Design eines FDM-fähigen Speichersystems

Dennis Wehrle,[1] Bernd Wiebelt,[2] Dirk von Suchodoletz[2]

Abstract: Es liegt in der Natur wissenschaftlicher Prozesse, dass viele Zwischenergebnisse, aber auch schlicht irrelevante Forschungsdaten gespeichert werden, die bei regelmäßigen Überprüfungen eigentlich gelöscht werden könnten. Weiterhin passiert es häufig, dass potentiell wertvolle Daten aufgrund von Platzmangel unwiederbringlich gelöscht werden. Ein nachhaltiges Forschungsdatenmanagement schafft den Spagat zwischen der Finanzierbarkeit einer wachsenden Datenmenge bei gleichzeitiger Optimierung der Qualität der Daten. Dieser Beitrag diskutiert, wie klassische technische Lösungen durch geeignete mit den einzelnen Wissenschafts-Communities abgestimmte Steuerungsrahmen ergänzt werden können. So ließe sich das hier beschriebene Speicherkonzept als FDM-Baustein im Nutzen verbessern, indem die Forschenden zu jedem Zeitpunkt qualifizierende Beschreibungen ihrer Daten hinzufügen, wobei die Angabe derselben eine Grundvoraussetzung für eine langfristige Speicherung darstellen. Für einen echten, den Rahmen der einzelnen Forschungsinstitution übergreifenden, Mehrwert können diese Metadaten über standardisierte Schnittstellen abgefragt und in bestehende und zu entwickelnde fachspezifische Workflows integriert werden. Ein FDM-fähiges Speichersystem muss berücksichtigen, dass Daten vieler Forschungsgruppen an verteilten Standorten liegen und von unterschiedlichen Wissenschafts-Communities verwendet werden.

Keywords: Forschungsdatenmanagement, Speichersystem, Storage, Governance, Data Life Cycle

1 Motivation

Sowohl die seit Jahrzehnten exponentiell steigende Verfügbarkeit von Rechenleistung, als auch die parallel dazu gewachsene Leistungsfähigkeit der Geräte zur wissenschaftlichen Datenerfassung hat zu einer Explosion der vorzuhaltenden Daten geführt. Mit gleichzeitig steigender Plattenkapazität wurden vielfach ad-hoc Lösungen etabliert, um für Arbeitsgruppen gemeinsamen Speicherplatz in der einfachst möglichen Form, als gemeinsam sichtbares Dateisystem, zur Verfügung zu stellen. Diese an sich vernünftige Vorgehensweise stößt inzwischen dort an ihre Grenzen, wo Kernleistungen der Wissenschaft tangiert sind, nämlich die erzielten Ergebnisse für andere Wissenschaftler nachvollziehbar und nachnutzbar zu machen (Review-Prozess) und für zukünftige Generationen zu konservieren (Archivierungsprozess). Das Thema Forschungsdatenmanagement (FDM) beschäftigt daher Wissenschaftler, Forschungseinrichtungen, Fördergeber und Politik seit geraumer Zeit. So hat die DFG bereits 1998 in ihrer Denkschrift „*Vorschläge zur Sicherung guter wissenschaftlicher Praxis*“ unter anderem Empfehlungen zur Sicherung und Aufbewahrung von Primärdaten veröffentlicht. Deshalb wird in zunehmendem Maße ein (institutionalisiertes) FDM zur Voraussetzung für die Bewilligung von Zuwendungen für neue

[1] Universität Freiburg, Professur für Kommunikationssysteme, Hermann-Herder-Str. 10, 79104 Freiburg, dennis.wehrle@rz.uni-freiburg.de

[2] Universität Freiburg, Rechenzentrum, Hermann-Herder-Str. 10, 79104 Freiburg, bernd.wiebelt@rz.uni-freiburg.de / dirk.von.suchodoletz@rz.uni-freiburg.de

Forschungsvorhaben. [3] Das IT-gestützte nachhaltige FDM [Ne12] [BHM11] rückt damit in den Fokus zukunftsorientierter Forschungsprozesse. [4] FDM muss sowohl starke organisatorische als auch technische Aspekte vereinen, die sich zudem durch unterschiedliche Zeiträume auszeichnen. Die mit einem wirkungsvollen und nachhaltigen FDM verbundenen Aufwendungen können kaum sinnvoll von einzelnen Forschungsgruppen oder Instituten geleistet werden, sondern sollten durch die Forschungseinrichtungen und ihre zentralen Einrichtungen wie Rechenzentren und Bibliotheken übergreifend koordiniert werden.

