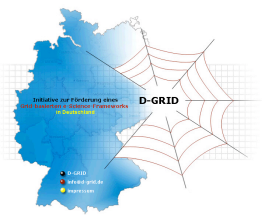
Generische Langzeitarchivierungs-architektur für D-Grid

Version - 14. Januar 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: Deliverable
Verfügbarkeit: vorläufig WissGrid-intern, Veröffentlichung im Februar 2010
Autoren:

Andreas Aschenbrenner, SUB
Frank Dickmann, UMG
Harry Enke, AIP
Bernadette Fritzsch, AWI
Michael Lautenschlager, DKRZ
Benjamin Löhnhardt, UMG
Jens Ludwig, SUB
Torsten Rathmann, DKRZ
Angelika Reiser, TUM
Florian Schintke, ZIB
Jens Stegmann, IDS
Stefan Strathmann, SUB

Revisionsverlauf

Datum	Autor	Kommentare
12.08.2009	AAsche, Jens Ludwig	Dokumentstruktur
4.9.2009	AAsche, Harry Enke, Bernadette Fritzsch, Jens Ludwig	Ergänzungen und Kommentierungen, erste AP interne Version
23.9.2009	Benjamin Löhnhardt, Jens Ludwig, Angelika Reiser, Florian Schintke, Stefan Strathmann	Ergänzungen bzgl. DGI und weitere Dienste, Use Case Medizin und LZA-Begriff, kleinere Korrekturen, Berücksichtigung von Kommentaren
18.10.2009	AAsche, Frank Dickmann, Jens Ludwig	Überarbeitung insb. Kapitel 3, Use Case Biostatistik
22.10.2009	AAsche, Harry Enke, Bernadette Fritzsch, Jens Ludwig, Jens Stegmann, Stefan Strathmann	Ergänzungen, Korrekturen, Klärungen, Grafiken: Version für Konsortiumsbegutachtung
6.11.2009	Michael Lautenschlager, Jens Ludwig, Torsten Rathmann, Jens Stegmann	Korrekturen aus der Konsortiumsbegutachtung
8.11.2009	Jens Ludwig	Erstellung der 1.0 Version
15.12.2009	AAsche	Kapitel 3.3 und 4.5 ergänzt
4.1.2010	Torsten Rathmann, Norman Fiedler	Fallstudien ergänzt
14.1.2010	Bernadette Fritzsch, Jens Ludwig	Korrekturen und Formatierung

Inhaltsverzeichnis

1. Einführung

> "Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly." [1]
>
> "Forschungsdaten müssen als Ressource begriffen werden, die sowohl für zukünftige Forschungen der eigenen als auch anderer Disziplinen wichtig ist und effizient genutzt werden sollte. [...] Um die Investitionen in die Erzeugung von Forschungsdaten voll auszuschöpfen, reicht es nicht, eine Langzeitarchivierung zu verfolgen, die als einmalige Aktivität ein statisches Ergebnis konserviert und wegschließt. Stattdessen muss eine Pflege der Daten über ihren gesamten Lebenszyklus betrieben werden [...] Neue wissenschaftliche Fragestellungen motivieren die Verwendung neuartiger Methoden und Instrumente, die z.B. durch virtuelle Forschungsumgebungen und Grid-Infrastrukturen unterstützt werden. Die Dateninfrastruktur muss damit schritthalten." [2]

Im Rahmen des WissGrid-Arbeitspaketes 3 (AP3) werden Konzepte und Werkzeuge erstellt, in D-Grid integriert und den Communities bzw. den Communty-Grids an die Hand gegeben, die zentrale Aufgaben einer Langzeitarchivierung erfüllen. Unter "Langzeitarchivierung" verstehen wir alle Aspekte, die für eine Nachnutzung von Forschungsdaten notwendig sind, selbst wenn ein langer Zeitraum zwischen der Erzeugung der Forschungsdaten und deren Nachnutzung liegt, und die ursprünglich für die Erzeugung verantwortlichen Systeme und Personen nicht mehr verfügbar sind. Wir untergliedern den Begriff Langzeitarchivierung in drei Ebenen: (1) Bit Preservation, (2) Content Preservation und (3) Data Curation. Für jede dieser Ebenen sind unterschiedliche Fähigkeiten notwendig, und je nach Anforderungen und Kontext können die Verantwortlichkeiten zwischen einer Daten-Infrastruktur und den

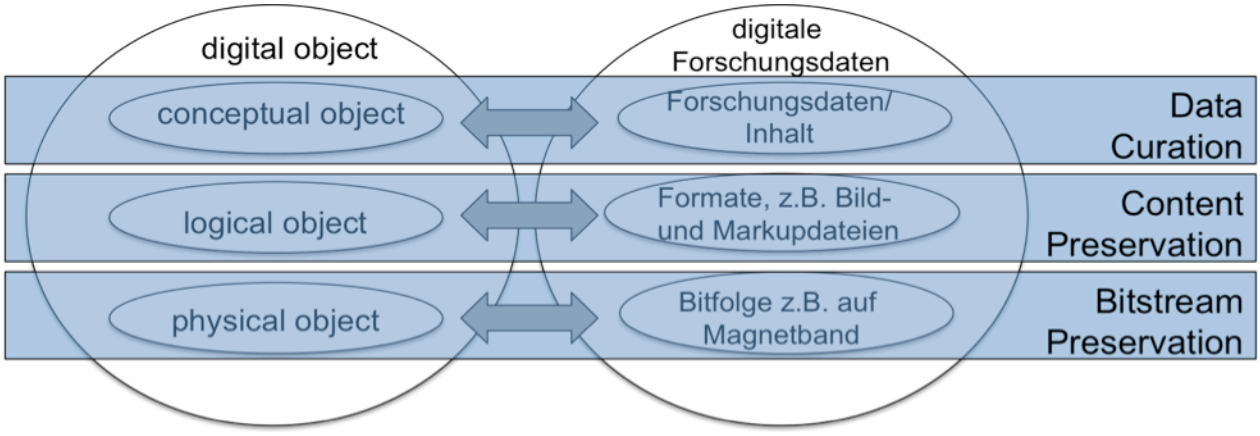


Abbildung 1: Die Ebenen der Langzeitarchivierung behandeln unterschiedliche Aspekte digitaler Objekte bzw. Forschungsdaten.

[1] Data's shameful neglect (Editorial), Nature 461, 145 (10 September 2009) | doi:10.1038/461145a, http://www.nature.com/nature/journal/v461/n7261/full/461145a.html

[2] nestor AG Grid/e-Science und Langzeitarchivierung: Digitale Forschungsdaten bewahren und nutzen – für die Wissenschaft und die Zukunft. nestor Bericht, 2009. http://nbn-resolving.de/urn:nbn:de:0008-2009071031

Communities unterschiedlich definiert werden. Eine detaillierte Diskussion zu diesen Ebenen der Langzeitarchivierung findet sich in Anhang 3, Kapitel 4.3.

Die Daten der Fach-Communities sind sehr heterogen und reichen von Klima- und Wetterdaten (nicht wiederherstellbare Daten und zum Teil Jahrhunderte alt), über aktuelle medizinische Daten (mit gesetzlichen Regelungen zur Archivierung), bis hin zu geisteswissenschaftlichen Daten (unwiederbringliche Zeugen unseres kulturellen Erbes; mit großem, manuellem Aufwand erstellt) und entsprechend unterschiedlich sind die Anforderungen. Trotz der verschiedenen Ausgangspunkte existieren in allen Communities große Datenmengen und die Notwendigkeit zur Langzeitarchivierung. Die Langzeitarchivierung/Datenpflege als Aufgabe, die den gesamten Lebenszyklus von Forschungsdaten betrifft, muss auch von den Grid-Infrastrukturen, die Teil dieses Lebenszyklus sind, unterstützt werden.

Und obwohl das Gebiet Langzeitarchivierung primär unabhängig von einer Grid-Infrastruktur existiert, so können doch einige Teilaspekte mit Hilfe von Gridtechnologien effizienter oder für den Nutzer transparenter gestaltet werden. Außerdem ist zu erwarten, dass durch die Etablierung kollaborativer, IT-basierter Arbeitsumgebungen in den Communities der Bedarf an einer engeren Anbindung von Datenarchiven an das Grid steigt: Daten, die im Grid bearbeitet und erzeugt werden, sollen auch „grid-nah“ abgelegt und für andere Nutzer verfügbar gemacht werden.

Jenseits der einzelnen Anwendungskontexte bauen multidisziplinäre Initiativen wie das Australian National Data Service (ANDS)[3], UKRDS (UK Research Data Service)[4] oder DataNet in den USA[5] auf die gemeinsamen Grundbedürfnisse und Synergien zwischen den Communities. In einzelnen Bereichen wie dem Aufbau und der kontinuierlichen Wartung von Format-Registern, kann eine einzelne Community allein kaum langfristig die Kosten der Dienste tragen.[6]

Dieses Dokument beschreibt eine modulare, generische Referenzarchitektur, die die notwendigen, recht unterschiedlichen Anforderungen abdeckt und Synergien zwischen den Communities ermöglicht. Es ist eine offene Architektur, sodass existierende Dienste iterativ erweitert und der aktuellen technologischen Umgebung angepasst werden können, sowie neue Dienste in die Architektur integriert und genutzt werden können. Dies ist nicht zuletzt deshalb notwendig, weil Langzeitarchivierung (LZA) im Moment zwar mit den aktuellen internationalen Erkenntnissen und Entwicklungen umgesetzt werden kann, sie aber nie „abgeschlossen“ sein kann und einzelne Komponenten immer wieder an den wissenschaftlich/technischen Fortschritt angepasst werden müssen.

Die Aufgabe von WissGrid liegt weniger in der Entwicklung neuer Technologien als vielmehr in der Anpassung der international vorhandenen Technologien und Konzepte an die speziellen und unterschiedlichen Anforderungen der Grid-Communities. Dieses Dokument entwickelt die dafür nötige generische und Community-übergreifende technische Referenz-Architektur.

---

[3] Australian National Data Service, ANDS. http://ands.org.au/

[4] UK Research Data Service, UKRDS. http://www.ukrds.ac.uk/

[5] DataNet (Sustainable Digital Data Preservation and Access Network Partners) ist mit 100.000.000 US Dollar über 5 Jahre dotiert. http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm

[6] Stephen L. Abrams and David Seaman. Towards a global digital format registry. In Proceedings of the 69th IFLA Conference, 2003. http://archive.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

Weiterhin identifiziert dieses Dokument einige grundlegende LZA-Dienste, die im weiteren Verlauf von WissGrid entwickelt, bzw. durch die Verknüpfung vorhandener Technologien für D-Grid adaptiert werden. Diese LZA-Dienste stellen keine umfassende Lösung für jeden möglichen Kontext dar, aber sie sind wesentliche Bausteine für den Aufbau von Community-spezifischen LZA-Systemen. Die detaillierte technische Dokumentation dieser LZA-Dienste erfolgt in separaten Dokumenten.

Obwohl eingebettet in ein technisches Umfeld, ist LZA doch eine primär organisatorische Herausforderung. Technische Systeme werden sich im Laufe der Zeit vielfach ändern, aber die Organisationsstrukturen, um mit diesen regelmäßigen Änderungen umzugehen, müssen stabil sein. Neben den in diesem Dokument beschriebenen technischen Arbeiten unterstützt WissGrid daher auch die organisatorischen Aspekte von LZA in den Community-Grids. Die dafür zu erstellenden "Blaupausen" werden ebenfalls in separaten Dokumenten vorgelegt.

Für die Entwicklung und Anpassung von LZA-Technologien und -Konzepten zielt WissGrid zu einem großen Teil auf die Content Preservation und zu geringeren Teilen auf die Bitstream Preservation und Data Curation. Dies sind Bereiche, die näher an der D-Grid-Infrastruktur oder stärker Community-spezifisch sind. Entsprechend müssen im D-Grid-Kontext Entwicklungen dieser Akteure für die Langzeitarchivierung herangezogen werden.

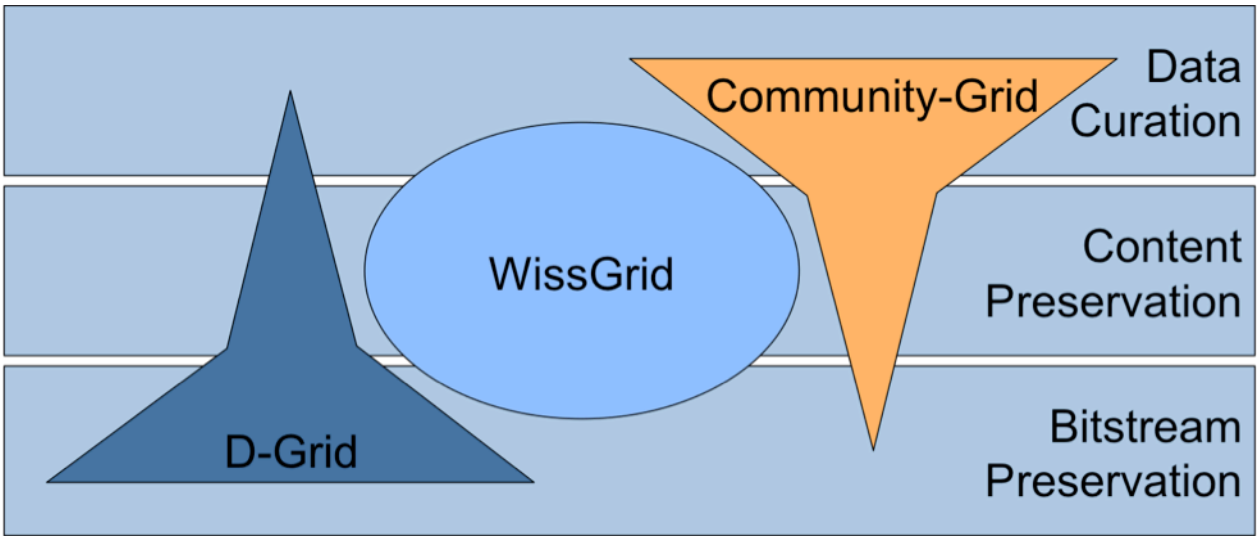


Abbildung 2: Die LZA-Entwicklungen können in unterschiedlichem Maße von D-Grid, WissGrid und den Community-Grids geleistet werden.

Von der Frage der Entwicklung ist die Frage des dauerhaften Betriebs zu unterschieden. Es mag einiges dafür oder dagegen sprechen den Betrieb analog zu der Entwicklung der Komponenten zu organisieren. Die Diskussion dieser Frage ist aber nicht Bestandteil der Architektur und wird daher in diesem Dokument nicht behandelt.

2. Vorgehensweise

Die Entwicklung der in diesem Dokument dargestellten LZA-Architektur basiert auf drei wesentlichen Quellen: (1.) internationalen Entwicklungen im Bereich LZA, (2.) den Erfahrungen der wissenschaftsbezogenen Communities in D-Grid und (3.) intensiver Diskussion mit neuen Communities.

1. Spätestens seit dem Bericht der "Task Force on Archiving Digital Information" im Jahr 1996 werden Fragestellungen der LZA strukturiert angegangen.[7] Nach seither mehr als 10 Jahren, in denen Konzepte und Standards entwickelt wurden, arbeiten derzeit diverse Initiativen am Aufbau von nachhaltigen Infrastrukturen für die Langzeitarchivierung (siehe Kapitel 2.1). Diese **existierenden LZA-Ansätze** bilden für WissGrid einen wesentlichen analytischen Ausgangspunkt für die Entwicklung der LZA-Architektur in D-Grid.

2. Obwohl diese vorhandenen Konzepte und Technologien ein guter Ausgangspunkt sind, ist ihre Übertragung in eine konkrete technische Anwendungsumgebung nicht trivial. Als Grid-Nutzer der ersten Stunde sind die wissenschaftsbezogenen Communities und WissGrid-Partner für diese Übertragung ideal positioniert. Alle haben bereits **langjährige Erfahrung** in der Abbildung von Community-Anforderungen und Konzepten auf Grid-Umgebungen, der kollaborativen, Community-übergreifenden Arbeitsumgebung in D-Grid sowie der Portierung von Diensten in das Grid.

3. Schließlich sind es auch die Anforderungen und Ansichten weiterer akademischer Communities als **zukünftige Nutzer**, die wesentlich für die Entwicklung der D-Grid LZA-Architektur sind. Einige der diskutierten Anforderungsprofile sind in Kapitel 4.1 kurz dargestellt. Die WissGrid LZA-Architektur wird u.a. in öffentlichen Workshops mit aktuellen und zukünftigen D-Grid Communities gemeinsam weiterentwickelt, und ist so flexibel konzipiert, dass sie auch in Zukunft an neue Anforderungen adaptiert werden kann.

2.1. Bisherige Ansätze für LZA-Architekturen

Die internationalen Entwicklungen im Bereich der Langzeitarchivierung sind für WissGrid ein wichtiger Ausgangspunkt, um von den langjährigen Erfahrungen und entwickelten Standards zu profitieren, aber auch um bereits existierende LZA-Dienste für WissGrid zu adaptieren bzw. Interoperabilität mit ihnen zu erreichen.

Zu den vielleicht wichtigsten Arbeiten derzeit im Bereich der LZA zählen der ISO Standard OAIS (Reference Model for an Open Archival Information System)[8] sowie die Arbeiten zur Definition von Kriterien für "Trusted Digital Repositories"[9]. Neben diesen und anderen eher

[7] RLG/CPA Task Force on Archiving of Digital Information: Preserving Digital Information: Final Report and Recommendations. May 1996.
http://www.worldcat.org/arcviewer/1/OCC/2007/08/08/0000070513/viewer/file2619.html

[8] OAIS, Reference Model for an Open Archival Information System. ISO 14721:2003. http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

[9] Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist. http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/

konzeptuellen Arbeiten, gibt es eine Vielzahl von LZA-Projekten und -Initiativen. Im Rahmen der WissGrid-Architektur wurden diejenigen analysiert, die ähnliche Anforderungen wie WissGrid ansprechen und Entwürfe veröffentlicht haben:

- Die Zusammenstellung von Curation Mikroservices der California Digital Library (in: "Preservation is not a place"[10]), die wie die vorliegende Architektur eine modulare und evolutionäre Entwicklung von Diensten und nicht ein monolithisches System zugrunde legt.
- Das EU-Projekt PLANETS[11], das eine Reihe von Werkzeugen für die Langzeitarchivierung entwickelt, die in einem Interoperabilitäts-Framework integriert werden.
- Der Australian National Data Service (ANDS)[12], der eine verteilte nationale Infrastruktur für Forschungsdaten aufbaut.
- Das EU-Projekt CASPAR[13], das mit dem Anspruch theoretischer Stringenz eine auch für Forschungsdaten geeignete Architektur entwirft.

Diese Projekte und Initiativen stammen aus Europa, Australien und den USA und decken ein breites Spektrum an unterschiedlichen Kontexten und Zielen ab: einzelne Institutionen mit einer spezifischen Zielgruppe (CDL), eine nationale, multi-disziplinäre Infrastruktur (ANDS), etc. Trotz ihrer unterschiedlichen Herkunft und Ziele stimmen sie in einer Vielzahl von Komponenten und Diensten überein, die sie für eine LZA-Infrastruktur als notwendig erachten. Ein tabellarischer Vergleich der Architekturansätze findet sich in Anhang 4.4.

