

REPOSITORIEN

Workshop der Arbeitsgruppe Elektronisches Publizieren der deutschen Akademien der Wissenschaften
vom 04. bis zum 06.10.2010 in der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste, Düsseldorf



UNION DER DEUTSCHEN AKADEMIEEN
DER WISSENSCHAFTEN

TextGrid - Virtuelle Forschungsumgebung, Forschungsdaten und Langzeitarchivierung (BMBF)



Dr. Heike Neuroth
Head Research & Development Department
SUB Göttingen, Germany

TEXT
GRID

ToC

- **Einführung**
- TextGridLab
- TextGridRep &
Langzeitarchivierung
- Ausblick

EU High-Level Experts Group on Scientific Data (Vision 2030 für Forschungsdaten)



- *“Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance”.*

Report wird im Oktober 2010 veröffentlicht



Definition Virtuelle Forschungsumgebung

Nach JISC (<http://www.jisc.ac.uk>)

- Zweck einer Virtuellen Forschungsumgebung (englisch: VRE) ist die größtmögliche Unterstützung von Forschungsprozessen aus allen Disziplinen während zunehmend komplexer werdenden Arbeitsabläufen
- Konzept umfasst vor allem Werkzeuge und Technologien, die von den Forschern benötigt werden, um kollaborativ, disziplinübergreifend, international und institutionell unabhängig interagieren zu können
- Der freie Zugriff auf Ressourcen (Daten, Dienste) und eine technische Infrastruktur (lokal, national) gehören ebenfalls dazu

Was ist TextGrid?



TEXT
GRID

- Ziel: Zugang und Austausch von Forschungsdaten in den Geistes- und Kulturwissenschaften mit Hilfe von Informationstechnologie (u.a. Grid-Technologie)
- TextGrid umfasst Werkzeuge, Ressourcen und Infrastrukturentwicklung. Es bietet flexible kollaborative Strukturen insbesondere für Forschungsverbünde (durch VREs)
- Ermöglicht die Zusammenarbeit in einer verteilten, sicheren, flexiblen und modularen Forschungsumgebung und die **gemeinsame** Nutzung von Werkzeugen, Daten und Methoden



Konsortium

TEXT
GRID



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

- **TextGrid Laboratory**

- Einstiegspunkt in die Virtuelle Forschungsumgebung
- stellt Werkzeuge, Dienste und Daten in einer intuitiv bedienbaren Umgebung zur Verfügung
- wird kontinuierlich weiterentwickelt

- **TextGrid Repository**

- Fachwissenschaftliches Langzeitarchiv (Grid-Infrastruktur)
- garantiert langfristige Verfügbarkeit und Zugänglichkeit der geisteswissenschaftlichen Forschungsdaten (DFG GwP)



ToC

- Einführung
- **TextGridLab**
- TextGridRep &
Langzeitarchivierung
- Ausblick

TextGridLab: Tools und Services (v1.0)

Tools im TextGridLab:



XML-Editor



Nutzer- und Projektverwaltung



Text-Bild-Link-Editor



Projektbrowser / Navigator



Wörterbuch-Recherche



Recherchetool



Workflow-Tool



Metadaten-Editor



Text Publisher Web



Aggregationen



Lemmatisierer



Upload Tool

Weitere Services:



Streaming Editor



Service-Registry



Tokenizer



Sortiertool

Tool	In aktueller Beta vorhanden, volle Funktionalität und Dokumentation zur v1.0 (2.Q 2011)
Tool	In Entwicklung, volle Funktionalität zur v1.0 (2.Q 2011)
Tool	Ergänzende Tools und Services mit eingeschränktem Support, bereits in aktueller Beta vorhanden

TextGridLab: Weitere Entwicklungen



Im Laufe der Projektlaufzeit bis spätestens Mai 2012:

- Musikwissenschaft: Noten-Editor
- Klassische Philologie: Glossen-Editor
- Kunstgeschichte: Integration Digilib
- Sprachwissenschaft: Integration LEXUS und COSMAS
- OCR für Frakturschrift: Erweiterung OCRopus

- Bibliographietool
- Kollationierer
- Text Publisher Print (DFG-Projekt)

