Langzeitarchivierung von Forschungsdaten: Konzepte und Infrastruktur

11. Oracle Bibliotheken Summit

28. Okt. 2010, Weimar

Jens Ludwig
ludwig@sub.uni-goettingen.de
SUB Göttingen

SPONSORED BY THE
Federal Ministry of Education and Research

Überblick

- WissGrid und die Fachdisziplinen
- grundlegende Konzepte
- Umsetzungsarbeiten

Überblick

- WissGrid und die Fachdisziplinen
- grundlegende Konzepte
- Umsetzungsarbeiten

Hintergrund von WissGrid

- WissGrid ist Teil der D-Grid-Initiative

- D-Grid: Infrastruktur zur verteilten und gemeinsamen Nutzung von Rechen- und Speicherressourcen
  - seit 2005
  - ca 35 Projekte
  - Wissenschaft, Industrie, Rechenzentren, ...
  - Fachdisziplinen haben dort virtuelle Forschungsumgebungen aufgebaut

- Auftrag von WissGrid: Die Interessen der wissenschaftlichen Nutzer vertreten

Ziele von WissGrid

Für die Fachdisziplinen die

- organisatorische und
- technische

Nachhaltigkeit im D-Grid sichern.

Drei Arbeitsbereiche:

- Betriebsmodell für akademische Grid-Nutzer
- Leitfäden (Blaupausen) und Beratung für neue Grid-Nutzer und Interessierte
- Langzeitarchivierung von Forschungsdaten

Die WissGrid-Partner

Repräsentieren die bisherigen fünf wiss. Grid-Projekte:

- HEP-Grid: Hochenergiephysik
- TextGrid: Geisteswissenschaften
- C3-Grid: Klimawissenschaften
- Medi-Grid: Medizin
- AstroGrid-D: Astronomie

Aber die interessierten, neuen Communities spielen eine wichtige Rolle:

- Sozialwissenschaften
- Photon Sciences
- ...

Wieso „Langzeitarchivierung im Grid“?

- Langzeitarchivierung ist keine Konservierung „toter“ Daten, sondern zielt auf Nachnutzung
- „Langzeitarchivierung“ als Überbegriff für
  - Sicherstellen der Nachnutzbarkeit
  - in einem anderen technischen, zeitlichen, fachlichen, organisatorischen oder sonstigen Kontext
- dadurch große Schnittmenge mit aktuellen Begriffen wie Data Driven Science, Data Sharing, etc.
- Nicht erst am Ende des Forschungsprozesses relevant, sondern während des ganzen Life Cycles (und damit auch im Grid)

Fragestellung

Wie kann die LZA von Forschungsdaten unterstützt werden? In einer Vielzahl von Disziplinen?

Und mit sehr unterschiedlichen Ausgangsbedingungen:

- große Repositorien und IT-Infrastruktur vorhanden vs noch nichts
- verschiedenste Daten
  - homogen vs heterogen
  - unveränderbar vs veränderbar und löschbar
  - open access vs vertraulich
- verschiedenste Organisationsformen

Beispieldisziplin: Photon Sciences

- Strahlung aus Teilchenbeschleunigern wird genutzt um Objekte zu analysieren (Moleküle, Kunstwerke, ...)
- Wissenschaftler sind nicht Teil der Institution, die die Instrumente betreibt
- hohe Konkurenz, deshalb hohe Sicherheitsanforderungen
- viele und teure Daten: von 1kB/h bis 1PB/Woche
- Archivierung: bisher Wissenschaftler verantwortlich, Datenverlust kommt regelmäßig vor

Beispieldisziplin: Germanistische Sprachwissenschaft

- Erforschung und Dokumentation der deutschen Sprache
- erstellt Textkorpora als Grundlage für Forschung
- große und heterogene Daten
  - derzeit 1,5 Mio Texte, mit detaillierten Metadaten und Annotationen, ca 5 TB
  - Audio, Video, etc. noch nicht abschätzbar
- rechtliche Situation
  - Verträge mit Urhebern
  - z.T. nachträgliches Löschen notwendig