Nach einer Phase der Konzept- und Strategieentwicklung haben Wissenschaftseinrichtungen erste Richtlinien erlassen und damit begonnen, Repositorien für Forschungsdaten einzurichten, welche zunehmend in Verbünde integriert werden und einen weltweiten Nachweis ihrer enthaltenen Daten erlauben [EU17]. Forschungsdaten auf nachhaltige Weise zu verwalten und in standardisierte Arbeitsabläufe zu integrieren erzeugt organisatorische Herausforderungen auf mehreren Ebenen. Zum einen sind die Wissenschafts-Communities selbst gefordert, ihre Prozesse entsprechend vorzubereiten und anzupassen. Zum anderen sind die Forschungsinstitution in der Pflicht, in Abstimmung mit den Forschenden passende Soft- und Hardware-Lösungen unter Finanzierungs- und Kosten-Nutzen-Abwägungen zu finden. Wegen der Verschiedenartigkeit der Anforderungen in den einzelnen Disziplinen [5] und schon bestehenden einrichtungs- und fächerübergreifenden Kooperationen lässt sich die Herausforderung effektiv nur in Zusammenarbeit realisieren. Bei einer Kooperation mehrerer Beteiligter auf Ebene der Forschungseinrichtung und in föderierten Verbünden müssen die Interessen der einzelnen Akteure im Gesamtsystem geeignet gegeneinander abgewogen werden. [6] Es werden Vorschläge aus Sicht des Rechenzentrums gemacht, wie ein Konzept- und Speichersystem für ein FDM an einer konkreten Forschungseinrichtung – hier der Universität Freiburg – mit Einbindung in kooperative und föderative Strukturen umgesetzt werden könnte. Diskutiert werden organisatorische und technische Fragen, insbesondere wie die Anreicherung mit fachspezifischen Metadaten gemeinsam mit den Communities realisiert und gesteuert werden kann. Die technische Umsetzung basierend auf einer Hierarchical Storage Management Architektur sollte Nutzern verschiedene Zugangswege zum Speichersystem erlauben und gleichzeitig standardisierte Schnittstellen für Replikation und Austausch der Daten und Metadaten anbieten.

2 Vorüberlegungen in Freiburg

Der Blick auf die Aktivitäten an anderen Forschungseinrichtungen und Gespräche mit einzelnen Forschungs-Communities an der Universität Freiburg ergibt ein breit gefächertes Bild an Anforderungen. Diese beinhalten den kurzfristigen Bedarf an Speicher im zwei bis dreistelligen Terabyte-Bereich mit file- oder objektbasierter Anbindung ebenso wie den Zugriff mittels Versionierungssystemen oder die Ablage in einem Repositorium. Einige

[3] Leitfaden der DFG für die Antragstellung: Projektanträge, `http://www.dfg.de/formulare/54_01/54_01_de.pdf`, S. 5

[4] Vgl. Grundsätze zum Umgang mit Forschungsdaten. Allianz der dt. Wissenschaftsorganisationen. `http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze` (2010)

[5] Ergebnisse einer qualitativen Befragung im Rahmen des bwFDM-Communities Projekts [Tr15].

[6] Vgl. „Überlegungen zu Steuerung und Governance von kooperativ betriebenen HPC-Infrastrukturen“, [vo16]

Communities können für ihre Datenerhebung auf bereits etablierte Forschungsinfrastrukturen zurückgreifen, benötigen jedoch Platz für eine langfristige Archivierung von Roh- oder Ergebnisdaten beispielsweise aus abgeschlossenen Promotionen. Im Bereich der Bioinformatik steigt der Bedarf allein durch immer höher auflösende Messinstrumente. Das Nutzungsprofil vieler Disziplinen lässt sich gut durch das Domain Model annähern. [7] Eine reine Speicherung von Forschungsdaten ist allerdings nicht ausreichend. Für ein effektives FDM-System werden zusätzliche Komponenten benötigt, wie in Abschnitt 3 erläutert.

