
Herausforderungen beim Aufbau eines föderierten Datenrepositoriums auf Basis von InvenioRDM

Dirk von Suchodoletz ¹, Jonathan Bauer ¹, Marcel Tschöpe¹, Holger Gauza ², Michael Derntl³, Steve Kaminski³

¹Universität Freiburg, Rechenzentrum;

²Universität Tübingen, Zentrum für Datenverarbeitung;

³Universität Tübingen, Digital Humanities Center

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

1 Einleitung

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

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: <https://doi.org/10.11588/heibooks.1288.c18069> (CC BY-SA 4.0)

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

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

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

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

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

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

2 Organisatorische Grundlagen

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

Nachhaltigkeit Die Frage nach einem nachhaltigen Umgang mit Forschungsdaten aus Betreibersicht hat mindestens zwei Dimensionen: Verfügbarkeit und Kosten. Die Zusage und Sicherstellung einer langfristigen Verfügbarkeit von Forschungsdaten resultiert zwangsläufig in Aufwendungen, die gegebenenfalls über die Projektlaufzeit hinaus gehen und trotzdem geplant werden müssen (Leendertse und Suchodoletz 2020). Relevante Kostenfaktoren sind dabei das Mengengerüst, der jeweils notwendige Aufwand und die Betreuung einer Datenpublikationsplattform. Da viele Forschungs-Communities diese Fragen im Rahmen der NFDI angehen, sind hier weitere Erkenntnisse zu erwarten. Gleichzeitig muss die Verfügbarkeit von Forschungsdaten bei technischem oder wirtschaftlichem Ausfall des Repositoriums mitgedacht werden. InvenioRDM unterstützt Betreiber hinsichtlich eines möglichst wirtschaftlichen Umgangs mit Ressourcen durch zwei zentrale Aspekte: Die Anbindung an das Speichersystem bwSFS und somit eine (geo-)redundante Speicherung von Daten mittels S3 sowie den Aufbau von Communities¹ innerhalb von InvenioRDM,

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

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

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

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

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

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

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

3 Technische Grundlagen

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

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

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

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

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

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

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

Forschung findet weltweit vernetzt in unterschiedlichsten Kooperations- und Interaktionsbeziehungen statt. Ergebnisse, die an anderen Einrichtungen produziert werden, bilden die Fragestellung für die eigenen Forschenden und umgekehrt. Diesem verteilten Charakter muss ein Datenpublikationssystem Rechnung tragen. Die eScience-Strategie des Landes Baden-Württemberg fordert deshalb unter anderem die Such- und Auffindbarkeit von Forschungsdaten über mehrere Repositorien hinweg, welche die Grenzen der einzelnen Einrichtung und Fachdisziplinen überwindet⁵. Für solche übergreifenden Initiativen zu Suchportalen existieren bereits verschiedene Ansätze, wie beispielsweise re3data.org⁶. Hierfür hat sich der gemeinsame Standard OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) mit einem technisch sehr einfachen Verfahren zwischen einem oder mehreren Data- und Service-Providern etabliert. Über die OAI-PMH- und REST-Schnittstellen von InvenioRDM können andere Forschungsinformationssysteme die publizierten Datensätze systematisch und automatisch sammeln und referenzieren. Durch Abfragen (*harvesting*) werden die Daten zusammengetragen und zu einem konsolidierten Suchindex aggregiert. Die Metadaten sind in ihrer Struktur nicht durch OAI-PMH spezifiziert, so dass beispielsweise verschiedene disziplinspezifische Datenformate angeboten werden können. Für ein Mindestmaß an Interoperabilität sollte jeder Daten-Provider sinnvollerweise Publikationsmetadaten z.B. nach Dublin Core (DC) unterstützen. InvenioRDM implementiert daher die Bereitstellung solcher Metadaten über OAI-PMH.