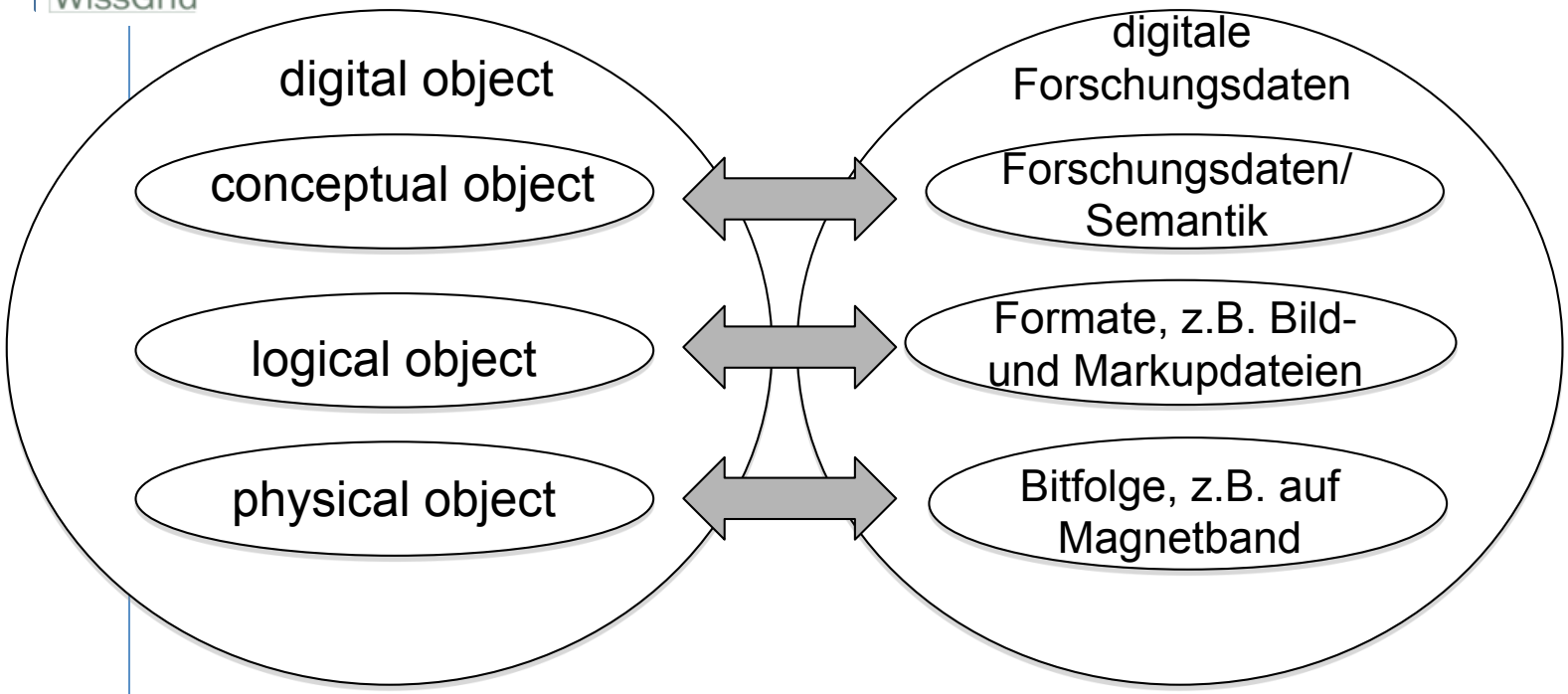
Ansatz

- Kein vollständiges System oder Archiv bauen, sondern generische Werkzeuge und Basispakete bereitstellen.
- Die Werkzeuge an die D-Grid-Umgebung anpassen und in die Forschungsumgebung der Communities integrieren.
- Jede Disziplin soll profitieren, ohne alles übernehmen zu müssen.
- Organisation und Finanzierung ausklammern, aber grundlegende LZA-Leitfäden anbieten.

