Kabinette des Wissens - Sammlungen der Humboldt-Universität

Ein Kooperationsprojekt der Technischen Abteilung und des Helmholtz-Zentrums für Kulturtechnik

Samobjects Redesign - ein Laborbericht

Geschichte des Projekts:

- Hermann von Helmholtz-Zentrum wurde als interdisziplinäres Zentrum gegründet im Jahr 2000
- Beginn des Projektes "Kabinette des Wissens - Sammlungen der HUB" 1998
- Entwicklung des Datenmodells und des Thesaurus, Erfassung, Digitalisierung
- Eingabe ab Mai 1999 (Access) bzw August 1999 (Imago-Access)
- Die vier zentralen Entitäten sind Dokument, Ereignis, Sammlung und Thesaurus
- Sammlungsgegenstand als dokumentarische Bezugseinheit (DBE)
- ca. 23500 Objekte in der Datenbank, davon 13000 online einsehbar
- Erschließungstiefe variiert nach Sammlung und Objekttyp, generell aber recht groß
- Auslaufen der wissenschaftlichen Stellen Anfang 2008 (Ende DFG-Förderung)

Technikgeschichte des Projekts:

- Anfänglich ein in C programmiertes Interface, im Hintergrund Access-Datenbank
- Aufgrund von Performanceproblemen Suche nach alternativer Datenbanklösung
- Portierung auf Sybase-Datenbank 2002, im Wesentlichen ohne große Modifikation der Datenbankstruktur
- Keine dauerhafte Lösung, besonders der immer mächtiger werdende Thesaurus machte Probleme
- zusätzlich machte proprietäres Entwicklungsmodell immer mehr Probleme
- ab 2004 daher Suche nach einer alternativen Software
- 2005 erfolgte Migration auf Java-basierte Archivierungssoftware, die an der TA entwickelt wurde
- 2006/2007 erfolgte die Entwicklung einer Webpräsenz, die im Mai 2007 online gestellt wurde

:::> 3 grundlegende Technologiesprünge im Verlauf des bisherigen Projektes, der vierte steht demnächst bevor und soll das System auf zeitgenössische Spezifikationen umstellen

Sichtbarkeit und Nutzung Content:

- Zugriffszahlen seit Jahren relativ hoch und konstant

- bei vielen Begriffen und Inhalten gutes Pageranking in Suchmaschinen

- dementsprechend generiert die Seite viele Anfragen:
  - Forschunginteresse am Content
  - zahlreiche Forschungsprojekte, besonders mit Lautarchivmaterial
  - Illustration für wissenschaftliche Publikationen
  - Featurebeiträge
  - Dokumentarfilme
  - Material für Ausstellungen

- HU-intern in der Vorbereitung von Jubiläen und Ausstellungen

Effekte - Referenzierung und Visibility

Zitate aus "Empfehlungen des Wissenschaftsrates zu Sammlungen als Forschungsinfrastruktur"

- **Sichtbarkeit generieren:**

"Mangels Sichtbarkeit stehen die entsprechenden Sammlungen für die wissenschaftliche Forschung nicht oder nur begrenzt zur Verfügung."

- **Interdisziplinäre Zugänge ermöglichen:**

"Objektbasierte Sammlungen sind auch unabhängig von ihrem originären Entstehungskontext prinzipiell offen für die Bearbeitung durch unterschiedliche Fachrichtungen und können dazu beitragen, aktuelle Theorien zu prüfen."

- **Infrastruktur bauen:**

"Sammlungen sind als Forschungsinfrastruktur „einzigartige Einrichtungen, Ressourcen und Dienstleistungen“, die die Forschung ganzer Wissenschaftsgemeinschaften ermöglichen und erleichtern."