Es lassen sich die folgenden **Arten von Diensten** unterscheiden:

- **Archiv- und Speicherdienste:** Systeme, die Daten für die weitere Verwendung bereitstellen wie z.B. Repositorien (Fedora, http://www.fedora-commons.org/)
- **Streaming-Dienste**: deren Funktionalität in einem bestimmten, klar definierten Einsatzgebiet zur Verwendung kommt, wie z.B. Formatvalidierung (JHove, http://hul.harvard.edu/jhove/) oder Daten-Konversion (Crib, http://crib.dsi.uminho.pt/). Zu den Streaming-Diensten zählen auch Funktionalitäten, die in einem ersten Schritt weitgehend automatisch abgearbeitet werden können, anschließend aber eine manuelle oder semi-automatische Nachbearbeitung erfordern.
- **Infrastrukturdienste:** unterstützende Funktionalitäten, die von anderen Diensten benötigt werden oder die die Interaktion zwischen unterschiedlichen Diensten ermöglichen, wie z.B. Authentifizierung, Identifizierung, etc.
- **Verzeichnisdienste**: aufgrund der in ihnen aggregierten Informationen relevant, wobei diese Inhalte von einer Institution oder von der Community kollaborativ erstellt,

[10] Stephen Abrams, Patricia Cruse, John Kunze: Preservation Is Not a Place. In: International Journal of Digital Curation, Vol 4, No 1 (2009). http://www.ijdc.net/index.php/ijdc/article/view/98
[11] PLANETS - Preservation and Long-term Access through Networked Services. http://www.planets-project.eu/
[12] ANDS - Australian National Data Service. http://ands.org.au/
[13] CASPAR. http://www.casparpreserves.eu/

erweitert und öffentlich angeboten werden, wie z.B. die Format Registry Pronom (http://www.nationalarchives.gov.uk/PRONOM/Default.aspx)

- **interaktive Dienste**: werden im Rahmen der Langzeitarchivierung unter anderem zum Preservation Planning (vgl. Funktion im OAIS Modell), wie z.B. Decision Support Systeme für Preservation Planning wie Plato (http://www.ifs.tuwien.ac.at/dp/plato/)

Die LZA-Architektur konzentriert sich auf die Einbindung von Streaming-Diensten und Archiv- und Speicherdiensten in D-Grid. Bei diesen besteht zum einen für eine D-Grid-LZA-Architektur der höchste Bedarf einer Anpassung an eine Grid-Umgebung. Auf Verzeichnisdienste und interaktive Dienste können Grid-Communities üblicherweise auch in ihrer bisherigen Form zugreifen und Infrastrukturdienste werden zum überwiegenden Teil schon im Rahmen von D-Grid angeboten. Zum anderen bietet bei den Streaming, Archiv- und Speicherdiensten die Nutzung der Grid-Technologien auch den größten Nutzen.

2.2. LZA-Anforderungen der Grid-Communities

Die Fallstudien zur Anforderungsanalyse (siehe oben bzw. Abschnitt 4.1) zeigen die Breite der Anforderungen aus unterschiedlichen Communities, in denen es zum Teil noch wenige Vorarbeiten gibt, zum Teil aber auch schon fortgeschrittene Konzepte zur LZA und Datenpflege entwickelt sind. Analog suchen neue Communities in D-Grid ohne Vorarbeiten LZA-Werkzeuge zum Aufbau einer kompletten LZA-Umgebung, während etablierte Communities mit LZA-Erfahrung mögliche Synergien zwischen ihren bestehenden Strategien und Systemen sowie den LZA-Infrastrukturen anderer Communities suchen.

Unterschiedliche Anforderungen aus den Fach-Communities können auf allen Ebenen der LZA bestehen: Bit Preservation, Content Preservation und Data Curation (vgl. Kapitel 4.3). Dabei sind speziell die höheren Ebenen und vor allem Data Curation in vielerlei Hinsicht eng mit den Community-spezifischen Anwendungen verknüpft. Nutzer arbeiten direkt an den Daten, und deren Modellierung und Pflege wird im Alltag zumeist von Nutzern durchgeführt oder zumindest eng von ihnen begleitet, wenn auch mit unterschiedlichem Professionalitätsgrad. Dementsprechend haben viele Communities - wenn nicht explizit, so doch implizit - Komponenten mit Relevanz für "Data Curation" in ihre Anwendungen eingebettet.

Die LZA-Strategie und die LZA-Systeme einer Community werden daher sowohl technisch als auch organisatorisch zu unterschiedlichem Grad von den Angeboten aus WissGrid unterstützt - je nach den Anforderungen aus der Community und den vorhandenen Strategien und Systemen. Die LZA-Blaupausen (beschrieben in einem parallelen WissGrid-Dokument) unterstützen die Formulierung von LZA-Strategien, wo sie noch nicht vorhanden sind, bzw. helfen vorhandene Strategien auf ihre Vollständigkeit hin zu prüfen. Analog zu diesem organisatorischen Rahmen kann ein LZA-System je nach Notwendigkeit aus generischen und aus spezifischen LZA-Diensten für eine Community angepasst werden.

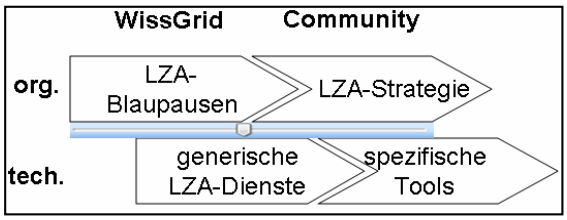


Abbildung 3: Der WissGrid- und der Community-Anteil an organisatorischen und technischen Aspekten der Langzeitarchivierung kann unterschiedlich groß sein. Communities können je nach Notwendigkeit stärker die "generischen Angebote" von WissGrid einsetzen, oder stärker auf eigene Entwicklungen setzen.

Die Anforderungen für eine LZA-Architektur in D-Grid ergeben sich auch aus dem aktuellen Umgang mit Datenpflege und dem Anwendungskontext von Nutzer-Communities. Im Rahmen von WissGrid wurden und werden fortlaufend Anforderungen in öffentlichen Workshops bzw. in Fallstudien mit potenziellen Nutzer-Communities gesammelt. Einige dieser Fallstudien sind in Abschnitt 4.1 exemplarisch aufgeführt. In ihnen werden unter anderem folgende LZA-relevante Aspekte diskutiert: Daten/Metadaten und Standards, Anwendungen und Anbindungen, organisatorischer Kontext und vorhandene Maßnahmen.

Die folgenden **generischen Anforderungen** konnten bisher aus Diskussionen mit den Community-Grids und den Fallstudien sowie aus der Analyse der LZA-Architekturen extrahiert werden:

- Modulare Plattform, so dass Dienste entsprechend der Community-spezifischen LZA-Strategie ausgewählt und zusammengestellt werden können.
- Evolutionäre Dienste, damit Dienste mit der sich ständig ändernden technischen Umgebung mitwachsen können.
- Heterogenität der Daten: unterschiedliche Datenmengen, Datenformate, Metadaten. Hier gibt es keinen Community-übergreifenden Trend; alle Varianten müssen unterstützt werden.
- Vorhaltung der Daten in Grid-Umgebung (Bit Preservation).
- Referenzierbarkeit der Daten, teilweise über Communitygrenzen hinweg, um die Nachnutzbarkeit der Daten weiter zu erhöhen.

3. LZA-Architektur für D-Grid

Dieses Kapitel beschreibt ein Referenz-Modell zur Einbindung von LZA-Diensten in D-Grid. Um den in Kapitel 2.2 beschriebenen heterogenen Anforderungen unterschiedlicher Communities gerecht zu werden, ist die im Folgenden definierte WissGrid LZA-Architektur weniger ein autarkes Komplettsystem, als vielmehr eine **modulare Toolbox von LZA-Diensten**, aus der jede Community die auf die eigene LZA-Strategie passenden Dienste zusammenstellen, vernetzen und in ihre Umgebung integrieren kann. So können auch für eine Aufgabe unterschiedliche Dienstvarianten vorhanden sein, die sich in Qualität, Anpassbarkeit oder Interoperabilität unterscheiden.

Diese modulare Zusammenstellung von LZA-Diensten hat auch den Vorteil, dass jeder Dienst an veränderte Anforderungen angepasst werden kann und jederzeit neue (generische oder Community-spezifische) Dienste in D-Grid integriert werden können. Dies stellt eine wesentliche Anforderung an eine übergreifende LZA-Infrastruktur dar, da sie sich mit dem technischen Fortschritt weiterentwickeln und sich einem veränderlichen organisatorischen Kontext anpassen können muss. Während sich also einzelne LZA-Dienste über einen Zeitraum von mehreren Jahren oder Jahrzehnten kontinuierlich ändern werden, soll das in diesem Kapitel beschriebene Referenz-Modell weitgehend stabil bestehen bleiben.

Das Ziel für die WissGrid LZA-Architektur ist eine optimale Einbindung in die D-Grid Infrastruktur - bei gleichzeitiger Offenheit gegenüber Community-spezifischen Systemen. Es muss z.B. genauso möglich sein, in C3-Grid erzeugte und im Grid verarbeitete Satellitendaten mit einem LZA-Dienst zu verarbeiten, wie auch geheime Daten einer pharmazeutischen Firma, die in einem proprietären Repository außerhalb des Grids verwaltet werden. Das bedeutet auch, dass in der offenen WissGrid LZA-Architektur neu entwickelte LZA-Dienste durch externe Spezialfirmen erzeugt werden und kommerziell angeboten werden können. Die im Rahmen des WissGrid-Projektes zu erzeugenden Dienste sind unter einer Open-Source-Lizenz und werden allen Communities in D-Grid zur freien Nutzung zur Verfügung gestellt.

WissGrid erzeugt während der Projektlaufzeit die **übergreifenden Kernfunktionalitäten** und einige ausgewählte **generische LZA-Dienste**, die so die Basis einer modularen, wachsenden Toolbox bilden. Wesentlicher Aspekt der übergreifenden Kernfunktionalitäten ist die in 3.1 beschriebene Interoperabilität und Einbindung von unterschiedlichen LZA-Diensten in die D-Grid Infrastruktur. Konkrete generische LZA-Dienste, die besonders sinnvoll und wichtig für eine LZA-Infrastruktur und weder im D-Grid-Kontext abgedeckt noch zu speziell oder umfangreich sind, wurden durch die Analyse der existierenden LZA-Architekturen (vgl. Kapitel 2.1, bzw. Anhang 4.4) identifiziert. Die meisten liegen bereits als Open-Source-Implementierung ohne Grid-Unterstützung vor. Eine detaillierte Beschreibung dieser Dienste findet sich im Anhang 2, Kapitel 4.2. Im Überblick sind es:

- Ein Charakterisierungsdienst, der Informationen wie z.B. technische Metadaten zur Interpretation und Nutzung der Daten extrahieren kann, sowie zur Validierung und Qualitätskontrolle von Daten.
- Ein Konvertierungsdienst, der die Umwandlung von Daten ermöglicht.

- Ein Forschungsdatenarchiv, das der Verwaltung von Informationsobjekten und Metadaten dient.
- Ein Provenienzdienst, der es ermöglicht Prozesse, die Daten prozessieren und verändern, langfristig nachvollziehbar zu dokumentieren, um die Authentizität später bewerten zu können.

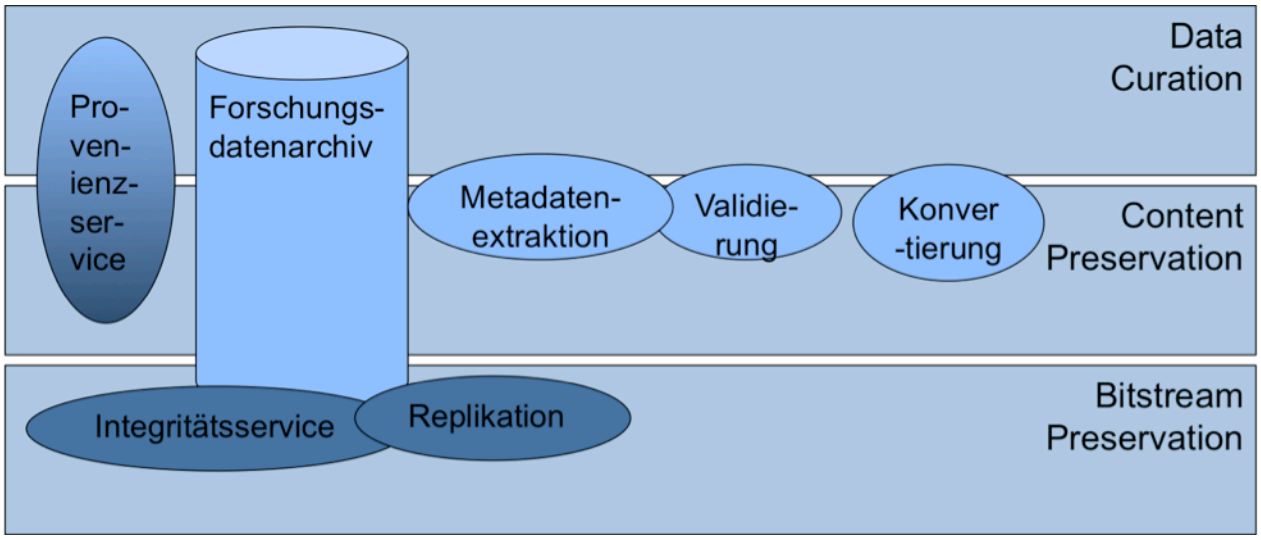


Abbildung 4: Zuordnung von LZA-Diensten zu den LZA-Ebenen. Hellblau sind Funktionalitäten im Aufgabenbereich von WissGrid. Dienste zur Integrität, Replikation und Provenienz sind ganz oder teilweise außerhalb der WissGrid-Aufgaben, da sie sehr nahe an der Infrastruktur sind. Vgl. Abbildung 2.

Alle diese Dienste und generell alle Dienstarten können in der WissGrid LZA-Architektur modular untereinander kombiniert werden. Z.B. kann ein Community-Grid mit bestehendem Datenmanagement im D-Grid lediglich einen vorhandenen Konversionsdienst in Anspruch nehmen und diesen in ihr Datenmanagement einbinden, während ein anderes Community-Grid ohne existierende Systeme ein komplettes Forschungsdatenarchiv mit automatischer Validierung der Daten beim Ingest-Prozess und ggf. automatischer Konversion in Standardformate benutzen kann.

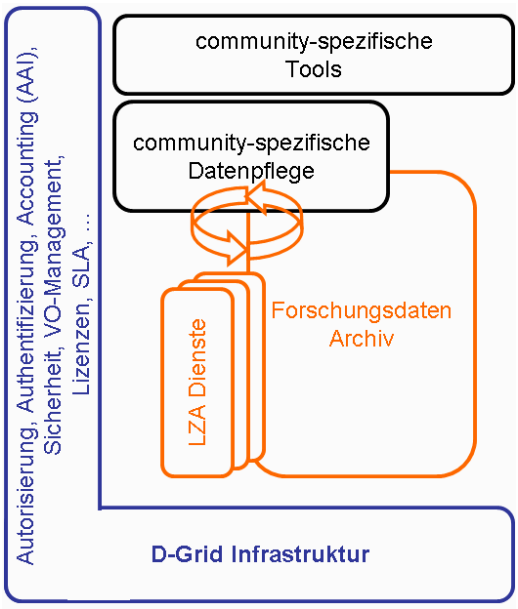


Abbildung 5: Vernetzung von WissGrid Forschungsdatenarchiv und LZA-Diensten mit community-spezifischer Datenpflege

Im Folgenden werden die Einbettung der Dienste in die D-Grid-Infrastruktur und die prinzipiellen Integrationsmöglichkeiten von Streaming-Diensten und einem Forschungsdatenarchiv beschrieben.

3.1. Einbettung in die D-Grid-Infrastruktur

Die D-Grid Infrastruktur unterstützt heterogene Grid Middleware zur Nutzung durch die Communities. Zur Einbettung von LZA-Diensten in D-Grid wird daher eine möglichst umfassende Interoperabilität mit diesen heterogenen Middleware-Systemen angestrebt. Durch die Middleware-übergreifende Standardisierung in Gremien wie dem Open Grid Forum (OGF), sowie den Integrationsarbeiten des D-Grid Integrationsprojektes (DGI) kann diese Interoperabilität durch einige dezidierte Standards und Systeme realisiert werden, die im Folgenden für die Bereiche AAI und Sicherheit, D-Grid Informationsdienste sowie Persistent Identifier behandelt werden.[14] Anforderungen in diesen Bereichen sind:

a. Interoperabilität mit GSI [AAI und Sicherheit]
b. optional, Nutzung von D-Grid-weiter Dienste-Registrierung oder Monitoring [D-Grid Informationsdienste]
c. optional, Einlesen von Daten aus einem überreichten Identifier [Persistent Identifier]

Bereits jetzt unterstützt das DGI in seiner Referenz-Installation[15] den dateiorientierten Speicherdienst dCache sowie das datenbankorientierte OGSA-DAI. Zusätzlich wird mit dem iRODS System zur Datei- und Metadatenverwaltung experimentiert. Die Rolle dieser und anderer Systeme zur Verwaltung von Forschungsdaten wird in Kapitel 3.3 weiter vertieft.

3.1.1. AAI und Sicherheit

In D-Grid baut die Sicherheitsinfrastruktur auf Verwendung von persönlichen X.509-Zertifikaten auf. Ein wesentlicher Standard hierfür ist die Grid Security Infrastructure (GSI)[16], die von Middleware-Systemen wie Globus, OGSA-DAI, iRODS, und anderen implementiert wird. Die Interoperabilität von LZA-Diensten mit der GSI ist daher eine Kernanforderung.

Die bisher erprobten X.509 Zertifikate sind für Einzelpersonen ausgestellt und Aktivitäten im D-Grid auf Basis dieser Zertifikate müssen auch von Einzelpersonen angestoßen werden. Um automatisierte LZA-Dienste wie eine Formatumwandlung einer Vielzahl von Objekten in einem ganzen Datenarchiv über das Grid ablaufen zu lassen (manchmal auch "job delegation" genannt), müssen andere Absicherungen gefunden werden. Hier bieten sich Service-Zertifikate an, die die Vertrauensbasis von der rein personellen Ebene hin zu den Diensten

[14] Die Beschreibungen in diesem Dokument dienen lediglich als Referenzmodell. Detaillierte Anleitungen für spezifische Konfigurationen, wie z.B. die Firewall-Konfiguration von Grid-Komponenten, finden sich für D-Grid spezifische Einstellungen auf den Support-Seiten des DGI, bzw. werden für LZA-Dienste im Rahmen des WissGrid Projektes entwickelt.
DGI: Firewall Configuration. http://dgiref.d-grid.de/wiki/Category:Firewall (viewed October 2009)
[15] D-Grid Referenz-Installation. http://dgiref.d-grid.de/wiki/Introduction (viewed October 2009)
[16] OGF GridForge: Grid Security Infrastructure (GSI). http://forge.ogf.org/sf/projects/gsi-wg

verlagern. Dazu laufen Arbeiten im Gap-Projekt GapSLC[17], die u.a. in Zusammenarbeit zwischen Nutzern und Ressourcenanbietern die technischen und organisatorischen Grundlagen für die Akzeptanz von Service-Zertifikaten anstreben.

Neben den GSI-basierten Systemen sind in anderen Umgebungen auch Shibboleth-basierte Authentifizierung und andere Systeme etabliert. Aktivitäten zum Aufbau von Interoperabilitäts-Gateways werden verfolgt und können je nach System eingesetzt werden, sind aber keine Kernanforderung für die WissGrid LZA-Architektur.

Analog ist auch die Abbildung von spezifischen Anforderungen je nach Community und System umzusetzen. Zum Beispiel werden je nach Kontext unterschiedliche Sicherheitsanforderungen bezüglich des Zugriffs auf die Daten zu erfüllen sein: medizinische Daten mit Personenbezug bedürfen stärkerer Autorisierungsmechanismen als Transkriptionen einer historisch interessanten Handschrift.

