Konzeptpapier

Forschungsdatenrepositorium für Thüringen (REFODAT)

Version: 1.5 (05.10.2023)

Autor*innen:
Kevin Lang (Bauhaus-Universität Weimar, TKFDM)
Roman Gerlach (Friedrich-Schiller-Universität Jena, TKFDM)
Nadine Neute (Universität Erfurt, TKFDM)
Silvio Hermann (Thüringer Universitäts- und Landesbibliothek Jena)
Jessica Rex (TU Ilmenau, TKFDM)
Olaf Schneider (URZ Jena, HS-ITZ)
Gerhard Vogt (TU Ilmenau, ThHoBi/BSC)

Inhalt

Allgemeine Definition eines Forschungsdatenrepositorium 3
Ist-Zustand: Was fehlt den Forschenden derzeit in Thüringen? 3
Soll-Zustand: Was sind die Anforderungen an ein neues System? 4
Wie würde ein für Thüringen entwickeltes System aussehen? 5
Ausblick – Schnittstellen zu weiteren Systemen 6
Wer ist an der Umsetzung beteiligt? 7
Benutzergruppenkonzept 7
- Gast 7
- Datengeber*in 8
- Eingeladene Datengeber*in 8
- Redakteur*in 8
- Administrator*in 8

Workflow-Modell 9
- Registrierung 9
- Workspace 9
- Eintrag 10
- Versionierung 12
- Archiv 12

Anhang 1: Begriffe 13
Anhang 2: Rechte der Benutzergruppen 14

Allgemeine Definition eines Forschungsdatenrepositorium

Ein Forschungsdatenrepositorium ist ein Ort zur digitalen Aufbewahrung und Verfügbarmachung von Forschungsdaten, die im Forschungsprozess entstanden sind. Der Begriff Forschungsdaten ist dabei weit gefasst und kann neben Daten u.a. Softwarecode, Methoden, Verfahren, Protokolle, und Umfragen beinhalten, die für das Verständnis, die Nachvollziehbarkeit und Reproduzierbarkeit der Forschungsergebnisse wichtig sind. Ein Repositorium wird typischerweise als Veröffentlichungsplattform für Forschungsdaten genutzt. Bei der Datenübergabe findet meist eine einfache Qualitätsprüfung statt. Eine umfangreiche Kuratierung erfolgt häufig nur bei fachspezifischen Repositorien. Die Sichtbarkeit von Datensätzen kann durch die Datenlieferanten (Autoren) bestimmt werden. In der Regel sind die Metadaten öffentlich zugänglich, so dass im Repositorium abgelegte Datensätze zumindest gefunden werden können. Der Zugang zu den eigentlichen Daten ist gegebenenfalls durch entsprechende Kontrollmechanismen beschränkt. Jedem Datensatz im Repositorium wird idealerweise ein persistenter Identifikator zugewiesen (z.B. DOI, URN). Ein Datensatz ist damit eindeutig identifizier- und zitierbar. Eine spätere Überführung ausgewählter Datenbestände in ein Langzeitarchivierungssystem ist möglich, benötigt aber entsprechende Schnittstellen bzw. zusätzliche Funktionalitäten.

Ist-Zustand: Was fehlt den Forschenden derzeit in Thüringen?

Bei der Wahl eines Datenrepositoriums wird Forschenden im Allgemeinen empfohlen, fachspezifische Repositorien zu wählen, da diese im Hinblick auf den Funktionsumfang und die Benutzerfreundlichkeit besser auf die Bedürfnisse einer Fachgemeinschaft abgestimmt sind. Falls kein fachspezifisches Repositorium existiert oder andere Gründe dagegensprechen, sollte auf generische / institutionelle Repositorien zurückgegriffen werden. Neben internationalen Angeboten wie Zenodo.org oder DataDryad.org bestehen an vielen deutschen Hochschulen fachübergreifende institutionelle Repositorien, die der Anforderung nachkommen, Daten am Ort ihrer Entstehung zu sichern und verfügbar zu machen.

Für die Veröffentlichung von Forschungsdaten in Thüringen kann aktuell die Digitale Bibliothek Thüringen (DBT) genutzt werden. Diese wurde ursprünglich als Publikationsserver für elektronische Dokumente, Abschlussarbeiten, Semesterapparate und Vorlesungsmitschnitte entwickelt und wird von der ThULB bereitgestellt. Prinzipiell ist dies ein thüringenweit nutzbarer Dienst, es sind jedoch nicht alle Hochschulen beteiligt. Um kurzfristig einen Dienst für die Publikation von Forschungsdaten anbieten zu können, wurde 2016/17 der pragmatische Ansatz gewählt, die DBT mit geringem Aufwand zu erweitern (u.a. Einführung des Objekttyps "Forschungsdaten" mit spezifischer Eingabemaske). Wie sich inzwischen gezeigt hat, funktioniert dieser Ansatz jedoch nur für einen Teil der Anwendungsfälle. Das System wurde für die Veröffentlichung der oben genannten Objekttypen konzipiert und bietet trotz der Erweiterung nicht den Funktionsumfang und die Handhabung eines Forschungsdatenrepositorium. Insbesondere bei den Upload- und Downloadmöglichkeiten und Kapazitäten entspricht es nicht den Anforderungen von Forschungsgruppen mit großen Datenvolumina. Das Metadatenschema ist nicht für Forschungsdaten konzipiert, und nur bedingt kompatibel mit anderen Forschungsdatenrepositorien, die dem DataCite-Schema folgen. Die Editierung von Forschungsdaten und Metadaten erfolgt auf Anfrage durch den DBT Support, da Forschende als Redakteure in der DBT nicht vorgesehen sind. Die DBT sichert aktuell eine Aufbewahrung von 5 Jahren zu. Der DFG Kodex von 2019 empfiehlt jedoch eine Archivierung von deutlich über 10 Jahren für die Dokumentation von Forschungsvorhaben und die aus den Vorhaben

