GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

UNIVERSITÄTSMEDIZIN
GÖTTINGEN : UMG

Einführung zu Langzeitarchivierung

Frank Dickmann
Universitätsmedizin Göttingen

Agenda

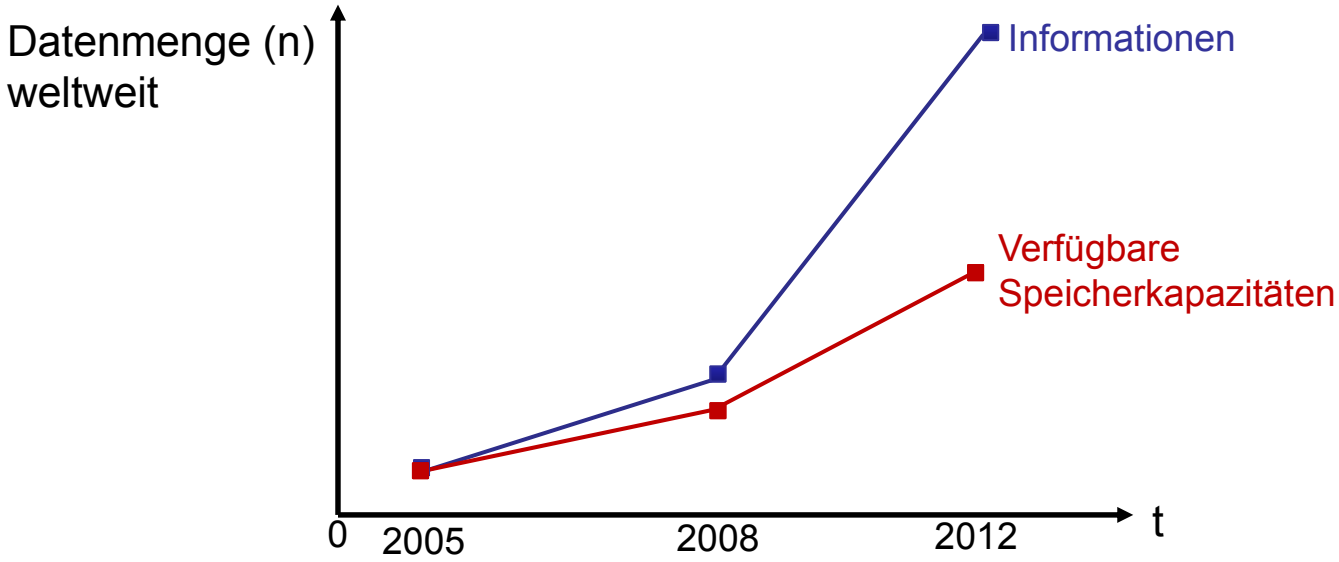
1. Was ist Langzeitarchivierung?
2. Ebenen der LZA
3. LZA-Projekte
   a) nestor
   b) Kopal
   c) KoLaWiss
   d) WissGrid
   e) SHAMAN
   f) e-Helvetica
   g) eArchivierung (TMF)
   h) Medical Archiv Grid

1. Was ist Langzeitarchiverung?

- Langzeitarchivierung bedeutet nachhaltige Bereitstellung von Informationen → 5 bis ∞ Jahre
- Langzeitarchivierung wird vornehmlich in Bibliotheken betrieben
  → erst Bücher, Zeitschriften und Kunstwerke (Bilder, Fotos, Noten)
  → später digitale Inhalte (ePublikationen)
- Langzeitarchivierungsprozesse gehen über übliche IT-Archivierungsprozeduren hinaus:
  - Berücksichtigen der Entwicklungen bei Datenformaten / Datenträgern
  - Einbeziehen von Metadaten
  - Anwenden von Standards des Datenmanagements: z.B. Workflows
  - Berücksichtigen von rechtlichen Rahmenbedingungen: z.B. Urheberrecht
  - **Nachnutzung** über einen langfristigen Zeitraum sicherstellen
  - **Intellektuelle Wiederverwendung** der Daten ermöglichen (kontextbezogen)

