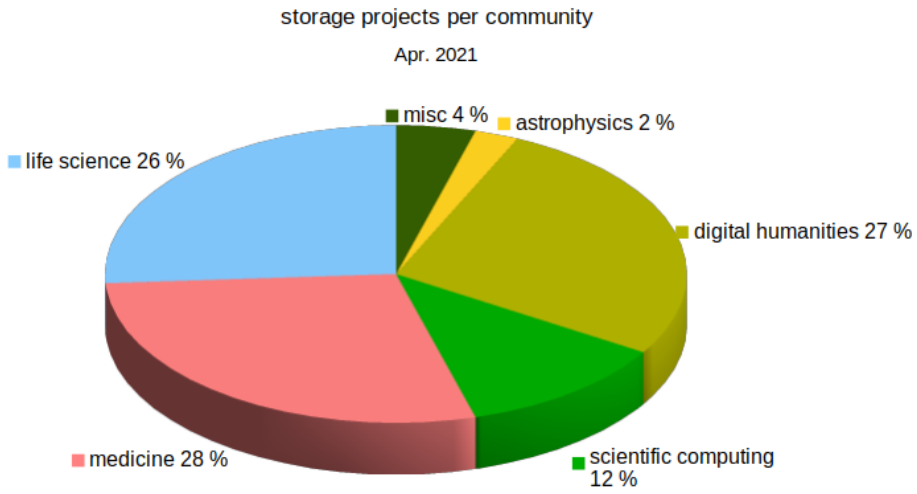


Figure 4: SDS@hd storage projects by communities.

4 Use-Cases

The scientific communities use SDS@hd for different workflows. Here we sketch a few exemplary use-cases for data-intensive research processes.

In life sciences and medical projects different types of microscopes and sequencers create data with high throughput. For example in the field of cryo electron microscopy new generation instruments are capable of producing several TB of data per day. The data are first cached locally and then transferred to SDS@hd. The original images must be available for several months for 2D and 3D structure analysis [9]. Medical projects with similar workflows deal with 3D images from expansion microscopy to identify structures in organs [10] or with high-resolution images of neuronal tissue samples for the analysis of neuronal structures and functions [11]. Other workflows in medical sciences use SDS@hd as download area and workspace for large scale genetic data from external sources.

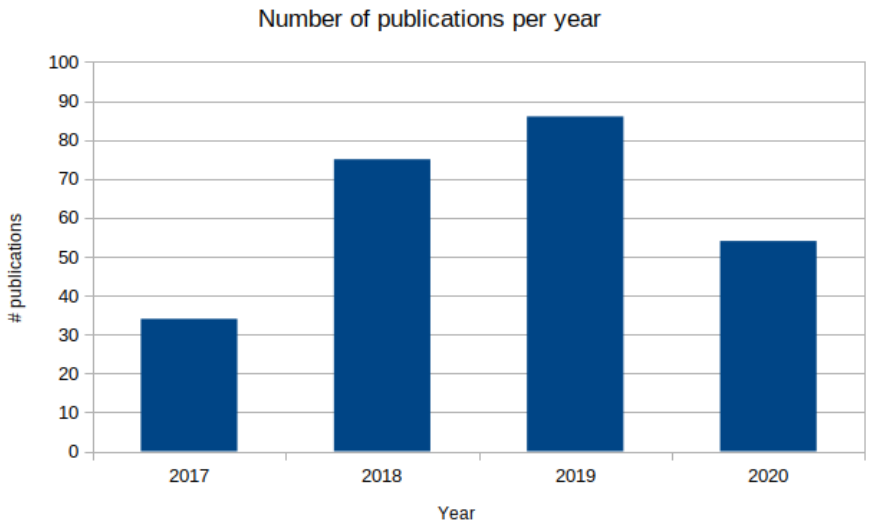


Figure 5: Number of reported publications 2017 - 2020.

In the humanities the digitalisation in historical sciences, archaeology, and linguistics creates big data and new challenges for data management. In this contect, a central storage solution as SDS@hd is essential for collaborative projects. For example, during an archaeological excavation digital and analog data are collected. Analog data like handwritten notes require digitalisation. All data are finally transferred to a storage project in SDS@hd for joint analysis.

In the field of scientific computing large scale data are produced in numerical simulations. Often it is required that the spacial distribution of physical quantities must be saved at many points in time to analyse the dynamic evolution of processes, for example in astrophysics [12].

For many use-cases the direct access to SDS@hd from the bwForCluster MLS&WISO becomes increasingly important for the production and post-processing of large simulation results as well as the analysis of images or genetic data. This development is reflected in a growing usage of SDS@hd in compute jobs.

5 References related to SDS@hd

Overall, a total of 249 publications related to SDS@hd were reported by users. Figure 5 shows the number of publications over the time. The rapid increase of publications shows a quick adoption of SDS@hd by its users and indicates the great value of this service for research projects. The amount of publications reported for 2020 is comparatively low, however it is expected to increase. The reason for this is that most publications are typically reported during the application of the storage project's extension which needs to be done once per year.