erwachsenen Daten. Schnittstellen, die eine nahtlose Übergabe von Forschungsdaten aus Arbeitsplattformen oder in eine LZA-Lösung bieten sind nicht im erforderlichen Umfang vorhanden (Langzeitspeicherung/Archivierung nach Forderungen von Drittmittelgebern). Zudem unterstützt die DBT die FAIR-Prinzipien (Findable, Accessible, Interoperable und Reusable) nicht in vollem Umfang.

Da die Funktionalität der DBT und ihr Rechte und Rollen Schema ihren ursprünglichen und weiter bestehenden Funktionsanforderungen besser entspricht als einem Forschungsdatenrepositorium ist es zweckmäßig die beiden Funktionalitäten zur trennen, die DBT in der bestehenden Form zu erhalten und ein eigenständiges Forschungsdatenrepositorium aufzubauen.

Soll-Zustand: Was sind die Anforderungen an ein neues System?

Das Forschungsdatenrepositorium dient primär der Veröffentlichung von Daten, die von Forschenden der Thüringer Hochschulen direkt übermittelt werden und für die kein geeignetes Fachrepositorium zur Verfügung steht. Darüber hinaus soll das Repositorium Schnittstellen bereitstellen, die es ermöglichen, gesammelte und ggf. kuratierte Bestände aus anderen Systemen (z.B. Arbeitsplattformen), die eine dauerhafte Verfügbarkeit nicht gewährleisten können, zu übernehmen. Das System soll neben den typischen Funktionen eines öffentlichen Repositoriums wie z.B. einer Suche, Ingest und Download Funktionen, einer Zugangs- und Zugriffsverwaltung, auch die Vergabe von persistenten Identifikatoren (kurz PID, meistens DOIs in der Praxis), die Verknüpfung von Datenautoren und Datensätzen mit zusätzlichen IDs und Ressourcen (u.a. ORCIDs, ROR, GND) sowie Export- und Zitierfunktionen (wie JSON-LD oder BibTeX) bereitstellen. Datensätze von Forschenden der Thüringer Hochschulen, die bereits in anderen Fachrepositorien veröffentlicht wurden, sollen in dem Repositorium registriert werden. Ein Abgleich mit der Hochschulbibliografie soll gewährleistet sein.

Die Erfüllung der FAIR-Prinzipien (Findable, Accessible, Interoperable und Reusable) soll durch das System unterstützt werden. Weiterhin sollten Metadaten (wie von DataCite) im- und exportierbar sein. Sie sind dauerhaft verfügbar und bleiben selbst nach der Löschung der eigentlichen Forschungsdaten erhalten. Bereits angelegte Metadaten sollen so weit wie möglich übernommen und weiterverwendet werden. Die Forschungsdaten im Repositorium werden dabei nach DFG Kodex 2019 mindestens 10 Jahre nach Herstellung des öffentlichen Zugangs vorgehalten. Autoren sollen in der Lage sein, ihre Datensätze erst nach einer Embargofrist frei zugänglich zu machen. Die Nutzungsrechte von Datensätzen bestimmen die Autoren mit der Wahl einer geeigneten Lizenz, die im System hinterlegt ist.

Das Forschungsdatenrepositorium soll sich nach Möglichkeit leicht in die bestehende Infrastruktur der betreibenden Anbieter integrieren lassen, so dass bestehende Kompetenzen genutzt und gestärkt werden können. Es werden keine Begrenzungen im Speicherplatzbedarf vorgegeben und größere Datenmengen können über externe Datenträger eingeliefert werden. Durch Guidelines und automatische Tests wie die Prüfung nach offenen Formaten oder das Erkennen von Duplikaten soll eine minimale Qualitätssicherung erreicht werden. Nach Ablauf von 10 Jahren soll, wenn möglich unter Beteiligung der Datengeber, anhand von noch festzulegenden Kriterien entschieden werden, ob die Daten in eine Langezeitarchivierung übergeben werden sollen.

Wie würde ein für Thüringen entwickeltes System aussehen?

Das in Thüringen zum Einsatz kommende System soll auf dem Open Source Repository-Framework *MyCoRe*[1] aufbauen. Ausgangspunkt für die Entwicklung kann bspw. ein *MODS Institutional Repository* (*MIR*[2]) sein oder *Collections*[3]. Beide basieren auf MyCoRe und bieten prinzipiell schon jetzt Möglichkeiten, Forschungsdaten zu erschließen und abzulegen. Generell ist die Verarbeitung beliebiger Datenmodelle, wie z.B. auch DataCite[4], möglich. Falls nötig kann man also ein bestehendes Datenmodell für Forschungsdaten erweitern. Das MyCoRe-Framework bietet sowohl einfache als auch komplexe Recherchemöglichkeiten, unterstützt die Abbildung von Nutzenden auf verschiedene Rollen und damit einhergehenden Rechten im Repository. Zudem ermöglicht es umfangreiche Dateioperationen auf die vom System verwaltete Daten. Daten sind mit einem Embargo versehbar.

