LABIMI/F

Deliverable

Romanus Grütz[a]
[a] Institut für Medizinische Informatik, Universitätsmedizin Göttingen

D5.2: Betriebskonzept

	D5.2: Betriebskonzept[1]
Autor(en)	Romanus Grütz, Svenja Wolff
Editor(en)	Thomas Steinke, Otto Rienhoff, Philipp Weil, Svenja Wolff
Datum	23.12.2013

A: Status des Dokuments

Version 1.0.1

[1]Dieses Dokument wurde im Rahmen des Projekts LABIMI/F erstellt. Das Projekt LABIMI/F wird gefördert von der Deutschen Forschungsgemeinschaft (DFG) unter dem Förderkennzeichen RI1000/2-1.

B: Bezug zum Projektplan

M5.2: Betriebskonzept

C: Abstract

Im Rahmen des DFG geförderten Projekts zur Langzeitarchivierung biomedizinischer Forschungsdaten LABIMI/F wurde eine verteilte generische IT-Infrastruktur für die Langzeitarchivierung von Forschungsdaten entwickelt und in Form einer Laborimplementierung als Prototyp realisiert.

Das in diesem Dokument vorliegende Betriebskonzept beschreibt sowohl die entwickelte IT-Infrastruktur zur Langzeitarchivierung biomedizinischer Forschungsdaten als auch die notwendigen technischen und organisatorischen Maßnahmen für deren Betrieb der oben genannten IT-Infrastruktur.

D: Änderungen

Version	Datum	Name	Kurzbeschreibung
0.0.1	08.05.2013	Romanus Grütz	Entwurf
0.0.2	22.05.2013	Romanus Grütz	Entwurf
0.0.3	09.06.2013	Thomas Steinke	Entwurf
0.0.4	02.07.2013	Romanus Grütz	Entwurf
1.0.0	17.09.2013	Romanus Grütz	Erste Version
1.0.1		Svenja Wolff	Revision Kapitel 3

E: Inhaltsverzeichnis

F: Abbildungsverzeichnis

G: Tabellenverzeichnis

1 Systemübersicht

Die IT-Infrastruktur zur Langzeitarchivierung von Forschungsdaten (im Folgenden als LZA-Infrastruktur bezeichnet) wird Forscher bzw. generell Dateneigentümer (Kapitel 1.2.3) dabei unterstützen, Forschungsprimärdaten (FPD)

1. zu sammeln,
2. zu annotieren,
3. zu archivieren und
4. in einer zentralen Infrastruktur durchsuchbar vorzuhalten, um sie für vom Dateneigentümer autorisierte Personen nachnutzbar zugänglich zu machen.

Diese Prozesse sind auf funktionelle Komponenten abgebildet und werden durch agierende Personen mit definierten Rollen durchgeführt. Im Folgenden werden die genannten Prozesse, funktionelle Komponenten und Rollen in einer Gesamtarchitektur vorgestellt, wie sie im Bereich der biomedizinischen Forschung relevant sind.

Die funktionellen Komponenten sind:

- eine Eingabeschnittstelle zur Erfassung der FPD und Metadaten,
- ein PID-Dienst zur Sicherung einer kohärenten Objektannotation,
- ein FPD-Repository (FPD-R) zur zugangskontrollierten Datenspeicherung,
- ein Metadaten-Repository (MDR) zur Speicherung und Verwaltung der Metadaten und
- eine gesicherte Netzwerkinfrastruktur zur geografischen Verteilung von Diensten.

Das Zusammenwirken der funktionellen Komponenten ist in Abbildung 1 schematisch dargestellt.

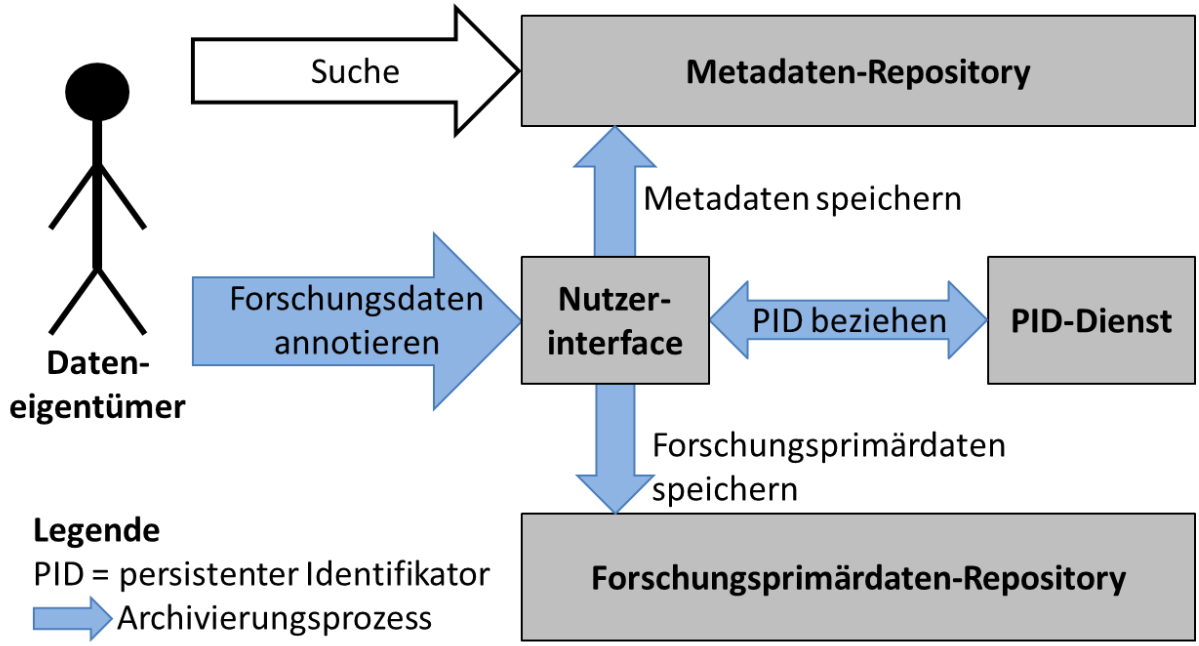

Abbildung 1: IT-Infrastruktur für die Langzeitarchivierung von Forschungsdaten. Aufgezeigt sind die notwendigen Komponenten der Infrastruktur und deren Abhängigkeiten. Die Selektion und Annotation der zu speichernden FPD nimmt der Dateneigentümer mit Hilfe eines Nutzerinterfaces (Kapitel 1.1.1) vor. Um die referenzielle Integrität der FPD – bspw. für die Erwähnung in Publikationen – sicherzustellen, werden persistente Identifikatoren (PID)s verwendet.

Anmerkung: Die Verwaltung der PIDs bzw. der Anschluss an das dafür notwendige Handle-System[2] ist nicht Teil dieses Betriebskonzepts, da hierfür bereits bestehende PID-Dienste verwendet werden können, wie zum Beispiel der PID-Dienst der Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), der in der prototypischen Implementierung verwendet wird.

Die Metadaten der FPD werden im MDR (Kapitel 1.1.2) gespeichert. Die PIDs verweisen auf die Metadaten der FPDs, in denen auch der Speicherort des FPD festgehalten ist. Auf diese Weise können die Metadaten von jedem interessierten Außenstehenden eingesehen und durchsucht werden. Die FPD selbst können unter organisatorischer und physikalischer Kontrolle des Dateneigentümers bleiben.

1.1 Komponenten

Nachfolgend werden alle in der LZA-Infrastruktur verwendeten Komponenten in ihrer Funktionalität und ihrer Interaktion detailliert beschrieben.

1.1.1 Nutzerinterface

Das Nutzerinterface wird am Arbeitsplatz des Dateneigentümers installiert und bietet die Möglichkeit seine FPD zu selektieren, zu annotieren und in der LZA-Infrastruktur zu archivieren. Um den zu archivierenden FPD persistente Identifikatoren zuweisen zu können, stellt das lokale Eingabewerkzeug eine Schnittstelle zu einem PID-Dienst bereit. Wenn möglich wird die URL der Metadaten von den FPD im MDR dem PID-

[2] http://www.handle.net

Dienst übergeben. Ist die endgültige URL nicht bekannt oder bestimmbar, wird eine Dummy-URL als URL-Wert der FPD beim PID-Dienst hinterlegt.

Neben dem Beziehen des PIDs von dem PID-Dienst übergibt das Nutzerinterface die Metadaten an das MDR. Hierbei ist zu beachten, dass die Metadaten in einem vom MDR verständlichen Format übermittelt werden. Die FPD werden in dem FPD-R abgespeichert. Abhängig von der Struktur des FPD-R wird die Verzeichnisstruktur oder die Projektzugehörigkeit mit angegeben.