3.1.2. D-Grid Informationsdienste

Das DGI betreibt eine Reihe von übergreifenden Informationsdiensten, die z.B. Nachweis (Monitoring) und Nutzung (Accounting) von verfügbaren Hardware-Ressourcen in D-Grid gewidmet sind. Die Registrierung und verteilte Nutzung von Streaming-Diensten (Grid Job, siehe 3.2) kann dadurch unterstützt werden.

Umgekehrt plant das DGI auch die Einführung eines Dienstes zur Überwachung von Diensten, um deren Verfügbarkeit zu sichern. Auch dieser Dienst kann für Streaming-Dienste ("Web Service", siehe 3.2) entsprechend eingesetzt werden.

Diese verfügbaren Infrastruktur-Dienste können von WissGrid LZA-Diensten eingesetzt werden, deren Einsatz ist aber keine Kernanforderung von Seiten des DGI oder von WissGrid. Dennoch ist es wichtig, zukünftige Entwicklungen weiter im Auge zu behalten und LZA-Dienste ggf. entsprechend anzupassen. Aktuelle Entwicklungen im Gap-Projekt "SLA4D-Grid" zur Einbettung einer Service-Level-Agreement-Schicht in D-Grid[18] könnten für die Vermittlung von LZA-Diensten relevant sein.

3.1.3. Persistent Identifier

Persistente Identifier (PI) dienen der dauerhaften Identifizierung von Dateneinheiten und sind daher für eine auf Nachhaltigkeit angelegte Grid-Infrastruktur unabdingbar. Beispiele sind DOI[19] oder URN[20]. In den einzelnen Communities werden bereits in den jeweiligen Datenarchiven persistente Identifier eingesetzt. LZA-Dienste können darauf zurückgreifen, in dem in der Jobbeschreibung eine Liste der PI's der zu bearbeitenden Daten spezifiziert wird, statt die Daten selbst vorher zu stagen und an den Dienst zu übergeben. Hierbei muss allerdings beachtet werden, dass der Identifierraum nicht einheitlich ist. Insofern müssen die Dienste dann entweder Schnittstellen zu unterschiedlichen lokalen Identifier-Systemen

[17] Nutzung von kurzlebigen Zertifikaten in portalbasierten Grids – GapSLC. http://www.d-grid-ggmbh.de/index.php?id=93&L=1
[18] SLA4D-Grid. http://www.sla4d-grid.de/
[19] http://dx.doi.org
[20] http://nbn-resolving.de

aufweisen oder ein Dienst zur systemübergreifenden Auflösung von PI‘s müsste aufgebaut werden. Letzteres wäre im Kontext einer Community-übergreifenden Identifikation von Entitäten und damit einer Nachnutzung von Daten über Communitygrenzen hinweg wünschenswert. Dies wird im Rahmen des Projekts jedoch höchstens exemplarisch für konkrete Beispiele an PI-Systemen zu leisten sein.

3.1.4. Organisatorische Aspekte

Um die WissGrid LZA-Architektur in die D-Grid-Infrastruktur zu integrieren, müssen neben den technischen auch eine Reihe von organisatorischen Punkten beachtet werden. Diese bedürfen einer weiteren Behandlung außerhalb des Architekturprozesses und können hier nur skizziert werden:

- Der D-Grid Software-Stack spezifiziert Grid-Distributionen und deren Konfiguration, über die auf Hardware-Ressourcen des D-Grid zugegriffen werden kann. Die WissGrid LZA-Architektur baut auf diesem Software-Stack auf. Außerdem bringt sich WissGrid in die Diskussionen im Rahmen des "Modifikationsprozesses" (regelmäßiger Review und Anpassung des Software-Stacks) mit ein.
- D-Grid unterstützt sowohl die Nutzung von Compute-Ressourcen als auch Storage-Ressourcen. Während vielfältige Konzepte für die Nutzung von Compute-Ressourcen des D-Grid schon vorliegen, existieren nur wenige für die von Storage-Ressourcen. Hier sind dCache und OGSA-DAI zu nennen, die jedoch über eine technologische Lösung noch nicht hinausgehen. Insbesondere existieren nur sehr allgemeine Policies für Grid-Storage, die hauptsächlich den intermediären Charakter behandeln. Im Rahmen des D-Grid Betriebskonzeptes für Storage wird derzeit keine Speicherdauer garantiert. Dies bedeutet im Extremfall auch, dass der Administrator eines D-Grid Knotens spontan entscheiden kann, ob Daten aufbewahrt oder gelöscht werden. Derzeit ist noch unklar, ob sich mittel- bis langfristig Storage-Knoten in D-Grid herausbilden, die spezielle Service Level Agreements (SLA) für langfristigen Storage anbieten, der ggf. auch Bit-Preservation (siehe Abschnitt 4.3.1) garantiert.
- Die weitere Rollenverteilung zwischen WissGrid, dem DGI und den Communities wird im Laufe des WissGrid Projektes weiter diskutiert. Generell ist das Ziel, dass die Communities selbst, mithilfe der von WissGrid entwickelten grundlegenden Konzepte, Architektur und funktionalen Dienste, ihre individuellen LZA-Strategien und entsprechenden Systeme errichten und nachhaltig betreiben können.
- Ebenso ist derzeit noch offen, ob die WissGrid LZA-Dienste mittel- bis langfristig in die Wartung des DGI überführt werden. Da LZA-Dienste ständig erweitert und dem aktuellen technischen Kontext angepasst werden müssen, wäre es sinnvoll, dass sich bei Erfolg des WissGrid LZA-Konzeptes ein D-Grid-weites LZA-Kompetenzzentrum herausbildet, das eng mit dem DGI zusammenarbeitet. Hierbei könnte ein LZA-Kompetenzzentrum sich vor allem auch der organisatorischen Fragestellungen in der Beratung von Communities annehmen, und als Community-übergreifender Mittler bezüglich Fragen, die technische Infrastruktur betreffen, fungieren. Fragen der

Organisation und des Betriebs eines solchen Zentrums wären zu einem späteren Zeitpunkt in einem separaten Dokument zu behandeln.

3.1.5. Metadaten

Metadaten sind notwendig, um langfristig Daten effektiv verwalten und nutzen zu können. Dies gilt insbesondere für die Langzeitarchivierung, wo darauf geachtet werden muss, dass zur Nutzung notwendiges technisches und semantisches Wissen (insbesondere Kontextwissen) durch Metadaten expliziert wird. Der maßgebliche Standard in diesem Bereich ist PREMIS[21], der Metadatenelemente für fünf miteinander verknüpfte Bereiche definierte: Intellectual Entities, Objects, Rights, Agents und Events. Diese Bereiche werden im D-Grid-Kontext durch verschiedene Akteure und Technologien abgedeckt (siehe Abbildung 6) und müssen für spätere Spezifikationen entsprechend berücksichtigt werden.

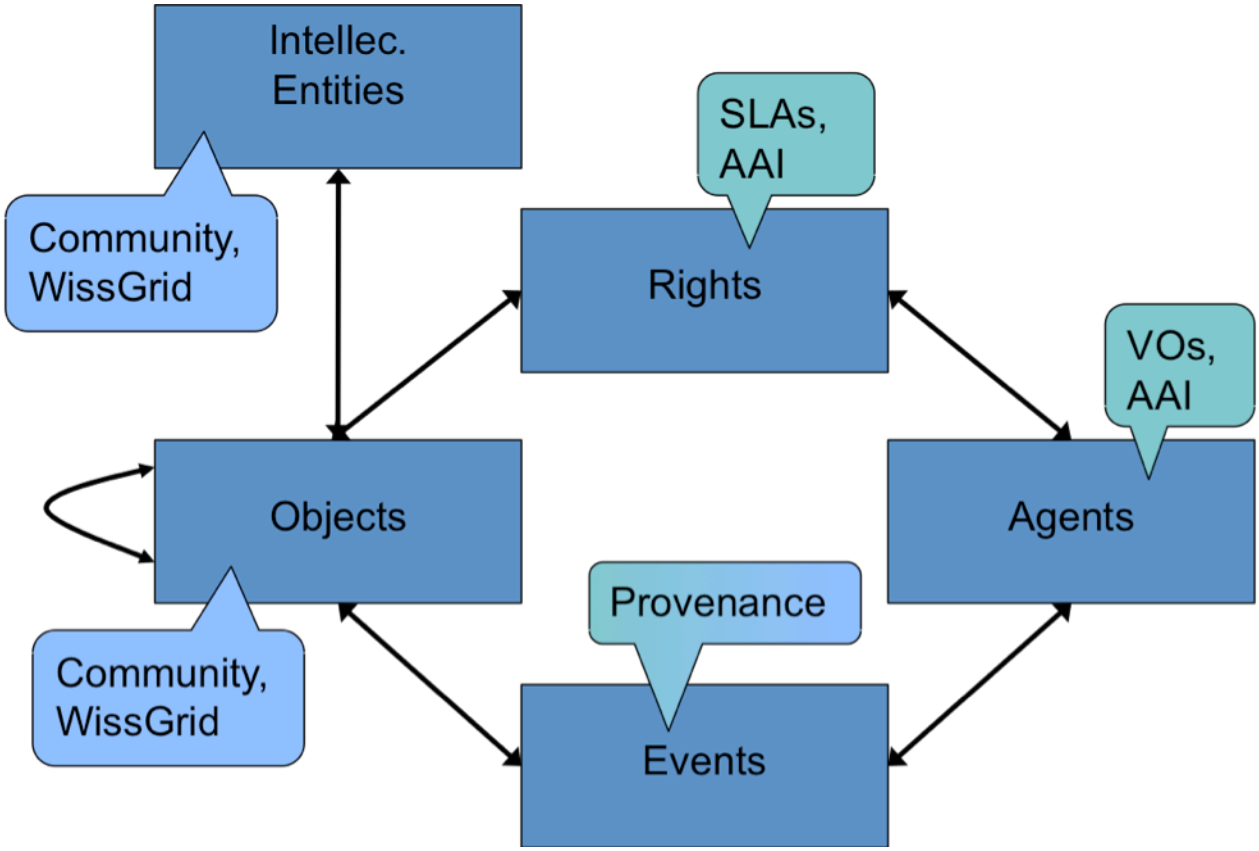


Abbildung 6: Die LZA-Metadatenkategorien von PREMIS werden im D-GRID durch verschiedene Akteure und Technologien abgedeckt.

[21] http://www.loc.gov/standards/premis/

3.2. LZA-Streaming-Dienste

Streaming-Dienste erfüllen dezidierte, abgeschlossene Aufgaben auf Daten oder Metadaten, die zur Langzeitarchivierung bestimmt sind. Dazu zählen z.B. die Validierung von Objekten (Formate oder Metadaten) oder die Format-Konversion (z.B. TIFF in JPEG2000, TeX in TEI/XML).

Für die Einbindung in Archive können primär zwei Herangehensweisen unterschieden werden: die Dienste werden direkt in das Archiv eingebettet oder die Daten werden zu Diensten auf ein externes System transferiert. Bei einer Einbettung ins Archiv entfällt der Datentransfer, was potenziell die Effizienz einer Aufgabe erhöht. Gleichzeitig wird dadurch aber auch eine mitunter erhebliche Last auf das Archivsystem gelegt. Speziell bei LZA-Aufgaben auf mehreren Objekten, bei denen z.B. eine umfangreiche Sammlung komplett in ein anderes Datenformat migriert werden muss, bietet sich die Verteilung des Dienstes auf fremde Systeme an.

Orthogonal dazu kann man unterschiedliche technische Einbindungsvarianten unterscheiden. Im Folgenden sind speziell zwei Varianten beschrieben: die Umsetzung als Web Service und als Grid Job. Diese technischen Einbindungsvarianten können dabei mit den oben beschriebenen Herangehensweisen (Daten zu Dienst oder Dienst zu Daten) beliebig kombiniert werden. So könnte zum Beispiel ein Web Service direkt in einem ansonsten nicht Grid-angebundenen Forschungsdatenarchiv installiert werden oder umgekehrt wäre für ein Archiv, dessen Storage-Bereich im Grid liegt, der Grid Job eine Variante, um den Dienst zu den Daten zu bringen.

Im Falle umfangreicherer Aufgaben mit mehreren Objekten kann man bei beiden im Folgenden beschriebenen Varianten entweder einzelne Objekte an den Dienst senden oder mehrere Objekte gleichzeitig verarbeiten (genannt "im Batch" verarbeiten). Die zweite Variante der Batch-Verarbeitung könnte dabei etwas instabiler sein, da sich Probleme gleich auf den gesamten Batch auswirken, nicht nur auf ein Objekt. Gleichzeitig ist die Batch-Verarbeitung aber potenziell auch effizienter, da weniger Kommunikations-Overhead nötig ist (dies macht sich z.B. besonders bemerkbar, wenn die Rechte über ein Zertifikat entweder mehrfach für jedes einzelne Objekt oder einmal für einen gesamten Batch an Objekten überprüft werden müssen).

3.2.1. Einbindungsvariante 1: Web Service

Die Umsetzung von LZA-Funktionalitäten als Web Services (SOAP, REST) wird in verschiedenen internationalen LZA-Initiativen unterstützt (z.B. PANIC, PLANETS, CRiB). Zum Beispiel bietet das PLANETS Projekt diverse Services zur Format-Validierung und Format-Migration an.

Extern laufende Web Services reduzieren den Installationsaufwand bzw. umgehen diesen komplett. Sie bieten sich besonders an, wo geringe Datenmengen im Spiel sind und somit der Transfer nicht schwer ins Gewicht fällt, also z.B. für einzelne Daten oder die Verarbeitung von kleineren Sammlungen bzw. für Tests. Bei größeren Datenmengen könnte der Service direkt oder in physischer Nähe des Archivs installiert werden, um den Zeitaufwand des Datentransfers zu verringern.

Die **technische Architektur** zur Einbettung von existierenden Diensten als Web Service in eine Grid-Umgebung ist inspiriert durch das Entwurfsmuster "Decorator"[22]. Hierbei kann jede Art von Dienst (egal ob er als Web Service oder Command-Line Tool vorliegt) als Web Service gekapselt werden. Der Decorator-Dienst ermöglicht hierbei auch die in Kapitel 3.1 formulierten Minimalanforderungen zur Einbindung in D-Grid, und dabei vor allem die Umsetzung der GSI-Fähigkeit sowie die Bereitstellung einer Service-Beschreibung, z.B. in WSDL.

Ein Client für einen solchen Dienst würde sich valide Credentials aus der Grid-Laufzeitumgebung holen oder selbst aus einem Grid-Zertifikat valide Credentials für die Kommunikation mit dem GSI-fähigen Web Service erzeugen, wenn der Client nicht in einer Grid-Umgebung läuft. Für die Erzeugung von Credentials können z.B. die MyProxy-Bibliotheken von Globus genutzt werden.

3.2.2. Einbindungsvariante 2: Grid Job

In Abhängigkeit vom Datenvolumen und der für den Dienst notwendigen Rechenkapazität können einige Aufgaben durch die Verteilung im Grid wesentlich effizienter bearbeitet werden. Gängige Grid-Funktionen stehen dafür je nach verwendeter Middleware zur Verfügung.

Hierzu muss der LZA-Streaming-Dienst bei allen beteiligten Computeprovidern installiert und via Command-Line abrufbar sein. Als **technische Architektur** bietet sich auch die Kapselung des Dienstes an, um wie auch bei der "Web Service"-Variante den Dienst mit GSI-Fähigkeit und gegebenenfalls weiteren Grid-spezifischen Fähigkeiten auszustatten.

Für einen Client bietet es sich an, ein Skript zur Verfügung zu stellen, das die entsprechenden Funktionen zur Job Submission übernimmt: von der Auswahl des Knotens, Staging der Daten bis hin zum Abholen der Resultate über die Job-ID. Die entsprechenden Grid-Funktionen muss der Client dafür aus der Grid-Laufzeitumgebung bzw. aus entsprechenden Bibliotheken einbinden.

3.3. Forschungsdatenarchive

Zur Erleichterung des Einstiegs von Communities mit wenig Erfahrung und wenigen existierenden Diensten im Bereich LZA erstellt WissGrid nachnutzbare Forschungsdatenarchive. Dabei handelt es sich explizit nicht um ein zentrales Archiv, sondern vielmehr um Softwarepakete im Stile der D-Grid Referenz-Installation, die jeweils für unterschiedliche Anforderungen empfohlen sind und deren Einbindung in D-Grid getestet ist. Alle diese Softwarepakete unterstützen LZA-Strategien und die Einbindung in D-Grid im gleichen Maß, sie unterscheiden sich lediglich in der Unterstützung möglicher Nutzungsszenarien und Community-spezifischer Anforderungen. Dieses Unterkapitel beschreibt unterschiedliche Ansätze zur Errichtung von Forschungsdatenarchiven und die derzeit von WissGrid unterstützten Anwendungsprofile. Die technische Umsetzung der

[22] Decorator Entwurfsmuster. http://en.wikipedia.org/wiki/Decorator_pattern

Anwendungsprofile und detaillierte Spezifikationen werden in einem parallelen Deliverable beschrieben.

Forschungsdatenarchive bewahren Daten über lange Zeiträume und unterstützen die Nachnutzung dieser Daten. Wesentlich dafür sind ein aktives Management der Daten (z.B. mit kontinuierlicher Überwachung der Formate) wie auch eine aussagekräftige Beschreibung der Daten, sodass sie auffindbar, administrierbar und interpretierbar sind, auch wenn die ursprüngliche Umgebung, in der die Daten erstellt wurden, nicht mehr vorhanden ist. Forschungsdatenarchive halten daher Daten zumeist in einem stabilen Format als Dateien, und verknüpfen die Daten eng mit beschreibenden Metadaten.

Auch für die Verwaltung von Datenbanken (z.B. relational) wird mitunter ein dateibasierter Ansatz empfohlen, da dieser vor allem bei langen Aufbewahrungszeiten als stabiler angesehen wird. Eine entsprechende Methode zur Überführung von Datenbanken in Dateien wird beispielsweise durch SIARD[23] beschrieben, das besonders auf die Beschreibung der Struktur und Inhalte der Daten zur späteren Interpretierbarkeit achtet.

Das DGI bietet in seiner Referenz-Installation für Storage derzeit die beiden Daten-Management Systeme dCache und OGSA-DAI an. Beide können nicht (allein) als Forschungsdatenarchive eingesetzt werden: OGSA-DAI ist auf die Virtualisierung von heterogenen Datenbanken spezialisiert und vernachlässigt dateibasiertes Datenmanagement. Bei beiden, OGSA-DAI und dCache, fehlt eine Metadatenverwaltung, die über Datei-Attribute hinausgeht.

Etablierte Systeme zur langfristigen Datenverwaltung sind Repository-Systeme wie Fedora[24]. Sie verwalten Daten in der Form von Objekten - der Verknüpfung aus (mitunter mehreren) Dateien mit Metadaten in einer Einheit. Die Einbettung von Repository-Systemen in Grid-Umgebungen wird in unterschiedlichen Arbeitsgruppen untersucht[25]; diesbezügliche Ansätze sind in Anhang 4.5 skizziert.

Obwohl Softwarepakete für Forschungsdatenarchive möglichst generische Dienste anbieten, gibt es doch Anforderungen, die so unterschiedlich sind, dass sie nicht von einem Systemtypus abgedeckt werden können. WissGrid unterscheidet daher unterschiedliche Anwendungsprofile, und bietet diesen jeweils eine passende Referenzsoftware. Diese Referenzsoftwarepakete werden im Rahmen des WissGrid Projektes in D-Grid eingebettet und die Interoperabilität mit den (oben beschriebenen) LZA-Diensten wird gesichert.

