WissGrid-Spezifikation: Grid-Repository

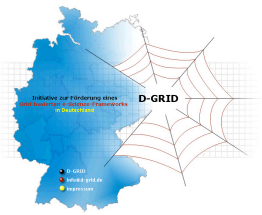
Version - 30. April 2010

Arbeitspaket 3

Verantwortlicher Partner - SUB

WissGrid

Grid für die Wissenschaft



Bundesministerium
für Bildung
und Forschung

Projekt: **WissGrid**

Teil des D-Grid Verbundes und der deutschen e-Science Initiative

BMBF Förderkennzeichen: 01|G09005A-G (Verbundprojekt)

Laufzeit: Mai 2009 - April 2012

Dokumentstatus: stabile Zwischenversion

Verfügbarkeit: öffentlich

Autoren:

Andreas Aschenbrenner, SUB

Harry Enke, AIP

Bernadette Fritzsch, AWI

Jens Ludwig, SUB

Torsten Rathmann, DKRZ

Angelika Reiser, TUM

Florian Schintke, ZIB

Frank Schluenzen, DESY

Jessica Smejkal, TUM

Stefan Strathmann, SUB

Revisionsverlauf

Datum	**Autor**	**Kommentare**
10.11.2009	AAsche, Jens Ludwig	Dokumentstruktur
25.11.2009	AAsche	erste Skizze
2.-14.12.2009	Jens Ludwig, Stefan Strathmann, Harry Enke, Angelika Reiser, Jessica Smejkal, AAsche	kleinere Korrekturen, Kommentare
9.12.2009	Bernadette Fritzsch	weiteres Profil C
15.12.2009	Florian Schintke	Ergänzungen, Korrekturen und Einbettung iRODS
17.12.2009	Bernadette Fritzsch	sprachliche Korrekturen, Ergänzung Fallbeispiel Klima
10.1.2010	AAsche	Überarbeitung, 1. Skizze zur tech. Umsetzung
11.1.2010	Jens Ludwig	Glossar, Formatierung
14.1.2010	Bernadette Fritzsch, Jens Ludwig, Angelika Reiser, Jessica Smejkal,	Einarbeitung der Konsortiumskommentare und Ergänzungen
8.4.2010	Bernadette Fritzsch	Überarbeitung Profil C
9.4.2010	AAsche	Profil A+B: Arbeitsplan technische Umsetzung
30.4.2010	Torsten Rathmann, Frank Schluenzen, Jens Ludwig	Korrekturen, Formatierungen

Inhaltsverzeichnis

0 Allgemeines Vorwort zu WissGrid-LZA-Spezifikationen

Die vorliegende Spezifikation für Grid-Repositorien ist Teil der Gesamtstrategie für Langzeitarchivierung (LZA) in D-Grid, entwickelt durch das Projekt WissGrid. Diese Spezifikation ist eine Komponente der WissGrid-LZA-Architektur und Teil eines wachsenden Pools an Angeboten zur LZA in WissGrid/D-Grid.

Jenseits der Einbettung in die LZA-Aktivitäten des WissGrid-AP3s ist diese Spezifikation darüber hinaus auch geprägt von den Konzepten der anderen WissGrid-APs, von den Entwicklungen von D-Grid Partnern sowie von verwandten Projekten. Alle Abhängigkeiten und Grundannahmen dieser Spezifikation sind in der WissGrid-LZA-Architektur ausführlich beschrieben; einige Punkte daraus sind im Folgenden hervorgehoben:

- Nachnutzung von existierenden Konzepten und Tools: Die Fachdisziplin "LZA" (z.B. Digital Curation Conference, iPRES) ist eine sehr aktive und verzweigte Fachdisziplin, ähnlich wie die Fachdisziplin "Grid" (z.B. Open Grid Forum). LZA im Grid ist nicht nur ein Versuch, die Erfahrungen zweier großer Fachdisziplinen zu verknüpfen. Vielmehr ist für die Nachhaltigkeit der LZA-Dienste und Repositorien in WissGrid ein langfristiger Austausch zwischen diesen Fachdisziplinen essenziell.
- Content Preservation ist der primäre Schwerpunkt des WissGrid-AP3s: Die darunter liegende Bit-Preservation wird idealerweise durch die (Grid-)Infrastruktur angeboten; darüber liegende Funktionen der Data Curation (z.B. inhaltliche Selektion der Daten, Beschreibung und wissenschaftliche Verknüpfung) können durch LZA-Dienste und Repositorien unterstützt werden, sind aber primär Aufgaben des jeweiligen Community Grids.
- Modulare Anpassbarkeit an den jeweiligen organisatorischen und technischen Kontext: Auch wenn die WissGrid-LZA-Architektur primär auf Content Preservation ausgerichtet ist, so werden die in AP3 erzeugten Dienste und Repository-Varianten anpassbar an Community-spezifische Anwendungen und Community-spezifische LZA-Strategien sein. In welcher Form die Fachberater und Blaupausen von WissGrid AP2 in Zukunft die Communities in der Entwicklung von LZA-Strategien und Repositorien unterstützen können, wird im weiteren Projektverlauf ausgearbeitet.
- Möglichkeit unterschiedlicher organisatorische Modelle für den Betrieb: Im Fall von Repositorien könnte man, beispielsweise, auf der einen Seite Softwarepakete unterscheiden, die von den Communities im Stile einer "Referenzarchitektur"

individuell installiert und angepasst werden können; auf der anderen Seite sind aber auch zentral gehostete generische Archive denkbar, die zwar Daten vertrauenswürdig aufnehmen können, aber ggf. nicht für die spezifischen Nutzungsszenarien einer Community optimiert sind.

In diesem Spezifikationsdokument werden nur die technischen Voraussetzungen für unterschiedliche Modelle beschrieben. Die entsprechenden Geschäftsmodelle und Einbettung in D-Grid werden im weiteren Verlauf von WissGrid AP1 ausgearbeitet.

- Ständige Weiterentwicklung der hier beschriebenen technischen Basis: Eine kontinuierliche Weiterentwicklung ist gerade im Bereich der LZA notwendig (die dem schnellen technischen Fortschritt innerhalb nur weniger Jahre voll ausgesetzt ist), aber auch um mit den sich ständig weiter entwickelnden Anwendungen und wissenschaftlichen Methoden der Nutzer mithalten zu können (wo Erfahrungen und Kritik an laufenden Implementierungen die Nutzererwartungen ständig antreiben). In diesem Sinne ist diese Spezifikation nur ein Schnappschuss in der fortlaufenden Evolution von LZA-Diensten bzw. Repositorien.

1 Einleitung

In allen wissenschaftlichen Fach-Communities spielen Daten eine wesentliche Rolle; die vertrauenswürdige Archivierung und Verfügbarkeit dieser Daten zur Verarbeitung ist eine der Grundvoraussetzungen des wissenschaftlichen Diskurses. Während die Verwaltung von binären Daten in Datenbanken und auch die Virtualisierung von Datenbanken in Grid-Umgebungen durch Technologien wie OGSA-DAI [1] weit fortgeschritten ist, ist die Verwaltung von digitalen Objekten und die Verknüpfung zwischen Objekten und Datenbanken vergleichsweise vernachlässigt worden. Dabei ist gerade die Verwaltung von dateibasierten Objekten eine weit verbreitete, wenn nicht gar universelle Herausforderung für Grid-Communities.[2]

Binäre Daten: In diesem Dokument werden strukturierte Daten, die üblicherweise in relationalen Datenbanken verwaltet werden, als "binäre Daten" bezeichnet.

Digitale Objekte: Digitale Objekte sind digitale Daten, die als intellektuelle Einheiten aus (einer oder mehreren) Dateien, zugehörigen Metadaten, sowie einem Netzwerk aus anderen Objekten bzw. referenzierbaren Informationen bestehen können. Objekte können alle Arten von Daten umfassen - strukturiert, semi-strukturiert (z.B. XML-basiert), oder unstrukturierte Daten wie z.B. Bilder oder Videos. Um sie explizit zu beschreiben, können so genannte Paketformate benutzt werden, die zugehörige Metadaten (z.B. deskriptiv, administrativ, Audit Trails) wie auch Relationen zu anderen Objekten und externen Erschließungsmaterialien enthalten.

Repositorien dienen primär der Langzeitarchivierung, der gemeinsamen Datenhaltung sowie dem Austausch und der kollaborativen Nutzung von digitalen Objekten innerhalb einer Community. Sie werden erst seit etwa Mitte der neunziger Jahre als eigenständiger Forschungszweig angesehen und finden sich in den unterschiedlichsten Communities und Anwendungsprofilen.[3] Es gibt daher keine universelle Definition und zeitlose Standards, auf die man zurückgreifen kann.

---

[1] OGSA-DAI, Open Grid Services Architecture - Data Access and Integration. http://www.ogsadai.org.uk/

[2] Philip Lord, Alison MacDonald. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. JISC e-Science Curation Report, November 2003. http://www.jisc.ac.uk/publications/documents/esciencefinalreport.aspx.

[3] Andreas Aschenbrenner, Max Kaiser. White Paper on Digital Repositorys. reUSE! Deliverable, March 2005. http://www2.uibk.ac.at/reuse/docs/reuse-d11_whitepaper_10.pdf.

Kernfunktionen von Repositorien beschreiben Heery und Anderson[4] als die technisch robuste sowie organisatorisch nachhaltige und vertrauenswürdige Verwaltung von (datei-basierten) Daten und zugehörigen Metadaten sowie die organisatorische und technische Einbettung der Schnittstellen für Ablage und Zugriff. In dieser Definition der Kernfunktionen wird das Zusammenspiel aus Technik und organisatorischen Maßnahmen deutlich. Wir unterscheiden daher in diesem Dokument zwischen Repositorien (der Technologie) und Forschungsdaten-archiven (Technologie und Organisation).

Repository: Softwaresystem zur Verwaltung von digitalen Objekten. Neben der Verwaltung von digitalen Objekten (speichern, abrufen, verändern bzw. neue Versionen anlegen) bieten Repositorien auch zumeist generische Mechanismen zur Einbettung der Objekte in wissenschaftliche und interaktive Workflows, z.B. für die kollaborative Bearbeitung von Objekten in interaktiven Editoren oder für automatisierte wissenschaftliche Berechnungen.

Forschungsdatenarchiv (bzw. synonym "Forschungsarchiv"): Ein Forschungsarchiv umfasst Technik und Organisation (z.B. Betrieb, Finanzierung, Verantwortlichkeiten). Dazu passt es die generischen Funktionalitäten von Repositorien an den spezifischen Kontext einer Community an (z.B. Anwendungsszenarien, organisatorischer Rahmen). Vor allem Vertrauenswürdigkeit und Langzeitarchivierung von Objekten bauen zwar auf die technische Basis von Repositorien und LZA-Diensten, können letztlich aber nur durch darüber liegende organisatorische Maßnahmen gewährleistet werden (z.B. finanzielle Stabilität, Rollen für Preservation Planning und Audit[5]).