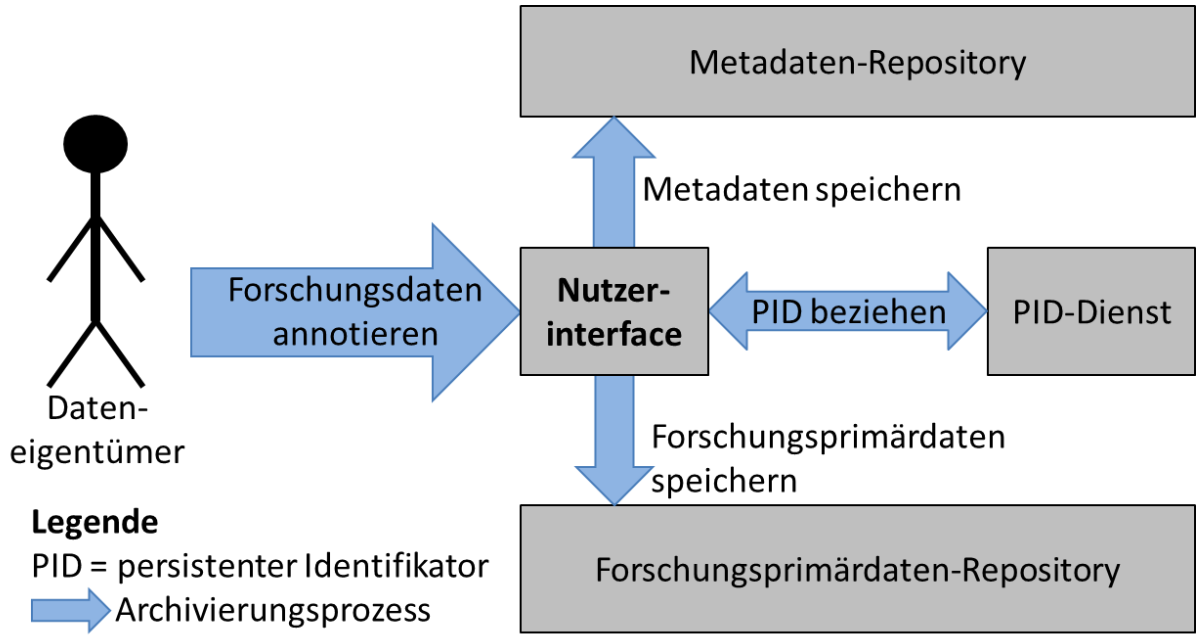

Abbildung 2: Die Schnittstellen des Nutzerinterfaces zu den Komponenten der LZA-Infrastruktur.

In der aktuellen Implementierung ist die Nutzerschnittstelle durch die grafische Oberfläche GeMeCo realisiert.

GeMeCo ist eine grafische Oberfläche mit deren Hilfe die Annotation von FPD mit relevanten Metadaten erfolgt. GeMeCo kann hierbei jedes nach Anhang A erstellte Metadatenschema (MDS) darstellen. Aufgrund der Unübersichtlichkeit bei einer großen Anzahl von Metadaten gruppiert GeMeCo die als zusammengehörig gekennzeichneten Eingabefelder des MDS und stellt diese als eigene Tabs dar.

Neben der Gruppierung bietet GeMeCo die Möglichkeit, externe Metadatenextraktoren einzubinden und zur automatisierten Metadateneingabe zu verwenden.

GeMeCo verfügt weiterhin über eine Sitzungsverwaltung, die es ermöglicht, den Annotations-Prozess jederzeit zu unterbrechen, explizit zu speichern und zu einem späteren Zeitpunkt wieder aufzunehmen. Nach Abschluss des Annotations-Prozesses bezieht GeMeCo vom PID-Dienst (Kapitel 1.1.4) einen PID für die annotierten FPD. Anschließend werden die Metadaten inkl. eindeutiger Verlinkung zu den FPD an das MDR verschickt und importiert. Die FPD selbst werden im FPD-R (Kapitel 1.1.3) abgelegt.

Zusätzlich können über GeMeCo die annotierten FPDs auch in externe Archivsysteme exportiert werden, wie bspw. das Extensible Neuroimaging Archive Tool (XNAT) im Falle von medizinischen Bilddaten.

1.1.2 Metadaten-Repository

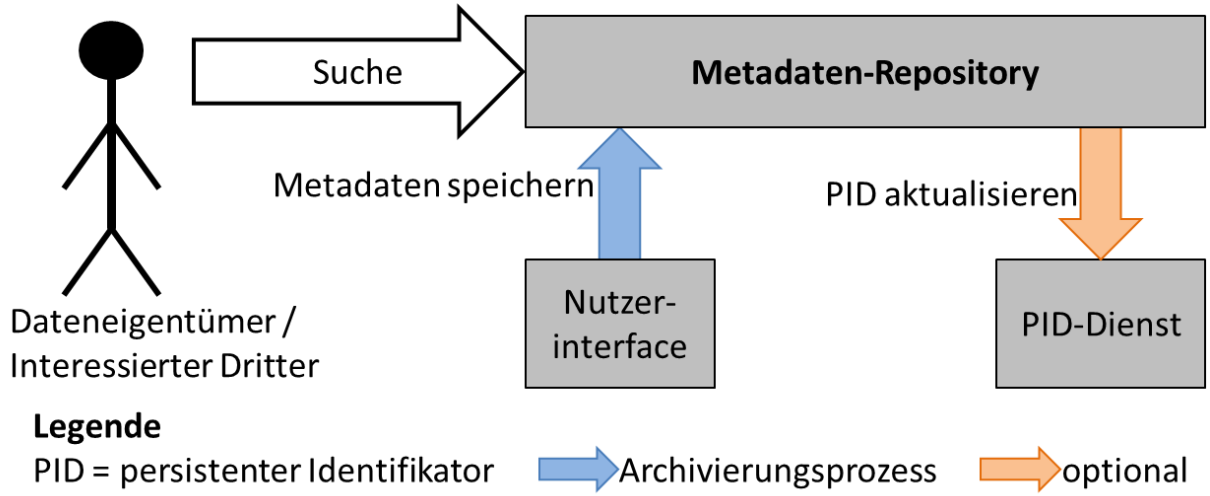

Abbildung 3: Die Schnittstellen des Metadaten-Repositoriums zu den Komponenten der LZA-Infrastruktur.

Das MDR verwaltet die Metadaten der FPD (Abbildung 3). In den Metadaten muss die eindeutige Adresse bzw. Verlinkung zum jeweiligen FPD enthalten sein. Diese Adressen müssen stets aktuell gehalten werden. Auf diese Weise bleibt der PID selbst beim Verschieben der FPD persistent bzw. gültig.

Das Nutzerinterface übermittelt die Metadaten an das MDR. Daher muss das MDR eine Schnittstelle bieten, Metadaten zu importieren. Wenn das Nutzerinterface den PID nicht bereits mit der richtigen URL zu den Metadaten bezieht, wird das MDR diesen aktualisieren (Abbildung 3).

Das MDR bietet den Dateneigentümern neben dem Metadateneingang eine Schnittstelle, um ihre Suchanfragen zu formulieren und ihre Ergebnisse einsehen und ggf. sortieren zu können.

In der aktuellen Implementierung wird DSpace[3] eingesetzt.

[3] http://www.dspace.org/

1.1.3 Forschungsprimärdaten-Repository

Das FPD-R dient der Speicherung der FPD. Das FPD-R enthält Mechanismen, die eine permanente Verfügbarkeit und eine abgestufte Zugriffskontrolle bspw. über einen Download-Server mit Nutzerverwaltung ermöglichen. Der Datentransport zum / vom Speichersystem erfolgt verschlüsselt.

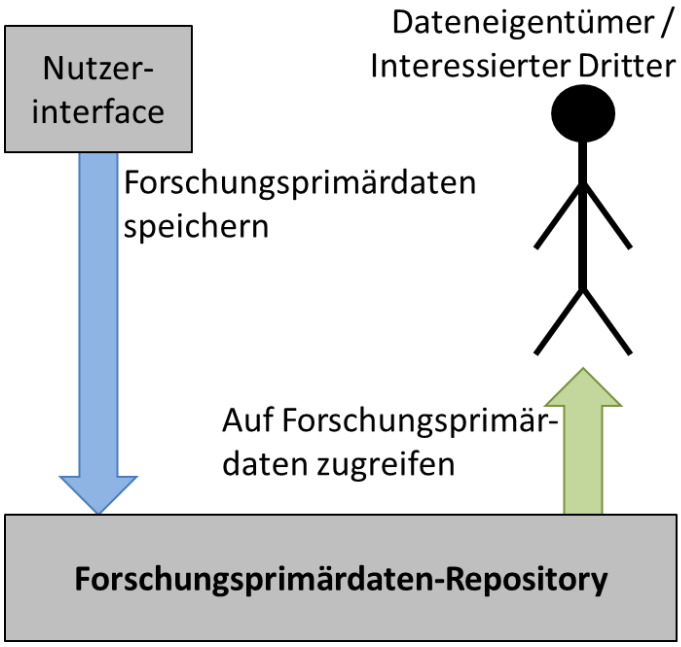

Abbildung 4: Die Schnittstellen des FPD-R zu den Komponenten der LZA-Infrastruktur. In der aktuellen Implementierung des FPD-R wird XtreemFS[4] eingesetzt. XtreemFS ist ein objekt-basiertes, verteiltes Dateisystem. Somit wird die Verteilung der Speicherressourcen über verschiedene Standorte hinweg ermöglicht (Abbildung 4). Es bietet Mechanismen zur Sicherstellung der Verfügbarkeit (Ausfalltoleranz) und der Zugriffskontrolle und ist für die Kommunikation via Wide Area Network (WAN) konzipiert. Für die notwendige Authentifizierung kommt über eine Public Key Infrastructure (PKI) mit X.509-Zertifikaten zum Einsatz. Die Verschlüsselung erfolgt über TLS/SSL.

Anmerkung: Im Rahmen der prototypischen Implementierung wird eine eigene PKI bzw. Certificate Authority (CA) erstellt und genutzt. Um eine bereits bestehende PKI nutzen zu können, müssen die einzelnen XtreemFS-Komponenten Zertifikate aus der gewünschten PKI erhalten und dessen CA XtreemFS als vertrauenswürdig bekanntgemacht werden (Installationsanleitung).

[4] http://www.xtreemfs.org/

1.1.4 PID-Dienst

Der PID-Dienst (Abbildung 5) weist den einzelnen FPDs einen PID zu. Die PIDs verweisen nicht direkt auf das jeweilige FPD im FPD-R sondern auf dessen Metadaten im MDR. Hierfür bietet der PID-Dienst dem Nutzerinterface eine Schnittstelle für das Beantragen von PIDs an (siehe Kapitel 1.1.1). Übergibt das Nutzerinterface dem PID-Dienst eine temporäre Dummy-URL, bietet der PID-Dienst dem MDR zusätzlich eine Schnittstelle zur Aktualisierung der URL-Einträge bereits existierender PIDs.