Nutzende greifen über eine Weboberfläche als Frontend auf das Forschungsdatenrepository zu. Das generelle Erscheinungsbild der Anwendung lässt sich unkompliziert an die Anforderungen eines Thüringer Systems anpassen.

Die eigentliche Verwaltung der Meta- und Forschungsdaten findet im Backendsystem statt, welches unter anderem die bereits oben genannten Anforderungen unterstützt. Insbesondere als Repositorium zur Veröffentlichung und ggf. Weitergabe von Meta- und Forschungsdaten an andere Systeme, kann eine auf MyCoRe-basierende Lösung seine Stärken ausspielen.

Meta- und Forschungsdaten sind über Standardschnittstellen (z.B. OAI-PMH[5]) schon jetzt abrufbar. Über IIIF[6] können Bilddaten standardisiert abgerufen werden. Die in MyCoRe vorhandene REST-API[7] unterstützt das Einreichen und das Hochladen von Forschungsdaten aus Drittsystemen. Funktionen, welche insbesondere gängige Workflows im Bereich Forschungsdatenmanagement unterstützen sollen, müssen aber sowohl im Frontend als auch im Backend noch implementiert werden. Hier wären besonders die Möglichkeit der Datenkuratierung und die Unterstützung eines Reviewprozesses (z.B. Generierung sicherer Links für Außenstehende) mit einem externen Gutachter zu nennen.

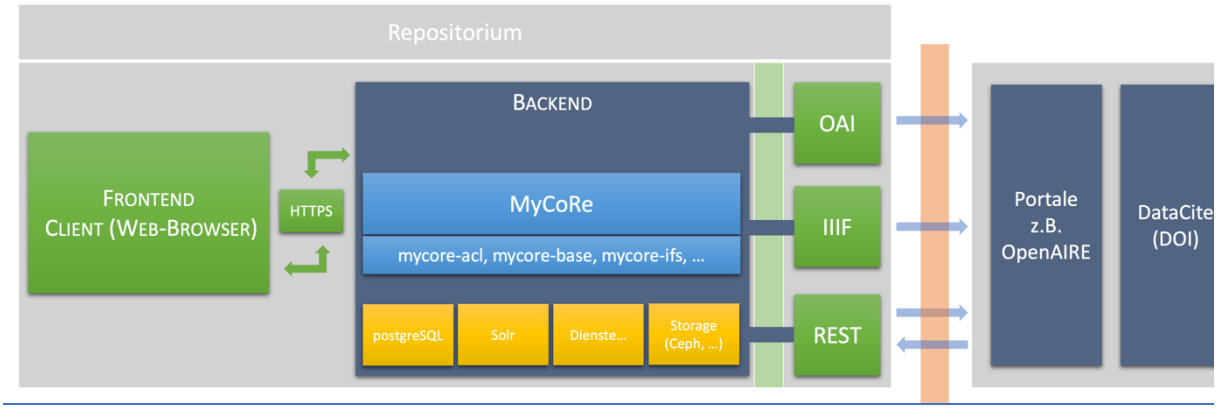


Abbildung 1: Forschungsdatenrepositorium (Architektur)

[1] https://www.mycore.de
[2] https://www.mycore.de/site/applications/mir/
[3] https://git.uni-jena.de/thulb/collections
[4] https://schema.datacite.org/meta/kernel-4.4/
[5] https://www.openarchives.org/pmh/
[6] https://iiif.io/
[7] https://www.mycore.de/documentation/interfaces/interface_rest_v2/

Ausblick – Schnittstellen zu weiteren Systemen

Das Forschungsdatenrepositorium wird mittelfristig Teil einer vielfältigen Landschaft weiterer Systeme sein und muss mit Hilfe von Schnittstellen und Protokollen Übergänge zu diesen Systemen bieten. Zu nennen sind hierbei Arbeitsplattformen (auch Datenmanagementsystem genannt), die u.a. in der aktiven Projektphase innerhalb eines Konsortiums zum Einsatz kommen sowie Systeme zur Langzeitarchivierung. Da sich diese Systeme von Repositorien unterscheiden, sollen sie zum besseren Verständnis hier skizziert werden. Die Implementierung solcher Systeme ist jedoch nicht Teil dieses Projektes.

Arbeitsplattformen (während der Laufzeit eines Forschungsprojektes)
Eine Arbeitsplattform gibt Forschenden die Möglichkeit, ihre Forschungsdaten (Daten, Code, Workflows, etc.) auf einer zentralen Plattform strukturiert innerhalb der Arbeitsgruppe oder eines Projektes abzulegen, zu verwalten und mit Partnern zu teilen. Die Forschungsdaten liegen mitunter in verschiedenen Versionen und Qualitätsstufen vor. Sie werden durch regelmäßige Backups gesichert und der Zugang wird durch Nutzerrechte geregelt. Dabei kann auch außenstehenden Projektpartnern anderer Organisationen der Zugriff gewährt werden. Die Daten werden nach einem standardisierten Metadatenschemata dokumentiert und nach Qualitätskriterien kuratiert. Nutzer haben dabei die Möglichkeit sowohl die Metadaten als auch die Daten zu durchsuchen. Die Arbeitsplattformen ermöglichen es Forschenden, Daten nach den FAIR-Prinzipien vorzuhalten und bietet standardisierte Schnittstellen für den maschinellen Austausch mit anderen Systemen. Sie bietet auch die Möglichkeit große Datenmengen zu verwalten, gegebenenfalls mit Hilfe verteilter Speichersysteme und einer zentralen Metadatenverwaltung. Die Arbeitsplattformen sind im Normalfall an das jeweilige Fachgebiet angepasst. Typische Anwendungen zum Verwalten von Forschungsdaten für Experimente in Laboren sind zum Beispiel elektronische Laborbücher.