Das Konzept für ein FDM bewegt sich im Spannungsfeld zwischen sehr vielfältigen und fachspezifischen Bedürfnissen der Wissenschafts-Communities einerseits und der praktischen Umsetzbarkeit einer technischen und organisatorischen Lösung andererseits. Um die Gefahr einer Insellösung zu vermeiden, ist auf einen möglichst standardisierten Ansatz zu achten und eine Umsetzung in föderativen Strukturen zu gewährleisten. [8] Ein implementiertes FDM einer Einrichtung muss im Einzelnen folgende Aspekte adressieren:
Forschungsdaten fallen durch die Vielfalt der Wissenschaftsdisziplinen in verschiedenen Formen an und sind vielfältiger Herkunft. Jede Disziplin hat eigene Vorstellungen ihrer beschreibenden Metadaten. Ein FDM deckt optimalerweise den gesamten Data-Lifecycle [Ba12] ab, von der Erhebung oder Erzeugung der Daten über die einzelnen Verarbeitungsschritte bis zu ihrer leichten Auffindbarkeit und langfristigen Bereitstellung für eine Nachnutzung durch Dritte.
Wissenschafts-Communities betreiben FDM auf unterschiedliche Weise [Kl13]. Viele Forschungsdaten laufen noch nicht in nationalen oder internationalen Datenzentren zusammen. Die Bedürfnisse unterscheiden sich durch die teilweise vor Ort oder weltweit verteilt schon vorhandenen Ressourcen, um Daten nach der Erzeugung oder Publikation geeignet abzulegen. Solche Strukturen sollen in Freiburg nicht dupliziert, sondern passend für die Forschenden vor Ort ergänzt werden und geeignete Schnittstellen für die Einbindung in übergeordnete Strukturen bieten.
Forschungseinrichtungen wie die Universität Freiburg müssen verbindliche Richtlinien erlassen und deren Umsetzung fördern. Hierzu zählen die Frage der Freigabe der Daten und Regelungen vergleichbar zu Publikationsverträgen mit der Universitätsbibliothek. Das Rechenzentrum der Universität hat sich aus diesen Gründen an Projekten wie bwFDM-Communities [Tr15] und „Landesweit koordinierte Strukturen für Nachweis und effiziente Nachnutzung von Forschungsdaten“ beteiligt, um schrittweise eine Policy für die Gesamteinrichtung zu entwickeln. Den Empfehlungen des RfII [Rf16] folgend liegt der Fokus auf der Einbindung in einrichtungsübergreifende Strukturen. Als Vorbilder [Er13] für eine konkrete Umsetzung dienen beispielsweise die Aktivitäten an der HU Berlin, der Universität Göttingen oder der RWTH Aachen [EMS].
Dienstanbieter müssen in die Lage versetzt werden, sowohl die kurz- und längerfristige Speicherung als auch den Nachweis der Forschungsdaten anzubieten. Da diese Aufgabe wegen der Vielfalt der Anforderungen und Disziplinen nicht von einer Einrichtung geleistet werden kann, sollte eine zuverlässige, verteilte Datenspeicherung in föderierten

[7] Daten entstehen oft durch Aktivitäten einzelner Forschender oder Gruppen (private Domäne), die in weiteren Schritten in Kooperationen geteilt (Gruppendomäne) und später je nach Art in die dauerhafte Domäne verschoben werden, wo eine Nachnutzung möglich wird (Zugriffsdomäne), vgl. [Kl13], S. 6

[8] Vgl. hierzu „Rahmenbedingungen einer disziplin-übergreifenden Forschungsdaten-Infrastruktur“, `http://www.forschungsdaten.org/index.php/Radieschen`

Verbünden erfolgen. [Kl13] [EU17]
Nachhaltige Finanzierung und Ausstattung des FDM sind wegen der erwartbaren erheblichen Zunahme der Datenmengen im Petabyte-Bereich pro Jahr eine nicht unerhebliche Herausforderung. Das FDM wird ebenso wie eine Bibliothek zu einer zentralen Infrastruktur (nicht nur am eigenen Standort) mit entsprechendem Finanzierungsbedarf [Rf16]. Da ein nicht unerheblicher Teil zukünftiger Kosten von der Datenmenge abhängt, sind Möglichkeiten zur Beteiligung der Nutzer insbesondere bei erheblichen Datenmengen vorzusehen. [9] Umso wichtiger werden geeignete Verfahren zur Qualifizierung der langfristig gespeicherten Daten, die ein ausgewogenes Verhältnis zwischen Quantität und Qualität der Forschungsdaten erzielen.

3 Vorschlag für eine FDM-Infrastruktur

Die für die Universität in der Diskussion befindliche FDM-Infrastruktur vereint eine Kombination aus Hard-, Software- und organisatorischen Bausteinen. Die bedarfsorientierte Realisierung orientiert sich gleichzeitig an den Lösungen von RADAR und am Service-Portfolio von EUDAT [10] sowie an aktuellen technologischen Umsetzungen der Storage-Hersteller für Massendaten. Das RZ koordiniert seine Aktivitäten mit Partnern in Baden-Württemberg. Für das Storage-System wird ein dreistufiges Konzept angestrebt, welches verschiedene Stadien des Data-Life-Cycle abdecken soll:

1. Layer I: Die oberste Ebene bietet verschiedene High-Level-Access-Varianten für Forschende an, über die sie direkt mit dem Storage-System interagieren können. Diese können lokale Filesystem-Caches an entfernten Standorten enthalten.
2. Layer II: Die mittlere Ebene kümmert sich um die primäre Datenhaltung für die aktuell im Zugriff befindlichen Daten.
3. Layer III: Die unterste Ebene übernimmt längerfristige Archivfunktionen. Hier werden Daten abgelegt, die nicht mehr aktiv bearbeitet werden und für bestimmte länger- und langfristige Zeiträume aufbewahrt werden sollen.