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

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

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

4 Aufbau und Betrieb institutioneller Repositorien

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

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

Einsatz an der Universität Tübingen Das Digital Humanities Center der Universität Tübingen⁷ betreibt das generische institutionelle Forschungsdatenrepository FDAT,⁸ das den Anforderungen von Drittmittelgebern entspricht. Der Anspruch liegt darin, als institutionelles Repository ein disziplinübergreifendes Angebot für Forschungsdatenmanagement samt Beratungsschwerpunkt auf den Geistes- und Sozialwissenschaften zu schaffen. Das Ziel ist es, einen nachhaltigen Umgang mit Forschungsdaten zu fördern (Abbildung 1 und 2). Eine Säule der Nachhaltigkeit liegt in der Anbindung an eine namhafte Institution und der Einsatz von bwSFS als technische Speicherschicht. Der nachhaltige und

⁷ <https://dh-center.uni-tuebingen.de>

⁸ <https://fdat.uni-tuebingen.de>

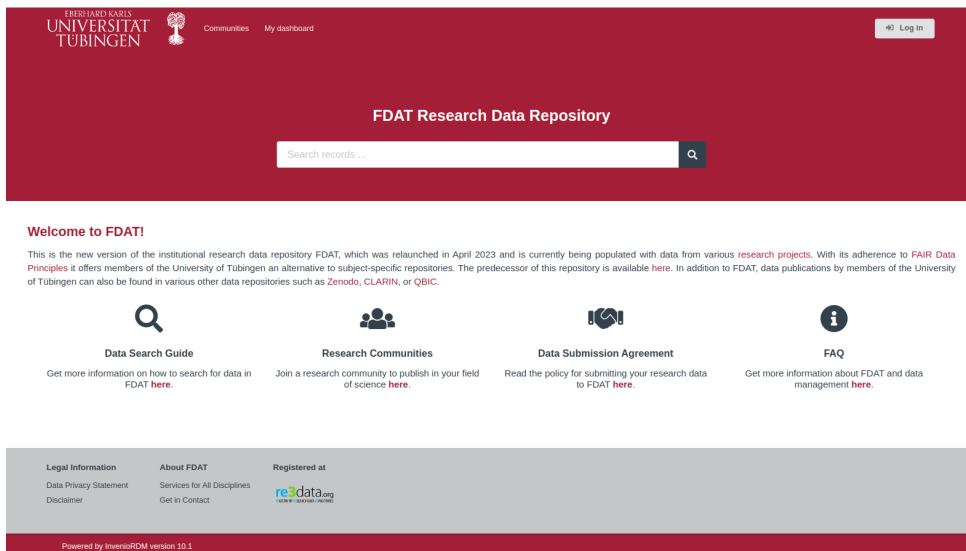


Abbildung 1: Startseite des Forschungsdatenrepositoriums FDAT.