Überblick

- WissGrid und die Fachdisziplinen
- grundlegende Konzepte
- Umsetzungsarbeiten

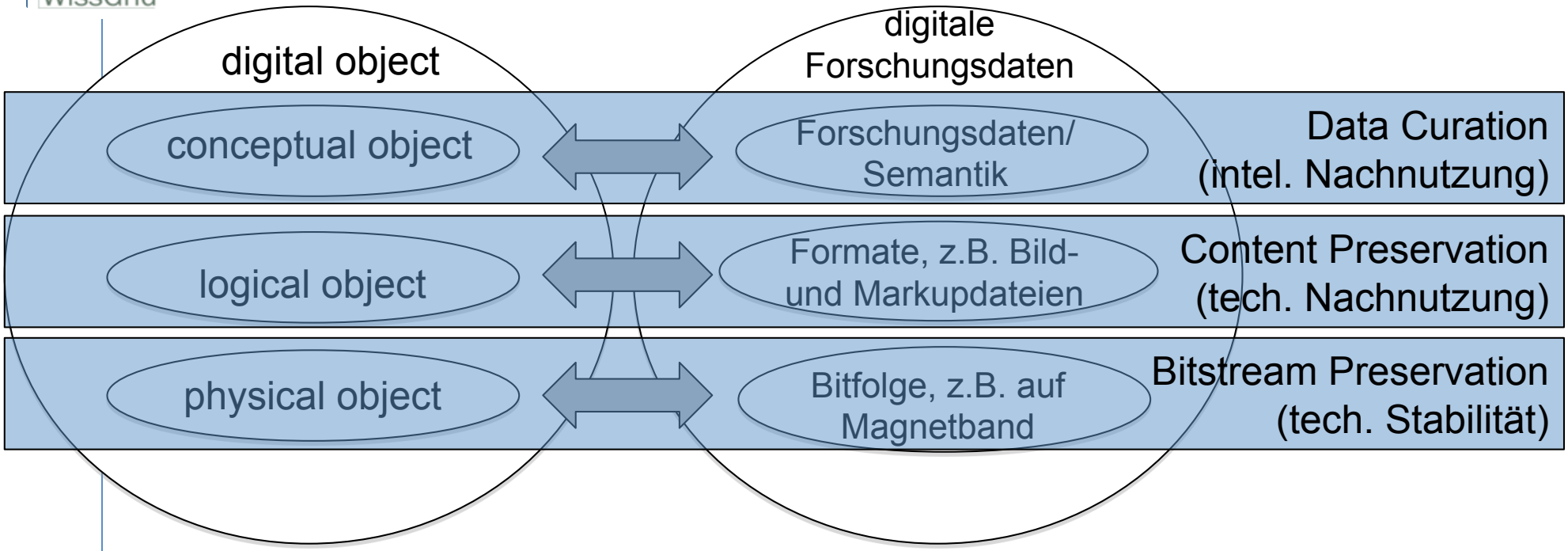
Drei Aspekte digitaler Objekte



- conceptual object: für Menschen bedeutsames Objekt
- logical object: von Software und Technik verstehbares Objekt
- physical object: Zeichen auf einem Trägermedium