[4] Rachel Heery, Sheila Anderson. Digital Repositorys Review. UKOLN, AHDS, 2005. http://www.ukoln.ac.uk/repositories/publications/review-200502/#review-200502

[5] Research Libraries Group. (2002). Trusted digital Repositorys: Attributes and responsibilities. An RLG-OCLC Report. http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

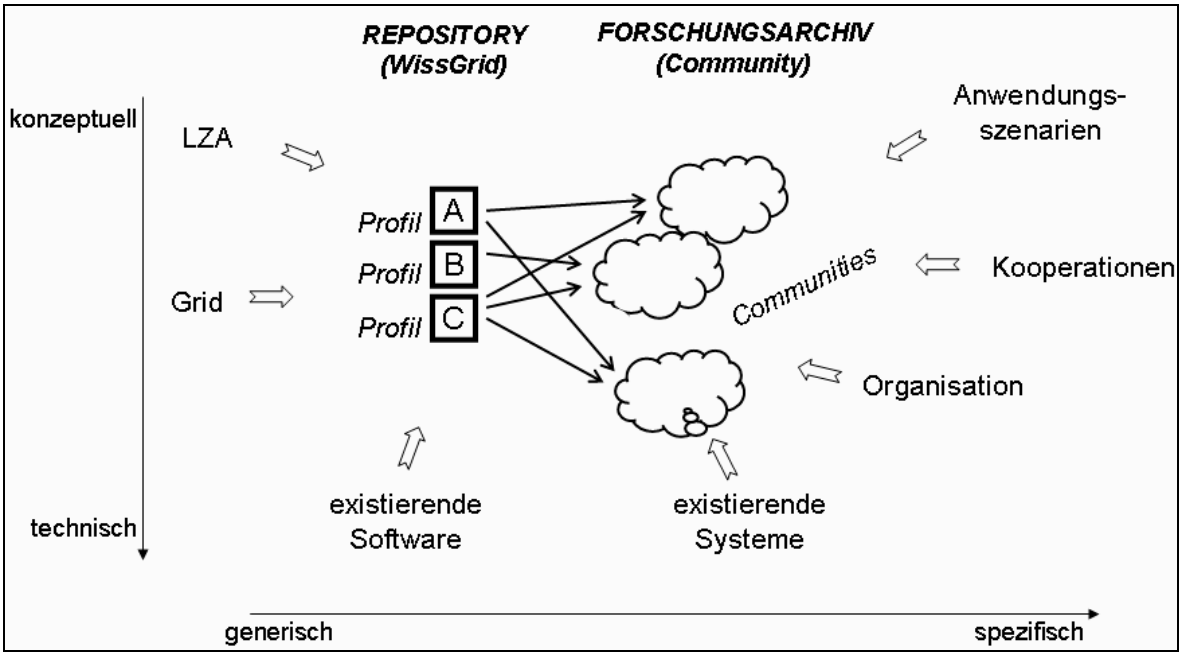


Abbildung 1 - Beziehung zwischen generischen Profilen
für Repositorien und deren (angepasster) Einsatz in Communities

Dieses Dokument konzentriert sich auf die technische Spezifikation eines generischen WissGrid Repositorys - technische Funktionen und Interoperabilität, sowie Einbindung in D-Grid und Verknüpfung mit anderen Komponenten der WissGrid LZA-Architektur.

Auch wenn dieses Dokument möglichst generisch gehalten ist, zeigt das oben beschriebene Verhältnis zwischen Repository-Technologien und Forschungsarchiven die mögliche Breite der denkbaren Anforderungen aus den Communities (z.B. existierende Systemumgebungen, Anwendungsszenarien, Unterstützung organisatorischer Strukturen, Preservation Planning). Um diesen breiten Anforderungen möglichst gerecht zu werden, unterscheidet WissGrid zwischen unterschiedlichen Anwendungsprofilen und damit auch - anstatt eines einzigen, generischen Software-Pakets - zwischen unterschiedlichen technischen Startpunkten zur Errichtung eines Community-spezifischen Forschungsarchivs. Anders gesagt, die Anwendungsprofile wurden gemeinsam mit den Communities entwickelt, um den Abstand zwischen den spezifischen Anforderungen aus den Communities und dem verfügbaren technischen Angebot aus WissGrid zu verringern.

Diese Spezifikation wird zuerst die Rolle von Repositorien in der WissGrid LZA-Architektur und die Einbettung in D-Grid abstrakt darlegen (Kapitel 2). Bereits in der WissGrid LZA-Architektur wurden die unterschiedlichen Anwendungsprofile für Repositorien beschrieben,

und diese werden in Kapitel 3 als Überblick iteriert. Ab Kapitel 4 folgt eine detaillierte technische Spezifikation der einzelnen Profile: eine Beschreibung des passenden Community-Kontextes, ein Überblick über Funktionen und Schnittstellen und schließlich auch die technische Umsetzung. Diese Spezifikation wird im Laufe von WissGrid kontinuierlich weiterentwickelt (zunächst in einer "zweiten" Iteration mit Erfahrungen aus technischen Prototypen) und muss auch jenseits der WissGrid Projektzeit einem ständigen Modifikations-prozess unterliegen.

2 Verortung in Architektur

Bei den in der WissGrid LZA-Architektur identifizierten Typen von Diensten lassen sich Repositorien als "Archiv- und Speicherdienste" einordnen. Jenseits dieser Kategorisierung bieten Repositorien diverse Funktionen und Schnittstellen. Als solche sind sie mehr als ein einzelner, klar definierter Dienst und sind vielmehr selbst ein abgestimmtes System aus mehreren einzelnen Diensten. Dieses Kapitel gibt einen Überblick über (1) die Dienste, aus denen ein Repository besteht mithilfe eines Mappings auf den OAIS Standard, über (2) die Einbettung in die D-Grid Infrastruktur, sowie über (3) die Interaktion mit anderen LZA-Diensten Diensten.

2.1 OAIS Mapping

Der ISO Standard OAIS[6] - a Reference Model for an Open Archival Information System - definiert die minimalen Funktionen eines generischen Repositorys. Dabei definiert das OAIS keine Architektur, sondern bietet vielmehr eine strukturierte Checkliste für die organisatorischen und technischen Aufgaben, die in einem Archiv bedacht werden müssen. Die Rollen für die Aufgaben können sich über mehrere Partner-Organisationen oder auch kommerzielle Anbieter verteilen, wenn dies aus dem organisatorischen Kontext heraus abgeleitet werden muss. Die Hauptkomponenten des OAIS sind: Ingest, Administration, Archival Storage, Data Management, Access, und Preservation Planning (vgl. Abbildung 2).

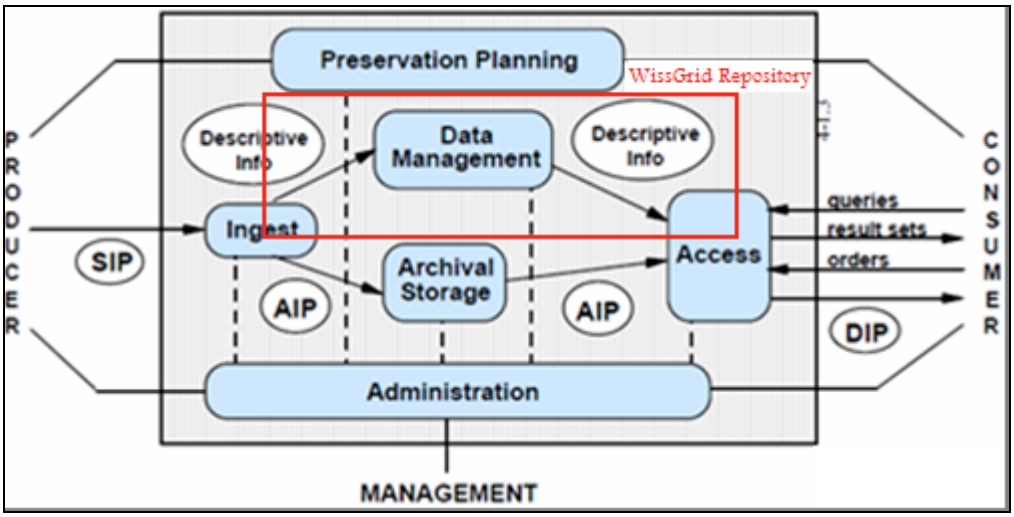


Abbildung 2 - OAIS Functional Entities (aus dem OAIS Standard)

[6] CCSDS - Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS), 2003. ISO 14721:2003.

Administration, Preservation Planning (OAIS): Administration und Preservation Planning umfassen mehr organisatorische Aufgaben, die in dieser Spezifikation zwar durch technische Dienste unterstützt, aber nicht direkt adressiert werden. Diese Spezifikation umfasst vor allem das Daten- und Metadatenmanagement sowie die generischen Aspekte der Schnittstellen Ingest und Access.

Ingest, Access (OAIS): Die Minimalanforderungen der Schnittstellen Ingest und Access werden durch CRUD-Funktionen abgedeckt; hierbei steht CRUD für Create/Read/Update/Delete und stellt die Grundbausteine von Datenverwaltungssystemen (z.B. Datenbanken und Repositorien) dar.[7] In den Kapiteln zu den Anwendungsprofilen (Kapitel 4-6) werden die technischen Komponenten und Schnittstellen unter anderem über ihre CRUD-Funktionalitäten spezifiziert.

Data Management (OAIS): Das Datenmanagement in Repositorien wird zumeist (aus Gründen der Robustheit und Skalierbarkeit) in der Form von Dateien durchgeführt, wobei Daten und Metadaten in XML Containern im Dateisystem abgelegt werden. Zusätzlich können diverse Datenbanken und Suchindices einen komfortableren und schnelleren Zugriff zu den Daten ermöglichen (z.B. Suche über Metadaten, Volltextsuche). Für die technische Umsetzung des Datenmanagements in den Anwendungsprofilen (Kapitel 4-6) werden existierende Softwarepakete herangezogen und für eine Nutzung im Grid angepasst. Unter den existierenden generischen Repository Software-Paketen sind z.B. Fedora, aDORe, Tupelo und andere Systeme zu nennen, die in der Open Repositorys[8] Konferenzreihe regelmäßig vertreten sind.

Archival Storage (OAIS): Archivdaten können auf den unterschiedlichsten Medien gespeichert werden. In einem Grid-Umfeld können existierende Storage-Ressourcen im Grid eine virtualisierte Speicherung ermöglichen, die nicht an die Hardware eines lokalen Systems gebunden ist (On Demand Storage). Für eine vertrauenswürdige Datenspeicherung im Sinne der LZA müssen dabei vor allem die Anforderungen zur Bit Preservation erfüllt sein, wie sie in der WissGrid Architektur beschrieben stehen. Auch die Errichtung von "Trust Zones", in denen ausgewählte Storage-Ressourcen besondere Qualitäts- und Sicherheitskriterien erfüllen, ist dafür denkbar. Die Bereitstellung von Bit Preservation auf D-Grid Ressourcen kann durch WissGrid angestoßen werden (in Bezug auf Anforderungen und mögliche

[7] CRUD. http://en.wikipedia.org/wiki/Create,_read,_update_and_delete
Macario Polo, Mario Piattini, and Francisco Ruiz. Reflective Persistence – Reflective CRUD: Reflective Create, Read, Update and Delete. http://hillside.net/europlop/HillsideEurope/Papers/ReflectivePersistence.pdf, 2001.

[8] Open Repositorys. http://openrepositories.org/

Herangehensweisen), muss aber letztlich durch D-Grid und mögliche andere Projekte (z.B. WisNetGrid) in den Betrieb umgesetzt werden.

2.2 Einbettung D-Grid

Wie das Mapping der OAIS-Komponenten zeigt, ist ein Repository kein einzelner, monolithischer Dienst, der sich über mehrere konzeptuelle Schichten erstreckt (Bit Preservation, Content Preservation, und Data Curation; siehe Definition im WissGrid Architektur-Dokument). Vielmehr können je nach Community-spezifischem Kontext die Funktionalitäten in den Schichten variieren.

Wie im WissGrid Architektur-Dokument beschrieben, konzentriert sich WissGrid auf die Content Preservation Schicht, und unterstützt die Communities (durch Beratung) ihre Community-spezifische Data Curation zu gestalten. Bit Preservation als Basis einer LZA-Strategie muss entweder durch die Community-eigenen Datenzentren sichergestellt werden, bzw. wird sich WissGrid gemeinsam mit D-Grid in der Entwicklung einer Musterlösung auf Basis von D-Grid Storage-Ressourcen engagieren, die auch von Communities genutzt werden kann (siehe speziell Kapitel 5).

