# Ein standortübergreifendes Speichersystem für Forschungsdaten

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

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

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

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

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

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

## 1 Einleitung

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


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

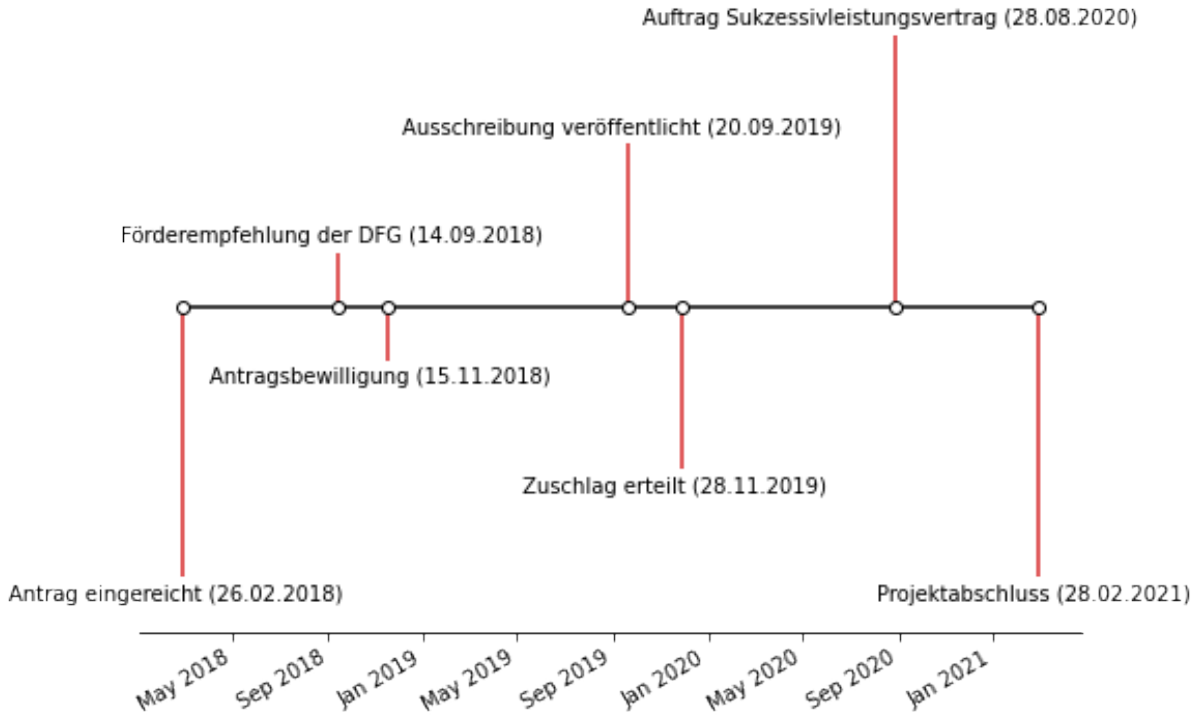


Abbildung 1: Zeitstrahl zum Projektverlauf.

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

Abbildung 1 zeigt den groben zeitlichen Verlauf des Projekts.

## 2 Ziele

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

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

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

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

# 3 Speichersystem

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

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

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

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

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

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

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

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

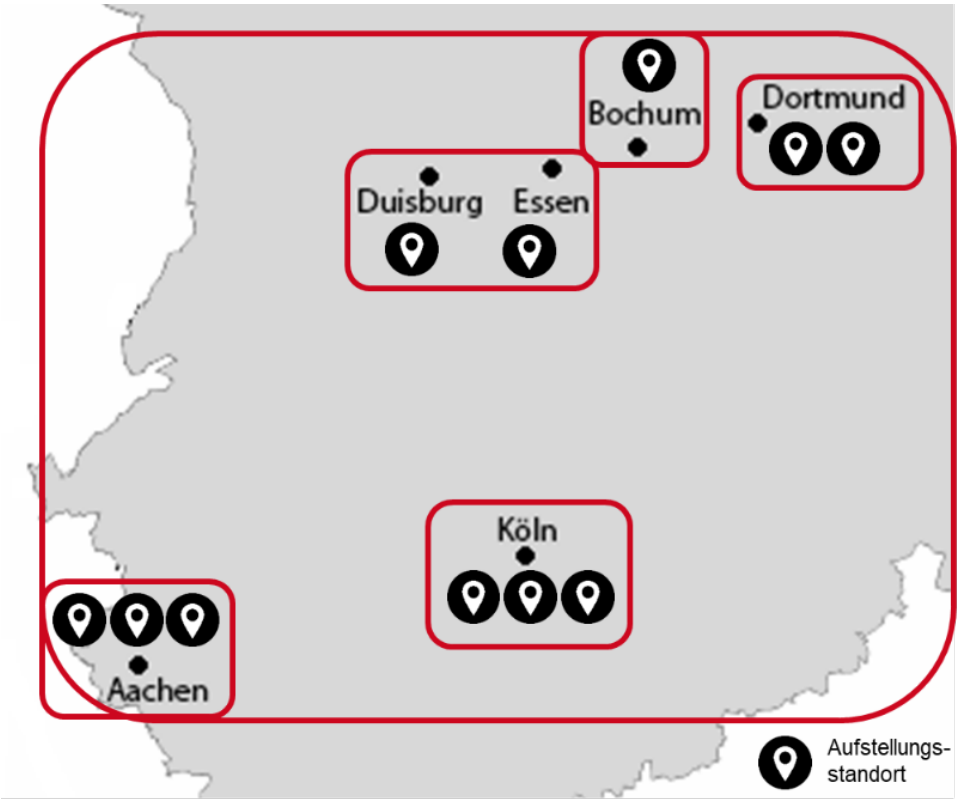


Abbildung 2: Aufstellungsstandorte des Speichersystems.