1. Was ist Langzeitarchivierung?

- Hintergrund: der steigende Einsatz von IT trägt zu einem exponentiellen Wachstum erzeugter Daten / Informationen bei
- In der Forschung werden Wissenschaftler daher immer häufiger mit Problemen des Datenmanagements konfrontiert



Wachstumstrend von Informationen und Speicherkapazitäten[1]

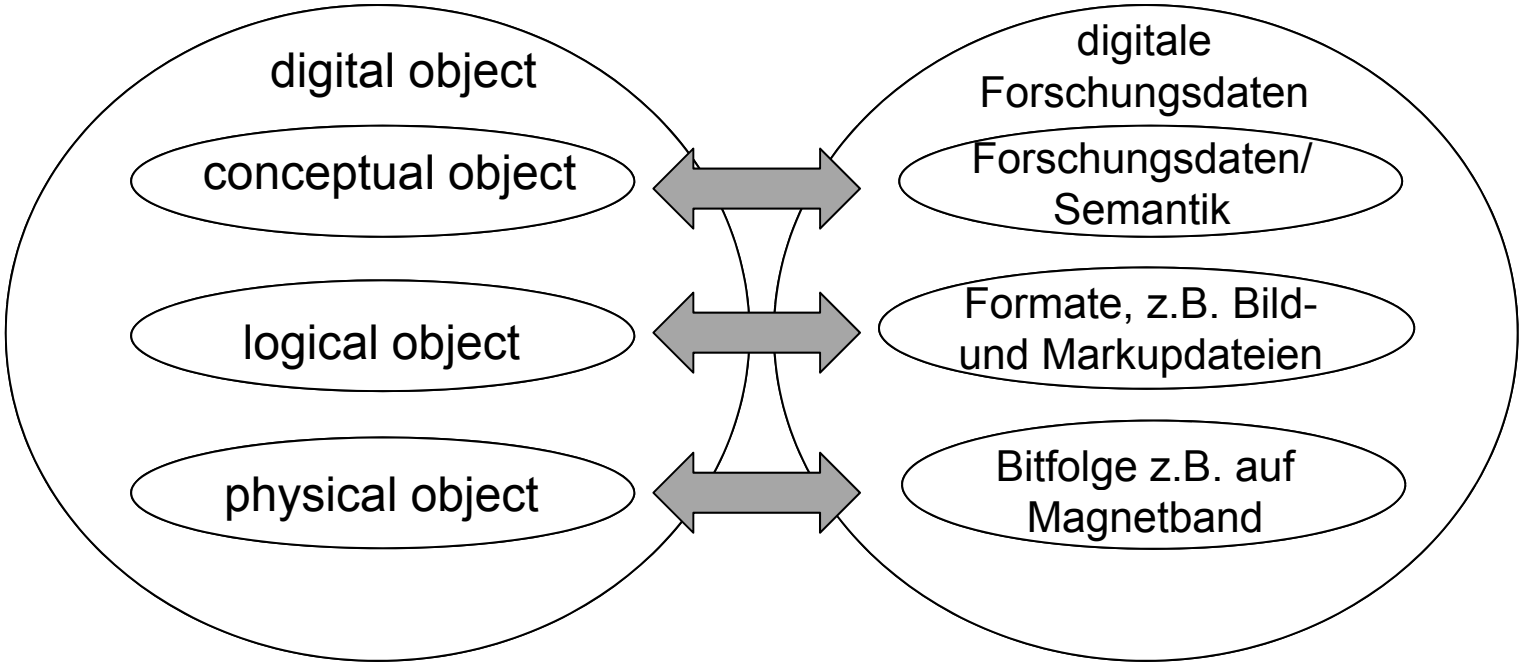
1) The Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2010): Sustainable economics for a digital planet: ensuring long-term access to digital information. Final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 116, The Blue Ribbon Task Force on Sustainable Digital Preservation and Access, February 2010, Final Report, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf, S. 10.

