

kobv



Kooperativer Bibliotheksverbund
Berlin-Brandenburg



Workshop

»Langzeitarchivierungsservices für KOBV-Bibliotheken«

13.11.2014

Wolfgang Peters-Kottig, Tim Hasler, Elias Oltmanns

Agenda

10:00 Uhr Begrüßung

10:10 – 11:15 Uhr

Teilnehmerrunde: Wie ist die Situation an Ihrer Einrichtung?

11:15 – 12:15

Status und Strategien am ZIB (*inkl. Kaffeepause*)

12:15 – 13:00

Diskussion: Fragen und Anregungen der Bibliotheken

Begleitfragen zum Status in Ihrer Einrichtung

- Gibt es ein **Mandat** / einen offiziellen Auftrag zur digitalen LZA in Ihrer Einrichtung?
- Welche **Medientypen** sind/wären bei Ihnen betroffen?
(Hochschulschriften, [Retro-]Digitalisate, lizenzierte E-Medien, Forschungsdaten)
- Welche **Nachweis-/ Präsentationssysteme** betreiben Sie?
(Dokumentenserver, Digitalisat-Repositoryn, Forschungsdatenreposit.,...)
- **Wer beschäftigt sich** in Ihrer Bibliothek mit dem Thema Langzeitarchivierung?
- Gibt es externe **Kooperationen oder (eigene) Projekte**?
- Bietet ein institutionelles **Rechenzentrum/IT-Service** für Ihre Bibliothek Speicherdienste?

Roadmap



- Kickoff-Workshop für KOBV-Bibliotheken...
- Online-Befragung zur Bedarfsanalyse: Januar 2015
- Auswertung: I. Quartal 2015
- Beratung im Kuratorium (AG Strategie) im Zuge der strategischen Neuausrichtung des KOBV: begleitend in 2015
- Zweiter Workshop: Mitte/Ende 2015?
- *Umsetzung schrittweise (Betriebs-Modell, Testpartner, kostenlose vs.kostenpflichtige Services, Zertifizierung...)*

Strategische Neuausrichtung des KOBV

Diskussion Bereich Langzeitarchivierung:

- Ausdifferenzierung nach Nutzeranforderungen
- gestaffelte Services (Basisdienst, erweiterter Dienst)
- Rechtemanagement (auch rechtl. Rahmenbedingungen)
- Schnittstellen zu externen Systemen / Access-Komponenten
- Richtlinien (Policies)
- Zertifizierung
- Geografisch getrennter zweiter Speicherort(?)
- Beratung zum Forschungsdatenmanagement im Hinblick auf Langzeitarchivierung

...Basisinfrastruktur am ZIB

- Hierarchisches Speichermanagementsystem:
Datenverteilung zw. Festplatten („online“) und Tape-Storage (Magnetbandkassetten, „offline“)
- Derzeit 90 Petabyte Kapazität (problemlos weiter ausbaubar:
18.000 Stellplätze für aktuell 5 TB-Bänder)

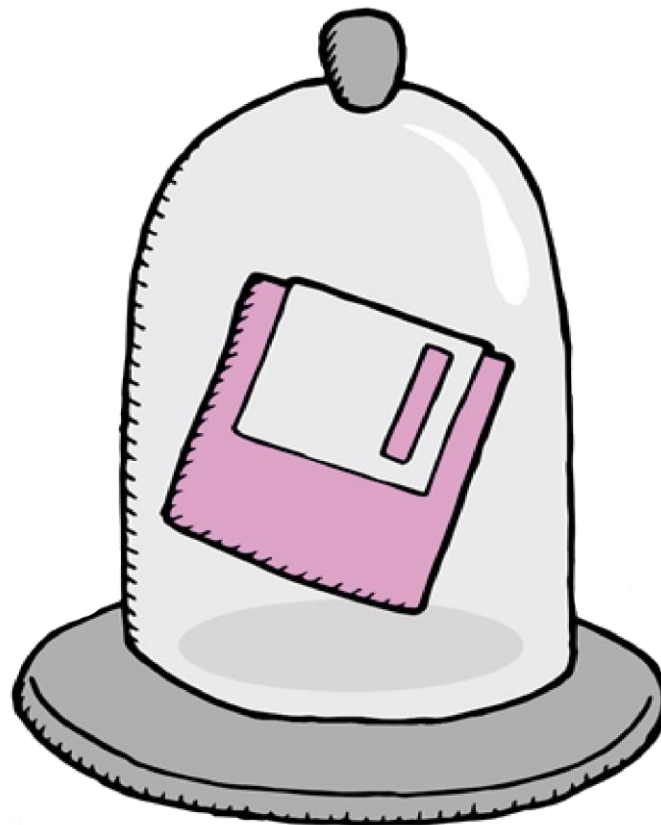
→ ZIB-Service zur sicheren Datenhaltung
(Bitstream-Preservation)