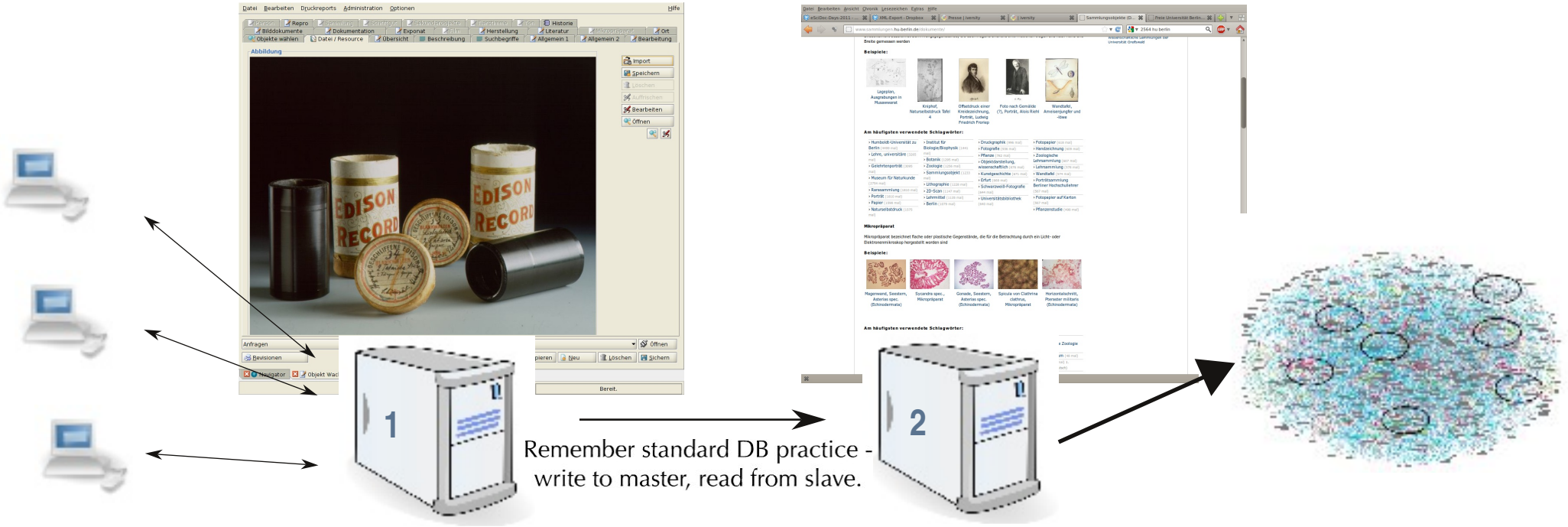
- **Überblick verschaffen:**

"Das Gros der Universitäten hat aufgrund eines unzureichenden analogen oder digitalen Erfassungs- und Erschließungsstandes keinen umfassenden Überblick über jeweils vorhandene Bestände."

:::> Kabinette erfüllen diese Prämissen, wenn auch mit Abstrichen

Technische Infrastruktur:

- Bearbeitung/Eingabe und Ausgabe mit je spez. Funktionalitäten erfolgt auf 2 getrennten Systemen
- Kabinette des Wissens teilt sich auf 2 Maschinen:
  - Bearbeitersystem Sam@Work: enthält alle Informationen und Objekte (1)
  - Webserver www.sammlungen.hu-berlin.de liefert für WWW freigegebenen Content (2)



Technische Kerndaten Sam@Work:

- Unterscheidung in 3 Layer: Infrastruktur, Implementierung und Inhalte
- RIAD: Java-zentriertes Rapid Application Development Framework
  - Eigenentwicklung der Technischen Abteilung (TA) der HU
  - gängige Toolchains: Java, Ant, Svn, Trac, Reporttools, etc
  - Corefunktionalität: End-zu-End-Verschlüsselung, Tunnelung, Hash-Generierung, Revisionssystem, Historisierung, Logging, Nutzer-, Ereignis-, Rollenverwaltung
  - Daten, Abhängigkeiten, Regeln und Metadaten werden unabhängig von der Datenbank in XML in der Middleware abgebildet
  - weitestgehende Herstellerunabhängigkeit
  - strictly Open Source: alle verwendeten Komponenten verwenden Lizenzen, die keine Abhängigkeiten von bestimmten Produkten erzeugen
  - die Software wird interessierten Projekten zur Verfügung gestellt
  - für eine eigene OpenSource-Lizenz sind wir prinzipiell offen

