

---

25. Juni 2010, Berlin

---



BBAW-Workshop  
Langfristarchivierung

**TextGrid - Vernetzte  
Forschungsumgebung in den  
eHumanities**



Heike Neuroth  
SUB Göttingen / MPDL München

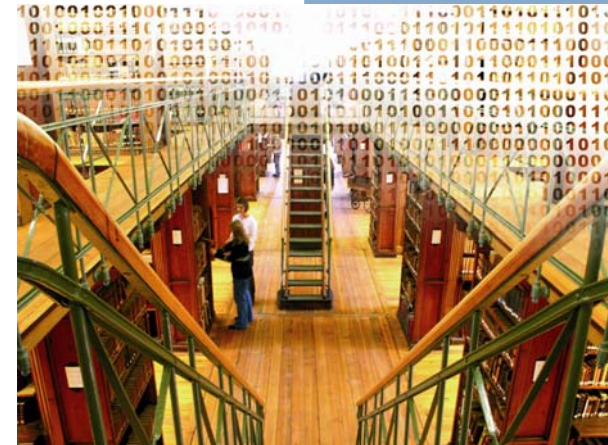
---

# Was ist TextGrid?



TEXT  
GRID

- Ziel: Zugang und Austausch von Informationen in den Geistes- und Kulturwissenschaften mit Hilfe von Informationstechnologie (Grid)
- TextGrid umfasst Werkzeuge, Ressourcen und Infrastrukturentwicklung. Es bietet flexible kollaborative Strukturen insbesondere für Forschungsverbünde
- Ermöglicht damit die Zusammenarbeit in einer verteilten, sicheren, flexiblen und modularen Forschungsumgebung und die **gemeinsame** Nutzung von Werkzeugen, Daten und Methoden
- 1. Phase: Grid Call I: 2006-09; 1,74 Mio € + Grid Storage / Knoten
- 2. Phase: Grid Call III: 2009-12; 3,18 Mio €



# Konsortium

TEXT  
GRID



GEFÖRDERT VOM



Bundesministerium  
für Bildung  
und Forschung

- **TextGrid Laboratory**

- Einstiegspunkt in die Virtuelle Forschungsumgebung
- macht vorhandene sowie neue Werkzeuge und Services in einer intuitiv bedienbaren Software verfügbar
- wird kontinuierlich weiterentwickelt



- **TextGrid Repository**

- Fachwissenschaftliches Langzeitarchiv, das in eine Grid-Infrastruktur eingebettet ist
- garantiert langfristige Verfügbarkeit und Zugänglichkeit der geisteswissenschaftlichen Forschungsdaten sowie eine optimale Vernetzung

# 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**

<b>Tool</b>	In aktueller Beta vorhanden, volle Funktionalität und Dokumentation zur v1.0 (Feb 2011)
<b>Tool</b>	In Entwicklung, volle Funktionalität zur v1.0 (Feb 2011)
<b>Tool</b>	Ergänzende Tools und Services mit eingeschränktem Support, bereits in aktueller Beta vorhanden

---

# TextGridLab: Weitere Entwicklungen

---



## Im Laufe der Projektlaufzeit bis spätestens Mai 2012:

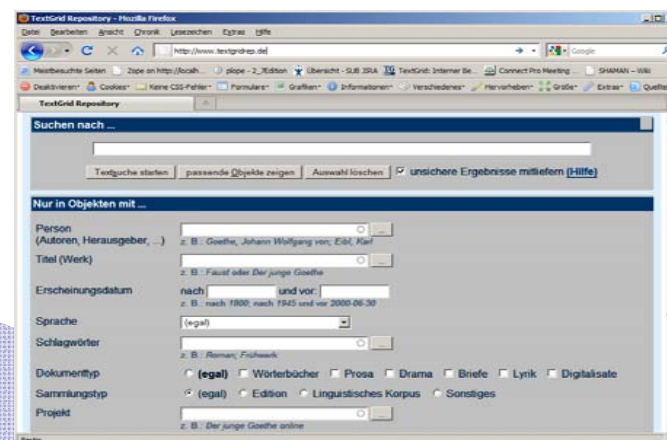
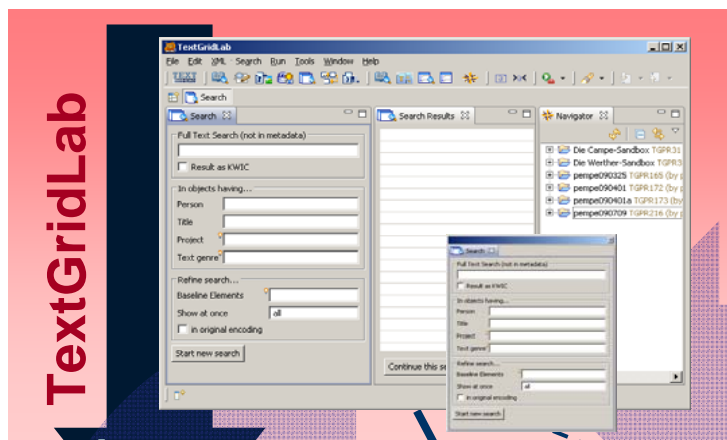
- Musikwissenschaft: Noten-Editor
- Klassische Philologie: Glossen-Editor
- Kunstgeschichte: Integration Digilib
- Sprachwissenschaft: Integration LEXUS und COSMAS
- OCR für Frakturschrift: Erweiterung OCRopus
  
- Bibliographietool
- Kollationierer
- Text Publisher Print (DFG-Projekt)

