HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

ein integrativer Ansatz für die Langzeitarchivierung einer Hochschule

Elisabeth Dregger-Cappel, Peter Schreiber, Hans-Dieter Weckmann

Digitale Langzeitarchivierung an Hochschulen
Humboldt-Universität Berlin 29./30.04.2013

Übersicht

- Struktur-Daten
- IKM-Strategie
- Vorprojekt Langzeitarchivierung
  - Bibliothek
  - Forschung
  - Klinik
- Strukturmodell
- OAIS
  - Producer
  - Consumer
- Ingest Beispiele
  - Bibliothek
  - Forschung (Genom)
- Fazit / weiteres Vorgehen
- Fundament – gemeinsame technische Infrastruktur

Heinrich-Heine-Universität Düsseldorf (HHU)

Struktur-Kennzahlen

- 5 Fakultäten
  - Mathematisch-Naturwissenschaftliche Fakultät
  - Philosophische Fakultät
  - Juristische Fakultät
  - Wirtschaftswissenschaftliche Fakultät
  - Medizinische Fakultät
- Haushaltsvolumen (2011) **304,705 Mio. €**
  - Landeszuschuss Universität 126,105 Mio. €
  - Landeszuschuss Fachbereich Medizin 115,488 Mio. €
  - Drittmittelaufwand 63,112 Mio. €

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Heinrich-Heine-Universität Düsseldorf (HHU)

Struktur-Kennzahlen

▪ Studierende (WS 2012/2013)	23.400
▪ Studienanfänger/-innen (Studienjahr 2012)	7.300
▪ Absolventen/-innen (Studienjahr 2012)	2.260
▪ Professuren (W3/W2/W1)	348
▪ Beschäftigte	4.200

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Universitätsklinikum Düsseldorf (UKD)

Struktur-Kennzahlen (2011)

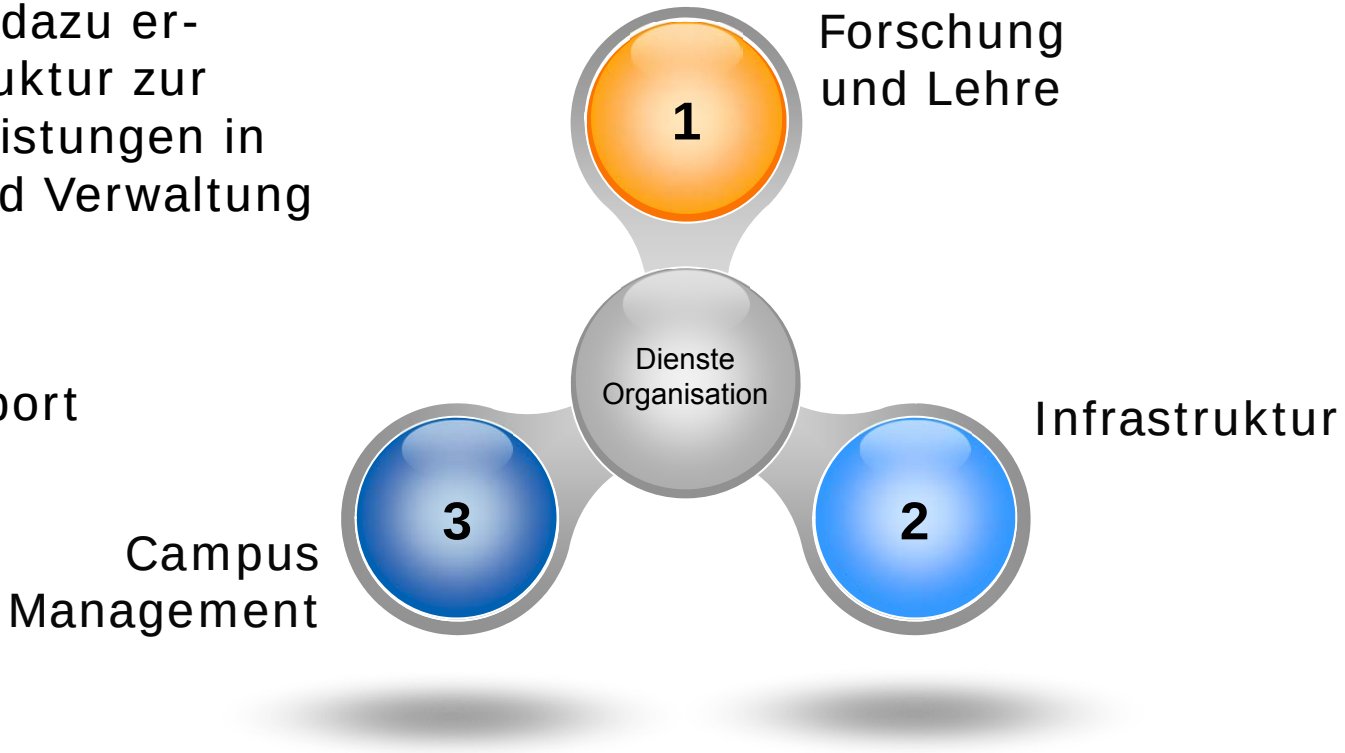
▪ Betten	1.180
▪ Mitarbeiter/-innen (Ärzte)	4.730 (817)
▪ Haushaltsvolumen	467 Mio. €
▪ Kliniken / Institute	30 / 30

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Die IKM – Strategie der HHU

Effiziente Nutzung von Informations- und Kommunikationstechnik sowie von Medien und der dazu erforderlichen Infrastruktur zur Verbesserung der Leistungen in Forschung, Lehre und Verwaltung

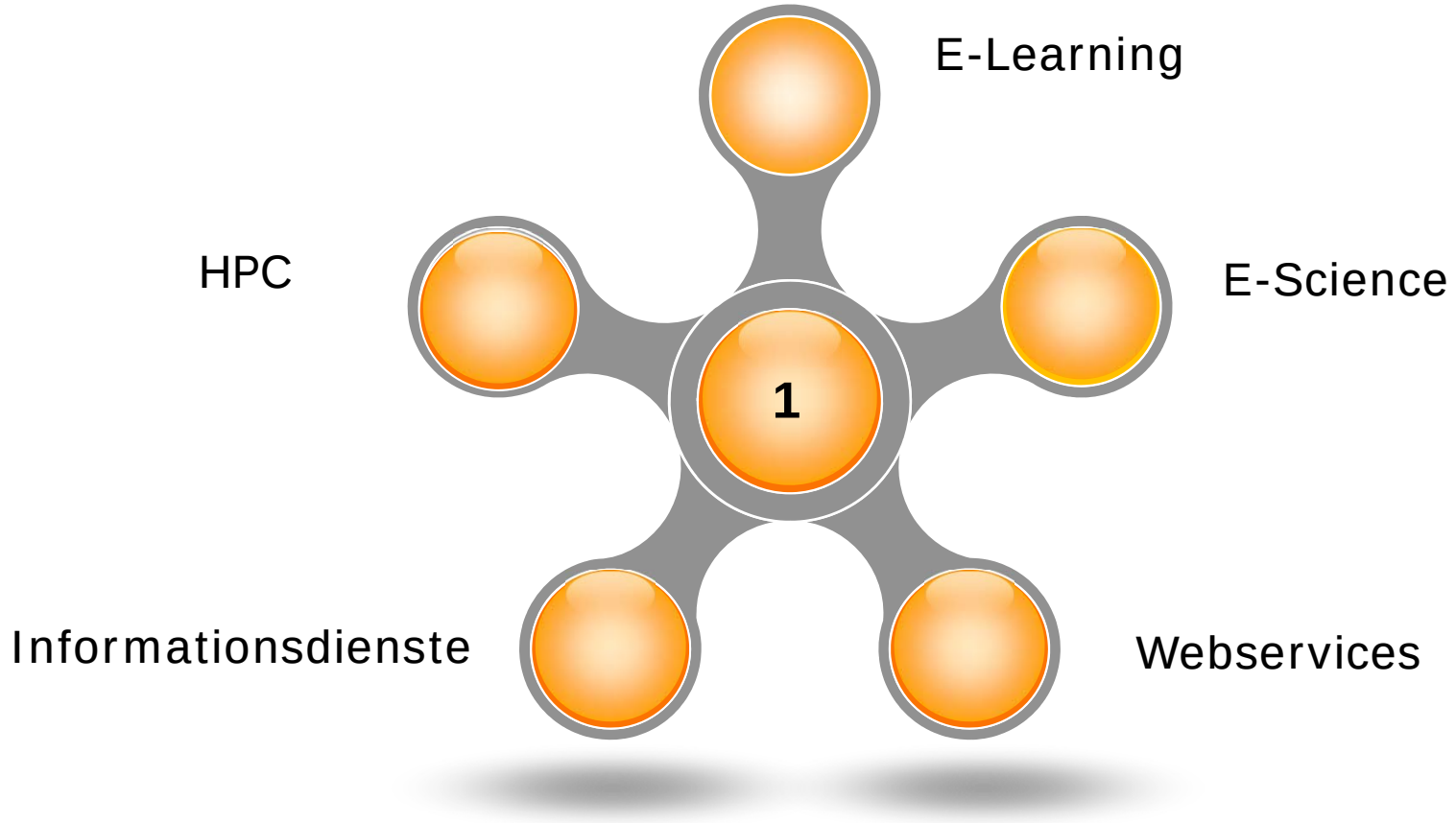
- Prozesse
- Organisation
- Technik und Support



HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

Die IKM – Strategie der HHU

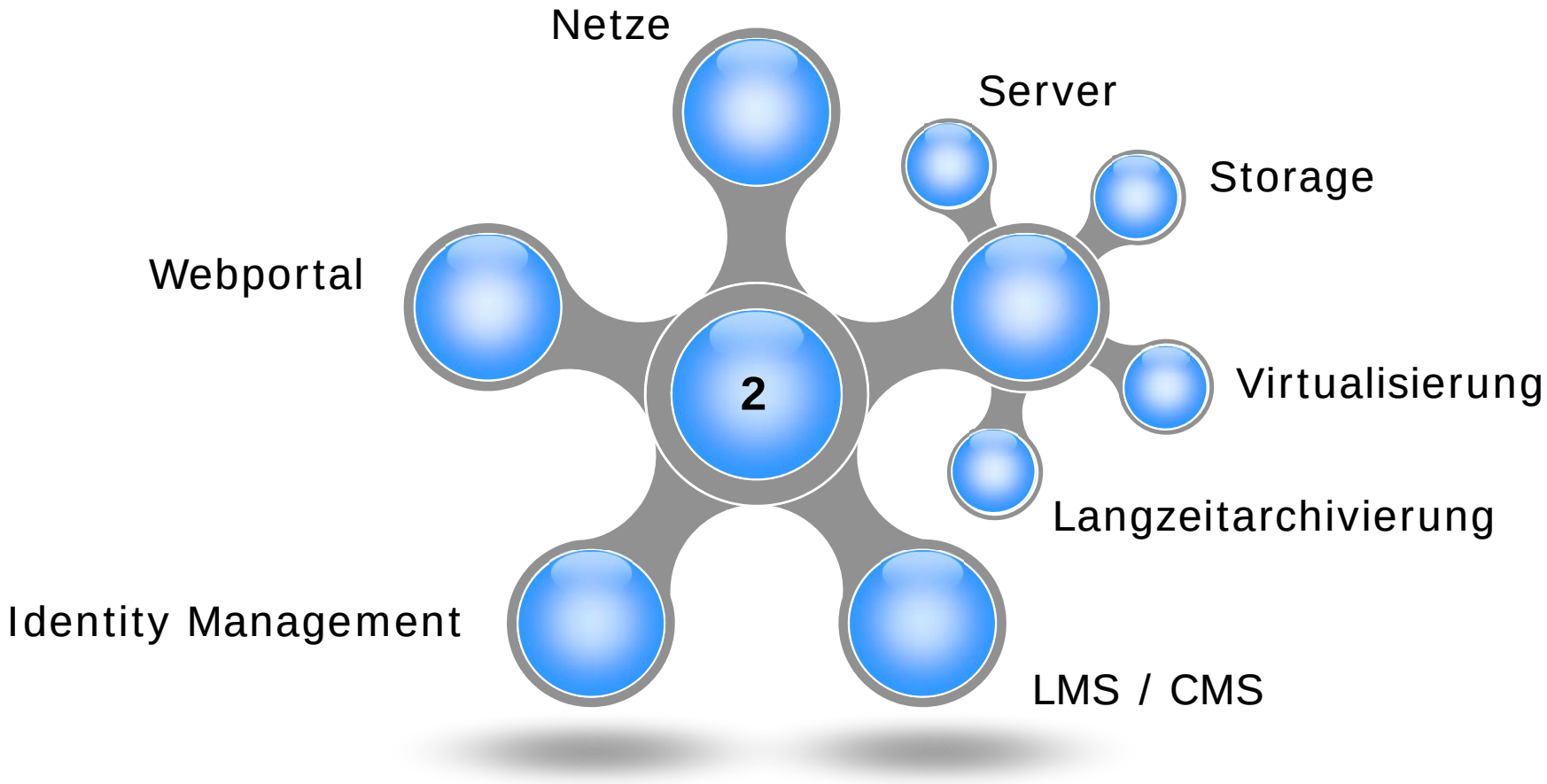
1 Forschung und Lehre



HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Die IKM – Strategie der HHU