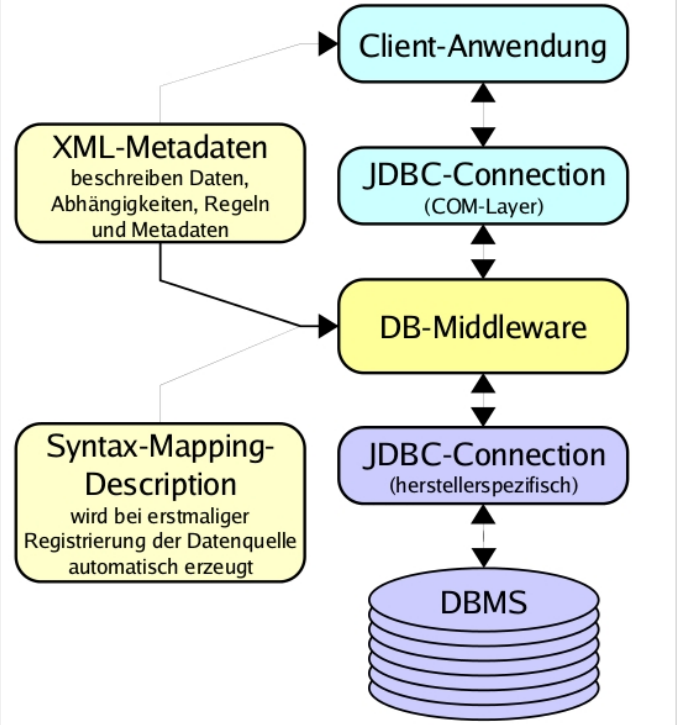


Abbildung 14: Struktur der DB-Middleware

Implementation Sam@Work:

- Client-Server-Architektur in Java
- Schwerpunkt auf Workflow und Effektivität
- Stabilität: Sam@Work seit Jahren im täglichen Produktivbetrieb
- Sicherheit: Java, Verschlüsselung, Trennung von Master- und Webserver
- Tunnelung der gesamten Applikation durch einen beliebigen Port
- Konfigurierbarkeit: Per XML in Entwicklungsumgebung
- Flexibilität: neue Module lassen sich schnell implementieren, dank Java Einbindung von 3rd-party Libs einfach
- Integriertes Digital Asset Management: automatisierte Generierung von Thumbnail, Preview und Webansichten
- Verwaltung beliebiger Mimetypen und Dateiformate
- Integration von OpenSource-Libraries zur automatisierten Contentverarbeitung
- Exportfunktionalität in Pdf, Xml, Excel, Html
- Librarymodule zur Erstellung spzifischer exportierbarer Abfragen an den Datenbestand
- Masseningest und multipler Bearbeitungsmodus für eine beliebige Anzahl an Datenobjekten
- und einiges mehr...

HERMANN VON
HELMHOLTZ-ZENTRUM
FÜR KULTURTECHNIK

Kabinette des Wissens
Kooperationsprojekt Technische Abteilung
Helmholtz-Zentrum für Kulturtechnik

Implementierung Webinterface

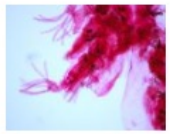
- Python basiertes Webdevelopment Framework
- Priorität Darstellung der Reichhaltigkeit der Beziehungen: "Rich Content"
- daher Aufarbeitung des DB-Contents für das Web und Vorkonfektionierung von Relationen

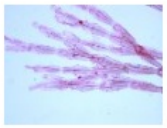
Mikropräparat

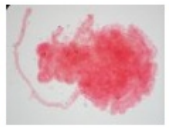
Mikropräparat bezeichnet flache oder plastische Gegenstände, die für die Betrachtung durch ein Licht- oder Elektronenmikroskop hergestellt worden sind

Beispiele:

Plumatella repens (Tentaculata), Mikropräparat


Zoobotryon pellucidum (Tentaculata), Mikropräparat


Gemelaria loriculata (Tentaculata), Mikropräparat


Asteridenlarven in Verwandlung, Seestern (Echinodermata)

Schnitt, Stachelhäuter (Psammechinus miliaris)

Am häufigsten verwendete Schlagwörter:

- Totalpräparat (418 mal) 1. Ganzkörperpräparat (Deutsch)
- Praktikum (360 mal)
- Lichtmikroskopie (250 mal)
- Digitalfoto (235 mal)
- Durchlichtverfahren (218 mal)
- Asteroidea (215 mal) 1. Seesterne (Deutsch)
- Institut für Biologie/Biophysik (181 mal)
- Vergrößerung (170 mal)
- Systematische Zoologie (151 mal)
- Zoologie (145 mal)
- Araneae (116 mal) 1. Webspinnen (Deutsch)
- Färbung, künstliche (93 mal)
- Mikrotomschnitt (90 mal)
- Kükenthal, Willy (84 mal)
- Glas (76 mal)
- Mikroskopie (69 mal)
- Berlin (69 mal)
- Systematik (69 mal)
- Taxonomie (69 mal)
- Morphologie (69 mal)
- Tentaculata (69 mal) 1. Kranzfühler (Deutsch)
- Echinodermata (69 mal) 1. Stachelhäuter (Deutsch)
- Querschnitt (66 mal)
- Naturwissenschaft (62 mal)
- Institut für Systematische Zoologie (58 mal)
- Lebensstadium (48 mal)
- Porifera (35 mal) 1. Schwämme (Deutsch)
- Zoologische Lehrsammlung (33 mal)
- Präparat, mikroskopisch (33 mal)
- Transversalschnitt (32 mal)
- Horizontalschnitt (24 mal)
- Biologie (23 mal)