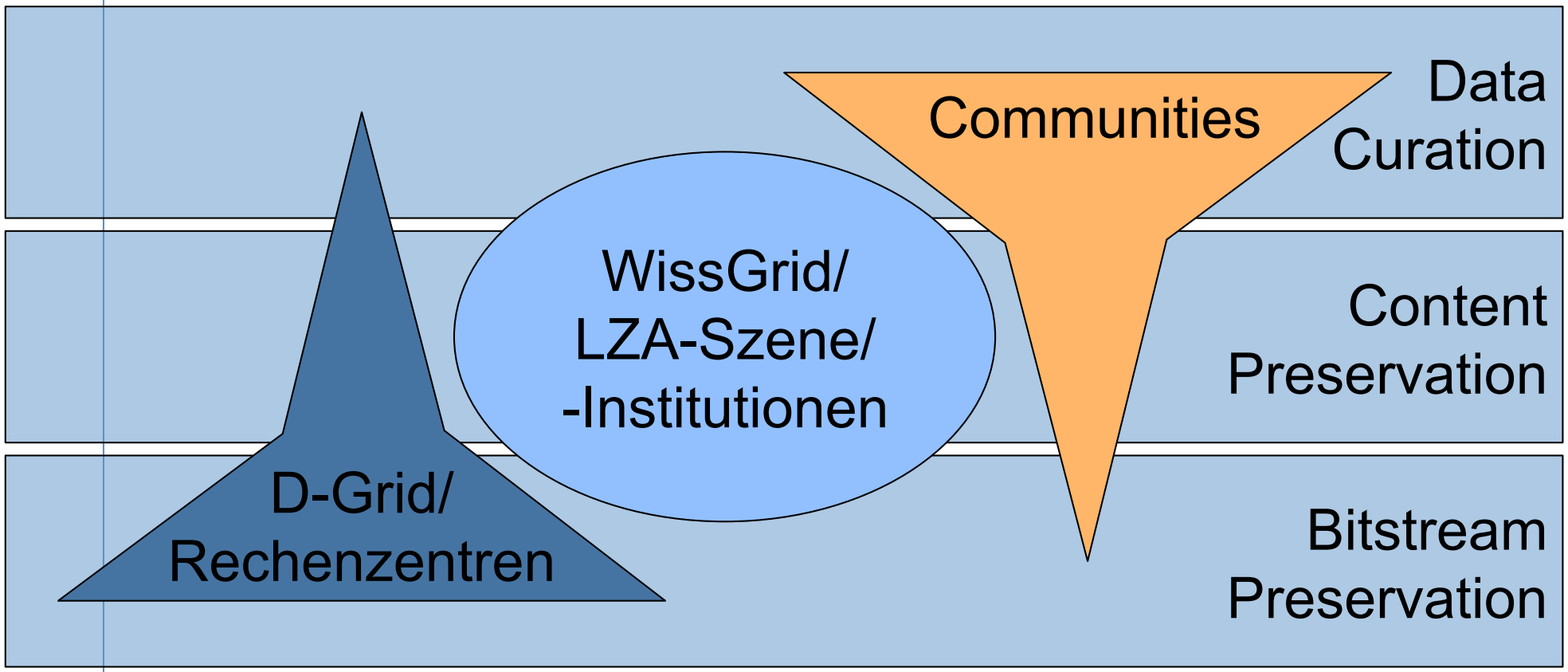
Angelehnt an Thibodeau: Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, CLIR 2002.