Die technische Umsetzung kann in einem hierarchischen Storage-Modell erfolgen, in dem die Daten je nach erwarteter Verfügbarkeit und Redundanzlevel abgelegt werden. Dieses wird durch einen Data-Mover orchestriert, der zusätzliche Informationen für seine Aktivitäten berücksichtigt, die aus den von den Forschenden gepflegten Metadaten gewonnen werden.

Layer I bietet verschiedene Dienste an, die von einem direkten Filesystem-Zugriff, über Versionierungsdienste wie beispielsweise GIT bis hin zum Repository- oder Object Store Zugriff reichen. Die Forschenden erhalten nach Beantragung Zugriff auf den Speicherplatz für einen gewissen Zeitraum. Für jedes Vorhaben wird hierzu ein (virtueller) Datencontainer im FDM-Storage-System erzeugt. Eine längerfristige Datenhaltung wird von

[9] Vgl. Vorteile der Nutzer in der Beteiligung und bestehenden Forschungsinfrastrukturen, Aufwuchsfinanzierung und andere Beteiligungsmodelle in [vo16].
[10] Vgl. `https://www.radar-projekt.org` bzw. in [EU17] Services und Support

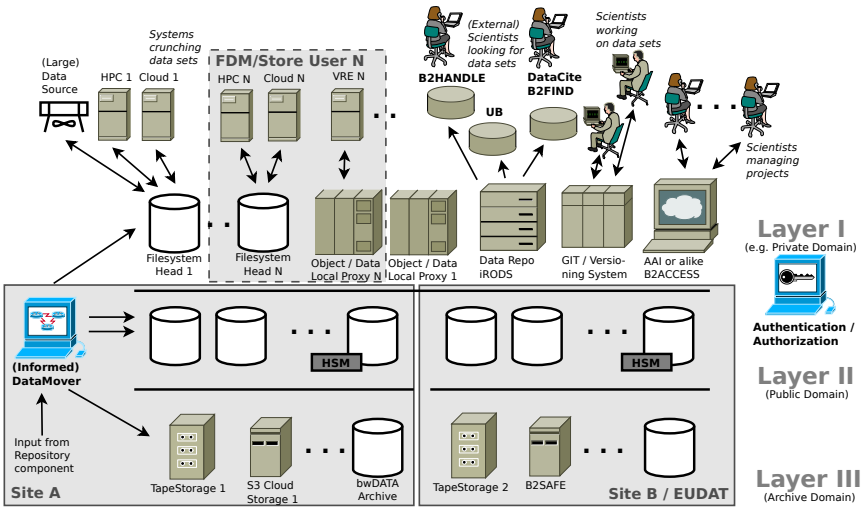

Abb. 1: Überlegungen zur FDM-Infrastruktur und Schnittstellen zu übergeordneten Systemen

der Qualifizierung der Daten und der Scientific Governance abhängig gemacht. Die technische Zugriffsebene besteht sowohl aus Hardwarekomponenten des FDM-Systems als auch aus einer Softwareschicht, die längerfristig angelegt ist und abstrakt auf der Hardwareschicht aufsetzt. Hierzu zählen die Bereitstellung von Versionierungsdiensten ebenso wie Repository- oder Object Store Komponenten. Hinzu kommt das Benutzerinterface (Beantragung, Metadaten) und Schnittstellen zu Management- sowie Authentifizierungskomponenten. Bausteine wie Versionierungssysteme oder Repositorien existieren bereits am Markt oder sind bereits in den beteiligten Einrichtungen im Einsatz. Sie sind jedoch geeignet in das Gesamtsystem zu integrieren.

Layer II – die mittlere Schicht – bildet die zentrale Komponente des FDM-Systems. Sie wird hauptsächlich durch die Hardwarekonzepte des zu beauftragenden Anbieters bestimmt. Sie bringt einen Data-Mover mit, der nach entsprechenden Kriterien die Ablage der Daten-Container steuert. Dieser könnte unterschiedliche Zahlen von Kopien ebenso wie die Lokalität der Daten berücksichtigen. Diese Schicht kümmert sich um die Integrität der Daten durch verschiedene technische Umsetzungen, wie beispielsweise Erasure Coding, RAID-Verbünde oder Kopien auf verschiedenen Ebenen des FDM-Systems. Ebenso könnte eine geografische Redundanz mit dem Partnerstandort in Tübingen in Betracht gezogen werden. Diese Elemente sind von verschiedenen Herstellern verfügbar, da eine breite Palette an Hierarchical Storage Management Systemen angeboten wird. Die Basis vieler solcher Systeme sind spezielle Filesysteme wie GPFS oder BeeGFS. Entsprechende Filesystem-Köpfe für SMB oder NFS werden üblicherweise angeboten.