- Das können wir sofort anbieten!



»Ich mache mir derzeit große Sorgen darüber, dass wir zwar ‚Bits‘ sichern können, nicht aber ihre Bedeutung, und dass wir am Ende auf einem Haufen Datenschrott sitzen werden.«

Vint Cerf, Chief Internet Evangelist, Google



Grundsätzliches...

Langzeitarchivierung bedeutet nicht
sichere Ablage in einem Tresor oder ein
zusätzliches Backup von E-Medien sondern
**Übernahme von Verantwortung für
zukünftige Nutzung**

Können wir *später* noch etwas damit anfangen?

```
A:TEST.TXT PAGE 1 LINE 4 COL 01 INSERT ON
< < < MAIN MENU > > >
--Cursor Movement--
^S char left ^D char right
^A word left ^F word right
^E line up ^X line down
--Scrolling--
^Z line up ^W line down
^C screen up ^B screen down
-Delete-
^G char
DEL chr lf
^I word rt
^Y line
-Miscellaneous-
^I Tab ^B Reform (from Main only)
^V INSERT ON/OFF
^L Find/Replce again
RETURN End paragraph
^H Insert a RETURN
^U Stop command
-Other Menus-
^J Help ^K Block
^Q Quick ^P Print
^O Onscreen
```

Die gerade in Verwendung befindlichen
Formate ändern sich permanent.

Es gibt Archiv-geeignete Formate, für die
einigermaßen sicher scheint, dass sie auch in
50 Jahren noch interpretierbar sind.

Andere Formate müssen beobachtet werden
und dann stellt sich die Frage, wie man
feststellt, ob eine Datei, die **automatisch
konvertiert** wurde, noch das gleiche Objekt
darstellt. **Wenn das überhaupt möglich ist.**

C1		25
UNIT		COST
12.9	556.85	
6.9	101.25	
49.9	12487.50	
4.9	9.90	
TOTAL	13155.50	
	1282.66	
	14438.16	

Grundsätzliches...

Es geht nicht um einmaligen
Erwerb eines fertigen Systems
für alle Assets

Nicht Technik ist das Kern-
Problem sondern Organisation

Grundsatzentscheidung für quelloffene Software

Aufbau des Langzeitarchivs mit Open-Source-Software als Garant für Transparenz und zur Verringerung von Risiken in der Langzeitverfügbarkeit.

- Archivemata (Archivierungssystem)
- iRODS (Datenmanagement-Schicht)
- Fedora Commons (Repository-Backend)
- Islandora (Repository-Frontend)

