Bausteine Forschungsdatenmanagement
Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern

Kriterien für die Auswahl einer Softwarelösung für den Betrieb eines Repositoriums für Forschungsdaten[i]

Alexandra Axtmann Felix Bach Jonathan Bauer André Blessing
Thomas Bönisch Nina Buck Holger Gauza Jan Hess Alexander Holz
Kerstin Jung Roland S. Kamzelak Andreas Kaminski
Heinz Werner Kramski Peter Krauß Jonas Kuhn Volodymyr Kushnarenko
Matthias Landwehr Jan Leendertse Björn Schembera
Claus-Michael Schlesinger Gabriel Schneider Lorena Steeb
Dirk von Suchodoletz Irene Schumm Džulia Terzijska Mona Ulrich
Gabriel Viehhauser

2021

Zitiervorschlag

Axtmann, Alexandra, Felix Bach, Jonathan Bauer, André Blessing, Thomas Bönisch, Nina Buck, Holger Gauza, Jan Hess, Alexander Holz, Kerstin Jung et al. 2021. Kriterien für die Auswahl einer Softwarelösung für den Betrieb eines Repositoriums für Forschungsdaten. *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern* Nr. 3/2021: S. 14-26. DOI: 10.17192/bfdm.2021.3.8348.

[i]Alexandra Axtmann (ORCID: 0000-0001-5303-5352), Felix Bach (ORCID: 0000-0002-5035-7978),

1 Zusammenfassung

Die zitierbare und öffentliche Bereitstellung von Forschungsdaten im Sinne von Open Science ist Bestandteil des Lebenszyklus von Forschungsdaten und wird als solche von immer mehr Forschungsförderern verlangt. Für die technische Umsetzung und den Betrieb eines Repositoriums ist die Auswahl einer geeigneten Software in Übereinstimmung mit verschiedenen Anforderungen notwendig. Die jeweiligen Anforderungen basieren dabei auf einer Analyse der vorhandenen Service- und Systemlandschaft. Die Relevanz der Anforderungen speist sich dabei auch aus den Zielen der jeweiligen betreibenden Einrichtung. Dieser Artikel sammelt mögliche Kriterien und diskutiert deren Relevanz mit dem Ergebnis, dass es keine allgemeingültige Checkliste für die Auswahl einer Softwarelösung gibt und diese von den individuellen Anforderungen der betreibenden Einrichtung abhängt. Der Artikel richtet sich an FDM-Praktikerinnen und FDM-Praktiker und Einrichtungen, die sich mit der Suche nach einer Softwarelösung für den Aufbau und Betrieb eines Repositoriums befassen.

2 Einleitung

Die öffentliche Bereitstellung von Forschungsdaten zur Nachnutzung im Sinne von Open Science ist Bestandteil des Lebenszyklus von Forschungsdaten und erlangt zunehmende Relevanz. Eine zitierbare Veröffentlichung dieser Daten zeugt von einer transparenten Forschung, belegt die Forschungsleistung von Forschenden sowie der jeweiligen Einrichtung und macht Forschung reproduzierbar und damit überprüfbar. Immer mehr Forschungsförderer erwarten bereits bei der Antragstellung die Dokumentation und Planung eines umsichtigen und nachhaltigen Umgangs mit

---

Jonathan Bauer (ORCID: 0000-0002-5624-2055), André Blessing (ORCID: 0000-0001-7573-578X), Thomas Bönisch (ORCID: 0000-0003-3108-8597), Nina Buck (ORCID: 0000-0002-4651-6040), Holger Gauza (ORCID: 0000-0003-0191-3680), Jan Hess (ORCID: 0000-0003-4162-2132), Alexander Holz (ORCID: 0000-0002-7465-2795), Kerstin Jung (ORCID: 0000-0002-9548-8461), Roland S. Kamzelak (ORCID: 0000-0003-4512-2047), Andreas Kaminski (Höchstleistungsrechenzentrum Stuttgart (HLRS)), Heinz Werner Kramski (ORCID: 0000-0001-5216-278X), Peter Krauß (ORCID: 0000-0002-5869-352X), Jonas Kuhn (ORCID: 0000-0003-2860-5960), Volodymyr Kushnarenko (ORCID: 0000-0001-7427-2410), Matthias Landwehr (ORCID: 0000-0001-9274-2578), Jan Leendertse (ORCID: 00000-0001-5676-493X), Björn Schembera (ORCID: 0000-0003-2860-6621), Claus-Michael Schlesinger (ORCID: 0000-0001-6718-5773), Gabriel Schneider (ORCID: 0000-0001-6573-3115), Lorena Steeb (ORCID: 0000-0002-7577-5412), Dirk von Suchodoletz (ORCID: 0000-0002-4382-5104), Irene Schumm (ORCID: 0000-0002-0167-3683), Džulia Terzijska (ORCID: 0000-0002-1698-6826), Mona Ulrich (ORCID: 0000-0001-9591-5614), Gabriel Viehhauser (ORCID: 0000-0001-6372-0337).