# TextGridRep: Gesamtübersicht



## Eclipse Frontend

## Portal (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

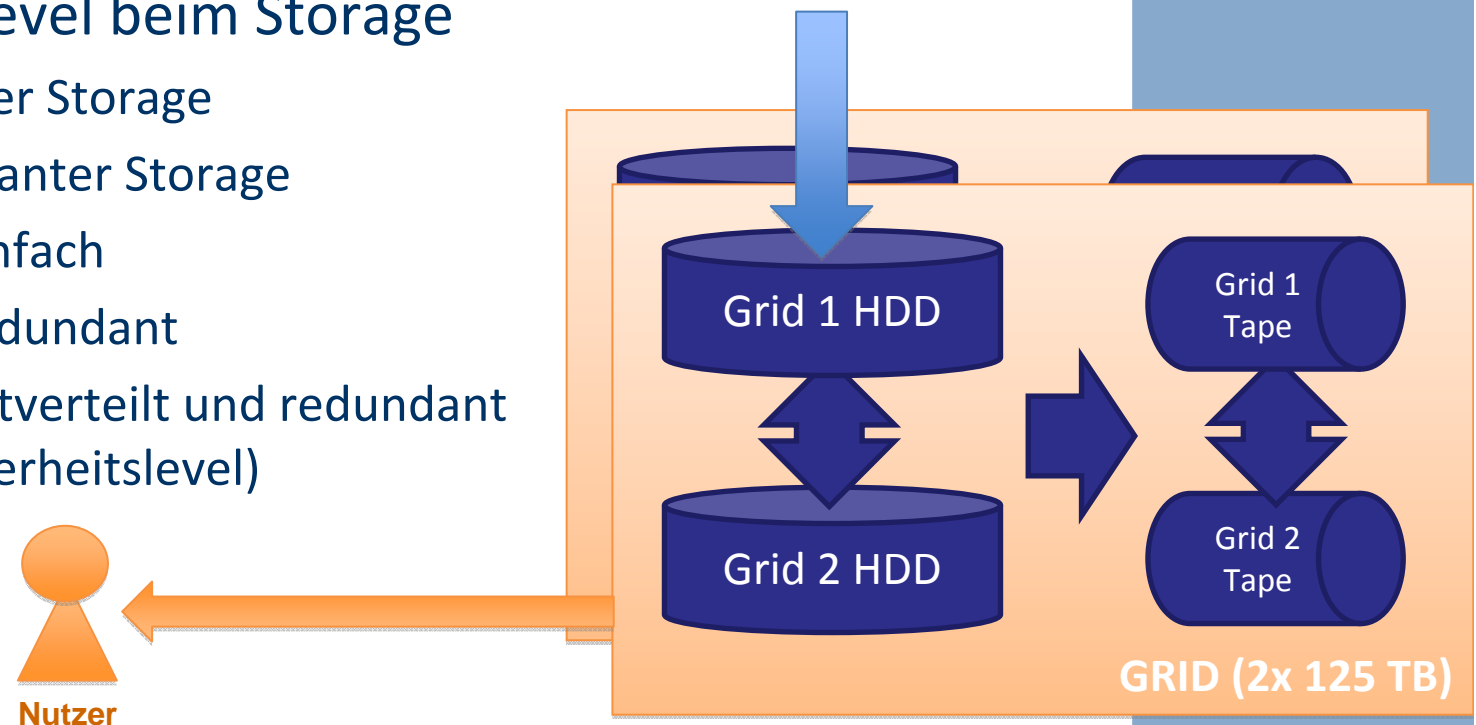
# TextGridRep: Umgang mit großen Datenmengen und -Portionen



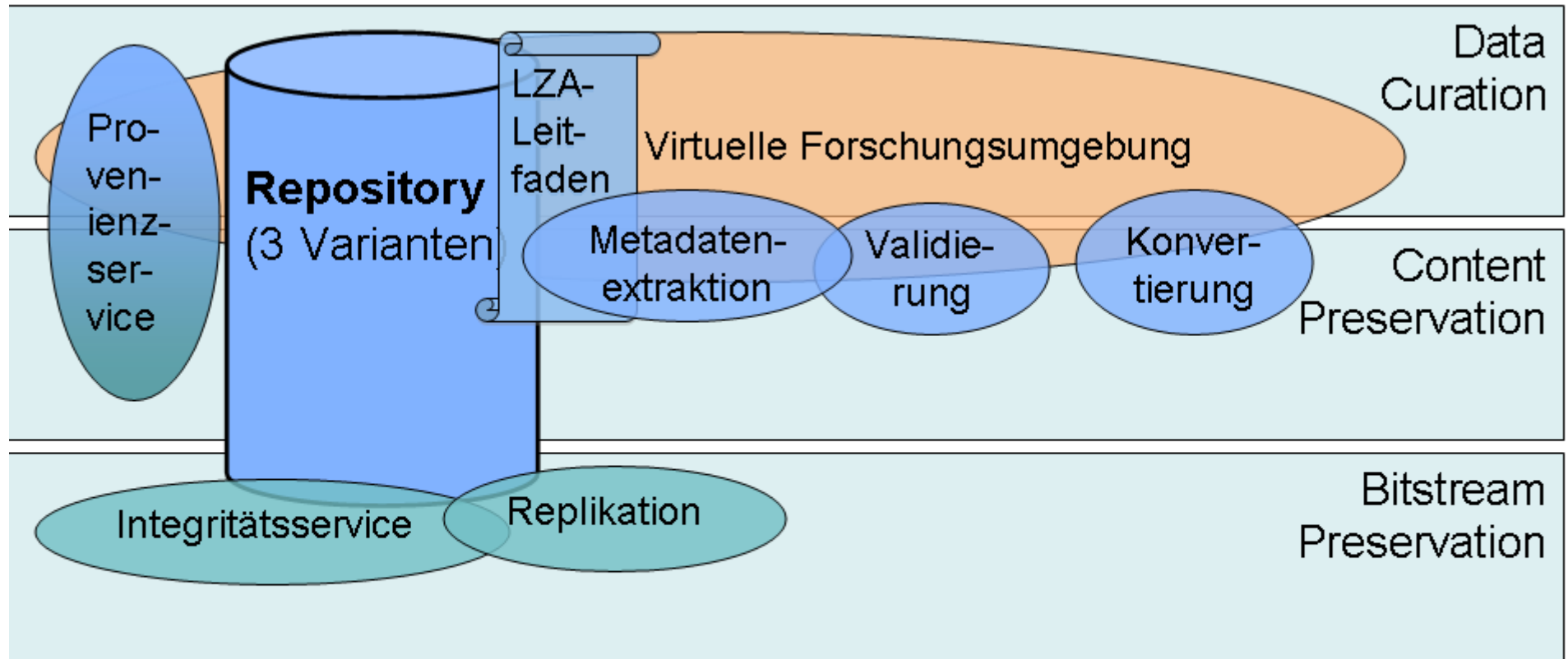
- Ingest großer Datenmengen („externer Content“)
  - zwei Routinen:
    - Via TextGridLab: Index1
    - Via koLibri: Index2
  - Verweisgenerierung, Auflösung interner Verweise
  - Unterstützung bestimmter Profile (z.B. DFG-Viewer METS, ...)
- zweiter Suchindex → „Portal“-Lösung
  - Ziel: Performante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

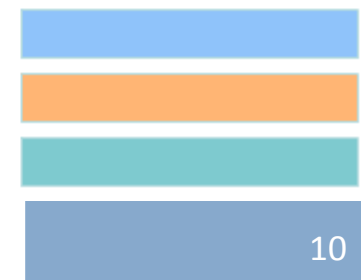
- 275 TB Grid-Storage, 250 TB Tape Storage
- zunächst mit bitstream preservation: redundanter Storage und Tape einfach (B+C)
- höherwertige LZA-Dienste später (WissGrid)
- Sicherheitslevel beim Storage
  - A) Einfacher Storage
  - B) Redundanter Storage
  - C) Tape einfach
  - D) Tape redundant
  - E)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