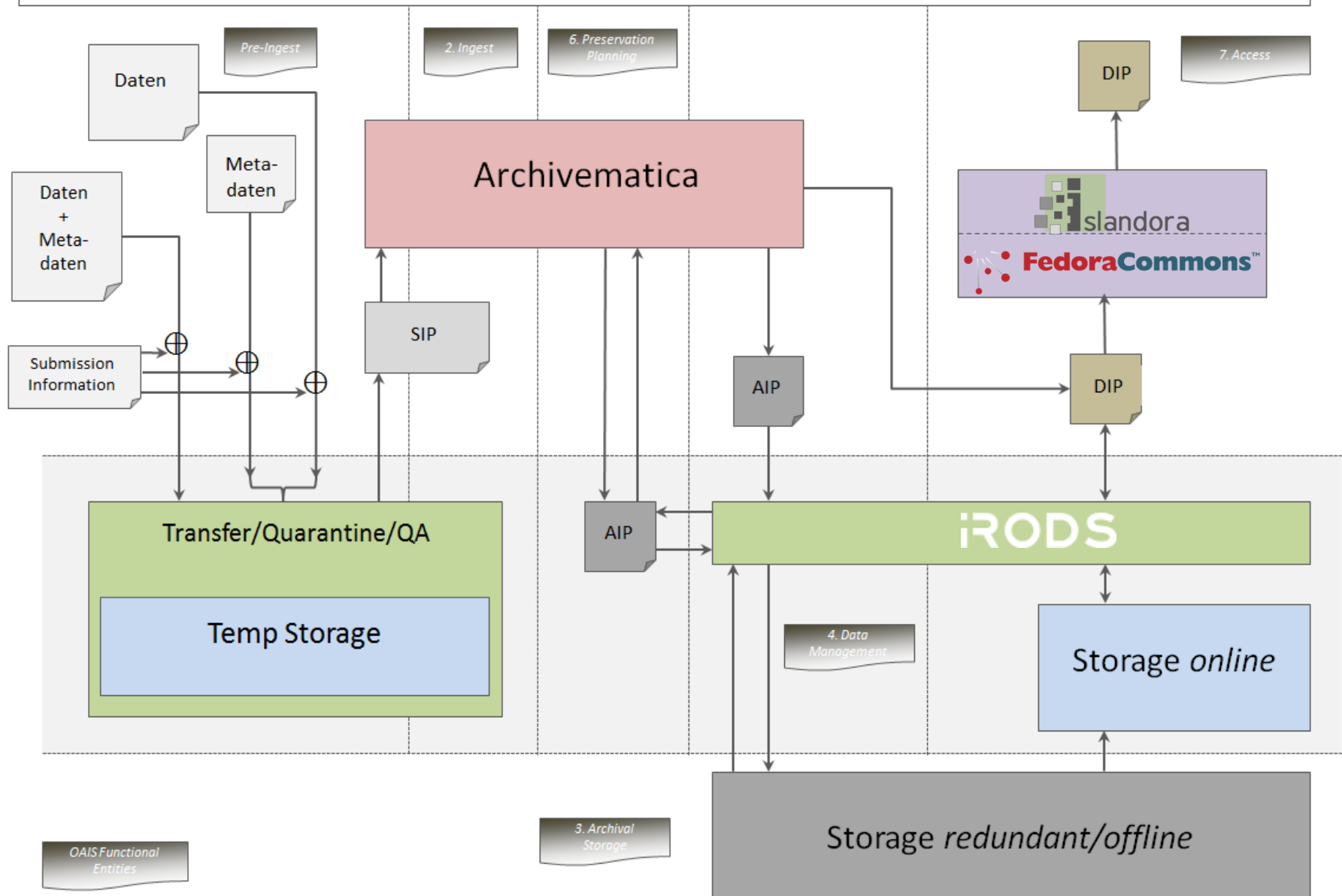


because 76% of time travellers
agree that the future sucked

Architekturkonzept digitales Langzeitarchiv

Policies/Standards/Access Control

5. Administration



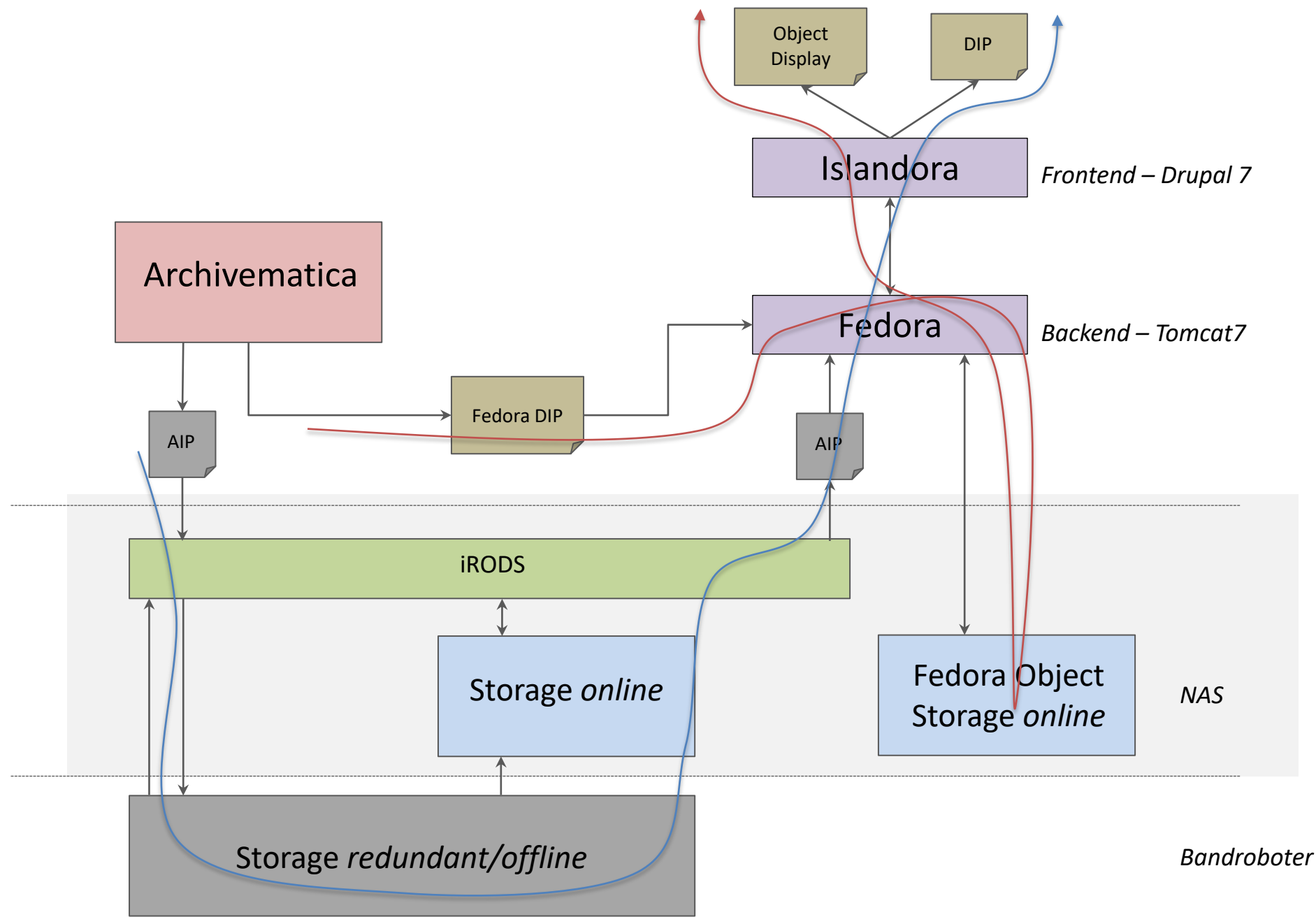
Access/Dissemination: Planung bei digiS

Fedora



- Digitales Repositorium
 - Ingest, Manage, Delivery
 - Beliebige Dateiformate
 - Datenhaltung: Dateisystem (XML + content)
- Modulare Architektur: *Digital Object Model*
 - Beziehungen untereinander möglich
- Ressourcen: internal XML, managed, externally- / redirect-referenced
- Zugang: API; REST-API [beta]; OAI_DC
- Rechteverwaltung mittels XACML

- (Rechtebasierter) Access für das LZA
 - Anzeige der Objekte (Derivate) in Collections
 - Durchsuchbarkeit
- Ausliefern einzelner Informationspakete vom Band (Master) mittels iRODS-Referenz
- Islandora:
 - Drupal Modul
 - (austauschbares) Frontend
 - konfigurierbare Anzeige (Solution Packs)
 - Modul-Konfiguration (z.B. Solr)
- Datendrehscheibe



Home

search

◦ [Advanced Search](#)

Navigation

- ▶ [Add Drupal content](#)
- ▼ [Islandora Repository](#)

User menu

- [My account](#)
- [Log out](#)