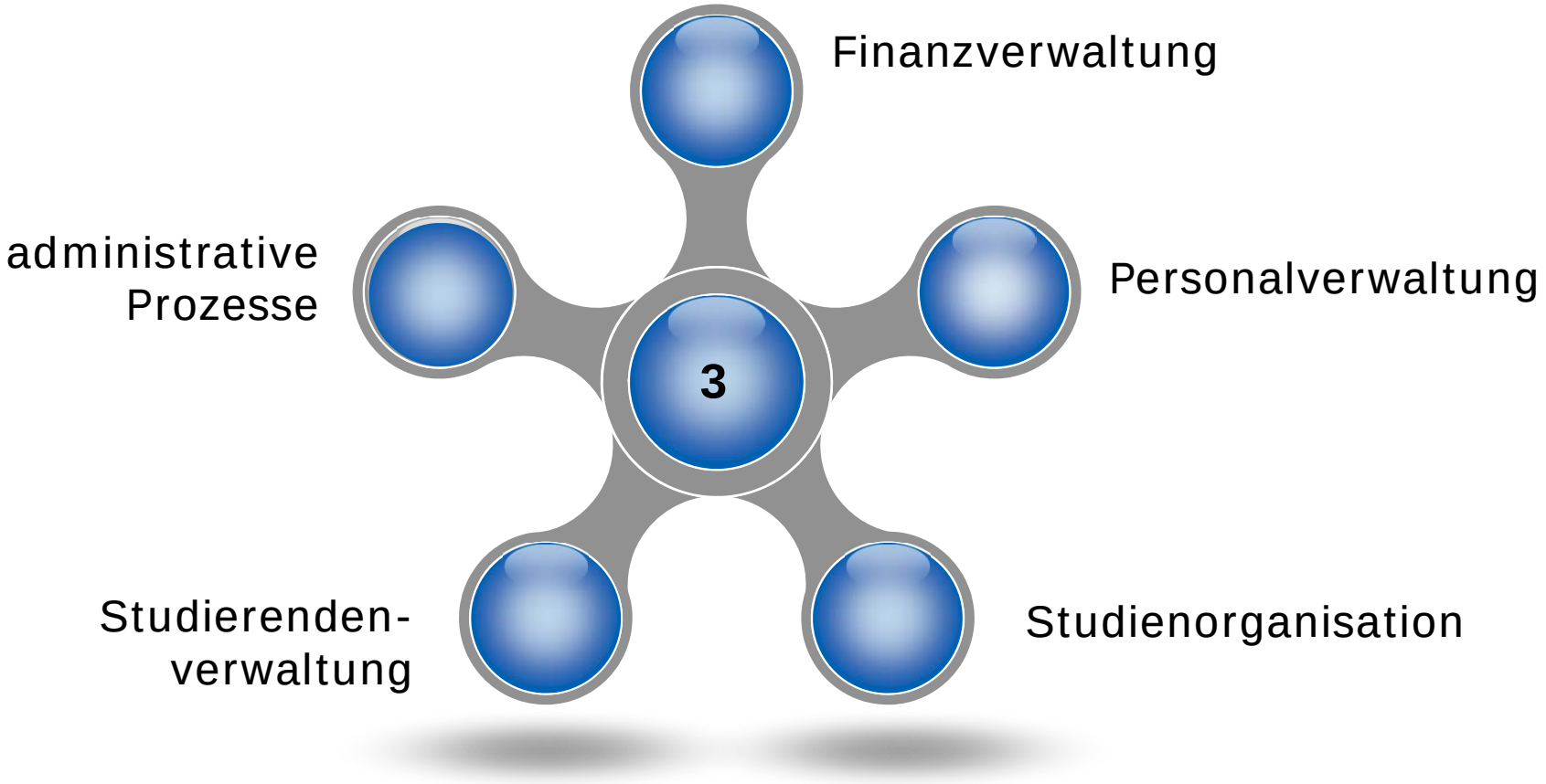
2 Infrastruktur



HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Die IKM – Strategie der HHU

3 Campus Management



HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Langzeitarchivierung an der HHU
Projektentwicklung

2010:

- Auftrag zu einer Studie durch Rektorat auf Anregung der Universitätsbibliothek
- Zusage des Rektorats zur Erhaltung von Forschungsdaten gemäß guter wissenschaftlicher Praxis

2011:

Bewilligung eines Vorprojekts

2013:

Abschluss Vorprojekt im Juni

Vorprojekt 2012

- Pilotanwendung für eine hierarchische Speicherverwaltung
- Proof of Concept mit Rosetta, OAIS konform
- Entwurf, Spezifizierung und Evaluierung einer geeigneten technischen Infrastruktur
- Datenlieferanten: Bibliothek, Forscher, Klinik
- Feststellen des Finanzbedarfs

Kooperationspartner:

Zentrum für **I**nformations- und **M**edientechnologie ←→

Universitäts- und **L**andes **B**ibliothek ←→

Universitäts**K**linikum **D**üsseldorf

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Bedarf aus Universitäts- und Landesbibliothek (ULB)

- Digitalisierung von z.B. Handschriften, Inkunabeln, besonderen Sammlungen
  Formate: tiff, jpeg, pdf
- Gesetz zur „Ablieferung des elektronischen Pflichtexemplars“
- *Düsseldorfer Dokumenten- und Publikationsservice* zur Veröffentlichung von Hochschulschriften, Aufsätzen etc. (Open Access)

Biologisch-Medizinisches Forschungszentrum (BMFZ)

			
SOLiD™ 5500XL Genetic Analyzer	Illumina HiSeq 2000	IonTorrent PGM™	*Ion Proton™ System*
Ca. 500 GB/ Lauf (2 Flowcells, 2 Wochen) .XSQ-Files (500 GB) .TIFF-Files (3 TB)	Ca. 1000 GB/Lauf (2 Flowcells, 1-2 Woche) .FASTQ-Files	Ca. 20 GB/Tag .FASTQ-Files .SFF-Files	*Ca. 200 GB/Tag*
Ca. 26 Läufe/Jahr 13 TB (88,4 TB TIFF-Files)	Ca. 20 Läufe/Jahr 20 TB	Ca. 200 Läufe/Jahr 4 TB	*Ca. 100 Läufe* *20 TB*

Bedarf aus Universitätsklinikum Düsseldorf (UKD)

Art der Daten

- Patientendaten, bestehend aus
  - Datenbank-Einträgen (werden i. d. R. nicht archiviert)
  - Dokumenten, wie Arztbriefe, Befunde, Behandlungsverträgen usw. die in der Patientenakte enthalten sind
  - Bildinformationen

- Im radiologischen Bereich (PACS) gibt es eine Aufbewahrungsfrist von 30 Jahren, in anderen Bereichen sind es i. d. R. 10 Jahre

- UKD: Projekt zur Digitalisierung der Patientenakten seit 2010
  - Eingescannt und in E-Archiv SHA gespeichert
  - Befunde der Subsysteme an Medico ➔ Archivierung über Medico

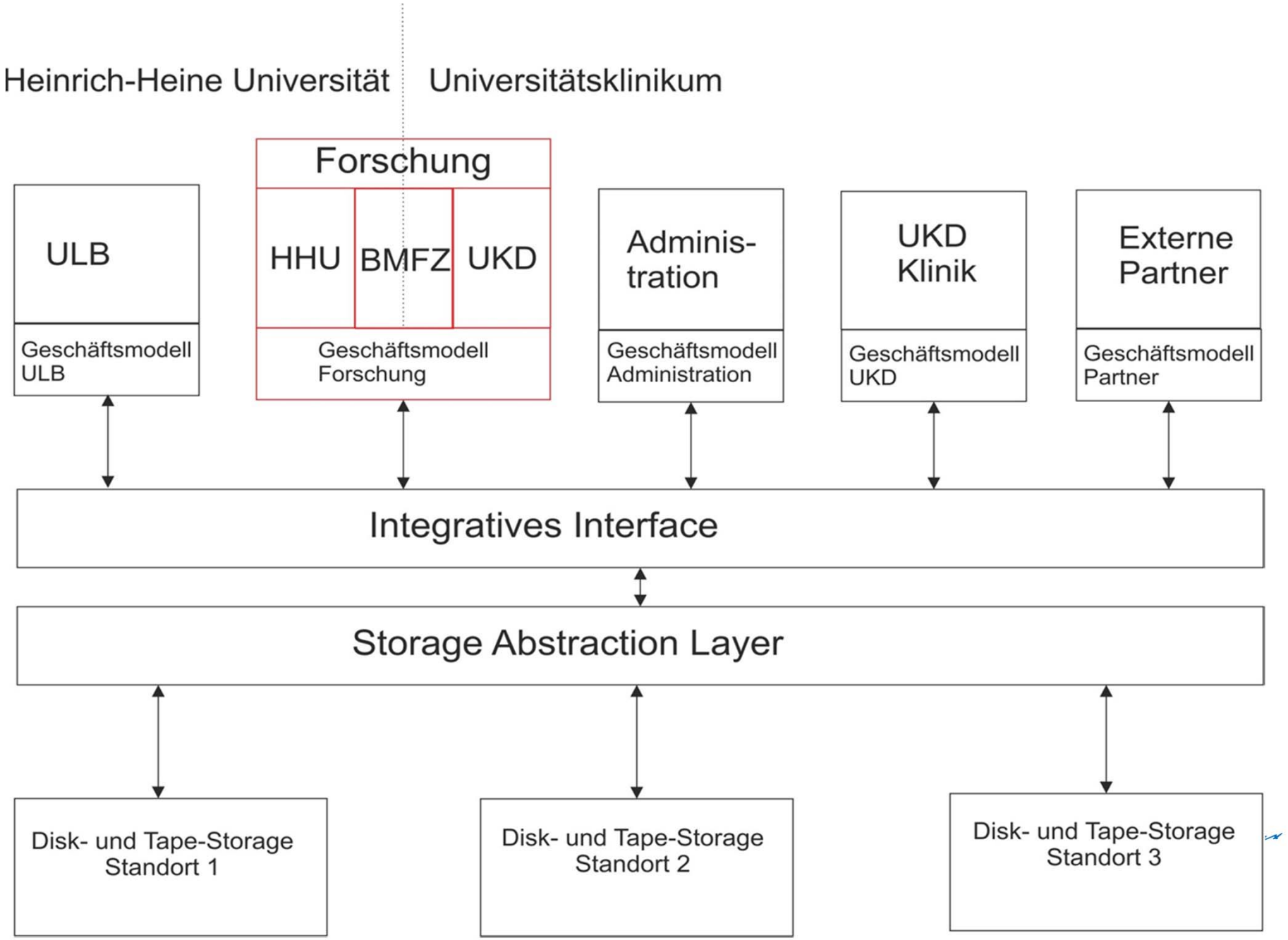
- SHA: Silent Cube-Technologie von FastLTA für Revisionssicherheit