Drei Aspekte der Langzeitarchivierung



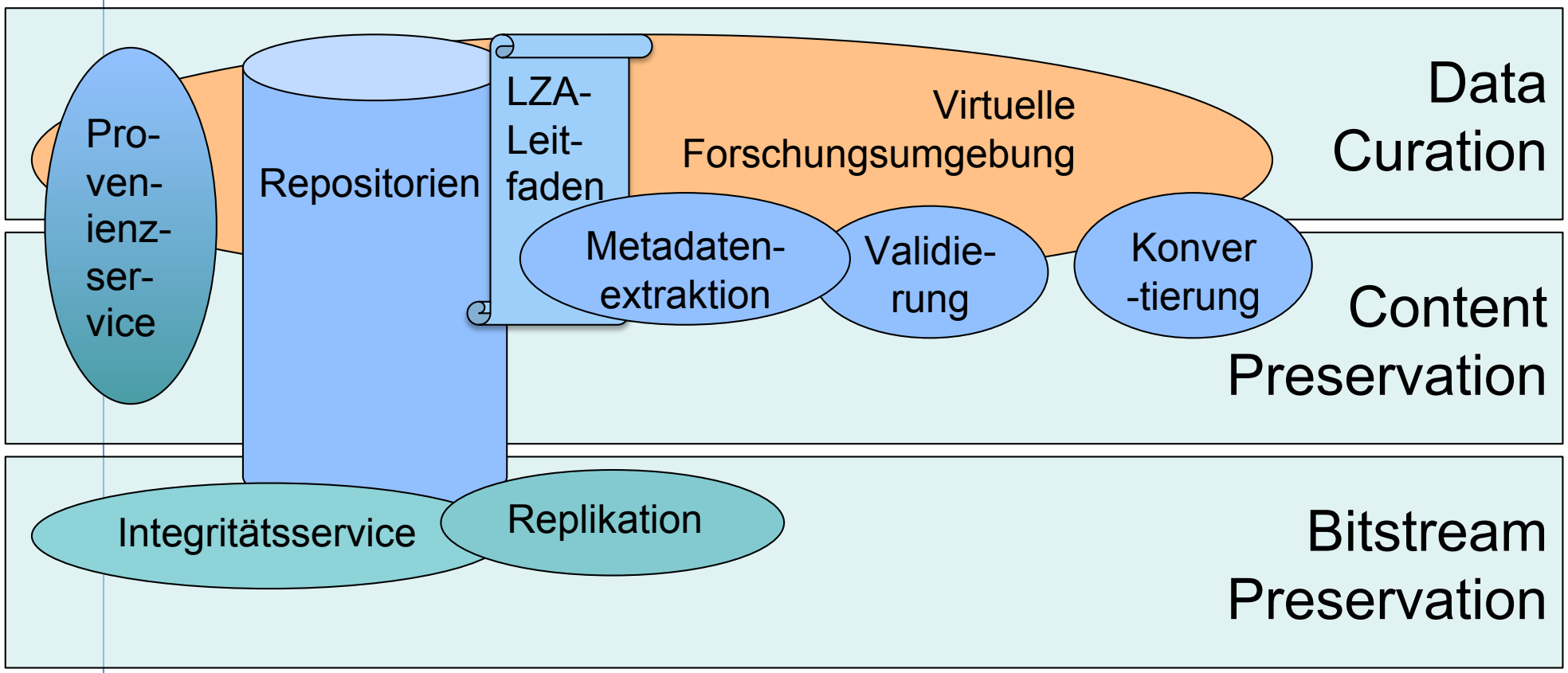
- Bitstream Preservation: technische Stabilität
  - genug unabhängige Kopien, Integritätsprüfung, ...
- Content Preservation: technische Nachnutzbarkeit
  - technische Qualitätskontrollen, Konvertierungen, ...
- Data Curation: intellektuelle Nachnutzbarkeit
  - Kontextinformationen, Objektmodelle, Versionierungen,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

Verantwortung für Entwicklung/Aufbau



- Betrieb vermutlich meist ähnlich (ohne Projekte)
- klassische bibliothekarische Tätigkeiten eher im Data Curation Bereich