Darüber hinaus muss ein System in der Installation an die speziellen Anwendungsbedürfnisse sowie den organisatorischen und technischen Kontext angepasst werden. Dies betrifft vor allem den Daten-Ingest und Access (siehe OAIS Referenzmodell) bzw. die Objektmodellierung. Diesen Prozess kann WissGrid durch Beratung unterstützen und technisch begleiten, jedoch stammt der wesentliche Beitrag dabei aus dem Anwendungs-

[23] Database Preservation: The international Challenge and the Swiss Solution. Digital Preservation Europe, Briefing Paper. Viewed August 2009.
http://www.digitalpreservationeurope.eu/publications/briefs/database_preservation.pdf
[24] Fedora - Flexible Extensible Digital Object and Repository Architecture. http://www.fedora-commons.org/
[25] OGF Digital Repositories Research Group. http://www.ogf.org/gf/group_info/view.php?group=dr-rg

kontext. Auch die Nutzung und der langfristige Betrieb des Forschungsdatenarchives liegt anschließend bei dem Community Grid, der Universität bzw. dem Anwendungspartner.

Die Analyse der wichtigsten Anforderungen lässt eine grobe Einteilung in die nachfolgend beschriebenen Anwendungsprofile zu. Für alle sind Aspekte wie LZA-Unterstützung und Sicherheit gleichermaßen relevant. Jedoch gibt es Unterschiede in den folgenden Dimensionen:

- Datenvolumen, Anzahl Objekte
- Zugriffsarten (Ingest, Access) und Interoperabilität
- Metadatenmodellierung

Neben diesen Dimensionen unterscheiden sich die nachfolgend beschriebenen Anwendungsprofile auch entlang der Integrationsmöglichkeiten von Grid und Repositorien-Technologien, die in Anhang 4.5 ausgearbeitet sind. Hierbei deckt Anwendungsprofil A die Integrationsvariante 1 ("Repositorien als Archiv-Backends für das Grid") ab, Anwendungsprofil B die Integrationsvariante 2 ("Daten-Grid als Repository-Storage"). Integrationsvariante 3 ("Virtualisierung von Repositorien") ist kein Kernbereich der Integration in D-Grid und kann daher nicht mit den in WissGrid vorhandenen Ressourcen abgedeckt werden. Dennoch könnten beide Anwendungsprofile A und B zusätzlich Integrationsvariante 3 unterstützen und dies wäre langfristig (jenseits der WissGrid-Projektphase) auch ein wichtiges Ziel.

3.3.1. Anwendungsprofil A: Grid-Workflow

Dieses Anwendungsprofil deckt speziell jene Community Grids ab, die große Mengen an Daten im Grid erzeugen und verarbeiten (vgl. Integrationsvariante 1, Anhang 4.5). Neben einer ständigen Verfügbarkeit für Verarbeitungen im Grid sollen diese Daten auch in einer vertrauenswürdigen Umgebung über lange Zeiträume aufbewahrt werden. 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.

- Datenvolumen, Anzahl Objekte - skaliert zu besonders großen Datenmengen im Terabyte-Bereich und unterstützt auch eine große Anzahl von einzelnen Objekten. Dies erfordert sowohl die Verwaltung und Wiederauffindbarkeit von einem speziellen Objekt in Sammlungen aus mehreren Millionen Objekten wie auch ein adäquates Datei-Management, das mit diesen Mengen umgehen kann.
- Zugriffsarten (Ingest, Access) und Interoperabilität - vor allem die technische Einbindung in die Grid-Umgebung ist wesentlich, damit Daten in Compute-Workflows eingebunden werden können (z.B. GridFTP, SRM).
- Metadatenmodellierung - eine effiziente Verknüpfung mit Metadaten ist wesentlich für die Einbindung in Workflows sowie die Verwaltung im Grid. Vor allem deskriptive Metadaten (z.B. zum schnellen Retrieval) sowie administrative Metadaten (z.B. für Administration von Daten-Replikation und Integritätsprüfung im Grid) sind hierbei gefragt.

3.3.2. Anwendungsprofil B: interaktive Forschungsumgebung

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 (vgl. Integrationsvariante 2, Anhang 4.5), 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.

- Datenvolumen, Anzahl Objekte - unterstützt insbesondere auch eine große Anzahl kleiner Objekte und Erschließungsmaterialien, die durch das kollaborative Arbeiten entstehen können (z.B. XML-Daten, Annotationen zu Bildern). Objekte mit großen Volumina können ebenso verwaltet werden, wobei die Zugriffsgeschwindigkeit abhängig von der Schnittstelle ist.
- Zugriffsarten (Ingest, Access) und Interoperabilität - eine nahtlose Einbindung in aktuelle Web-Technologien bietet sich für interaktive, kollaborative Nutzerumgebungen an, zumal dadurch die Verknüpfung mit existierenden Forschungsumgebungen oftmals unterstützt wird. Mit dem Fortschritt von Web-Technologien und möglicherweise anderen Technologien mag sich dies ändern, aber aktuell ist eine Unterstützung von HTTP/REST-basierten Technologien essenziell, um möglichst nah an den Nutzer heranzukommen.
- Metadatenmodellierung - für eine kollaborative Arbeitsumgebung ist vor allem die Flexibilität und Reichhaltigkeit der Metadaten relevant. Dazu ist auch notwendig, dass Nutzer selbst Metadatenmodelle definieren können und Metadaten und Relationen zwischen Objekten selbst zuweisen können.

3.3.3. Anwendungsprofil C: föderierte Archive

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[26] können diese bestehenden Archive aber untereinander verbunden werden, sodass die (eigentlich verteilten) Bestände wirken als lägen sie in einem einzigen virtuellen Gesamtarchiv (vgl. Integrationsvariante 3, Anhang 4.5). Auch die Einbettung der Objekte in Repositorien in wissenschaftliche Workflows im Grid kann durch dieses Profil gefördert werden.

Durch diese Verknüpfung von bestehenden Archiven garantiert diese Integrationsvariante per se keine Vertrauenswürdigkeit, sondern bietet nur in dem Maße langfristige LZA, wie sie durch die einzelnen Archive angeboten wird. Dennoch können LZA-Strategien durch die Virtualisierungsschicht (Föderierung) unterstützt werden.

[26] z.B. Open Archives Initiative. http://www.openarchives.org/, Open Repositorys. http://openRepositorys.org/

Da eine Repositorien-Föderation eine Integrationsschicht über existierenden Archiven darstellt, wird an dieser Stelle keine Analyse des Anwendungsprofils durchgeführt (wie in den anderen Profilen oben anhand der Eigenschaften: Datenvolumen, Anzahl Objekte / Zugriffsarten / Metadatenmodellierung). Auch wird WissGrid für dieses Anwendungsprofil keine Software entwickeln, sondern vielmehr den Zugang zu bestehen Erfahrungen erleichtern und an D-Grid anpassen.

4. Anhänge

4.1. Anhang 1: Fallstudien

Die Fallstudien dienen dazu, die hier vorgelegte WissGrid LZA-Architektur an dem Ist-Stand der Communities zu überprüfen und sichtbar zu machen, wo Änderungen an der Architektur oder ggf. sogar in der Praxis der Communities sinnvoll sind. Sie werden zusammen mit der Architektur fortgeschrieben. Neben den derzeitigen Fallstudien aus dem Bereich Klima (Deutsches Klimarechenzentrum, DKRZ), Medizin (Medizinische Anwendungen, TMF) und Biostatistik sind u.a. Untersuchungen der Bereiche germanistische Linguistik und Sozialwissenschaften in Vorbereitung.

4.1.1. Offene Vorlage für Fallstudien

Da sich eine Fallstudie primär an der Situation der Community orientieren muss, gibt es keine verbindliche Vorgabe für die zu behandelnden Themen. Im Folgenden sind aber einige üblicherweise relevante Themen für den Ist- und Soll-Zustand der Langzeitarchivierung aufgeführt, um auf gegebenenfalls unbeachtete Aspekte aufmerksam zu machen:

- Motivation zur Archivierung, Wert der Daten, gesetzliche Bestimmungen,
- Inhalt und Qualität der Daten, Datenselektion, Aufbewahrungsdauer und evtl. Notwendigkeit zur Löschung, Qualitätssicherungsworkflow
- Art der Daten, Größe der Files, Menge (Anzahl, Size)
- Einbringen ins Archiv, wie/wo liegen die Daten vor (Bulk im Einmaltransfer oder stetig anfallende Einzeldaten), wer führt sie ein (Erzeuger, Archivist)
- Zugang, Nutzen und Nachnutzung: online in Webseite, offline; Beschreibung der Applikation: Werkzeuge, Suchmechanismen; Teamarbeit, Veröffentlichung der Daten
- Datendokumentation und Metadaten, Standards, Relationen zwischen den Daten (Verknüpfungen, Versionen, Übersetzungen, etc.), Provenienz und Prozessdokumentation, ständige Erweiterung
- Speicherung und Langzeitarchivierung, redundante Ablage, Datenkonversion (on-ingest, on-request), Migration/Emulation der Daten
- Datenschutz, Datensicherheit, Eigentum/Copyright, Forschungsethik, Zugriffsrechte, gesetzliche Bestimmungen
- Zusammenfassung der allgemeinen und dienstspezifischen Anforderungen an WissGrid

4.1.2. Fallstudie Klima

Motivation zur Archivierung und Wert der Daten

In der Klimacommunity gibt es drei Weltdatenzentren, die Daten für die gesamte Community speichern und bereitstellen und die folgenden thematischen Schwerpunkte setzen: Für Simulationsdaten zur Klimaentwicklung ist das World Data Center Climate (WDCC) verantwortlich, das WDC Mare stellt vor allem Daten aus ozeanographischen Messungen bereit und im WDC RSAT werden Daten aus Satellitenmessungen und davon abgeleitete Datenprodukte zur Verfügung gestellt. Daneben gibt es noch umfangreiche Datenarchive an einzelnen anderen Einrichtungen, da die Daten z.B. von der Einrichtung erhoben wurden und dort die zugehörige Expertise liegt (Bsp. Messreihen des Deutschen Wetterdienstes).

Die in den Archiven vorgehaltenen Messdaten repräsentieren den Zustand der Umwelt zum jeweiligen Zeitpunkt und können im Gegensatz zu vielen Experimentdaten nicht durch Wiederholung eines Versuchs erneut erhoben werden. Daher ist ihre Archivierung essentiell.

Die gespeicherten Daten sind für die Klimaforschung selbst aber auch für die Klimafolgenforschung (Ökonomie, Landwirtschaft,...) von Interesse und besitzen gesamtgesellschaftliche Relevanz.

Die folgenden Ausführungen beziehen sich vor allem auf das WDCC, das hier als ein Beispiel für wichtige Daten in der Klimaforschung dienen soll.

Rechtliche Bestimmungen

Klimadaten unterliegen keiner gesetzlichen Archivierungspflicht. Das WDCC bekennt sich aber zu den Regeln zur Sicherung guter wissenschaftlicher Praxis[27], aus denen sich eine Selbstverpflichtung ergibt, Primärdaten mindestens zehn Jahre lang aufzuheben. Eine Verpflichtung zur Langzeitarchivierung ergibt sich auch aus den Prinzipien[28], denen die Weltdatenzentren unterliegen. Hier ist z:B. vorgeschrieben, dass die Daten im Falle der Schließung des Weltdatenzentrums an ein anderes Weltdatenzentrum weitergegeben werden müssen. Außerdem garantieren die WDC-Prinzipien den (fast) kostenfreien Zugang zu den Daten für Wissenschaftler aller Länder.

Inhalt und Qualität der Daten

Datenselektion

Das Archiv des WDCC ist beschränkt auf verarbeitete Klimadaten, mit Schwerpunkt auf Daten aus Klimamodellrechnungen und verwandten Beobachtungen, die in ganz bestimmten Dateiformaten vorliegen. Erlaubt sind ASCII, GRIB und NetCDF. Die beiden letzten sind zwei in der Klimaforschung übliche Formate. Die Archivierung von Rohdaten aus Satelliten ist nicht vorgesehen, da sie in den anderen WDCs gespeichert werden.

Aufbewahrungsdauer und eventuelle Notwendigkeit zur Löschung

Mindestens zehn Jahre, eine Höchstdauer ist nicht festgelegt.

---

[27] (a) Deutsche Forschungsgemeinschaft, Sicherung guter wissenschaftlicher Praxis, Wiley-VCH, 1998, ISBN 3-527-27212-7; (b) Max-Planck-Gesellschaft, Verantwortliches Handeln in der Wissenschaft, ISSN 1438-8715, Anhang 1.

[28] Principles and Responsibilities of ICSU World Data Centers, http://www.ngdc.noaa.gov/wdc/guide/gdsystema.html

Qualitätssicherungsworkflow

Für eine Publikation mit DOI/URN Registrierung von Primärdaten (bibliographische Referenz ISO 690-2, siehe unten) werden an diese Datensätze und beschreibende Daten (Metadaten) eine Reihe von formalen Anforderungen gestellt, denen sie entsprechen müssen, um eine fest definierte Qualität für die Registrierung aufzuweisen. Die am WDCC übliche Qualitätskontrolle umfasst zwei unterschiedliche Prüfbereiche: Erstens den syntaktischen Bereich, hier werden die Daten und Metadaten auf Vollständigkeit überprüft, sowie getestet, ob die Daten für Nutzer verfügbar sind. Zweitens den semantischen Bereich, hier werden Metadaten und Daten auf korrekte Inhalte überprüft.

Art der Daten: Größe der Files, Menge (Anzahl, Size)

Strikte Vorgaben bezüglich der Dateigröße und –anzahl gibt es zwar nicht, es werden aber wenige große Dateien von den Datenbereitstellern erwartet, weil die Anzahl der in einer Containerdatei vorhandenen Dateien aus technischen Gründen begrenzt ist.

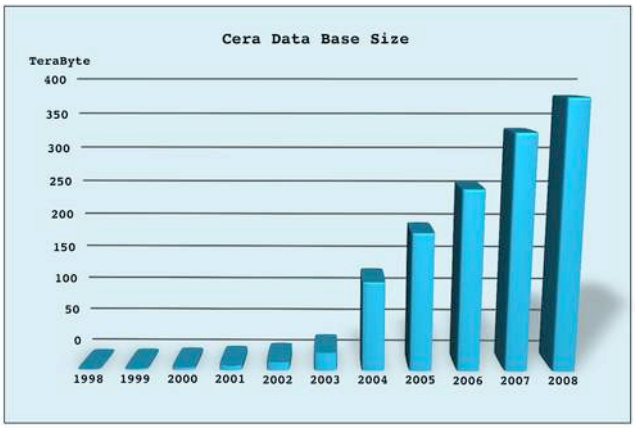


Abbildung 7: Datenmenge im WDCC (World Data Center Climate)

Im Jahr 2008 hat das WDCC ohne Kopien 380 TB an Daten enthalten. Die Datenmenge wächst von Jahr zu Jahr.

Einbringen ins Archiv: wie/wo liegen die Daten vor (Bulk im Einmaltransfer oder stetig anfallende Einzeldaten), wer führt sie ein (Erzeuger, Archivist)

Daten werden als Bulk im Einmaltransfer eingebracht.

Meistens werden die Daten von den Erzeugern beim DKRZ abgeliefert und durch einen Archivisten ins Archiv gebracht. Ein derzeit noch in Entwicklung befindliches Portal auf Basis von Geonetwork[29], das u.a. auch einen Metadateneditor enthält, der auf das Metadatenprofil des WDCC abgestimmt wurde, soll den direkten Ingest durch die Erzeuger erleichtern.

[29] http://www.osgeo.org/geonetwork

Zugang, Nutzen und Nachnutzung: online in Webseite, offline; Beschreibung der Applikation: Werkzeuge, Suchmechanismen

Der Zugang zum WDCC kann online über grafische Benutzerschnittstellen auf zwei verschiedenen Wegen erfolgen.

Der eine Weg führt über das CERA-Portal[30]. CERA steht für „Climate and Environment data Retrieval and Archiving system“. Das CERA-Portal wird wie das WDCC vom Deutschen Klimarechenzentrum betrieben. Über das CERA-Portal sind alle Datensätze des WDCC erreichbar. Gesucht werden kann in den Metadaten und zwar nach folgenden Strategien:

- Über eine Liste der Experimente[31]
- Über Begriffe, die aus einer Liste ausgewählt werden können
- Über den Namen der Modell-Software (Code-Suche)
- Volltextsuche
- Hierarchische Suche in einer Baumstruktur aus Oberbegriffen und Begriffen
- Suche in einer Tabelle mit Autorennamen, Titel und DOI

Nach erfolgreicher Suche kann auf die Daten per Download zugegriffen werden. Für die Nachbearbeitung stehen Zeit- und Datenformatkonvertierer sowie fachspezifische Berechnungswerkzeuge zur Verfügung.

Der zweite Weg zu Daten des WDCC führt über das C3-Portal[32]. Das C3-Portal ist die grafische Benutzerschnittstelle zum C3Grid[33] (**C**ollaborative **C**limate **C**ommuntiy Data and Processing **Grid**), einem echten Datengrid, in dem die Daten über verschiedene Standorte verteilt liegen. Neben Daten aus dem WDCC werden über das C3-Portal auch Daten des WDC Mare, des WDC RSAT, des Deutschen Wetterdienstes und zahlreicher anderer Organisationen bereitgestellt.

[30] http://cera-www.dkrz.de/CERA/
[31] Der Begriff „Experiment“ umfasst hier auch Modellrechnungen. Modellrechnungen sind sozusagen numerische Experimente.
[32] http://www.c3grid.de/portal/
[33] http://www.c3grid.de/

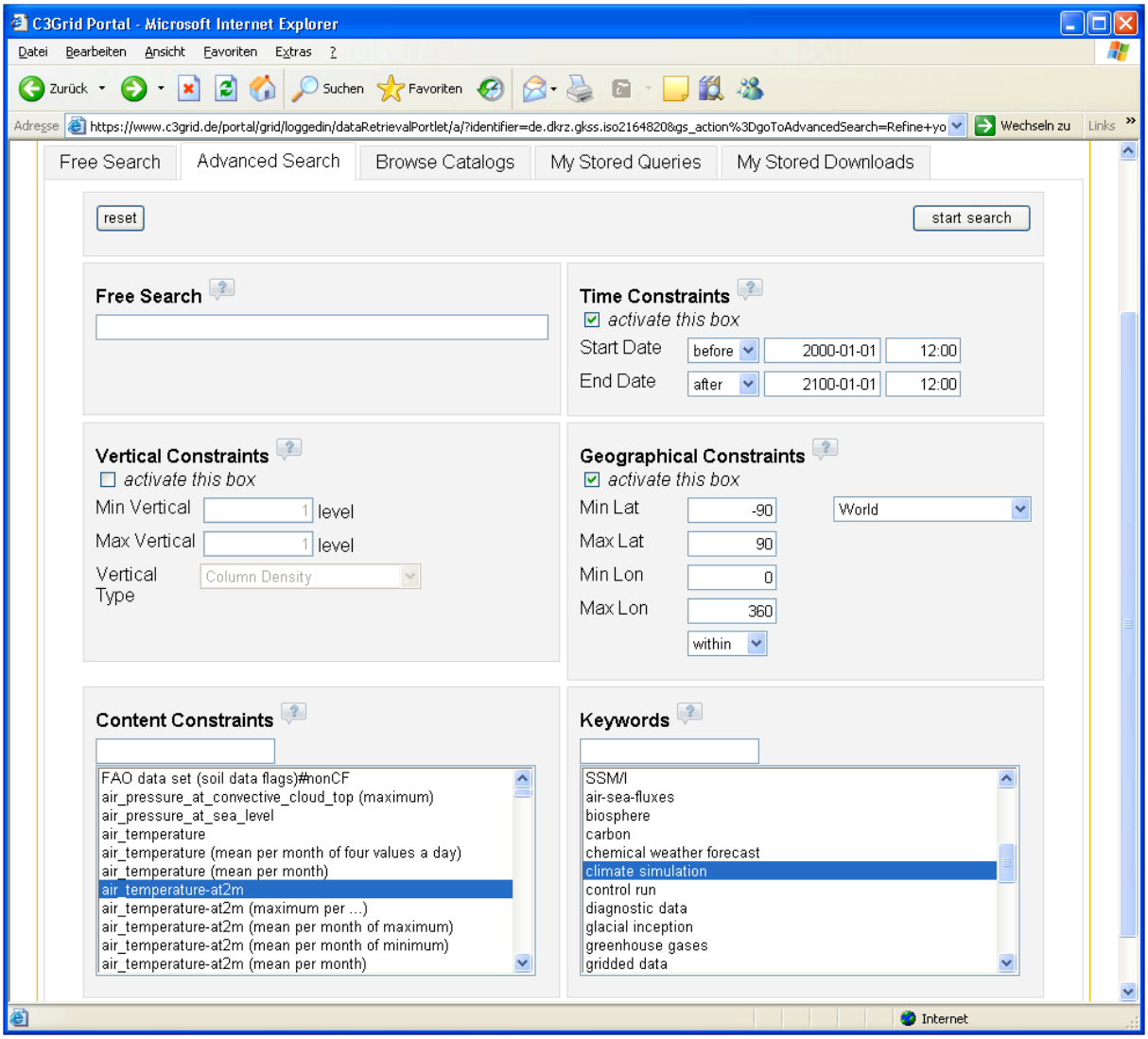


Abbildung 8: Web-Formular zur Datensuche im C3-Portal

Über das gezeigte Web-Formular kann in den Metadaten gesucht werden. Nach erfolgreicher Suche kann über das C3-Portal ein Grid-Job für den Download abgeschickt werden.