Forschungsdaten, beispielsweise in Form eines Datenmanagementplans, der unter anderem Angaben zu geplanten Lizenzen für Forschungsdaten, Rechten an Daten etc. enthält. Die Umsetzung des Datenmanagementplans ist ein kontinuierlicher Prozess im Laufe eines Projekts und nicht auf eine Datenveröffentlichung zum Projektende hin beschränkt. Der Umgang mit Forschungsdaten wird unter anderem in den Richtlinien Guter Wissenschaftlicher Praxis[1], den Open-Access-Policies von Hochschulen, Forschungsinstituten und Forschungsförderern sowie in den „Data Policies" von Zeitschriften adressiert, wobei Repositorien eine zentrale Rolle spielen. Im allgemeinen Sprachgebrauch bezieht sich der Begriff *Repositorium* auf einen Speicherort oder eine Informationsinfrastruktur für die Speicherung, Organisation und Bereitstellung von (Forschungs)Daten.[2] Im Folgenden bezieht sich der Begriff *Repositorium* auf eine Softwarelösung, die - eingebettet in eine Organisationsstruktur und gegebenenfalls im Kontext weiterer Systeme - Forschungsdaten übernimmt, verwaltet und publiziert. Repositorien bilden bei diesem Sprachgebrauch das technische Grundgerüst für das Forschungsdatenmanagement, da sie den gesamten Prozess von der Übernahme über die Qualitätskontrolle bis hin zur zitierfähigen Veröffentlichung unterstützen. Softwarelösungen für Repositorien sind für unterschiedliche Zwecke und Einsatzszenarien verfügbar. Zu den verbreitetsten zählen beispielsweise Fedora, DSpace, MyCoRe, Islandora, EPrints, Dataverse, Rosetta, Archivematica und Invenio. Die Bestimmung von Kriterien für die Auswahl eines Repositoriums ist nicht trivial und es müssen neben Aspekten der Wirtschaftlichkeit, Skalierbarkeit und Funktionalität noch weitere wie die Dokumentation, Verbreitung, Entwicklungsperspektive sowie das Daten- und Lizenzmodell berücksichtigt werden. Der Aufwand für die Erarbeitung eines Kriterienkatalogs darf nicht unterschätzt werden. Zwangsläufig ergeben sich Abhängigkeiten zur betreibenden organisatorischen Einheit und der grundlegenden technischen Infrastruktur für den Betrieb der Software und die Speicherung der Daten. Die in diesem Artikel präsentierten und diskutierten Aspekte basieren auf den Erfahrungen der beteiligten Autorinnen und Autoren und werden von diesen als besonders relevant eingeschätzt.[3] Die Absicht ist es, FDM-Praktikerinnen und FDM-Praktikern und Institutionen, die auf der Suche nach einer Softwarelösung für den Aufbau und Betrieb eines Repositoriums

[1] Siehe auch Kodex der DFG zur guten wissenschaftlichen Praxis: https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp/ [Letzter Aufruf 10.06.2021].

[2] Siehe hierzu die Definitionen unter https://www.forschungsdaten.org/index.php/Repositorium [Letzter Aufruf 10.06.2021] und https://www.forschungsdaten.info/themen/veroeffentlichen-und-archivieren/repositorien/ [Letzter Aufruf 02.06.2021]. Eine Diskussion verschiedener Begriffsverwendungen bietet beispielsweise Nicole Offhaus, „Institutionelle Repositorien und Universitätsbibliotheken - Entwicklungsstand und Perspektiven" (Bachelor Thesis, Fachhochschule Köln, Fakultät für Informations- und Kommunikationswissenschaften, Institut für Informationswissenschaft, 2012) urn:nbn:de:hbz:79pbc-20120712112 .

[3] Die Autorinnen und Autoren arbeiten im Kontext der vom Land Baden-Württemberg geförderten vier Science Data Center am Aufbau und Betrieb von fachspezifischen Repositorien mit.

- Literaturforschung: SDC4Lit, https://www.sdc4lit.org [Letzter Aufruf 10.06.2021]
- Wirtschaftswissenschaften: BERD@BW, https://www.berd-bw.de/ [Letzter Aufruf 10.06.2021]
- Molekulare Materialforschung: MoMaF, https://momaf.scc.kit.edu/ [Letzter Aufruf 10.06.2021]
- Bioinformatik: BioDATEN, https://portal.biodaten.info/ [Letzter Aufruf 10.06.2021].

sind, Impulse und Denkanstöße zu liefern.