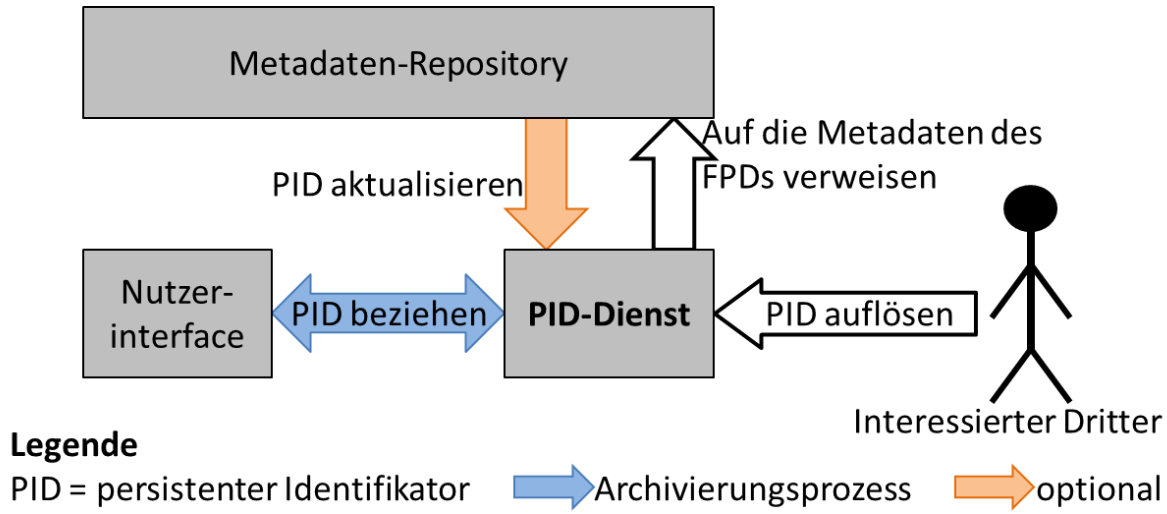

Abbildung 5: Der PID-Dienst der LZA-Infrastruktur.

Anmerkung: In der LZA-Infrastruktur kann ein externer PID-Dienst verwendet werden. So ist der PID-Dienst der GWDG (Mitglied des European Persistent Identifier Consortium[5] (EPIC)) beispielhaft angebunden.

1.2 Rollen, Rechte und Pflichten

In diesem Kapitel werden die Rollen innerhalb der LZA-Infrastruktur mit den damit verbundenen Rechten und Pflichten beschrieben.

1.2.1 System-Administrator

Der System-Administrator ist für die Verfügbarkeit und Instandhaltung der Hard- und Software innerhalb der LZA-Infrastruktur verantwortlich. Es ist seine Pflicht, die LZA-Infrastruktur zu aktualisieren, gegen Angreifer zu sichern und nach außen hin verfügbar zu machen. Der System-Administrator teilt allen beteiligten Personen Störungen nach Bekanntwerden bzw. geplante Wartungszeiten mit. Weiterhin ist der System-Administrator für die Benutzerverwaltung und das Hinzufügen neuer Projekte (siehe Kapitel 1.2.2) verantwortlich.

Es wird empfohlen die Rolle des System-Administrators im Falle der Archivierung geschützter Daten, die zu unabhängigen und nicht kooperierenden Projekten bzw. Standorten gehören, auf verschiedene organisatorisch getrennte oder weisungsbefreite Personen aufzuteilen. Ohne die Aufteilung dieser Rolle auf verschiedene voneinander unabhängige Personen kann eine unerlaubte

[5] http://www.pidconsortium.eu/

Zusammenführung dieser Daten nicht garantiert werden. Abhängig von der Organisationsstruktur bzw. Verteilung der Kompetenzen des LZA-Betreibers ist eine weitere Aufteilung der Rolle des System-Administrators auf die einzelnen Hard- und Software-Komponenten zu empfehlen.

1.2.2 Projekt und Projektleiter

Ein Projekt ist innerhalb der LZA-Infrastruktur eine eigene Organisationsstruktur, welche eigene – von anderen Projekten unabhängige – Speicherbereiche inkl. Nutzerverwaltung verwendet.

Der Projektleiter ist für sein Projekt verantwortlich. Er muss sein Projekt zu Beginn an der LZA-Infrastruktur anmelden und entscheidet über den Datenmanagementplan sowie mögliche Daten-Policies für die FPD. Dies beinhaltet auch, Zugriffsrechte für die Daten des Projektes einzustellen und auch projektfremden Personen Zugriff darauf zu gewähren bzw. zu entziehen.

1.2.3 Dateneigentümer

Die Rolle des Dateneigentümers wird i.d.R. dem Forscher zugewiesen, der die LZA-Infrastruktur zur Archivierung seiner FPD nutzt. Er muss Mitglied eines Projektes sein und hat das Recht, neue FPD und Metadaten der Infrastruktur hinzuzufügen und auf seine eigenen FPD zuzugreifen. Weiterhin hat der Dateneigentümer das Recht, anderen Benutzern der LZA-Infrastruktur Lese- und/oder Schreibrechte an seinen eigenen als geschützt gekennzeichneten FPD zu erteilen oder zu entziehen.

Wenn der Dateneigentümer seine FPD geschützt auf eigenen Ressourcen speichert, verpflichtet er sich beim Verschieben dieser Daten deren Speicherort im MDR zu aktualisieren und konsistent zu halten.

1.2.4 Datenkurator

Der Datenkurator stellt die Interpretierbarkeit der archivierten FPDs und Metadaten sicher. Er ist verpflichtet, die domänen-spezifischen Veränderungen der Technologie, der Software und der Datenformate zu beobachten und zu analysieren. Ziel dieser Analyse ist das frühzeitige Erkennen neuer relevanter Metadaten, Dateiformate und Metadatenextraktoren und die abnehmende Nutzbarkeit bestehender Dateiformate.

Zusammen mit dem System-Administrator muss der Datenkurator eine Strategie entwickeln und umsetzen, um auf den erkannten Wandel rechtzeitig und adäquat zu reagieren.

Der Datenkurator erstellt in Absprache mit den Administratoren und Anwendern Standard Operation Procedures (SOP) zur Festlegung einheitlicher Vorgehensweisen bei der Nutzung der LZA-Infrastruktur. Die SOPs definieren als Datenmanagementplan die Prozesse

1. zur Erzeugung und Dokumentation von FPD gemäß fachwissenschaftlicher Standards,
2. zur Archivierung von FPD in der LZA-Infrastruktur und
3. zum Zugriff auf FPD aus der LZA-Infrastruktur.

Die Rolle des Datenkurators erfordert Kenntnis innerhalb der Nutzer-Domänen. Daher sollte für jede Domäne mindestens eine Person, die in dieser Domäne tätig ist, als Datenkurator ernannt werden.

1.2.5 Interessierter Dritter

Der interessierte Dritte ist eine an den Metadaten und FPD interessierte – ggf. außenstehende – Person. Der interessierte Dritte hat das Recht, im MDR nach FPD zu suchen. Er kann den Dateneigentümer kontaktieren, um einen FPD-Zugriff zu beantragen.

1.3 Prozesse

Im Folgenden werden die drei wichtigsten und am häufigsten durchzuführenden Prozesse: das Hinzufügen neuer Projekte (Abschnitt 1.3.1), die Dateneingabe (Abschnitt 1.3.2) und der Datenzugriff (Abschnitt 1.3.3) beschrieben. Jeder dieser Prozesse wird zu Beginn mit einer vereinfachten Grafik und anschließend mit einer detaillierten Beschreibung erläutert. Um die Zuständigkeit für die einzelnen Schritte innerhalb der Prozesse darzulegen, wird jeweils eine verantwortliche Rolle oder Softwarekomponente genannt.

System-Administratoren sind für den korrekten Betrieb ihres jeweiligen Systems (Kapitel 1.2.1) verantwortlich. Damit sind diese auch indirekt für alle Prozessschritte innerhalb ihres zu betreuenden Systems verantwortlich.

1.3.1 Workflow A: Hinzufügen neuer Projekte

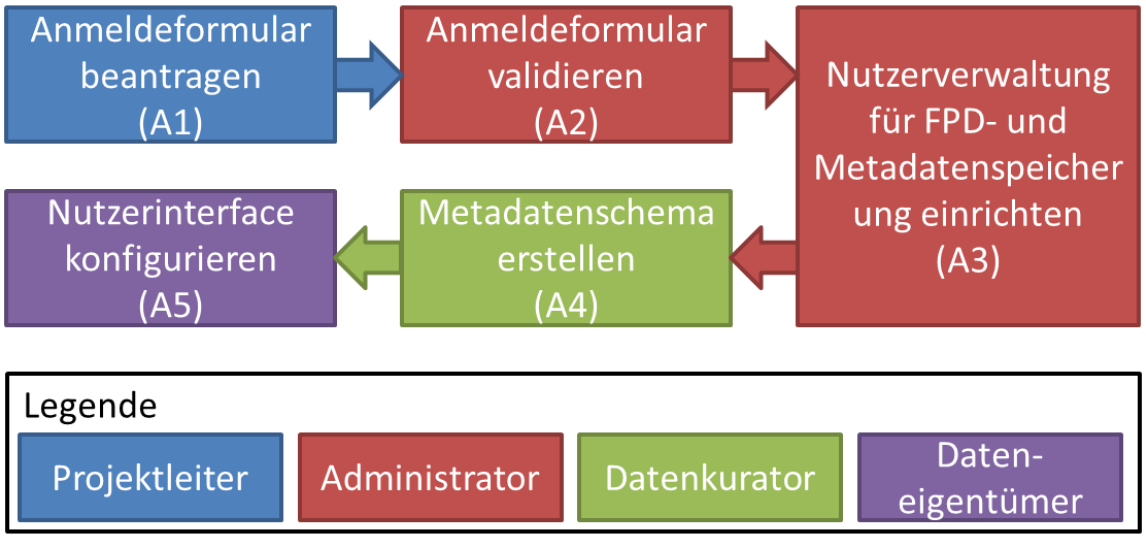

Abbildung 6: Vereinfachter Prozess zum Hinzufügen neuer Projekte.