ToC

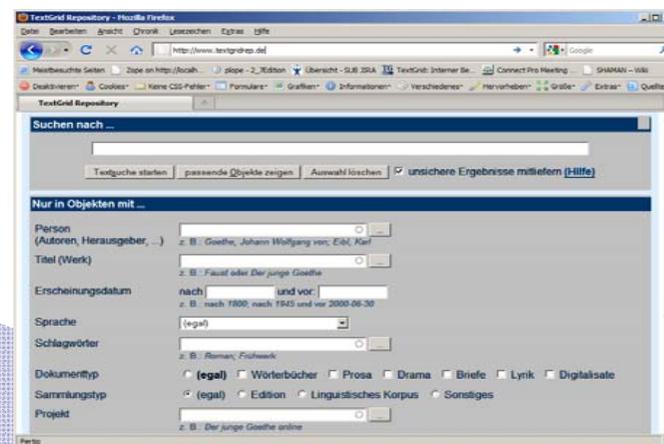
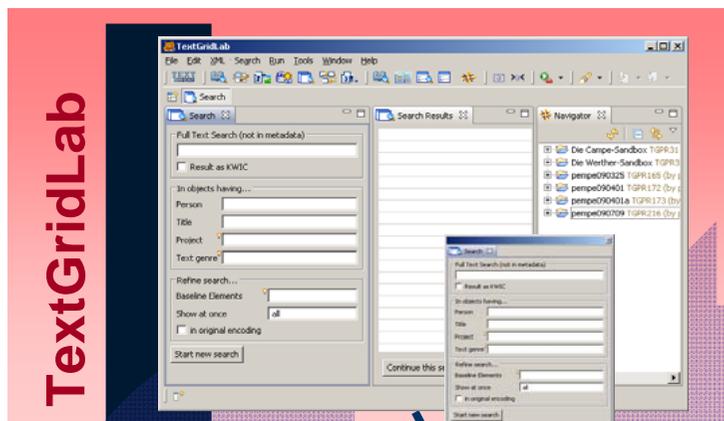
- Einführung
- TextGridLab
- **TextGridRep &
Langzeitarchivierung**
- Ausblick

TextGridRep: Gesamtübersicht



Eclipse Frontend

Portal(e) Suche + Anzeige



TextGridLab

Ingest 1

TextGridRep

Rechtemanagement

TG-auth*

Such-Index 1

isPublic – TG-publish

Such-Index 2

Ingest 2
TG-publish

- + Metadaten-Validierung /QA
- + Persistent Identifier
- + ggf. LZA-MD
- + ggf. LZA-Services

dynamisch

Grid Storage

statisch

LZA

- Offene Schnittstelle zum Such-Index 2
- Sammlungs-spezifische Portale möglich

- große Datenmengen
- individuell angepasst
- + ggf. Metadaten-Validierung
- + ggf. Persistent Identifier
- + ggf. LZA-MD
- + ggf. LZA-Services