3 Vorgaben an Repositorien durch den Anwendungsbereich

Für die erfolgreiche Implementierung und Nutzung eines Repositoriums ist die genaue Definition des Anwendungsbereichs notwendig. Hieraus lassen sich die Anforderungen der Einrichtung und die Auswahlkriterien an die Software formulieren. Hinzu kommen noch Schätzungen hinsichtlich der prospektiven Entwicklung unter anderem von Datenmengen, Objektanzahl und Nutzerzahlen. Zusätzlich muss die generelle Verortung des Repositoriums als organisatorische Einheit im Lebenszyklus von Forschungsdaten betrachtet werden. Hierbei unterscheiden sich die Anforderungen von ‚kalten', ‚lauwarmen' und ‚heißen' Forschungsdaten.[4] Letztere erfordern ein höheres Maß an Bearbeitbarkeit von Metadaten und Datenobjekten und in der Regel ist kein Repositorium gesucht, sondern eine Umgebung für kollaboratives Arbeiten. Ein weiterer, wichtiger Aspekt ist die Frage, ob ein community- beziehungsweise disziplinspezifisches, ein institutionelles oder ein generisches und allgemein zugängliches Repositorium ohne spezifische Zielgruppe aufgebaut werden soll. Das Repositorium darf hinsichtlich der Unterstützung von Dateiformaten, Dateigrößen und Dateianzahl nicht zu unerwünschten Einschränkungen bei der Nutzung führen. Die hier genannten Anforderungen berücksichtigen keine organisatorischen und infrastrukturellen Anforderungen, wie sie beispielsweise für das CoreTrustSeal umgesetzt werden müssen.[5] Folgende Punkte sind relevant:

- In welcher Umgebung (andere Systeme, Organisationsstruktur etc.) soll das Repositorium eingesetzt werden und welches Anwendungsgebiet soll abgedeckt werden?
- Soll ein breit aufgestellter Publikationsdienst an einer Universitätsbibliothek (UB) oder institutionsübergreifend aufgebaut werden?
- Soll ein eng definierter Publikationsdienst mit Fokus auf eine Fachrichtung oder ein allgemeiner Publikationsdienst aufgebaut werden? Je spezifischer die Ausrichtung ist, desto eher gibt es bereits wohldefinierte Anforderungen innerhalb einer Fachcommunity.
- Soll das Repositorium mit oder ohne Weboberfläche für Endnutzerinnen und Endnutzer betrieben werden?

[4] Das Konzept der Unterscheidung zwischen heißen und kalten Daten wurde wahrscheinlich aus der Informationswissenschaft entlehnt. Auf Forschungsdaten übertragen hat sich folgende Verwendung etabliert: Kalte Daten: Die Forschungsdaten wurden publiziert. Lauwarme Daten: An den Forschungsdaten wird weitgehend nicht mehr aktiv gearbeitet. Heiße Daten: An den Forschungsdaten wird aktiv gearbeitet.

[5] Für mehr Informationen über das CoreTrustSeal siehe https://www.coretrustseal.org/ [Letzter Aufruf 10.06.2021].

- Mit welcher Zugriffshäufigkeit auf die Forschungsdaten ist zu rechnen?
- Ist eine rechtssichere Zugriffsverwaltung auf die Forschungsdaten notwendig?
- Mit welchen Datenvolumina und jährlichem Datenzuwachs ist zu rechnen?
- Sollen nur die Daten abgeschlossener oder auch die laufender Projekte übernommen werden?
- Welche Daten sollen gespeichert werden und wie sollen diese organisiert sein? Sollen Verzeichnisstrukturen oder gepackte Dateien (zip-File, Tarball, BagIT, OCFL etc.) eingesetzt werden?

4 Kosten- und Lizenzmodelle

Ein wichtiges Entscheidungskriterium für die Auswahl eines Repositoriums ist das dahinterstehende Kosten- bzw. Lizenzmodell. Neben kommerziellen Anbietern, die ein Produkt gegen eine Lizenzgebühr zur Verfügung stellen und dessen Quellcode nicht öffentlich ist, existieren Akteure, die für ein kostenlos nutzbares Open-Source-Produkt Dienstleistungen wie Auftragsentwicklung und Schulungen anbieten.[6] Das Kosten- bzw. Lizenzmodell für die reine Nutzung der Softwarelösung hat somit direkte Auswirkungen auf die Betriebskosten. Eine kostenfreie Nutzung von Software ist häufig mit dem Begriff *Open Source* verknüpft, jedoch ist die Verbindung nicht zwangsweise gegeben. Zu den relevanten und zu prüfenden Punkten zählen:

- Fallen (einmalige oder jährliche) Kosten bzw. Lizenzgebühren für die Nutzung der Softwarelösung an?
- Ist (evtl. kostenpflichtiger) Support über Anbieter oder eine Community verfügbar?
- Welche Leistungen, z. B. Bedingungen für die Weiterentwicklung, umfasst der Support?
- Ist der Quellcode öffentlich zugänglich und unter welcher Lizenz wird dieser bereitgestellt?