Thesaurus (Schlagworthierarchie)

Fachgebiet (69 direkte Unterbegriffe, 222 insgesamt)
Verknüpfte Dokumente (5043)

Körperschaft (41 direkte Unterbegriffe, 1304 insgesamt)
Verknüpfte Dokumente (11452)

Lebewesen (4 direkte Unterbegriffe, 16672 insgesamt)
Verknüpfte Dokumente (2711)

Material (44 direkte Unterbegriffe, 119 insgesamt)
Verknüpfte Dokumente (10564)

Ort (7 direkte Unterbegriffe, 1268 insgesamt)
Verknüpfte Dokumente (10659)

Person (10 direkte Unterbegriffe, 1908 insgesamt)
Verknüpfte Dokumente (10174)

Repro/ Medium (9 direkte Unterbegriffe, 14 insgesamt)
Verknüpfte Dokumente (5489)

Sammlung (43 direkte Unterbegriffe, 59 insgesamt)
Verknüpfte Dokumente (12010)

Sammlungsobjekt (20 direkte Unterbegriffe, 323 insgesamt)
Verknüpfte Dokumente (12150)

Sprachen (20 direkte Unterbegriffe, 395 insgesamt)
Verknüpfte Dokumente (6238)

Technik (22 direkte Unterbegriffe, 252 insgesamt)
Verknüpfte Dokumente (10586)

Thema/ Forschung (167 direkte Unterbegriffe, 1517 insgesamt)
Verknüpfte Dokumente (10815)

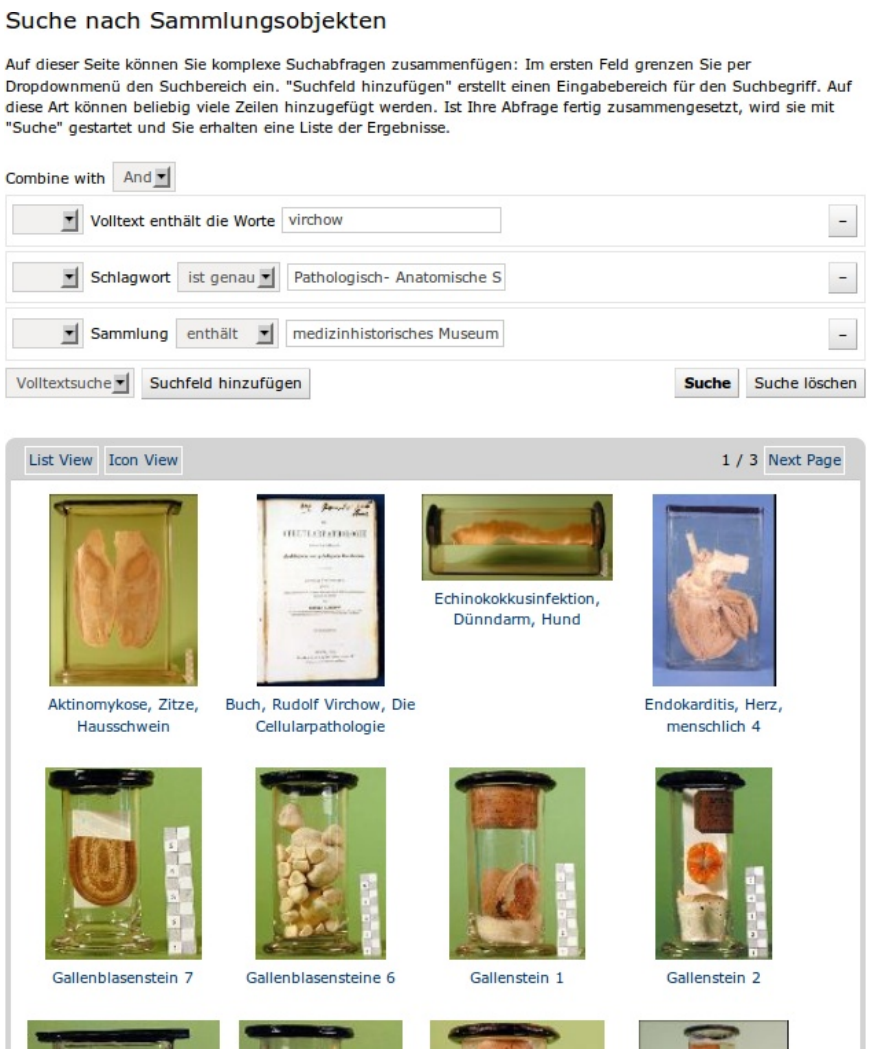
Zeit (12 direkte Unterbegriffe, 190 insgesamt)
Verknüpfte Dokumente (6678)