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

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

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

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

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

# 4 FDM-Konzept

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

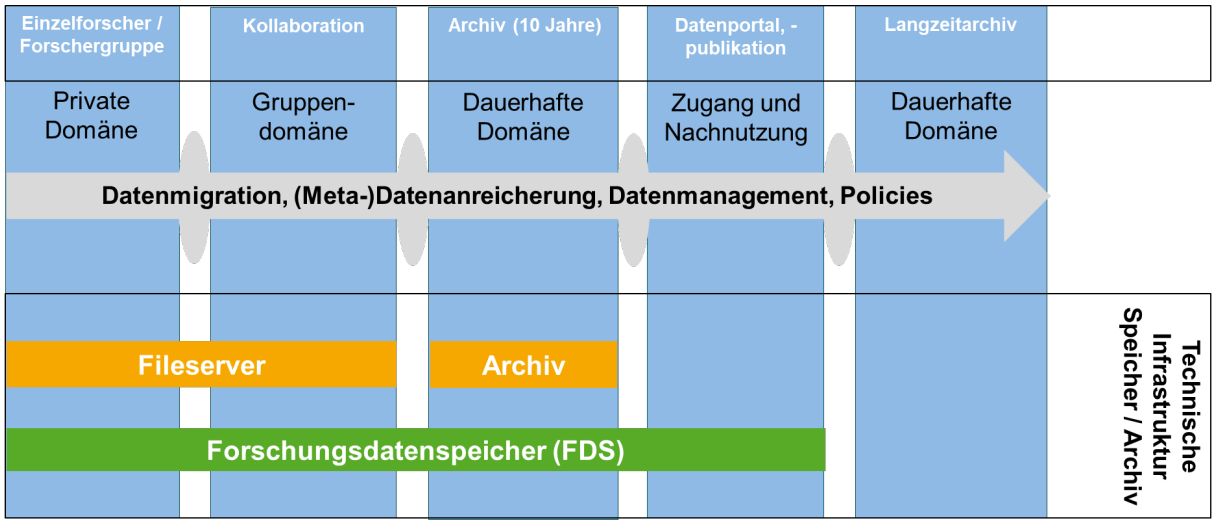


Abbildung 3: Verortung des FDS im Domänenmodell.

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

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

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

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

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

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

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

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

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

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

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

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

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

# 5 Offenheit des Konsortiums

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

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

Zudem besteht die Möglichkeit für weitere Hochschulen, sich mit eigener Hardware an dem Speichersystem zu beteiligen. Allerdings erfordert dies die Nutzung der gleichen Hardware wie sie bereits verwendet wird, konkret DELL ECS- und Isilon-Systeme. Die lokale Hardware kann dann in bestehende oder neu zu definierende Replikationsgruppen eingebunden werden, so dass die Daten georedundant gesichert werden.

# 6 Fazit

Das Speichersystem Forschungsdatenspeicher (FDS) stellt einen wichtigen Baustein bei der Entwicklung hin zu einem bewussten und nachhaltigen Umgang mit Forschungsdaten dar. Es wirkt transformierend auf die Art und Weise, wie Speicherplatz für Forschungsdaten zur Verfügung gestellt wird und bietet einen Kristallisationspunkt für die Entwicklung von Anwendungslandschaften für die Forschung.

Objektorientierter Speicher ist eine technologische Grundlage, die für die Ablage von Forschungsdaten hervorragend geeignet ist. Sie reduziert gegenüber klassischem Blockspeicher den Overhead an technischen Metadaten deutlich und ermöglicht so nahezu unbegrenzte Skalierung bei geringen Kosten. Durch die Isilon bietet des FDS aber auch die Möglichkeit, für bestimmte Anwendungsfälle Blockspeicher über die klassischen Fileserver-Protokolle zur Verfügung zu stellen.

Das gesamte Projekt über die Beantragung, Beschaffung, Installation bis hin zum Betrieb des Systems hat den Beteiligten großen Einsatz abverlangt und eine intensive Zusammenarbeit nötig gemacht. Gerade diese intensive Zusammenarbeit und der enge Austausch haben sich jedoch als Wert an sich erwiesen. Das stark zusammengewachsene Team konnte sehr viel technische Expertise aufbauen und durch den intensiven Austausch auch die FDM-Konzepte an den Standorten weiterentwickeln.

## Acknowledgements

Research Data Storage (RDS) wurde unter der Fördernummer 124-4.06.05.08-139057 vom Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen finanziert.

## Literaturverzeichnis

[1] Eifert, T., Claus, F., and A. Lopez. “Research Data Storage (RDS): Verteilte Speicherinfrastruktur für Forschungsdatenmanagement: Gemeinsamer Antrag (öffentliche Fassung) im DFG-Programm ”Großgeräte der Länder”: RWTH Aachen University (Konsortialführer), Fachhochschule Aachen, Ruhr-Universität Bochum, Technische Universität Dortmund, Universität Duisburg-Essen, Universität zu Köln” Veröffentlicht auf dem Publikationsserver der RWTH Aachen University, (2018): doi: https://doi.org/10.18154/RWTH-2021-04541.

[2] Politze, M., Claus, F., Brenger, B., Yazdi, M. A., Heinrichs, B., and A. Schwarz. “How to Manage IT Resources in Research Projects? Towards a Collaborative Scientific Integration Environment”. European Journal of Higher Education IT 2020-1. Paris, France.