Der in Abbildung 6 vereinfacht dargestellte und in Tabelle 1 detailliert beschriebene Prozess wird immer dann ausgeführt, wenn ein neues Projekt in der LZA-Infrastruktur angelegt wird.

Tabelle 1: Prozessschritte zum Anlegen eines neuen Projektes in der LZA-Infrastruktur.

Beschreibung des Prozessschrittes		Verantwortlicher/ Software
A1.	Der Projektleiter beantragt mit dem Anmeldeformular beim System-Administrator Zugang zur LZA-Infrastruktur für sein Projekt.	Projektleiter
A2.	Der System-Administrator prüft alle Angaben. Wenn die Angaben unvollständig oder missverständlich sind, wird der Projektleiter darüber informiert und der Workflow endet hier.	System-Administrator
A3.	Der System-Administrator erstellt im MDR (z.B. DSpace) das Projekt. Er erstellt sowohl im MDR als auch im FPD-R einen Benutzer für den Projektleiter und gibt bei Bedarf Hilfestellung beim Konfigurieren eigener Speicher-Komponenten.	
A4.	Der Datenkurator erstellt ein oder mehrere neue MDS und ggf. eine projektspezifische Konfigurationsdatei für das Nutzerinterface und stellt sie dem Dateneigentümer des Projektes zur Verfügung.	Datenkurator
A5.	Der Dateneigentümer lädt das Nutzerinterface herunter und importiert die vom Datenkurator erstellte Konfigurationsdatei und das bzw. die für ihn relevanten MDS.	Dateneigentümer

1.3.2 Workflow B: Dateneingabe

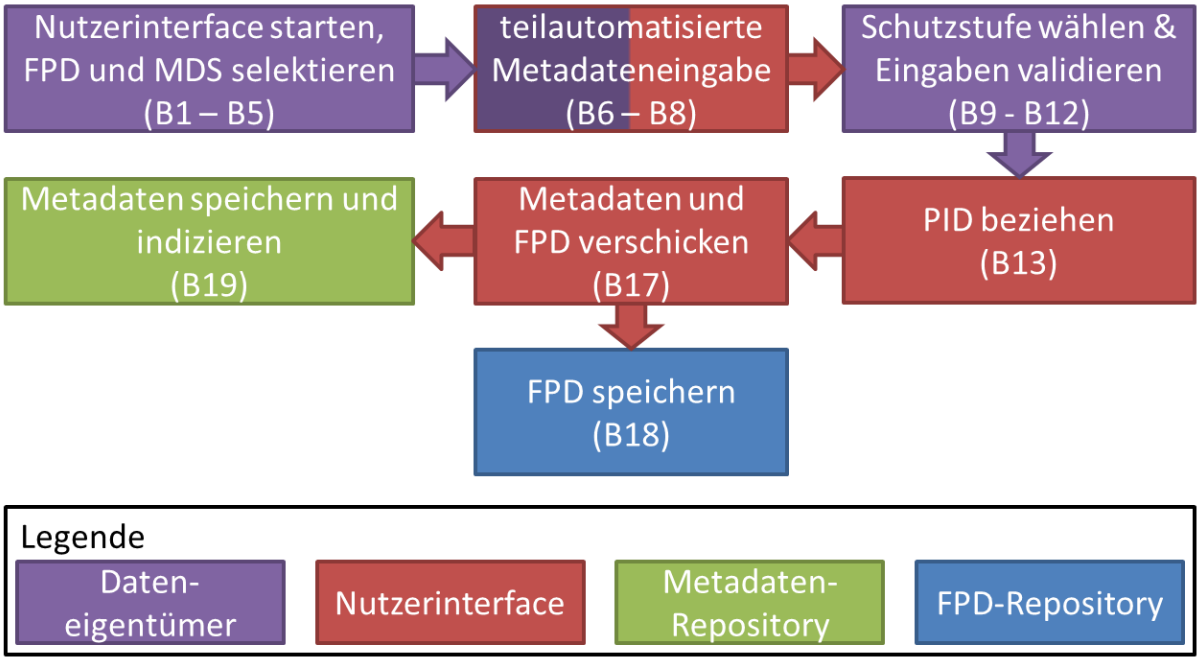

Abbildung 7: Vereinfachter Prozess zur Dateneingabe in die LZA-Infrastruktur.

Der in Abbildung 7 vereinfacht dargestellte und in Tabelle 2 detailliert beschriebene Prozess wird immer dann ausgeführt, wenn FPD in der LZA-Infrastruktur abgelegt werden.

Tabelle 2: Prozessschritte zur Dateneingabe in der LZA-Infrastruktur.

Beschreibung des Prozessschrittes	Verantwortlicher/ Software
B1. Der Dateneigentümer startet das Nutzerinterface (z.B. GeMeCo). **B2.** Der Dateneigentümer selektiert den zu archivierenden Datensatz.	Dateneigentümer
B3. Das Nutzerinterface extrahiert aus den selektierten Datensätzen Metadaten mit Hilfe externer Werkzeuge wie bspw. das *File Information Tool Set* (FITS)[6]. **B4.** Diese Metadaten werden zur Vorauswahl eines MDS verwendet.	Nutzerinterface
B5. Der Dateneigentümer überprüft, korrigiert und bestätigt anschließend die Selektion des zu verwendenden MDS.	Dateneigentümer
B6. Das Nutzerinterface stellt dem Dateneigentümer Eingabemöglichkeiten für die im ausgewählten MDS spezifizierten Metadaten dar. Die Darstellung und die erlaubten Eingaben hängen von den im MDS angegebenen Typen bzw. Wertebereichen ab. **B7.** Das Nutzerinterface hebt die notwendigen Metadaten hervor. **B8.** Das Nutzerinterface trägt die zuvor extrahierten, validierten Metadaten (A.3) in die korrespondierende Eingabemöglichkeit (falls vorhanden) ein und markiert diese als automatisch extrahiert bzw. blendet sie aus.	Nutzerinterface
B9. Der Dateneigentümer wählt die Schutzstufe / Sichtbarkeit der FPD. [public oder protected] **B10.** Wurde das Nutzerinterface mit mehr als einem Archivsystem konfiguriert, kann der Dateneigentümer das Ziel-Archivsystem auf eines der vorkonfigurierten Archivsysteme ändern. **B11.** Der Dateneigentümer validiert und vervollständigt die Metadaten.	Dateneigentümer
B12. Sobald eine Metadateneingabe vorgenommen wurde, wird hierfür die im jeweiligen MDS spezifizierte Plausibilitätsprüfung durchgeführt und dem Dateneigentümer Rückmeldung darüber gegeben, ob diese erfolgreich war oder warum diese nicht erfolgreich war. **B13.** Das Nutzerinterface beantragt einen PID beim PID-Dienst.	Nutzerinterface
B14. Der PID-Dienst generiert einen neuen PID, speichert die Dummy-URL als Referenz und antwortet dem Nutzerinterface mit dem neuen PID.	PID-Dienst
B15. Das Nutzerinterface fügt den Datensätzen den PID als Metadatum hinzu, erstellt das Submission Information Package (SIP) entsprechend dem (Import-)Format des ausgewählten Archivsystems und sendet es an das Archivsystem. **B16.** Das Nutzerinterface führt beim PID-Dienst ein Update für den erhaltenen PID aus und aktualisiert dessen URL auf die in der	Nutzerinterface

[6] http://code.google.com/p/fits/.

Beschreibung des Prozessschrittes	Verantwortlicher/ Software
Konfigurationsdatei angegebene Base-URL plus PID (z.B.: www.dspace.uni-goettingen.de/q=12.12134.213948).	
B17. Falls der Dateneigentümer in Schritt A.9 die Sichtbarkeit „public“ gewählt hat, verschiebt das Nutzerinterface die FPD in FPD-R für öffentliche Daten (Vol_{pub}.).	
B18. Das FPD-R speichert und verwaltet die FPD.	FPD-R
B19. Das MDR empfängt und importiert das SIP.	MDR

1.3.3 Workflow C: Datenzugriff

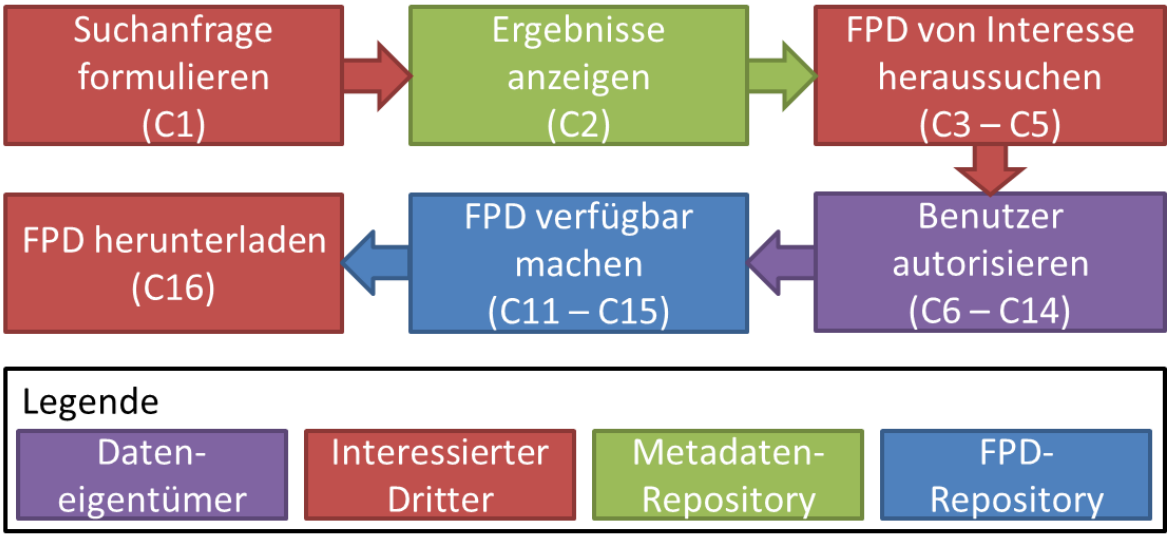

Abbildung 8: Vereinfachter Prozess zum Datenzugriff in der LZA-Infrastruktur.