Suchoptionen:

- **Einfache Suche für den schnellen Einstieg**
- **Komplexe Suchmöglichkeiten durch Generierung von Searchstrings und deren Facettierung:**
  Volltext, Schlagwort (Thesaurus), Dokumenttyp, Feldinhalt, Sammlung
- **oder Suche speziell im Thesaurus**

Technik:

- **Django-Project Webdevelopment-Framework**
- **nur die öffentlich sichtbaren Inhalte werden importiert und für die Darstellung aufbereitet**
- **enthält redaktionelles Modul für Webbeiträge und virtuelle Rundgänge durch die Objektbestände**



Schwerpunkt in den letzten beiden Jahren:

- Umbau zu Portal der Sammlungsaktivitäten und Sammlungserschließungen
- Nutzung für die Arbeit des Sammlungsbeauftragten
- Nutzung für die Belange der Sammlungsleiter: Modul zur Repräsentation aktiver Sammlungen
- Mehrsprachigkeit deutsch / englisch
- "Kabinette des Wissens" trat in den Hintergrund zugunsten allgemeiner Belange der Sammlungen

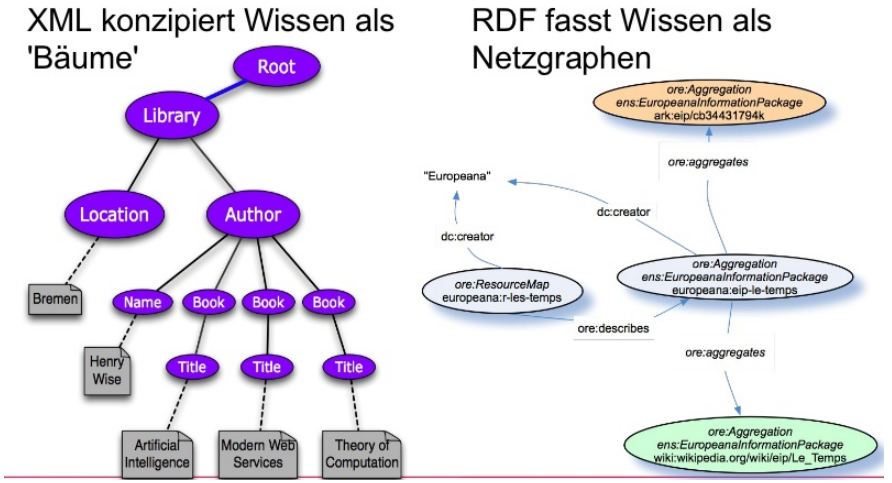
Pläne für die nahe Zukunft

- Eigenes Modul zur Präsentation des Lautarchivs
  - Forschungsdokumentation und Rückfluss der wiss. Beschäftigung mit dem Material
  - Login-basierter Zugang zum Audiomaterial
  - Virtuelle Forschungsumgebung zur Transkription und Übersetzung
- mögliche weitere Module zur Präsentation spezifischer Bestände

Rewrite Samobjects

- Integration neuer Komponenten:
  - Entwicklung von Schnittstellen für Aggregatoren:
    - OAI-PMH wurde als Proof of Concept bereits implementiert:
      http://141.20.145.212/oaiprovider
  - Umstellung des Datenmodells
  - Anbindung kontrollierter Vokabulare über Schnittstellen