Layer III implementiert die langfristige Datenpublikations- und Archivebene. Sie könnte in unterschiedlicher Form umgesetzt werden. So sollte mindestens eine S3-Schnittstelle für eine potenzielle Nutzung von internen und externen Cloud-Services bereitgestellt werden. Durch die Repository-Komponente des Gesamtsystems sollte ein durchgängig transparenter Zugriff (durchgängige Referenzierung) auf die Daten unabhängig vom tatsäch-

lichen physikalischen Lagerort sichergestellt werden. Hierbei sollte das jeweilige Nachweis- und Zugriffssystem damit umgehen können, dass die Daten unter Umständen erst mit Zeitverzögerung bereitgestellt werden.

Der Übergang von Daten in die Archivschicht sollte durch einen formalen Vorgang begleitet werden, an dem sowohl der Eigentümer die Daten abschließend durchsieht als auch ein Storage-Gremium der jeweiligen Community die Archivwürdigkeit bestätigt. Spätestens mit der Übernahme in den Layer III sind Daten mit geeigneten Persistent Identifiern [11] zu versehen und können damit permanent referenziert werden. Die Referenzierung könnte beispielsweise aus den Systemen der jeweiligen Universitätsbibliotheken heraus erfolgen. Ebenso wäre eine Einbindung in die DataCite-Infrastruktur [BD11] denkbar.

Data-Mover sind zentrale Bestandteile von modernen hierarchischen Storage-Systemen. Sie sorgen für eine geeignete Verteilung und Redundanz der Daten über das Gesamtsystem hinweg. Diese Komponente sollte dahingehend erweiterbar sein, dass sie zusätzliche Kriterien, die beispielsweise den Metadaten der Datensätze entnommen werden, in ihren Entscheidungen berücksichtigen kann. Ziel sollte es dabei sein, dem System Informationen in ausreichendem Umfang und Qualität bereit zu stellen, dass die geeignete Platzierung der Daten weiterhin weitgehend automatisch erfolgen kann. Manuelle Eingriffe sollten lediglich über die entsprechende Anpassung der Metadaten durch Nutzer oder Storage-Gremien laufen. Der Data-Mover muss Zugriff auf alle Layer des FDM-Storage-Stacks haben. Sollte die Archivebene nicht direkt Bestandteil des Systems sein, sollte er diese zumindest geeignet ansprechen können.

4 Mehrwert durch Qualifizierung der Daten

Die Speicherplatzbedürfnisse der einzelnen Forschungsprojekte können sehr verschieden ausfallen, weshalb einerseits unterschiedliche Arten des Zugriffs auf die FDM-Ressource als auch andererseits verschiedene Zustände der Datenhaltung [TGHR07] angenommen werden. Im Rahmen dieses Vorhabens wird grob von drei Schritten in der Datenhaltung ausgegangen (Abbildung 2). Der Einstieg (Schritt 1) erfolgt über die formale Anmeldung eines Forschungsvorhabens. Die Einstiegshürden hierfür werden bewusst niedrig gesetzt. An dieser Stelle werden eine Reihe grundsätzlicher Metadaten zum Projekt erhoben. Der beantragte Speicherplatz wird als (virtueller) Storage-Container mit den gewünschten Zugriffsmethoden für einen begrenzten Zeitraum [12] bereitgestellt. Dieses Verfahren stellt sicher, dass regelmäßig neuen Forschergruppen Speicher zugeteilt werden kann. Die Storage-Container werden mit den bereitgestellten wissenschaftlichen Metadaten und den vom System generierten technischen Metadaten verknüpft. Für den geforderten Workflow kann auf existierende Implementierungen im Land zurückgegriffen werden, die bereits seit einiger Zeit im bwHPC-Projekt produktiv im Einsatz sind. Die dort eingesetzte Plattform könnte für die Zwecke angepasst und erweitert werden, da sie viele notwendige Komponenten wie beispielsweise die Authentifizierung über bwIDM

[11] Vgl. [EU17], B2Handle

[12] Dieser Zeitraum könnte beispielsweise 100 Tage betragen und mit geringen formalen Hürden nochmal um diese Zeitspanne verlängert werden. Dieses wird im Rahmen der Governance vereinbart.