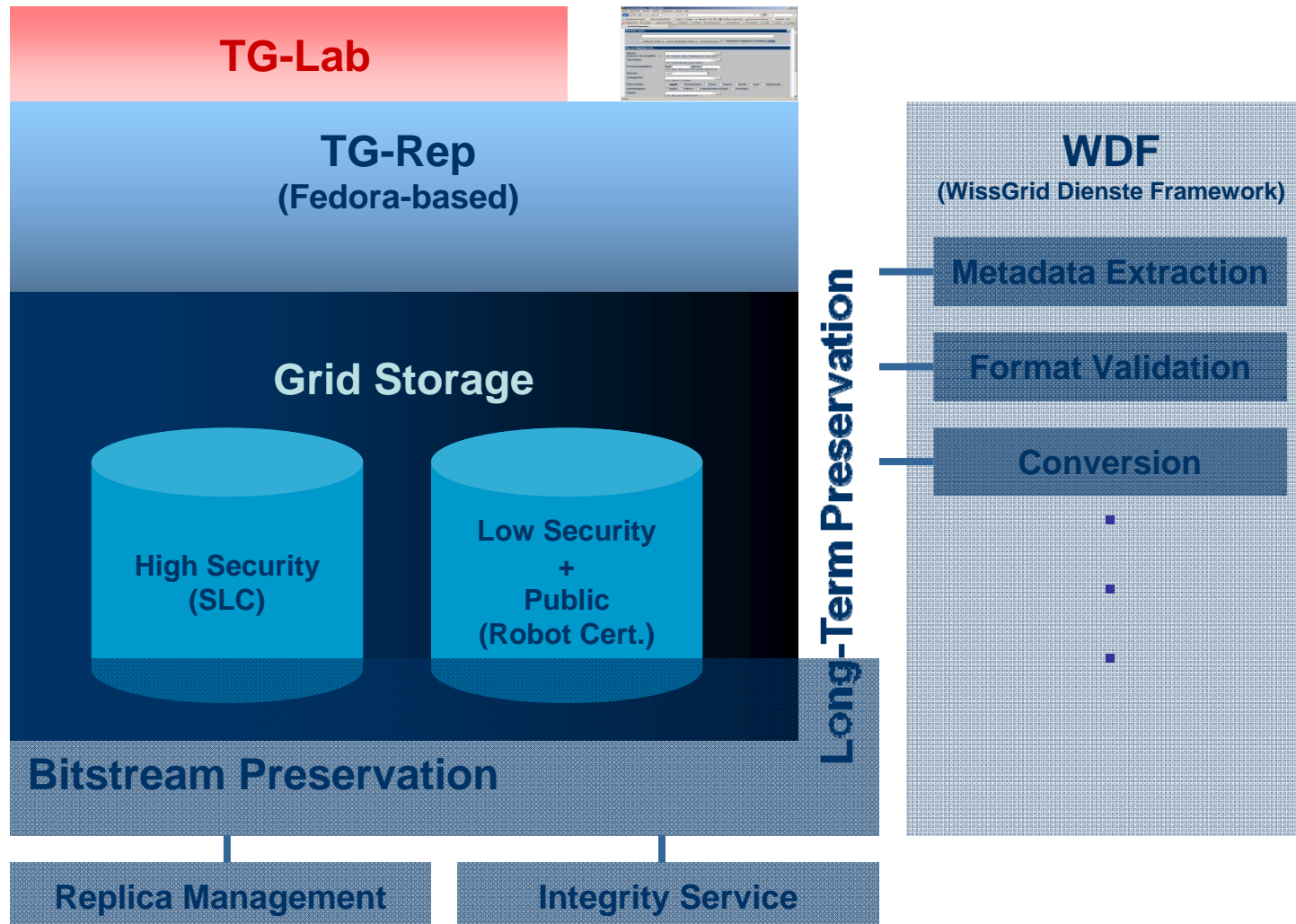
# TextGridRep: LZA-Dienste (WissGrid)



WissGrid =  
Community =  
D-Grid/Infrastrukturanbieter =



# LZA und Grid Storage (ab 2011)



# TextGridRep: Basisdienste

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

\* 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	€€€	€€€€	€€€€€

# Funktionen von Metadaten in TextGrid

- **strukturelle Annotationen** zur Auszeichnung von Inhalten (Absätzen, Seitenumbrüche usw.)
- **bibliographische Metadaten** zur Beschreibung von TextGrid-Objekten (Autor, Titel usw.)
- **Beziehungsdaten** zur Beschreibung von Beziehungen zwischen Objekten, Sammlungen etc.
- **technische Metadaten** zur Verwaltung von Benutzern und Projekten

*Umsetzung eines neuen Metadatenschemas  
zur Version 1.0 (Februar 2011)*

# Persistent Identifier

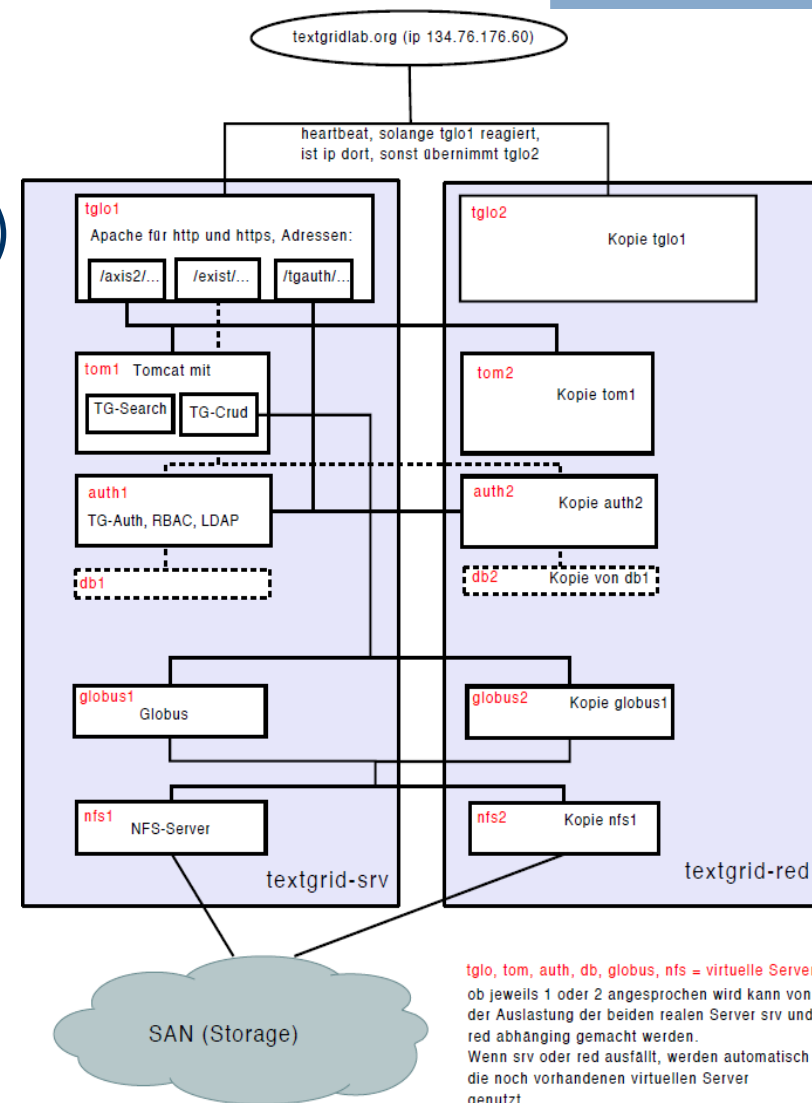
- PID-Service der GWDG:  
<http://handle.gwdg.de/PIDservice/>  
(in Zusammenarbeit mit bzw. für die MPG)
- GWDG ist Partner in *EPIC - European Persistent Identifier Consortium* - <http://www.pidconsortium.eu/>
- Resolving: Splash Page mit Metadaten und Informationen z.B. über Replikate
- Verweise auf 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