Langzeitarchivierungssystem
Einige Forschungsdaten werden als so wertvoll erachtet, dass sie über die Zeit von 10 Jahren hinaus archiviert werden sollen. Da aus Kostengründen nicht alle Forschungsdaten archiviert werden können, ist es notwendig eine Bewertung vorzunehmen. Für die ausgewählten Datensätze sind zusätzliche Maßnahmen der Qualitätssicherung erforderlich, wie beispielsweise die Umwandlung der Daten in archivierbare Formate oder die Erzeugung standardisierter Archivpakete. Ein Langzeitarchivierungssystem bietet die entsprechenden Funktionen und nimmt die Einlagerung vor.

Für Forschungsdaten existiert bislang kein solches System in Thüringen. Allerdings bestehen Erfahrungen aus dem Projekt „Aufbau eines Systems der elektronischen Langzeitarchivierung / Ausbau des Digitalisierungszentrums Friedenstein Gotha“ (EFRE-Förderung) sowie einem weiteren Projekt zum "Aufbau einer digitalen Langzeitarchivierung für Publikationen und Daten aus dem Lehrbetrieb" (Bachelor-, Master- und Promotionsarbeiten), gefördert mit Mitteln des Hochschulpaktes 2020. Das Ergebnis der genannten Projekte ist die Etablierung eines Speicherverbunds für die Digitale Langzeitarchivierung von Kulturdaten in Thüringen. Im Verbund archivierte Daten sind bei mehreren, institutionell unabhängigen Serviceprovidern gespeichert. Durch automatisierte Software-Prozesse werden sowohl die „Bitstream Preservation“, also der Erhalt der digitalen Informationen als solcher, als auch das „Preservation Planning“ zur Erhaltung des Informationsgehaltes der Archivalien sichergestellt. Koordiniert wird der Verbund vom IT-Zentrum der Thüringer Hochschulen. Neben der

Koordination des Verbundes betreut das IT-Zentrum auch die Software auf den Archivknoten der Serviceprovider (durch entsprechende Auftragsvergabe an Dienstleister). Als Software kommt eine für das Digitale Archiv NRW (DA-NRW) entwickelte Software-Suite auf der Basis der Open-Source-Software iRods zum Einsatz.

Diese Lösung soll perspektivisch auch für Forschungsdaten zur Verfügung stehen. Dazu sind neben einer Erweiterung des Betriebskonzeptes auch Anpassungen am System nötig, die eine Überführung von Datenbeständen aus dem oben beschriebenen Repositorium in das digitale Langzeitarchiv ermöglichen. Aus diesem Grund ist eine enge Abstimmung und Koordinierung zwischen den Projekten erforderlich.

Wer ist an der Umsetzung beteiligt?

Der Strategierat des Thüringer Kompetenznetzwerkes hat in seiner Sitzung am 04.05.2022 den Aufbau eines Forschungsdatenrepositoriums beschlossen. Über die Umsetzung wird der Strategierat in den regelmäßigen Sitzungen informiert.

Das Thüringer Kompetenznetzwerk Forschungsdatenmanagement (TKFDM) erhielt vom Strategierat den Auftrag für die Ausarbeitung eines Umsetzungskonzeptes. Dieses wurde in der gemeinsamen AG Forschungsdaten des TKFDM, ThHoBi/BSC und des HS-ITZ[8] erstellt. Die AG hat zudem die fachlich-inhaltliche Leitung des Projekts inne und begleitet die konkrete Ausgestaltung des neu zu schaffenden Systems während der gesamten Projektlaufzeit.

Die technische Umsetzung erfolgt durch Mitarbeiter*innen der ThULB und des Universitätsrechenzentrums der Uni Jena (URZ/HS-ITZ) und ggf. durch eine Auftragsvergabe an externe Auftragnehmer. Die technische Infrastruktur wird durch das HS-ITZ bereitgestellt und betrieben.

Benutzergruppenkonzept

Das Benutzergruppenkonzept benennt die vorgesehenen Benutzergruppen und beschreibt diese. Eine Visualisierung dieses Rechte/Rollenkonzeptes mit einer abschließenden Auflistung der Features des Repositoriums erfolgt im Anhang. Hier werden ebenfalls die Optionen zur Vergabe von Bearbeitungsrechten an den Daten dargestellt. Die Rollen des Gasts und der Datengeber*in orientieren sich an Zenodo.

Gast

Als Gast werden anonyme Benutzer*innen der Plattform bezeichnet, die sich nicht angemeldet haben. Diese können alle Einträge zu den Forschungsdaten sehen und im Rahmen der durch die Datengeber*innen festgelegten Zugangsbedingungen und Lizenzen nutzen sowie über ein Kontaktformular Zugang zu den Daten oder weitere Nutzungsrechte anfragen. Für die Downloads wird nur die allgemeine Downloadanzahl der Datensätze erfasst. Das Speichern von z.B. Suchfiltern erfolgt über Cookies bei dem Gast selber.