1. Was ist Langzeitarchiverung?

- Realisiert Anforderungen an die gute wissenschaftliche Praxis[1]
  - Aufbewahrung wissenschaftlicher Primärdaten >= 10 Jahre
  - Bislang nicht standardisiert implementiert
- Langzeitarchive für Forschungsdaten sollen in Deutschland aufgebaut werden (DFG mit aktueller Ausschreibung)
  - USA/UK haben bereits Anstrengungen in dieser Hinsicht unternommen
  - Der Erfolg hängt stark von der Berücksichtigung der Nutzeranforderungen ab → Anreize schaffen
  - Der Erfolg setzt ebenso klare Richtlinien für die Forschung voraus
  - Sonst → „Empty Archives“[2]

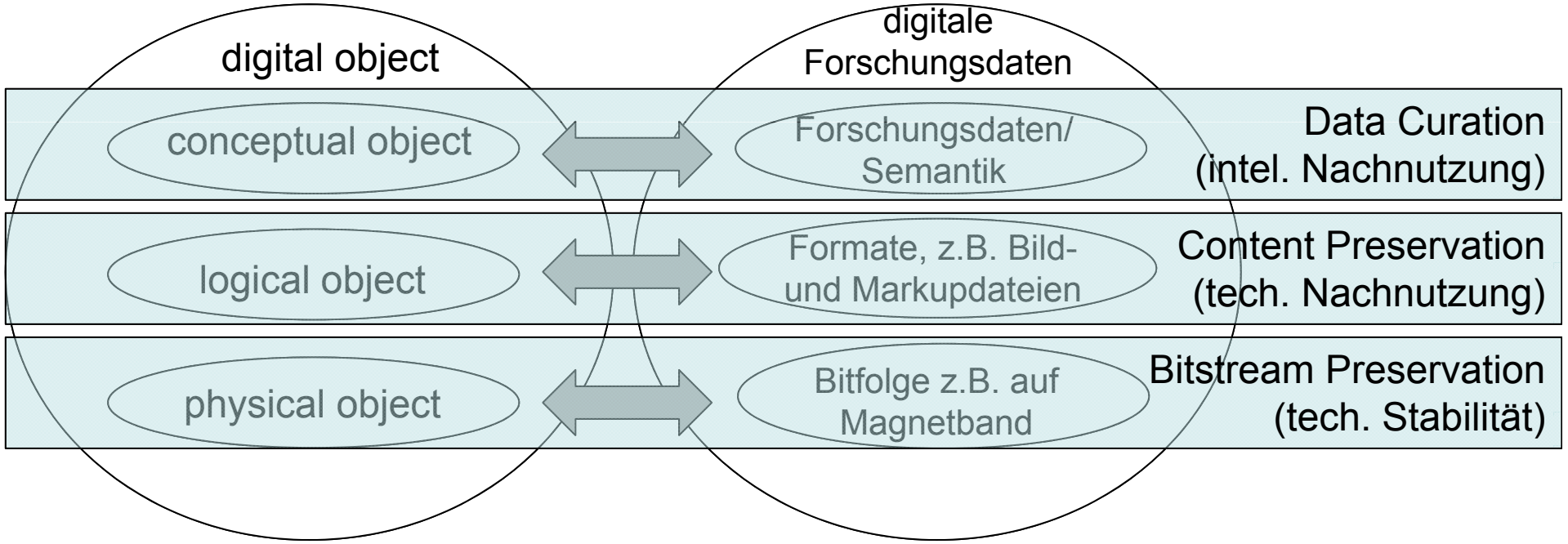
1) Deutsche Forschungsgemeinschaft (1998): Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“, Denkschrift, WILEY-VCH, Weinheim.
2) Nelson, B. (2009): Data sharing: Empty archives, Nature, 461 [7261], pp. 160-163.

