



3. DH-Rundgang

Digitale Langzeitarchivierung am ZIB

18.11.2014

Konrad-Zuse-Zentrum für Informationstechnik Berlin

Wolfgang Peters-Kottig, Tim Hasler, Elias Oltmanns

Konrad-Zuse-Zentrum für Informationstechnik



Ein *außeruniversitäres* Forschungsinstitut des Landes Berlin für anwendungsorientierte Mathematik und Informatik.

ZIB

SEARCH 

AA  Imprint

RESEARCH | INSTITUTE | RESEARCH SERVICES | MEMBERS | TRAINING & OUTREACH | CONTACT

The Zuse Institute Berlin (ZIB)

is an interdisciplinary research institute for applied mathematics and data-intensive high-performance computing. Its research focuses on modeling, simulation and optimization with scientific cooperation partners from academia and industry.

Basisinfrastruktur: Supercomputer

- Das ZIB betreibt einen der zwei Knoten des Norddeutschen Hochleistungs-Rechnerverbundes (HLRN) für 7 von 16 Bundesländern.



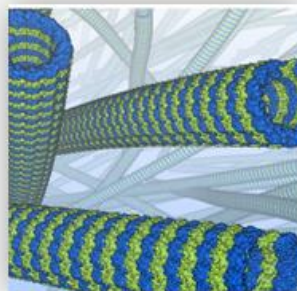
Exemplarische Projekte:

- Konstruktive Verbesserung von Verkehrsflugzeugen und Flugzeugtriebwerken
- Dreidimensionale Modelle von extrasolaren Gasplaneten
- Modellierung paläoklimatischer Bedingungen

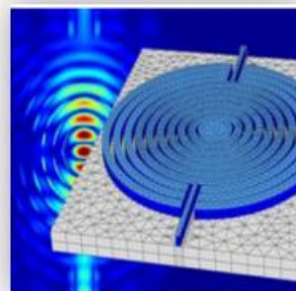
Research Focus Areas



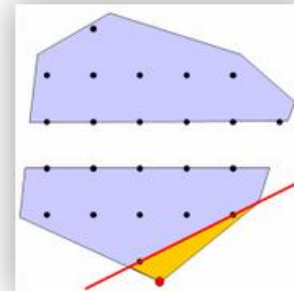
Solutions for Individualized Medicine



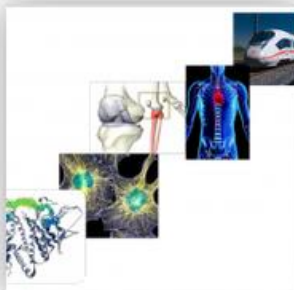
Molecular and Biological Processes



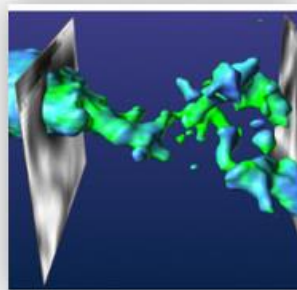
Materials and Optical Processes



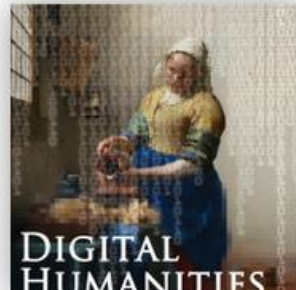
Combinatorics and MINLP



MSO for Multiscale Systems



Robust Solutions under Uncertainty



Digital Humanities



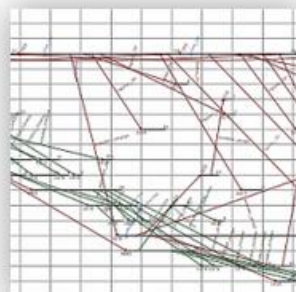
Efficient Infrastructure and Logistics



Large-Scale Data Management, Curation & Analysis



Highly Parallel Computing

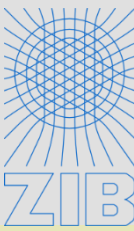


Scalable and Fault-Tolerant Algorithms



Publications

Digital Humanities im ZIB: Arbeitsgruppen



INFORMATION
TECHNOLOGY
TOOLS FOR
MUSEUMS

C. Saro



Beratung und Services für Museen

DIGITAL
PRESERVATION

W. Peters-Kottig



**Beratung und Services zum Thema
Langzeitarchivierung**

SERVICE CENTER
DIGITIZATION
BERLIN

A. Müller



**Beratung und Services spartenübergreifend
Bibliotheken, Archive, Gedenkstätten, Museen**

KOBV LIBRARY
NETWORK -
RESEARCH &
DEVELOPMENT

B. Rusch



Beratung und Services für Bibliotheken

...Basisinfrastruktur: Datenhaltung

- Sicherung auf Magnetband
- Extrem robustes Datenhaltungskonzept
- 90 Petabyte Kapazität derzeit



→ Bitstream-Preservation

»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

Langzeitarchivierung bedeutet nicht
Backup oder Datensicherung sondern
**Übernahme der Verantwortung für die
zukünftige Nutzung von Inhalten**

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
L-----!