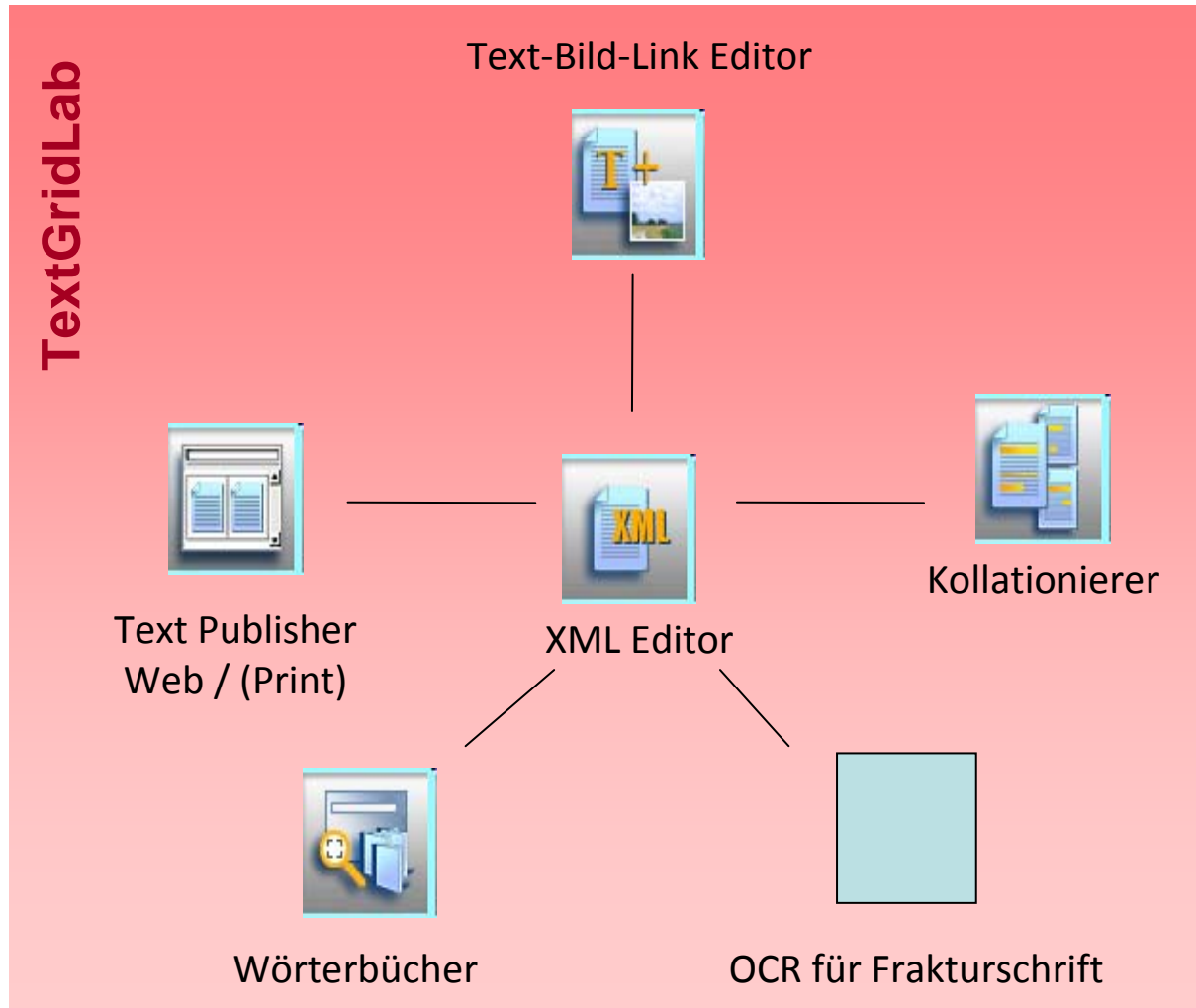
## Konzept zur Ausfallsicherheit

- 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  
(Februar 2011)