[8] https://thhobi.de/bsc/ag-forschungsdaten.html

Datengeber*in

Als Datengeber*innen werden Benutzer*innen bezeichnet, welche sich registrieren und die Nutzungsbedingungen für das Anlegen von Workspaces für Projekte akzeptieren. Durch die Registrierung über die Logindaten einer Thüringer Hochschule werden die Benutzer*innen automatisch für die Rolle freigeschaltet. Erfolgt die Registrierung über das Repositorium selber oder durch eine ORCID-ID, muss die Rolle erst durch die Redakteur*innen freigeschaltet werden. Die Datengeber*innen können in einem selbst angelegten Workspace Forschungsdaten hochladen, über einen Datenträger einreichen, über einen externen Link laden oder durch einen externen Identifier (wie DOI) registrieren lassen. Weiterhin können und müssen die Datengeber*innen Metadaten zu den Forschungsdaten angeben, die für den Eintrag benötigt werden. Wurde der Datensatz als Eintrag im Repositorium von den Redakteur*innen freigegeben, steht es den Datengeber*innen offen, Metadaten des Eintrags zu ändern und eine neue Version des Datensatzes anzulegen, was zu einem neuen Workspace führt, welcher aber mit dem Eintrag des Vorgängerdatensatzes verbunden ist. Sie können weiterhin Freigaben für einen "Closed Access"-Datensatz erteilen, wenn dieser selber angelegt wurde.

Eingeladene Datengeber*in

Als "Eingeladene Datengeber*innen" werden Benutzer*innen beschrieben, die selber die Rolle "Datengeber*in" haben (im System also kein Unterschied besteht), aber zu einem fremden Workspace bzw. Eintrag eingeladen wurden, um an diesem kollaborativ mitzuarbeiten. Je nachdem wie die Einladung der Eigentümer*innen des Workspace bzw. Eintrags konfiguriert wurde, können die eingeladenen Datengeber*innen die Forschungsdaten und Metadaten im Workspace nur lesen oder auch umschreiben. Weitere Features, die von den einladenden Datengeber*innen explizit für die eingeladenen Datengeber*innen freischalten müssen, sind die Fähigkeit weitere Einladungen verschicken zu dürfen, einen Workspace zum Review durch die Redakteur*innen freizugeben oder von einem fertigen Eintrag eine neue Version des Datensatzes anzulegen.

Redakteur*in

Redakteur*innen sind Benutzer*innen, die sich um die Qualitätssicherung der Datensätze eines abgegrenzten Personenkreises kümmern (z.B. Hochschule). Werden Forschungsdaten und Metadaten in einem Workspace von Datengeber*innen angelegt und zur Prüfung freigegeben, wird die entsprechende Redakeur*innen für die Kontrolle informiert. Neben automatischen Tests müssen Redakteur*innen auch formal prüfen, ob Forschungs- und Metadaten den Leitlinien entsprechen. Ist alles in Ordnung, können die Redakteur*innen den Datensatz als Eintrag im Repositorium freigeben. Redakteur*innen sind weiterhin zuständig für Fragen von den Datengeber*innen bezüglich der Veröffentlichung der hochgeladenen Datensätze und können gegebenenfalls die Zugriffsrechte auf die Workspaces und Einträge von Datensätzen der eigenen Hochschule managen.

Administrator*in

Die Administrator*innen haben Zugriffrechte auf alle Daten des Systems. Sie kümmern sich um die technischen Einstellungen und Probleme des Repositoriums und sollten nur im Notfall Forschungsdaten und Metadaten verändern dürfen, wenn diese nicht die Betriebsregeln (z.B. durch DataCite, Veränderung von Forschungsdaten nach Registrierung der DOI) brechen. Notfälle können z.B. der Verlust der Zugangsdaten von Benutzer*innen sein.

Workflow-Modell

Das Workflow-Modell beschreibt den Umgang mit einem Datensatz im Thüringer Forschungsdatenrepositorium in verschiedenen Zuständen aus Sicht der Benutzergruppen. Die Zustände spiegeln dabei primär den Veröffentlichungsprozess von Forschungsdaten wieder. Als Vorbild dient der Workflow von Zenodo. Dieser besteht aus den Zuständen „Deposit“ (hier "Workspace") und „Record“ (hier "Eintrag") bezüglich eines Datensatzes. Um den Kontext des Workflows zu verdeutlichen, werden über den eigentlichen Workflow hinaus auch die „Registrierung“ der Benutzer*innen und die Übergabe der Forschungsdaten in das „Archiv“ beschrieben.

Registrierung

Personen, die das Repositorium ohne eine Registrierung nutzen, werden als "Gast" bezeichnet. Die Veröffentlichung von Daten im Repositorium setzt eine vorherige Registrierung und Anmeldung voraus. Erfolgt die Registrierung über den Login einer Hochschule in Thüringen, kann man als "Datengeber*in" ohne weitere Prüfungen sofort eigene Workspaces anlegen oder sich zum kollaborativen Einreichen und Verwalten von Datensätzen bei fremden Workspaces einladen lassen. Erfolgt die Registrierung hingegen über das Repositorium selber, muss die Person erst manuell frei geschaltet werden, indem ihr ein Institut von einer Redakteur*in zugeteilt wird. Mit der Registrierung stimmt die Person den Nutzungsbedingungen des Repositoriums (insbesondere dem Umgang mit dem Datenschutz) zu. Redakteur*innen und Administrator*innen können sich auf die gleichen Wege registrieren, deren Rollen müssen aber noch zugeordnet werden.