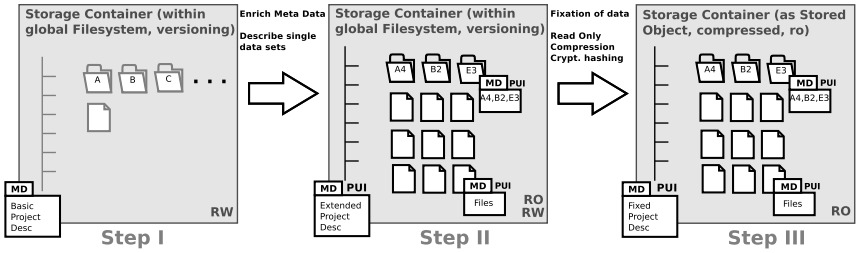

Abb. 2: Anreicherung von Metadaten, wenn Daten länger im System vorgehalten werden sollen (Schritt 1 zu 2) sowie Übergang zur Langzeiterhaltung (Schritt 2 zu 3)

bereits mitbringt. In diesem Stadium werden Daten als vorübergehender Arbeitsbereich („Scratch“) betrachtet. Auf diese Weise können Forschende mit wenig Aufwand Speicher einrichten. Dieses ist insbesondere für die erste Datenaufnahme und Verarbeitung relevant.

Der Data-Mover wird den Benutzer über den Ablauf der Haltefrist rechtzeitig informieren, so dass dieser die Chance hat, die Daten entweder geeignet zu klassifizieren oder selbst zu sichern. Ein Nutzer kann für einen bestimmten Zeitraum die Haltefrist der Daten durch einfachen Antrag verlängern, allerdings nicht beliebig oft. Dadurch wird sichergestellt, dass das System nicht mit Scratch-Daten überläuft, die keinen langfristigen Wert über den konkreten Arbeitsschritt des Nutzers hinaus besitzen.

Zu dem Augenblick, an dem Daten eine bestimmte Qualität erreicht haben und für einen längeren Zeitraum vorgehalten werden sollen, können sie für den längerfristigen Verbleib im FDM-System qualifiziert werden (Schritt 2). Dazu kann der Eigentümer sowohl weitere Projekt-Metadaten ergänzen als auch seine Daten mit zusätzlichen fachspezifischen Metadaten (als auch um Tags und Persistent Unique Identifiers) anreichern, so dass sie sich potenziell (öffentlich) referenzieren und zitieren lassen. Die Forschenden sollten im Sinne von Open Access verpflichtet werden, ihre Daten nach einer gewissen Zeit – welche vom Projekt, den Zuwendungsgebern oder der Community-Policy abhängen können – bereitzustellen. Dieses kann auch im Rahmen einer Publikation erfolgen. Auf diese Weise sollen die Forschenden ermutigt werden, eine hohe Qualität sowohl der Daten als auch der Metadaten sicherzustellen. Dieser Vorgang könnte beispielsweise Komponenten [13] analog zu einem „Publikationsvertrag“ für die Veröffentlichung eines Werkes über eine Universitätsbibliothek beinhalten.

Mit Erreichung bestimmter Projektziele beziehungsweise nach dem Projektabschluss ist eine längerfristige Datenhaltung vorzusehen (Schritt 3), so diese nicht in Community-eigenen Repositorien langfristig abgelegt werden. Eine geforderte, langfristige Datenhaltung könnten beispielsweise die von der DFG geforderten zehn Jahre für die Überprüfbarkeit von Forschungsergebnissen sein. Ebenso könnten die Daten wiederum für weitere Vorhaben (auch für externer Nutzer) von Interesse sein. Je nach Art und Häufigkeit des Zugriffs auf die Daten könnte der Data-Mover unterschiedliche Speichersysteme im Hintergrund vorsehen.

[13] Abtretung langfristiger Rechte, automatischer Rechteübergang nach einer bestimmten Zeit o.ä.

Jederzeit bei Bedarf, spätestens jedoch im Übergang von Schritt 2 zu 3 - dem Übergang in die dauerhafte Domäne - werden die Daten in der vorliegenden Form „eingefroren“ (Sicherstellung der Unveränderlichkeit). Sie werden dann üblicherweise aus dem System für den laufenden Zugriff in ein Object Store überführt. Sollte für eine spätere Anbindung und Nachnutzung der Daten ein Dateisystemzugriff erforderlich sein, ließe sich dieses über geeignete technische Maßnahmen [14] abbilden.

Weiterhin können beim Übergang in die dauerhafte Domäne automatische Prozesse im Schritt von Layer 2 zu 3 ausgelöst werden, die von Kompression über die Bildung von Prüfsummen bis hin zur Verschlüsselung der Daten reichen. Auf diese Weise sollen sowohl effiziente Datenhaltung als auch Integrität und Unveränderbarkeit der Datensätze sichergestellt werden. Die Daten bleiben weiterhin für die Nutzung verfügbar, werden jedoch dann als abgeleitete Versionen genutzt. [15] Hierdurch wird die notwendige Persistenz der Daten für ihre Zitier- und Referenzierbarkeit sichergestellt. In dieser Stufe wird der Data-Mover keine Daten automatisch löschen, sondern anhand der Metadaten (bspw. Hinweis auf eine Publikation), Zugriffsrechten und Zugriffshäufigkeit festlegen, wo die Daten geeignet lagern und wieviele (geo-)redundante Kopien erzeugt werden sollen.