Verortung zentraler Entwicklungen



WissGrid

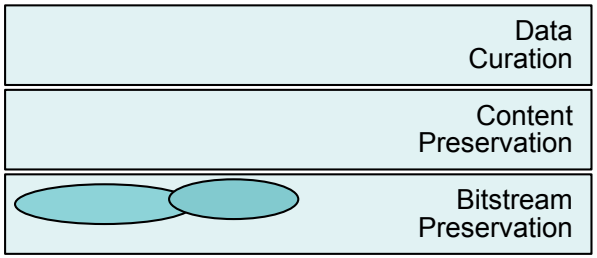
Community

D-Grid/Infrastrukturanbieter

Überblick

- WissGrid und die Fachdisziplinen
- grundlegende Konzepte
- Umsetzungsarbeiten

Bitstream Preservation

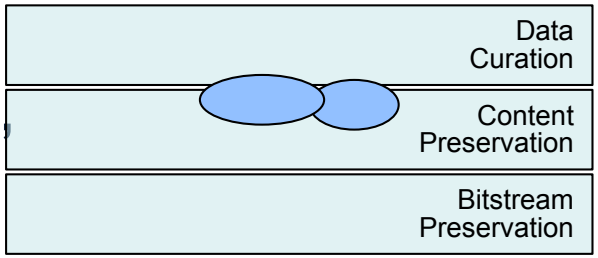


- David Rosenthal: „Bit Preservation A Solved Problem?“ (No...)

- Nicht die absolute Sicherheit der Bits ist das Problem, sondern verlässliche Angaben für rationale Kosten-Nutzen-Bewertung

- Kriterien und Service Level Agreement (elek. Verträge) für Speicher ausarbeiten zusammen mit
  - Anwendern
  - Rechenzentren
  - SLA-Experten

Charakterisierung

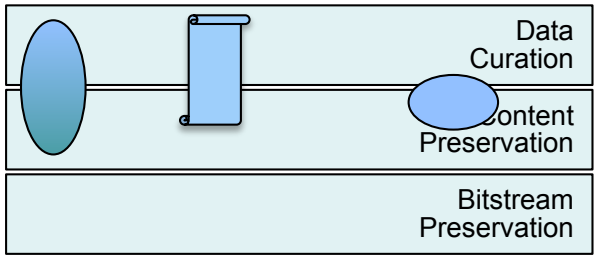


- JHOVE2: Identifizierung, Validierung, Metadatenextraktion, ...

- um wissenschaftliche Dateiformate erweitern
  - NetCDF, Klimaforschung (fertig)
  - ROOT, Teilchenphysik
  - FITS, Astronomie
  - ggf. weitere

- parallele, skalierbare Verarbeitung im Grid

Konvertierung, Provenienz, LZA-Leitfäden



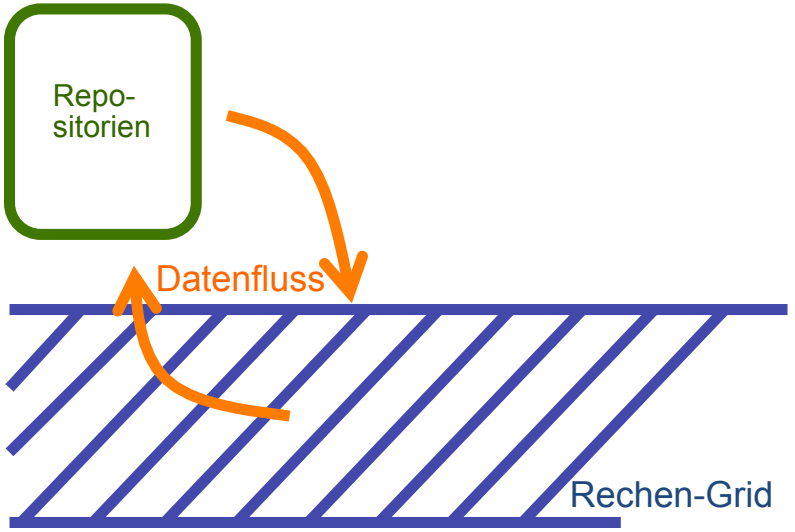
- Konvertierung: Framework um Dateiformate zu konvertieren (analog zu JHOVE)

- Provenienz: Geschichte eines Objekts dokumentieren (Wann wurde das Objekt wie verarbeitet und verändert?), Anforderung und System definieren

- LZA-Leitfäden:
  - Communities an LZA, Data Curation und organisatorische Aufgaben heranführen
  - Grid-Besonderheiten berücksichtigen

Grid-Repositorien-Typ A: Rechen-Grid (Repositorien als separates Archiv)