Workspace

Datengeber*innen haben eine Übersicht über alle selbst angelegten oder explizit mit ihnen geteilten Datensätzen in Form von „Workspaces“ und „Einträgen“. Workspaces (Datensätze vor der Veröffentlichung) erhalten bei der Erzeugung eine eindeutige interne ID. Eine Datengeber*in kann aber Rechte an andere Datengeber*innen vergeben, so dass diese einen Workspace sehen und bearbeiten können. Workspaces sind ausschließlich für die anlegenden Datengeber*innen, die Redakteur*innen und die Administrator*innen sowie durch die Erstellenden explizit eingeladene weitere Datengeber*innen einsehbar und veränderbar.

Ein Datensatz besteht aus zwei Komponenten, den Forschungsdaten und den sie beschreibenden Metadaten, wie Titel des Datensatzes, Namen der Autor*innen, Datum der Erhebung, Fachgebiet, Typ, Sprache oder Schlüsselwörter. Angegebene Metadaten können abgespeichert werden, müssen aber noch nicht vollständig sein. Es findet eine formale Prüfung der Eingaben statt, z.B. ob sie einem Standard entsprechen (wie Datumformat nach ISO-Norm). Die Einbindung kontrollierter Vokabulare (z.B. GND, gemeinsame Normdatei) sowie von Ontologien ist anzustreben. Mit der Erstellung des Workspaces wird ein Digtial Object Identifier (DOI) für den Workspace reserviert und eindeutig mit ihm verknüpft. Die DOI kann z.B. in Publikationen verwendet werden, auch wenn der Datensatz selbst noch nicht veröffentlicht ist.

Es gibt vier Möglichkeiten Dateien der Forschungsdaten einem Workspace hinzuzufügen:

(1) Upload der Dateien von eigenem Rechner
(2) Übergabe der Dateien auf einem externen Datenträger (z.B. Festplatte)
(3) Übertragung der Dateien eines externen Servers
(4) Registrierung eines bereits veröffentlichten Datensatzes

Im ersten Fall werden die Dateien durch einen Upload vom eigenen Computer über den Webbrowser zum Workspace hinzugefügt.

Im zweiten Fall übergibt die Datengeber*in in Absprache mit der zuständigen Redakteur*in die Dateien der Forschungsdaten auf einem zulässigen Datenträger. Die Redakteur*in kopiert die Dateien vor Ort und fügt sie dem entsprechenden Workspace hinzu. Die Benutzer*in wird informiert, wenn die Dateien im Workspace vorliegen.

Im dritten Fall lädt die Datengeber*in die Dateien nicht von ihren eigenen Rechner in den Workspace, sondern gibt eine URL als Bezugsquelle der Dateien an. Das System lädt die Dateien dann von dort aus. Dies geschieht asynchron im Hintergrund, so dass die Datengeber*in nicht aktiv im Repositorium am Browser eingeloggt bleiben muss. Sie bekommt eine Benachrichtigung, sobald das Repositorium alle Dateien komplett in den Workspace überführt hat.

Im vierten Fall wurde der Datensatz bereits auf einem anderen Repositorium veröffentlicht und soll lediglich im Thüringer Forschungsdatenrepositorium registriert werden. Hierfür gibt die Datengeber*in den eindeutigen Identifier (z.B. DOI) an und das System übernimmt die verfügbaren Metadaten aus dieser Quelle. Die Forschungsdaten verbleiben beim ursprünglichen Repositorium und werden anhand des eindeutigen Identifikators verlinkt. Metadaten, die im ursprünglichen Repositorium fehlen, aber im Thüringer Forschungsdatenrepositorium Pflichtangaben sind, sind zu ergänzen.

Die Datengeber*in hat jederzeit die Möglichkeit eine automatische Qualitätskontrolle durchführen zu lassen. Diese beinhalten z.B. eine Prüfung auf Dubletten, die Verwendung proprietärer Dateiformate oder das Vorhandensein einer Readme-Datei in den Forschungsdaten. Bei einem negativen Prüfergebnis wird die Datengeber*in informiert und gebeten, Änderungen vorzunehmen. Hat die Datengeber*in das Ergebnis der automatischen Qualitätskontrolle bestätigt und sind alle notwendigen Metadaten vorhanden, kann die Veröffentlichung des Datensatzes über eine Benachrichtigung an die zuständige Redakteur*in veranlasst werden.

Die Redakteur*in hat die Aufgabe den Datensatz zu betrachten und nach Qualitätskriterien zu bewerten, die nicht durch die automatische Qualitätskontrolle erkannt werden konnten. Dazu gehören z.B. die Vollständigkeit der Dokumentation in der Readme-Datei oder die Konsistenz der Angaben in den Metadaten. Sollten Mängel vorliegen, kann die Redakteur*in die Datengeber*in auffordern, diese zu beheben oder sie führt Änderungen selbst durch. Nach Abschluss der Prüfung, erfolgt die Freigabe durch die Redakteur*in. Der Workspace wird zu einem Eintrag umgewandelt.

Eintrag

Ein Eintrag ist ein auffindbarer Datensatz anhand von Metadaten. Während der Workspace-Phase wird entschieden auf welche Art der Datensatz mit seinen Forschungsdaten zugänglich gemacht werden soll. Dies kann einer der folgende Fälle sein: (1) „Open Access", (2) „Embargoed Access", (3) „Closed Access" und (4) „External Access" (siehe Abb. 2). In allen Fällen sind die Metadaten des Datensatzes öffentlich sichtbar.