Über das C3-Portal sind nicht alle Daten des WDCC verfügbar. Nicht gefunden werden können Daten mit Nutzungsbeschränkungen. Gemeint sind solche, vor deren Nutzung eine Erklärung unterschrieben werden muss, dass die Daten nur für wissenschaftliche Zwecke verwendet werden. Ebenfalls nicht gefunden werden können Daten mit Metadaten, die nicht dem CERA-2-Datenmodell[34] entsprechen. Solche Daten sind im WDCC teilweise noch vorhanden.

Statt eines Downloads kann auch ein Workflow zur rechnerischen Weiterverarbeitung über das C3-Portal gestartet werden. Hierfür sucht sich der Nutzer einen der vorgegebenen, fachspezifischen Workflows aus einer Liste aus und startet den Grid-Job über das Portal. Die Workflow-Software kann nicht über das Portal verändert werden. Zusätzlich benötigte Informationen können aber – wie in der folgenden Abbildung gezeigt – über ein Formular mitgegeben werden.

[34] http://www.mad.zmaw.de/wdc-for-climate/cera-data-model/

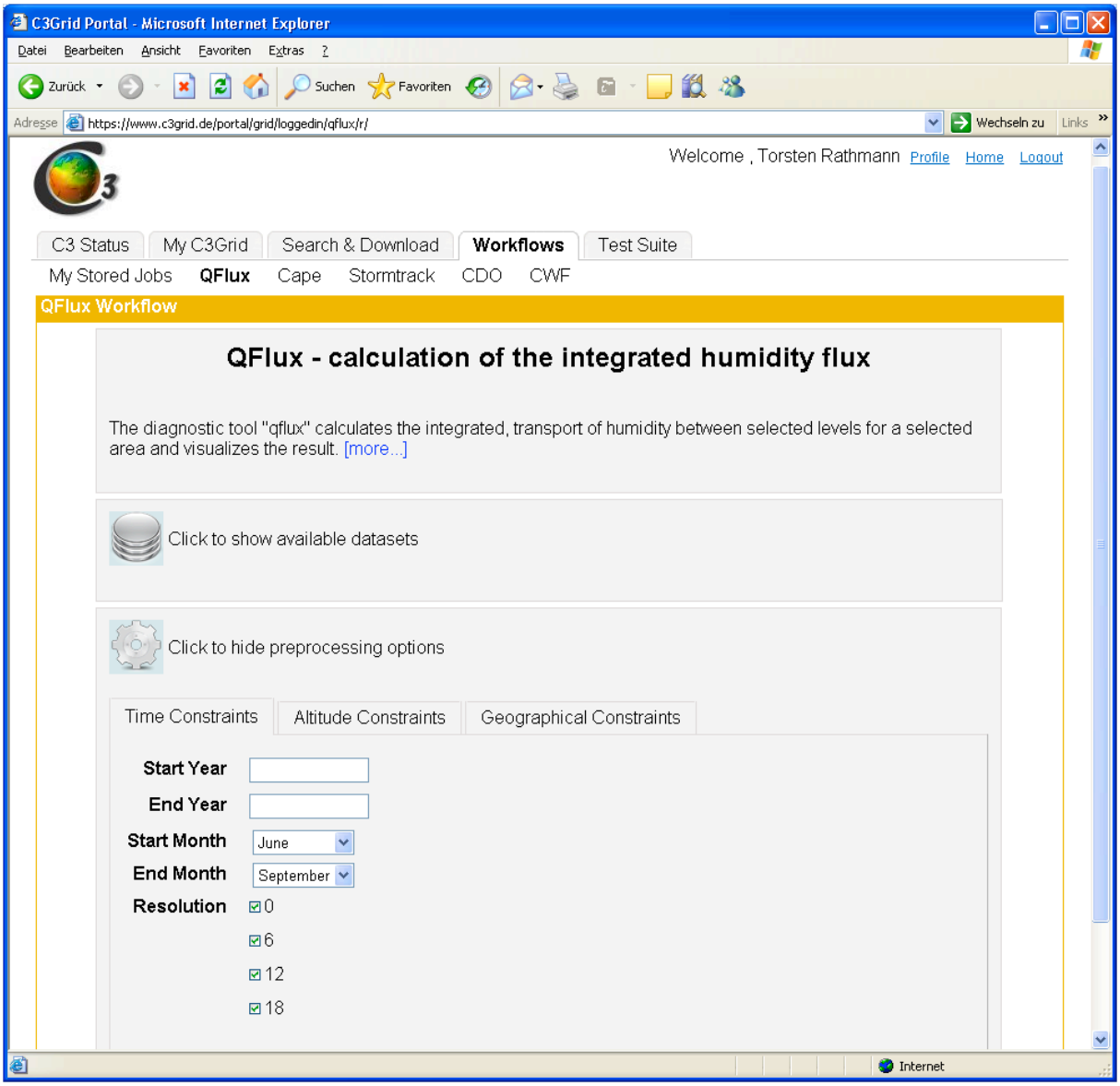


Abbildung 9: Web-Formular für den Feuchtefluss-Workflow

Erwartungen an WissGrid bestehen im Bereich Werkzeuge. Weder im WDCC noch im C3Grid gibt es ein Werkzeug zur Metadatenextraktion aus den Daten. Dabei sind in den Headern der NetCDF- und GRIB-Dateien viele relevante Informationen vorhanden. Diese müssen zurzeit von Hand extrahiert oder neu eingegeben werden. Fehler bei der Metadaten-Eingabe haben schon zu Widersprüchen zwischen Headern und Metadaten geführt. Hier wird dringender Bedarf für ein automatisch arbeitendes Werkzeug gesehen.

Für die Übernahme ins Archiv werden zudem Werkzeuge zur automatischen Format-validierung benötigt. Zurzeit geschieht die Formatvalidierung weder automatisch noch vollständig.

Der Wunsch nach Werkzeugen zur Datenformatkonversion kommt vor allem aus der Klima-folgenforschung. Diese arbeitet gewöhnlich mit anderen Datenformaten als die Klima-forschung. Große Bedeutung haben hier GIS (Geografische InformationsSysteme) und Excel. Es wird dringender Bedarf für Werkzeuge gesehen, die NetCDF oder GRIB in diese Formate konvertieren können.

Veröffentlichung der Daten

Publizierte Primärdaten am WDCC werden mit den persistenten Identifiern DOI und URN versehen und sind im Bibliothekskatalog der Deutschen Nationalen Bibliothek für Wissenschaft und Technologie (TIB) in Hannover verzeichnet. Der Eintrag bei der TIB erfolgt nach der bibliographischen Referenz ISO 690-2, die unter anderem den Titel und die Autorenliste aufführt und es den Wissenschaftlern ermöglicht, ihre Daten zitierfähig aufzuarbeiten und zu veröffentlichen. Die TIB gehört zu einem weltweiten Netz von Registrierungsagenturen, die der Internationalen DOI Foundation[35] angeschlossen sind.

Datendokumentation und Metadaten, Standards

Für Metadaten gibt es im WDCC Regeln, die im CERA-2-Datenmodell[36] zusammengefasst sind. Das CERA-2-Datenmodell ist konform mit dem internationalen Metadaten-Standard: ISO 19115/19139.

Relationen zwischen den Daten (Verknüpfungen, Versionen, Übersetzungen, etc.)

Es gibt zwei Kategorien von Metadaten entsprechend der Struktur der Datenbank des WDCC (drei Kategorien):

1. Metadaten des Experimentes (Zusammenfassung von Datensätzen)

   Hier finden sich alle Metadaten zum Experiment, insbesondere Titel und Autoren.

2. Metadaten der Datensätze

   Hier finden sich alle notwendigen Metadaten, die jeden einzelnen Datensatz beschreiben.

3. Datensätze selbst

Im Langzeitarchiv des WDCC können Daten nicht verändert werden. Ist eine Änderung notwendig, wird dem Archiv ein Erratum oder eine komplette neue Version hinzugefügt. Im wissenschaftlichen Datenarchiv (Bearbeitungsphase) findet dagegen keine Versionierung statt.

Provenienz und Prozessdokumentation, ständige Erweiterung

Dokumentiert wird, wie die Daten erzeugt wurden. Diese Information ist in den Metadaten enthalten. Eine Qualitätscharakterisierung befindet sich ebenfalls in den Metadaten. Die Metadaten werden über den gesamten Lebenszyklus mitgepflegt.

Erweiterungswünsche kommen von der Klimafolgenforschung-Community. Diese wendet häufig mehrere Nachbearbeitungsschritte auf vorhandene Daten an. Gewünscht wird eine genaue Protokollierung dieser Nachbearbeitungsschritte im Workflow. Häufig ist schon die Reihenfolge der Bearbeitungsschritte von Bedeutung, z.B. ob erst gemittelt und dann das Maximum bestimmt wird oder umgekehrt. Bisher wird nur die Erzeugung der Daten protokolliert.

[35] http://www.doi.org
[36] http://www.mad.zmaw.de/wdc-for-climate/cera-data-model/

Speicherung und Langzeitarchivierung: redundante Ablage, Datenkonversion (on-ingest, on-request), Migration/Emulation der Daten

Im Archivbereich am Deutschen Klimarechenzentrum ist die redundante Ablage der Daten realisiert.

Abbildung 10: Ein Teil dieses High Performance Storage Systems am Deutschen Klimarechenzentrum wird für das WDCC verwendet. Das Bild zeigt drei der sechs vorhandenen Bibliotheken mit je 10000 Ablageplätzen für Bandkassetten. Jede Bandkassette hat zur Zeit eine Kapazität von einem Terabyte.

Bei einem Wechsel auf einen neuen Rechner werden statt der Datenformate die wenigen Zugriffswerkzeuge portiert. Dies ist Standard in der Klimaforschung. Eine umfassende Formatkonvertierung wäre wegen der großen Datenmengen nicht möglich.

Datenschutz, Datensicherheit, Eigentum/Copyright, Forschungsethik, Zugriffsrechte, gesetzliche Bestimmungen

Der Zugang zum Katalog (zu den Metadaten) ist frei. Für den Zugang zu den Daten (Download) ist eine Registrierung erforderlich (Benutzername und Passwort). Für einen Teil der Daten besteht ein weitergehender Schutz. Hier muss vom Nutzer eine Erklärung unterschrieben werden, dass die Daten nur für wissenschaftliche Zwecke genutzt werden. Zum Beispiel fallen die Daten des Europäischen Zentrums für mittelfristige Wettervorhersage (ECMWF) unter eine solche Beschränkung.

Im WDCC sind die Produzenten/Autoren Eigentümer der Daten, können die Daten aus der Langzeitarchivierung aber nicht löschen. Dies ist insbesondere wichtig, wenn Datenentitäten mit einem Persistent Identifier versehen sind, um den gesicherten Zugriff auf diese Daten zu garantieren.

Zusammenfassung der allgemeinen und dienstspezifischen Anforderungen an WissGrid

Sowohl Klimaforschung als auch Klimafolgenforschung sind an Werkzeugen im Bereich Content Preservation interessiert. Dringender Bedarf besteht für Werkzeuge zur Metadatenextraktion, Formatvalidierung und Formatkonversion. Vor allem die Klimafolgenforschung hat darüber hinaus Interesse an einem Werkzeug zur Protokollierung von Verarbeitungsschritten.

4.1.3. Fallstudie Medizin

Archivierung ist im medizinischen Umfeld sowohl in der Versorgung als auch in klinischen Studien bzw. in der klinischen Forschung ein Thema. Zu den Anwendungsgebieten in einem Krankenhaus zählen die Archivierung von Krankenunterlagen (Aufbewahrung und Verwaltung von Patientenakten und weiteren aus der Patientenversorgung resultierenden Dokumenten), die Archivierung von Verwaltungsunterlagen (sowohl die Unterlagen zu patientenbezogenen Verwaltungsvorgängen als auch die Dokumente des Krankenhauses als Unternehmen, wie Personalverwaltung oder Finanzbuchhaltung) sowie weitere technische Archive. In dieser Fallstudie wird die Archivierung von Forschungsunterlagen (Unterlagen zu klinischen Studien und sonstige medizinische Forschungsdokumentation) im Rahmen von klinischen Studien betrachtet.

Klinische Studien

Das Arzneimittelgesetz (AMG) definiert eine klinische Studie als „jede am Menschen durchgeführte Untersuchung, die dazu bestimmt ist, klinische oder pharmakologische Wirkungen von Arzneimitteln zu erforschen oder nachzuweisen oder Nebenwirkungen festzustellen oder die Resorption, die Verteilung, den Stoffwechsel oder die Ausscheidungen von Arzneimitteln zu untersuchen, mit dem Ziel, sich von deren Unbedenklichkeit oder Wirksamkeit zu überzeugen“.

Motivation zur Archivierung

In der klinischen Forschung bedeutet Archivierung die lang dauernde Aufbewahrung der Studiendokumentation nach der Beendigung oder dem Abbruch einer klinischen Studie. Notwendig wird die Archivierung durch gesetzliche Bestimmungen wie das Arzneimittelgesetz (AMG). Weitere nationale (Good Clinical Practice (GCP)-Verordnung, Signaturgesetz), sowie internationale (International Conference on Harmonisation (ICH), GCP, Food and Drug Administration (FDA), für digitale Studiendokumente zusätzlich Code of Federal Regulations (CFR) Part 11, FDA Guidance for Industry, EU Annex 11) Vorgaben finden Anwendung. Zusätzliche Vorgaben gibt es beispielsweise für Röntgenbilder und Strahlenanwendungen mit der Strahlenschutzverordnung (StrlSchVO) und der Röntgenverordnung (RöV).

Archivierung ist mehr als nur die langfristige Speicherung von Daten und Dokumenten. Sie beinhaltet u.a. die Möglichkeit, einzelne Daten und Dokumente gezielt wieder auffinden und nutzen zu können. Man spricht dabei von der Möglichkeit der Nachnutzung. Die archivierten Dokumente sollen es erlauben, den Verlauf einer klinischen Studie zu rekonstruieren und die Qualität der erhobenen Daten und verwendeten Methoden bzw. Verfahren auch nachträglich zu evaluieren. Die **Provenienz** der Daten, sowohl deren Authentizität ist wesentlich. Somit

wird der Prozess zum Ergebnis nachvollziehbar und kann ggf. zur Verifikation der Ergebnisse genutzt werden.

Die Archivierung in der medizinischen Forschung und der Einsatz von elektronischen **Datenarchiven** ist noch weitgehend ungelöst bzw. wird derzeit noch uneinheitlich gehandhabt. Für viele Felder in der medizinischen Forschung ist unklar, ob und ggf. welche Vorgaben und Verpflichtungen es gibt. Entsprechend wird vielfach gänzlich ohne Archivierungskonzept gearbeitet oder weitgehend regellos papierbasiert aufbewahrt. Eine elektronische Speicherung findet oft nur rudimentär statt. Eine Anforderung an ein Datenarchiv umfasst, dass eine große Menge an Daten bei möglichst ständiger Verfügbarkeit langfristig in einer vertrauenswürdigen Umgebung abgelegt werden kann. Die Möglichkeit der Weiterverarbeitung innerhalb des Grids ist dabei anzustreben. Aufgrund der hohen Datenschutzanforderungen in der Medizin, sind hierbei zukünftig weitere Datenschutzvorkehrungen, wie beispielsweise ein Pseudonymisierungsdienst, zu integrieren.

Inhalt und Qualität der Daten

Gemäß GCP müssen umfangreiche Quelldaten wie z.B. Laborberichte, Krankenakte (digitale Sammlung medizinischer Daten eines Patienten), Zentrumsordner (zentrumsweite Dokumentation), Identifizierungs- und Pseudonymisierungslisten (Datenschutz, siehe auch Abschnitt Zugang, Nutzung und Nachnutzung), Einwilligungserklärungen (Unterschrift bzw. Signatur), ausgefüllte Case Report Forms (CRF) (elektronischer oder papierbasierter Erhebungsbogen in dem Prüfungsdaten eines Patienten festgehalten werden), dokumentierte Rückfragen (Queries), Nebenwirkungsmeldungen nach Beendigung der Prüfung für 10 Jahre gespeichert und für ein eventuelles Audit bereit gehalten werden. Bei der Studienarchivierung unterscheidet man zwei grundlegende Aspekte der Archivierung. Zum Einen erfolgt die Archivierung aller wesentlichen Dokumente der Studie in einem Trial Master File (TMF) oder Investigator Site File (ISF). Dabei handelt es sich überwiegend um Dokumente, die im Original in Papierform vorliegen (wie bspw. Verträge, Emails, Notizen, etc.). Zum Anderen erfolgt eine Archivierung der Studien-Datenbanken, welche in unterschiedlichen Formaten wie bspw. Bilddaten, PDF (Portable Document Format) oder XML (Extensible Markup Language), vorliegen.