Development

- [Clear cache](#)
- [Devel settings](#)
- [Entity info](#)

Top-level Collection

View

Manage



[Grid view](#) [List view](#)



[Basic Image
Collection](#)



[Bröhan Museum
Berlin](#)



[Compound Collection](#)



[Large Image
Collection](#)



[ZIB Compound
Collection](#)



[digiS](#)



[digiS Image
Collection](#)

EWIG

Entwicklung von Workflowkomponenten für die Langzeitarchivierung von Forschungsdaten im Bereich Erd- und Umweltwissenschaften

- Gefördert durch die DFG
- 0,5 Stelle für 3 Jahre
- Drei Projektpartner



Drei Komponenten

Entwicklung von institutionellen **Policies**

Exemplarische **Workflows** zur Übernahme von Daten
in ein Langzeitarchiv analog zur DIN 31645

Entwicklung von **Lehrkomponenten**

Entwurf einer institutionellen Policy

Policy über die Organisation der Datenprojekte

(Stand Juli/2013, Version 3)

§1

Dieses Dokument regelt die Verantwortlichkeiten bei der Verwaltung von Datenprojekten (DP) und enthält Empfehlungen zur Organisation der Daten innerhalb eines DP. Änderungen an diesem Dokument müssen dem IT-Aktiv vorgelegt und entschieden werden.

§2

Ein DP versteht sich als eine zusammenhängende Partition innerhalb des zentralen Datenbereiches mit einer eigenen UNIX-Gruppe und einem ausgezeichneten Nutzer, dem Datenprojektadministratoraccount (DPAA). Das DP hat einen eindeutigen und aussagekräftigen Datenprojektnamen (DPN). Der DPN wird ebenso als UNIX-Gruppenname, DPAA und als Ordnername im Verzeichnissystem benutzt. Das DP wird auf Systemebene der IT-Infrastruktur unter dem Pfad „/daten/[DPN]/“ sichtbar gemacht.

§3

Ein neues DP kann eingerichtet werden, sobald eine inhaltliche und technische Abgrenzung von Daten notwendig wird. Wenn der dafür notwendige Bereich mehr als 5% des noch verfügbaren Plattenplatzes oder mehr als 1% des Gesamtbereiches im zentralen Speicherpool beträgt, muss die Einrichtung durch das IT-Aktiv oder den geschäftsführenden Direktor bestätigt werden. Gleiches trifft bei einer Vergrößerung eines bestehenden DP zu. Das IT-Aktiv kann die Einrichtung eines DP empfehlen.

§4

Das Datenprojekt enthält die zwei Unterordner „work“ und „arch“. Der Ordner „work“ dient als zeitweilig zu nutzender Arbeitsbereich für die Vorbereitung und Bearbeitung von Daten, die nach Abschluss der Arbeiten durch den Projektadministrator in den Ordner „arch“ einsortiert werden. Eine Beschreibung des DP erfolgt in der Textdatei „/daten/[DPN]/README“.

§5

Ein DP wird von mindestens einem Wissenschaftler genutzt. Ein Wissenschaftler aus der Nutzergruppe muss als Datenprojektverantwortlicher (DPV) bestimmt werden. Der DPV kann einen oder mehrere Stellvertreter bestimmen. Dies ist in der „README“ des DP zu dokumentieren.

§6

Aufgaben des DPV:

- sorgt für eine optimale Nutzung des DP durch regelmäßige Überprüfung der DP-Aktivitäten
- dokumentiert das DP in einer README-Datei nach den vorgegebenen Empfehlungen (siehe dazu das Template im Intranet <http://...>)
- organisiert die Daten im „arch“ Bereich und dokumentiert diese DP-spezifischen Regeln
- sorgt für eine konsistente Überleitung von Daten aus dem „work“ in den „arch“ Bereich
- nutzt den DPAA für administrative Arbeiten im DP
- legt die Verfallszeit von Daten im „work“-Bereich fest
- zeigt Änderungen zum DP (Größenanpassungen, Verantwortlichkeiten) in schriftlicher Form der IT-Administration an

§7

Jede Person mit einem Institutsaccount kann die Schreibrechte für ein DP bekommen, wenn dies im Rahmen seiner Tätigkeit notwendig wird. Der DPV des DP muss für einen Institutsaccount die Aufnahme in oder das Entfernen aus der DP Nutzergruppe bestätigen.