Im ersten Fall von „**Open Access**" werden die Forschungsdaten frei zur Verfügung gestellt. Die Nutzungsbedingungen werden durch eine Lizenz geregelt (z.B. CC, GPL, ODC oder CDLA). Nach der Veröffentlichung können Gäste alle Dateien des Datensatzes herunterladen und nach den Nutzungsbedingungen der Lizenz verwenden.

Im zweiten Fall von „**Embargoed Access**" wird ebenfalls von der Datengeber*in eine Lizenz gewählt, aber zusätzlich ein zweites in der Zukunft liegendes Datum nach der Veröffentlichung, bei dem die Forschungsdaten des Datensatz zu „Open Access" überführt werden. Bis dahin sind nur die Metadaten des Datensatzes einsehbar, aber es gibt keinen Zugang für Gäste zu den Forschungsdaten. Sobald das

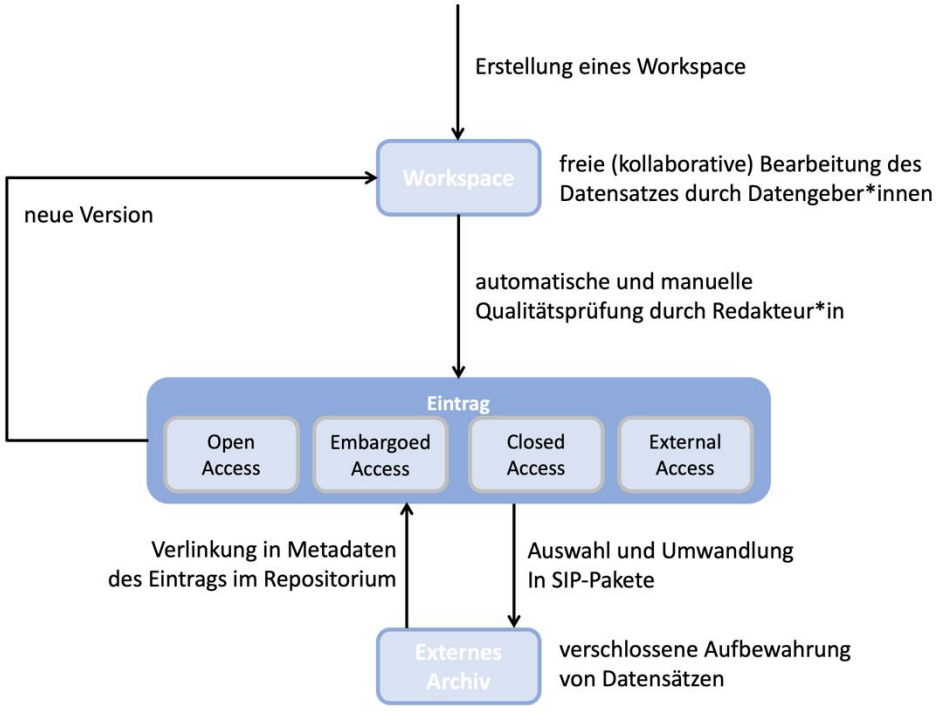
Datum erreicht ist und die Forschungsdaten öffentlich verfügbar sind, greift auch die gewählte freie Lizenz für die Nutzungsbedingungen der Forschungsdaten.

Im dritten Fall von „**Closed Access**" kann keine freie Lizenz gewählt werden. Die Forschungsdaten sind nur auf Anfrage verfügbar und alle Rechte bleiben vorbehalten. Gäste haben die Möglichkeit über ein Kontaktformular der Datengeber*in, welche den Datensatz erstellt hat, anzufragen, ob sie Zugang zu den Forschungsdaten bekommen können. Hierzu geben sie in einem Feld ihre E-Mail-Adresse an und in einem freien Textfeld kann ein Grund angegeben werden, warum sie gern Zugriff auf die Forschungsdaten hätten. Die Bedingungen für den Zugriff auf Forschungsdaten kann durch einen Vortext des Benutzers spezifiziert werden. Wird die Anfrage von einem Gast abgeschickt, erhält die Datengeber*in eine Benachrichtigung. Die Datengeber*in kann dann entscheiden, ob sie den Zugriff des Datensatzes an den Gast gewährt, indem sie von dem System einen Token generieren lässt, der dem Gast einen zeitlimitierten Zugriff auf die Forschungsdaten erlaubt. Andernfalls kann die Datengeber*in den Gast aber auch über die angegebene E-Mail-Adresse kontaktieren, um Bedingungen weiter auszuhandeln.

Der vierte und letzte Fall „**External Access**" ist ein Sonderfall, bei der die Forschungsdaten nicht im Repositorium selber liegen, sondern bereits bei einer anderen Plattform veröffentlicht wurden. Über einen Identifikator (bspw. eine DOI) werden die zugehörigen Metadaten des Datensatzes von der anderen Plattform importiert und gegeben falls von der Datengeber*in ergänzt. Das Repositorium registriert also nur die Metadaten, während der eigentliche Zugang zu den Forschungsdaten von der externen Plattform geregelt wird.