Art der Daten

Je nach klinischer Studie kann der Umfang (z.B. der Korrespondenz) und die Art der Daten (Bilddaten, Fragebögen, Laborwerte, etc.) variieren. Im Rahmen eines Gutachtens im Auftrag der Telematikplattform für Medizinische Forschungsnetze e.V. (TMF) wurden konventionelle Dateiformate untersucht und hinsichtlich ihrer Eignung für die Archivierung im Kontext klinischer Studien bewertet. Unter konventionellen Dateiformaten werden die Formate verstanden, die bereits seit mehreren Jahren im Rahmen der elektronischen Archivierung verwendet werden. Eine Umfrage bei 24 TMF-Verbünden vom Januar 2007 zeigte einen hohen Bedarf an folgenden konventionellen Dateiformaten:

- PDF und TIFF (Tagged Image File Format) als „klassische“ Dateiformate für die Archivierung

- Office-Formate wie z.B. Microsoft Word, OpenDoc und EPS (Encapsulated PostScript) als Textformate
- ASCII (American Standard Code for Information Interchange) und CSV (Character Separated Values) als Dateiformate für Rohdaten
- JPEG (Joint Photographic Experts Group) und DICOM (Digital Imaging and Communications in Medicine) für die Archivierung von Bilddaten
- HTML (Hypertext Markup Language) als Dateityp des Internets
- S/MIME (Secure Multipurpose Internet Mail Extensions) als Dateiformat für gesicherte Emails

Insgesamt handelt es sich dabei um Standard-Formate. Einzig das Bilddatenformat DICOM ist ein spezielles Datenformat aus dem medizinischen Bereich. Eine **Validierung** der Daten bei Einfuhr ins Archiv ist derzeit nicht verbreitet im Einsatz. Um sicher stellen zu können, dass der Nutzer das Dokument in seiner Nutzerumgebung öffnen kann, ist eine „Conversion on access“ erforderlich.

Daneben gibt es das von dem Clinical Data Interchange Standards Consortium (CDISC) entwickelte Operational Data Model (ODM). ODM ist ein Datenstandard, welcher neben dem Austausch von Studiendaten auch der Archivierung von kompletten Studien dient. Dabei handelt es sich um eine Software-unabhängige Archivierung von Studien. ODM speichert Daten und Metadaten gemeinsam, ist hierarchisch, gemäß der CRF-Metapher aufgebaut, besitzt einen Audit Trail und speichert Daten im offenen Format XML. Somit beinhaltet eine ODM Instanz alle Informationen einer Studiendatenbank in maschinenlesbarer Form. Eine in ODM gespeicherte Studie kann jederzeit rekonstruiert werden und ermöglicht somit weiterhin das Audit der Studie durch eine Behörde. Die Vertraulichkeit kann dabei durch die „XML Encryption“ gewährleitet werden, die Integrität und Authentizität durch „XML Signatures“, die auf kryptographischen Verfahren basieren.

Einlagerung ins Archiv

Zunehmend wird eine spezielle Studiensoftware für klinische Studien eingesetzt, um in Zukunft Daten aus unterschiedlichen Quellen, z.B.: Papier-CRF, Electronic Data Capture (EDC), Patiententagebücher, Röntgenbilder, Labordaten und Emails, zusammenzuführen und gemeinsam zu archivieren. Bei papiergebundenen Dokumenten ist zunächst eine Digitalisierung der Dokumente erforderlich. Für den Scanvorgang ist entscheidend, dass die Informationen des Papieroriginals in die elektronische Form übernommen werden. Eine qualifizierte elektronische Signatur (siehe auch Abschnitt Zugang, Nutzung und Nachnutzung) ist dabei nur erforderlich, wenn die gesetzliche Schriftform ersetzt und die Beweisqualität der Urkunde nach § 37 1a ZPO erreicht werden soll. Die Archivierung erfolgt nach Abbruch oder nach Ende der Studie und ist ein einmaliger Prozess. Ein wichtiger Aspekt ist die verteilte Verantwortlichkeit bei der Archivierung. Sowohl der Sponsor als auch der einzelne Prüfarzt haben Archivierungsverpflichtungen. Der Prüfarzt muss persönlich die Aufbewahrung seiner prüfungsbezogenen Unterlagen (Prüfarztordner inkl. der Patienten-

identifikationsliste) gewährleisten, der Sponsor die Archivierung seiner gesamten Studienunterlagen (TMF).

Zugang, Nutzung und Nachnutzung

Die Konferenz der Datenschutzbeauftragten des Bundes und der Länder hat grundlegende Sicherheitsziele für die elektronische Aufbewahrung definiert. Hierzu zählen die Vertraulichkeit der Daten, ihre Authentizität (Zurechenbarkeit), die Integrität, die Verfügbarkeit, die Revisionsfähigkeit, die Validität, die Rechtssicherheit und konkrete Nutzungsfestlegung. Das Vorhandensein einer Signatur (Einfach- oder Mehrfachsignatur) spielt hierbei eine wichtige Rolle. Bei elektronischen Dokumenten unterscheidet das Signaturgesetz (SigG) zwischen verschieden Formen von elektronischen Signaturen. Es gibt die einfache elektronische Signatur, die fortgeschrittene elektronische Signatur und die qualifizierte elektronische Signatur.

Für die medizinische Forschung gilt nach § 40 Abs. 1 BDSG, dass die Daten nur für die Zwecke der Forschung verarbeitet und genutzt werden dürfen. Sobald dies nach dem Forschungszweck möglich ist, sind die Daten nach § 40 Abs. 2 S. 1 BDSG zu anonymisieren. Anonymisierung von Daten bedeutet, dass der Personenbezug der gespeicherten Daten (nachträglich) verloren geht (§ 3 Abs. 6 Bundesdatenschutzgesetz (BDSG)). Der Personenbezug darf unter normalen Umständen nicht oder nur mir einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft wieder herstellbar sein.

In der klinischen Versorgung muss z.T. sehr häufig täglich auf die archivierten Daten zugegriffen werden. Hingegen ist in der klinischen Forschung nur selten (z.B. für die Audits) ein Zugriff erforderlich.

Datendokumentation und Metadaten

XML als Import-/Export-Format für Metadaten in der Archivierung wird bereits heute eingesetzt. Hierbei werden Daten aus Dokumenten (z.B. Health Level 7 (HL7)-Messages) für die Indizierung der Dokumente verwendet.

Wichtige Metadaten sind Informationen über den Personenbezug, die Dokumenterzeugung, die Signierer und die Urheber. Bei dem Personenbezug ist das Verhältnis des Dokuments zum Patienten / Probanden von Interesse, inwiefern dieser personalisiert, pseudonymisiert, anonymisiert ist oder kein Personenbezug besteht. Bei der Dokumentenerzeugung ist wichtig, wie das Dokument erstellt worden ist (Papier, elektronisch) und ob es sich um ein Originaldokument oder eine Kopie handelt. Die Signierer sind diejenigen Personen (Prüfer, Leiter der klinischen Prüfung (LKP), Mitglied des Studienteams, Sponsor, Ethikkommission, Labor, Apotheker, Hersteller (Arzneimittel), Datenmanager, Monitor, Versicherer, Statistiker, Patient, Behörde, keinen, andere Quelle der Signatur), die das Dokument signiert haben. Der Urheber eines Dokuments kann beispielsweise der Sponsor, Prüfer, LKP, Studienteam, Ethikkommission, Apotheker, Statistiker oder anderer Urheber sein.

Metadaten werden zu einem Grossteil im Anwendungskontext erzeugt - teilweise automatisch, zum Teil aber auch manuell.

Datenschutz, Datensicherheit, Zugriffsrechte, etc.

Nach ICH-GCP Guideline E6 (5.1.1) ist jeder Sponsor, der eine klinische Studie durchführt, verpflichtet, mit Hilfe von Audits eine unabhängige Qualitätssicherung zu gewährleisten. Laut ICH-GCP Guideline E6 (5.19) wird durch ein Audit kontrolliert, ob studienbezogene Aktivitäten nach den geltenden gesetzlichen Bestimmungen und GCP, gemäß Prüfplan und nach den geltenden SOPs durchgeführt und Dokumente ordnungsgemäß erstellt werden. Die DIN EN ISO 9000 beschreibt Grundlagen für Qualitätsmanagementsysteme und legt die Terminologie für Qualitätsmanagementsysteme inkl. Audits fest. Als Anleitung zur Planung und Durchführung von Audits kann die Norm DIN EN ISO 19011 angewandt werden.

Interne Audits überprüfen die Einhaltung und Sinnhaftigkeit der eigenen Festlegungen und internen Regelungen einer Einrichtung und dienen vor allem der Verbesserung. Die Durchführung von internen Audits stellt sicher, dass ein Qualitätsmanagementsystem aufrechterhalten und weiterentwickelt wird. Es werden Abweichungen von Anforderungen festgestellt und Verbesserungspotenziale eruiert, sowie Korrektur- bzw. Verbesserungsmaßnahmen überwacht. Der Auditor darf nicht in seinem eigenen Zuständigkeitsbereich auditieren, um die Unabhängigkeit des Audits sicherzustellen. Ein Auditor ist grundsätzlich berechtigt, Einblick in die Patientenakte und in alle Quelldaten zu erhalten. Der Patient unterschreibt in seiner Einwilligung, dass diese autorisierten Personen Einblick in seine Krankengeschichte erhalten dürfen.

Schlussfolgerung

In klinischen Studien gibt es einen großen Bedarf an vertrauenswürdigen LZA-Lösungen bedingt durch das gesetzliche Umfeld und die medizinisch-gesellschaftlich Wichtigkeit der Daten. Trotz dieses vorhandenen Drucks gibt es noch keine einheitlichen Lösungen im medizinischen Umfeld - heterogene, lokale Lösungen existieren, während eine übergreifende Infrastruktur interoperabler Forschungsdatenarchive noch gesucht wird. Zur Sicherstellung der Authentizität sind unter anderem Dienste zur Formatvalidierung und Formatkonvertierung notwendig. Neben den Anforderungen zur Bewahrung der Authentizität der Daten, ist eine Anforderung für eine LZA-Strategie die Einbindung in Workflows innerhalb der MediGRID-Umgebung.

Abkürzungen

AMG	Arzneimittelgesetz
ASCII	American Standard Code for Information Interchange
BDSG	Bundesdatenschutzgesetz
CDISC	Clinical Data Interchange Standards Consortium
CFR	Code of Federal Regulations
CRF	Case Report Forms
CSV	Character Separated Values
DICOM	Digital Imaging and Communications in Medicine
EDC	Electronic Data Capture
EPS	Encapsulated PostScript
FDA	Food and Drug Administration
GCP	Good Clinical Practice
HL7	Health Level 7

HTML	Hypertext Markup Language
ICH	International Conference on Harmonisation
ISF	Investigator Site File
JPEG	Joint Photographic Experts Group
KKS	Koordinierungszentrum für Klinische Studie
LKP	Leiter der klinischen Prüfung
ODM	Operational Data Model
PDF	Portable Document Format
RöV	Röntgenverordnung
S/MIME	Secure Multipurpose Internet Mail Extensions
SigG	Signaturgesetz
SOP	Standard Operating Procedure
StrlSchVO	Strahlenschutzverordnung
TIFF	Tagged Image File Format
TMF	Telematikplattform für Medizinische Forschungsnetze e.V.
TMF'	Trial Master File
XML	Extensible Markup Language
ZPO	Zivilprozessordnung

Quellen

[1] Brandner, Antje; Brandner, Ralf: Gutachten – Konventionelle Dateiformate für die Archivierung im Kontext klinischer Studien, 2007

[2] CDISC: http://www.cdisc.org, letzer Zugriff am 03.09.2009

[3] Kuchinke, Wolfgang: Bericht für Arbeitspaket 0 des TMF-Projektes eArchivierung, 2007

[4] Semler, Sebastian Claudius; Ripkens-Reinhard, Anita: Archivierung von klinischen Forschungsunterlagen – Rechtsgrundlagen, Bezüge zur Krankenaktenarchivierung und elektronische Verfahrensweisen, 2006

[5] TMF: Rechtsgutachten zur elektronischen Archivierung, 2008

[6] TMF: http:// www.tmf-ev.de, letzer Zugriff am 03.09.2009

4.1.4. Fallstudie Biostatistik

Die Biostatistik wertet hauptsächlich Daten aus, die im Rahmen klinischer Studien erhoben werden. Klinische Studien sind sehr stark regulierte Forschungsprojekte, bei denen neue Behandlungsmethoden und Medikationen direkt an Patienten betrachtet und ausgewertet werden. Dabei entstehen Daten zu Genetik, Biomaterial, Bildmaterial und Krankheitsbildern. Es besteht dabei ein Unterschied zwischen klinischen Studien und klinischer Forschung. Während die klinischen Studien aufgrund der direkten Arbeit mit Patienten einer hohen Regulation unterliegen und Daten in pseudonymisierter Form verwendet werden, nutzt die klinische Forschung Daten in anonymisierter Form und ohne direkten Bezug zu Patienten.

Bei der Pseudonymisierung wird mittels eines Pseudonymisierungsdienstes sichergestellt, dass ein Rückbezug von den erhobenen Daten auf einen Patienten möglich ist. Ein

Pseudonym stellt dabei die Verbindung zwischen medizinischen Datensätzen (MDAT) und identifizierenden Datensätzen (IDAT) her. Die identifizierenden Daten sind nur dem Studienarzt bekannt, der den entsprechenden Patienten untersucht hat.

Die Ergebnisse der klinischen Forschung sind nicht direkt klinisch übertragbar und spielen meist erst im Rahmen weiterer klinischer Studien eine Rolle. Des Weiteren sind die untersuchten Fragestellungen der klinischen Forschung weiter gefasst, was in klinischen Studien aufgrund der hohen Regulierung nicht möglich ist.

Die Biostatistik arbeitet i.d.R. in der klinischen Forschung und wertet anonymisierte Daten entsprechend der Methoden der Biostatistik aus. Das Ergebnis der Forschungsarbeit der Biostatistik sind u.a. neue Erkenntnisse zu Korrelationen zwischen Krankheitsbildern und genetischen Merkmalen.

Langzeitarchivierung in der Biostatistik

Ein standardisiertes Vorgehen zur Langzeitarchivierung der Forschungsdaten ist in der Biostatistik gegenwärtig nicht implementiert. Die Archivierung der Forschungsdaten obliegt momentan den Wissenschaftlern selbst und führt insbesondere bei dem Wechsel von Institutionen zu Problemen, da die Wissenschaftler ihre Datenbasis mitnehmen. Ein zentrales Repository der Daten existiert nicht, wobei jedoch offene Datenbanken z.B. zu Genomdaten[37] genutzt werden können. Diese folgen i.d.R. den hauptsächlich verwendeten Datenformaten in der Bioinformatik bzw. Biostatistik[38]. Erstrebenswert ist der Einsatz eines gemeinsamen Forschungsdatenarchivs, in dem die Forscher der Biostatistik ihre Forschungsdaten ablegen können. Hierbei sollen die Daten langfristig in einer vertrauenswürdigen Umgebung gespeichert werden können.

Während der Forschungsarbeit sollen mit dem Datenbestand Berechnungen im Grid durchgeführt werden. Hierzu ist geplant, dass relevante Applikationen der Biostatistik, wie die Statistikanwendung R, mittelfristig auf sämtlichen MediGRID-Grid-Ressourcen installiert sein wird. Nach der Berechnung im Grid, werden die Ergebnisse als neue Daten in das Forschungsarchiv eingeführt.

Weiterhin sind interaktive Umgebungen von Interesse, in denen bestimmte Daten beispielsweise in einer Web-Anwendung virtuellen Teams zur Verfügung gestellt werden können. Somit hätten die Forscher die Möglichkeit die Veröffentlichung ihrer Ergebnisse zentral vornehmen zu können. Wünschenswert hierbei ist, dass feingranulare Berechtigungen für den Zugriff auf die Datenbestände eingerichtet werden können.

Ein weiterer Vorteil dieser zentralen Datenspeicherung ist zudem die Etablierung von derzeit noch nicht vorhandenen standardisierten Metadatenbeschreibungsformaten für die Daten.

Die Anforderungen der DFG bezüglich guter wissenschaftlicher Praxis mit einer Aufbewahrungszeit der Forschungsdaten von 10 Jahren werden seitens der Biostatistik gegenwärtig noch nicht einheitlich umgesetzt.

---

[37] European Bioinformatics Institute (EBI): http://www.ebi.ac.uk/.
[38] Zu Formaten siehe Abschnitt Verwendete Daten.

Anforderungen an die Langzeitarchivierung

Bezüglich Sicherheit ergibt sich aus dem Forschungscharakter der Biostatistik, dass die Biostatistik nicht den restriktiven Anforderungen bezüglich der gesetzlichen Vorgaben aus dem Arzneimittelgesetz (AMG) oder der Strahlenschutzverordnung (StrlSchVO) unterliegt. Aufgrund der engen Einbindung in klinische Studien ist es jedoch nicht per se dauerhaft auszuschließen, dass die restriktiven gesetzlichen Vorgaben in Zukunft Anwendung finden werden.

Es besteht jedoch ein hohes Interesse der Biostatistik an der Wahrung der Intellectual Property Rights (IPR) bezüglich der Forschungsdaten. Hierzu ist die Sicherstellung der Provenienz der Daten von entscheidender Rolle. Das Interesse an IPR besteht, da die Forschungsdaten erheblichen Einfluss auf den Erfolg der wissenschaftlichen Arbeit der einzelnen Biostatistiker haben. Der Grund hierfür liegt darin, dass sich aus den Forschungsergebnissen Patente ergeben können und eine Patentierung nur möglich ist, wenn die zugrundeliegende Datenbasis bei Beantragung des Patents nicht öffentlich zur Verfügung steht. Aus diesem Grund ist es aus Sicht der Biostatistik sinnvoll, dass Verträge den Umgang mit den Forschungsdaten regeln. Dies könnte im Rahmen von Service Level Agreements (SLA) erfolgen, die entsprechende Rahmenbedingungen definieren und zusichern. Innerhalb von Forschungsprojekten sollten die IPR im Rahmen der Kooperationsverträge geregelt sein. Es bestehen dabei jedoch nicht immer klare Vereinbarungen. Aufgrund einer höheren Sensibilität bezüglich des Themas IPR ist jedoch zu erwarten, dass diesbezügliche Regelungen in Kooperationsverträgen in Zukunft verstärkt Beachtung finden. Im Rahmen der UMG ist dies bereits der Fall.

Eine Änderungshistorie zur Nachvollziehbarkeit von Änderungen ist aus Sicht der Biostatistik notwendig. Die Herkunft der Daten, sowohl deren Authentizität ist dabei wesentlich. Der Einsatz der qualifizierten elektronischen Signatur ist jedoch nicht erforderlich, sofern der MDAT-Datenbestand reinen Forschungscharakter besitzt.

Bezüglich der Nutzerverwaltung wird die Unterstützung von Gruppen als Zugriffsberechtigte für Daten angesehen. Dies soll den kollaborativen Arbeitsaspekt unterstützen.

Verwendete Daten

Die Biostatistik erhält ihre Daten i.d.R. bereits gefiltert und anonymisiert. Die Filterung bzw. Anonymisierung erfolgt durch die Partner in den klinischen Studien. Ein Zugriff auf so genannte Studiendatenbanken ist demnach nicht direkt möglich sondern geschieht über eine in der jeweiligen klinischen Studien angebundenen Person (Medizinische Dokumentare) oder Institution, die die Daten extrahiert und anonymisiert.

Verwendete Datenformate in der Biostatistik sind:

- CSV bzw. TXT
- XLS
- MySQL-dumps (TXT)
- XML

die als Basisdaten für den Input in die Statistikprogramme „R“ und SAS verwendet werden. Hierzu sind zusätzlich noch die Skripte für die Statistiksoftware relevant. Zur vollständigen und verlässlichen späteren Verifikation ist es notwendig, dass die komplette Statistikumgebung mit archiviert wird, da sich die Rechenweise der Software im Zeitablauf durch Updates und Upgrades ändern kann. Eine Konvertierung der Forschungsdaten ist in diesem Fall nicht notwendig. Eine Validierung der Daten bei der Einfuhr in das Archiv erfolgt derzeit nicht generell.

Die klinische Forschung verwendet für die MDAT i.d.R. die gleichen Datenformate wie sie auch in klinischen Studien eingesetzt werden.