5 Der grundlegende Aufbau

Repositorien als Software unterscheiden sich hinsichtlich ihres Aufbaus (monolithisch vs. modular) und Selbstanspruchs (Framework vs. Turnkey).[7] Monolithische Repositorien vereinen alle Funktionen „am Stück“, während bei modularen Repositorien funktionale Module durch Schnittstellen wie REST oder SWORD miteinander verbunden werden und dadurch viele Anpassungsmöglichkeiten bieten. Zudem skalieren

[6] Diese Anbieter sind häufig auch an der Entwicklung des Open-Source-Produkts stark beteiligt. Ein Beispiel hierfür ist Artefactual Systems mit Archivematica.

[7] Beispiele für die Einteilung: Monolithisch: Rosetta. Modular: Archivematica. Framework: Invenio. Turnkey: Dataverse.

sie mit der Anzahl der Nutzenden und dem Datenumfang besser als monolithische Repositorien. Im Aufbau sind sie allerdings komplexer. Frameworks bieten lediglich ein Gerüst, und es obliegt den betreibenden Einrichtungen, eine vollständige Lösung zu implementieren. Hierdurch ergeben sich zahlreiche Anpassungsmöglichkeiten. Im Gegensatz dazu sind Turnkey-Lösungen direkt nach der Installation betriebsbereit, haben aber wesentlich begrenztere Konfigurationsmöglichkeiten. Um flexible Lösungen aufbauen, anpassen und betreiben zu können, sind Programmierfertigkeiten beim eigenen Personal erforderlich. Folgende Fragen sind hilfreich:

- Welche Architektur (monolithisch vs. modular) implementiert die Software?
- Wie erfolgt die Installation der Software?
- Werden moderne container- oder cloud-basierte Betriebskonzepte unterstützt?[8]
- Auf welchen Programmiersprachen basiert die Software?
- Kann die Software durch selbstentwickelte Funktionalitäten erweitert werden?
- Wie hoch ist der Installations-, Konfigurations- sowie Wartungsaufwand der Software?

6 Kontextanalyse

Ein Repositorium wird in eine bestehende infrastrukturelle und organisatorische Umgebung eingebettet und löst möglicherweise ein bestehendes System ab. Es hat sich daher in bestehende Workflows einzufügen und mit existierenden Systemen zu kommunizieren. Es muss somit genau ermittelt und bei der Auswahl beachtet werden, welche Funktionen und Schnittstellen ein Repositorium mitbringen muss, beziehungsweise wie hoch die Entwicklungs- und Anpassungsarbeiten sein werden. Relevante Schnittstellen umfassen unter anderem Metadatenharvesting via OAI-PMH[9], die Anbindung an ORCID und andere Identity-Provider zur Authentifizierung oder die Anbindung an einen PID-Registrar. Folgende Aspekte sind relevant:

- Ist die Übernahme von Bestandsdaten notwendig und wie groß ist der Migrations- und Integrationsaufwand in diesem Falle?
- Welche Schnittstellen zu bestehenden Diensten sind beispielsweise für Authentifizierung, Identifikation[10], Übernahme von (Bestands)Daten und Metadaten[11], Datenweitergabe oder Recherche notwendig?

[8] Ein Container enthält Software einschließlich aller Abhängigkeiten, die zur Ausführung notwendig sind. Dadurch wird sichergestellt, dass containerisierte Software in verschiedenen Umgebungen eingesetzt werden kann. Container können in Cloud-Infrastrukturen eingesetzt werden, wodurch eine höhere Verfügbarkeit und Skalierbarkeit erreicht werden kann.

[9] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), https://www.openarchives.org/pmh/ [Letzter Aufruf 10.06.2021].

[10] Persistente Identifikatoren (PIDs) für Datensätze und -objekte wie DOIs und Handle, für Personen wie ORCID ID oder für Institutionen wie ROR ID.

[11] Z. B. aus anderen Repositorien oder aus Arbeitsumgebungen.

7 Verbreitung, Nachhaltigkeit und Transparenz

Die Verbreitung und die Größe der nutzenden Einrichtungen einer Software lassen meistens Rückschlüsse auf die Entwicklungsperspektive zu. Eine große Verbreitung deutet auf breite Akzeptanz hin und spricht für eine – zumindest momentan – gesicherte Entwicklungsperspektive. Besonders bei Open-Source-Produkten ist die Größe der Entwicklercommunity oder die Beteiligung großer und namhafter Einrichtungen dafür entscheidend, dass Probleme schnell behoben und Funktionen (weiter)entwickelt werden. Gegebenenfalls wird diese Community von einzelnen Institutionen oder Serviceanbietern dominiert, wodurch der Spielraum für das Einbringen eigener Anforderungen eingeschränkt wird. Eine nachhaltige Entwicklung zeichnet sich durch eine konsequente Open-Source-Policy und die Verfügbarkeit von umfassender und aktueller Dokumentation aus. Eine transparente und kontinuierliche Entwicklung zeigt sich durch regelmäßige Commits in öffentlichen Code-Repositories sowie entsprechenden Changelogs und Wikis. Die relevanten Punkte umfassen:

- Wie viele Institutionen nutzen die Software?
- Welche Institutionen (Größe und Impact) nutzen die Software?
- Sind diese Institutionen an der Entwicklung selbst beteiligt?
- Wie groß ist die Entwicklercommunity und wie breit ist das Projekt aufgestellt?
- Wie transparent ist die Entwicklung und die Dokumentation, z. B. auf GitHub?
- Sind Gremien wie wissenschaftliche Beiräte und Nutzerforen eingerichtet?