2. Ebenen der LZA



Grafik angelehnt an Thibodeau: Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, 2002. http://www.clir.org/pubs/reports/pub107/thibodeau.html

2. Ebenen der LZA



Grafik übernommen von: Jens Ludwig et al.: Generische Langzeitarchivierungsarchitektur für D-Grid, 2010. http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.1-LZA-Architektur-v1.1.pdf

- Data Curation: intellektuelle Nachnutzbarkeit
  - Kontextinformationen, Objektmodelle, Versionierungen, ...
- Content Preservation: technische Nachnutzbarkeit
  - technische Qualitätskontrollen, Konvertierungen, ...
- Bitstream Preservation: technische Stabilität
  - genug unabhängige Kopien, Integritätsprüfung, ...

3. LZA-Projekte

Eine Übersicht

3. a) nestor

- Network of Expertise in long-term Storage and availability of digital Resources in Germany
- Kompetenznetzwerk in Deutschland für Langzeitarchivierung digitaler Inhalte → Plattform mit europaweiten Kooperationen
- Schwerpunkt: LZA allgemein
- Förderung
  - BMBF
  - 2003 bis 2009
  - Seit 2009 weitergeführt durch beteiligte und weitere Institutionen
- Zielsetzungen
  - **Standardisierung** → Persistent Identifier, Datei-Formate, Metadaten, Referenzmodelle (z.B. OAIS)
  - **Qualifizierung** → Kooperation mit Hochschulen für Weiterbildung und Publikationen
  - **Vernetzung** → Kooperation von Arbeitsgruppen und Projekten
- http://www.langzeitarchivierung.de

3. a) nestor

- Partner
  - Bayerische Staatsbibliothek
  - Deutsche Nationalbibliothek
  - FernUniversität Hagen
  - Niedersächsische Staats- und Universitätsbibliothek Göttingen
  - Humboldt-Universität zu Berlin
  - Landesarchiv Baden-Württemberg
  - Institut für Deutsche Sprache
- Arbeitsgruppen
  - Media (nicht-textuelle Medien)
  - Recht (rechtliche Rahmenbedingungen)
  - Digitale Bestandserhaltung
  - Vertrauenswürdige Archive – Zertifizierung
  - Langzeitarchivierungsstandards
  - Grid / eScience und Langzeitarchivierung

3. b) Kopal

- Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen
- Schwerpunkt: LZA allgemein
- Förderung
  - BMBF
  - 2004 bis 2007
  - Bis dato größtes Verbundprojekt der digitalen Langzeitarchivierung
- Zielsetzungen
  - Aufbau eines **Archivsystem**s zur Sicherung der Langzeitverfügbarkeit digitaler Dokumente → **organisatorisch** & **technisch**
  - Physischen **Erhalt** und **Interpretierbarkeit** von Daten sicherstellen
- http://kopal.langzeitarchivierung.de

3. b) Kopal

- Partner
  - Deutsche Nationalbibliothek (Gesamtprojektleitung)
  - Niedersächsische Staats- und Universitätsbibliothek Göttingen
  - IBM Deutschland GmbH
  - Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
- Implementierung
  - Berücksichtigung internationaler Standards → OAIS
  - Ausgerichtet auf unterschiedliche Bedürfnisse
  - Basiert auf Arbeiten von IBM und der Königlichen Bibliothek der Niederlande → Digital Information Archiving System – DIAS
    - IBM Content Manager
    - IBM Tivoli
    - Implementierung von OAIS
  - Eigene Open Source-Entwicklung “Softwarebibliothek koLibRI - kopal Library for Retrieval and Ingest“ → betrieben von GWDG

3. c) KoLaWiss

- Kooperative Langzeitarchivierung für Wissenschaftsstandorte
- Schwerpunkt: LZA allgemein / Biomedizinischer Kontext eingebracht
- Förderung
  - DFG
  - 2008 bis 2009
- Zielsetzungen
  - Entwickeln eines **Organisations- und Geschäftsmodells** für eine kooperative Langzeitarchivierung → am Beispiel Göttingen
  - Erarbeiten von bundesweiten **Förderempfehlungen** für eine konkrete Umsetzung der Organisations- und Geschäftsmodelle von LZA-Knoten
- Arbeitspakete
  - Technik
  - Recht
  - Kosten
  - Organisation und Fördermaßnahmen
- http://kolawiss.uni-goettingen.de