Im Sinne von Open Science und im Hinblick auf die Leitlinien der Fördermittelgeber wird der erste Fall bevorzugt und als Standardfall betrachtet. Die anderen Fälle sollten nur mit Begründung durch die Datengeber*in Anwendung finden. Damit wird der Leitsatz: „Zugänglich, wenn möglich, eingeschränkt, wenn nötig" umgesetzt.

Abbildung 2: Grafische Darstellung des Workflows



Versionierung

Sowohl die Datengeber*in als auch die eingeladenen Datengeber*innen des ehemaligen Workspace, können zu dem Eintrag eines Datensatzes eine neue Version erstellen. Wenn eine neue Version angelegt werden soll, so wird ein Workspace mit einer neuen ID erstellt und die Metadaten und Forschungsdaten aus der alten Version übernommen. Die Datengeber*innen können die Forschungsdaten und ggf. die Metadaten aktualisieren und werden zur Dokumentation der Veränderungen aufgefordert. Danach reichen sie die neue Version des Datensatz zur automatischen und manuellen Begutachtung und anschließenden Veröffentlichung ein. Nach der Freigabe durch die zuständige Redakteur*in ist der Datensatz als neuer Eintrag öffentlich sichtbar und verknüpft mit seiner Vorgängerversion.

Archiv

Datensätze können neben ihrem Eintrag im Repositorium auch an eine Archivierungslösung übergeben werden. Diese erfüllt die Qualitätsmerkmale einer Langzeitarchivierung. Das Thüringer Forschungsdatenrepositorium bietet dazu eine Schnittstelle, über die Forschungsdaten in entsprechende SIP-Pakete (inkl. Metadaten) umgewandelt und überführt werden können. Die Metadaten des Datensatzes und der Status der Zugänglichkeit bleiben im Repositorium erhalten, aber es wird eine Information abgelegt, dass die Forschungsdaten auch in einem Archiv gesichert wurden.

Anhang 1: Begriffe

Begriff	Kurzbeschreibung	Alternativen
Workspace	Ort im System zur temporären Speicherung und Bearbeitung eines vorläufigen Datensatzes vor der Veröffentlichung	Staging Area Bearbeitungsbereich
Vorläufiger Datensatz	Daten im Workspace	Entwurf/Draft Einreichung
Metadaten	Beschreibende Daten zu Forschungsdaten (wie Veröffentlichungsdatum, Autor*innen, Identifikatoren, Schlüsselwörter, …)	
Eintrag	Auffindbarer Datensatz mit Metadaten	Metadaten und Publikation/Veröffentlichung
Datensatz	Forschungsdaten + Metadaten	Forschungsdatenpaket Datenpaket
Forschungsdaten	Daten (ohne Metadaten)	Primärdaten
Open Access	Offener Zugang zu Forschungsdaten	Public Access
Embargoed Access	Offener Zugang zu Forschungsdaten erst nach Ablauf einer Embargofrist	
Closed Access	Kein Zugang zu Forschungsdaten, Zugang zu Forschungsdaten nur nach expliziter Freischaltung	
External Access	Zugang zu Forschungsdaten wird über externe Plattform geregelt.	
Gast	Keine Registrierung nötig; Zugang nur zu öffentlichen Forschungsdaten (Open Access)	Anonymer Nutzer
Datengeber*in	Registrierte*r Benutzer*in; kann Datensätze anlegen und veröffentlichen; kann für die eigenen Datensätze Rechte vergeben	
Redakteur*in	Ansprechperson für Gäste und Benutzer; prüft und veröffentlicht Datensätze	Kurator*in
Administrator*in	Technischer Betrieb	

Anhang 2: Rechte der Benutzergruppen

	Öffentliche Datensätze anschauen	Recherche über öffentliche Datensätze	Zugang mittels Uni-Login	Zugang mittels System-Account	Dashboard (berechtigte Datensätze anzeigen lassen)	Workspace erstellen	Freigabe Workspace zu Eintrag erteilen	Freigabe Datensatz bei "Closed Access" erteilen	Systemprotokolle anzeigen lassen	Kontaktformular zu Benutzer des Datensatzes	Kontaktformular allgemein
Gast	X	X								X	X
Datengeber*in	X	X	X	X	X	X	X	X		X	X
Eingeladene Datengeber*in	X	X	X	X	X					X	X
Redakteur*in	X	X	X	X	X	X	X	X		X	X
Administrator*in	X	X	X	X	X	X	X	X	X	X	X

Benachrichtigungen

	Änderung eigener Datensatz	Änderung Datensatz einer Hochschule	Änderung Datensätze	Freigabeanforderung Workspace	Anfrage zu "Closed Access" Datensatz	Auftreten von Systemproblemen
Gast					X	
Datengeber*in	X				X	
Eingeladene Datengeber*in	X					
Redakteur*in	X	X		X	X	
Administrator*in	X	X	X	X	X	X

	Lesezugriff auf eigene Workspaces	Schreibzugriff auf eigene Workspaces	Rechtevergabe für eigene Workspaces	Lesezugriff auf Workspaces der Hochschule	Schreibzugriff auf Workspaces der Hochschule	Rechtevergabe auf Workspaces der Hochschule	Lesezugriff auf alle Workspaces	Schreibzugriff alle Workspaces	Rechtevergabe alle Workspaces
Gast									
Datengeber*in	X	X	X						
Eingeladene Datengeber*in	X	X							
Redakteur*in	X	X	X	X	X	X			
Administrator*in	X	X	X	X	X	X	X	X	X