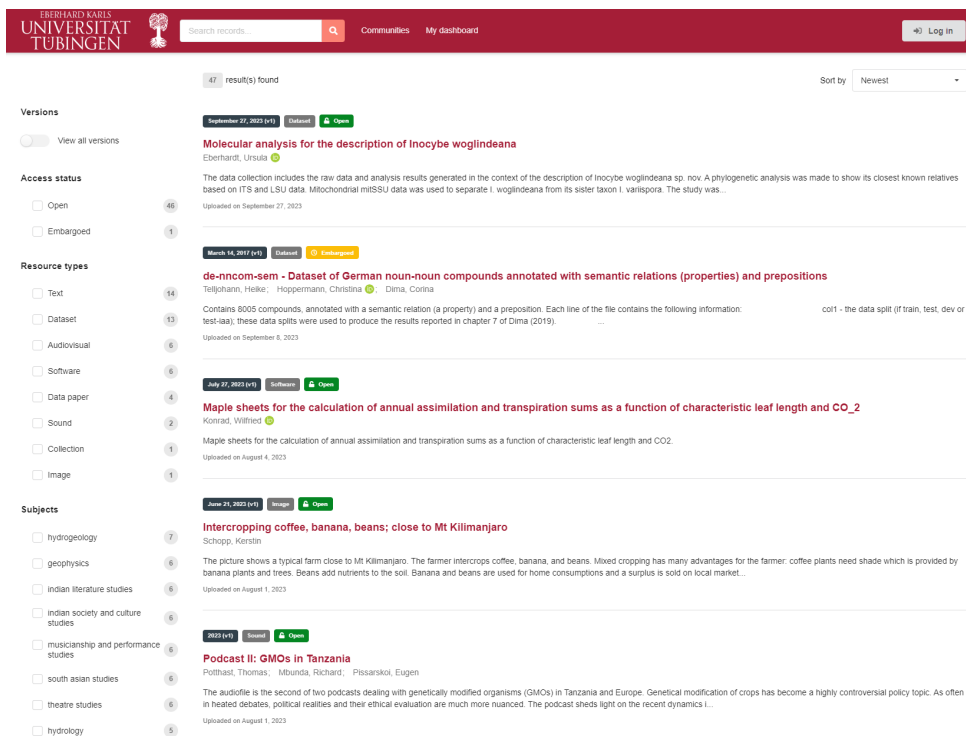


Abbildung 2: Listenansicht publizierter Forschungsdaten.

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

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

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

4.1 Workflow-Integration

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

⁹ Für weitere Informationen zum Zertifizierungsprozess siehe <https://www.coretrustseal.org>.

¹⁰ <https://dh-center.uni-tuebingen.de/fdat-policy/agreement.html>

¹¹ <https://rdmg.uni-freiburg.de>

¹² Für darüberhinausgehende weitere Informationen vergleiche <https://www.nfdi4plants.de/content/learn-more/annotated-research-context.html> bzw. Suchodoletz u. a. (2020).

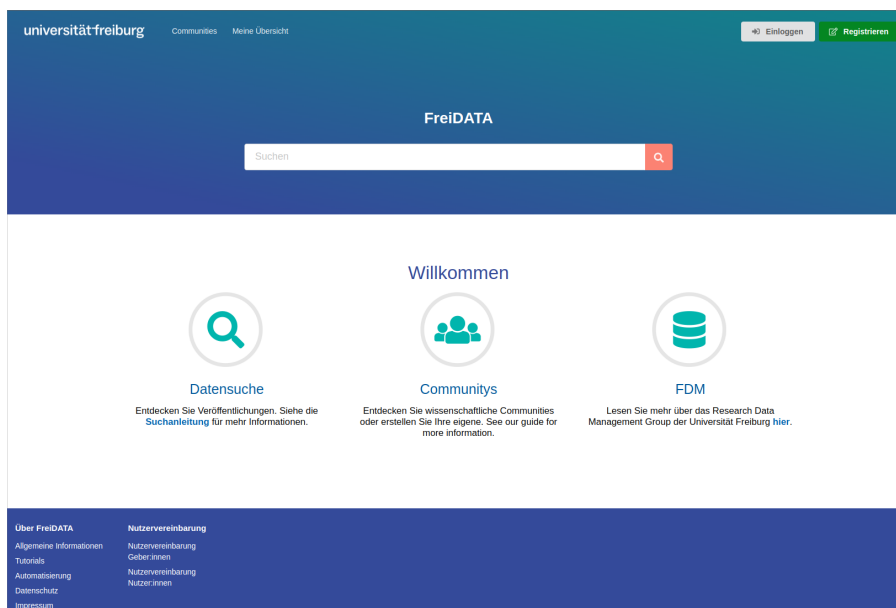


Abbildung 3: FreiData als institutionelles Repository der Universität.