- Orientierung an der Entwicklung großer Portale wie Europeana und DDB
  - Diskussion bei Europeana (Switch von ESM zu EDM) zeigt dass die Dinge im Fluss sind
  - Orientierung auf semantische Technologien und "Unschärfe" bei heterogenen Metadaten



Quelle: Gradmann, Linked Open Europeana: Das Europeana Data Model

Relationales Datenmodell Kabinette:

- **Redundanz der abgebildeten Informationen:**

Beispiel Deskriptor- und Dokumenttabelle

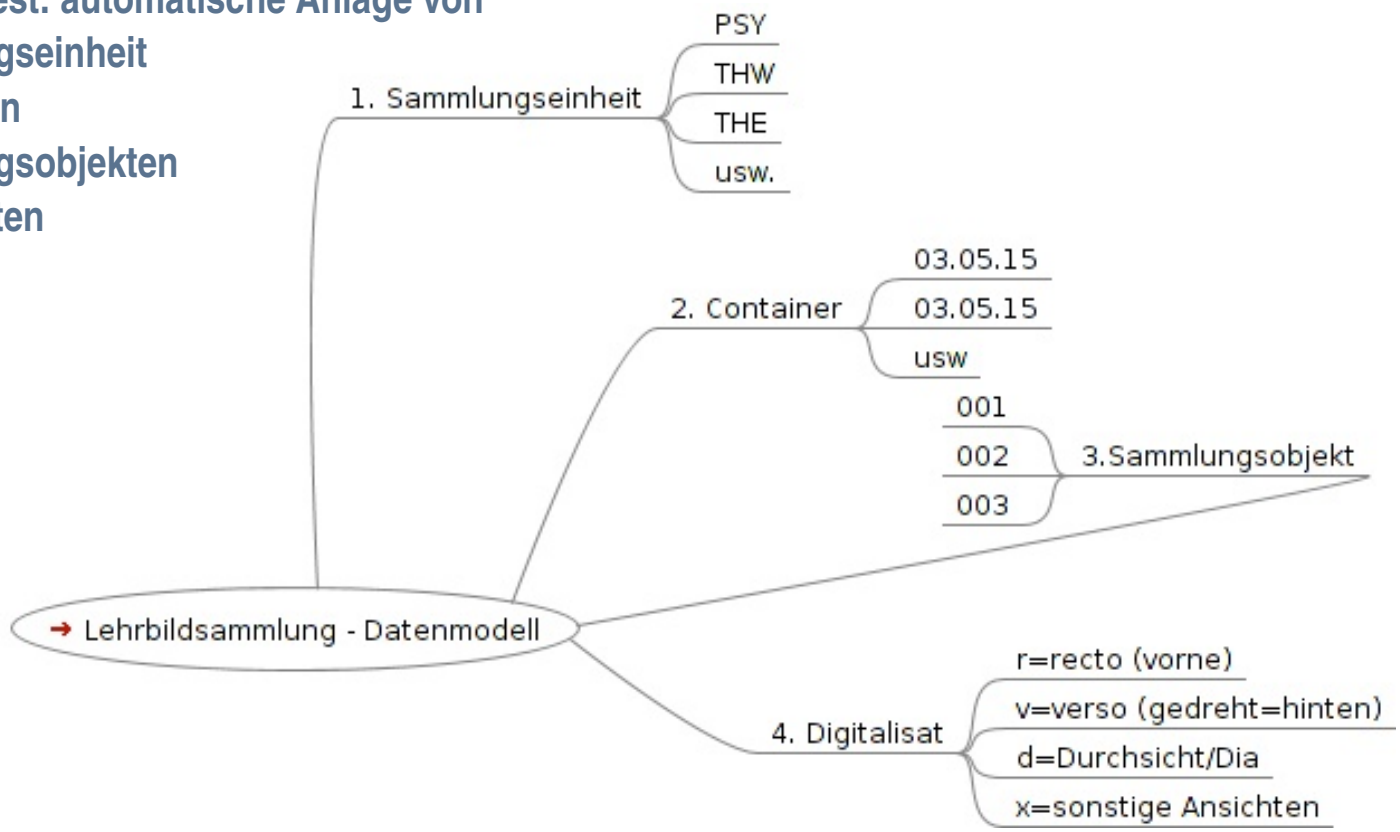
```
[sammlung]> select dokument.id, dokument.name as dokumentname, thesaurus_desc.name as descriptor, thesaurus.name as description from
dokument,thesaurus as thesaurus_desc, thesaurus, dokumentdescriptor where thesaurus_id_descriptor=thesaurus_desc.id and
thesaurus_id_description=thesaurus.id and dokument_id=dokument.id and dokument.id=322;
+-----+-----------------------------------------------------------------------+---------------+-----------------+
| id  | dokumentname                                                          | descriptor    | description     |
+-----+-----------------------------------------------------------------------+---------------+-----------------+
| 322 | Schottisches Englisch (Großbritannien), Der Verlorene Sohn - PK 40  | Sammler       | Brandl, Alois   |
| 322 | Schottisches Englisch (Großbritannien), Der Verlorene Sohn - PK 40  | Projektleiter | Doegen, Wilhelm |
| 322 | Schottisches Englisch (Großbritannien), Der Verlorene Sohn - PK 40  | Aufnahmeort   | Dyrotz          |
| 322 | Schottisches Englisch (Großbritannien), Der Verlorene Sohn - PK 40  | Herkunftsland | Großbritannien  |
+-----+-----------------------------------------------------------------------+---------------+-----------------+

[sammlung]> select * from dokumentdescriptor where dokument_id = 322;
+----+-------------+-------------------------+--------------------------+
| id | dokument_id | thesaurus_id_descriptor | thesaurus_id_description |
+----+-------------+-------------------------+--------------------------+
| 17 |         322 |                     164 |                     1033 |
| 18 |         322 |                    6718 |                     1631 |
| 19 |         322 |                    1710 |                     2064 |
| 20 |         322 |                    6055 |                     2065 |
+----+-------------+-------------------------+--------------------------+
```