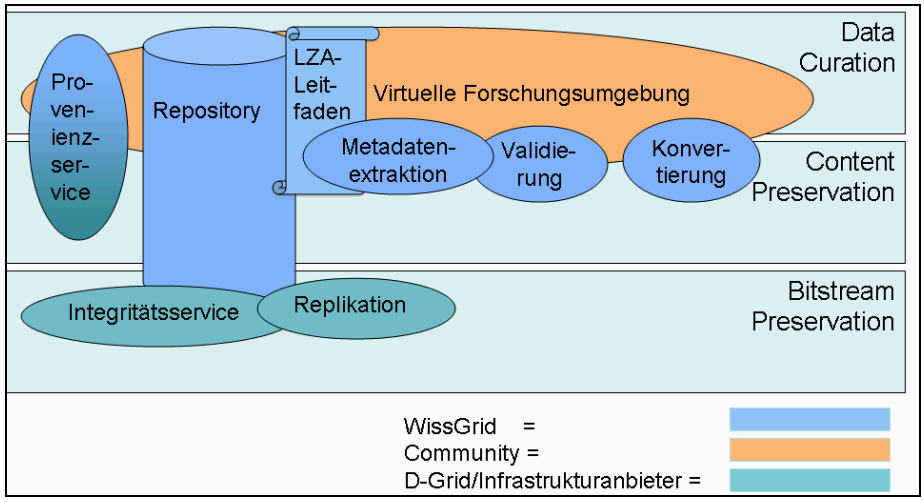


Abbildung 3 - Strukturierung einer LZA-Strategie nach 3 konzeptuellen Ebenen, und Zuteilung der WissGrid Aktivitäten nach diesen Ebenen bzw. Interaktion mit D-Grid und Communities.

Neben der Nutzung von D-Grid Storage-Ressourcen unterscheiden wir im Anhang 5 des WissGrid Architektur-Dokumentes mehrere Varianten zur Einbettung eines Repositorys in die Grid-Umgebungen.

1. Repositorien als Archiv-Backends für das Grid - die im Repository gespeicherten Objekte können direkt in wissenschaftliche oder administrative Workflows im Compute-Grid eingebunden werden (gewählter Ansatz: "Daten zu den Diensten")
2. Daten-Grid als Storage für Repositorien - Nutzung der D-Grid Storage Ressourcen für Bit Preservation (ggf. in expliziten Trust Zones, die Vertrauenswürdigkeit hinsichtlich von Kriterien zur Bit Preservation und Sicherheit garantieren)
3. Virtualisierung von Repositorien - Vernetzung von mehreren (technisch und organisatorisch getrennten) Repositorien in ein "virtuelles" Repository, eine so genannte "Repository Föderation".

Jede dieser Einbettungs-Varianten erfordert diverse Interoperabilitätsmechanismen zwischen den involvierten Komponenten. Hierzu zählt auch die Interoperabilität zu horizontalen Diensten aus D-Grid und den Community Grids (z.B. Authentifizierung/Autorisierung und Sicherheit, und D-Grid Informationsdienste). Nicht alle diese Interoperabilitätsmechanismen können in einem Gesamtsystem und im Rahmen der WissGrid Projektzeit komplett umgesetzt werden, aber die folgende Spezifikation beschreibt, wie die Vorteile jeder Integrationsvariante für spezifische Community-Anforderungen im Rahmen der verfügbaren Ressourcen möglichst optimal umgesetzt werden können.

2.3 Interaktion Repository / LZA-Dienste

Neben der Einbettung zwischen den konzeptuellen Schichten (Bit Preservation: D-Grid, Content Preservation: WissGrid, und Data Curation: Communities), ist für WissGrid Architektur vor allem die Interaktion zwischen dem Repository und den LZA-Diensten in der Content Preservation Schicht relevant, wie es in Abbildung 4 angedeutet ist.

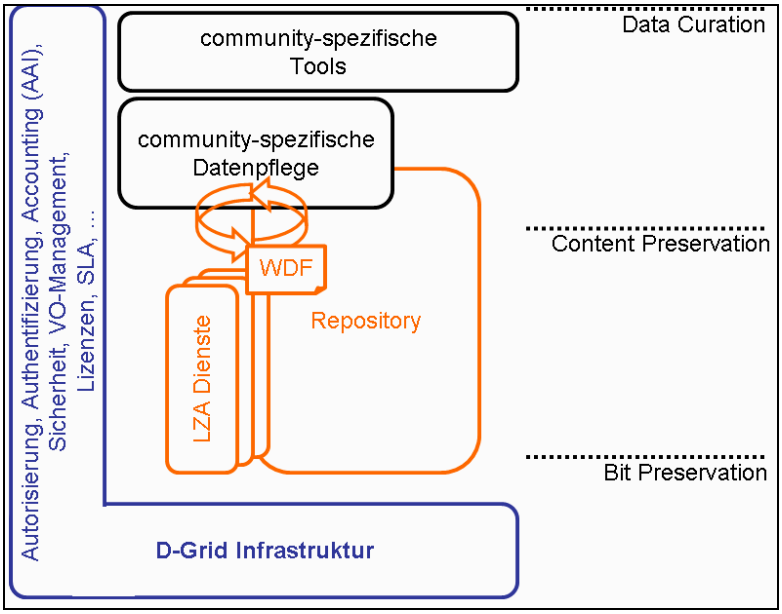


Abbildung 4 - Komponenten in allen konzeptuellen LZA-Schichten, und speziell die Vernetzung des Repositorys und den LZA-Diensten in der Content Preservation Schicht

Die Interaktion zwischen einem Repository und LZA-Diensten kann je nach den Community-spezifischen Anforderungen unterschiedlich eingesetzt werden: So könnte beispielsweise eine Community als Teil ihrer LZA-Strategie alle Daten beim Einbringen ins Archiv (dem so genannten "Ingest") in ein Standard-Format überführen [Conversion on Ingest], eine andere nur beim Zugriff [Conversion on Access], und davon abgesehen könnten alte Bestände systematisch von einem Format in ein anderes überführt werden [Migration als Ergebnis des Preservation Planning]. - Alle diese Fälle sind beispielhafte Ausprägungen der Interaktion zwischen dem Repository und dem LZA-Dienst zur Formatkonvertierung; analog können alle LZA-Dienste wie in einem Dienste-Baukasten nach Anforderungen (die sich für eine Community auch im Laufe der Zeit ändern können) eingesetzt werden.

Die technische Komponente, die die Interaktion zwischen einem Repository und den LZA-Diensten ermöglicht, ist ein WissGrid Dienste Framework (WDF) und in der WissGrid LZA-Dienste Spezifikation beschrieben. Hierbei hat perspektivisch jedes Repository einen angepassten WDF-Dienst, der die Interaktion des Repositorys mit D-Grid-weiten und Community-spezifischen LZA-Diensten koordiniert.

Zur Sicherstellung der Kommunikation zwischen einem Repository und einem zugehörigen WDF-Dienst, muss das Repository folgende Funktionen anbieten:

- CRUD-Schnittstelle (vgl. Ingest/Access in Kapitel 2.1), zur Extraktion und Einfuhr von Daten

- Filterung, zur Selektion von Daten (z.B. nach Formaten, Zeit letzte Aktualisierung)
- Möglichkeit zur automatischen Rückführung der (z.B. konvertierten) Daten, Versionierung, Setzen der Metadaten, Audit Trail für Provenance, Rechte-Abgleich bei Zugriff über Service (z.B. Robot-Zertifikat)
- Trust Level: durch die Metadaten sollte der erwartete Trust Level für ein Objekt klar werden. Eine Policy könnte z.B. sein, dass private Daten anders behandelt werden als Open Access Daten

Die genaue technische Ausprägung dieser Schnittstelle (z.B. welche Metadaten sind zur Filterung relevant, Protokolle) wird im Laufe der Weiterentwicklung dieser Spezifikation ausgeführt.

3 Repository Anforderungen - Anwendungsprofile

Fach-Communities unterscheiden sich zum Teil grundlegend darin, wie Repositorien in ihre Forschungsumgebungen eingebunden werden können, deren Anbindung an das Grid, und wie Community-spezifische Services und (interaktive) Tools beide verknüpfen. Die optimale Anpassung an Nutzer-Anforderungen der jeweiligen Fach-Community muss Priorität haben.

Trotz der unterschiedlichen Anforderungen sind Repositorien in allen Communities einige Kernfunktionalitäten gemein. Dazu zählen

- die Verwaltung von datei-basierten digitalen Daten (digitale Objekte),
- die Verknüpfung von Metadaten mit diesen Objekten sowie Relations-Netzwerke zwischen diesen Objekten,
- der ortsungebundene Zugriff auf Objekte durch virtuelle Teams, die über institutionelle oder geographische Grenzen hinweg miteinander kooperieren,
- sowie die Unterstützung von Strategien zur Erhaltung der Daten und ihrer Interpretierbarkeit und damit die Interoperabilität des Repositorys mit anderen WissGrid LZA-Diensten.

Jenseits dieser generischen Kernfunktionalitäten ist die Einbettung eines Repositorys in den spezifischen Community-Kontext von zentraler Bedeutung. Dabei sind u.a. folgende Aspekte von Bedeutung:

- Wie sehen die Nutzungsszenarien für die Daten im Repository bzw. Forschungsarchiv aus? In welche (automatischen, interaktiven und manuellen) Workflows sind die Daten einzubinden? Technisch betrifft dies z.B. die Referenzierbarkeit der Daten (durch z.B. Persistent Identifier), mögliche Schnittstellen (z.B. FTP, GSIFTP, WebDAV, Storage Cloud REST Interface) oder die unterstützte Zugriffsgeschwindigkeit (z.B. offline, near-line, oder unverzögerter online Zugriff). [Nutzungsanforderungen]
- Welche Arten von Daten und Metadaten sind in der Community verbreitet bzw. in welchen Volumina treten sie auf? [Skalierbarkeit des Systems]
- Welchen Objekt-Modellen liegen die Daten zugrunde? Einige Punkte der Daten-Modellierung umfassen: Datenformate, Komplexität der Objekt-Aggregationen, Beschreibung der Daten und Objekte sowie die Verknüpfung der Daten untereinander.

Fragen der Daten-Modellierung müssen Communities selbst entscheiden können, aber auf einer Interoperabilitätsebene ist es möglich, komplexe Community-spezifische Datenmodelle auf simplere, generische Datenmodelle abzubilden (im Dublin Core[9] Jargon: Dumbing-Down). [Daten-Modellierung]

- Welchen Anforderung muss die Datenverwaltung genügen in Hinblick auf z.B. Datenschutz und Datensicherheit, Zugriffsbeschränkungen und Lizenzen, Daten-Integrität und redundanter Speicherung, und anderen? [administrative Anforderungen]

Die Heterogenität der möglichen Anforderungen aus den Communities können kaum alle sinnvoll von einem einzigen monolithischen System umgesetzt werden. Gerade das auf dem Paradigma der Service-Orientierung aufbauende technische Umfeld der Grid-Technologien legt daher nahe, dass unterschiedliche Anforderungen auch von unterschiedlichen Diensten adressiert werden. Um spezifischen Anforderungen gerecht zu werden, können daher Repositorien aus unterschiedlichen technischen Systemen, Modulen bzw. aus unterschiedlichen Konfigurationen von Modulen bestehen und auch organisatorisch getrennt sein.

Im Folgenden werden einige typische Anwendungsprofile unterschieden und hinsichtlich ihrer technischen Umsetzung auf Basis der oben genannten drei Grid/Repository Integrationsvarianten (vgl. Kapitel 2.2) spezifiziert.

- Profil A / Variante 1: ein Software-Stack für ein Repository, dessen Inhalte über Grid-Protokolle exportiert und importiert werden können, um z.B. Datenanalysen in einem Compute-Grid durchzuführen
- Profil B / Variante 2: ein Software-Stack für ein Repository, das seine Inhalte auf Grid Ressourcen abspeichern bzw. auch vertrauenswürdig aufbewahren kann (wo entsprechende Bit Preservation durch Trusted Zones im Grid angeboten wird)
- Profil C / Variante 3: Adaptoren für existierende Repositorien, die die Vernetzung (auch genannt: "Föderierung") von Repositorien untereinander ermöglicht, bzw. auch den Datenaustausch mit anderen Komponenten (z.B. Visualisierungscluster, Analyse im Grid, föderierte Such-Portale) unterstützt.