# Anwendungsfall: Editionsphilologie

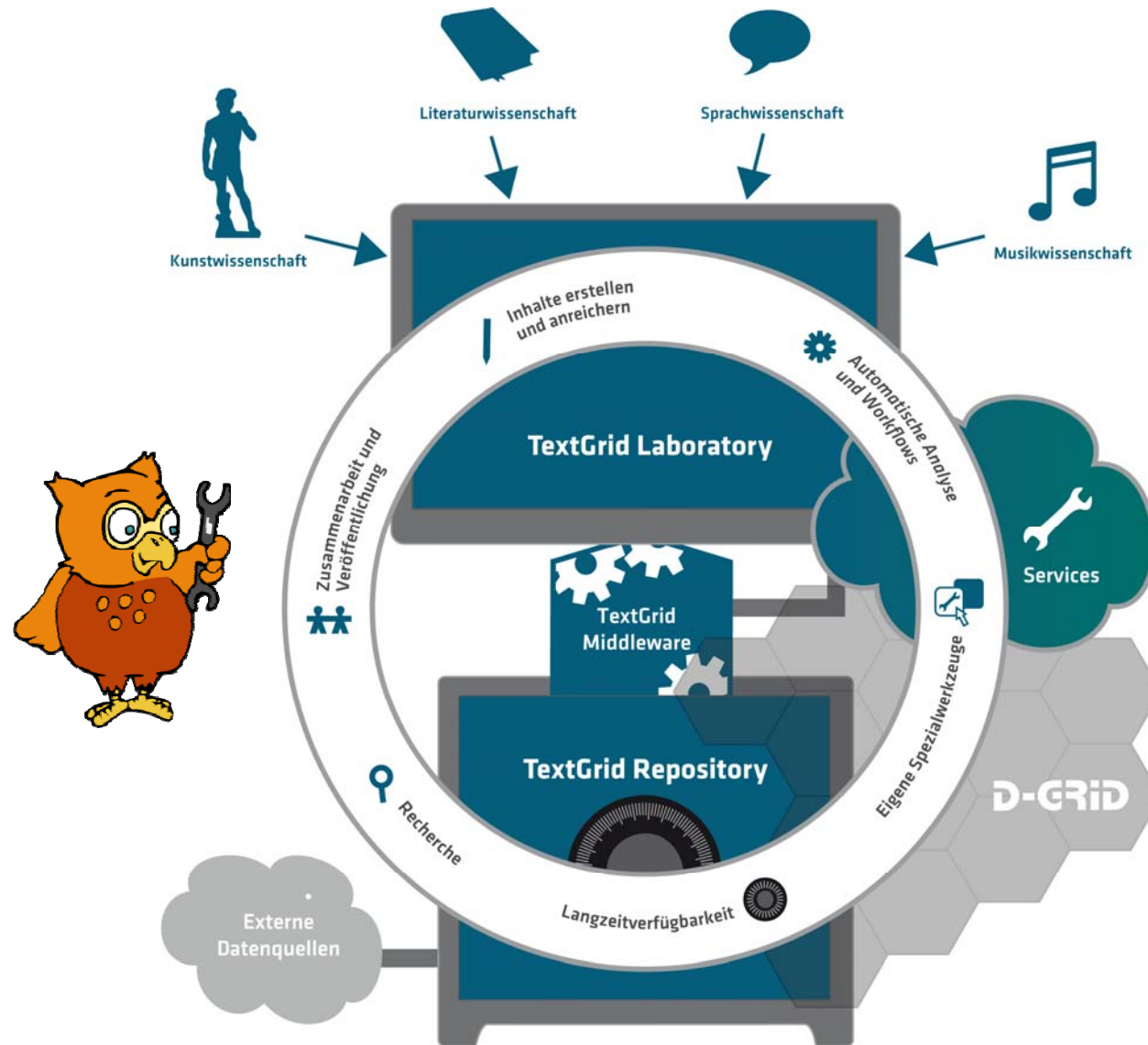


## TextGridRep

- Grid Storage für die Forschungsdaten
- Suche
  - Metadaten
  - Beziehungen
  - Volltext
  - XML-Strukturdaten
- Publikation und Nachweis im sammlungs-spezifischen Portal
- Persistent Identifier
- Metadaten-Validierung
- LZA-Dienste

- TextGrid Version 1.0: Februar 2011
  - Release einer stabilen, einsatzfähigen Version 1.0: 31. Januar 2011
  - Intensivtests 1. Dezember 2010 bis 31. Januar 2011
  - Release-Workshop 23./24. Februar 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)

# Infrastruktur für die eHumanities





**\*\* Vielen Dank \*\***

**Fragen, Anmerkungen ?**

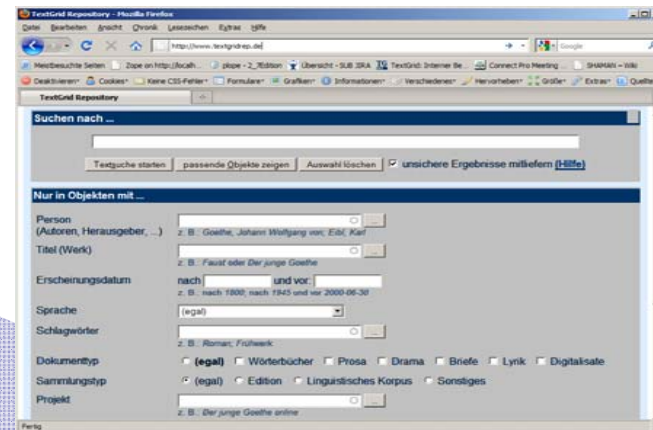
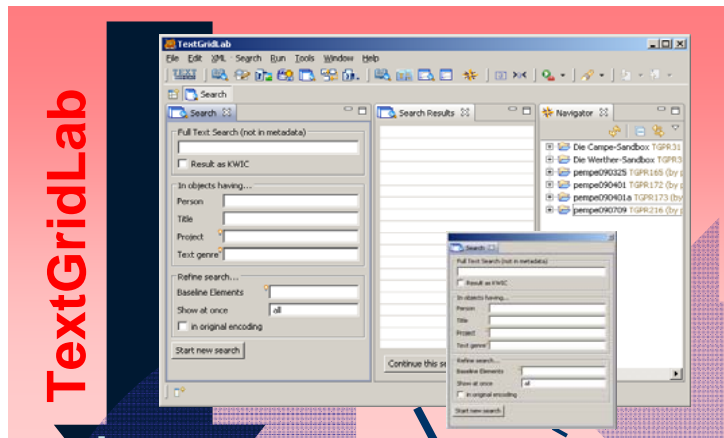
## Backup-Folien...