Schichten unterschiedlicher Versuche, Aussagen zu modelieren, finden sich an verschiedenen Stellen und zeigen, dass die Datenbank als Testlabor diente

Flexibilität beim Masseningest ermöglicht effektiven Workflow: Stell Dir vor Du nimmst 1000 Objekte auf einmal auf und alle haben anschliessend Namen und sind hierarchisch abgelegt

`THW-03.05.15-003-r.tif = Theaterwissenschaft - 3. Einheit, 5. Kasten, 15. Schublade - 3. Dia rekto`

- **Masseningest: automatische Anlage von**
  - **Sammlungseinheit**
  - **Containern**
  - **Sammlungsobjekten**
  - **Digitalisaten**



Datenstatistik:

- Dokumente: 23357
  - Person 800
  - Sammlung 77
  - Film 5
  - Bild 4267
  - Plastisches Objekt 6746
  - Ton 6609
  - Schrift 1971
  - Sekundärobjekt 1119
  - Tierstimme 1080
  - Mikropräparat 401
- Sammlungen 80
- Ereignisse 6414
- Thesaurus 24284
- Thesaurus Synonym: 21916
- Dokumentbeziehung: 30515
- Dokumentdescriptor: 14542

:::> Beziehungen und Ereignisse spielen eine große Rolle und könnten noch besser abgebildet werden

Generelle Umbauarbeiten am Datenmodell:

- Sekundärobjekt auflösen
- Objekt-Dokument-Hierarchie einführen
- Verschlagwortungsredundanz eliminieren
- Deskriptoren überprüfen und integrieren oder eliminieren
- inflationären Sammlungbegriff auflösen bzw reduzieren (aktiv., als Dokument, als Sammlung, als Schlagwort)

Generelle Designziele für den nächsten Technologiesprung:

- Orientierung an Datenmodell für Universitäre Sammlungen (in the making)
- Anbindung kontrollierte Vokabulare (GND,...)
- Implementierung von Mindeststandards und Standard-Austauschformaten
- Aufarbeiten und Vorbereiten von Metadaten und Digitalisaten für zentrale Aggregatoren
- Berücksichtigung der Kriteriensammlungen von domainspezifischen Akteuren

:: Dank an die Koordinierungsstelle!

HERMANN VON
HELMHOLTZ-ZENTRUM
FÜR KULTURTECHNIK

Kabinette des Wissens
Kooperationsprojekt Technische Abteilung
Helmholtz-Zentrum für Kulturtechnik

Danke für Ihre Aufmerksamkeit!

Kontakt:

- **michael.willenbuecher@hu-berlin.de**
- **+49 30 2093 99969**
- **kabinette@hu-berlin.de**
- **http://www.sammlungen.hu-berlin.de**