Je nach Anforderungen kann eine Community eines dieser Anwendungsprofile und die zugehörige technische Realisierung heranziehen, um es für sich zu installieren und

[9] Dublin Core: Dumb Down Principle. DCMI Glossary.
http://dublincore.org/documents/usageguide/glossary.shtml

anzupassen. In diesem Sinne unterstützen die Profile die Anforderungsanalyse und verbessern die Passgenauigkeit des WissGrid-Angebots für Repositorien. Auch Mischformen und die Kombination mehrer Profile sind im Prinzip denkbar, erhöhen aber potenziell den Aufwand für die Umsetzung überproportional.

4 Profil A - Grid-Workflow

Dieses Anwendungsprofil deckt speziell jene Community Grids ab, die große Mengen an Daten im Grid erzeugen und verarbeiten. Neben einer ständigen Verfügbarkeit für Verarbeitungen im Grid sollen diese Daten auch in einer vertrauenswürdigen Umgebung über lange Zeiträume aufbewahrt werden.

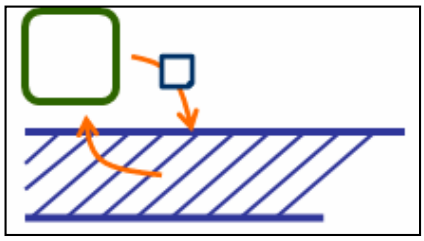


Abbildung 5 - Schnittstelle zwischen persistentem Repository und Compute-Grid. (vgl. Integrationsvariante 1, Anhang 5, WissGrid Architektur)

Zur Verarbeitung im Grid können Daten in einem durchgehenden Workflow direkt aus einem Repository extrahiert, in einem Compute-Grid verarbeitet und schließlich die Ergebnisse wieder ins Repository zurückgespielt werden.

In diesem Profil erarbeitet WissGrid eine konkrete technische Umsetzung (auf Basis der Grid/Repository Integrationsvariante 1, siehe Abbildung 5) für Communities, die derzeit noch kein Grid-Repository besitzen.[10] Den dafür von WissGrid entwickelten Repository Software-Stack können Communities für sich in D-Grid installieren und anpassen. Organisatorische Aspekte zur Einbettung eines Repositorys in den Community-Kontext und zur langfristigen Wartung werden in diesem technischen Dokument nicht abgedeckt.

4.1 Spezifische Anforderungen

Dieses Anwendungsprofil ist vor allem für Communities mit u.a. folgenden spezifischen Anforderungen geeignet:

- Skalierbarkeit auch bei großen Datenmengen im Terabyte-Bereich und großer Anzahl von einzelnen Objekten.

  Dies fordert die Verwaltung und Wiederauffindbarkeit von einem speziellen Objekt in Sammlungen aus einer großen Anzahl Objekten sowie ein adäquates Datei-Management. Bei häufigem Transfer und Verarbeitung der Daten im Grid ist je nach Transfer-Rate die Größe einzelner Objekte zu beachten, da sie die Gesamt-Performance der Anwendung einschränken kann.

[10] Während die Profile A und B also die technische Basis für neue Repositorien schafft, wird in Profil C (Archiv Föderationen) ein Ansatz zur Verknüpfung von existierenden Repositorien verfolgt.

- Technische Einbindung in die Grid-Umgebung.

  Die Unterstützung wesentlicher Protokolle (z.B. GridFTP, SRM) sichert bei den Zugriffsarten (Ingest, Access) die Integration der Daten in die Grid-Workflows und unterstützt die Interoperabilität

- Effiziente Verknüpfung mit Metadaten für die Einbindung in Workflows sowie die Verwaltung im Grid.

  Vor allem genau vorstrukturierte deskriptive Metadaten (z.B. zum schnellen Retrieval) sowie administrative Metadaten (z.B. für Administration von Daten-Replikation und Integritätsprüfung im Grid) sind hierbei gefragt.

Für die Einbindung in D-Grid wird bei diesem Anwendungsprofil speziell die Integrationsvariante 1 ("Repositorien als Archiv-Backends für das Grid", vgl. WissGrid LZA-Architektur Anhang 5) genutzt. In der technischen Umsetzung liegen daher die Schwerpunkte auf den Grid-Schnittstellen des Repositorys und der direkten Einbettung von Objekten aus dem Repository in Grid-Workflows.

4.2 Aus den Fallstudien

In der **Biostatistik** werden die im Rahmen klinischer Studien erhobenen Daten (z.B. über Behandlungsmethoden und Medikationen) ausgewertet, um z.B. Korrelationen zwischen Krankheitsbildern und genetischen Merkmalen zu erforschen. Die Basis dafür sind anonymisierte bzw. pseudonymisierte Daten zu Genetik, Biomaterial, Bildmaterial und Krankheitsbildern, auf denen z.B. Mustererkennungsalgorithmen und statistische Analysen im Grid durchgeführt werden. Ein typisches Szenario ist hierbei der Erhalt bestimmter, in einer Studiendatenbank vorgehaltenen Daten und die anschließende Auswertung mithilfe von statistischen Entwicklungsumgebungen wie R oder SAS. Diese wissenschaftlichen Daten (vom Ausgangsmaterial bis zu den Ergebnisdaten) sind z.B. im Rahmen von Patentierungen wesentlich, aber auch zur Nachnutzung und für gute wissenschaftliche Praxis langfristig wertvoll.

Derzeit gibt es in der Biostatistik keine Forschungsdatenarchive, bzw. werden lediglich vereinzelt und lokal Repositorien gehalten. Eine solche ad hoc Datenverwaltung genügt allerdings nicht den Ansprüchen eines vertrauenswürdigen Forschungsarchivs und kann die Authentizität und Provenienz von Forschungsdaten nicht langfristig sichern. Die WissGrid Repository-Angebote können die technische Basis für Forschungsarchive in der Biostatistik

schaffen; speziell Profil A ist auf die Aufbewahrung von heterogenen Objekten und deren Nutzung in wissenschaftlichen Berechnungen im Grid ausgelegt.

Hintergrund und Details zu Kontext und Anforderungen der Biostatistik finden sich im WissGrid Architekturdokument (Anhang Fallstudien).

4.3 Schnittstellen und Funktionalität

In diesem Abschnitt werden die Anforderungen für das "Profil A" auf technische Schnittstellen und grundlegende Funktionalitäten abgebildet. Die technische Umsetzung wird dann im folgenden Abschnitt beschrieben.

Wie in Kapitel 2 beschrieben, stellen die CRUD-Schnittstellen die Basis für die Einbettung des Repositorys in D-Grid dar.

- Create
    - *Funktion:* Ingest (OAIS), Erzeugen einer Grid-ID, Registrierung in relevanten Grid Information Services und Setzen der Rechte
    - *technische Schnittstelle:* der Storage Resource Manager (SRM)[11] gilt als Interoperabilitäts-Protokoll im Storage-Bereich; iRODS i-commands
- Read
    - *Funktion:* Access (OAIS), möglichst direkter Zugriff aus einem Compute-Grid, nach Überprüfung der Rechte
    - *technische Schnittstelle:* SRM - evt. wahlweise Ausgabe des nativen Files oder als OAI-ORE bzw. METS Container mit den Metadaten
- Update
    - *Funktion:* vgl. Ingest (OAIS), abhängig von Preservation Policy (z.B. Überschreiben nicht möglich, nur Einspielen einer neuen Version), und nach Überprüfung der Rechte
    - *technische Schnittstelle:* (siehe "Create")
- Delete

[11] Storage Resource Manager (SRM). https://sdm.lbl.gov/srm-wg/

- *Funktion:* Löschen von Objekten abhängig von Preservation Policy bzw. Rechten; in jedem Fall sollte das Löschen ein "Sonderfall" sein und nicht aus dem wissenschaftlichen Workflow heraus möglich sein
- *technische Schnittstelle:* iRODS i-commands ("Löschen" darf nur mit speziellen Rechten möglich sein, und darf nicht "zufällig" aus dem wissenschaftlichen Workflow heraus "passieren", daher ist hier keine SRM-Schnittstelle vorgesehen; außerdem ist in SRM ein Löschen von Objekten je nach SRM-Version gar nicht vorgesehen)

Neben CRUD-Schnittstellen sind in Kapitel 2 zur Verknüpfung der WissGrid LZA-Dienste mit einem Repository einige Schnittstellen und Funktionalitäten identifiziert worden:

- Filter für Abfragen
  - *Funktion:* Abfrage von Objekten nach Metadaten (z.B. alle Objekte von Autor XY, oder alle Objekte im Format YZ)
  - *technische Schnittstelle:* CQL oder OpenSearch
- Rechte
  - *Funktion:* Kompatibilität mit Sicherheitsinfrastruktur in Compute Umgebung
  - *technische Schnittstelle:* GSI (Grid Security Infrastructure)[12]
- Trust Level
  - *Funktion:* Ausgabe des Trust Levels (Minimalanforderungen bei externer Verarbeitung eines Objektes) als Teil der Metadaten
  - *technische Schnittstelle:* iRODS-Metadaten, bzw. eingebettet in OAI-ORE Container

Folgende weitere Funktionalitäten könnten potenziell die Einbettung des Repositorys in wissenschaftliche Grid-Workflows erleichtern und erweitern. Deren Umsetzbarkeit ist noch zu prüfen:

- Bulk Import und Export

[12] Ian Foster et al.: A Security Infrastructure for Computational Grids. ftp://ftp.globus.org/pub/globus/papers/security.pdf

- *Funktion:* ganze Sammlungen können in einem in das Archiv eingeführt bzw. aus dem Archiv ausgelesen werden

- OAI-ORE Export
  - *Funktion:* Exposure von einzelnen Objekten oder ganzen Objektgruppen via ein Container-Format wie METS[13] oder OAI-ORE[14]. Dies betrifft vor allem das "Read" der CRUD-Schnittstelle.

- Tests aus unterschiedlichen Grid Middlewares
  - *Funktion:* Evaluation der Nutzung aus unterschiedlichen Grid Middlewares. Falls z.B. der SRM durch SAGA [15] auch im Storage-Bereich als Interoperabilitätsstandard abgelöst wird, müsste entsprechend reagiert werden.

4.4 Auswahl der Technologien und technische Umsetzung

Die "technische Umsetzung" wird im Projektverlauf fortgeschrieben bis zur abschließenden Dokumentation der Produkte.

iRODS hat sich als Basistechnologie für die oben beschriebenen Anforderungen und in Gesprächen mit den Communities, innerhalb von D-Grid und mit externen Experten herauskristallisiert. Dies hat vor allem zwei Gründe:

(1) Das Daten-Grid-System iRODS wird bereits vom DGI unterstützt.[16] Über das virtuelle Supportzentrum des DGI erhalten Nutzer Beratung und Hilfestellung für den Einsatz von iRODS.[17] - Die enge Zusammenarbeit mit dem DGI ist wesentlich für die WissGrid LZA-Architektur.

(2) Die Nutzung von iRODS alleine ist nicht hinreichend für die Umsetzung eines LZA-Konzepts, allerdings gibt es eine aktive Community[18], die iRODS entsprechend anpassen und

---

[13] Metadata Encoding and Transmission Standard (METS). http://www.loc.gov/standards/mets/

[14] OAI-ORE, Object Reuse and Exchange. http://www.openarchives.org/ore/

[15] Simple API for Grid Applications (SAGA). http://www.ogf.org/documents/GFD.90.pdf

[16] Es ist zwar (noch) nicht im offiziellen D-Grid Software Stack, hat aber de facto das Vorgängersystem SRB, Storage Resource Broker, abgelöst.

[17] Die Homepage des virtuellen Supportzentrums, dessen Dienste auch in die DGUS- Supportinfrastruktur eingebunden sind, bietet unter http://dgi2.d-grid.de/index.php?id=454 auch Einstiegsmaterial und Installationsanleitungen bezüglich iRODS an.

[18] Eines der diesbezüglich wichtigsten Projekte ist SHAMAN.
SHAMAN - Sustaining Heritage Access through Multivalent ArchiviNg. http://shaman-ip.eu/shaman/

in den spezifischen Kontext einbetten wird(?). - Die Einbettung in die Arbeiten der LZA-Community ist wesentlich für die Nachhaltigkeit eines LZA-Konzeptes, und damit auch für die WissGrid LZA-Architektur.