8 Skalierbarkeit

Es ist damit zu rechnen, dass in Zukunft immer mehr und größere Datenpakete von einem Repositorium verwaltet und bereitgestellt werden müssen. Bei der Auswahl eines Repositoriums ist deshalb darauf zu achten, dass das technische Grundgerüst bezüglich möglicher zukünftiger Entwicklungen geeignet ist. Hier sind zwei Komponenten zentral: Die zugrundeliegende Datenbank für die Verwaltung von Daten und Metadaten sowie die Speicherkomponenten für die Datenobjekte. Hier gilt momentan die Verwendung von Objektspeichern als besonders zukunftsfähig gegenüber klassischen Dateisystemen, da sie eine (hohe) Georedundanz und effektive Speicherung über Einrichtungsgrenzen hinweg ermöglichen. Der Einsatz von Message Queues und Load Balancing ermöglicht den verteilten Betrieb des Repositoriums auf mehreren Knoten bzw. Servern, eine größere Leistungsfähigkeit und eine höhere Verfügbarkeit. Für die Bewertung der Skalierbarkeit bezüglich der Anzahl verwaltbarer Objekte und Datenmengen, Metadaten, Zugriffszahlen etc. ist eine Einschätzung von Expertinnen und Experten unerlässlich. Folgende Punkte sind relevant:

- Welche Art von Datenbank wird verwendet?
- Welche Speicher sind nutzbar?
- Sind mehrere Speicherschichten parallel nutzbar?

9 Metadaten und Datenmodell

Grundlegend für den Einsatz des Repositoriums im Sinne von Open Science ist die Berücksichtigung und Umsetzbarkeit der FAIR-Prinzipien, welche auf die Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendbarkeit von Forschungsdaten abzielen.[12] Welche Angaben dabei in den Metadaten erfasst werden müssen, richtet sich sehr stark nach dem zu beschreibenden Datenobjekt. Die Verwaltung von klassisch bibliothekarischen Daten benötigt andere Metadaten als Forschungsdaten im Kontext der Strömungssimulation. Daher sollte ein Repositorium es entweder ermöglichen, eigene Metadatenschemata anzulegen, zu versionieren, durchsuchbar zu machen und wohlgeformt zu exportieren oder (community-spezifische) Services einzubinden. Die Integration von Vokabularien unterstützt Datenerzeugerinnen und Datenerzeuger in der Vergabe von Metadaten, erhöht deren Qualität und verbessert die Interoperabilität. Flexibilität muss ebenfalls bei der Zuordnung von Metadatum und (Forschungs)Datum gewährleistet sein. Eine 1:1-Relation von Forschungs- und Metadatum ist eher die Ausnahme als die Regel. Häufiger tritt der Fall auf, dass ein Metadatensatz mehrere Datensätze beschreiben kann (1:n). Entsprechend muss ein Repositorium beispielsweise mehrere Datenobjekte einem Metadatensatz zuordnen oder diese logisch gruppieren. Ebenfalls kann es wichtig sein, Forschungsdaten nachträglich unter einem Meta-Metadatum zu gruppieren, um zum Beispiel alle Projekte eines Sonderforschungsbereichs zusammenzufassen. Je nach Ausrichtung des Repositoriums kann es zu Änderungen an den Metadaten und den Forschungsdaten selbst kommen, besonders dann, wenn Daten laufender Projekte übernommen werden. Hierbei kann es zu Änderungen, wie Ergänzungen/Löschungen und Korrekturen an den Daten und Metadaten, kommen.[13] Das Repositorium muss in diesem Fall Objekte verwalten können, die sich bei notwendigen Änderungen versioniert speichern lassen ohne vorherige Versionen zu verlieren. Die relevanten Fragen lauten also:

- Wird die Umsetzung der FAIR-Prinzipien unterstützt?
- Welche Metadatenschemata werden unterstützt?
- Können eigene Metadatenfelder definiert werden?
- Gibt es Pflichtfelder, die immer ausgefüllt werden müssen?
- Können Metadaten nachträglich geändert werden?
- Können Forschungsdaten nachträglich hinzugefügt werden?
- Wird eine Versionierung von Daten und Metadaten unterstützt?
- Können Forschungsdaten nachträglich gruppiert oder verknüpft werden?
- Können Lizenzen auf Daten und Datensatz-Ebene sowie Freigabe-Möglichkeiten wie Embargo, Zugriffsberechtigungen für bestimmte Personen(-kreise) umgesetzt oder erfasst werden?