Schlussfolgerung

In der Biostatistik besteht ein Bedarf an LZA-Lösungen. Aufgrund der auf die Forschung ausgerichteten Arbeit und der i.d.R. anonymisierten Daten, spielt die Datensicherheit in der Biostatistik eine geringere Rolle als beispielsweise bei klinischen Studien. Derzeit obliegt es dem einzelnen Forscher sein Datenrepository zu pflegen. Ein übergreifendes Forschungs-datenarchiv würde die Arbeitsprozesse in Bezug auf die Berechnungen mit den Daten im Grid und der anschließenden Veröffentlichung in interaktiven Web-Anwendungen optimieren. Die Anforderungen aus der Biostatistik sind dabei insbesondere die Validierung und Authentizität der Forschungsdaten, sowie deren Provenienz. Zur erfolgreichen Einführung eines solchen LZA-Dienstes muss die Akzeptanz in der Biostatistik vorhanden sein. Hierzu sollte der Mehrwert für den einzelnen Forscher klar herausgestellt werden.

4.1.5. Fallstudie germanistische Sprachwissenschaft

Motivation zur Archivierung, Wert der Daten, gesetzliche Bestimmungen

Im Rahmen des Arbeitspaktes 3 kooperieren das Verbundprojekt WissGrid mit dem Institut für Deutsche Sprache (IDS), um innerhalb der linguistischen Community Potenziale für den Bedarf nach Langzeitarchivierung auszuloten. Das Forschungsmaterial, das am IDS anfällt, kann zwar nicht als prototypisch für die gesamte Gemeinschaft der Sprachwissenschaftler gelten. Doch soll im Folgenden zumindest exemplarisch ein Grundbedarf eruiert werden, der für eine Kooperation mit der Community als Schablone dienen könnte.

Als eine der zentralen Aufgabengebiete des IDS kann die Dokumentation der Sprache der Gegenwart angesprochen werden. Um diesen Auftrag zu erfüllen, bedarf es einer möglichst breiten empirischen Basis an Sprachdaten, welche für die Germanistik den wichtigsten Rohstoff für ihre Arbeit darstellen. So stützt sich beispielsweise das IDS seit Beginn seines Bestehens 1964 auf gespeicherte Sprachressourcen (DeReKo) und unterhält die dazu notwendige digitale Infrastruktur. Der Wert des hierbei zugrunde liegenden Materials kann nicht in Zahlen bemessen werden. Es handelt sich vielmehr um eine maßgebliche Ressource, auf deren Auswertung weltweit 20.000 Germanisten dringend angewiesen sind. Ihr Verlust würde folglich der gesamten Community die Grundlage für ihre Forschung entziehen.

Da allerdings urheberrechtliche Bestimmungen sowie bestehende Lizenzverträge eine Weitergabe an Dritte unmöglich machen, können die Archivalien nur singulär und zentralisiert abgelegt werden. Schon dies führt dazu, dass ein Verlust nicht durch Redundanzen aufgefangen werden könnte und wirkt sich insbesondere dann erschwerend aus,

berücksichtigt man den Umstand, dass es unmöglich ist, dieses einmalig akquirierte Material zu reproduzieren.

Vornehmlich erzeugt eine Weitergabe oder Veröffentlichung solcher Ressourcen einen elementaren Grundrechtskonflikt zwischen dem Recht auf Eigentum und dem Recht auf informelle Selbstbestimmung. Es ergeben sich folglich deutliche Einschränkungen hinsichtlich der freien Entfaltung der Wissenschaft, im Zuge derer die Bewahrung der rechtlichen Integrität der Urheber von Sprachdaten mit deren Weiterverarbeitung zu wissenschaftlichen Zwecken kollidiert.

Um derartige Dilemmata meistern zu können, ist die sprachwissenschaftliche Community dringend auf nachhaltige Lösungsstrategien angewiesen. Zum Aufbau eines Kompetenzzentrums für die Langzeitarchivierung innerhalb der Germanistik wurden am IDS bereits erste Schritte in diese Richtung unternommen, z.B. durch eine Mitgliedschaft in dem Verbundprojekt Nestor. Weiterhin muss in absehbarer Zeit ein verbindlicher Standard im Umgang mit Sprachdaten hinsichtlich der geschilderten juristischen und ethischen Aspekten geschaffen werden.

Eine Institutionalisierung dieses Vorhabens im Rahmen von WissGrid wäre vor diesem Hintergrund für den Fortbestand der germanistischen Forschungsarbeit von existenzieller Notwendigkeit.

Inhalt und Qualität der Daten, Datenselektion, Aufbewahrungsdauer und evtl. Notwendigkeit zur Löschung, Qualitätssicherungsworkflow

Die Sprachdaten liegen in erster Linie in Form von Textkorpora vor, die zum Teil selbst akquiriert, selektiert und aufbereitet wurden. Hierbei bestehen zum Teil Auflagen, die nach einer gewissen Nutzungsdauer eine Löschung vorschreiben. Als Beispiel wären hier die umfangreichen Textkorpora oder Lexika zu nennen. Problematisch wird eine nachhaltige Aufbewahrung der in immer stärkerem Maße herangezogenen fremden Daten, sowie der noch unbearbeiteten Rohdaten, für die nur teilweise eine einheitliche Policy des gesicherten Umgangs gefunden wurde. Sehr viel stärker werden solche Anforderungen bei den perspektivisch in ihrer Bedeutung mehr und mehr zunehmenden multimodalen Daten ins Gewicht fallen, für die entsprechende Lösungsstrategien noch nicht einheitlich erarbeitet wurden.

Mit Blick auf die zuvor geschilderte Problematik können sich Bearbeiter solcher Korpora die Nichtexistenz eines standardisierten Qualitätssicherungsworkflows nicht mehr länger erlauben.

Art der Daten, Größe der Files, Menge (Anzahl, Size)

Die Korpora umfassen vornehmlich ganze Texte geschriebener Sprache, aber auch Audio-Daten, multimodale Daten, Rohdaten oder versionierte Text-Korpora. Hinzu kommt ein beachtlicher Bestand an Metadaten sowie manuelle und automatische Annotationen und virtuelle Kollektionen (Liste von PID-Verweisen u.U. mehreren Millionen Einzelressourcen).

Allein der Gesamtbestand aller Textdaten und der zugehörigen Annotationen – es liegen schätzungsweise 1000 Dateien von jeweils bis zu 32 GB Umfang vor – umfasst derzeit etwa 5TB. Im Zuge der weiteren Aufbereitung der Daten wird diese Zahl in zwei Jahren mit Sicherheit auf 10TB angewachsen sein. Völlig unberücksichtigt bleibt dabei die ungleich viel

größere Zahl an multimodalen Daten, deren Umfang sich derzeit noch nicht voll überblicken lässt.

Einbringen ins Archiv, wie/wo liegen die Daten vor (Bulk im Einmaltransfer oder stetig anfallende Einzeldaten), wer führt sie ein (Erzeuger, Archivist)
Für die Textkorpora wäre hinsichtlich der momentanen Arbeitsweise ein schrittweises Einbringen sinnvoll. Zunächst wäre ein Einmaltransfer des schon bestehenden Datenkorpus in die Archivierung angeraten. Darüber hinaus fallen täglich noch unbearbeitete Rohdaten an, die etwa jedes Halbjahr in die zu archivierende Form des Korpus überbracht werden. Eventuell wäre es angesichts dieser dichten Folge an neuem Material zu empfehlen, auch die Rohdaten in einem wöchentlichen Turnus zu sichern.

Innerhalb dieses Prozesses greifen jedoch diverse rechtliche Bestimmungen ein, wie sie etwa in Gestalt von Absprachen mit den Urhebern vorliegen, welche den Bearbeiter dazu nötigen, die Daten auch nach abgeschlossenem Archivierungsvorgang noch abändern, erweitern oder nachweislich löschen zu müssen. Eine Versionierung der Daten zu gewährleisten, wäre vor diesem Hintergrund unabdingbar.

Für diesen Prozess wäre es dennoch vorzuziehen, wenn alle Ressourcen durch den Archivisten mit entsprechendem fachlichen Hintergrund in das Archiv überführt werden könnten, die vom Erzeuger für den Transfer in einem gesonderten Verzeichnis zwischenlagern. Eventuelle Garantien und Fristen wären demnach gesondert auszuhandeln.

Zugang, Nutzen und Nachnutzung: online in Webseite, offline; Beschreibung der Applikation: Werkzeuge, Suchmechanismen; Teamarbeit, Veröffentlichung der Daten
Zugang und Nutzung aller Daten unterliegen im Wesentlichen den Übereinkünften der Lizenzverträge und Einwilligungserklärungen zwischen den Urhebern und dem IDS. Alle darüber hinausgehenden, ungeregelten oder allgemeinen Nutzungsrechte unterliegen dem Bundesdatenschutzgesetz sowie dem Urheberrecht. Ähnliches trifft auf die den Datensätzen beigeordneten Annotationen zu, deren Verwendung in gesonderten Lizenzabsprachen mit den jeweiligen Softwareentwicklern geordnet wurde. In aller Regel bedeutet dies, dass eine Nachnutzung nur durch die Vertragnehmer des IDS erfolgen kann.

Technisch gesehen wird auf die Daten durch das vom IDS eigens zu diesem Zweck entwickelte Tool COSMAS II zugegriffen.

Datendokumentation und Metadaten, Standards, Relationen zwischen den Daten (Verknüpfungen, Versionen, Übersetzungen, etc.), Provenienz und Prozessdokumentation, ständige Erweiterung
Für die Textkorpora gelten folgende Standards und Formate:

- Metadaten: Basisformat XML; XCES/TEI-Header.
- Relationen in Metadaten über PIDs (z.B. Versionsrelationen, hierarchische Relationen, Ähnlichkeitsrelationen).
- Provenienz: XCES im TEI-Header angegeben (dokumentiert Aufarbeitung durch das IDS und Textquelle).
- Prozessdokumentation: für Textdaten im TEI-Header enthalten; für multimodale Daten in nichtstandardisierten Metadatenformaten.
- Geplant ist eine Überführung der Korpora von XCES nach TEI P5.

Datenschutz, Datensicherheit, Eigentum/Copyright, Forschungsethik, Zugriffsrechte, gesetzliche Bestimmungen

Wie eingangs erwähnt, steht die Community vor der Schwierigkeit, dass ein Großteil der Textkorpora Eigentum ihrer jeweiligen Urheber ist. Die Bearbeiter, im konkreten Fall das IDS, verfügen i.d.R. lediglich über ein einfaches Nutzungsrecht, das in ca. 130 Lizenzverträge genauer spezifiziert ist. Ein geringerer Anteil der Ressourcen ist frei von Rechten Dritter oder unterliegt GPL- oder CC-Bedingungen.

Einen Sonderfall stellen Video- und Audiodaten dar. Sie unterliegen sämtlich Datenschutzbestimmungen und den Angaben in den für jeden Datensatz gesondert verhandelten Einwilligungserklärungen.

Zusammenfassung der allgemeinen und dienstspezifischen Anforderungen an WissGrid

Die zuvor dargestellte Sachlage hat es bisher stets verunmöglicht, bestehende LZA-Dienste zu nutzen. Insbesondere haben es datenschutzrechtliche Vereinbarungen immer vorausgesetzt, dass sämtliche digitalen Ressourcen rein physisch die Räumlichkeiten des IDS nicht verlassen dürfen, sondern vor Ort archiviert werden müssen. Durch die Modularität der WissGrid LZA-Architektur und ihre Einbettbarkeit in die spezifischen Anforderungen am IDS, böte eine Teilnahme nun erstmals die Möglichkeit, das dringliche Desiderat nach einer dauerhaften Aufbewahrung zu erfüllen.

Bezogen auf die angebotenen Dienste steht zunächst ein eigenes Forschungsdatenarchiv im Fordergrund, worin die Textkorpora eingebracht werden könnten. Von besonderem Interesse ist hierbei das Anwendungsprofil B, das eine interaktive und kollaborative Bearbeitung von Daten unterstützt und insbesondere die Erfassung von Metadaten begünstigt.

Auf den übrigen Arbeitsgebieten der Konvertierung, Metadatenextraktion und Validierung hat das IDS zwar schon einige Vorarbeit geleistet, doch steht zu erwarten, dass aus den vorhandenen Fähigkeiten Synergien gebildet werden können. So wäre es sehr wünschenswert, wenn WissGrid generische Dienste und mögliche Infrastrukturressourcen für die Konvertierung der am IDS erstellten Datenmengen in bekannte Dateiformate anbieten würde. Eine Reversibiliät solcher Prozesse sollte jedoch stets gewährleistet sein.

Mit Blick auf die Sicherheit der Daten, die durch strikte Lizenzvereinbarungen besonderer Garantie bedarf, sind Leistungen wie der Provenzienzservice ein unverzichtbarer Bestandteil jeder Langzeitarchivierung dar.

Gleiches gilt für eine zuverlässige Validierung. Aufgrund der schon angesprochenen Unwiederbringlichkeit des Materials ist eine beständige Überprüfung auf technische Fehler Grundvoraussetzung für eine dauerhafte Archivierung.

Zusammenfassend lässt sich also feststellen, dass ein Langzeitarchiv, welches der Community eine technisch dauerhafte und juristisch wasserdichte Lösung zu bieten vermag, die Sicherheit wissenschaftlichen Arbeitens in der germanistischen Sprachwissenschaft für die Zukunft gewährleisten würde.

4.2. Anhang 2: Dienste

Die im Rahmen von WissGrid zu implementierenden Dienste werden (1) aus der Analyse existierender LZA-Architekturen, sowie (2) den Diskussionen mit Communities abgeleitet.

4.2.1. Extraktion von Metadaten und Validierung

Dienste zur Extraktion von Metadaten und Validierung digitaler Objekte und ihrer Formate (wie z.B. JHOVE[39]) sind für ein langfristiges Datenmanagement und die Qualitätssicherung notwendig. Wenn Daten z.B. in ein Langzeitarchiv eingespeist werden, können diese automatisch das zugrunde liegende Datenformat erkennen; ist das Datenformat bekannt, validieren sie die Korrektheit des Datenformats und extrahieren Metadaten. Die Extraktion von Metadaten ist notwendig, um Daten effektiv verwalten und nutzen zu können. Dies gilt insbesondere für die Langzeitarchivierung, wo darauf geachtet werden muss, dass zur Nutzung notwendiges technisches und semantisches Wissen (insbesondere Kontextwissen) durch Metadaten expliziert wird. Z.B. kann eine effektive Formatkonvertierung (siehe Dienst Konvertierung 4.2.2) von mit fehlerhafter Software produzierten Datenobjekten nur durch eine granulare Datenselektion realisiert werden, für die technische Metadaten notwendig sind.

Die fortgeschrittensten Projekte zur Charakterisierung, Validierung und Identifizierung von Dateiformaten und digitalen Objekten sind JHOVE, DROID[40] und Entwicklungen im Rahmen des EU-Projekts PLANETS[41]. Viele aktuelle Langzeitarchivierungsprojekte bauen inzwischen auf Web-Service-Technologien, die auch direkt in Grid Workflows eingebunden werden und als vertrauenswürdige Dienste den Communities angeboten werden können. Insbesondere aktuelle Arbeiten im Rahmen von JHOVE2[42], dem PLANETS Testbed[43] und verwandten Initiativen sind hinsichtlich der Fragestellung zu evaluieren, inwiefern auf ihrer Basis Formatverarbeitung als Grid-Job bzw. als Streaming-fähiger Web Service Implementierungen in einen Grid-Workflow möglich ist.

4.2.2. Konvertierung

Die Überführung von Daten aus einem in ein anderes Format ist zentral, um Forschungsdaten nach wie vor in geänderten Umgebungen (z.B. Software, Hardware) lesen und interpretieren zu können. Dienste zur Formatkonvertierung ermöglichen nicht nur, wichtige Daten von veralteten in aktuelle Formate zu konvertieren, sondern dienen auch zur Anpassung von Daten an fremde Schnittstellen beim Datenaustausch und zum Qualitätsmanagement von Daten über ihre gesamte Existenz hinweg. Für einen Migrationsdienst sind u.a. die Entwicklungen in den internationalen Projekten CriB[44], source[45] und DExT[46] und im

[39] http://hul.harvard.edu/jhove/
[40] http://droid.sourceforge.net/
[41] http://www.planets-project.eu/
[42] http://confluence.ucop.edu/display/JHOVE2Info/JHOVE+Project+Scope
[43] http://www.dcc.ac.uk/resource/technology-watch/planets-testbed/
[44] http://crib.dsi.uminho.pt
[45] http://www.source.bbk.ac.uk/
[46] http://www.data-archive.ac.uk/dext/

deutschen kopal-Projekt[47] zu nennen. Sie liefern unterschiedliche architektonische Ansätze zur allgemeinen Realisierung von Migrationen und stellen zum Teil auch prototypische Implementierungen bereit. Formatkonvertierung wird in den unterschiedlichsten Archiven und Kontexten gebraucht. Für eine Einbettung ins Grid gibt es unterschiedliche Szenarien:

- Die Konvertierung einzelner Dateien läuft unabhängig ab und eine verteilte Bearbeitung auf Grid-Ressourcen kann daher problemlos durchgeführt werden, insbesondere bei großen Sammlungen.
- Formatkonvertierung ist in einer Vielzahl von Archiven aller Communities ein wichtiges Thema. Um zu vermeiden, dass jede Community oder gar jedes Archiv für sich die dafür notwendigen Schnittstellen bzw. Dienste entwickelt, könnten diese Schnittstellen zu den definierten Diensten kollaborativ entwickelt und ins Archiv eingebettet werden.
- Formatkonvertierung wird als Teil von (standardisierten) Migrationspfaden durchgeführt und kann gemeinschaftlich verwaltet und nachgenutzt werden.

Für die Implementierung eines Workflows zur Formatkonvertierung im Grid sind die genannten, international verfügbaren Werkzeuge auf ihre Einsetzbarkeit in verschiedenen Grid- Szenarien zu prüfen und, wo nötig, zu erweitern. Bei der Realisierung der unterschiedlichen Verarbeitungsszenarien kann auf die Erfahrungswerte der Communities mit Infrastrukturen für skalierbares Datenmanagement zurückgegriffen werden.

4.2.3. Forschungsdatenarchiv

In allen wissenschaftlichen Communities spielen Daten eine wesentliche Rolle; die vertrauenswürdige Archivierung und Verfügbarkeit von Daten zur Verarbeitung ist die Basis 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 weit fortgeschritten ist, ist die Verwaltung von digitalen Objekten und die Verknüpfung zwischen Objekten und Datenbanken vergleichsweise vernachlässigt worden. Werkzeuge, die über die simple (redundante) Speicherung von Dateien hinausgehen und die die Verwaltung von Objekten, Metadatenverwaltung, Einbettung in Workflows und Schnittstellen nach außen übernehmen, werden oft mühsam zusammengestellt bzw. zu einem Großteil selbst entwickelt. Diese Situation führt zu vielen Entwicklungen, die oftmals an den heutigen Möglichkeiten vorbei gehen oder proprietär sind. Für viele Communities fehlt ein entsprechendes Instrument in einer Grid-Umgebung und für manche Communities kann das Fehlen dieser Instrumente ein starkes Hindernis für die Verwendung von Grid-Technologien darstellen.