Im Folgenden werden die oben (siehe 4.3) beschriebenen Anforderungen auf iRODS abgebildet, um fehlende Komponenten zu identifizieren. Hierbei werden nicht nur die existierenden Komponenten vom iRODS Kernsystem betrachtet, sondern auch andere (obwohl zum Teil noch ungetestete) Vorarbeiten aus internationalen Projekten. WissGrid wird diese existierenden Vorarbeiten in ein Gesamtpaket für das "Profil A" integrieren, und wo notwendig fehlende Komponenten in Eigenentwicklung nachführen.

4.4.1 SRM-Schnittstelle

Eine SRM-Schnittstelle für iRODS existiert bisher nicht.

Für den eigentlichen Zugriff existiert eine C-basierte Programmierschnittstelle sowie diverse Clients, die letztendlich auch wieder die C-basierte Schnittstelle des Systems verwenden. Um SRM-Kompatibilität zu erreichen, bieten sich die sogenannten iCommands an - kleine Kommandozeilenbasierte Programme, die mit iRODS mitgeliefert werden und das System über Befehlsaufrufe nutzbar macht.

Der Vorgänger von iRODS - der SRB, Storage Resource Broker - konnte ein SRM-SRB Interface[19] vorweisen. Mit der Ablösung von SRB durch iRODS gibt es Interesse an einem SRM-iRODS Interface,[20] derzeit sind aber keine diesbezüglichen Entwicklungsarbeiten bekannt.

4.4.2 Retrieval-Filter: CQL/OpenSearch

Für Anfragen an Repository-Kataloge haben sich vor allem die Contextual Query Language (CQL) und OpenSearch etabliert,[21] die von der konkreten Implementierung des Metadaten-

---

... Für eine möglichst umfassende Abdeckung von LZA-Funktionalitäten in iRODS arbeitet SHAMAN auch an einer Abbildung der LZA-Kriterien der TRAC-Checkliste auf das iRODS System:
Perla Innocenti, Seamus Ross, Elena Maceviciute, Tom Wilson, Jens Ludwig, Wolfgang Pempe: Assessing Digital Preservation Frameworks: the approach of the SHAMAN project. In: Proceedings of the ACM MEDES. France, October 2009.

[19] SRM-SRB Interface: A new milestone for grid interoperation
http://www.beliefproject.org/news/srm-srb-interface-a-new-milestone-for-grid-interoperation

[20] EGEE Introduction for SRM-SRB interface. See "Future".
http://www2.twgrid.org/APTeam/images/1/14/SRM-SRB_Intro.pdf

[21] Ray Denenberg: OASIS Search Web Services. In: D-Lib Magazine, January/February 2009.
http://www.dlib.org/dlib/january09/denenberg/01denenberg.html

Managements im Repository abstrahieren (z.B. relationale Datenbank, RDF Triple Store). CQL wurde ursprünglich für das SRU Protokoll entwickelt, das sowohl als SOAP- als auch als REST-Variante vorliegt.

Für iRODS direkt gibt es bisher keine bekannte Implementierung von SRU oder OpenSearch. Allerdings bietet das "Cheshire3 Information Framework" sowohl SRU, als auch OAI-PMH[22] Interfaces, die hierfür wieder verwendet werden können. Cheshire wird im Rahmen des EU-Projektes SHAMAN prototypisch auf ein iRODS System aufgesetzt.

4.4.3 GSI-Schnittstelle

Die Nutzung von iRODS über GSI-Authentifizierung wurde nachträglich in den iRODS Release aufgenommen[23] und wurde bereits vom DGI getestet.[24]

4.4.4 Bulk Import: koLibRI

Gerade in wissenschaftlichen Umgebungen liegen oftmals größere Datensammlungen in Verzeichnisstrukturen zur Einfuhr in ein Repository bereit. Obwohl man über manuelle Eingabe oder über handgefertigte Skripte die Sammlungen in das Repository einführen könnte, empfiehlt sich durch die schiere Datenmenge ein entsprechendes Framework zum automatischen Bulk Import.

Im Rahmen des LZA-Projektes kopal wurde eine generische Engine zum Bulk Ingest geschrieben: koLibRI.[25] WissGrid wird die Möglichkeiten zur Anpassung von koLibRI für iRODS prüfen.

4.4.5 OAI-ORE Export - Read-Schnittstelle

In digitalen Objekten sind (mitunter mehrere) Dateien und zugehörige Metadaten eng mit einander verknüpft. Oft erlauben Repositorien daher auf der einen Seite den Zugriff auf einzelne Dateien, auf der anderen Seite auch auf Container, die (mehrere) Dateien und Metadaten verknüpfen. Verbreitete Container-Formate sind METS und OAI-ORE.

---

[22] Open Archives Initiative - Protocol for Metadata Harvesting.
http://www.openarchives.org/OAI/openarchivesprotocol.html

[23] iRODS unterstützt GSI seit Release 1.1
https://www.irods.org/index.php/Release_Notes_1.1

[24] DGI-Bericht M18 „iRODS mit GSI“ auf der iRODS Seite des DGI (http://dgi2.d-grid.de/index.php?id=456)

[25] kolibri, kopal Library for Retrieval and Ingest. http://kopal.langzeitarchivierung.de/index_koLibRI.php.de

Zur Auswahl des richtigen Formates würde man in einer iRODS-Umgebung jeweils unterschiedliche Identifier vergeben und diese mit unterschiedlichen Format-Metadaten suchbar machen.

Derzeit ist kein OAI-ORE oder METS Export für iRODS verfügbar.

4.4.6 Verknüpfung mit Profil C: OAI-PMH Schnittstelle

Uns ist derzeit kein stabiler OAI-PMH Provider Service für iRODS bekannt. Um im Sinne von Profil C auch Daten über OAI-PMH aus iRODS extrahieren zu können, müsste diese Funktionalität in iRODS nachimplementiert werden.

4.5 Überblick über Aufgaben für Profil A

- [alle, Moderation UGOE] Anleitung zum Aufbau eines Repositorys nach Profil A; primär als technische Anleitung, aber auch unter Erwähnung organisatorischer Aspekte zur Nutzung als Blaupausen von Fachberatern, AP2:
    - Design und Umsetzung: Daten-Modelle, Einbettung in Community Workflows
    - Planung Hardware-Infrastruktur, Personal (intern/extern?)
    - Migration existierender Tools und Daten
- [ZIB, SUB] Interaktion mit D-Grid und verwandten Projekten für Ebene Bit Preservation
    - Aufbau von Bit Preservation (in dezidierten Trusted Zones) in D-Grid als Angebot für Communities ohne Hardware
    - Aufnahme von iRODS in den D-Grid Software Stack
    - Verknüpfung der Arbeiten des DGI2 FG4 Supportzentrums
- Entwicklung Schnittstellen iRODS
    - SRM Schnittstelle - Entwicklung
    - GSI Schnittstelle - Tests und Verknüpfung mit SRM
    - CQL/OpenSearch - Entwicklung (evt Cheshire)
    - OAI-PMH Schnittstelle - Entwicklung
    - OAI-ORE Export - Entwicklung [auf Nachfrage mit Community]

    - Bulk Import - Entwicklung (evt koLibRI) [auf Nachfrage mit Community]

- Tests zur Interaktion mit WDF

5 Profil B - interaktive Forschungsumgebung

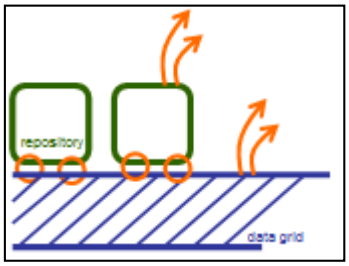


Abbildung 6 - Grid Storage Infrastruktur für Repositorien (vgl. Integrationsvariante 2, Anhang 5, WissGrid Architektur)

Dieses Anwendungsprofil wurde speziell für jene Community Grids entwickelt, deren Nutzer vor allem Daten in interaktiven Umgebungen wie z.B. Web-Anwendungen erzeugen und kollaborativ bearbeiten, deren Daten aber vertrauenswürdig und langfristig im Grid bewahrt bleiben sollen. Vor allem die Modellierung von Metadaten und die Abbildung von Forschungsprozessen und Daten-Lebenszyklen auf diese Daten sind in diesem Anwendungsprofil von Relevanz.

In diesem Profil erarbeitet WissGrid eine konkrete technische Umsetzung (auf Basis der Grid/Repository Integrationsvariante 2, siehe Abbildung 6) für Communities, die derzeit noch kein Grid-Repository besitzen.[26] Den dafür von WissGrid entwickelten Repository Software-Stack können Communities für sich in D-Grid installieren und anpassen. Organisatorische Aspekte zur Einbettung eines Repositorys in den Community-Kontext und zur langfristigen Wartung werden in diesem technischen Dokument nicht abgedeckt.

5.1 Spezifische Anforderungen

Dieses Anwendungsprofil ist vor allem für Communities mit u.a. folgenden spezifischen Anforderungen geeignet:

- Skalierbarkeit auf eine hohe Anzahl einzelner Objekte.

  Vor allem in kollaborativen Umgebungen kann eine große Anzahl kleiner Objekte und verknüpfter Erschließungsmaterialien entstehen (z.B. XML-Daten, Annotationen zu Bildern), die durch das Repository und evt. auch durch vernetzte Dienste verwaltet werden müssen. Objekte mit großen Volumina können ebenso verwaltet werden, wobei die Zugriffsgeschwindigkeit abhängig von der Schnittstelle variieren kann.

- Nahtlose Einbindung in existierende, vor allem web-basierte Nutzerumgebungen.

[26] Während die Profile A und B also die technische Basis für neue Repositorien schafft, wird in Profil C (Archiv Föderationen) ein Ansatz zur Verknüpfung von existierenden Repositorien verfolgt.

Aktuell werden für interaktive, kollaborative Nutzerumgebungen vielfach HTTP/REST-basierte Technologien genutzt, und diese müssen im Sinne der Nutzerfreundlichkeit integriert werden.

- flexible Modellierung von reichhaltigen Metadatenmodellen.

  Dazu ist auch notwendig, dass Nutzer selbst Metadatenmodelle definieren können, dass sie Metadaten und Relationen zwischen Objekten selbst zuweisen können, und dass Metadaten im Laufe des Lebenszyklus von Informationen wachsen können.

Für die Einbindung in D-Grid wird bei diesem Anwendungsprofil speziell die Integrationsvariante 2 ("Daten-Grid als Storage für Repositorien", vgl. WissGrid LZA-Architektur Anhang 5) genutzt. In der technischen Umsetzung liegt daher ein Schwerpunkt auf der Grid-Storage-Anbindung des Repositorys sowie den Web-Schnittstellen hin zu den Nutzungsumgebungen.

5.2 Aus den Fallstudien

Für wissenschaftliche Arbeit im Bereich der **germanistischen Sprachwissenschaft** ist bisher eine Datenmenge von schätzungsweise 15 Millionen Einzeltexten in Gestalt von XML und TEI-Datensätzen sowie einer vergleichbaren Menge an Annotationen angefallen. Diese Daten sollen Forschern im Rahmen der Lizenzbestimmungen in einer interaktiven Umgebung zur kollaborativen Bearbeitung zur Verfügung gestellt werden. Für eine sinnvolle Ablage der Textkorpora ist darüber hinaus ein hohes Maß an Flexibilität hinsichtlich der Implementierung von Metadatenmodellen erforderlich, werden doch sämtliche Datensätze im Zuge der Bearbeitung bereits mit einem standardisierten Metadatenkatalog versehen. Da alle Ressourcen strengen Lizenzauflagen unterliegen, ist neben der Regelung fester Zugriffsrechte und der Beschränkung der physischen Vorhaltung der Daten auf ausgewählte Zentren auch die Möglichkeit des dauerhaften Löschens einzelner Texte prioritär.