6 Future Work

In the future, the interplay of SDS@hd and services for publication repositories and long-term archives will be further developed and is planned to be the topic of a subsequent publication. In the context of the project bwHPC-S5, a data federation between the heteorogeneous and distributed systems and services within Baden-Württemberg is in development [3]. SDS@hd will be under further development and is planned to be integrated into this emerging federation which promises simplified technical and organizational transitions from one IT service to another and by this increases its value for research.

Acknowledgements

The authors gratefully acknowledge the large scale data facility (LSDF2) supported 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 and INST 35/1503-1 FUGG. Federated user support for SDS@hd is funded via bwHPC-S5 by MWK.

Bibliography

[1] Baumann, M., V. Heuveline, O. Mattes, S. Richling, S. Siebler. "SDS@hd – Scientific Data Storage." Proceedings of the bwHPC Symposium 2017, Universität Tübingen (2018).

[2] Baumann, M., O. Mattes, S. Richling, S. Siebler, A. Balz. "SDS@hd – Scientific Data Storage." Science-Tage 2019 – Data to Knowledge, Heidelberg: heiBOOKS (2020). https://doi.org/10.11588/heibooks.598.c8444

[3] Hartenstein, H., T. Walter, and P. Castellaz. Schneider, G., V. Heuveline, K.H. Horstmann, B. Neumair, T. Hätscher, A. Pfister, J. Beutner, M. Resch, T. Walter, S. Wesner, R. Dorn, M. Holst, M. Steuert, and J. Last. "Rahmenkonzept der Hochschulen des Landes Baden-Württemberg für datenintensive Dienste – bwDATA Phase III (2020-2024)." Publikationssystem Universitätsbibliothek Tübingen. http://dx.doi.org/10.15496/publikation-55923

[4] Richling, S., M. Baumann, S. Friedel, and H. Kredel. "bwForCluster MLS&WISO." Proceedings of the 3rd bwHPC-Symposium: Heidelberg 2016, pp. 103-107 (2017). https://doi.org/10.11588/heibooks.308.418

[5] Baumann, M., F. Heß, L. Maylein, T. Mechler, B. Scherbaum, and E. Volkmann. "heiARCHIVE, a long-term preservation service at Heidelberg University." E-Science-Tage 2021 – Share Your Research Data, Heidelberg: heiBOOKS (2021).

[6] Baumann, M., F. Bösert, S. Siebler, P. Skopnik, and J. E. Sundermann. "Entwurf einer Infrastruktur für den Datenaustausch großer Forschungsdatenmengen mittels WebDAV, FTS3 und OIDC." E-Science-Tage 2021 – Share Your Research Data, Heidelberg: heiBOOKS (2021).

[7] Föderiertes Identitätsmanagement der baden-württembergischen Hochschulen. http://bwidm.de/.

[8] DFN-AAI - Authentication and authorization infrastructure. https://www.aai.dfn.de/en/.

[9] Zupa, E., A. Zheng, A. Neuner, M. Würtz, P. Liu, A. Böhler, E. Schiebel, and S. Pfeffer. "The cryo-EM structure of a γ-TuSC elucidates architecture and regulation of minimal microtubule nucleation systems." Nature Communications 11, 5705 (2020). https://doi.org/10.1038/s41467-020-19456-8

[10] Cinzia, B., A. u. M. Khan, T. Picascia, Q. Sun, V. Heuveline, and N. Gretz. "New technical approaches for 3D morphological imaging and quantification of measurements." The Anatomical Record 3, 10 (2020). https://doi.org/10.1002/ar.24463

[11] Klevanski, M., F. Herrmannsdoerfer, S. Sass, V. Venkataramani, M. Heilemann, and T. Kuner. "Automated highly multiplexed super-resolution imaging of protein nanoarchitecture in cells and tissues." Nature Communications 11, 1552 (2020). https://doi.org/10.1038/s41467-020-15362-1

[12] Pellegrini, E. W., S. Reissl, D. Rahner, R. S. Klessen, S. C. O. Glove, R. Pakmor, R. Herrera-Camus, and R. J. J. Grand. "Warpfield population synthesis: the physics of (extra-) Galactic star formation and feedback-driven cloud structure and emission from sub-to-kpc scales." Monthly Notices of the Royal Astronomical Society 498, 3 (2020). https://doi.org/10.1093/mnras/staa2555

Veranstalter

Die **E-Science-Tage** werden vom Projekt bw2FDM unter Beteiligung des Karlsruher Instituts für Technologie, der Universität Konstanz und der Universität Heidelberg veranstaltet und vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg gefördert.

Druck und Bindung
Books on Demand GmbH
In de Tarpen 42, 22848 Norderstedt

Durch die offene Bereitstellung von Forschungsdaten profitieren Wissenschaftler:innen ungemein: vorhandene Datensätze können wiederverwendet und die Effizienz des wissenschaftlichen Fortschritts gesteigert werden. Der Datenaustausch schafft mehr Transparenz, macht die Zusammenarbeit effizienter und kann als Qualitätssicherungsmaßnahme betrachtet werden. Im Rahmen der E-Science-Tage 2021 wurden die Bedeutung, die Risiken und die Vorteile der gemeinsamen Nutzung von Forschungsdaten erörtert. Der vorliegende Tagungsband ist eine Sammlung von Vorträgen und Postern zum Thema „Share Your Research Data“.

ISBN 978-3-948083-55-7