3. c) KoLaWiss

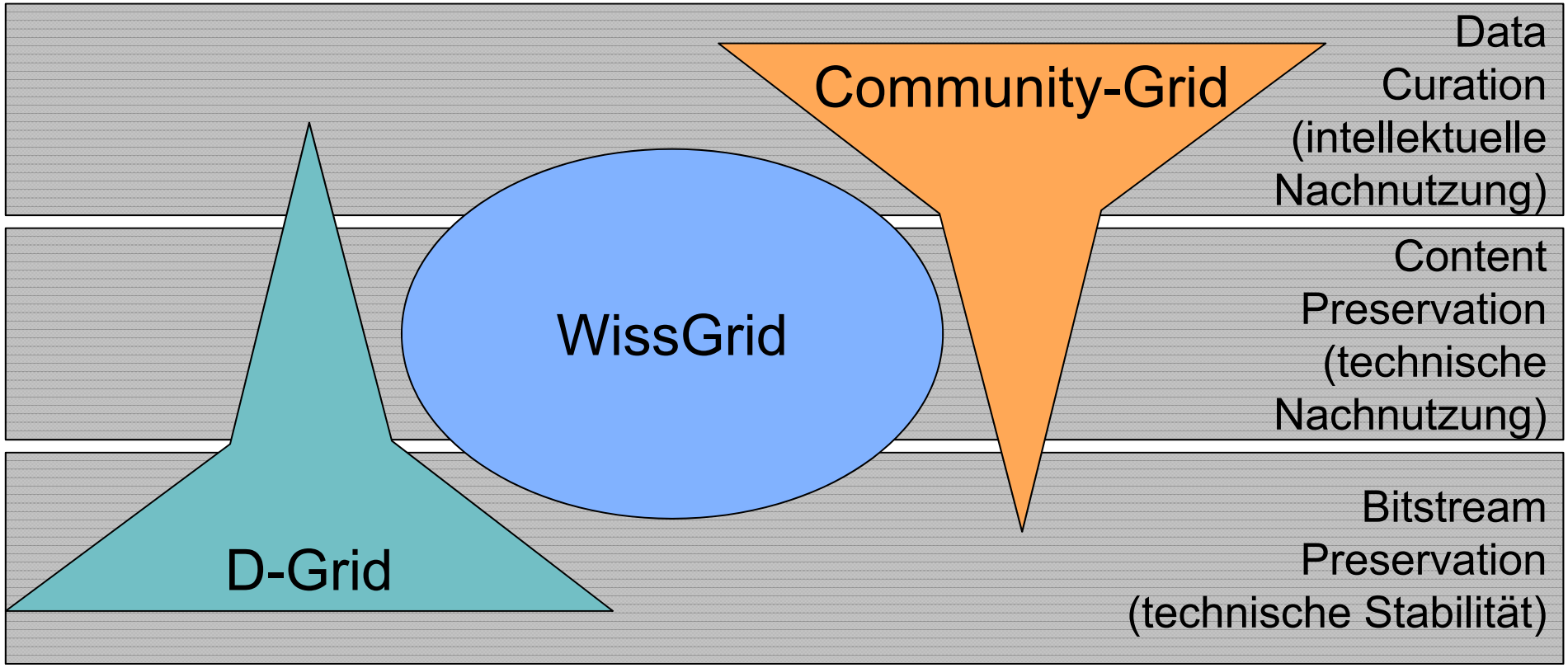
- Partner
  - Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
  - Niedersächsische Staats- und Universitätsbibliothek Göttingen
  - Universitätsmedizin der Georg-August-Universität Göttingen Geschäftsbereich Informationstechnologie
  - Medizinische Informatik, Universität Göttingen
- Spezifische Inhalte „Biomedizinische Forschungsdaten“
  - Erste Analyse von Anforderungen an LZA durch die biomedizinische Forschung
  - Behandlung von Fragestellungen zu Anforderungen der biomedizinischen Forschung in einem Rechtsgutachten
  - Forschungsprimärdaten sind nicht durch Urheberrecht geschützt → „gewisse geistige Schöpfungshöhe“ notwendig
    - CSV-Daten (z.B. reine Messwerte) sind nicht geschützt
    - Durch Anwenden wissenschaftlicher Verfahren modifizierte Datenbestände sind geschützt
    - Methodisch abgeleitete Erkenntnisse sind geschützt