Konzeptioneller Entwurf und technische Umsetzung

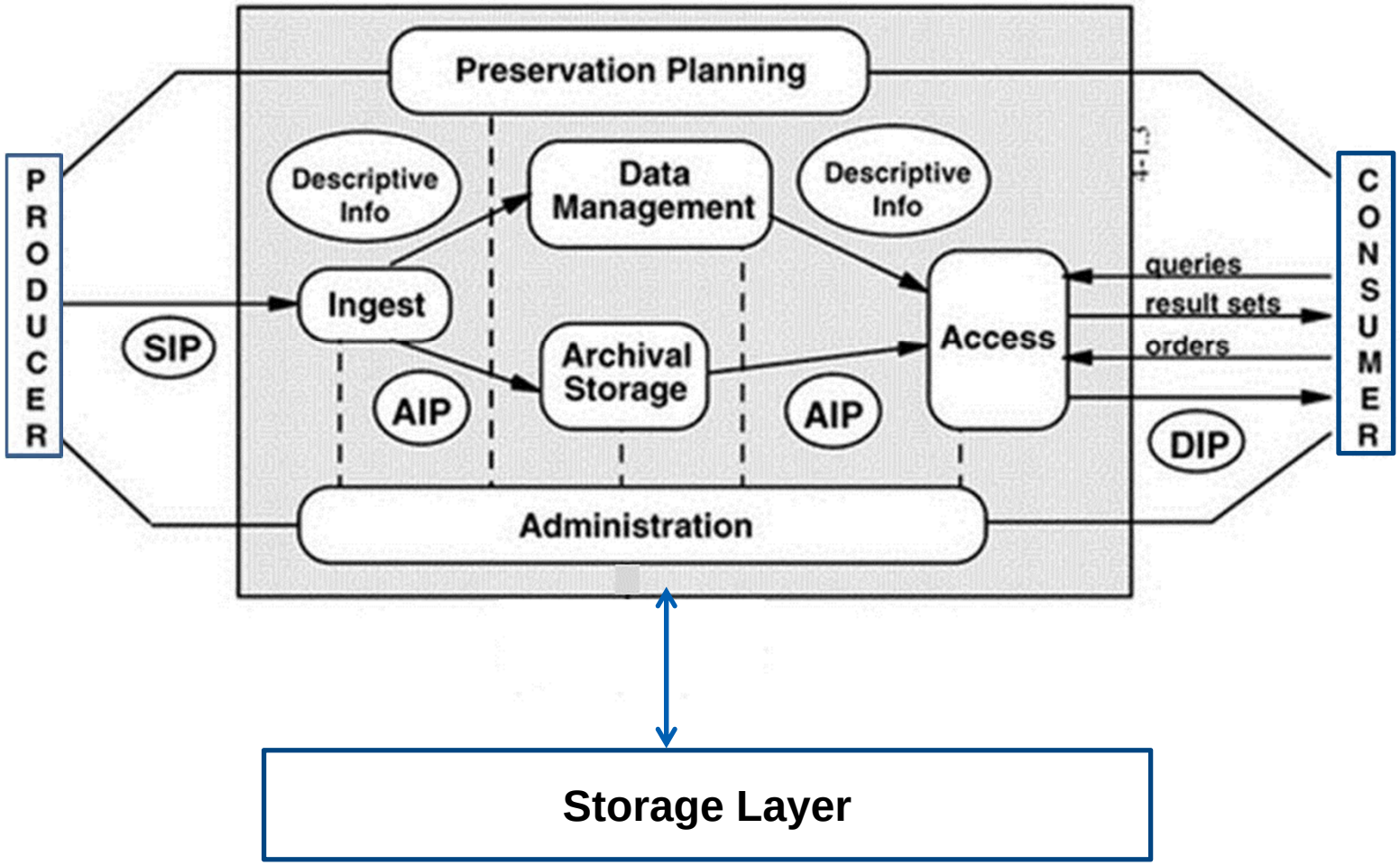
- Integratives Interface:
  Einheitliches Interface für Speicherung und Wiedergewinnung von Langzeitdaten

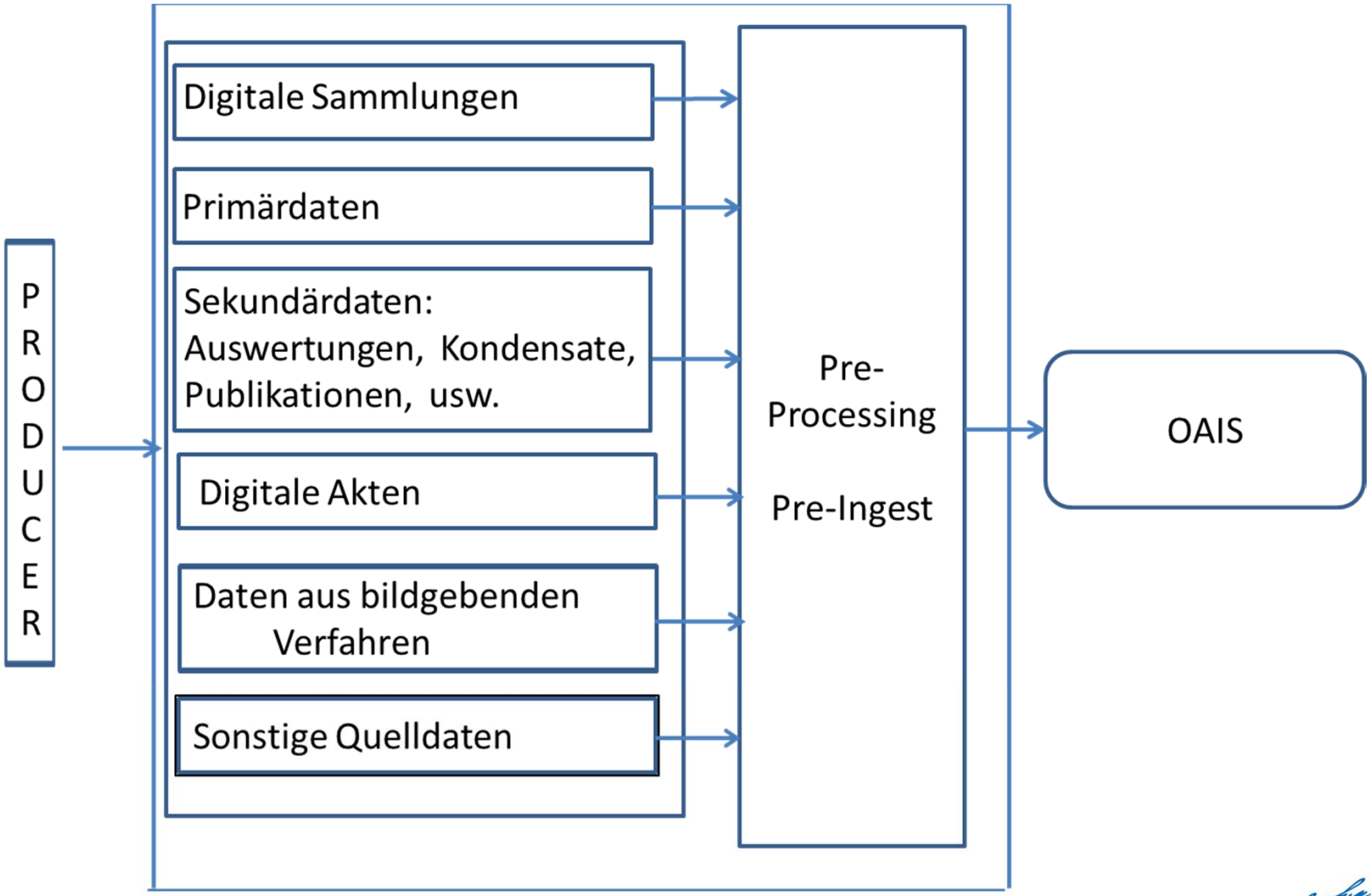
- Heterogene Datenbestände und Technologien:
  - Verteilte Archiv-Datenbestände
  - Unterschiedliche Datenformate
  - Technologisch heterogene Backend-Plattformen
  - Mehrere lokale oder externe Standorte