WissGrid kann die germanistischen Sprachwissenschaften bei der Errichtung eines flexiblen Grid-Repositorys unterstützen. Speziell Profil B der Repository-Arbeiten in WissGrid bietet die technische Basis für ein interaktives Repository nach den Maßgaben der Community.

Hintergrund und Details zu Kontext und Anforderungen der germanistischen Sprachwissenschaft finden sich im WissGrid Architekturdokument (Anhang Fallstudien).

5.3 Schnittstellen und Funktionalität

In diesem Abschnitt werden die Anforderungen für das "Profil B" auf technische Schnittstellen und grundlegende Funktionalitäten abgebildet. Die technische Umsetzung wird dann im folgenden Abschnitt beschrieben.

Wie in Kapitel 2 beschrieben, stellen die CRUD-Schnittstellen die Basis für die Einbettung des Repositorys in D-Grid dar.

- Create
    - *Funktion:* Ingest (OAIS): ggf. Community-spezifische Validierungen, umwandeln in internes Format (AIP - Archival Information Package)
    - *technische Schnittstelle:* HTTP/REST
- Read
    - *Funktion:* HTTP/Web-Zugriff auf die Objekte bzw. auf Kontext-spezifische Ableitungen von den Objekten; Zugriff auf OAI-ORE
    - *technische Schnittstelle:* HTTP/REST
- Update
    - *Funktion:* vgl. Ingest (OAIS), abhängig von Preservation Policy (z.B. Überschreiben nicht möglich, nur Einspielen einer neuen Version), und nach Überprüfung der Rechte
    - *technische Schnittstelle:* (siehe "Create")
- Delete
    - *Funktion:* Möglichkeit zum Löschen von Objekten abhängig von Preservation Policy bzw. Rechten
    - *technische Schnittstelle:* HTTP/REST

Neben CRUD-Schnittstellen sind in Kapitel 2 zur Verknüpfung der WissGrid LZA-Dienste mit einem Repository einige Schnittstellen und Funktionalitäten identifiziert worden:

- Filter für Abfragen
    - *Funktion:* Abfrage von Objekten nach Metadaten (z.B. alle Objekte von Autor XY, oder alle Objekte im Format YZ)
    - *technische Schnittstelle:* CQL oder OpenSearch

- Trust Level
    - *Funktion:* Ausgabe des Trust Levels (Minimalanforderungen bei externer Verarbeitung eines Objektes) als Teil der Metadaten
    - *technische Schnittstelle:* in Metadaten des Objekts

Folgende weitere Funktionalitäten könnten potenziell die Einbettung des Repositorys in Web-Umgebungen erleichtern und erweitern. Deren Umsetzbarkeit ist noch zu prüfen:

- Interoperabilität mit X.509 Grid Zertifikaten
    - *Funktion:* Nutzer hat die Möglichkeit ihr/sein Zertifikat hochzuladen, und aus dem Repository heraus wissenschaftliche Workflows anzustoßen (Verknüpfung Profil A)
- Einbettung von Web Services (SOAP/REST)
    - *Funktion:* Objekte können aus dem Repository direkt in wissenschaftliche Workflows eingebettet werden, evt. sogar über existierende Workflow-Engines
- Verbreitung von öffentlichen Repository-Inhalten
    - *Funktion:* Möglichkeit zur Weiterleitung an und Vernetzung mit relevanten Portalen (med.info, arXiv, Driver, Europeana) über OAI-PMH.

5.4 Auswahl der Technologien und technische Umsetzung

Die "technische Umsetzung" wird im Projektverlauf fortgeschrieben bis zur abschließenden Dokumentation der Produkte.

Obwohl iRODS inzwischen auch ein Web Interface[27] anbietet, sind die Möglichkeiten mit iRODS für viele web-basierte Forschungsplattformen unzureichend entlang zweier Dimensionen: Offenheit und Mächtigkeit der Web Umgebung, sowie Komplexität der verarbeitbaren Objekte. Existierende web-basierte Repositorien gehen entlang beider Dimensionen weit über die Funktionalitäten von iRODS und anderen grid-basierten Systemen

[27] iRODS Web Interface. https://www.irods.org/index.php/iRODS_Browser

hinaus.[28] Die Entwicklung von Repositorien mit Web Portalen - wie z.B. DSpace, Fedora, EPrints - wird international von einer aktiven Community im Rahmen der OpenRepositories[29] vorangetrieben.

Die Nutzung von Daten-Grids durch Web-orientierte Repositorien (nach Integrationsvariante 2, "Daten-Grid als Storage für Repositorien") wurde bereits erprobt. Eines der ersten diesbezüglichen Systeme war die Kombination von DSpace mit dem Storage Resource Broker (dem Vorgänger von iRODS). [30] Obwohl im Rahmen der OpenRepositories Cloud Technologien stärker diskutiert werden,[31] gibt es diverse Projekte, die an der Schnittstelle zwischen iRODS und Fedora[32] arbeiten. Das größte bekannte Projekt ist dabei ADONIS[33], das iRODS als nationale Speicherinfrastruktur nutzt, und Nutzern über Fedora Zugriff auf die Daten gibt bzw. in Fedora die Metadaten und Objektmodelle verwaltet. Auch direkt vom iRODS Team wird diese Verknüpfung verfolgt.[34]

Nach Sichtung einer Reihe potenzieller Systeme - z.B. DSpace[35], Tupelo[36], aDORe[37] - sind die Vorteile der Nutzung durch eine starke Community für die Nachhaltigkeit eines Systems kaum aufzuwiegen. Wegen ihrer starken Community-Unterstützung wird das WissGrid "Profil B" daher auf einer iRODS/Fedora Kombination aufbauen. Derzeit gibt es zwar noch kein produktives iRODS/Fedora System, aber gemeinsam mit den internationalen Partnern scheint diese Umsetzung möglich.

Im Folgenden werden die in 5.3 beschriebenen Anforderungen auf eine iRODS/Fedora Kombination abgebildet, um fehlende Komponenten zu identifizieren. WissGrid wird existierende Vorarbeiten in ein Gesamtpaket für das "Profil B" integrieren, und wo notwendig fehlende Komponenten in Eigenentwicklung nachführen. Besonders die Konsistenz zwischen

---

[28] Forschungsinfrastrukturen wie z.B. eSciDoc (www.escidoc.org) und das Australian National Data Service ANDS (www.ands.org.au) haben sich aus diesen Gründen von Grid Technologien entfernt und arbeiten auf Basis einer Repositorien Infrastruktur.

[29] OpenRepositories. http://www.openrepositories.org/

[30] DSpace/SRB Integration Project

[31] z.B. DuraCloud. http://duracloud.org/

[32] Fedora. http://www.fedora-commons.org/

[33] ADONIS. http://www.tge-adonis.fr/

[34] Bing Zhu, Richard Marciano, Reagan Moore: Enabling Inter-repository Access Management between iRODS and Fedora. Accepted by the 4th International Conference on Open Repositories. Atlanta, Georgia, USA. May 18-21, 2009. http://smartech.gatech.edu/dspace/handle/1853/28494

[35] DSpace. http://www.dspace.org/

[36] Tupelo. http://tupeloproject.ncsa.uiuc.edu/

[37] aDORe. http://african.lanl.gov/aDORe/projects/adoreArchive/

iRODS und Fedora - technisch (Daten immer in beiden registriert) und organisatorisch (welche Aufgaben werden auf welcher Ebene behandelt) - wird dabei zu behandeln sein.

5.4.1 Schnittstelle zwischen iRODS und Fedora

Gemäß der Grid/Repository-Integrationsvariante 2 fungiert iRODS als Speicherinfrastruktur, und Fedora verwaltet die Metadaten und Objektmodelle bzw. gibt Web-Zugriff auf die Daten. Hierbei sind die Systeme nach unten und nach oben hin frei konfigurierbar: Obwohl iRODS als Storage-Infrastruktur verwendet wird, können sich Communities darin unterscheiden, ob sie eine offene iRODS Zone verwenden oder eine separate, geschlossene Zone aufsetzen, die die spezifischen Anforderungen für Vertrauenswürdigkeit (vgl. Anforderungen Bit Preservation und Sicherheit) umsetzt. Nach oben hin zum Nutzer-Interface eröffnet Fedora sehr breite Möglichkeiten zum Aufsetzen von Forschungsumgebungen, Katalogsystemen, oder anderen Applikationen (z.B. aufbauend auf Fedora-Frameworks wie Fez, Muradora, oder eSciDoc).

Die technische Schnittstelle zwischen iRODS und Fedora kann auf unterschiedlichen Datenprotokollen basieren:

- FUSE
  https://www.irods.org/index.php/iRODS_FUSE
- WebDAV
  https://projects.arcs.org.au/trac/davis/wiki/WikiStart
- Apache VFS

  http://www.omii.ac.uk/wiki/CommonsVFSExtensionsForGrids
- JSAGA, Java implementation of the Simple API for Grid Applications
  http://grid.in2p3.fr/jsaga/

In ersten Tests schien die Implementierung von FUSE in iRODS und Fedora am weitesten fortgeschritten, und die schnelle Umsetzung eines Prototyps erwiesen sich als vielversprechend. Weitergehende Tests haben aber Probleme mit größeren Dateien zu Tage gebracht, die derzeit noch nicht gelöst werden konnten. Obwohl für Apache VFS und JSAGA noch substanzielle Implementierungen sowohl in iRODS als auch Fedora nötig wären, versprechen diese Ansätze daher derzeit die stabilere Lösung.

Die Rechteverwaltung wird aus heutiger Sicht von Fedora übernommen. Obwohl derzeit sowohl für Fedora[38] als auch für iRODS[39] Erweiterungen für eine Rechteverwaltung mittels XACML/SAML entwickelt werden, wird ein zu großer Performance-Verlust durch eine tiefer gehende Synchronisierung der beiden Rechteverwaltungen erwartet. Für den Web-Nutzer wird die Rechteverwaltung allerdings transparent sein. Auch hierüber werden weiterführende Tests im Laufe von WissGrid Aufschluss geben.

5.4.2 HTTP/REST Schnittstelle

Fedora bietet eine HTTP/REST Schnittstelle[40] und engagiert sich auch in der Umsetzung von Standard Interfaces wie SWORD[41]. Hier sind derzeit keine Weiterentwicklungen im Rahmen von WissGrid notwendig.

5.4.3 Retrieval-Interface (CQL oder OpenSearch)

Generische Standards für Suche und Filterformate sind - wie oben beschrieben - CQL und OpenSearch. Fedora bietet OpenSearch als Ausgabeformat an, aber derzeit keine CQL-basierte Suche.

Im Rahmen des BMBF-Projektes eSciDoc wurde eine CQL-basierte Suchschnittstelle implementiert.[42] Ob und wie diese Entwicklungen übernommen werden können, muss evaluiert werden.