3. d) WissGrid

- Grid für die Wissenschaft
- Drei Hauptarbeitspakete
  - AP1 – Betriebsmodell
  - AP2 – Blaupausen
  - AP3 – Langzeitarchivierung
- Schwerpunkt: LZA allgemein
- Zielsetzungen der Langzeitarchivierung
  - Integration zwischen Grid und LZA herstellen
  - **Grid als Basistechnologie** für LZA verwenden
  - Entwickeln **modularer LZA-Dienste**
  - Bereitstellen von Dokumentation zur Unterstützung von Nutzern
  - Entwickeln einer **generischen Grid-LZA-Architektur**, die durch wissenschaftliche Communities adaptiert werden kann
- http://www.wissgrid.de

3. d) WissGrid

- Ebenen der Langzeitarchivierung



Grafik übernommen von: Jens Ludwig et al.: Generische Langzeitarchivierungsarchitektur für D-Grid, 2010.
http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.1-LZA-Architektur-v1.1.pdf

3. d) WissGrid

- Partner (ohne Unteraufträge)
  - Universität Göttingen
  - Astrophysikalisches Institut Potsdam
  - Alfred-Wegener-Institut Bremerhaven
  - Deutsches Elektronen Synchrotron
  - Deutsches Klimarechenzentrum GmbH
  - Konrad-Zuse-Zentrum für Informationstechnik Berlin
  - Niedersächsische Staats- und Universitätsbibliothek
  - Technische Universität Dortmund
  - Universitätsmedizin Göttingen
  - Universität Heidelberg
  - Universität Trier
  - Universität Wuppertal

3. e) SHAMAN

- Sustaining Heritage Access through Multivalent Archiving
- Schwerpunkt: LZA allgemein
- Förderung
  - Europäische Kommission im Rahmen des 7. Rahmenprogramms
  - 2008 bis 2011
- Zielsetzungen
  - Entwicklung konzeptioneller und technischer Grundlagen für die neue Generation **vernetzter Langzeitarchivierungssysteme**
  - **Analyse** bestehender Systeme und institutioneller Ansätze, Technologien und Archivierungsprozesse
  - Fokus auf
    - wissenschaftliche Publikationen in Bibliotheken
    - Dokumente in behördlichen Sammlungen
    - digitale Objekte aus industriellem Design und Produktionstechnik
    - Datenquellen aus e-Science Anwendungen
- http://shaman-ip.eu/shaman

3. e) SHAMAN

- Partner, u.a.
  - Deutsche Nationalbibliothek
  - FernUniversität Hagen
  - Niedersächsische Staats- und Universitätsbibliothek
  - Philips
  - University of Glasgow
  - University of Illinois
  - University of Liverpool
  - Universität Magdeburg
  - Xerox
- Langzeitarchivierungsrahmenkonzept
  - Auf Grundlage des OAIS
  - Grid-Technologie als Basis für eine verteilte Archivierungsinfrastruktur
  - Protoypische Evaluierung in Testumgebungen und Praxisszenarien