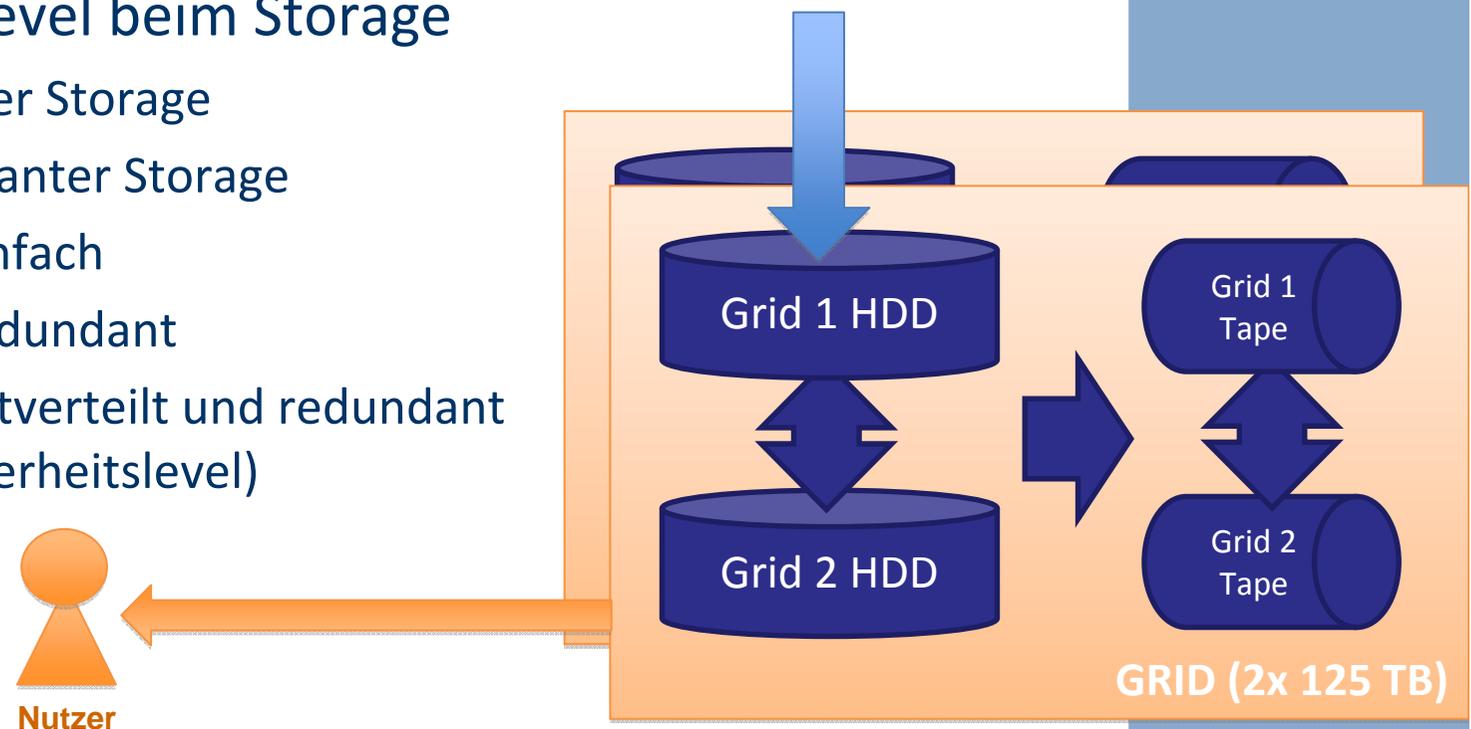
TextGridRep: Umgang mit großen Datenmengen

- Ingest großer Datenmengen („externer Content“)
 - zwei Routinen:
 - Via TextGridLab: Index1
 - Via koLibri: Index2
 - Verweisgenerierung, Auflösung interner Verweise
 - Unterstützung bestimmter Profile (z.B. DFG-Viewer METS, ...)

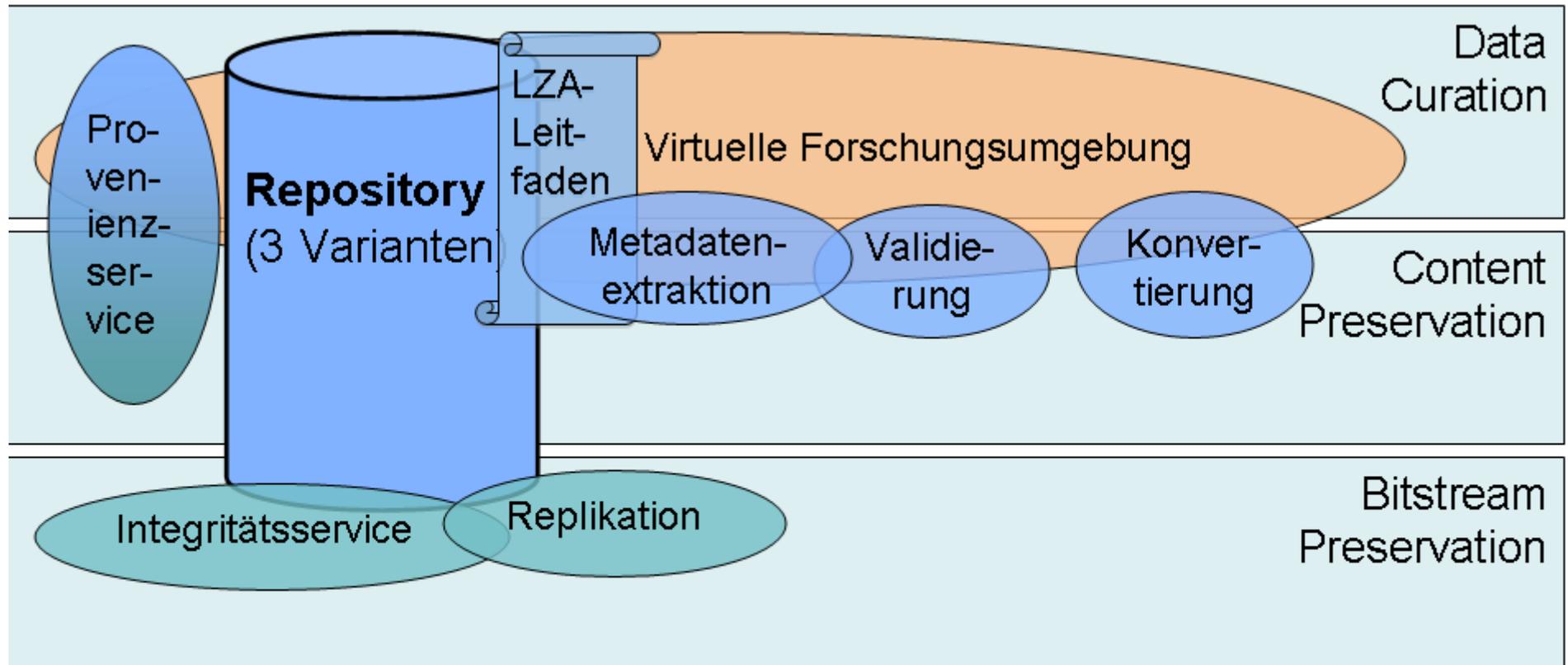
- zweiter Suchindex → „Portal“-Lösung
 - Ziel: Performante Suche für publizierte Daten
 - Umsetzung: Zweite Instanz des Suchdienstes (TG-search) ohne Verbindung zum Rechtemanagement (TG-auth*)
 - Browser-basierte Suche
 - REST-Schnittstelle für externe / individuelle Portal-Lösungen