5.4.4 Einbettung von Web Services

Fedora bietet neben der HTTP/REST Schnittstelle (siehe Abschnitt 5.4.2) auch eine SOAP-Schnittstelle.[43] Die Verknüpfung von Repository-Diensten mit Workflow Engines wird z.B. im myExperiment-Projekt evaluiert.[44]

---

[38] Shibboleth-Erweiterungen für Fedora:
* http://www.educause.edu/Resources/ShibbolethIdentityManagementan/163727
* http://www.ramp.org.au/drama/documents/extended-abstract.pdf
* im Rahmen des eSciDoc-Projektes umgesetzt

[39] ASPIS. http://www.jisc.ac.uk/whatwedo/programmes/einfrastructure/aspis.aspx

[40] Fedora API. http://www.fedora-commons.org/definitions/1/api/

[41] Simple Web-service Offering Repository Deposit, SWORD. http://www.swordapp.org/

[42] Michael Hoppe, Matthias Razum: A SRW/U-compliant Search Service for Fedora. In: Proceedings of the OpenRepositories 2008. http://pubs.or08.ecs.soton.ac.uk/106/1/submission_133.pdf

[43] Fedora Web Service Interfaces.
http://www.fedora-commons.org/confluence/display/FCR30/Web+Service+Interfaces

5.4.5 Verknüpfung mit Profil A

Obwohl die beiden Profile A und B auf iRODS aufsetzen, ziehen sich mögliche Synergien aber nicht durch alle Schnittstellen. Wichtig ist vor allem anzumerken, dass es beim derzeitigen Stand des Wissens in dieser Konfiguration nicht möglich ist, Profil A und Profil B direkt zu verknüpfen. Dies betrifft mehrere Punkte:

- Die GSI-Anbindung und Einbettung von Daten in iRODS in Grid Workflows ist nur möglich, wenn auch der Nutzer x.509 Zertifikate zur Verfügung stellt. Der entsprechende Mechanismus über Fedora existiert noch nicht.
- Die Daten werden durch Fedora in einem speziellen METS Container abgelegt, der nicht unähnlich zum OAI-ORE Format ist. Daher ist es nicht möglich durch GSI-FTP direkt auf die Daten (als Files) zuzugreifen, sondern es müsste ein eigener File-Export geschrieben werden.
- Die Verwaltung der Zugriffsrechte erfolgt in Fedora über SAML/XACML. Auch wenn in Projekten (unabhängig vom Kern-iRODS-Team) Shibboleth in iRODS eingebettet wird,[45] und damit auch eine SAML-Schnittstelle in iRODS aufgebaut wird, müssten doch die beiden Rechte-Management-Systeme von iRODS und Fedora explizit parallel aktiviert sein und ständig synchronisiert werden. Ob dies über iRODS Rules und Fedora Disseminators effizient gemacht werden kann, muss erst getestet werden.

Aufgrund dieser Ungewissheiten und des relativ hohen Aufwandes entsprechende Experimente durchzuführen, wird WissGrid zwar diesbezügliche externe Aktivitäten verfolgen und wo möglich begleiten, kann aber selbst keine Lösung der Situation versprechen.

5.4.6 Verknüpfung mit Profil C: OAI-PMH Schnittstelle

Für Fedora gibt es eine OAI-PMH Schnittstelle.[46]

---

[44] De Roure, D. and Goble, C. (2009) Lessons from myExperiment: Research Objects for Data Intensive Research. In: eScience Workshop 2009, October 15-17, 2009, Pittsburgh, US. (Submitted) http://eprints.ecs.soton.ac.uk/17744/

[45] ASPIS - Architecture for a Shibboleth-Protected iRODS System. http://www.jisc.ac.uk/whatwedo/programmes/einfrastructure/aspis.aspx

[46] Fedora: OAI Provider Service. http://www.fedora-commons.org/confluence/display/FCSVCS/OAI+Provider+Service+1.2

5.5 Überblick über Aufgaben für Profil B

- Anleitung zum Aufbau eines Repositorys nach Profil B: primär als technische Anleitung, aber auch unter Erwähnung organisatorischer Aspekte zur Nutzung als Blaupausen von Fachberatern, AP2:
    - Design und Umsetzung: Daten-Modelle, Einbettung in Community Workflows
    - Planung Hardware-Infrastruktur, Personal (intern/extern?)
    - Migration existierender Tools und Daten
- Verknüpfung iRODS / Fedora
    - Evaluation (vor allem Apache VFS und JSAGA), Implementierung, und Tests
- Schnittstellen Fedora
    - CQL/Opensearch - Entwicklung (evt eSciDoc)
- Tests zur Interaktion mit WDF

6 Profil C - föderierte Archive

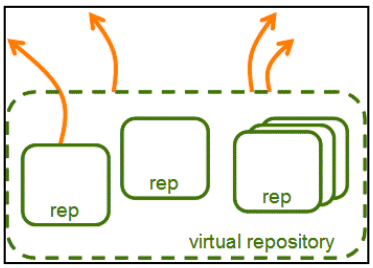


Abbildung 7 - Repository-Virtualisierung (vgl. Integrationsvariante 3, Anhang 5, WissGrid Architektur)

In Communities, die bereits über ein oder mehrere Repositorien verfügen, können diese (aus organisatorischen, finanziellen und anderen Gründen) nicht durch ein „Grid-Repository“ (vgl. Profil A und B) ersetzt werden. Nach dem Muster des Grid Paradigmas und mit den Technologien aus dem Repositorien Umfeld[47] können diese bestehenden Archive aber untereinander verbunden werden, so dass die (eigentlich verteilten) Bestände wirken, als lägen sie in einem einzigen virtuellen Gesamtarchiv (Integrationsvariante 3, Anhang 5, WissGrid Architektur). Auch die Einbettung der Objekte in Repositorien in wissenschaftliche Workflows im Grid kann durch dieses Profil gefördert werden.

Die Vertrauenswürdigkeit dieser Integrationsvariante bzgl. LZA ergibt sich aus der Vertrauenswürdigkeit der einzelnen Datenarchive. Die Virtualisierungsschicht (Föderation) kann zudem die vorhandenen LZA-Strategien durch die Bereitstellung von archivübergreifenden Diensten unterstützen. Dieses Profil skizziert eine Situation, wie sie in einigen technisch fortgeschrittenen Communities existiert. Es bietet keine generische technische Umsetzung, da diese vom technischen Kontext in der Community abhängt. Die folgenden Hinweise und technische Anleitung zur Föderierung von existierenden Repositorien (auf Basis der Grid/Repository Integrationsvariante 3, siehe Abbildung 7) können aber für die Umsetzung in einer Community herangezogen werden.

6.1 Spezifische Anforderungen

Dieses Anwendungsprofil ist speziell für Communities geeignet, in denen bereits ein oder mehrere Repositorien mit Datei-basierten Daten und Metadaten (digitale Objekte) bestehen. Die Eigenschaften einer Repository-Föderation (die teilweise je nach Sichtweise gleichzeitig als Vor- oder Nachteile gesehen werden können) beinhalten:

- Die Archive bleiben als selbständige Entitäten bestehen und werden in ihrer Verantwortlichkeit für spezielle Daten nicht berührt. Der politisch-administrative Aufwand ist daher gering.

[47] z.B. Open Archives Initiative. http://www.openarchives.org/
Open Repositorys. http://openRepositories.org/

- Der Kreis der Datenanbieter kann relativ einfach erweitert werden, da der Aufwand für die Datenprovider für ihre Gridanbindung überschaubar ist. Gleichzeitig erreichen Archive mit ihrer Teilnahme am Grid eine große und wachsende Gruppe von potentiellen Interessenten für ihre Daten.
- Auch sehr heterogene Datenquellen können mit diesem Konzept relativ einfach miteinander verknüpft werden, da die Spezifika der einzelnen Archive weiterhin erhalten bleibt. Die Nutzer profitieren daher besonders bei Fragestellungen, die sehr unterschiedliche Daten miteinander in Beziehung bringen.
- Die Archive agieren weiterhin eigenständig. Daher können evtl. unterschiedliche Policies in der Langzeitverfügbarkeit oder der Versionierung von Daten Probleme bereiten.
- Änderungen an den zentralen Schnittstellen müssen jeweils bei den Archiven möglichst zeitnah nachgeführt werden (Konsistenz).
- Daten haben in diesem Szenario im Grid selbst eine nur begrenzte Lebensdauer. Die Aspekte der Langzeitarchivierung selbst bleiben in der Verantwortung der Archive.

Für die Einbindung in D-Grid wird bei diesem Anwendungsprofil speziell die Integrationsvariante 3 ("Repository-Virtualisierung", vgl. WissGrid LZA-Architektur, Anhang 5) genutzt.

6.2 Aus den Fallstudien

Die beschriebene Verbindung von unterschiedlichen Forschungsarchiven zu einer integralen Arbeitsumgebung ist beispielhaft im Collaborative Climate Community Data and Processing Grid (**C3Grid**)[48] umgesetzt worden. Dabei werden wesentliche Datenarchive der deutschen Klimaforschung vernetzt. Der Wissenschaftler erhält dadurch einen einheitlichen Zugriff auf die Daten der beteiligten Einrichtungen, unabhängig von deren konkreten Spezifika der Ablage im Datenarchiv. Ausgangspunkt waren die schon bestehenden Datenarchive in der Klimacommunity, die aber wegen der jeweils sehr spezifischen Fragestellungen recht unterschiedlich aufgebaut sind. Für den Wissenschaftler bedeutet dies, dass er zum Einen nur schwer einen Gesamtüberblick über alle verfügbaren Daten erhalten kann und zum Anderen die Modalitäten für den Zugriff auf die jeweiligen Daten jeweils spezifisch für das angefragte

[48] Collaborative Climate Community Data and Processing Grid (C3Grid). http://www.c3grid.de

Archiv sind und damit einen Mehraufwand für den Nutzer bedeuten. Die Schaffung eines neuen einheitlichen Datenablagesystems wäre wegen der Vielzahl der Daten und der hohen Volumina keine akzeptable Alternative. Daher wird durch das Grid eine Abstraktionsschicht über alle Datenarchive geschaffen, die die Spezifika der Archive vor dem Nutzer verbirgt.

Hintergrund und Details zu Kontext und Anforderungen der Klimaforschung finden sich im WissGrid Architekturdokument (Anhang Fallstudien).

6.3 Schnittstellen und Funktionalität

WissGrid entwickelt keine dezidierte Software zum Aufbau von Föderationen von Forschungsarchiven, da diese zu stark auf den Kontext (technisch und organisatorisch) in der speziellen Community zugeschnitten werden müsste. Jedoch können die bisher im D-Grid gewonnenen Erfahrungen weitergegeben werden. Die eingesetzten Komponenten basieren auf den Arbeiten der Open Archives Initiative (OAI)[49] und der OpenRepositories[50] und sind zu einem großen Teil recht generisch, so dass sie sich auf andere Communities anpassen und übertragen lassen. Auch die Erfahrungen und entwickelten Systeme innerhalb von D-Grid (siehe z.B. Abschnitt 6.2) können für unerfahrene Communities herangezogen werden.

Die folgenden Punkte skizzieren die nötigen Schritte zum Aufbau von Repository-Föderationen.

1. Einigung auf ein **einheitliches Metadatenprofil** zwischen allen beteiligten Archiven: Dies muss nicht zwangsläufig darin münden, dass danach alle Archive intern auch ein identisches Metadatenschema verwenden, da insbesondere bei der Verknüpfung von sehr heterogenen Datenquellen sinnvollerweise nicht alle Spezifika der einzelnen Beschreibungen in das gemeinsame Profil übernommen werden können. Vielmehr muss eine möglichst große gemeinsame Schnittmenge gefunden werden, so dass alle Datenprovider ihre Beschreibungen möglichst automatisiert auf das gemeinsame Metadatenprofil abbilden können (z.B. mit Hilfe von Thesauren). Soweit als möglich sollten dabei natürlich vorhandene Standards beachtet werden, falls sie in der jeweiligen Communities bereits definiert sind.
2. **zentraler Dateninformationsdienst** zur Suche nach Datenobjekten: Die Metadateninformationen müssen für den Nutzer über (zentrale oder verteilte)

[49] Open Archives Initiative, OAI. http://www.openarchives.org/

[50] Open Repositories. http://openrepositories.org/

Metadatenkataloge durchsuchbar sein. Dafür können übliche Standards wie OAI (Open Archives Initiative) genutzt werden. Die Metadaten der einzelnen Archive werden dabei über das OAI Protocol for Metadata Harvesting (OAI-PMH)[51] jeweils zusammengefasst und können dann indiziert werden. Von Seiten der Archive ist dazu ein entsprechender OAI-Server zu installieren.

3. Definition einer generischen **Zugriffs-Schnittstelle** zwischen Archiv und Grid-Umgebung: Die meist sehr spezifische Ablage der Daten in den Archiven kann durch die Definition einer möglichst generischen Schnittstelle für den Datenzugriff vor dem Nutzer verborgen werden. Von der Grid-Seite aus werden die Daten immer in einheitlicher Weise angesprochen, die Datenprovider übernehmen dann die Umsetzung des Datenrequests auf ihre konkrete Speichermimik. Wünschenswert ist dabei, dass nicht nur die angeforderten Daten, sondern auch gleich die zugehörigen Metadaten an den Nutzer geliefert werden. Dies ist insbesondere dann von Bedeutung, wenn nach einer eventuellen Modifikation der Daten diese wieder in die Langzeitarchivierung abgelegt werden sollen.