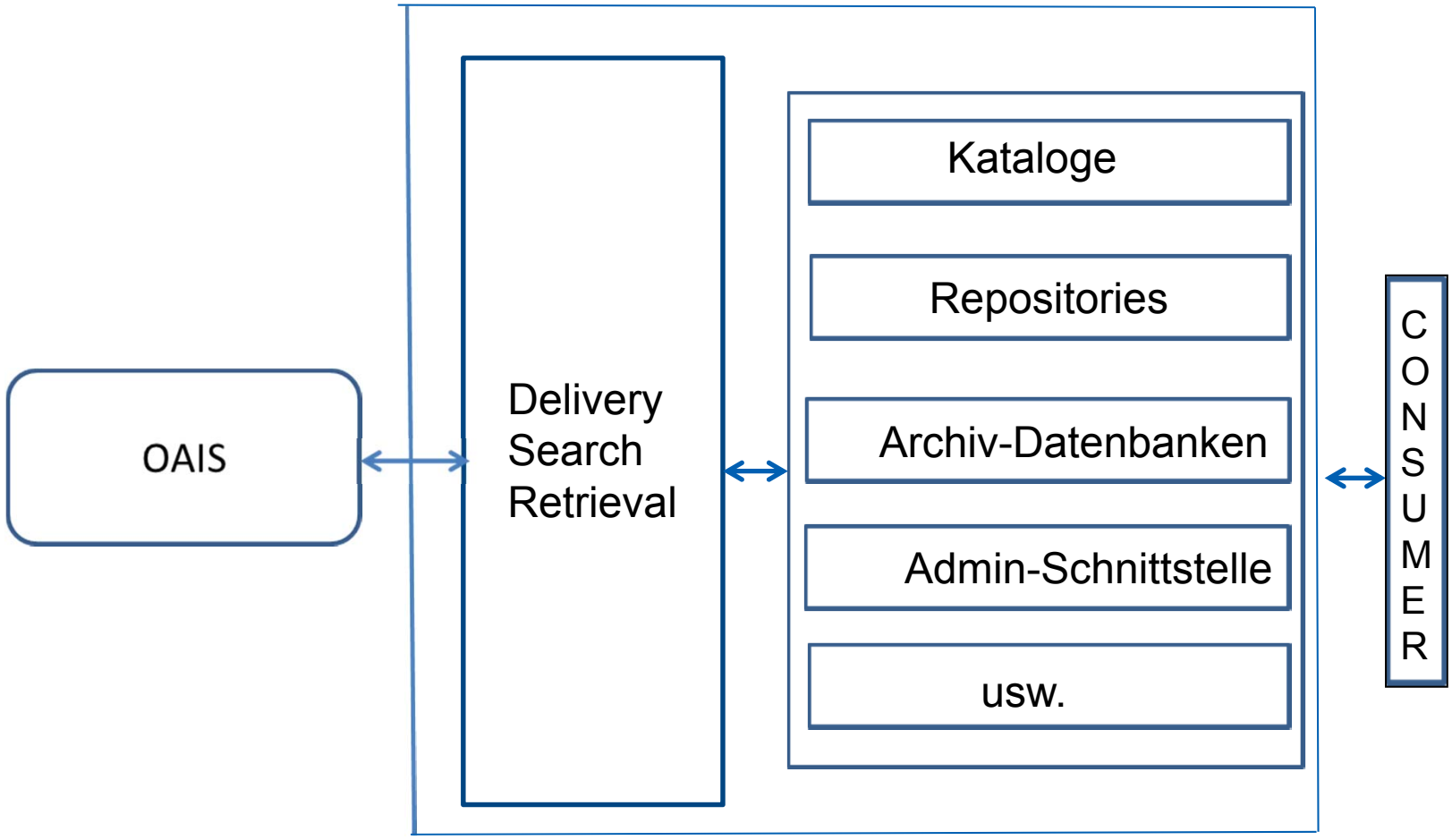
- Komfort und Transparenz:
  - Speicherung oder Rückgewinnung selbst großer Datenmengen
  - Komfortabel und transparent für Anwender (Web-Oberfläche)


Heinrich-Heine Universität
Universitätsklinikum
Forschung
ULB
HHU
BMFZ
UKD
Adminis-
tration
UKD
Klinik
Externe
Partner
Geschäftsmodell
ULB
Geschäftsmodell
Forschung
Geschäftsmodell
Administration
Geschäftsmodell
UKD
Geschäftsmodell
Partner
Integratives Interface
Storage Abstraction Layer
Disk- und Tape-Storage
Standort 1
Disk- und Tape-Storage
Standort 2
Disk- und Tape-Storage
Standort 3

Referenzmodell: Open Archival Information System (OAIS)



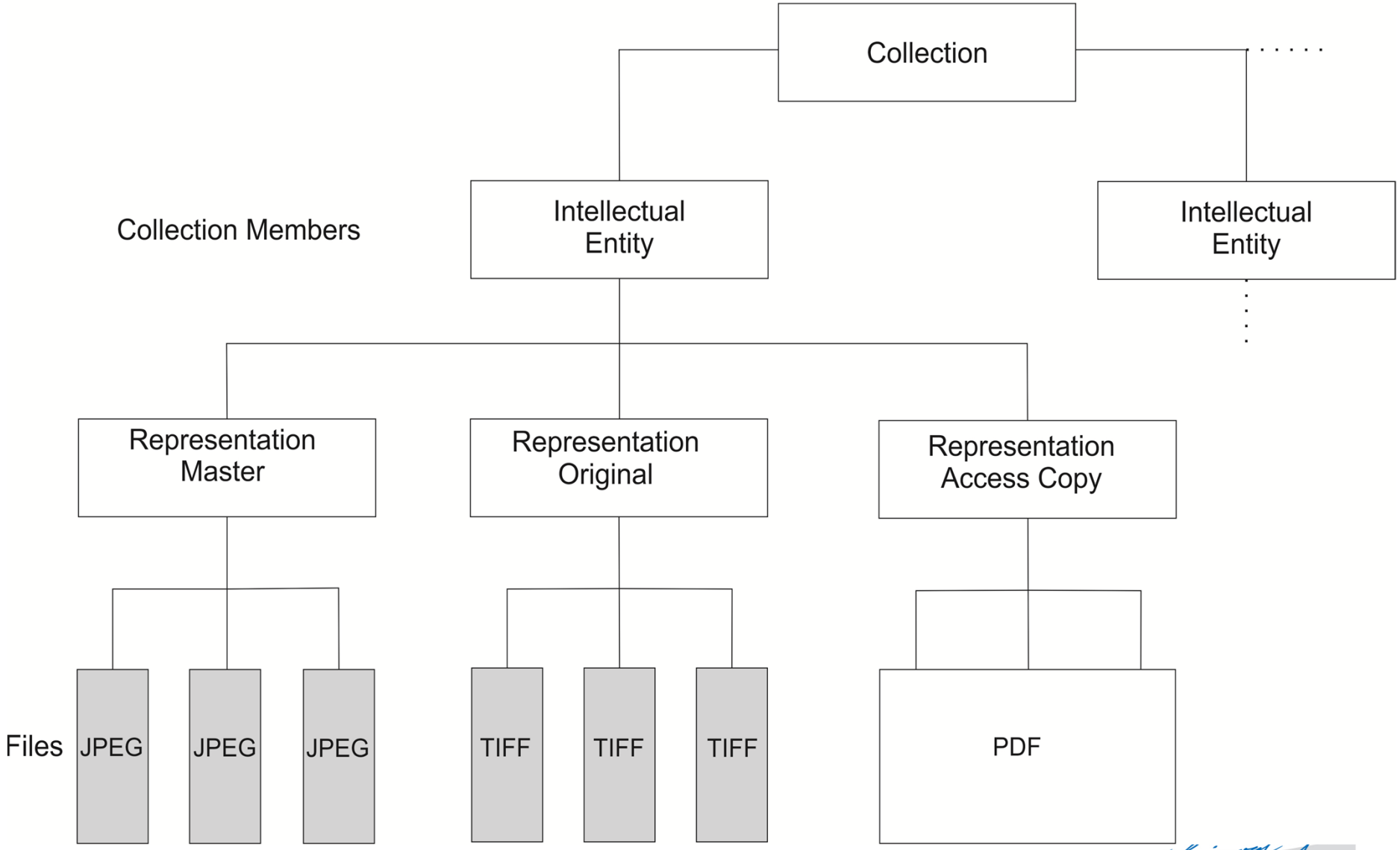



OAIS
Delivery
Search
Retrieval
Kataloge
Repositories
Archiv-Datenbanken
Admin-Schnittstelle
usw.
CONSUMER