Eine automatische Löschung könnte beispielsweise für Daten vorgesehen werden, deren Haltefrist abgelaufen ist und für die keine Zugriffe über einen bestimmten Zeitraum registriert wurden. Diese Entscheidung könnte an ein entsprechendes Governance-Gremium delegiert werden. Ebenso wären Fragen der Finanzierung langfristiger Datenhaltung mit den Heimat-Institutionen und einzelnen Communities abzustimmen. Unabhängig von der aktuellen Haltung der Daten sollten diese transparent mindestens ab Schritt 2 referenziert sein und beispielsweise via OAI-PMH oder entsprechender anderer geeigneter Systeme und Standards auffindbar sein. Die Referenzierbarkeit der Daten sollte unabhängig von der konkreten Ablage (auch bei externen Diensten wie bwDATA-Archiv, EUDAT) erhalten bleiben und je nach Festlegung durch das FDM-System gesteuert werden.

Qualifizierende Metadaten an Storage-Container und später den Datensätzen dienen der Steuerung des FDM-Gesamtsystems, der Wiederauffindbarkeit und einer guten Nachnutzbarkeit. Metadaten geben üblicherweise die jeweiligen Fach-Communities oder auch die Standards der Bibliotheken und (Daten-)Archive vor. Für die Zwecke des FDM-Systems kommen einige technische Metadaten hinzu, die entweder automatisch ermittelt oder vom Nutzer abgefragt werden. Hierzu zählen (insbesondere für Schritt 1): *Projektbeschreibung* die sich an etablierten Standards der Communities und Fördergebern wie der DFG orientiert und die Zuordnung von Forschungsprojekten erleichtern; *Eigentümer des Projekts*, üblicherweise Leitung einer Arbeitsgruppe oder eines Forschungsprojekts. Dies soll sicherstellen, dass Daten auch nach Weggang von Arbeitsgruppenmitglieder zugreifbar bleiben. Diese sind zudem Ansprechpartner für das Steuerungsgremium der betroffenen Wissenschafts-Community; *Erwartete Laufzeit des Projekts und erwarteter Umfang der Daten*, dient einerseits der Steuerung und andererseits der Abschätzung zukünftigen Speicherbedarfs im Zeitablauf; *Bevorzugter Typ der Nutzung des Speicher-Containers* (File-

[14] An dieser Stelle werden häufig sog. Fuse-Layer eingesetzt.
[15] Dieses wird technisch beispielsweise über Copy-on-Write Mechanismen umgesetzt.

system, Object Store, Versioning, ...) wird von der Workflow-Engine genutzt, um nach formaler Freigabe die notwendigen Ressourcen und Schnittstellen bereitzustellen.

Die spätere Qualifizierung der Daten in Schritt 2 wird insbesondere durch die Standards und Vereinbarungen der jeweiligen Communities bestimmt. Hier sollte das FDM-System in der Lage sein, verschiedene Metadaten-Standards [16] zu unterstützen.

Governance Zwischen den verschiedenen Wissenschafts-Communities, den Betreibern und den Zuwendungsgebern muss ein sinnvoller Ausgleich der Interessen und Kosten organisiert werden. [17] Storage-Kapazität unterscheidet sich insofern fundamental von Compute-Kapazität, dass nach Abschluss eines Projekts die Ressource nicht wieder automatisch frei wird und an weitere Nutzer weitergegeben werden kann. Es muss sichergestellt werden, dass auch später hinzukommende Forschungsgruppen das System für ihre Zwecke nutzen können. Während es für kurze Zeiträume kein Problem darstellt, auch größere Datenmengen vorzuhalten, wird dieses für lange Fristen kostenintensiv. Um einen guten Kompromiss aus Menge und Qualität der Daten zu erhalten, sind verschiedene Herangehensweisen denkbar. Eine Variante setzt auf einen höheren Grad des Selbstmanagements der Nutzer-Community vor dem Hintergrund vorliegender Forschungsdatenmanagementkonzepte. [18]

Der erste Schritt der Datenhaltung setzt weitgehend auf Self-Service und orientiert sich an den eingeführten Prozessen vergleichbarer Verbundprojekte. [19] Mit der Verfügbarkeit einer gewissen Basisberechtigung soll es einfach möglich sein, für zunächst begrenzte Zeit Speicherplatz für ein Forschungsprojekt zu beantragen. Um eine effiziente Nutzung insbesondere bei längerfristiger Belegung der Ressource zu erreichen, werden Schwellen eingebaut, die schrittweise eine hohe Qualität der Daten sicherstellen sollen. Diese beinhalten automatisierbare Anteile (Nutzung der Informationen aus den Metadaten) und Review-Prozesse durch (gewählte) Gremien, die beispielsweise fest in der akademischen Selbstverwaltung der einzelnen Fakultäten angesiedelt sein könnten. Diese könnten so Einfluss nehmen, für welche langfristige Datenhaltung Fakultätsbeiträge für ein einrichtungsweites FDM eingesetzt würden.