-Delete-
^G char
DEL chr lf
^I word rt
^Y line

-Miscellaneous-
^I Tab ^B Reform
^V INSERT ON/OFF
^L Find/Replce again
RETURN End paragraph
^H Insert a RETURN
^U Stop command

-Other Menus-
(from Main only)
^J Help ^K Block
^Q Quick ^P Print
^O Onscreen

```

Hallo Welt!
Dies ist ein

Die gerade in Verwendung befindlichen
Formate ändern sich permanent.

Alle 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.

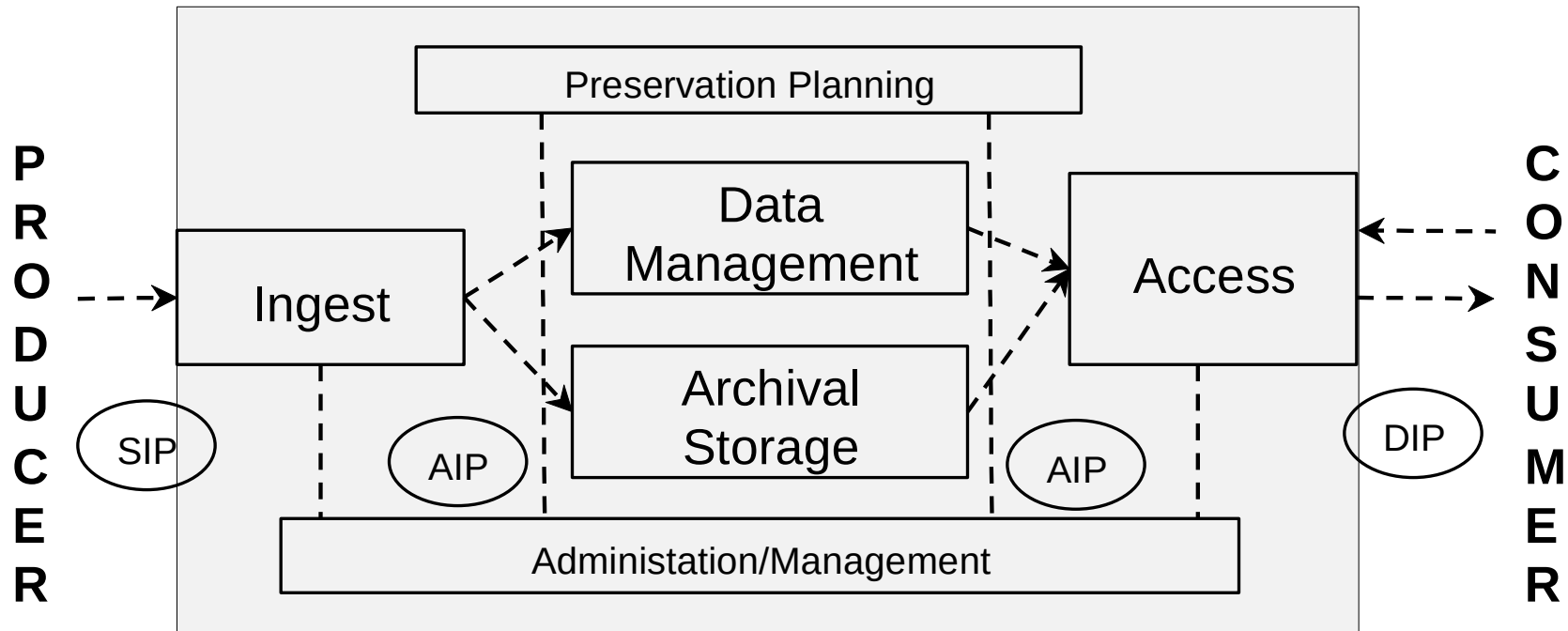
...wenn das überhaupt möglich ist.

...wie funktioniert das für komplexe Formate
oder vollständige Applikationen?

UNIT		COST
12.9	556.85	
6.7	101.25	
49.9	12487.50	
4.9	9.90	
TOTAL	13155.50	
	1282.66	
	14438.16	

18
19
20

Der Rahmen: OAIS-Referenzmodell



Sinn des Ganzen ist vor allem die Verständigung über gemeinsame Terminologie-

Grundsatzentscheidung für quelloffene Software



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

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

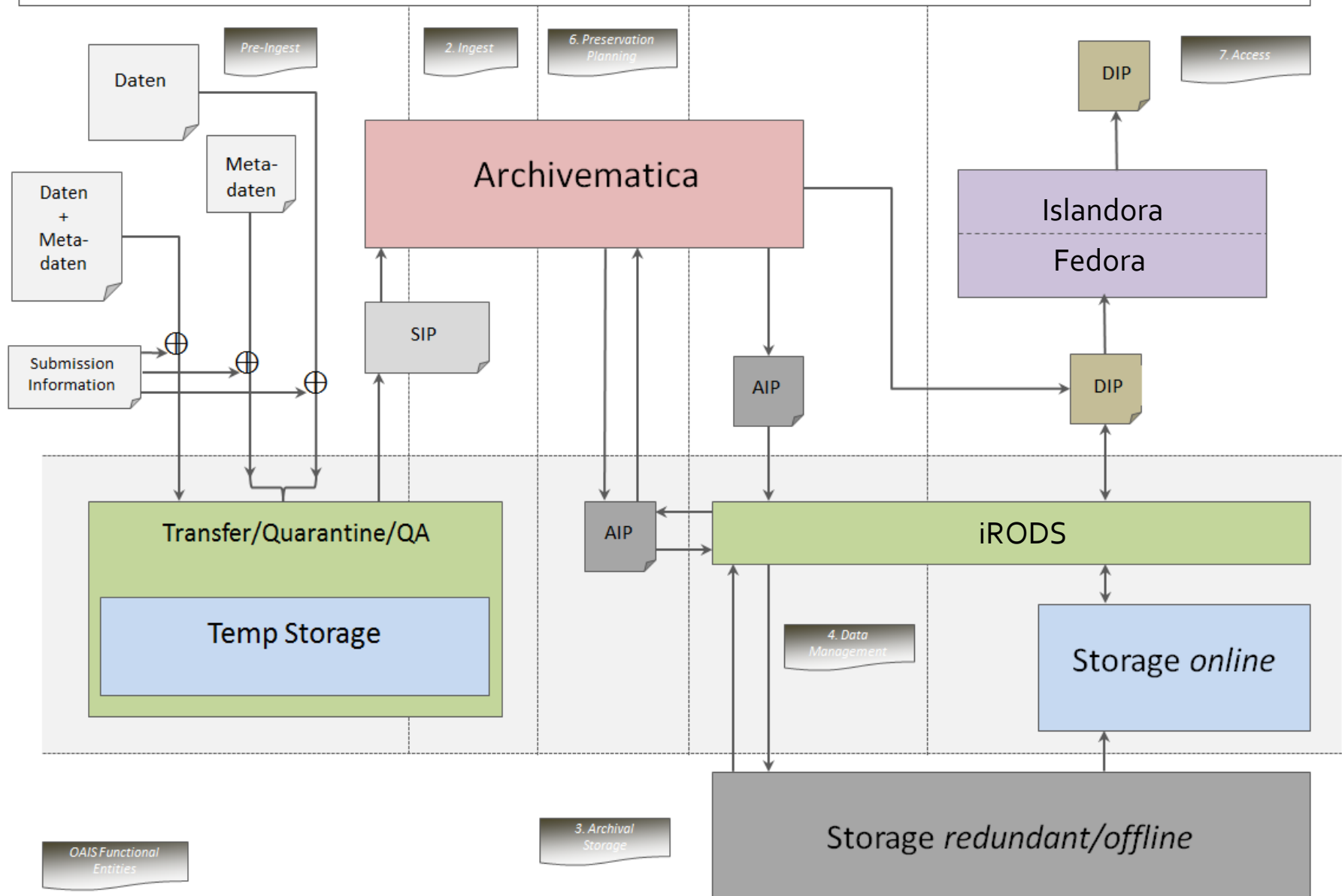
 archivematica[®]

because 76% of time travellers
agree that the future sucked

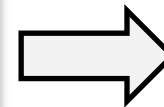