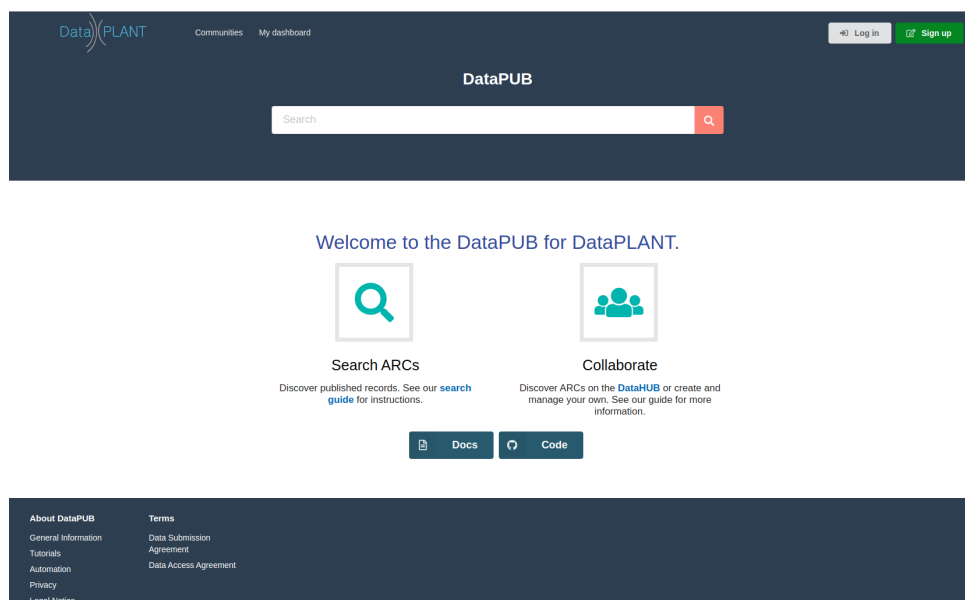


Abbildung 4: DataPUB als Community-Repository für DataPLANT.

wendung der GitLab-CI werden ARCs automatisch auf die Vollständigkeit der zu den Daten zugehörigen Metadaten geprüft und den Nutzenden wird ein Feedback in Form eines Badges auf der Webseite angezeigt. Die Nutzenden können validierte ARCs für die Veröffentlichung vorbereiten, was zur Erstellung eines *Drafts* in InvenioRDM führt. Die Drafts werden automatisch durch ein sogenanntes *Submission Request* einer Community zugeordnet und müssen von einem zuständigen Data Steward in der Rolle *Curator* akzeptiert werden. ARCs werden auf diese Weise nochmal einer manuellen Qualitätskontrolle unterzogen. Sobald die Daten für die Publikation freigegeben wurden, werden diese veröffentlicht und ein DOI registriert. Hierdurch wird der Stand der Daten fixiert und referenzierbar. Außerdem können über die OAI-PMH Schnittstelle die Metadaten der veröffentlichten Datensätze in Forschungsinformationssysteme übertragen werden. Der Einsatz der REST-API ermöglicht darüber hinaus die Anbindung an weitere Plattformen im Kontext von bwHPC und Galaxy¹³. Hierdurch werden Plattformen für die Datenproduktion in Form von wissenschaftlichen Workflows und Versionierungsplattformen mit einer Plattform für die Datenpublikation und langfristige Datenhaltung verknüpft.


5 Bisherige Erfahrungen und Ausblick

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

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

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

¹³ <https://galaxyproject.org>



☰

Communities

Organize, curate and collaborate on records for your institution, project, topic or event.

🔍
+ New community

My communities [See all](#)



BioDATEN
Life Cycle

SDC Bioinformatics
DATA ENV...


The center will support
bioinformatics workflows...

New communities [See all](#)




Troy Project

Research data from
excavations at the...



Faculty of Science

Research data originating from
the faculty of science at the...



Faculty of Humanities

Research data originating from
the faculty of humanities at t...

Abbildung 5: Darstellung und Auswahl von Communities in FDAT.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

BioDATEN Life Cycle SDC Bioinformatics DATA ENVIRONMENT **Change** ✕ Remove

Files

Metadata-only record ⓘ Storage available 0 out of 100 files 0 bytes out of 100.00 GiB

Drag and drop files - or - **Upload files**

⚠ File addition, removal or modification are not allowed after you have published your upload.

Basic information

||| Digital Object Identifier*

Do you already have a DOI for this upload? Yes No

Copy/paste your existing DOI here...