4.2.4. Provenienzdienst

Ein Provenienzdienst ermöglicht es, datenverarbeitende und -verändernde Prozesse langfristig nachvollziehbar zu dokumentieren, um die Authentizität später bewerten zu können. Es handelt sich dabei um eine Querschnittsfunktionalität, die von allen WissGrid-Diensten und

[47] http://kopal.langzeitarchivierung.de/

dem Forschungsdatenarchiv unterstützt werden muss. Für die Umsetzung sind Entwicklungen und Know-How aus den Projekten AeroGrid und "The EU Provenance Project: Enabling and Supporting Provenance in Grids for Complex Problems"[48] zu untersuchen. Aufgrund des Infrastrukturcharakters eines Provenienzdienstes liegt eine vollständige Implementierung vermutlich außerhalb der Möglichkeit des WissGrid-Projekts und eine Beschränkung auf konzeptionelle Arbeiten und die Interoperabilität der WissGrid-Entwicklungen ist deshalb notwendig.

[48] http://www.gridprovenance.org/

4.3. Anhang 3: Bedeutung des Begriffs "Langzeitarchivierung"

Der Begriff Langzeitarchivierung (LZA) wird in verschiedenen Kontexten verschieden definiert. Im Rahmen von WissGrid werden drei Varianten unterschieden, die zusammen die Langzeitarchivierung insgesamt definieren. Sie ergeben sich aus der analytischen Unterscheidung von folgenden drei Ebenen digitaler Objekte:[49]

- der physikalischen Ebene (digitale Objekte werden auf physikalischen Medien gespeichert),
- der logisch-technischen Ebene (digitale Objekte werden in bestimmten Formaten kodiert) und
- der intellektuellen Ebene (digitale Objekte erfüllen einen bestimmten Sinn für Menschen).

Die dazugehörigen Arten der Langzeitarchivierung sind die Bitstream Preservation, die Content Preservation, und Data Curation. Auch wenn ihre Funktionalitäten unabhängig voneinander und auch für Zwecke außerhalb der Langzeitarchivierung benutzt werden können, bauen sie sinnvollerweise aufeinander auf. Jede Variante ist unabhängig von der Dauer der Archivierung und auch eine Veränderung der Art der Langzeitarchivierung eines Objekts kann im Lauf der Zeit sinnvoll sein, wenn z.B. ein Objekt nicht mehr aktiv genutzt wird.

Die folgende Darstellung will das Wesentliche der drei Varianten bündig illustrieren und Begriffe klären. Sie hat keinen Anspruch auf Vollständigkeit, Endgültigkeit und beschreibt nicht die Entwicklungsaufgaben des WissGrid-Projekts. Insbesondere werden die organisatorischen Konsequenzen und notwendigen Verpflichtungen für die Langzeitarchivierung wie langfristige Zusagen, Finanzierung, Regelungen zur Rechtsnachfolge, etc. nicht behandelt.

4.3.1. Bitstream Preservation

Diese Stufe der Langzeitarchivierung gewährleistet, dass jedes Bit eines Datenobjekts ohne unbeabsichtigte Veränderungen verfügbar ist, und begegnet so zum Beispiel dem Verfall der Speichermedien und -technologien. Die wesentlichen Faktoren einer Bitstream Preservation sind die Anzahl der Kopien, die Verteilung und Unabhängigkeit der Kopien (geographisch, aber auch organisatorisch, finanziell, technologisch, politisch), die Haltbarkeit der Speichermedien und regelmäßige Integritätstests. Die Notwendigkeit zur Bitstream

[49] Die Unterscheidung der drei Objektebenen orientiert sich an: Kenneth Thibodeau, Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, 2002, http://www.clir.org/pubs/reports/pub107/thibodeau.html . Für die Unterscheidung und Zuordnung der Qualitätsstufen gibt es zwar bisher international keine Einigung, die hier vorliegende Darstellung beruht aber auf Vorarbeiten verschiedener Projekte wie KoLlaWiss (http://kolawiss.uni-goettingen.de/), der Kenntnis des internationalen „state of the art“ und eigener Expertise.

Preservation ergibt sich u.a. auch aus den "Vorschlägen zur Sicherung guter wissenschaftlicher Praxis" der DFG.[2]

Als primäre Erbringer der Bitstream Preservation sind Daten-/Rechenzentren prädestiniert, die definierte Qualitätskontrollen, Risikoabschätzungen und Garantien erbringen können.

4.3.2. Content Preservation

Um digitale Objekte zitieren zu können, aber auch um rechtliche/vertragliche Auflagen zu erfüllen, reicht es nicht, dass die Bits noch vorhanden sind. Die aktuelle Technik muss das Objekt auch inhaltlich im Wesentlichen identisch wiedergeben können, selbst wenn die ursprüngliche Technik (Formate, Software, Speicherort, etc.) veraltet ist, um z.B. das Objekt überprüfen zu können. Die wesentlichen Faktoren für diese Qualitätsstufe sind die persistente Identifikation der Objekte, eine kontinuierliche Beobachtung der Technologieentwicklung, technische Qualitätskontrollen sowie Erhaltungsmaßnahmen wie Formatkonvertierungen/Migrationen oder die Bereitstellung von Emulatoren.

Ein Beispiel für Institutionen, die eine Langzeitarchivierung dieser Art anstreben, sind klassische Gedächtnisinstitutionen wie Bibliotheken oder Archive, die publizierte und statische Dokumente aufbewahren.

4.3.3. Data Curation

Während Content Preservation im Wesentlichen die technische Nutzbarkeit fertiger und statischer Objekte behandelt, geht Data Curation über die verbreitete Vorstellung von Archivierung als statische Konservierung nach Ende der aktiven Nutzung hinaus. Data Curation betrachtet die gesamte Lebensdauer und ist die langfristige Erhaltung der intellektuellen Nutzbarkeit und Nachnutzbarkeit. Sie beinhaltet insbesondere die Konzeption von Daten und Metadaten, die Integration von Funktionalitäten in virtuelle Forschungsumgebungen, Versionierung von Objekten, Pflege der Zugriffsberechtigungen, Bewertung der Aufbewahrungswürdigkeit, Sammlungsbildung sowie inhaltliche Anreicherungen und Verknüpfungen. Diese Aufgaben werden z.B. wichtig, wenn man selbst Einfluss auf die Erschaffung der Objekte hat oder wenn die Objekte für die Nutzung fortlaufend ergänzt oder aktualisiert werden sollten. Ein Beispiel wären Messdaten, die fortgeführt, nachträglich von Messfehlern bereinigt und weiter prozessiert werden. Data Curation ist ebenfalls notwendig, falls zur Nutzung von Objekten Hintergrund- und Kontextinformationen bereitgestellt werden müssen, wie z.B. Wörterbücher ausgestorbener Sprachen oder die Dokumentation eines Experimentaufbaus.

Die erwähnten Aufgaben können in der Regel nicht ausschließlich durch Dienstleister erbracht werden, da deren Komplexität die Expertise und direkte Beteiligung der Communities und individuellen Wissenschaftler erfordert. Dienstleister für diese

[2] Siehe Empfehlung Nummer 7 in: Deutsche Forschungsgemeinschaft, Vorschläge zur Sicherung guter wissenschaftlicher Praxis, Bonn 1998, http://tinyurl.com/DFG-gwp98 . Inzwischen wurden die Empfehlungen der DFG sogar deutlich erweitert, sind aber noch nicht verbindlich: siehe Deutsche Forschungsgemeinschaft, Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, Bonn 2009. http://tinyurl.com/DFG-Empfehlungen09 .

Qualitätsstufe werden voraussichtlich insbesondere wissenschaftsnahe Institutionen im Zusammenspiel mit den Wissenschaftlern sein, wie derzeit z.B. die World Data Centers.

4.4. Anhang 4: Bisherige Ansätze für LZA-Architekturen

Vier ausgewählte Architekturentwürfe wurden auf ihre Funktionalitäten hin untersucht. Ähnliche oder vergleichbare Funktionaliäten befinden sich in der beigefügten Tabelle in derselben Zeile und sind gruppiert. Das Verhältnis zu der WissGrid-Architektur ist farblich gekennzeichnet. Da die verschiedenen Projekte und Entwürfe unterschiedliche Ansprüche und Herangehensweisen haben, sind die Gruppierungen und farblichen Zuordnungen zwangsläufig z.T. wage und im Detail diskutabel.

[siehe Datei "Anhang4-data-curation-architecture-matrix.xls"]

4.5. Anhang 5: Ansätze zur Grid/Repositorien-Integration

Die Grid- und die LZA-Community haben sich über lange Zeit weitgehend unabhängig voneinander entwickelt. Technologien für LZA und Repositorien basieren nicht auf Grid-Technologien, und umgekehrt entwickelten Grid-Communities eigene spezialisierte Archivsysteme. Doch welche Möglichkeiten zur technologischen Verknüpfung von Grid und Repositorien-Technologien gibt es?

Dieser Abschnitt listet eine Reihe unterschiedlicher Perspektiven auf ein Grid-Repository. Diese Auflistung stammt aus Diskussionen in beiden Communities[50], und sie versucht möglichst umfassend existierende Erfahrungen und mögliche Perspektiven in dem Bereich abzudecken. Mit diesem Anspruch auf Vollständigkeit geht diese Auflistung auch darüber hinaus, was WissGrid oder ein beliebiges Einzelprojekt leisten kann. Andererseits bietet WissGrid mit der Entwicklung eines Frameworks für unabhängige LZA-Dienste einige Ansätze, die über die reine Interaktion zwischen Grid und Repositorien hinausgehen. Obwohl die folgende Analyse also keinen "Aufgabenkatalog" darstellt, bietet sie die konzeptuelle Basis zur Definition und Abgrenzung der Aufgaben in WissGrid.

Zu den Ansätzen zur Grid/Repository Integration zählen:

1. Repositorien als Archiv-Backends für das Grid
2. Daten-Grid als Repository-Storage
3. Virtualisierung von Repositorien
4. Einbettung von Repositorien in wissenschaftliche Workflows
5. Einbettung von Repository-Modulen in Grid Technologien

Im Folgenden werden die Ansätze 1 bis 3 näher betrachtet. Die (Ansatz 4) Einbettung von Repositorien in wissenschaftliche Workflows ist nicht per se ein Grid-bezogenes Thema; alle uns bekannten Repositorien bieten Web Service Schnittstellen an, und auch eine tiefere Integration wird bereits umgesetzt, wie z.B. die Verknüpfung von Fedora und Taverna im myExperiment Projekt.[51] Auf der anderen Seite (Ansatz 5) deckt die Einbettung von Repository-Modulen in Grid-Technologien zwar ein breites Feld an Themen ab, vom automatisierten Hosting bis hin zur On-the-Fly Instanziierung von virtuellen Repositorien, liegt aber noch in der IT Grundlagenforschung und kann daher im Rahmen von WissGrid (noch) nicht behandelt werden.

[50] Andreas Aschenbrenner (SUB), Tobias Blanke (KCL), Neil P Chue Hong (OMII), Nicholas Ferguson (OGF Europe), Mark Hedges (KCL): A Workshop Series for Grid/Repository Integration. In: D-Lib Magazine, January/February 2009.http://www.dlib.org/dlib/january09/aschenbrenner/01aschenbrenner.html

[51] David DeRoure, Carol Goble: Research Objects for Data Intensive Research. Submitted to IEEE eScience 2009, Oxford. http://wiki.myexperiment.org/index.php/Research_Objects_for_Data_Intensive_Research

4.5.1. Repositorien als Archiv-Backends für das Grid

In diesem Ansatz dienen Repositorien zur persistenten Speicherung von Forschungsdaten, die gezielt für Compute-Prozesse im Grid genutzt werden können. Die Verknüpfung des Repositories mit dem Grid konzentriert sich dabei auf die Daten-Schnittstelle, wobei das Grid nur eines der denkbaren Werkzeuge ist, die über die Schnittstelle auf die Daten im Repository zugreifen könnten.

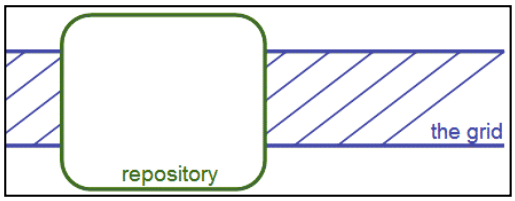


Abbildung 11 - Schnittstelle zwischen persistentem Archiv und Compute-Grid

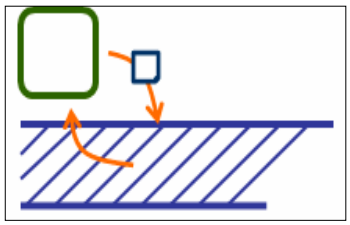
Da das Grid und das Repository technologisch weitgehend unabhängig voneinander sind (außer in Bezug auf die Daten-Schnittstelle) und die Daten zur Prozessierung im Grid zuerst aus dem Repository herausgeholt werden müssen, sprechen manche bei diesem Ansatz weniger von einer "Integration" als vielmehr von einer "Verknüpfung" unterschiedlicher Umgebungen.

Wenn man allerdings technisch tiefer in diese Verknüpfung einsteigt, sieht man vielfältige Abhängigkeiten zwischen den beiden Infrastruktur-Umgebungen, speziell mit Bezug auf AAI und Sicherheit. Authentifizierung der Nutzer und Rechtemanagement ist bei jenen Archiven unerlässlich, in denen Daten bzw. Subsets der Daten veränderbar sind oder diese gar (nach Ablauf einiger Jahre) gelöscht werden können. Auch da, wo Archive nicht-öffentliche Daten aufbewahren, eventuell mandantenfähig für unterschiedliche Nutzergruppen, kommen Sicherheitsaspekte ins Spiel. An der Schnittstelle zwischen dem Repository und dem Grid müssen dazu ggf. unterschiedliche AAI- und Rechtemanagement-Technologien aufeinander abgebildet werden. Je nachdem ist denkbar, dass (a) ein "Repository unter dem Grid versteckt" ist, also nur über die im Grid vorhandenen VO-Mechanismen auf ein Repository zugegriffen werden kann, oder (b) Interoperabilität zwischen den VO- bzw. Rechtesystemen über z.B. die in IVOM[52] entwickelten Technologien hergestellt werden kann.

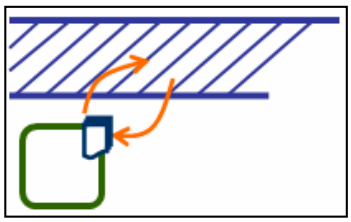
[52] D-Grid Projekt IVOM: Interoperabilität und Integration der VO-Management Technologien im D-Grid. http://www.d-grid.de/index.php?id=314

Obwohl Repositorien und Grid Technologien in diesem Ansatz daher lediglich über eine Schnittstelle miteinander "verknüpft" werden, sprechen wir also durch die tiefer gehende Interoperabilität der beiden Infrastrukturumgebungen durchaus von einer "Integration".

Es können zwei unterschiedliche Ansätze für diese Integration unterschieden werden:



1. Daten zu den Diensten: die Daten werden aus dem Repository extrahiert und in eine Grid-Umgebung für Verarbeitungsaufgaben transferiert. Bei rechenintensiven Jobs können mit diesem Ansatz die verfügbaren Compute-Ressourcen optimal genutzt werden, allerdings kann gerade bei großen Datenmengen durch den Datentransfer ein spürbarer Overhead entstehen.



2. Dienste zu den Daten: Grid-Jobs werden direkt im Repository verarbeitet, wo die Daten aufliegen. Dieser Ansatz verringert den Aufwand zum Datentransfer, legt aber Last für die Berechnungen auf den/die Repository-Server. Außerdem wird durch die von Nutzern getriggerte Ausführung von Jobs, die nicht zur Administration des Repositories sondern vielmehr anwendungsspezifisch sind, potenziell eine wesentliche Sicherheitslücke geöffnet. Im Rahmen von WissGrid wird dieser Ansatz daher vorläufig nicht weiter verfolgt, mögliche Aktivitäten in Partnerprojekten (z.B. AstroGrid) werden aber begleitet.

4.5.2. Daten-Grid als Repository-Storage

In diesem Ansatz nutzt ein Repository das Grid als Storage-Infrastruktur. Eine nationale Grid-Infrastruktur wie D-Grid könnte somit für zahlreiche Repositorien die Infrastruktur stellen, und dadurch die Storage-Verwaltung bündeln. Wesentlich dafür ist eine nachhaltige Strategie zur Datenvorhaltung in der Grid-Infrastruktur.

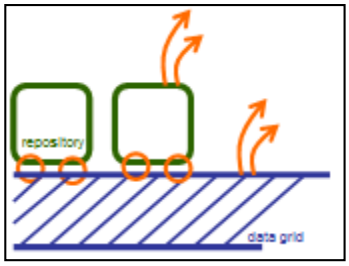


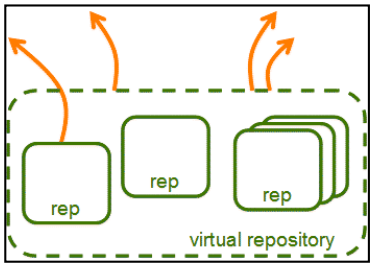
Abbildung 14 - Grid Storage Infrastruktur für Repositorien

Als Teil einer solchen Strategie können von einer Storage Infrastruktur diverse Aufgaben übernommen werden, die für Bitstream Preservation notwendig sind: von der Datenreplikation und regelmäßigen Integritätsprüfungen bis hin zu transparenter Migration von Storagemedien. Da Repository-Software dem üblichen Software-Lebenszyklus folgt und voraussichtlich etwa alle 5 Jahre migriert werden müsste, würde es dem Gesamtsystem zusätzliche Stabilität verleihen, wenn die Daten in einer nachhaltigen Grid-Infrastruktur verwaltet werden.

Zusätzlich zu den im vorigen Ansatz angesprochenen Fragestellungen zur Integration der Nutzer- und Rechteverwaltung, gibt es in diesem Ansatz Optimierungspotenzial in der Definition der Schnittstellen zwischen dem Datengrid und dem Repository. Idealerweise würde die Skalierbarkeit des verteilten Grids auch dem Grid-Repository zugute kommen,

indem auch aus Nicht-Grid-Umgebungen direkt auf Objekte im verteilten Datengrid zugegriffen werden könnte. Für die Definition einer solchen generischen, verteilbaren Schnittstelle müssen Grid-Konzepte und Repository-Konzepte (z.B. das OAI-ORE Format[53]) zusammengebracht werden.

4.5.3. Virtualisierung von Repositorien



Das zentrale Charakteristikum von Grid-Technologien ist die Virtualisierung von Ressourcen (Storage- und Compute-Hardware), bzw. auch z.B. die Virtualisierung von Datenbanken (cf. OGSA-DAI[54]). In diesem Ansatz werden Mechanismen zur Virtualisierung von existierenden Repositorien angeboten, und dadurch Grid Technologien auf einer Schicht, die auf der Hardware-Virtualisierung aufbaut, erweitert. (Einen ähnlichen Vorschlag, der durch eine internationale Expertengruppe der Europäischen Kommission eingebracht wurde, gibt es auch für Dienste.[55])

Die inhaltliche Vernetzung von Repositorien ist ein wichtiges Thema in der Repository Community, wo viele Repositories zu einem einzigen virtuellen Repository föderiert werden. Etablierte Standards zur Repository Föderierung wie OAI-PMH werden von internationalen Repositorien-Infrastrukturen wie Driver[56] und auch in den unterschiedlichsten Communities eingesetzt. Eine Einbettung dieser Standards in Grid-Umgebungen steht allerdings noch aus.

[53] OAI-ORE - Open Archives Initiative, Object Re-Use and Exchange. http://www.openarchives.org/ore/
[54] OGSA-DAI. http://www.ogsadai.org.uk/
[55] Next Generation Grids Expert Group (NGG): Service Oriented Knowledge Utilities (SOKU). Januar 2006. http://www.semanticgrid.org/NGG3/
[56] Driver, Digital Repository Infrastructure for European Research. http://www.driver-repository.eu/