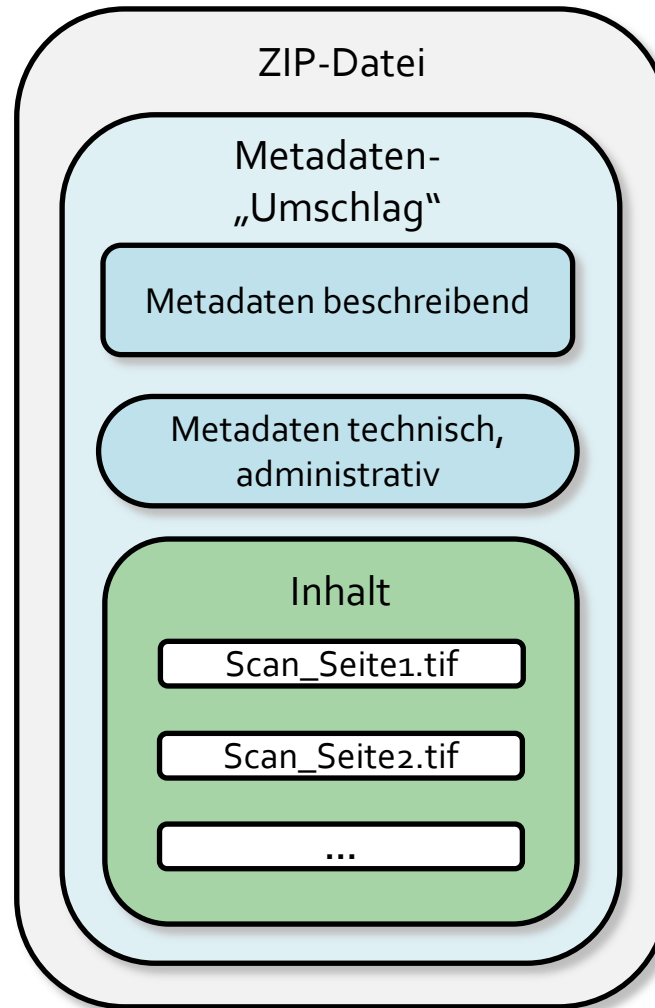
Architekturkonzept digitales Langzeitarchiv

Policies/Standards/Access Control

5. Administration



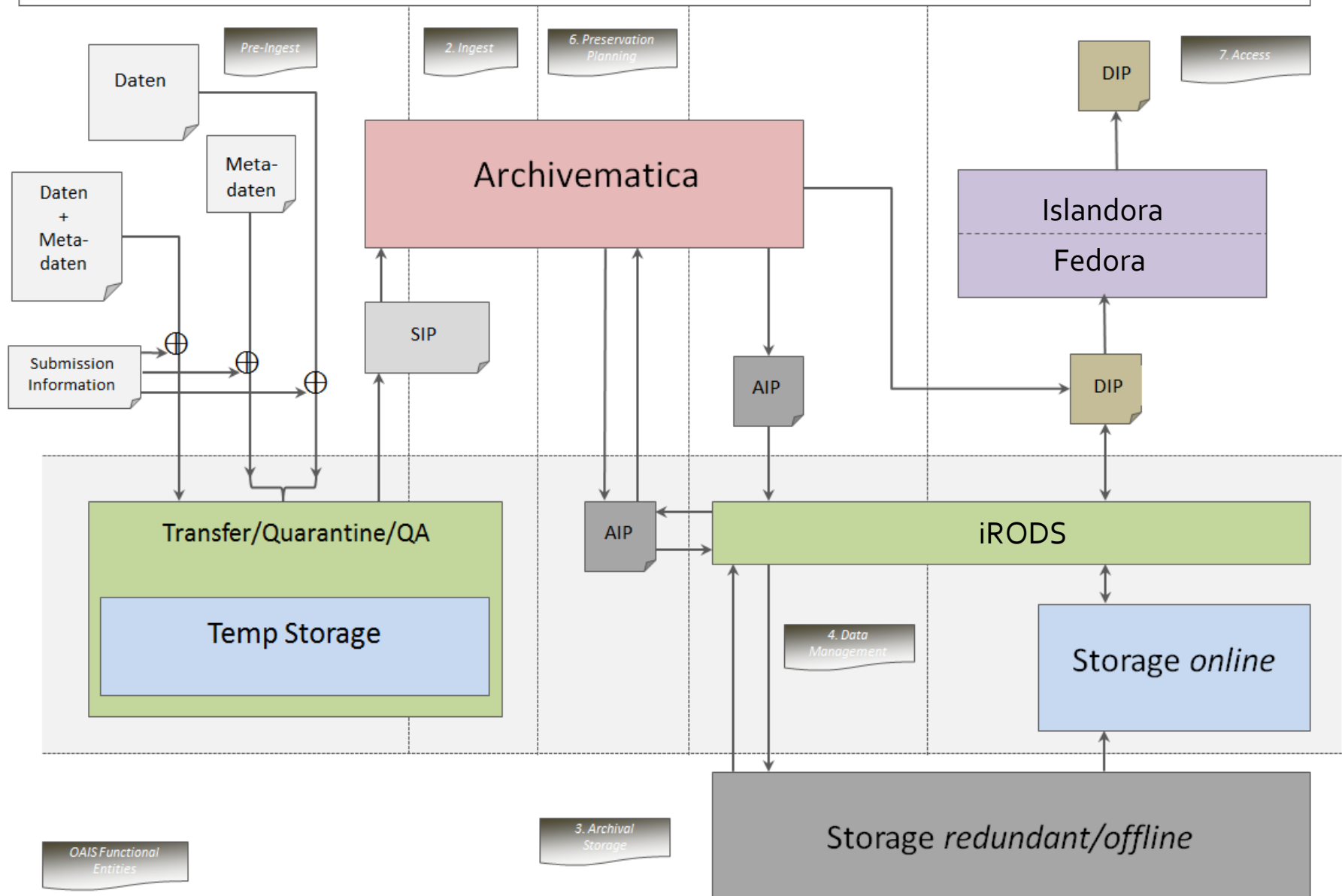
Speicherung als Archivpaket



Architekturkonzept digitales Langzeitarchiv

Policies/Standards/Access Control

5. Administration

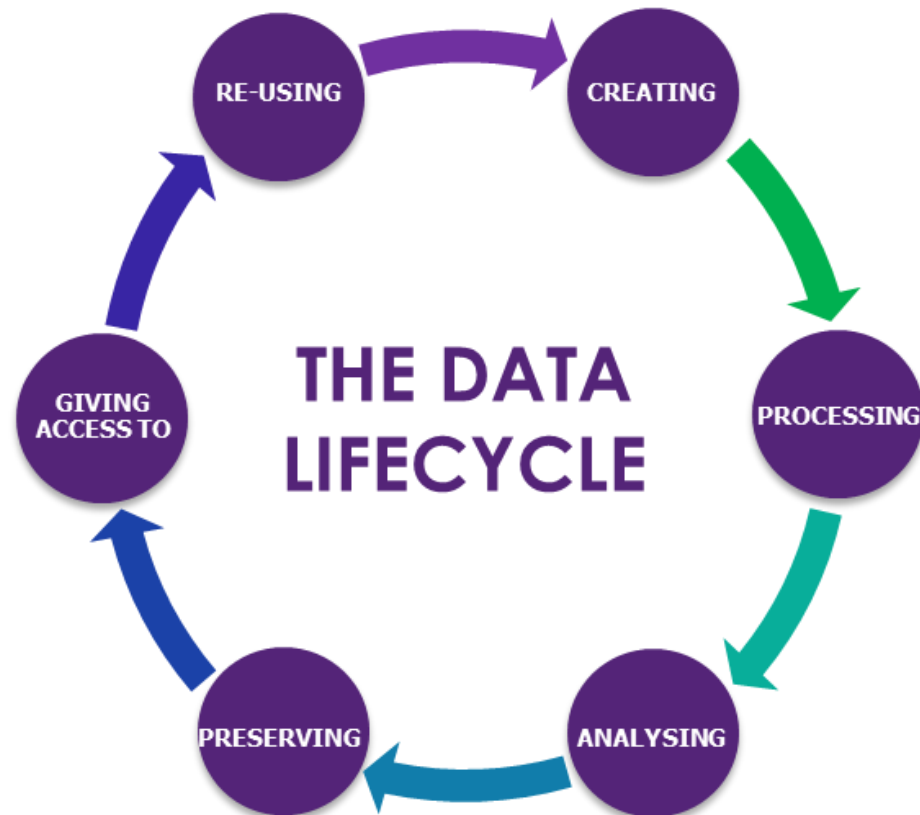


OAIS Functional Entities

3. Archival Storage

Storage redundant/offline

...ein Schritt zurück



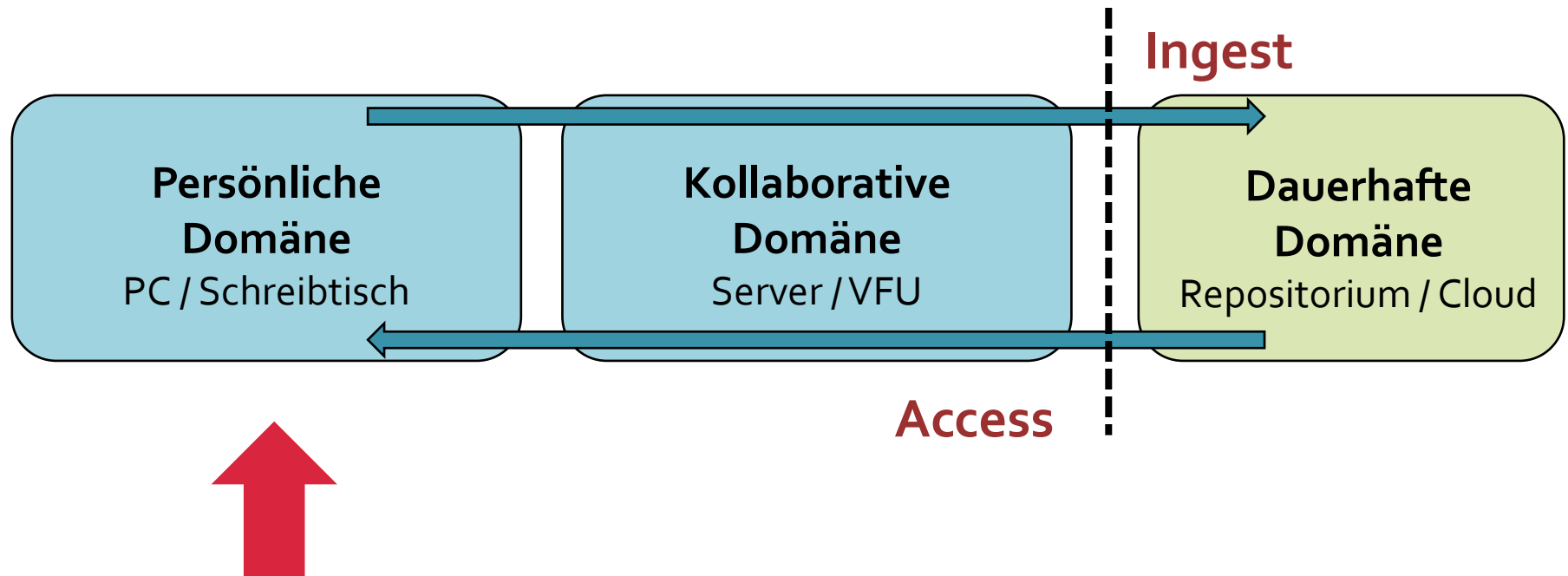
EWIG

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

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

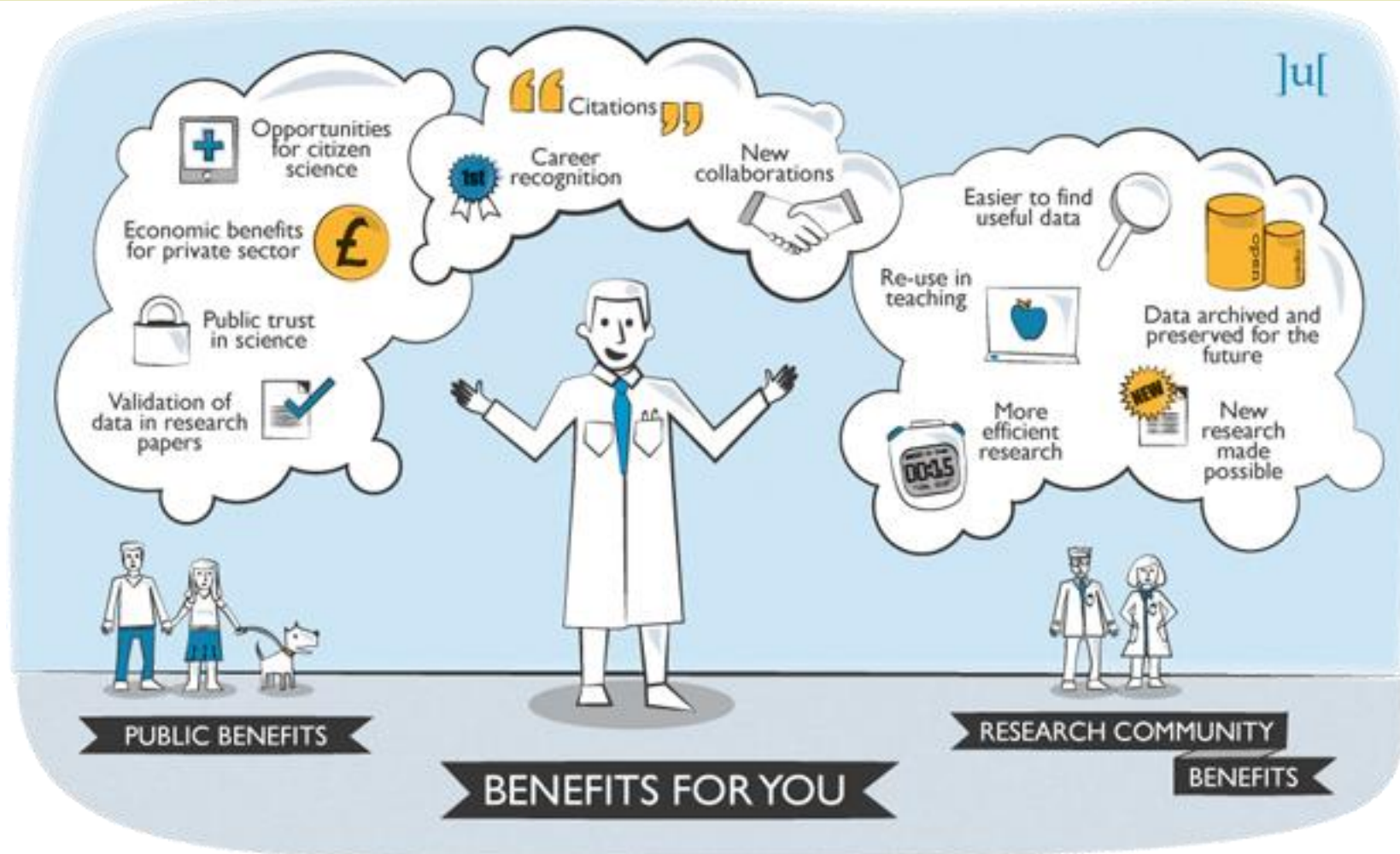


Langzeitarchivierung – so früh wie möglich



Langzeitarchivierung beginnt beim
Forschungsdatenmanagement

Bekannte Vorteile



Noch besser bekannte Vorbehalte

- Meine Daten gehören mir.
- Meine Daten sind so speziell, die kann eh' niemand anderes verwenden.
- Der Beweis ist zu unbeholfen um ihn jemand anderem zu zeigen.
- Was ist, wenn Fehler drin sind?
- ... oje oje oje oje ...

... oje oje oje oje

A STORY TOLD IN FILE NAMES:

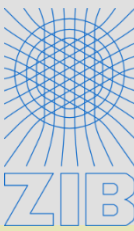