[12] https://www.go-fair.org/fair-principles/ [Letzter Aufruf 10.06.2021].

[13] Die Auswahl der Speicherschicht hat direkte Auswirkungen auf die Umsetzung einer Versionierung beziehungsweise der Durchführung von Änderungen. Im Falle einer Änderung werden vom Objektspeicher neue Objekte angelegt, sodass der ursprüngliche Zustand erhalten bleibt. Bei konventionellen Dateispeichern können Dateien direkt geändert werden, sodass gegebenenfalls Diffs (Änderungen im Vergleich zu vorheriger Version) erzeugt und gesichert werden müssen, welche die Unterschiede zu vorherigen Versionen enthalten.

10 Authentifizierung und Rechteverwaltung

Für die alltägliche Anwendung und Betreuung etc. werden Repositorien in die IT-Infrastruktur integriert und sollen auf vorhandene Mechanismen zur Authentifizierung zurückgreifen. Entsprechend ist eine Anbindung vorhandener Authentifizierungs- und Autorisierungssysteme (AAI) wie LDAP oder Shibboleth notwendig. Während die Integration in vorhandene Systeme in vielen Fällen leicht umsetzbar ist, erfordert ein damit verknüpftes Rechte- und Rollenmanagement Klarheit über konzeptionelle Fragen. So sollte es möglich sein, dass Forschende, die ihre Einrichtung verlassen haben, nach wie vor bestimmte Aktionen mit ihren publizierten Datensätzen durchführen können (Ergänzen, Aktualisieren von Metadaten, versionierte Fehlerbereinigung, Kommunikation mit Dritten bei Anfragen). Neben datenbezogenen Rechten sind auch die Rechte und Rollen zu beachten, welche für die umzusetzenden Workflows notwendig sind. Hierzu zählt unter anderem die Umsetzung der Rolle einer Data Stewardess oder eines Data Stewards für die Qualitätskontrolle, einen Peer-Review-Prozess oder ein Vier-Augen-Prinzip. Folgende Anforderungen müssen geklärt werden:

- Gibt es ein Konzept der umzusetzenden Workflows beziehungsweise Use Cases hinsichtlich notwendiger Rollen und Rechte?
- Welche Rechte und Rollen sind software-seitig vorgesehen und können diese erweitert werden?
- Auf welchen Ebenen (Projekt, Datenobjekt etc.) können Rechte beziehungsweise Rollen festgelegt werden?
- Wie können Embargofristen umgesetzt werden?
- Müssen Datenschutzbeschränkungen umgesetzt werden?
- Welche AAI-Lösungen bzw. welche Schnittstellen werden unterstützt (OpenID connect/oAuth/Shibboleth, ORCID)?

11 Ingest

Der Begriff *Ingest* wird hier breit gefasst und bezeichnet den gesamten Prozess der Datenübernahme. Die Anforderungen an diesen Prozess sind von dem Einsatzszenario der betreibenden Institution abhängig. Übergeben Anwenderinnen und Anwender ihre Daten selbst, dann sind entsprechende Interfaces und (Web-)Formulare notwendig. Zum Tragen kommen hier Fragen zur Annotation mit Metadaten auf verschiedenen Ebenen (Projekt, Datenobjekt etc.). Bei einer Übernahme von Bestandsdaten sind Schnittstellen wie REST, SOAP oder SWORD notwendig und es müssen wahrscheinlich Plug-ins verwendet und gegebenenfalls angepasst oder eigenständige Apps erstellt werden, welche die Bestandsdaten übertragen, Metadaten von Quell- auf Zielschemata mappen oder Datenobjektstrukturen anpassen. Ingest-Workflows sollten immer eine Qualitätssicherung durch Validatoren und Identifikatoren für Dateitypen wie VeraPDF oder FITS beinhalten und vorhandene Prüfsummen abgleichen, um frühzeitig die Les- und Interpretierbarkeit von Dateien zu gewährleisten. Zu fragen ist:

- Welche Möglichkeiten bzw. Schnittstellen zum Ingest (manuell vs. automatisch) sind vorhanden?
- Wie ist der Ingest-Prozess für die Anwenderinnen und Anwender angelegt?
- Sind Plug-ins zur Validierung von Datensätzen vorhanden?
- Werden Prüfsummen generiert und mit bereits vorhandenen abgeglichen?