# TextGrid VRE – Black Content

## Eclipse RCP

## Portal (search + display)



TextGridLab

Ingest 1

TextGridRep

Rechtemanagement

TG-auth\*

Such-Index 1

isStable

Such-Index 2

TG-auth\*

Ingest 2



+ MD-Validierung  
+ PID  
+ ggf. LZA-MD  
+ ggf. LZA-Services



dynamisch

Grid Storage

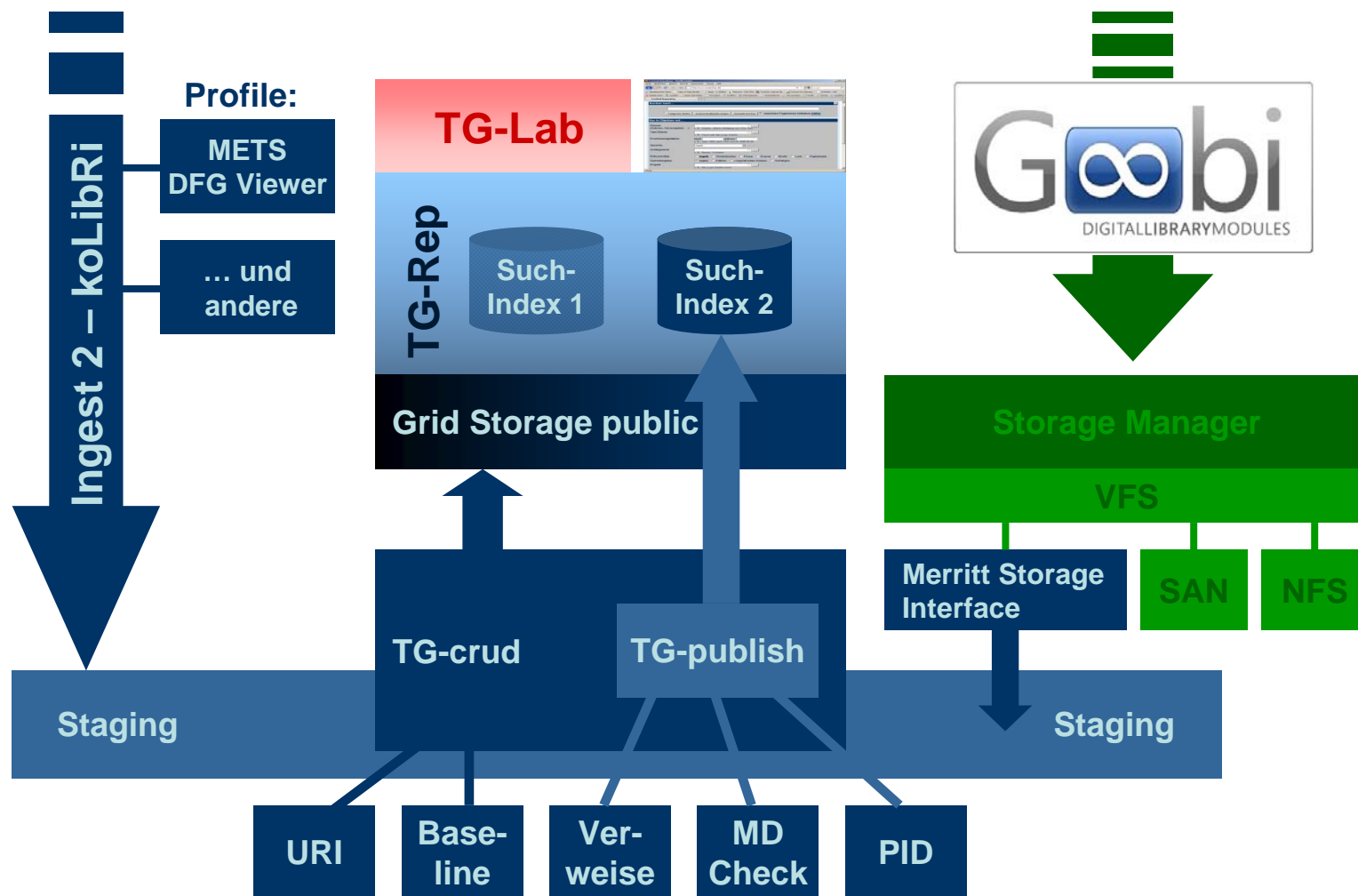
statisch

LZA

- große Datenmengen
- individuell angepasst
- + **Zugriffsrechte**
- + ggf. MD Validierung
- + ggf. PID
- + ggf. LZA-MD
- + ggf. LZA-Services