Der in Abbildung 8 vereinfacht dargestellte und in Tabelle 3 detailliert beschriebene Prozess wird immer dann ausgeführt, wenn ein interessierter Dritter FPD in der LZA-Infrastruktur sucht und darauf zugreifen möchte. Eine detaillierte Beschreibung der einzelnen Schritte wird nachfolgend gegeben.

Tabelle 3: Prozessschritte zum Datenzugriff in der LZA-Infrastruktur.

Beschreibung des Prozessschrittes	Verantwortlicher/ Software
C1. Der interessierte Dritte formuliert eine Suchanfrage an das MDR.	Interessierter Dritter
C2. Das MDR listet dem interessierten Dritten alle Ergebnisse.	MDR
C3. Der interessierte Dritte kann die Ergebnisse gruppieren und umsortieren.	Interessierter Dritter

Beschreibung des Prozessschrittes	Verantwortlicher/ Software
C4. Der interessierte Dritte selektiert ein oder mehrere Objekt(e). **C5.** Der interessierte Dritte versucht das / die selektierten Objekt(e) herunterzuladen.	
C6. Das FPD-R überprüft, ob der interessierte Dritte berechtigt ist, den / die angefragten Datensätze herunterzuladen. Wenn dieser berechtigt ist, weiter ab C.14.	FPD-R
C7. Wenn der interessierte Dritte dazu nicht berechtigt ist, kann dieser über die Kontaktdaten des Dateneigentümers eine Zugriffserlaubnis beantragen. **C8.** Der interessierte Dritte authentifiziert sich gegenüber dem Dateneigentümer bzw. nennt diesem seine Identität und formuliert sein Interesse an den Datensätzen.	Interessierter Dritter
C9. Der Dateneigentümer entscheidet nun, ob und in welchem Umfang dem anfragenden interessierten Dritten Zugriff auf die selektierten Datensätze gewährt wird.	Dateneigentümer
C10. Gewährt der Dateneigentümer dem anfragenden interessierten Dritten keinen Zugriff, wird dieser via E-Mail darüber informiert. (In der prototypischen Implementierung wird dies manuell vom Dateneigentümer durchgeführt.) **C11.** Wird dem anfragenden interessierten Dritten der Zugriff auf die Datensätze gewährt, überprüft das System ob dieser bereits einen Account für das Speichersystem hat. Falls dies nicht der Fall sein sollte, **C12.** wird Ihm ein solcher Account erstellt. (In der prototypischen Implementierung wird dies manuell vom System-Administrator durchgeführt.)	FPD-R
C13. Der Dateneigentümer vergibt die Zugriffsrechte (Read, Write, Execute) auf seine Daten an den Account des interessierten Dritten.	Dateneigentümer
C14. Das FPD-R benachrichtigt den interessierten Dritten darüber, dass und wie der Zugriff erfolgen kann. **C15.** Das FPD-R bietet dem interessierten Dritten eine sichere Download- bzw. Zugriffsmöglichkeit.	FPD-R
C16. Der interessierte Dritte lädt den selektierten Datensatz herunter.	Interessierter Dritter

2 Betrieb

Neben den infrastrukturellen Anforderungen sind für den Betrieb einer LZA-Infrastruktur verschiedene organisatorische Bedingungen zu erfüllen. Im Folgenden werden die notwendigen Anforderungen beschrieben.

2.1 Nutzerverwaltung und Authentifizierung

Jeder Dateneigentümer, der die LZA-Infrastruktur nutzen möchte, muss einen Benutzeraccount im MDR und einen im FPD-R erhalten. Die meisten Systeme bieten mehrere Authentifizierungsmechanismen. In der Laborimplementierung wird zur Authentifizierung am FPD-R eine eigene PKI mit X.509-Zertifikaten verwendet. Dies bedeutet, dass die Erstellung eines Benutzeraccounts für das FPD-R die Beantragung eines Benutzerzertifikats bei der CA der PKI beinhaltet.

Die Authentifizierung am MDR erfolgt in der Laborimplementierung – um eine höhere Benutzerakzeptanz zu erreichen – durch „Benutzername/Passwort“. Diese kann jedoch – um ein höheres Maß an Sicherheit zu etablieren – auf die bereits für das FPD-R genutzten X.509-Zertifikate konfiguriert werden.

Innerhalb des MDR können verschiedene Bereiche für die unterschiedlichen Standorte und Projekte angelegt werden. Den Nutzern werden Rollen und damit verbundene Rechte innerhalb dieser Bereiche zugewiesen.

2.2 Verfügbarkeit

Für einen geregelten Betrieb einer LZA-Infrastruktur muss die Verfügbarkeit definiert sein. Wenn Umstände, die nicht vom Infrastrukturanbieter zu vertreten sind, eine Nutzung unmöglich machen, gilt die LZA-Infrastruktur als verfügbar, wenn gleichzeitig gilt:

- das Metadatenrepositorium ist über das Internet erreichbar,
- die Eingabe von Metadaten ist mittels der entsprechenden Werkzeuge möglich und
- der Zugriff auf die Datenzugriffsdienste sowie den öffentlichen Speicherbereich ist möglich.

Weitere Details zur Definition der Verfügbarkeit sind dem SLA zu entnehmen.

Die Verfügbarkeit kann durch Monitoring der LZA-Infrastruktur, bspw. durch die Monitoringsoftware Nagios, überprüft werden. Alle Tests, die eine Anmeldung an der LZA-Infrastruktur benötigen, müssen mit einem eigens dafür eingerichteten Test-Account durchgeführt werden. Neben der reinen Erreichbarkeit der einzelnen Server und Dienste ist auch die Überprüfung auf korrekte Verarbeitung der Funktionen zu empfehlen. Die Monitoring-Intervalle und die Form der Dokumentation des Monitorings bzw. der Verfügbarkeit liegen im Ermessen des LZA-Infrastrukturanbieters bzw. sind mit dessen Kunden abzustimmen.

2.3 Backup

Die Erstellung von Backups und Migrationen auf andere Speichertechnologien dienen der Sicherstellung von Datenverfügbarkeit und Bitstream Preservation.

Die Sicherung und bei Bedarf die Wiederherstellung der FPD in den öffentlichen Speicherbereichen und der Metadaten erfolgt durch den System-Administrator. Die Datenübertragung bei der Sicherung oder im Falle einer Wiederherstellung darf nur mittels einer gesicherten Verbindung (Kapitel 2.4, BSI TR-02102) durchgeführt werden. Die zu verwendenden Maßnahmen sind dem IT-Grundschutzhandbuch (http://www.bsi.bund.de/gshb) zu entnehmen.

Die Backup- und Wiederherstellungsstrategie wird von den Administratoren zusammen erarbeitet, beschlossen und dokumentiert. Diese Strategie muss das MDR und die öffentlichen FPD-R-Komponenten beinhalten. Für geschützte FPD, die den Standort des Eigentümers nicht verlassen, muss von den dort eingesetzten Administratoren eine standortbezogene Backup- und Wiederherstellungsstrategie erstellt und bei Bedarf umgesetzt werden.

2.4 Sicherheit

Die Kommunikation innerhalb der LZA-Infrastruktur findet verschlüsselt statt. Hierzu kommen X.509-Zertifikate und eine Public-Key-Infrastructure zum Einsatz. Das verwendete Verschlüsselungsverfahren und die Schlüssellänge sind hiervon unabhängig und können frei gewählt werden. Für einen optimalen Schutz wird empfohlen, die technische Richtlinie des Bundesamts für Sicherheit in der Informationstechnik zu befolgen.

Nach BSI TR-02102 ist ein als sicher geltendes Verschlüsselungsverfahren RSA mit einer Schlüssellänge von ≥2.000 Bit. Um die Angriffsfläche bzw. die Anzahl der nach außen offenen Ports zu minimieren kann die LZA-Infrastruktur – mit Ausnahme des MDR – optional innerhalb eines VPNs betrieben werden.

Die Kommunikation der einzelnen Komponenten erfolgt standardmäßig über die in Tabelle 4 aufgeführten Ports:

Tabelle 4: Auflistung aller LZA-Komponenten inkl. verwendeter Ports.

Komponente	Protokoll	Verwendete Ports
DSpace	TCP	80 / 443
PID-Dienst	TCP	80
XNAT	TCP	80 / 443
XtreemFS – DIR	TCP	32638
XtreemFS – MRC	TCP	32636
XtreemFS – OSD	TCP & UDP	32640

2.5 Datenschutz

Die über den Dateneigentümer erhobenen Kontaktdaten werden a) für die interne Benutzerverwaltung und b) als Kontaktmöglichkeit, die einem interessierten Dritten bereitgestellt wird, verwendet.

Die FPD bleiben rechtliches Eigentum des Dateneigentümers, welcher die FPD in die LZA-Infrastruktur hinzufügt und werden im Falle von öffentlichen FPD nur durch Beauftragung des LZA-Infrastrukturbetreibers in den öffentlichen Speicherbereichen verwaltet.