12 Bearbeitung und Kuratierung

In Abhängigkeit von vorhandenen Workflows und dem geplanten Einsatzfeld muss über die Kuratierung von Metadaten und Datenobjekten nachgedacht werden. Bei Selbsteingabe durch die Forschenden sollten Metadaten und Datenobjekte auf Vollständigkeit, Korrektheit und Lesbarkeit überprüft werden. Die Prüfung der Lesbarkeit (lässt sich eine Datei öffnen und anzeigen) kann automatisch mit Validatoren erfolgen, während die inhaltliche Überprüfung Fachwissen voraussetzt. Die Repositorien-Software sollte daher entsprechende Workflows unterstützen, um eine Datenfreigabe nach einer Überprüfung vorzunehmen. Hierbei können gegebenenfalls datenschutzrechtliche Problemfälle vor einer Veröffentlichung abgefangen werden. Bei Änderungen von Datenobjekten (Ergänzungen, Korrekturen oder Löschung von einzelnen Dateien) müssen Abläufe und Vereinbarungen definiert werden, welche mit Hilfe des Repositoriums umgesetzt werden sollen. Eine Versionierung der Datenobjekte und Metadaten bei eindeutiger Referenzierbarkeit mittels PIDs ist hierbei anzustreben.

- Welche Workflows sind zur Überprüfung von abgegebenen Metadaten verfügbar?
- Können Metadaten und Datenobjekte nachträglich korrigiert oder erweitert werden? Sind diese Änderungen transparent (Versionierung) für die Nutzenden?
- Lassen sich Peer-Review-Workflows und gestufte Freigaben umsetzen?
- Wie sehen die Möglichkeiten zur Qualitätssicherung (Validierung, etc.) aus?

13 Suche und Anzeige

Damit Forschungsdaten im Sinne von FAIR nachhaltig und Forschungsergebnisse reproduzierbar sind, müssen die Daten suchbar, auffindbar und im Idealfall direkt abrufbar sein. Entsprechend ist neben der Bereitstellung über PIDs auch eine Suchfunktion innerhalb eines Repositoriums notwendig, welche die Eingrenzung und Suche beispielsweise anhand von Schlagwörtern, Datentypen, klassischen und fachspezifischen Metadaten erlaubt. Neben Fragen zum Metadatenmodell und der Indexierung von Metadatenfeldern geht es hierbei um die integrierten Suchfunktionalitäten wie Volltext- oder Facettensuche.

Forschungsdaten und Suchfunktion müssen den Nutzerinnen und Nutzern auf eine intuitive Art angeboten werden. Hierfür ist die Einbindung des Repositoriums in ein Portal oder eine Webpage über entsprechende Schnittstellen denkbar. Haben Forschende relevante Forschungsdaten identifiziert, sollten sie sich diese herunterladen oder anzeigen lassen können. Für Ersteres ist es abhängig von der Größe des Datensatzes sinnvoll, nur einzelne Dateien aus einem Datensatz herunterladen zu können. Für Letzteres ist die Möglichkeit zur Datenvisualisierung oder die Übergabe an einen externen Service wünschenswert bzw. notwendig. Jedoch ist es gerade bei Forschungsdaten eher wahrscheinlich, dass es für proprietäre oder seltene Formate keine geeigneten Visualisierungswerkzeuge gibt.
Hinsichtlich der Präsentation sollte ein entsprechendes Branding durch die betreibende Institution und gegebenenfalls weiterer Untereinheiten wie Institute, Abteilungen oder Projekte möglich sein. Zusammenfassend sind folgende Fragen relevant:

- Welche Metadatenfelder oder Datei-Inhalte werden für die Suche indexiert?
- Welche Suchfunktion kommt zum Einsatz?
- Lassen sich Beziehungen zwischen Datensätzen darstellen?
- Wie lässt sich das Frontend des Repositoriums in bestehende Seiten einbinden?
- Wird gegebenenfalls bereits eine Datenvisualisierung unterstützt? Wenn ja, für welche Dateiformate?
- Welche Möglichkeiten in Bezug auf das institutionelle Branding existieren?

14 Langzeitarchivierung und Integrität

Grundlage aller Bemühungen im Sinne der Langzeitarchivierung von Forschungsdaten ist der Erhalt der Datenstromintegrität und die Vermeidung von proprietären Dateiformaten bei der Datenübernahme.[14] Ein Repositorium kann hierzu einen wichtigen Beitrag leisten, indem es die Integrität des Datenstroms technisch schützt und damit Datenverlust verhindert. Auf die Vermeidung von proprietären Dateiformaten sowie auf die verstärkte Nutzung offener Formate muss hingegen auf organisatorischer Ebene hingearbeitet werden. Ein weiterer Ansatzpunkt für die Langzeitarchivierung von Forschungsdaten ist die Erfassung von Angaben zur Provenienz, eingeräumten Rechten an Forschungsdaten wie zum Beispiel die Erlaubnis zur Umwandlung in langzeitstabile Dateiformate und die Erfassung von Angaben zur verwendeten Hard- und Softwareumgebung. In der Langzeitarchivierungscommunity ist PREMIS hierfür der De-facto-Standard.[15] Die relevanten Angaben können bereits beim Ingest unter anderem durch Validatoren gewonnen werden. Folgende Punkte sollten berücksichtigt werden:

[14] Für weitere Informationen siehe https://www.langzeitarchivierung.de/ [Letzter Aufruf 10.06.2021].
[15] **PRE**servation **M**etadata: **I**mplementation **S**trategies: https://www.loc.gov/standards/premis/ [Letzter Aufruf 10.06.2021].