Pilot-Projekt für die Langzeitarchivierung

- Erstellung eines ersten Prototypen als Kern eines Langzeitarchivierungssystems
- Archivierung und Wiedergewinnung von Beispieldaten
- Beispieldaten:
  - Biologisch-Medizinisches Forschungszentrum (BMFZ) / Genomics & Transcriptomics Laboratory (GTL): *Sequenzierdaten*
  - Center for Advanced Imaging (CAi): *Mikroskopiedaten*
  - Universitätsklinikum: *CT/MR-Daten*
  - Universitäts- und Landesbibliothek: *Digitalisate von Inkunabeln*
- Basis ExLibris Rosetta – Proof of Concept

Beispiel Digitalisate ULB (Datenmodell)

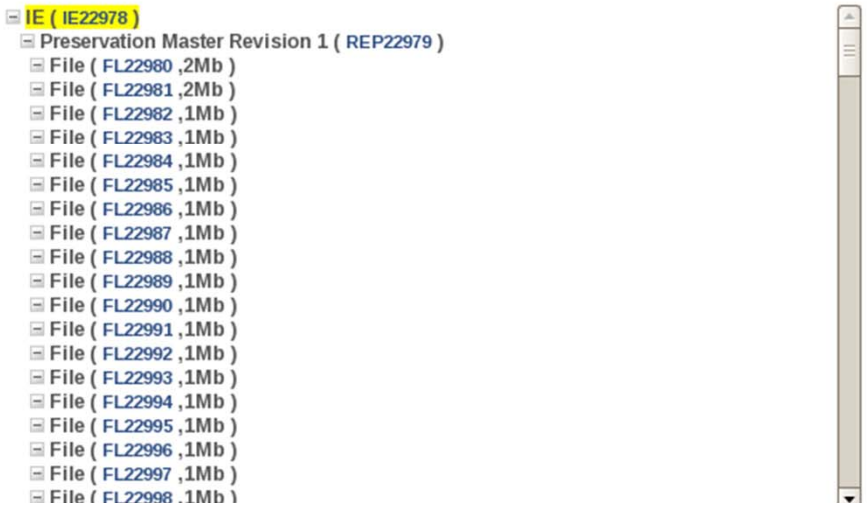


Collections List

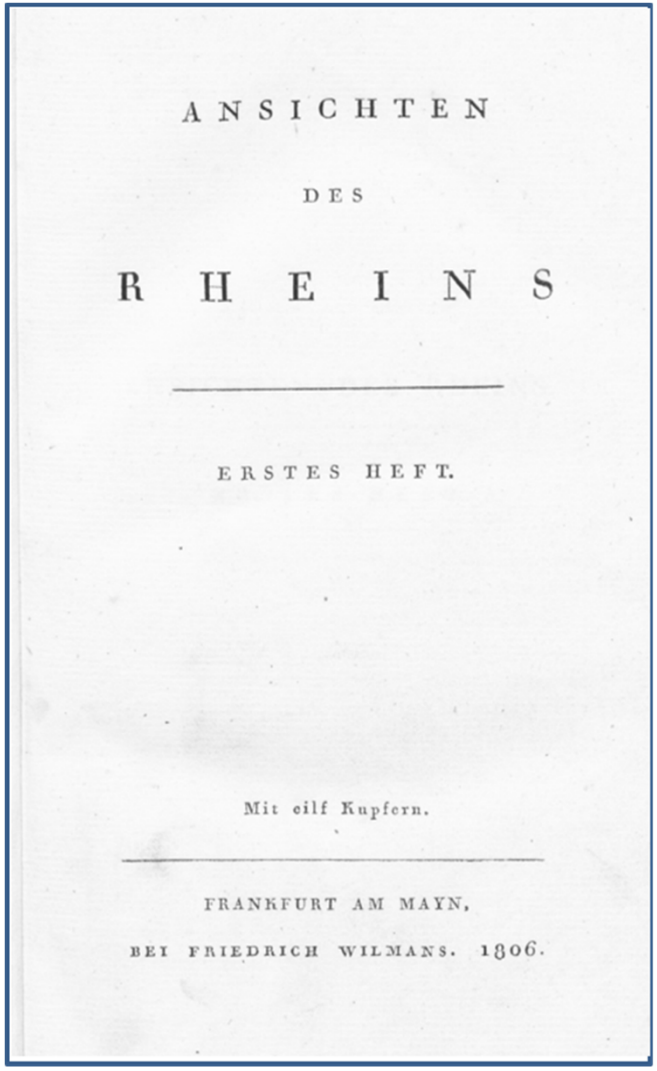
Collection Name	Description	Creation Date	Number of IEs
UKD Collections	Created by AssignCollectionByDCTask (DC source: 'UKD Collections')	11.02.2013 14:27	0
patient_CT	Created by AssignCollectionByDCTask (DC source: 'patient_CT')	12.12.2012 16:45	4
patient_MR	Created by AssignCollectionByDCTask (DC source: 'patient_MR')	18.12.2012 19:14	2
BMFZ Collections	Created by AssignCollectionByDCTask (DC source: 'BMFZ Collections')	11.02.2013 15:21	0
study 1	Created by AssignCollectionByDCTask (DC source: 'study 1')	11.02.2013 15:21	1
CAi Collections	Created by AssignCollectionByDCTask (DC source: 'CAi Collections')	12.02.2013 16:20	0
Wintersemester 2012	Created by AssignCollectionByDCTask (DC source: 'Wintersemester 2012')	12.02.2013 16:20	0
IIF	Created by AssignCollectionByDCTask (DC source: 'IIF')	12.02.2013 16:20	0
Gruppe 1	Created by AssignCollectionByDCTask (DC source: 'Gruppe 1')	12.02.2013 16:20	1
ULB Collections	Created by AssignCollectionByDCTask (DC source: 'ULB Collections')	12.02.2013 19:06	4

Collection Members

IE PID	IE Title	IE Creation Date
IE21727	Statistisches Jahrbuch Nordrhein-Westfalen ...	18.12.2012 14:52
IE21737	Jubel-Kalender zur Erinnerung an die Völkerschlacht bei Leipzig vom 16. - 19. October A. D. 1813	18.12.2012 14:53
IE21886	Amtsblatt für den Regierungsbezirk Düsseldorf	18.12.2012 14:54
IE22978	Katechismus für den teutschen Kriegs- und Wehrmann, worin gelehret wird, wie ein christlicher Wehrmann seyn und mit Gott in den Streit gehen soll	24.01.2013 15:38



IRICH HEINE
SITÄT DÜSSELDORF



ANSICHTEN

DES

RHEINS

ERSTES HEFT.

Mit eilf Kupfern.