4. temporäres **Datenmanagement** im Grid: Die Kontrolle über die im Grid befindlichen Daten obliegt einem Datenmanagementsystem: Solange die Daten im Grid bekannt sind und dort bearbeitet werden, sorgt es für das Staging der Daten von den entsprechenden Providern und für deren Transport zu den jeweils prozessierenden Sites. Dabei werden die geltenden Zugriffsrechte im Grid abgebildet. Dabei sind die Anwendungsprofile der jeweiligen Community (z.B. hinsichtlich der Zahl der Datenzugriffe, deren Volumina und Caching-Möglichkeiten) zu beachten. Sie entscheiden wesentlich darüber, welche vorhandenen Datenmanagementsysteme verwendet werden können und inwiefern community-spezifische Erweiterungen notwendig sind.

5. Definition einer generischen **Ablage-Schnittstelle** zwischen Grid und Archiv: Um Daten aus der Arbeitsumgebung des Wissenschaftlers in die Langzeitarchivierung zu transportieren, muss der für die Archive jeweils gültige Ingest-Workflow durchlaufen werden, der üblicherweise Schritte wie Metadatenextraktion und Formatvalidierung enthält. Dabei können wesentliche Voraussetzungen schon automatisiert erfüllt werden, wenn die notwendigen Metadaten bereits im Grid verfügbar gemacht werden. Eventuell muss dann noch das Grid-Metadatenprofil lokal durch weitere Angaben ergänzt werden. Beim Ingest-Workflow sind Daten mit längerer Verfügbarkeit auch mit persistenten

[51] OAI Protocol for Metadata Harvesting - OAI-PMH.
http://www.openarchives.org/OAI/openarchivesprotocol.html

Identifiern zu versehen. Dabei sollte möglichst auch schon eine fachübergreifende Verwendbarkeit der Daten in Betracht gezogen werden.

Mit diesen Voraussetzungen können vorhandene Archive durch Grid-Technologie miteinander verbunden werden, so dass für den Wissenschaftler eine homogene Arbeitsumgebung entsteht, in der Daten aus unterschiedlichen Quellen miteinander verknüpft werden können und eine glatte Verbindung zwischen Langzeitarchivierung und Grid besteht. Das föderierte System von Repositorien bietet dann folgende Funktionalität an:

- einheitliche Suche über alle verknüpften Datenbestände hinweg

  Mit dem zentralen Dateninformationsdienst liegen alle dazu notwendigen Informationen in dem Metadatenkatalog vor. Bei der Konzeption des Katalogs ist die Mächtigkeit des in der Community eingesetzten Metadatenprofils zu beachten. U.U. ist nur ein subset der Metadateninformationen für eine allgemeine Suche notwendig („search metadata“).

- einheitlicher Zugriff auf alle Daten

  Nutzerseitig werden im Grid die Parameter für den Datenzugriff einheitlich vorgegeben, unabhängig davon, wie die Daten nun konkret in dem jeweiligen Datenarchiv abgelegt sind. Die gemeinsame Datenzugriffsschnittstelle sorgt dann für die Umsetzung des Requests auf die jeweilige archivspezifische Zugriffsmimik. Beipielimplementierungen für den Zugriff auf unterschiedliche Datenbanksysteme und Flat File Archives sind vorhanden und können nachgenutzt werden.

6.3.1 Verknüpfung mit Profil A und B

Das Konzept zur Verknüpfung von einzelnen Archiven kann übertragen werden. So können die im Profil A und B skizzierten Repositorien mit analogen Methoden als neue Komponenten in der Föderation integriert werden.

6.4 Überblick über Aufgaben für Profil C

- Anleitung zum Aufbau von Repositorien-Föderationen primär als Anleitung für Community-spezifische Implementierungen

  - Design und Umsetzung: Daten-Modelle, Einbettung in Community Workflows
  - Planung Kompetenzen und Personal
  - Überblick über existierende Komponenten (z.B. Dateninformationsdienst, Mediation PID-Systeme)
- Interaktion mit WDF
  - Evaluation und Empfehlung von einheitlichen CRUD-Interfaces

7 Anhänge

7.1 Anhang 1: Glossar

Bitstream Preservation: Bitstream Preservation ist eine Form der Langzeitarchivierung, die darauf zielt, dass jedes Bit eines Datenobjekts ohne unbeabsichtigte Veränderungen verfügbar ist. Sie begegnet so zum Beispiel dem Verfall der Speichermedien und Speichertechnologien und beinhaltet Aktivitäten wie regelmäßige Integritätstests und das Anlegen von verteilten und unabhängigen Kopien. In einem Grid-Umfeld ist die Schaffung spezifischer "Trust Zones" denkbar, in denen ausgewählte Storage-Ressourcen besondere Qualitäts- und Sicherheitskriterien erfüllen. (Siehe auch D3.1 "Generische Langzeitarchivierungsarchitektur für D-Grid", Anhang 3.)

Charakterisierung: siehe Formatcharakterisierung

Content Preservation: Content Preservation ist eine Form der Langzeitarchivierung, die darauf zielt, die technische Nutzbarkeit von Daten zu erhalten. Sie umfasst Aktivitäten wie eine kontinuierliche Beobachtung der Technologieentwicklung, technische Qualitätskontrollen und Erhaltungsmaßnahmen wie Formatkonvertierungen/Migrationen oder die Bereitstellung von Emulatoren. (Siehe auch D3.1 "Generische Langzeitarchivierungsarchitektur für D-Grid", Anhang 3.)

Data Curation: Eine Form der Langzeitarchivierung, die darauf zielt, die intellektuelle Nutzbarkeit von Daten zu erhalten. Sie umfasst Aktivitäten wie die Konzeption von Daten und Metadaten, Versionierung von Objekten, Bereitstellung von notwendigen Hintergrund- und Kontextinformationen, etc. (Siehe auch D3.1 "Generische Langzeitarchivierungsarchitektur für D-Grid", Anhang 3.)

Digitale Objekte: Digitale Objekte sind digitale Daten, die als intellektuelle Einheiten aus (einer oder mehreren) Dateien, zugehörigen Metadaten sowie einem Netzwerk aus anderen Objekten bzw. referenzierbaren Informationen bestehen können. Objekte können alle Arten von Daten umfassen - strukturiert, semi-strukturiert (z.B. XML-basiert) oder unstrukturierte Daten wie z.B. Bilder oder Videos. Um sie explizit zu beschrieben, können so genannte Paketformate benutzt werden, die zugehörige Metadaten (z.B. deskriptiv,

administrativ, Audit Trails) wie auch Relationen zu anderen Objekten und externen Erschließungsmaterialien enthalten.

Formatcharakterisierung: Ein Dienst zur Formatcharakterisierung extrahiert aus digitalen Objekten Metadaten wie z.B. die eindeutige Bezeichnung des Formats und der Formatversion. Die Metadaten können technischer Natur sein, wie z.B. Auflösung und Farbrauminformationen bei Bildformaten oder Erstellungssoftware und -hardware, aber auch deskriptive Metadaten, die das Objekt intellektuell beschreiben, wenn sie eingebettet sind. Metadaten sind notwendig, um Daten effektiv verwalten und nutzen zu können.

Formatkonvertierung: Dienste zur Formatkonvertierung überführen digitale Objekte von einem Format möglichst verlustfrei in ein anderes. Damit ist es nicht nur möglich wichtige Daten in veralteten Formaten durch die Umwandlung in aktuelle Formate nutzbar zu halten, sondern auch der Datenaustauch kann durch die Anpassung von Daten an fremde Schnittstellen und Software erleichtert werden.

Formatvalidierung: Formatvalidierung ist die Prüfung eines digitalen Objekts auf seine technische Korrektheit, ob die syntaktischen und ggf. semantischen Vorschriften des Formats eingehalten werden, und stellt einen Teil einer Qualitätssicherung dar.

Forschungsdatenarchiv (bzw. synonym "Forschungsarchiv"): Ein Forschungsarchiv umfasst Technik und Organisation (z.B. Betrieb, Finanzierung, Verantwortlichkeiten). Dazu passt es die generischen Funktionalitäten von Repositorien an den spezifischen Kontext einer Community an (z.B. Anwendungsszenarien, organisatorischer Rahmen). Vor allem Vertrauenswürdigkeit und Langzeitarchivierung von Objekten bauen zwar auf die technische Basis von Repositorien und LZA-Diensten, können letztlich aber nur durch darüber liegende organisatorische Maßnahmen gewährleistet werden (z.B. finanzielle Stabilität, Rollen für Preservation Planning und Audit[52]).

Langzeitarchivierung (LZA): Alle Aktivitäten, die darauf abzielen die Nutzbarkeit digitaler Daten angesichts eines sich verändernden Kontextes (zeitlich, technisch, intellektuell, etc.) zu erhalten. Formen der Langzeitarchivierung sind Bitstream Preservation, Content Preservation und Data Curation. (Siehe auch D3.1 "Generische Langzeitarchivierungsarchitektur für D-Grid", Anhang 3.)

[52] Research Libraries Group. (2002). Trusted digital repositories: Attributes and responsibilities. An RLG-OCLC Report. http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

Metadatenextraktion: siehe Formatcharakterisierung

Konvertierung: siehe Formatkonvertierung

Provenienzdienst: Ein Provenienzdienst erfasst und speichert Informationen über Prozesse, die Daten verarbeiten und verändern. Er ermöglicht es dadurch die Verarbeitungsschritte langfristig nachvollziehbar zu dokumentieren, um die Authentizität von Daten später bewerten zu können. Es handelt sich dabei um eine Querschnittsfunktionalität, die z.B. in der Middleware implementiert sein muss.

Repository: Softwaresystem zur Verwaltung von digitalen Objekten. Neben der Verwaltung von digitalen Objekten (speichern, abrufen, verändern bzw. neue Versionen anlegen) bieten Repositorien auch zumeist generische Mechanismen zur Einbettung der Objekte in wissenschaftliche und interaktive Workflows, z.B. für die kollaborative Bearbeitung von Objekten in interaktiven Editoren, oder für automatisierte wissenschaftliche Berechnungen.

Trust Zone: Gerade in der LZA ist die Integrität und Sicherheit von Daten oft besonders relevant (siehe Glossar-Eintrag zu "Vertrauenswürdigkeit"). Dies verträgt sich mitunter nicht mit der Offenheit und Verfügbarkeit von Ressourcen in einer Grid-Infrastruktur. Zur Abhilfe können "Vertrauenszonen" (Trust Zone) auf speziellen Grid-Ressourcen geschaffen werden, die durch besondere technische und organisatorische Maßnahmen einen hohen Grad an Vertrauenswürdigkeit sicherstellen können (z.B. Datensicherheit, Datenintegrität, Bit Preservation). Trust Zones können Teil des Konzeptes einer Grid-Infrastruktur wie D-Grid sein, können aber technisch auch explizit getrennt von der offenen Compute-Grid-Infrastruktur sein, wo das notwendig ist.

Validierung: siehe Formatvalidierung

Vertrauenswürdigkeit: "Eigenschaft eines Systems, gemäß seinen Zielen und Spezifikationen zu operieren (d.h. es tut genau das, was es zu tun vorgibt bzw. was seine Betreiber versprechen, dass es tut). Aus der Sicht eines Benutzers ist ein System vertrauenswürdig, wenn seine Erwartungen erfüllt werden." (Entnommen aus "Kriterienkatalog zur Prüfung der Vertrauenswürdigkeit von PI-Systemen", nestor (Hrsg.), 2009, urn:nbn:de:0008-20080710140). In der Langzeitarchivierung wird die Vertrauenswürdigkeit üblicherweise durch Kriterien geprüft, die die Wahrscheinlichkeit der Nachnutzbarkeit der Daten erhöhen.