Da die System-Administratoren Zugriff auf Teile der FPD haben (siehe Kapitel 1.1.3), müssen diese im Falle der unverschlüsselten Speicherung der FPD verpflichtet werden, die FPD nicht zu lesen, zu analysieren oder weiterzugeben.

Nicht zugriffsberechtigte Personen dürfen nicht auf die gespeicherten Daten zugreifen können. D.h. weder lesend, analysierend noch weitergebend.

Die FPD, die die Dateneigentümer am eigenen Standort verwalten, obliegen den Datenschutzmaßnahmen des Betreibers der entsprechenden Speicherressource.

Wird an den jeweiligen Standorten XtreemFS verwendet, besteht die jedoch Möglichkeit, interessierten Dritten einen auf das Lesen beschränkten Zugriff zu gewähren.

2.6 Unterhaltung und Wartung

Die für den täglichen Betrieb notwendigen Aufgaben sind im Kapitel 1.2 beschrieben. An dieser Stelle werden ebenfalls die jeweils verantwortlichen Personen bzw. Rolleninhaber definiert.

Die Wartungsarbeiten werden möglichst in nutzungsarmen Zeiten durchgeführt, es sei denn, es handelt sich um dringende Arbeiten (z.B. Notfall-Patches, um die Sicherheit der angebotenen LZA-Infrastruktur und der aufbewahrten Metadaten und FPD zu gewährleisten). Die nutzungsarmen Zeiten sind mit Hilfe von Nutzungsstatistiken oder durch Absprache mit den Anwendern zu ermitteln. Unterbrechungen werden in angemessener Zeit im Voraus den Benutzern per Email gemeldet.

In nachweislich dringenden Fällen kann gegen geplante Wartungstermine Widerspruch eingelegt werden. Nach Prüfung durch zuständiges Personal (siehe Kapitel 1.2) werden die Wartungsarbeiten ggf. auf andere Termine verlegt, wenn dadurch nicht die Betriebssicherheit für andere Benutzer beeinflusst wird oder unzumutbarer Zusatzaufwand entsteht (Weiteres regelt das SLA).

2.7 Mögliche Restriktionen und Abhängigkeiten

Die im Abschnitt 2 beschrieben funktionellen Komponenten sind im LABiMi/F Projekt in einer prototypischen Infrastruktur umgesetzt worden. Die folgenden Einschränkungen und Abhängigkeiten der Software-Komponenten beziehen sich daher auf die aktuelle Implementierung.

Um die beschriebene LZA-Infrastruktur im vollen Umfang nutzen zu können, muss der Dateneigentümer GeMeCo und den XtreemFS-Client verwenden.

Die LZA-Infrastruktur ist von der Kommunikationsfähigkeit der einzelnen Komponenten abhängig. Kann bspw. GeMeCo keine Verbindung zu DSpace

herstellen, können die Daten nicht importiert werden. Eine externe Abhängigkeit besteht in der Verwendung des PID-Dienstes.

2.7.1 DSpace-Archivsystem

Vorkommen:	einmalig beim Infrastrukturanbieter
Abhängigkeiten:	Oracle Java JDK 6 oder 7 oder OpenJDK 6 oder 7
	Apache Maven 2.2.x oder höher
	Relationale Datenbank: PostgreSQL oder Oracle
	Servlet Engine (Apache Tomcat 5.5 oder höher, Jetty, Caucho Resin oder ähnliches)
	Internetverbindung

2.7.2 GeMeCo

Vorkommen:	bei jedem Teilnehmer, der seine Daten annotieren und in der LZA-Infrastruktur hinzufügen will
Abhängigkeiten:	Oracle Java JRE
	Verbindung zum DSpace-Archivsystem (i.d.R. Port 80/443)
	Verbindung zum PID-Dienst (i.d.R. Port 80/443)
	Verbindung zum verwendeten Speicher (XtreemFS-Client oder XNAT)

2.7.3 PID-Dienst

Vorkommen:	bei externem Anbieter
Abhängigkeiten:	abhängig vom externen Anbieter

2.7.4 XtreemFS – Client

Vorkommen:	bei jedem Teilnehmer, der auf die verteilte Speicherstruktur zugreifen will
Abhängigkeiten:	Verbindung zu der XtreemFS-Speicherstruktur (i.d.R. Port 32636, 32638 und 32640)
	Unter Linux: FUSE 2.6 oder höher
	boost 1.35 oder höher
	openSSL 0.9.8 oder höher

	libattr
	Linux 2.6 kernel

2.7.5 XtreemFS – DIR, MRC, OSD

Vorkommen:	einmalig beim Infrastrukturanbieter für öffentliche Daten und bei jedem Teilnehmer, der einen geschützten Bereich betreiben möchte
Abhängigkeiten:	Oracle Java JRE 1.6.0 oder höher

2.8 Haftung und Support

Die Haftung ist im SLA geregelt. Der Support wird per Email oder Telefon von Administratoren, Datenkuratoren und geschulten Dritten im Rahmen des SLA geleistet.

3 Finanzierung

Eine langfristige und nachhaltige Nutzung der LZA-Infrastruktur ist nur über ein qualifiziertes Finanzierungskonzept zu realisieren.

3.1 Grundlegende Überlegungen

Der Betrieb einer LZA-Infrastruktur ist geprägt durch das Volumen der zu speichernden Daten, ihrer Sicherheit und ihrer Reproduzierbarkeit sowie der Verfügbarkeit von Diensten. Die Reproduktion z.B. von genomischen Daten kann bei längeren Aufbewahrungsdauern sehr komplex werden. Die Aufbewahrungsdauer richtet sich wiederum nach der Absicht, die mit der Langzeitarchivierung verfolgt wird (z.B. Backup).

Die DFG Empfehlung zur Sicherung guter wissenschaftlicher Praxis gibt eine Aufbewahrungsdauer von 10 Jahren an [1]. Im Rahmen des Projektes „Kooperative Langzeitarchivierung für Wissenschaftsstandorte“ (KoLaWiss) wurden verschiedene Service Level für die Aufbewahrungsdauer darlegt. Hier wird eine Aufbewahrungsdauer von mehr als 30 Jahren als Langzeitarchivierung definiert [2].

Der Betrieb und die Nutzung der LZA-Infrastruktur über einen Zeitraum von 10 bis 30 Jahren setzten eine entsprechend langfristige Finanzierung voraus. Ein zugrunde liegendes Finanzierungskonzept muss diesen Zeithorizont mit zukünftigen Kosten und Preisentwicklungen berücksichtigen. Ferner muss berücksichtigt werden, dass die Aufbewahrungsdauern weit über die Projekt- und damit über die Förderlaufzeiten hinausgehen.

3.2 Kostenarten

Für den Betrieb der LZA-Infrastruktur fallen Investitionskosten für Server, für Netzwerkinfrastruktur und für bauliche Infrastruktur an. Ebenso müssen in der Kalkulation Personal- und Qualifizierungskosten, Betriebskosten für Strom, Wartungskosten sowie Kosten für Ersatzinvestitionen berücksichtigt werden.

Die in der Laborimplementierung verwendeten Softwareprodukte für das MDR, FPD-R und Nutzerinterface sind gegenwärtig frei verfügbar und verursachen keine zusätzlichen Kosten. Bei einer langfristigen und realistischen Betrachtung über 30 Jahre muss aber berücksichtigt werden, dass hier Änderungen eintreten können und die Produkte unter Umständen nicht mehr frei verfügbar sein werden.

In der folgenden Tabelle sind die zu berücksichtigenden Kostenarten mit derzeitigen Kostenschätzungen aufgeführt. Die Tabelle stellt die Werte exemplarisch zusammen. Je nach konkreter Fragestellung mögen reale Kostenanalysen durchaus davon abweichen.

Tabelle 5: Übersicht über die zu berücksichtigenden Kostenarten mit Kostenschätzung.

Investitionskosten		**Kostenschätzung**	**Anmerkung**
Hardware (IT)	Server	2.500 € / Server / Jahr	Kosten für Servermiete laut Angebot GWDG
	Speichermedium (extern)	490 € / TB / Jahr	Kostenabschätzung der GWDG; der Bedarf des

		Backup (externes Speichermedium)	160 € / TB / Jahr	Sequenzierzentrums des IKMB der Universität Kiel wird mit ca. 1 TB alle 20 Tage angegeben

		Client-PCs (Standard Desktop)	900 € / PC	durchschnittlicher Wert laut Rahmenvertrag mit Hardwareanbieter
	Software / Lizenzen (Neubeschaffungen / Upgrades)	Nutzerinterface	kostenlos	Open Source
		MDR	kostenlos	Open Source
		FPD-R	kostenlos	Open Source
Sachkosten				
	Software / Wartung	Software- und Lizenzkosten / Updates	8-16% des Anschaffungspreises / Jahr	Wartung der Hardware ist meist in den Anschaffungskosten enthalten
	Customizing	Beratungsleistungen zur Implementierung oder späterer Anpassungen	800 € - 1.200 € / Tag	Dienstleistungstag eines Projektleiters, Informatikers oder Applikationsspezialisten
	Schulungen	Schulungen für Endanwender und IT-Personal	ca. 800 € / Person / Tag	Kosten variieren z.B. je nachdem, ob es sich um interne oder externe Schulungen handelt
	Entsorgung	Anfallende Kosten für die Entsorgung von IT-Hardware oder ggf. auch Software / Lizenzen		Kosten für die Entsorgung der Client-PCs am Ende der Laufzeit; Abschreibungsdauer 3 Jahre
	Dienstleistung	PID-Dienst	150 € / 500 PIDs / Jahr	bei mehr als 500 PIDs fallen hier variable Kosten pro zusätzlichem PID an; Nutzung für akademische Einrichtungen und bei forschungsbezogener Verwendung ist allerdings meist kostenlos;
	Betriebskosten	Strom	ca. 120 Watt / Stunde x 0,26 € / kWh = 0,031 € / Stunde	Stromkosten müssen in Abhängigkeit der Anzahl der Client-PCs und ihrer Betriebszeit berücksichtigt werden
Personal				
	System-Administrator	Technischer Mitarbeiter	ca. 50.000 € / Jahr	Personalkosten sollten verursachungsgerecht zugerechnet werden; Kapazitäten müssen jederzeit vorgehalten werden
	Support	Technischer Mitarbeiter	ca. 50.000 € / Jahr	

3.3 Anwendungsszenarien und Kostenblöcke

In Deliverable 5.4 ist die Installation einer LZA-Infrastruktur beschrieben. Auch einzelne Forschungseinrichtungen sind damit prinzipiell in der Lage, eine eigene LZA-Infrastruktur aufzubauen und zu betreiben.