3. f) e-Helvetica

- Schwerpunkt: LZA für Bibliotheken
- Zeitliche Einordnung
  - Start in 2001
  - Ende 2008 wurde die strategische Planung für 2009-2015 festgelegt
- Zielsetzung
  - Im Kontext des gesetzlichen Auftrags der Schweizerischen Nationalbibliothek, gedruckte oder auf anderen Informationsträgern gespeicherte Informationen, die einen Bezug zur Schweiz haben
    - sammeln
    - erschließen
    - erhalten
    - vermitteln
  - Sammlung und Archivierung von **elektronischen Publikationen**
- http://www.e-helvetica.admin.ch

3. f) e-Helvetica

- Partner
  - Schweizerische Nationalbibliothek
  - Hochschul- und Universitätsbibliotheken
  - Kantonsbibliotheken und weitere Partnerinstitutionen
  - Verlage: momentan Karger-Verlag
  - Schweizerisches Bundesarchiv

3. g) eArchivierung (TMF)

- Schwerpunkt: Elektronische Archivierung von klinischen Studien
- Förderung
  - 2006 bis 2007
  - Im Rahmen der TMF-Förderung
- Zielsetzung
  - Eruieren der Bedarfslage und **Anforderungen** an **Prozesse**
  - Bestimmen der **rechtlichen Rahmenbedingungen** für die elektronische Aufbewahrung
  - Untersuchen von Möglichkeiten für **Archivformate** für die revisionssichere Archivierung (XML, CDISC)
  - Analyse von Archivierungsszenarien nach **betriebswirtschaftlichen Gesichtspunkten**
  - Anschluss an übergreifende Langzeitarchivierungsprojekte herstellen
- http://www.tmf-ev.de/Themen/Projekte/V042_01_eArchivierung.aspx

3. g) eArchivierung (TMF)

- Partner
  - KKS Düsseldorf, Halle, Heidelberg, Köln, Leipzig, Münster
  - TMF e.V.
  - Universitätsklinikum Freiburg, Zentrum Klinische Studien
  - Universität Göttingen, Medizinische Informatik
- Arbeitspakete
  - AP0: Arbeitsprozesse der Archivierung
  - AP1: Rechtliche Rahmenbedingungen
  - AP2: XML als neues Archivformat
  - AP3: CDISC-Format für die Archivierung
  - AP4: Konventionelle Formate
  - AP5: Wirtschaftlichkeitsanalyse
  - AP6: Beurteilung und Handlungsempfehlungen

3. h) Medical Archive Grid

- Gemeinsamer Förderantrag auf Beschaffung eines Speicher-systems für Langzeitarchivierung und Backup
- Schwerpunkt: LZA für klinischen Betrieb / biomedizinische Forschung
- Förderung
  - Fachkonzept → DFG
  - Hardware → Art. 143c GG „Länderförderung“
  - Beantragt Ende 2009
- Zielsetzung
  - Einrichtung eines **gemeinsamen virtuellen Speicherpools** für die beteiligten **zwei Kliniken → Data Center** basierend auf **Virtualisierung**
  - Auslagerung der „Archiv Disaster Recovery Zweitkopie“ auf dem entfernten Standort → Erhöhung der **Hochverfügbarkeit** / Reduzierung des **Platzbedarfes** der Kliniken für die Langzeitarchivierung
  - „**Speicher on-demand**“ soll auch kleineren Kliniken angeboten werden können

3. h) Medical Archive Grid

- Partner
  - Technische Universität München (TU), Klinikum rechts der Isar
  - Ludwig-Maximilians-Universität, Klinikum der Universität München
- Inhalte
  - Anforderungen an die Speicher Infrastruktur
  - Performance Anforderungen
  - Anforderungen Backup
  - Anforderungen Archivierung
  - Security Anforderungen
  - Management Anforderungen
- Vision
  - Alle bayerischen Universitätskliniken bezüglich LZA miteinander verbinden

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

UNIVERSITÄTSMEDIZIN
GÖTTINGEN UMG

Kontakt

Universitätsmedizin Göttingen

Medizinische Informatik

http://www.mi.med.uni-goettingen.de/

Frank Dickmann

Computational Medicine und Grid-Computing

fdickmann@med.uni-goettingen.de

Tel.: (0551) 39 - 14355