Location:  C:\user\research\data

Filename	Date Modified	Size	Type
 data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
 data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
 data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
 data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
 data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
 data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
 data_2010.05.29_aaarrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
 data_2010.05.29_#&@*!&!.dat	2:40 AM 5/29/2010	0 KB	DAT file
 data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
 data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
 data_2010.05.29_woohoo!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
 data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
 analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
 ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
 Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
 JUNK...	2:45 PM 5/29/2010		Folder
 data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

- Entwicklung von institutionellen **Policies**
- Exemplarische **Workflows** zur Übernahme von Daten in ein Langzeitarchiv analog zur DIN 31645
- Entwicklung von **Lehrkomponenten**

Policy - Verfahrensbasiert



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 Datenprojektname (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

Entwicklung eines Übergabeworkflows:

- Verantwortlichkeiten bei der Übernahme
- Signifikante Eigenschaften
- Formate
- Art und Format der Metadaten
- Rechte
- Definition der Transferpakete
- Struktur der Pakete
- Abschluss einer Übernahmevereinbarung

Handreichung zum Forschungsdatenmanagement

Handreichung:

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



- Humanities Data Centre
- „Kooperationsprojekt“ mit Berlin-Brandenburgischer Akademie der Wissenschaften (BBAW)

Designstudie „Humanities Data Centre“ (HDC)



HUMANITIES DATA CENTRE



Projektziele:

- Leistungsangebot eines HDC definieren,
- geeignete Arbeitsabläufe und Prozesse entwickeln,
- technischen und organisatorischen Grundriss liefern.

Perspektive:

- Laufzeit der Designstudie zunächst bis 2016.
- Im Anschluss dreijährige Aufbauphase eines Datenzentrums geplant.