TextGridRep: Grid-Storage

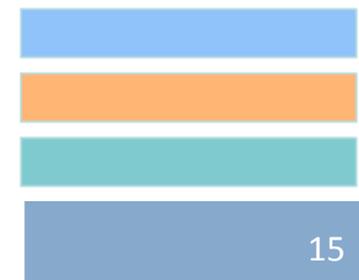
- 275 TB Grid-Storage, 250 TB Tape Storage
- zunächst mit bitstream preservation: redundanter Storage und Tape einfach (B+C)
- höherwertige LZA-Dienste später (WissGrid)
- Sicherheitslevel beim Storage
 - A) Einfacher Storage
 - B) Redundanter Storage
 - C) Tape einfach
 - D) Tape redundant
 - E) Standortverteilt und redundant (max. Sicherheitslevel)



TextGridRep: LZA-Dienste (WissGrid)



WissGrid =
Community =
D-Grid/Infrastrukturanbieter =



TextGridRep: Basisdienste

	Rechte- management	Metadaten- Validierung	Persistent Identifizier- Vergabe	LZA (-Dienste)
Ingest 1 (via TG-Lab)	+	-	-	-
Publizieren (isPublic)	- *	+	+	(+)
Ingest 2 (via koLibRi)	- *	(+)	(+)	(+)

* Freier Lesezugriff

TextGridRep: LZA-Stufen

	Bitstream Preservation für max. 10 Jahre (gem. DFG- Richtlinien)	Bitstream Preservation längerfristig	zusätzlich Höherwertige LZA-Dienste und SLAs
Redundanter Storage und Tape einfach	€	€€	€€€
Redundanter Storage und Tape redundant	€€	€€€	€€€€
zusätzlich standort-verteilt	€€€	€€€€	€€€€€