FRANKFURT AM MAYN,

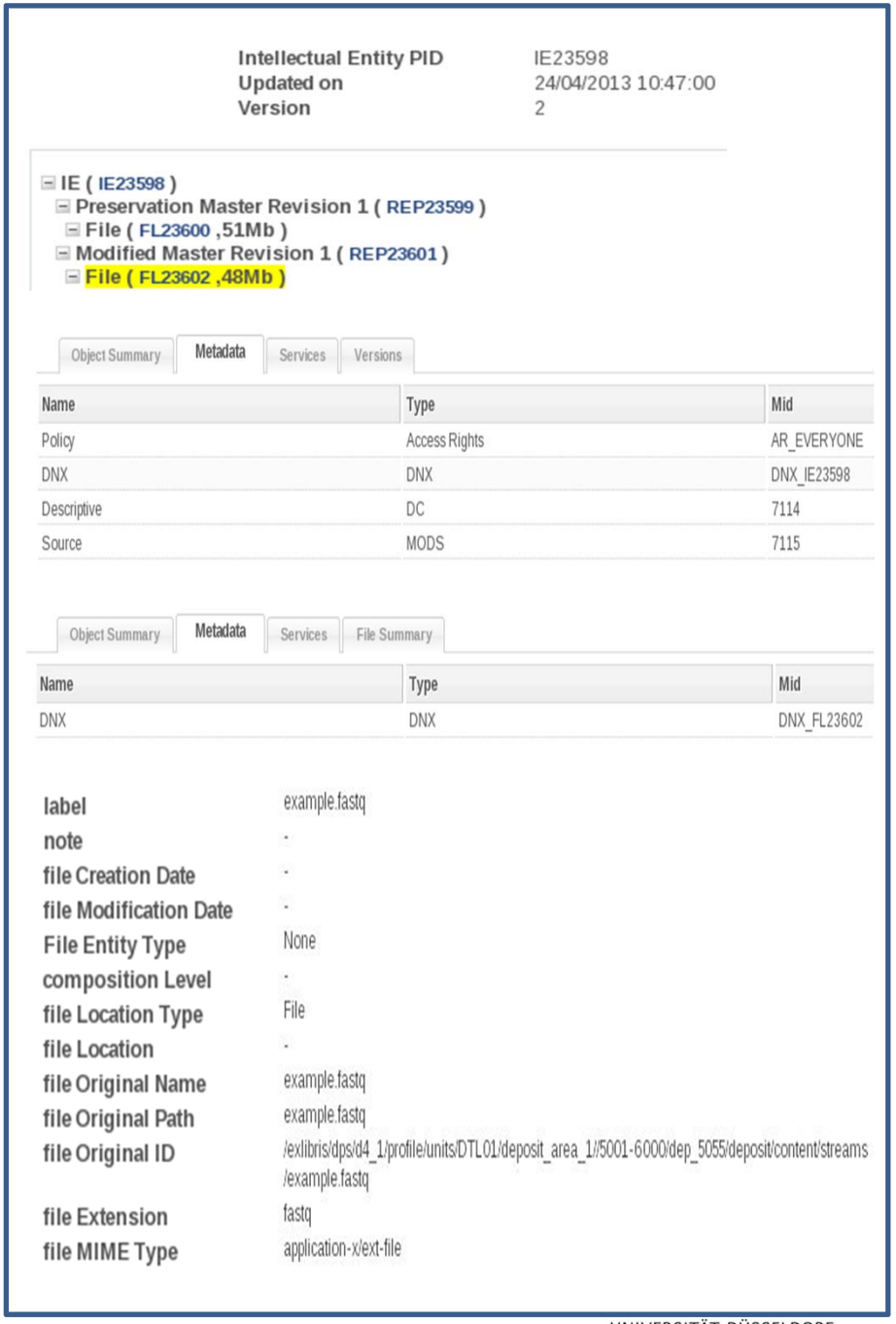
BEI FRIEDRICH WILMANS. 1806.

ULB: Digitalisat


```
<?xml version="1.0"?>
- <mets:mets LABEL="Digitale Sammlungen" OBJID="116" xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/version18/mets.xsd" xmlns:vls="http://semantics.de/vls"
xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:mets="http://www.loc.gov/METS/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  - <mets:metsHdr CREATEDATE="2011-05-31T16:02:17">
    - <mets:agent OTHERTYPE="SOFTWARE" ROLE="OTHER" TYPE="OTHER">
        <mets:name>vls/2.10.1</mets:name>
      </mets:agent>
    - <mets:agent OTHERTYPE="INSTANCE" ROLE="OTHER" TYPE="OTHER">
        <mets:name>ulbd-master-1</mets:name>
      </mets:agent>
    - <mets:agent OTHERTYPE="BUILDER" ROLE="OTHER" TYPE="OTHER">
        <mets:name>vdzip</mets:name>
      </mets:agent>
  </mets:metsHdr>
  - <mets:dmdSec ID="md140839">
    - <mets:mdWrap MDTYPE="MODS" MIMETYPE="text/xml">
      - <mets:xmlData>
        - <mods xsi:schemaLocation="http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-
          3-4.xsd" version="3.4" xmlns="http://www.loc.gov/mods/v3">
          - <titleInfo>
              <title>Mahlerische Ansichten des Rheins von Mainz bis Düsseldorf</title>
              <subTitle>mit 3 Kupfern u. 1 Karte</subTitle>
            </titleInfo>
          - <titleInfo type="alternative">
              <title>Ansichten des Rheins</title>
            </titleInfo>
          - <name type="personal" valueURI="http://d-nb.info/gnd/100400159"
            authorityURI="http://d-nb.info/gnd/" authority="gnd">
              <namePart>Schreiber, Alois Wilhelm</namePart>
            - <role>
                <roleTerm type="code" authority="marcrelator">aut</roleTerm>
              </role>
            </name>
          - <name type="personal">
              <namePart>Vogt, Nikolaus</namePart>
            - <role>
                <roleTerm type="code" authority="marcrelator">aut</roleTerm>
              </role>
            </name>
            <typeOfResource>text</typeOfResource>
            <genre authority="marcgt">book</genre>
          - <originInfo>
            - <place>
                <placeTerm type="text">Frankfurt am Mayn</placeTerm>
              </place>
              <publisher>Wilmans</publisher>
              <dateIssued encoding="w3cdtf" keyDate="yes">1806</dateIssued>
              <issuance>monographic</issuance>
            </originInfo>
```

```
@M4GC8:4:14
AATGGCATTAAAAATAGTTATTTAAGAACAATAATGCCTTATTTTGGTGCTTGTGAGCCTTCTTTCACACATGGG
GAAAGTGAAGCTCTGTAGGGTAAAGTCCAGAAGATCATGGGCCTGCTTAAGTGTCAGAGCTGGATTCAAACCA
AATCTCAGGCTACCTGCTGAAGGTTCTGTTTACCTCCACTTTAGCATGCAGATGACAGAGA
+
A8::18<6.<<<<*<;<@9ACC4>7>A7@<+1,1<57260+000&+156371/0+;<<2:2699/:;9>9=;<>>,>>1;@@>3
<7;;?<::8<19>?27<52656188114599/8-826:2<8@@?@AAA?558@:A<5899&.(-1&099;>@-
82,0(+4//13+.*.032**-&*1-..+./02)..4011533:449222;;;
@M4GC8:4:21
AGCTCAGAACCATGCTTTATTCATAGTAGGGGCTCAACTCATGTTGATTAGTAATAAGCAAGATAAGTGTCTGA
CCGCTAACACCTAATTATTAGATTCCAATGCTGTGCTAGGCCTGTGAGATAATCAGTAACCATCTTTTGTCAATG
ATAATCCTTATAAAGCTAGTTTGACCTTTTATTCATAATGTAAAACTGCTTTGCATTATATTTCG
+
CCCCCC@@;?<??=>>@3@C?CDDC22.?A?.=>=?8=CCCCBE?EEE@CCCC?CE@CCC@ECCC@CCCCDDDEE
@C??@>CCC?CC?C=CC??@?C?D:?;>==;666?B==8=;@@@@@=<;<8:::<=?8=7<;8:<<,:>;;4;9755.33,328
77?-6387-.2&-32,2<<,;;6:6660553388)3....=099==7=;;;>/7.
@M4GC8:4:32
CAGAGCTGGATTCAAACCCAAATCTCAGGCTACCCTGCTGAAGGTTCTGTTTTACCTCCACTTTTAGCTAGACAG
ATGACAGAGTCACATCTTAGTTCCCAAGAAGAGGCATTTTCTGAATTACTTTCCCCCTTTCTCAACTGTTGTATTG
AAAGTACTTTCTTCTGAAAAAGTAATTAACAAAA
+
86<8>:6=4551164'..%-2)1263152;<635-
547306/3,2.13++..%+4.54*//155*.1./2+++043778911199::77764/423-55*2,--(*22+/1113*0...-
5126:=19::9(99-4144++.*.+***/.,,0(.20114+++%-/,,000',44/1,413366'
@M4GC8:4:42
GAAAACATGCAGGTGATTGTCAGAATGCCCTAATGAAAGAATGCCAATGCAGTTAATATACTTCTATTCCCCTCC
CATAGCTCACCCCTTGAAGCTTGGGAAGCATCTCGTAGCAAGTCTTGTGTTCAAGTGCTTCCGATGGCTACTTTC
TAGGTTCTTCAAAGCCAGGCAATA
+
0044*.9<<899.518503=????85978,02*024:.26(568264;:7;;;/5/8:88::188,,*+97(++.&+5111473333(0.62
&+-7147)3(341761/374+-+04433-****%-*(*0,,.,404521/110/36*//32/1+0,(-*-'**(**(**).*
@M4GC8:5:9
TCTGAGTGAATTTCCTCCCGGAAAGACCTGCCTAGAAGCATTCATGCATAGTTTTCAGGCAATAATTTCAAGAAA
TGTAAAACAAAAATTATGGCGTCAGCAAAAATCTATCTGCTGATAGATTTTCCCTAACTAAAAAAAAAAAGCAA
T
+
:EEEDDCBB088.85>?@2@8CC5CCE?DCC@CCCC?CCDE>BACCDAA?.222'2:?9BA6;;(22'-
>8?EE5DCB@A@.8????-?9CCB<D?==C?>>EEE3B@22.8688>9@@::;466(84*,0(++*---------%*22(*
```