Einige der in Tabelle 5 genannten Kostenblöcke zählen zu den Fixkosten, da sie unabhängig von der Nutzungsintensität sind. Hierunter fallen die Kosten für Server, Personal, Räume und Customizing-Dienstleistungen. Je mehr Daten gespeichert werden, desto geringer werden die Kosten pro gespeicherter Einheit.

Andere Kostenblöcke sind abhängig von der Nutzungsintensität. Es handelt sich um variable Kosten. Hierzu zählen zum Beispiel Kosten für Lizenzen und Betriebskosten.

Kosten für den Speicher sowie für den PID-Dienst gehören zu der Kategorie der sprungfixen Kosten. Bis zu einer bestimmten Speichermenge bzw. bis zu einer bestimmten Anzahl an genutzten PIDs sind die Kosten fix. Wird das Volumen überschritten, fallen weitere sprungfixe Kosten bzw. variable Stückkosten an.

Hardware / Server

Die Kosten für Server sind nutzungsintensitäts- und kapazitätsunabhängig. Um ein hohes Maß an Ausfallsicherheit und Skalierbarkeit zu gewährleisten, wird sowohl für die Speicherung der FPD als auch der Metadaten die Verwendung von mindestens drei Speichermedien im RAID-Modus empfohlen. Weiterhin sollten auch das MDR und das FPD-R auf unterschiedlichen Servern betrieben werden.

Sofern der Anwender geschützte Bereiche innerhalb des eigenen Standortes verwenden möchte, muss dieser mindestens einen weiteren Server inkl. Speichermedium für die LZA-Infrastruktur bereitstellen. Zudem sind nicht nur geschützte Bereiche innerhalb eines Standortes sondern auch geschützte Bereiche zwischen Standorten möglich.

Das Backup kann auf einem externen Speichermedium mit ausreichend Speichervolumen durchgeführt werden.

Der Speicherbedarf hängt vom Datenaufkommen ab und ist somit variabel. Speicher wird häufig in größeren Einheiten zu einem Gesamtpreis angeboten – unabhängig davon, wie viele Speichereinheiten tatsächlich genutzt werden. Je mehr Daten gespeichert werden, desto niedriger sind die Kosten pro gespeichertem Datum. Dies gilt jedoch nur so lange, bis der Speicher belegt ist und ein neues Speichermedium angeschafft werden muss. Diese so genannten sprungfixen Kosten sind bei der Kalkulation zu berücksichtigen.

PID-Dienst

Für die Referenzierung der Metadaten in der LZA-Infrastruktur ist das Betreiben eines PID-Dienstes vorgesehen. Dabei kann hier auf den Dienst Dritter zurückgegriffen werden (Kapitel 1.1.4). Die Nutzung des PID-Dienstes der GWDG ist auf Forschung beschränkt und wird Forschungseinrichtungen kostenlos zur Verfügung gestellt. Auch Anbieter ähnlicher Systeme, z.B. Digital Object Identifier (DOI), bieten akademischen Einrichtungen ihren Service in der Regel kostenlos an. Nicht-akademische Einrichtungen müssen für diesen Service zahlen. Der DOI-

Service der Deutschen Zentralbibliothek für Medizin verlangt für bis zu 500 DOIs im Jahr 150 €. Jeder weitere DOI wird pro Stück abgerechnet.

Personal

Der Infrastrukturanbieter benötigt mindestens einen System-Administrator für die Installation, den Betrieb und den Support der LZA-Infrastruktur. Sofern Anbieter und Nutzer nicht zur selben Einrichtung gehören, muss die Betreuung der Infrastruktur auf der Anwenderseite ebenfalls durch einen System-Administrator erfolgen. Wie viel Kapazitäten für die Installation, den Betrieb, den jeweiligen Support und die Betreuung der Nutzer benötigt werden, hängt ebenfalls von der Nutzungsintensität der LZA-Infrastruktur ab. Davon unabhängig müssen jedoch gewisse Personalkapazitäten ständig vorgehalten und finanziert werden. (Siehe hierzu auch Kapitel 1.2.1 und 1.2.4.)

Die obigen Ausführungen zeigen, dass der Betrieb der LZA-Infrastruktur einen – im Vergleich zu den variablen Kosten – hohen Anteil an fixen bzw. sprungfixen Kosten aufweist. Ein hoher Anteil an fixen Kosten bedeutet, dass bei einer höheren Nutzung Kostenvorteile, sogenannte Skaleneffekte, entstehen können (Fixkosten pro gespeichertem Datum sinken).

Obwohl die LZA-Infrastruktur auch von einzelnen Forschungseinrichtungen betrieben werden kann, kann sich dies aufgrund zu geringer Datenmengen als unwirtschaftlich erweisen, da Skaleneffekte nicht genutzt werden können. Ein externer Dienstleister, der die LZA-Infrastruktur betreibt und mehreren Nutzern gegen ein Entgelt zur Verfügung stellt, hat die Möglichkeit Skaleneffekte optimal zu nutzen.

Die im Rahmen dieses Projektes ebenfalls erstellte Dienstgütevereinbarung regelt einen solchen Betrieb einer LZA-Infrastruktur. Sie unterscheidet zwischen „Auftragnehmer“ (Anbieter der LZA-Infrastruktur) und „Auftraggeber“ (Nutzer der LZA-Infrastruktur). Der Auftragnehmer stellt die LZA-Infrastruktur bereit, die der Auftraggeber als Dienstleistung einkauft.

3.4 Finanzierungsmodell

Hohe Fixkosten, Skaleneffekte sowie ein kontinuierlicher personeller und technologischer Betreuungsbedarf sprechen dafür, dass die LZA-Infrastruktur von einem Dienstleister betrieben wird, der die Leistung langfristig gegen ein zu entrichtendes Entgelt zur Verfügung stellt.

Die Frage der Finanzierung nach Projektende ist damit allerdings nicht geklärt. Der Wissenschaftsrat betont jedoch in seinen Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020, dass eine (Langzeit-)Archivierung zu den ständigen Aufgaben der Wissenschaft gehört und sie somit durch eine Grundfinanzierung sichergestellt werden muss [3].

Die DFG hat diese Empfehlung in ihrem Leitfaden für die Antragstellung bereits berücksichtigt und ermöglicht die Beantragung der Übernahme von fachspezifischen Kosten zur Sicherung, Aufbewahrung und nachhaltigen Verfügbarkeit von Forschungsdaten. [4]

Dies stellt sowohl die Forschungseinrichtungen als auch die Fördermittelgeber bzw. die Projektträger vor Herausforderungen, denen in diesem Modell[7] begegnet werden soll.

Falls Forschungseinrichtungen für abgeschlossene Projekte weiterhin Zahlungen des Projektträgers (bzw. des Fördermittelgebers) über einen Zeitraum von bis zu 30 Jahren für die Langzeitarchivierung ihrer Forschungsdaten erhalten sollten, würde dies einen erheblichen organisatorischen Mehraufwand für beide Seiten bedeuten. Ferner ist die Entwicklung der Kosten über einen Zeitraum von 10 bis zu 30 Jahren bei der Antragstellung meist schwer abzuschätzen.

Daher wird empfohlen, dass Forschungseinrichtungen zukünftig den Speicherbedarf für ihr Projekt ermitteln und in ihrem Projektantrag die entsprechende Speicherkapazität beantragen.

Der Projektträger schließt im Gegenzug einen Vertrag mit dem LZA-Infrastrukturbetreiber über die Sicherung, Aufbewahrung und nachhaltige Verfügbarkeit der Forschungsdaten. Zahlungen für die Dienstleistung erfolgen direkt vom Projektträger an den LZA-Infrastrukturbetreiber.

Der Projektträger setzt ein Expertengremium ein, das diese Angaben den Antrag auf seine Schlüssigkeit prüft. Dieses Gremium entscheidet im Anschluss darüber, ob die Einrichtung den beantragten Speicherplatz erhält und über welchen Zeitraum die Forschungsdaten aufzubewahren sind. Vor Ablauf der zunächst vereinbarten Speicherdauer entscheidet das Gremium darüber, ob die Aufbewahrungsdauer der Forschungsdaten verlängert wird.

In Projekten zum Thema „Virtuelle Forschungsumgebungen“ wurden bereits ähnliche Fragestellungen zur Finanzierung diskutiert und es wurden entsprechende Finanzierungsmodelle entwickelt und vorgestellt.[5] Das hier vorgestellte Modell reduziert den organisatorischen Aufwand und stellt die Verfügbarkeit von Forschungsdaten sowie die Finanzierung der LZA-Infrastruktur durch den Projektträger über die Projektlaufzeit hinaus sicher. Vorteil dieses Modells ist außerdem, dass der Projektträger die Kontrolle über die Art und Menge der gespeicherten Daten sowie die Kosten der Sicherung, Aufbewahrung und nachhaltigen Verfügbarkeit der Forschungsdaten erhält.