Persistent Identifier

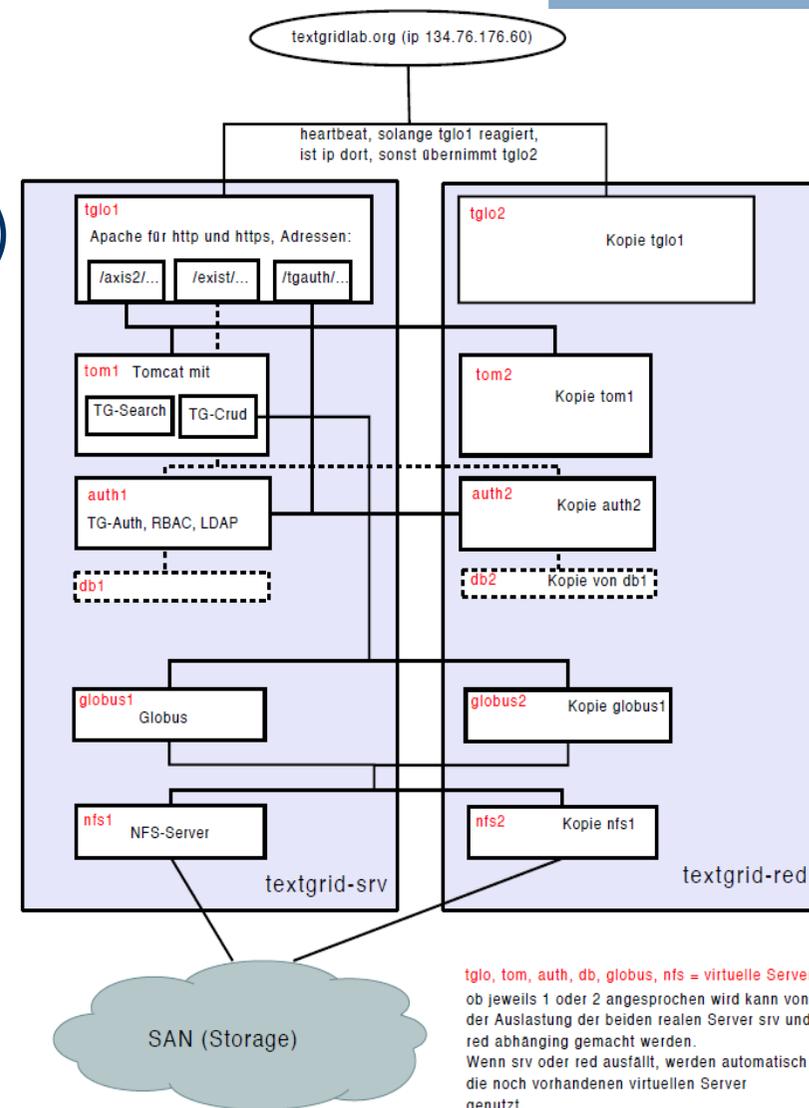
- PID-Service der GWDG:
<http://handle.gwdg.de/PIDservice/>
(in Zusammenarbeit mit bzw. für die MPG)
- GWDG ist Partner in *EPIC - European Persistent Identifier Consortium* - <http://www.pidconsortium.eu/>
- Verweise auf Teilbereiche von Objekten. Umsetzung mit an URI angehängte XPath-Ausdrücke, z.B.:
 - `textgrid:djgoethe:Faust:20070231T012345#xpath(/div[4]/div[6]/p[3])`
- Metadaten sollten auf Seiten des PID-Services auf das absolute Minimum beschränkt werden
(Synchronisationsaufwand)

Stabilität, Ausfallsicherheit

Konzept zur Ausfallsicherheit

- zweiter, redundanter Server (räumlich getrennt von Server 1)
- Performance-/Belastungs-Tests
- HA / Failover mit Heartbeat
- Backup-Konzept gemeinsam mit GWDG entwickelt
- zwei neue Fileserver
→ LZA, redundante Datenhaltung

Umsetzung zur Version 1.0
(Beginn 2011)



ToC

- Einführung
- TextGridLab
- TextGridRep &
Langzeitarchivierung
- **Ausblick**

- TextGrid Version 1.0: Frühsommer 2011
 - Release einer stabilen, einsatzfähigen Version 1.0: Beginn 2011
 - Intensivtests 1. Dezember 2010 bis 31. März 2011
 - Release-Workshop Mai/Juni 2011 in Göttingen
- Weitere Entwicklungen während der Projektlaufzeit (bis Mai 2012):
 - Höherwertige LZA-Dienste
 - Fachspezifische Tools für die Musikwissenschaft, Klassische Philologie, Kunstgeschichte und Sprachwissenschaft
 - OCR für Frakturschrift: Erweiterung OCRopus
 - Bibliographietool, Kollationierer
 - Text Publisher Print (DFG-Projekt)

Neue Rolle für wiss. Bibliotheken (und Akademien)?

- Unterstützung des gesamten Forschungsprozesses
 - bereits bei der Fragestellung bzw. Antragstellung, z.B. für Datenmanagement
 - durch die kooperative Entwicklung von Diensten (generisch, z.B. PID) und Werkzeugen (fachspezifisch, z.B. Lemmatisierer)
 - durch Unterstützung und Verantwortung im Bereich der digitalen Langzeitarchivierung inklusive Publikation auch von Forschungsdaten
 - durch den nachhaltigen Betrieb von Diensten, Werkzeugen, Daten-Repositories etc.
 - durch Beratung, Vernetzung, Schulung etc.
 - ..?



**** Vielen Dank ****

Fragen, Anmerkungen ?