- Überführung in den Regelbetrieb und Zertifizierung als vertrauenswürdiges Langzeitarchiv (z.B. nach ISO 16363) wird angestrebt.

Beteiligte Institutionen

■ Partner:

- Akademie der Wissenschaften zu Göttingen (ADWG)
- Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)
- Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG), Konsortialleitung
- Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften Göttingen (MPI MMG)
- Georg-August-Universität Göttingen, Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)
- Assoziierter Partner: Max-Planck-Institut für Wissenschaftsgeschichte (MPIWG)

■ Gefördert:

vom

Niedersächsischen Ministerium für Wissenschaft und Kultur

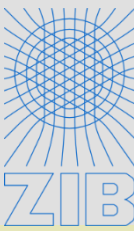
mit

600.000 € (nur Personalmittel)

für

zwei Jahre

Kooperationsprojekt BBAW ZIB



■ Herausforderung

Zur Langzeitverfügbarkeit gehören:

- Erhalt der Daten und
- Erhalt der „significant properties“ einer Präsentationsschicht zur geeigneten Darstellung der Daten.

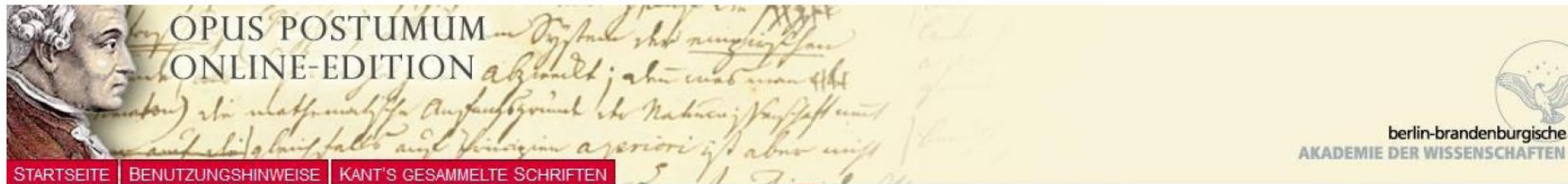
■ Projektvorhaben

An 2 Beispielprojekten der BBAW aus den digitalen Geisteswissenschaften:

- „significant properties“ der Präsentationsschicht identifizieren und möglichst formalisiert beschreiben,
- Testmigration in definierte Umgebung mit Ziel der Langzeitverfügbarkeit durchführen,
- Robustheit des Ergebnisses durch (Teil-)Reimplementierung oder den Austausch von Komponenten testen.



Beispiel: Kants „Opus postumum“ im Web-Angebot der BBAW



DIGITALE EDITION

Convolut I
Convolut II
Convolut III
Convolut IV
Convolut V
Convolut VI
Convolut VII
Convolut VIII
Convolut IX
Convolut X
Convolut XI
Convolut XII
Convolut XIII

SUCHE

Metadaten:
Convolut II, Seite 5
Alte Zählung:
Conv. II, Halbbogen I, 1
Akademieausgabe:
XXI, 161,02-162,32
Auf dem Manuskript:
Neue Zählung: 5
Alte Zählung: 2. Conv. I, 1.
Signatur Kant: No. 1.
Andere Signatur: 2
Bildnachweis:
Ms. germ. fol. 1702,
Conv. II, S. 5, SBB-PK

Convolut II P_5

TRANSKRIPTION

2 Von den metaphysischen Anfangsgründen der Naturwissen-

3 schaft (*philosophia naturalis*) *¹ geht die Tendenz i<der Philosophie> auf die *Physik*

4 Naturforschung (*Physica*) als einem System der empirischen

5 Naturkunde, als auf welche jene abzweckt; den was man die