BMFZ: Sequenzierdaten (FASTQ-Format)



Intellectual Entity PID	IE23598
Updated on	24/04/2013 10:47:00
Version	2

- IE (IE23598)
  - Preservation Master Revision 1 (REP23599)
    - File (FL23600 ,51Mb)
  - Modified Master Revision 1 (REP23601)
    - File (FL23602 ,48Mb)

Object Summary | **Metadata** | Services | Versions

Name	Type	Mid
Policy	Access Rights	AR_EVERYONE
DNX	DNX	DNX_IE23598
Descriptive	DC	7114
Source	MODS	7115

Object Summary | **Metadata** | Services | File Summary

Name	Type	Mid
DNX	DNX	DNX_FL23602

label	example.fastq
note	-
file Creation Date	-
file Modification Date	-
File Entity Type	None
composition Level	-
file Location Type	File
file Location	-
file Original Name	example.fastq
file Original Path	example.fastq
file Original ID	/exlibris/dps/d4_1/profile/units/DTL01/deposit_area_1/5001-6000/dep_5055/deposit/content/streams/example.fastq
file Extension	fastq
file MIME Type	application-x/ext-file

Pre-Processing / Pre-Ingest

- SIP Package Handler (z.B. Docupack)
- Viewer und Editor für Strukturierung und Metadatenerfassung (halbautomatisch)
- SIP (Submission Information Package) für den Ingest des LZA-Systems
- Erzeugung von Struktur und Metadaten für automatische Verarbeitung durch das LZA-System

Ergebnis Vorprojekt
Vorschlag an die Universitätsleitung für das weitere Vorgehen

- Aufbau eines Kompetenz-Zentrums für langfristige Digitale Datenerhaltung
- Entwicklung einer Service-Schnittstelle
  - zum Management und der Archivierung von Forschungsdaten
  - Unterstützung des Enterprise Content Management (ECM) der Hochschule insbesondere der Universitätsverwaltung
  - Unterstützung beim Management und der langfristigen Erhaltung der digitalen Sammlungen der ULB
  - Ressourcen-Sharing mit dem UKD

Konkrete Arbeitsschritte

- Entwicklung und Einsatz eines Web-basierten halb-automatischen Pre-Ingests
- Auswahl und Implementation einer geeigneten LTP-Software (gemäß OAIS)
- Ausbau der notwendigen Infrastruktur
- Entwicklung von Repräsentations- und Auslieferungsmethoden

Infrastruktur – gemeinsame technische Basis

- Storage verteilt über **3** Standorte
- Verwendung von Block- und NAS-Storage für Datenbanken und Pre-Ingest

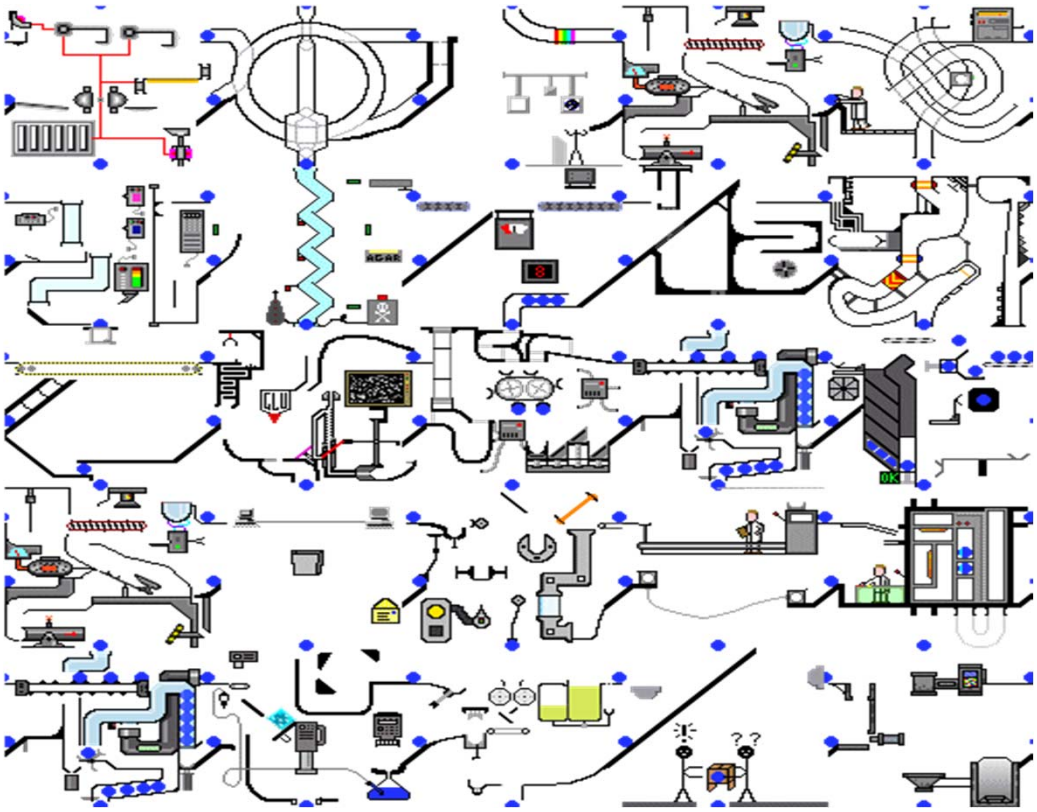
NAS - Filer

- Einsatz von über die Standorte verteilten und sich synchronisierendem Objekt-Storage
- Disk-Storage mit mindestens 1 PByte, Wachstumsraten jährlich mindestens 100 TBytes steigend

System Storage

- Tape Libraries an zwei Standorten als zusätzliches langfristiges Sicherungsmedium
  1,5 PByte, Erweiterung 2013 um 1,75 PByte

Bandroboter



Dr. Walther UKD

Danke für die Aufmerksamkeit!