[7] Die Idee des im Folgenden vorgestellten Finanzierungsmodells basiert auf den Ausführungen der Kapitel 3.1 bis 3.3 sowie auf Erkenntnissen der im Rahmen des Projektes durchgeführten Experteninterviews und Workshops.

4 Ausblick

Die entwickelte Archivinfrastruktur bietet zurzeit alle Basisfunktionen, die für den Betrieb eines Archivs notwendig sind. Darüber hinaus ist für den Produktivbetrieb die Konsolidierung der Benutzerverwaltungen der verschiedenen Subsysteme und das Erstellen einer geeigneten grafischen Administrationsschnittstelle zu empfehlen.

Durch den modularen Aufbau der Archivinfrastruktur sind der Austausch von Komponenten und die Anpassung an die eigenen Anforderungen und Bedürfnisse möglich. Auch der entwickelte GeMeCo ist modular entwickelt worden und bietet damit vor allem viele Anpassungs- bzw. Weiterentwicklungsmöglichkeiten. Neue Datentypen, Metadatenschema und neue FPD-R (inkl. Speicherstruktur) lassen sich über diesen modularen Aufbau mit wenigen Eingriffen umsetzen.

Die Verfügbarkeit der LZA-Infrastruktur kann durch eine langfristige Finanzierung über den Projektträger langfristig sichergestellt werden. Ein mögliches Lösungsszenario ist in Kapitel 3.4 gezeigt. Die enge Zusammenarbeit zwischen Experten und IT ist dabei essentiell. Weitere Lösungen können so unter Berücksichtigung der Forschungsinteressen und der technischen Machbarkeit erarbeitet werden.

5 Definitionen und Abkürzungen

API	Programmierschnittstelle
BSI	Bundesamt für Sicherheit in der Informationstechnik
CA	Certificate Authority
DIR	Directory Service
DOI	Digital Object Identifier
EPIC	European Persistent Identifier Consortium
FPD	Forschungsprimärdaten
FPD-R	Forschungsprimärdaten-Repository
GeMeCo	Generic Metadata Collector
GUI	Grafische Benutzerschnittstelle
GWDG	Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
IP	Internetprotokoll
JRE	Java Runtime Environment
LABIMI/F	Langzeitarchivierung biomedizinischer Forschungsdaten
MDR	Metadaten-Repository
MDS	Metadatenschema
MRC	Metadata and Replica Catalog
NAS	Network Attached Storage
OSD	Object Storage Device
PID	Persistenter Identifikator
PKI	Public-Key-Infrastructure
RAID	Redundant Array of Independent Disks
REST	Representational State Transfer
SAN	Storage Area Network
SIP	Submission Information Package
SLA	Service Level Agreement
SSH	Secure Shell
TLS/SSL	Transport Layer Security / Secure Sockets Layer
URL	Uniform Resource Locator
UUID	Universally Unique Identifier
WAN	Wide Area Network
XML	Extensible Markup Language

5.1 DSpace

DSpace ist eine OpenSource Software Plattform zur Archivierung digitaler Objekte. Momentan sind 1430 DSpace Installationen bei www.dspace.org registriert[8]. DSpace wird von einer großen Community getragen und weiter entwickelt. Dies verhindert Herstellerabhängigkeiten und lässt auf einen langen Fortbestand der Software schließen. Die aktuelle Version 3.1 ist auf den 30. Januar 2013 datiert.

5.2 LABIMI/F

In dem DFG-geförderten Forschungsprojekt LABIMI/F werden interdisziplinär Handlungsempfehlungen für die biomedizinische Community bzgl. der Langzeitarchivierung digitaler medizinischer Bilddaten und Genomdaten formuliert.

Die Handlungsempfehlungen basieren auf zuvor durchgeführten Anforderungsanalysen und durch eine Laborimplementierung der aus den Anforderungsanalysen entwickelten Infrastruktur am Standort Göttingen.

5.3 Persistenter Identifikator (PID)

Ein persistenter Identifikator (PID) ist eine eindeutige Zeichenkette, welche ein (digitales) Objekt weltweit eindeutig identifiziert. Diese PIDs referenzieren auf den Speicherort ihres Objektes oder auf dessen weiterführende Informationen. PID-Dienste verwalten die Abbildung von PID auf dessen Objekt / Informationen und sind daher auch für das Auflösen der PIDs verantwortlich. Durch das Aktualisieren der am PID-Dienst hinterlegten Objekt-URL kann der Speicherort eines Objektes verändert werden, ohne dass sich dessen PID ändert.

5.4 XtreemFS

XtreemFS ist ein objekt-basiertes, föderiertes Dateisystem, das für den Einsatz in verteilten und Cloud-Umgebungen konzipiert ist. Es bietet eine POSIX-kompatible Nutzerschnittstelle an. Sämtliche Kommunikation zwischen den einzelnen Speicherressourcen kann verschlüsselt werden. XtreemFS ist durch verschiedene Replikationsstrategien in der Lage, die Daten ausfallsicher bereit zu stellen und entsprechend der Nutzerzugriffe schnell verfügbar zu machen.

[8] Stand: 8. Mai 2013, http://www.dspace.org/whos-using-dspace

6 Anhang

Anhang A – Strukturvorgabe für Metadatenschemata (XSD)

```
<?xml version="1.0" encoding="UTF-8"?>
<schema xmlns="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://gemeco.labimi-f.med.uni-
goettingen.de/schema/metastructure" xmlns:tns="http://gemeco.labimi-
f.med.uni-goettingen.de/schema/metastructure"
elementFormDefault="qualified">

  <complexType name="schemaType">
    <sequence>
      <element name="name" type="string"></element>
      <element name="version" type="string"></element>
      <element name="author" type="string"></element>
      <element name="language" type="language"></element>
      <element name="description" type="string"></element>
      <element name="remark" type="string" maxOccurs="unbounded"
minOccurs="0"></element>
      <element name="source" type="tns:sourceDefType" maxOccurs="unbounded"
minOccurs="0"></element>
      <element name="collection" type="tns:collectionType"></element>
    </sequence>
  </complexType>

  <complexType name="collectionType">
    <sequence>
      <element name="section" type="tns:sectionType" maxOccurs="unbounded"
minOccurs="0"></element>
    </sequence>
  </complexType>

  <complexType name="sectionType">
    <choice maxOccurs="1" minOccurs="1">
      <element name="element" type="tns:elementType" maxOccurs="unbounded"
minOccurs="1"></element>
      <element name="section" type="tns:sectionType" maxOccurs="unbounded"
minOccurs="1"></element>
    </choice>
      <attribute name="id" type="string" use="required"></attribute>
      <attribute name="name" type="string"></attribute>
      <attribute name="optional" type="boolean" use="optional"></attribute>
    </complexType>

    <element name="schema" type="tns:schemaType"></element>

    <complexType name="elementType">
      <all>
        <element name="name" type="string"></element>
        <element name="description" type="string" maxOccurs="1"
minOccurs="0"></element>
        <element name="type" type="string"></element>
        <element name="source" type="tns:sourceRefType" maxOccurs="1"
minOccurs="0"></element>
        <element name="constraints" type="tns:constraintsType"
maxOccurs="1" minOccurs="0"></element>
      </all>
      <attribute name="id" type="string" use="required"></attribute>
      <attribute name="optional" type="boolean" use="optional"></attribute>
    </complexType>
```

```
    <complexType name="constraintsType">
      <all>
        <element name="minValue" type="string" minOccurs="0"></element>
        <element name="maxValue" type="string" minOccurs="0"></element>
        <element name="pattern" type="string" minOccurs="0"></element>
        <element name="values" type="tns:stringListType"
minOccurs="0"></element>
      </all>
    </complexType>

    <complexType name="stringListType">
      <sequence>
        <element name="item" type="string" maxOccurs="unbounded"
minOccurs="0"></element>
      </sequence>
      <attribute name="clusivity" type="tns:clusivityType"
use="optional"></attribute>
      <attribute name="external" type="string" use="optional"></attribute>
    </complexType>

    <simpleType name="clusivityType">
      <restriction base="string">
        <enumeration value="inclusive"></enumeration>
        <enumeration value="exclusive"></enumeration>
      </restriction>
    </simpleType>

    <complexType name="sourceRefType">
      <choice>
        <element name="id" type="string"></element>
          <sequence>
            <element name="ref" type="string"></element>
            <element name="name" type="string"></element>
          </sequence>
      </choice>
    </complexType>

    <complexType name="sourceDefType">
      <sequence>
        <element name="package" type="string"></element>
        <element name="version" type="string" maxOccurs="1"
minOccurs="0"></element>
        <element name="internalRef" type="string" maxOccurs="1"
minOccurs="0"></element>
        <element name="remark" type="string" maxOccurs="unbounded"
minOccurs="0"></element>
      </sequence>
      <attribute name="id" type="string"></attribute>
    </complexType>
</schema>
```

Literaturverzeichnis

[1] DFG (Hrsg.). Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission “Selbstkontrolle in der Wissenschaft” ; Proposals for safeguarding good scientific practice : recommendation of the Commission on Professional Self Regulation in Science. Weinheim: Wiley-VCH; 1998.

[2] Dickmann F. KoLaWiss Kooperative Langzeitarchivierung für Wissenschaftsstandorte, AP5 - Kosten der elektronischen Langzeitarchivierung. Göttingen: Abteilung Medizinische Informatik; 2009.

[3] Wissenschaftsrat (Hrsg.). Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020. Berlin: 2012.

[4] DFG (Hrsg.). Leitfaden für die Antragstellung; Projektanträge. Bonn: 2013.

[5] Dickmann F, Fiedler N, Kaspar M, Falkner J. Bedarf, Ausrichtung und Finanzierung Virtueller Forschungsumgebungen. PIK - Prax Informationsverarbeitung Kommun 2012;35.