- Datenverarbeitung im Grid
- Speicherung im Repositorium



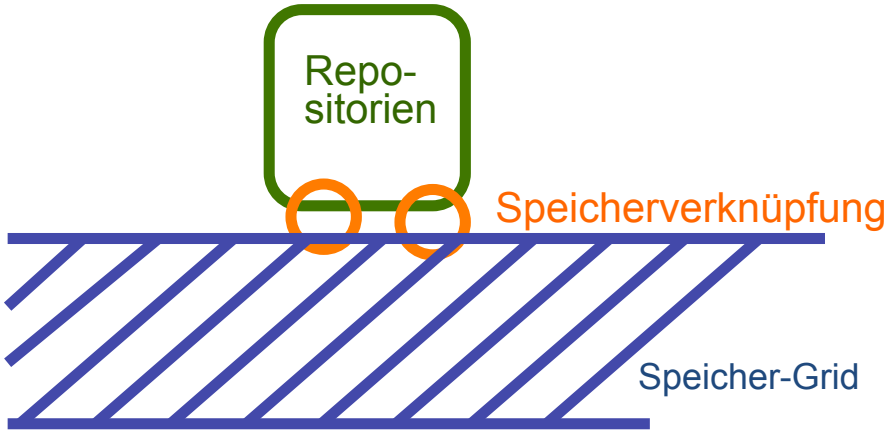
Aufgaben:

- Repositorium braucht Grid-Schnittstellen
- Abbildung der Rechte zwischen Grid und Repositorium
- Daten zu den Diensten vs Dienste zu den Daten

Grid-Repositorien-Typ B: Speicher-Grid

- Repositorien speichern direkt im Grid.

- Grid und externe Dienstleister können effizienter sein und mehr Funktionen anbieten.



Aufgaben:

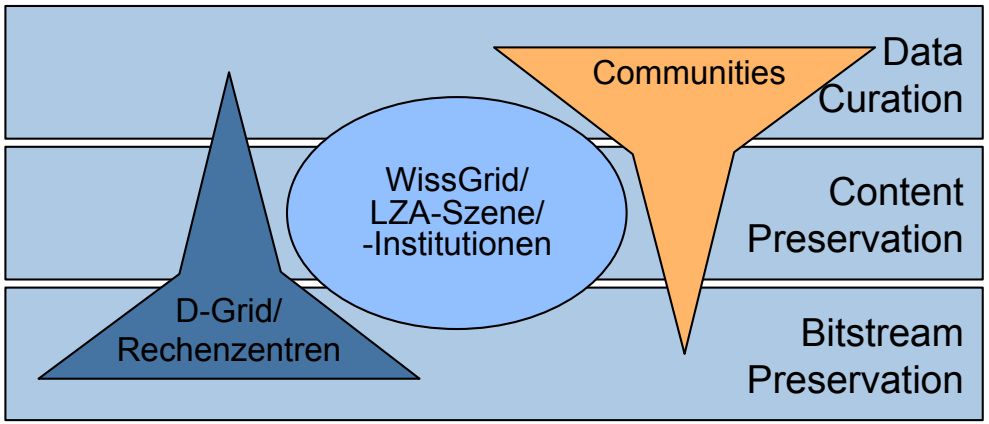
- Verknüpfung Fedora mit iRODS (erster Prototyp fertig)

- Sicherheit der Daten im Grid

- Daten auch direkt durch Grid-Mechanismen zugänglich (aber Synchronisierung erforderlich)

Zusammenfassung und Fazit

- Unterschiedliche Akteure habe unterschiedliche LZA-Aufgaben für Forschungsdaten:



- Disziplinübergreifende Zusammenarbeit ist möglich und die Bibliotheksperspektive kann etwas beitragen.

- Es kann generische Werkzeuge/Dienste geben, die aber immer Anpassungen und Integrationsaufwand erfordern.

Vielen Dank!