§8

Die Nutzung eines DP setzt die Anerkennung dieser Policy voraus. Bei Missachtung der Policy kann die Nutzung des DP durch das IT-Aktiv unterbunden werden. Darüber hinaus gelten die DP-spezifischen Organisationsregeln, deren Einhaltung in der Verantwortung des DPV liegt.

Empfehlungen für den DPV und die Nutzer von DPs

- Der DPV hat die Verantwortung für einen sorgfältigen Umgang mit dem DPAA. Es sollten nicht mehr als zwei Personen diesen Account benutzen. Es wird empfohlen, das Passwort des DPAA regelmäßig zu wechseln.
- Zur Überwachung der Platzaufteilung innerhalb eines DP werden einmal täglich die DP ausgewertet und eine Belegungsliste für jeden Nutzer des DP erzeugt. Diese Liste befindet sich im Datenprojekt unter folgendem Namen „/daten/[DPN]/[USERNAME]“ und enthält in der ersten Spalte die Dateigröße, in der zweiten den kompletten Namen der Datei. Diese Dateien sind Grundlage der Statistik unter <http://...>
- Der DPV sollte vor Benutzung eines DP die Organisationsregeln des „arch“-Bereiches in der „README“ dokumentieren. Beispiele für eine Gestaltung der Ordnerstruktur befinden sich auf den Webseiten des Intranet (<http://...>).
- Der DPV kann jederzeit Daten aus dem Datenprojekt löschen. Dies sollte jedoch immer nach Absprache mit den DP-Nutzern erfolgen.

Übernahmevereinbarung -> DIN 31645

- *SubmissionManifestVersion*: 1.0
- *SubmittingOrganization*: Name der einliefernden **Institution**
- *OrganizationIdentifier*: ISIL-Nummer der einliefernden Institution (oder ähnliches, sofern vorhanden)
- *ContractNumber*: Vom ZIB vergebene **Vertragsnummer**
- *Contact*: Name des administrativen **Ansprechpartners**, der bei Fragen oder Zugriffsanfragen kontaktiert wird
- *ContactEmail*: Adresse von *Contact*, an die Übernahmeprotokolle geschickt werden sollen
- *SubmissionIdentifier*: Eindeutiger, von der einliefernden Institution vergebener **Identifier**, der die Lieferung kennzeichnet.
- *SubmissionDescription*: Selbstgewählte inhaltliche **Beschreibung** der Lieferung
- *Rights*: urheberrechtliche, datenschutzrechtliche, persönlichkeitsrechtliche Angaben und Embargozeiten; z.B. „public domain“, „freier Zugriff – Rechte vorbehalten“ oder konkrete **Lizenzangaben**.
- *DataSourceSystem*: Software, aus der die Daten **exportiert** wurden (mit Versionsangabe)
- *MetadataFile*: Name und Pfad zur **Metadaten-datei** (nach Klärung der Verzeichnisstruktur des Transferpakets)
- *MetadataFileFormat*: Dateiformat der Metadaten-dateien. Beispiele: csv, dc, MARC, dcterms, datacite, lido, ead, mods
- *DataCurator*: Name des Ansprechpartners, der bei Fragen oder Zugriffsanfragen zu dieser Lieferung kontaktiert/genannt wird (wenn abweichend von *Contact*)
- *DataCuratorEmail*: Adresse *DataCurator*
- *DataEncoding*: Beschreibung der verwendeten **Dateiformate** (optional), z.B. TIFF

Rückkopplung in die Lehre

Workshop „Wege in die Köpfe“



Handreichung zum Forschungsdatenmanagement

Handreichung:

- Wie organisiere ich Daten?
- Welche Werkzeuge helfen mir?
- Warum muss ich mich um die Kosten kümmern?
- Weiterführende Literatur