5 Vorläufiges Fazit

Der zunehmende Bedarf, Forschungsdaten zu speichern, zu qualifizieren und in wissenschaftliche Arbeitsabläufe einzubinden, muss adressiert werden. Dieser sollte nicht mehr wie bisher durch einzelne dezentrale Storage-Lösungen ohne wirkliche Langfristperspektive und nachhaltige Nutzung beziehungsweise Austausch der Daten bedient werden. So würden lediglich vorhandene Lösungen und Infrastrukturen mit einem hohen Kosten- und Management-Aufwand dupliziert. Die Diskussions- und Abstimmungsprozesse zwischen Universitätsleitung, zentralen Einrichtungen und den Wissenschafts-Communities laufen.

[16] Vgl. beispielsweise [EU17], Dienste des B2FIND
[17] Vgl. hierzu Darstellungen in [vo16], S. 281ff., S. 344ff.
[18] Vgl. Stand in Baden-Württemberg, `www.forschungsdaten.info`
[19] Vgl. Darstellungen in [vo16] bzw. „Zentrale Antragsseite“ (ZAS), `https://www.bwhpc-c5.de/ZAS`

Weitgehender Konsens besteht, dass ein (abstrakt) übergreifendes FDM-System angestrebt wird, welches sich in Verbundstrukturen einordnet und Schnittstellen und Austausch mit europäischen Lösungen sucht. Das System soll mit einer Kernkomponente, welche die wichtigen Dienste bietet, starten. Es soll durch regelmäßige Erneuerungen unter Beteiligung der Communities ausgebaut werden. Um sowohl eine hohe Qualität der Daten als auch eine effiziente Nutzung des Systems zu erreichen, werden die Wissenschafts-Communities in den Fakultäten und Instituten in den Aufbau der Workflows zur Qualifizierung der Daten und den Aufbau der Governance-Strukturen eingebunden. Das System wird gleichfalls versuchen, durch einen hohen Selfservice-Anteil den Forschenden weitgehende Freiheiten beispielsweise bei der Wahl der Metadaten und der Qualitätskriterien einzuräumen. Eine frühestmögliche Beteiligung der Communities ist vor allem im Hinblick auf die Akzeptanz und des Nutzen des FDM-Systems notwendig. Gleichzeitig müssen die Hürden für die Nutzung des Systems so gering wie möglich gehalten werden.

Literaturverzeichnis

[Ba12] Ball, Alexander: Review of Data Management Lifecycle Models. February 2012.

[BD11] Ball, Alex; Duke, Monica: How to cite datasets and link to publications. Digital Curation Centre, 2011.

[BHM11] Büttner, Stephan; Hobohm, Hans-Christoph; Müller, Lars: Handbuch Forschungsdatenmanagement. Bock+ Herchen, 2011.

[EMS] Eifert, Thomas; Muckel, Stephan; Schmitz, Dominik: Introducing Research Data Management as a Service Suite at RWTH Aachen. In: Ges. für Informatik eV (GI). S. 55.

[Er13] Erway, Ricky: Starting the Conversation: University-Wide Research Data Management Policy. ERIC, 2013.

[EU17] EUDAT Collaborative Data Infrastructure, 2017. `https://www.eudat.eu/`.

[Kl13] Klar, Jochen; Enke, Harry: Report „Organisation und Struktur“, 2013.

[Ne12] Neuroth, Heike; Strathmann, Stefan; Oßwald, Achim; Klump, Jens; Ludwig, Jens: Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. 2012.

[Rf16] RfII – Rat für Informationsinfrastrukturen: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, 2016. `http://www.rfii.de/?wpdmdl=2075`.

[TGHR07] Treloar, Andrew; Groenewegen, David; Harboe-Ree, Cathrine: The data curation continuum: Managing data objects in institutional repositories. D-Lib magazine, 13(9):4, 2007.

[Tr15] Tristram, Frank; Bamberger, Peter; Cayoglu, Ugur; Hertzer, Jörg; Knopp, Johannes; Kratzke, Jonas; Rex, Jessica; Schwabe, Fabian; Shcherbakov, Denis; Svoboda, Dieta-Frauke; Wehrle, Dennis: Öffentlicher Abschlussbericht von bwFDM-Communities, 2015. `http://bwfdm.scc.kit.edu/downloads/Abschlussbericht.pdf`.

[vo16] von Suchodoletz, Dirk; Schulz, Janne; Leendertse, Jan; Hotzel, Hartmut; Wimmer, Martin: Kooperation von Rechenzentren Governance und Steuerung – Organisation, Rechtsgrundlagen, Politik. de Gruyter, 2016.