6 (mit Newton) die mathematische Anfangsgründe der Naturwissenschaft nent

7 ist zwar auch die *² gleichfalls auf Principien *a priori* ist aber nicht

8 ein Erkenntnis aus Begriffen sondern aus der Construction derselben

9 (den jenen correspondirenden reinen Anschauungen) und macht keinen

10 Theil der Naturphilosophie aus sondern ist *³ hiebey

11 $\emptyset$ ä $\emptyset$ verwischet $\emptyset$ nur ein

12 $\emptyset$ ie $\emptyset$ wie wohl vortrefliches ja unentbehrliches Instrument zu dieser

13 ihrem Zwecke zu gelangen; $\emptyset$ welche $\emptyset$ welcher die empirische Naturforschung

14 (*Physica*) ist.

15 Es ist aber unmöglich aus bloß empirischen Erkenntni

Beariffen

TRANSKRIPTION FAKSIMILE ?

- Von der ZIB-Leitung gibt es ein starkes Commitment für weiteres Engagement im Bereich Digital Humanities
- Es werden derzeit Dienstleistungen für Bibliotheken, Museen und Archive weiterentwickelt
- Anforderungen aus den Fachdisziplinen werden in der ZIB-BBAW-Kooperation und im Rahmen des HDC-Projekts ausgelotet
- Umsetzung / Angebote für die DH in Berlin? Schrittweise...