A DOI allows your upload to be easily and unambiguously cited. Example: 10.1234/foo.bar

📁 Resource type*

📄 Title*

Abbildung 6: Umsetzung eines Publikationsworkflows der BioDATEN-Community.




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

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

Danksagung

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

ORCID:

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

Literaturverzeichnis

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

- Axtmann, Alexandra, Felix Bach, Jonathan Bauer, André Blessing, Thomas Bönisch, Nina Buck, Holger Gauza u. a. 2021. „Kriterien für die Auswahl einer Softwarelösung für den Betrieb eines Repositoriums für Forschungsdaten“. *Bausteine Forschungsdatenmanagement*, Nr. 3: 14–26. DOI: <https://doi.org/10.17192/bfdm.2021.3.8348>. <https://bausteine-fdm.de/article/view/8348>.
- Bauer, Jonathan, Marcel Tschöpe, Julian Weidhase, Timo Mühlhaus, Christoph Garth, Gajendra Doniparthi, Holger Gauza, Louisa Perelo, Cristina Martins Rodrigues und Dirk von Suchodoletz. 2023. „From DataPLANT’s DataHUB to DataPUB(lication)“. In *International Workshop on Science Gateways*. Accepted for publication.
- Brettschneider, Peter, Alexandra Axtmann, Elisabeth Böker und Dirk von Suchodoletz. 2021. „Offene Lizenzen für Forschungsdaten: Rechtliche Bewertung und Praxistauglichkeit verbreiteter Lizenzmodelle“. *O-Bib. Das Offene Bibliotheksjournal* 8 (3): 1–22. DOI: <https://doi.org/10.5282/o-bib/5749>. <https://www.o-bib.de/bib/article/view/5749>.
- Deutsche Forschungsgemeinschaft e.V. 2013. *Sicherung guter wissenschaftlicher Praxis*. Wiley Online Library. ISBN: 978-3-527-33703-3. DOI: <https://doi.org/10.1002/9783527679188>. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527679188>.
- Leendertse, Jan, Susanne Mocken und Dirk von Suchodoletz. 2019. „Datenmanagementpläne zur Strukturierung von Forschungsvorhaben“. *Bausteine Forschungsdatenmanagement*, Nr. 2: 4–9. DOI: <https://doi.org/10.17192/bfdm.2019.2.8003>.
- Leendertse, Jan, und Dirk von Suchodoletz. 2020. „Kosten und Aufwände von Forschungsdatenmanagement“. *Bausteine Forschungsdatenmanagement*, Nr. 1 (1): 1–7. DOI: <https://doi.org/10.17192/bfdm.2020.1.8246>. <https://bausteine-fdm.de/article/view/8246>.
- Martins Rodrigues, Cristina, Dirk von Suchodoletz, Timo Mühlhaus, Jens Krüger und Björn Usadel. 2021. „DataPLANT – Ein NFDI-Konsortium der Pflanzen-Grundlagenforschung“. *Bausteine Forschungsdatenmanagement*, Nr. 2 (2): 46–56. DOI: <https://doi.org/10.17192/bfdm.2021.2.8335>. <https://bausteine-fdm.de/article/view/8335>.
- Suchodoletz, Dirk von, Elisabeth Böker, Peter Brettschneider und Franziska Rapp. 2020. „Entwicklung in Baden-Württemberg: ORCID und ROR IDs als Standard für langfristige Personen- und Institutionen-Identifizierung“. *Bausteine Forschungsdatenmanagement*, Nr. 2: 80–88. DOI: <https://doi.org/10.17192/bfdm.2020.2.8272>. <https://doi.org/10.17192/bfdm.2020.2.8272>.
- Suchodoletz, Dirk von, Peter Brettschneider, Elisabeth Böker, Jochen Apel, Dorothea Iglezakis, Karsten Schmidt und Gabriel Schneider. 2021. *Leitfaden – Verantwortungsvoller Umgang mit Forschungsdaten*. DOI: <https://doi.org/10.5281/zenodo.4907422>. <https://doi.org/10.5281/zenodo.4907422>.