Neues Projekt: HDC – Humanities Data Centre

HUMANITIES DATA CENTRE



- Projekt-Partner: GWDG (Konsortialleitung), SUB, BBAW, Akademie der Wiss. Göttingen, Max-Planck Inst. f. multiethn. u. multireligiöse Gesellschaften, assoziiert: Max-Planck-Inst. f. Bildungsgeschichte
- Förderung: Niedersächs. Ministerium f. Wiss. und Kultur; vorerst nur initiale Design-Phase bis 2016: 600.000 € Personalmittel insgesamt (keine Sachmittel)
- ZIB ist „Infrastrukturpartner“ (0,5 Stelle E13)
- Projekt-Ziele:
 - Sicherstellung der langfristigen Speicherung und Nachnutzbarkeit von geisteswissenschaftlichen Forschungsdaten
 - Aufbau einer nachhaltigen geisteswissenschaftlichen Forschungsdateninfrastruktur
 - Klärung der technischen, organisatorischen und rechtlichen Voraussetzungen

Neue Kooperation ZIB-BBAW im Bereich Digital Humanities

- Kooperationsvertrag
- Zusammenarbeit z.B. im HDC-Projekt
- Forschungsthema: Wie können Projektergebnisse der BBAW langfristig zugänglich bleiben?
- ZIB ist außerdem am interdisziplinären Forschungsverbund Digital Humanities in Berlin ifDHB beteiligt: → Vernetzung in der Region

if|DH|b

nestor-Partner:

- * Bayerische Staatsbibliothek
- * Bibliotheksservice-Zentrum Baden-Württemberg
- * Bundesarchiv
- * Computerspiele Museum Berlin
- * Deutsche Kinemathek – Museum für Film und Fernsehen
- * Deutsche Nationalbibliothek
- * FernUniversität Hagen
- * GESIS - Leibniz-Institut für Sozialwissenschaften
- * Goportis - Leibniz-Bibliotheksverbund Forschungsinformation
- * Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen
- * Georg-August-Universität Göttingen / Niedersächsische Staats- und Universitätsbibliothek Götting
- * Humboldt-Universität zu Berlin
- * Institut für Deutsche Sprache
- * Institut für Museumsforschung (Stiftung Preußischer Kulturbesitz)
- * Konrad-Zuse-Zentrum für Informationstechnik Berlin
- * Landesarchiv Baden-Württemberg
- * Landesarchiv Nordrhein-Westfalen
- * PDF/A Competence Center
- * Rechenzentrum der Universität Freiburg
- * Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden

nestor-Arbeitsgruppen:

- AG Digitale Bestandserhaltung
- AG Emulation
- AG Kooperation und Vernetzung
- AG Kosten
- ✗ AG Media
- AG Policy
- AG Recht
- AG Zertifizierung
- ✗ AG Formaterkennung
- ✗ AG Forschungsdaten

Zusammengefasst:

Was bedeutet das alles für den KOBV?

- KOBV-Zentrale profitiert von ZIB-Infrastruktur, Forschung und Kooperationen (v.a. digiS)
- Starkes Commitment der ZIB-Leitung zum Thema Langzeitarchivierung und Datenmanagement („Research Fokus Area Digital Humanities“)
- LZA-Architektur modular, flexibel, anpassbar an Anforderungen und Bedarfe (kein „Monolith“)
- Integration/Anpassung an weitere Services der KOBV-Zentrale (Hosting, Repositories/Dokumentenserver...)

...Diskussionsrunde

- Wie viele Medien sind eigentlich betroffen? (*Datenmenge...*)
- Was sind denkbare „Trigger-Events“, d.h. in welcher Situation würden Sie Daten aus dem Langzeitarchiv holen wollen?
- Wieviele der langzeitarchivierten Medien müssen immer kurzfristig zugänglich sein?
- Welche Zugriffsbeschränkungen gibt es ggf.?
- Ist Langzeitarchivierung zukünftig als Standard-Bibliotheksservice zu sehen (=mit vorhandenem Personal umzusetzen) oder weiterhin als zusätzliche Aufgabe, für die Projektgelder o.ä. notwendig bleiben?
- Ist LZA eher eine Notwendigkeit oder eine Chance zur „Profilierung“ der Bibliothek, oder ...?