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

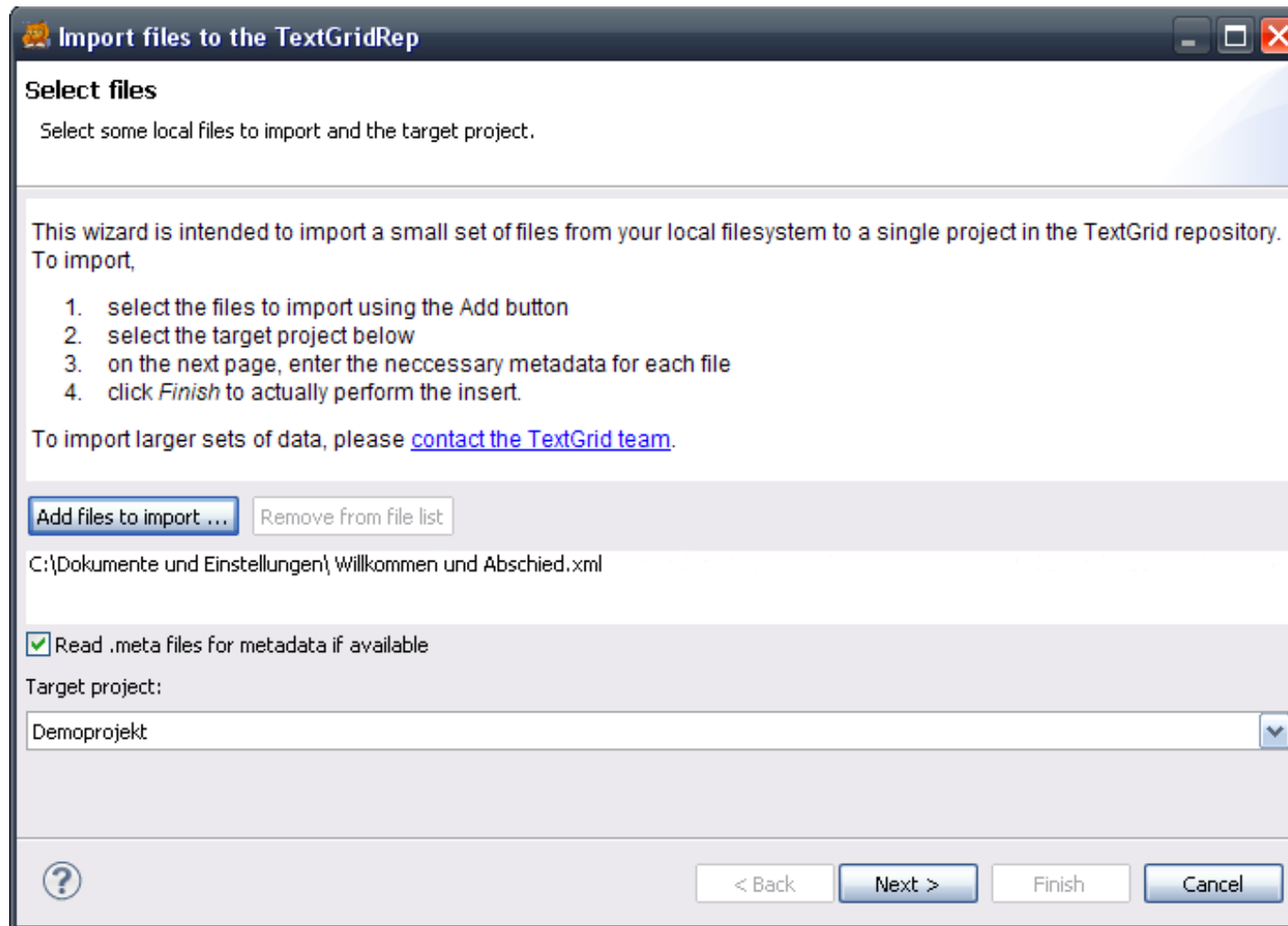
# Ingest 2 und TG-publish



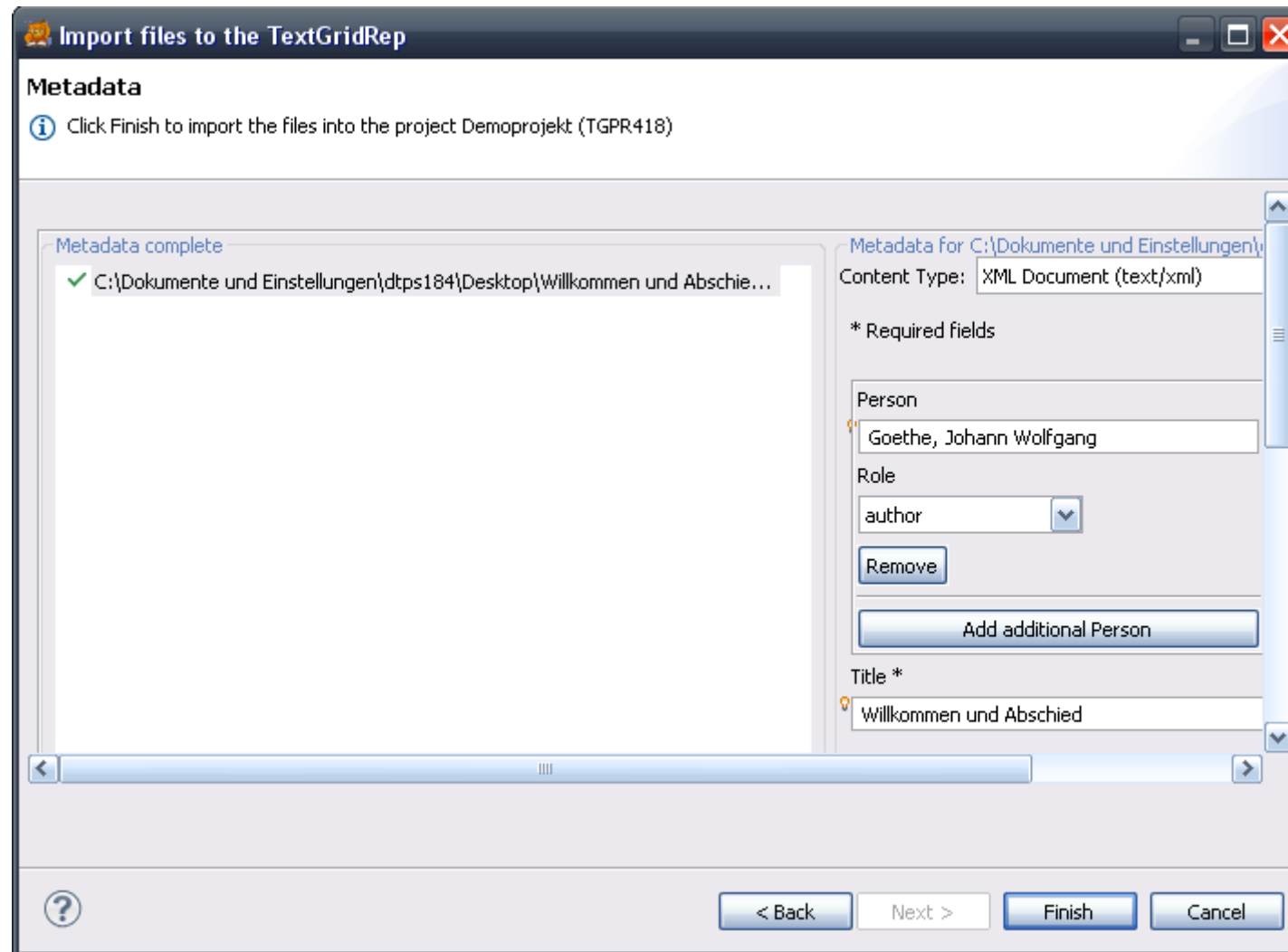
# Kernmetadaten-set

Item	Ausgabe	Werk
* title	* title	* title
	* contributor	* creator
identifier	identifier	identifier
* rightsHolder	* licence	
* format	description	abstract
	formOfNotation	* created
	language	date
		subject / temporal / spatial
		* type
		type