- Orientiert sich das Repositorium am Open Archival Information System Modell (OAIS)?[16]
- Werden vom Repositorium regelmäßig Prüfsummen berechnet und mit vorhandenen abgeglichen, um Datenverlust frühzeitig zu erkennen?
- Wird die Verwendung moderner Dateisysteme wie ZFS und btrfs unterstützt, welche Fehler im Datenstrom selbstständig korrigieren?
- Werden Metadaten speziell für die Langzeitarchivierung erfasst?
- Bietet das Repositorium Schnittstellen zu Langzeitarchivierungssystemen?

15 Risiko und Exitstrategie

Ein Repositorium ist eine komplexe Software und auf das Zusammenspiel mehrerer Komponenten angewiesen, um definierte oder selbst angestrebte Servicelevel zu erreichen. Ein erhebliches Risiko für den Betrieb resultiert aus fehlender Weiterentwicklung, sodass beispielsweise nach notwendigen Software- und Sicherheits-Updates des Basissystems Komponenten nicht mehr funktionieren. Ebenso könnten Schnittstellen veralten und nicht mehr zur Systemlandschaft einer Einrichtung passen, die sich auch ständig weiterentwickelt. Datenverlust kann durch (geo-)redundante Speicherung und angemessene Backupmechanismen kompensiert werden. Neben dem Verlust der Daten ist der Verlust der Zuordnung von Daten zu Metadaten oder der Metadaten selbst besonders kritisch. In diesem Worst-Case-Szenario ist es vorteilhaft, wenn das Repositorium Datenpakete erzeugt, welche neben den enthaltenen Daten auch die zugehörigen Metadaten speichern. Durch solche selbstbeschreibenden Pakete kann ein Repositorium gegebenenfalls leichter wiederaufgebaut werden. Ein Beispiel hierfür sind Datenpakete beziehungsweise Container im BagIT Format oder mit dem neuen Oxford Common File Layout (OCFL). Ein weiterer Vorteil dieser in sich geschlossenen Pakete ist, dass ein Wechsel auf ein anderes System oder die Übergabe an eine Langzeitarchivierungssoftware vereinfacht wird, sofern keine adäquaten Schnittstellen oder Exportmechanismen vorhanden sind.

16 Fazit

Die Suche nach der geeigneten Softwarelösung für den Betrieb eines Repositoriums erfordert eine genaue Analyse der bereits vorhandenen Service- und Systemlandschaft einer Einrichtung hinsichtlich notwendiger Funktionen und Schnittstellen, die Ermittlung der verfügbaren personellen und finanziellen Ressourcen, die Definition umzusetzender Workflows und eine Abstimmung mit dem Infrastrukturbetreiber.

[16] Das OAIS dient als Referenzmodell und beschreibt ein digitales Langzeitarchiv als organisatorische Einheit aus Menschen und Systemen, siehe hierzu: „Das Referenzmodell OAIS - Open Archival Information System“ (Version 2.0): urn:nbn:de:0008-20090811179 .

Viele Fragen erfordern zudem Expertenwissen, gut abgewogene Annahmen und vertiefte Kenntnisse der in Betracht kommenden Repositorien. Neben Introspektion ist hier der Austausch mit Kolleginnen und Kollegen sinnvoll. Es wird deutlich, dass die Auswahl einer Software nicht anhand einer einfachen Checkliste erfolgen kann, da sie keine Allgemeingültigkeit für alle Szenarien und Einrichtungen haben kann. Die im Text dargestellten Punkte sollen diesen Sachverhalt verdeutlichen. Als Ergänzung zu diesem Artikel wurde exemplarisch eine Bewertungsmatrix auf Basis von Überlegungen mehrerer Institutionen erstellt, die weitere Anhaltspunkte für die Auswahl einer geeigneten Software für ein Repositorium bietet. Diese Matrix ist unter https://doi.org/10.5281/zenodo.5562885 abrufbar. Die Gewichtung und Bewertung der Relevanz der einzelnen Kriterien müssen jedoch von der jeweiligen Institution geleistet werden.

Förderung

Dieser Artikel wurde in Kooperation der Projekte bw2FDM, BERD@BW, BioDATEN, MoMaF und SDC4Lit erstellt. Die Förderung der genannten Projekte erfolgt durch das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg im Rahmen der Digitalisierungsstrategie digital@bw.