# Der Import-Dialog II



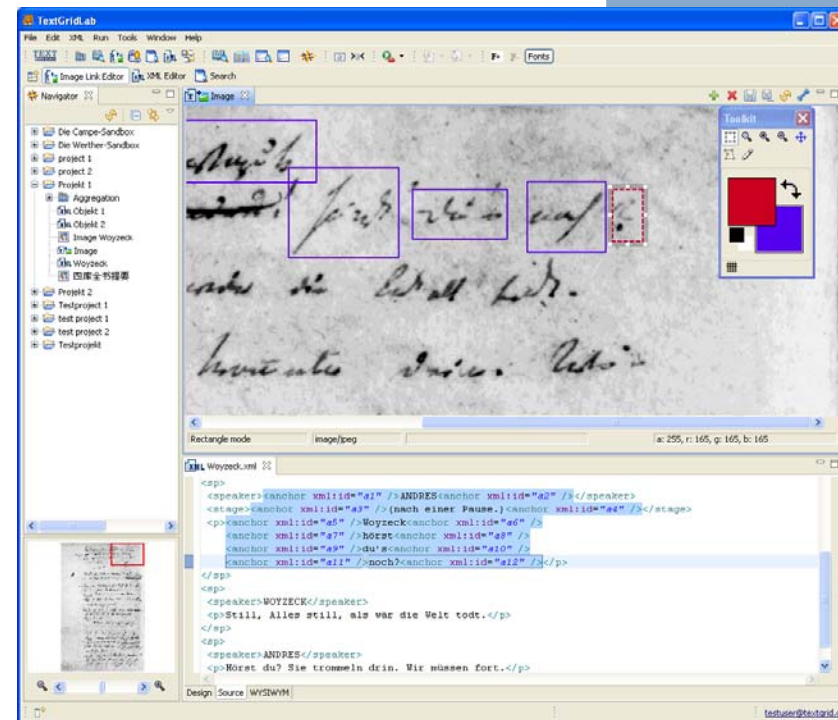
# Der Import-Dialog III



# Text-Bild-Link-Editor



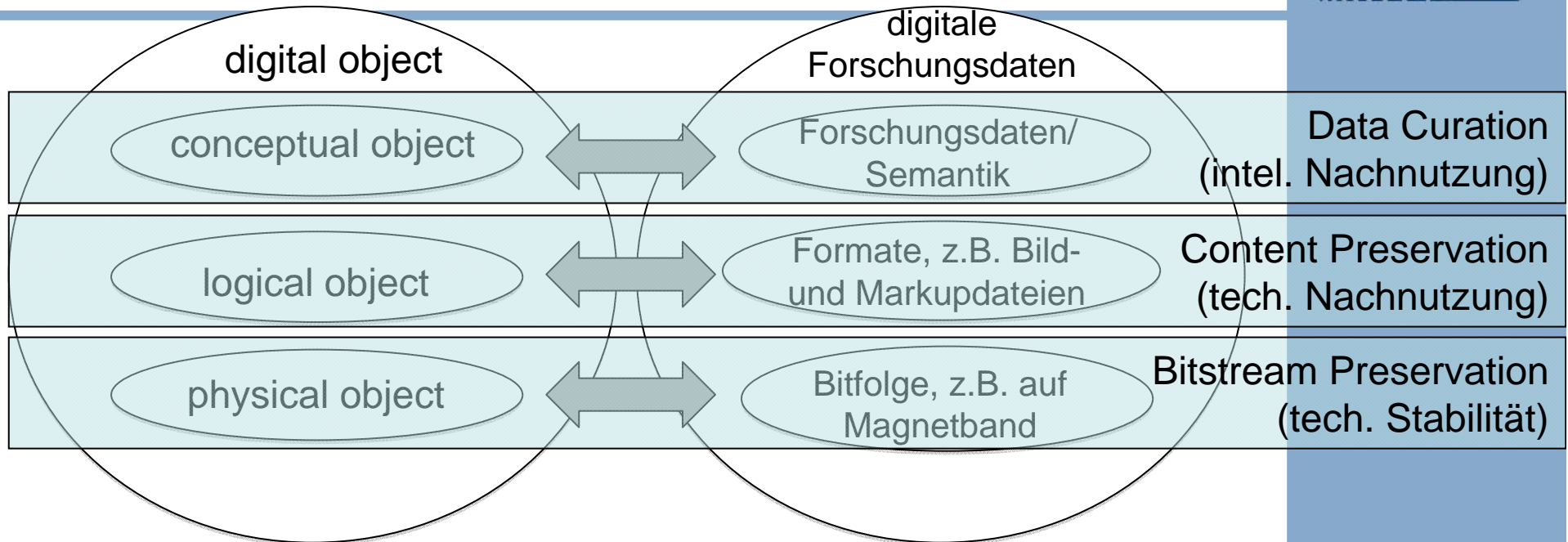
- Der **Text-Bild-Link-Editor** (Text ↔ Bild) unterstützt den XML-Editor bei der Alignierung von Text und Bildelementen. Ziel ist die Erstellung einer Ausgabedatei, die die Textelemente und die topographische Beschreibung enthält.



## Definition Virtuelle Forschungsumgebung

- Nach JISC (<http://www.jisc.ac.uk>)
- Zweck einer VRE ist die größtmögliche Unterstützung von Forscher/innen und ihren 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.

# Konzepte zur LZA (WissGrid)



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

# Bibliografische Metadaten

Metadaten, die für ein TextGrid-Objekt relevant sind, gibt es auf unterschiedlichen Ebenen:

Itemebene:

Digitalisat einer Seite aus Goethes *Faust* (2009)



Ausgabenebene:

*Reclam Universal-Bibliothek, Nr. 1, Faust I* (1996)



Werkebene:

*Goethe: Faust* (1808)

# Metadaten: Erfassungsmaske (Entwurf)

To create a new TextGridObject, please select its project first and fill in the following fields (fields with an asterisk \* are mandatory). If your object is part of an edition, you can allocate it to an existing one, search for or create a new edition.



A tree view showing project selection. 'project 1' is expanded to show 'edition 1'. 'project 2' is also visible. A red callout box points to 'edition 1' with the text 'Tip: will be allocated to this edition'.

Tip: will be allocated to this edition

Edition

source	<u>Dickens, Charles: Große Erwartungen. Frankfurt am Main: Insel Verlag 2010, Auflage: 1, ISBN 3458352384</u>	
title	Größe Erwartungen digital	+   -
	▶ more	
object name*	Seite 1	
type*	JPEG image (image/jpeg) ▼	
note		
	▶ more	
	< back	next >
	finish	cancel

[+ Fortschrittsbalken]

# Weitere Metadaten – Item

additional metadata

Item

object name*	<input type="text" value="Seite 1"/>		
type*	<input type="text" value="JPEG image (image/jpeg)"/>		▼
note	<input type="text"/>		
rightsholder*	<input type="text" value="SUB Göttingen"/>		
identifier	<input type="text" value="1234567"/>	type	<input type="text" value="PID"/> <input type="text" value="URI"/> ...

◀ less


< back    next >    finish    cancel  
[+Fortschrittsbalken]

# Weitere Metadaten – Edition

Edition

source Dickens, Charles: Große Erwartungen. Frankfurt am Main: Insel Verlag 2010, Auflage: 1, ISBN 3458352384

title \* Große Erwartungen digital

further title [Eine digitalisierte Ausgabe](#)   

contributor

identifier  type

language German (ger)  script Latin (Latn)

licence \* [cc by-sa](#)

note

released in \*

work

! Changes in the metadata will affect all objects associated with this edition

[+Fortschrittsbalken]

# Weitere Metadaten – Werk

work

Work

creator \* author: Dickens, Charles (<http://d-nb.info/gnd/11>) + -

uniform title\* ► search for / edit uniform title

identifier  type  ▼

date of creation\*

approx. date of creation  switch to date range

not before\* 1860 not before approx. Ostern 1834

not after\* 1862 not after approx.

back

type of work \* prose ▼ specific genre Roman

keywords subject Gesellschaftskritik type  + -

time Viktorianische Ära type thesaurus - -

time 19. Jahrhundert type  + -

place London type Getty + -

abstract Der Roman handelt von...

note

ok cancel

# Mengenberechnung / Skalierung (Beispiel: GDZ)

- Derzeitige Produktion: 150.000 Seiten pro Monat. Realistisch in 2010: 200.000
- Farbe, 300 dpi, TIFF uncompressed: 25 MB pro Seite
- x2 für Master / optimiertes TIFF
- $25 \times 2 \times 200.000 = 10\text{TB}$  pro Monate „worst case“. „best case“: 5TB
- Tageslast: 333 GB pro Tag, 13,8 GB pro Stunde, 3,8 MB/s bei 24/7 System
  
- Belastung auf Netzstruktur (Bandbreite bei 8h-Schicht):
  - Ca. 5 MB/s best case (alle 8 Scanstationen schreiben sequentiell)
  - 16 MB/s worst case (alle 8 Scanstationen schreiben parallel)
  
- Faustregel für große Einrichtung: 1 Image mit 25 MB alle zwei Sekunden produziert
- 8 Bände (Objekte) pro Stunde zu archivieren mit 250 Seiten (6,25 GB x 8 Bände x 2 Master/Optimiert = 100GB/h)
- 24/7 Betrieb: 33 GB/h, 2,7 Objekte/h