

# TextGrid

wissenschaftliche Textdatenverarbeitung –  
ein Community-Grid für die Geisteswissenschaften

## Technische Aspekte

Vortrag von Johannes Dörr

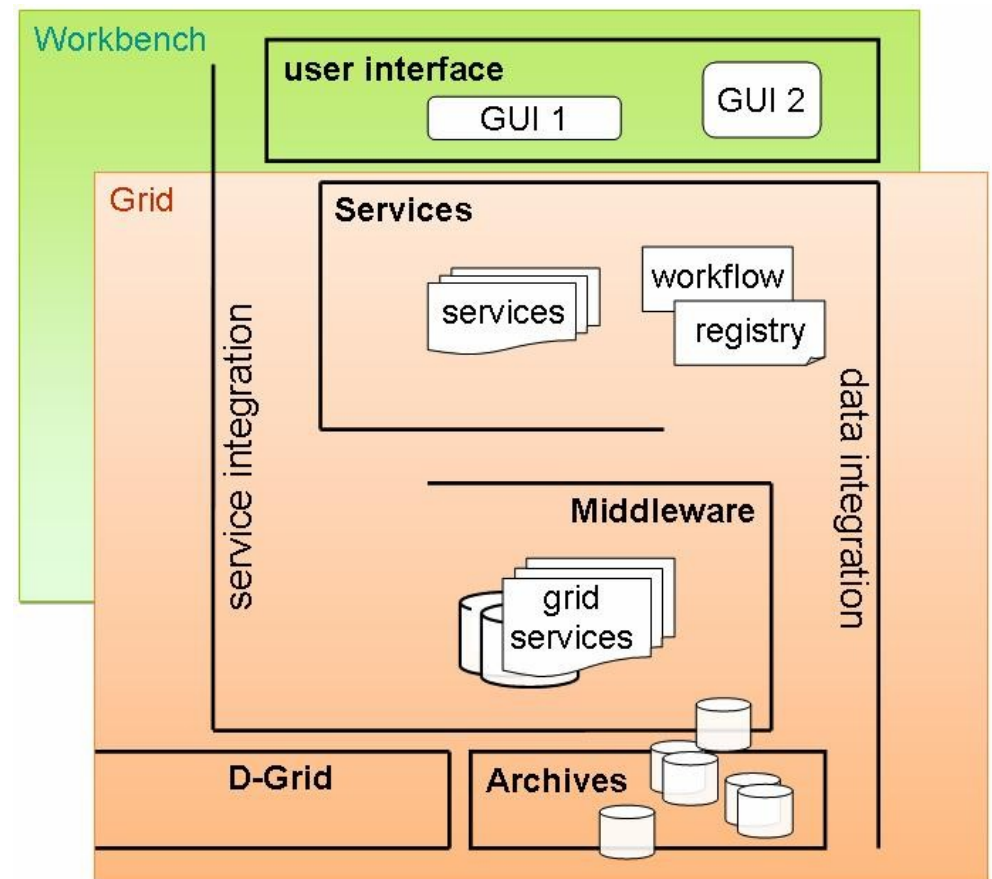


- Anforderungen
- Aufbau von TextGrid
  - **User Interface**
  - **Service Layer**
  - **Middleware**
  - **Archive**
- Wie kommen Daten ins TextGrid?
- Security
- Demonstration von TextGridLab
- Geplante Tools
- Zusammenfassung

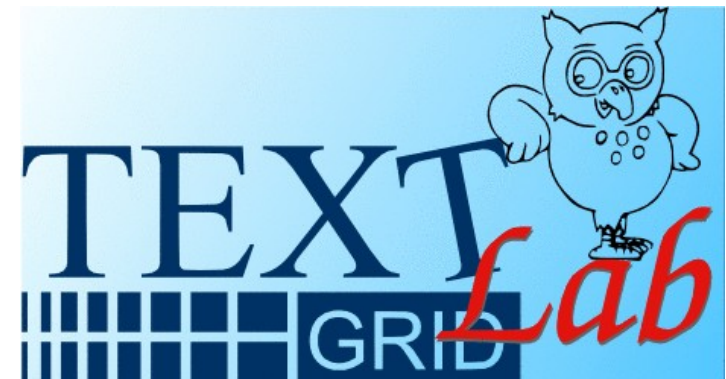
- Zusammenbringen spezialisierter Anwendungen der Textwissenschaft
- Erweiterbarkeit für Belange anderer Geisteswissenschaften
- Wenige Vorgaben an den Benutzer
- Einfachheit [1]: Geisteswissenschaftler sind keine Informatiker
- Einfachheit [2]: Unkomplizierte API erleichtert Implementierung von neuen Funktionen

# Aufbau von TextGrid

- Aufteilung in Schichten
  - **User Interface**
  - **Service Layer**
  - **Middleware**
  - **Archives**

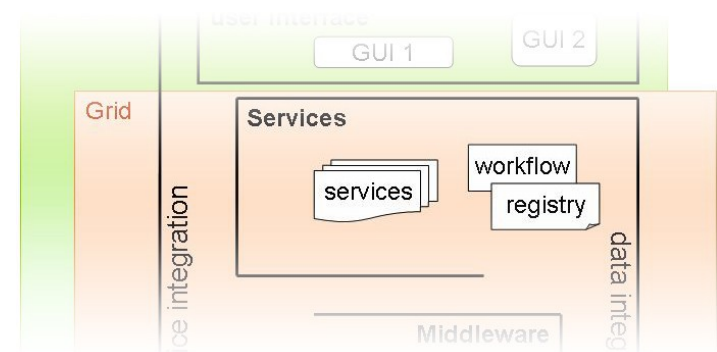


- Theoretisch mehrere, spezifische GUIs denkbar
- Basierend auf Eclipse Rich Client Platform
  - **Sehr gut erweiterbar**
  - **Kernaspekt von TextGrid: Einfache Bedienung**
- Bisher ist nur eine GUI (*TextGridLab*) in Entwicklung für:
  - **Edition, Annotation**
  - **Analyse**
  - **Projektmanagement**
- Offline-Bearbeiten von Dokumenten



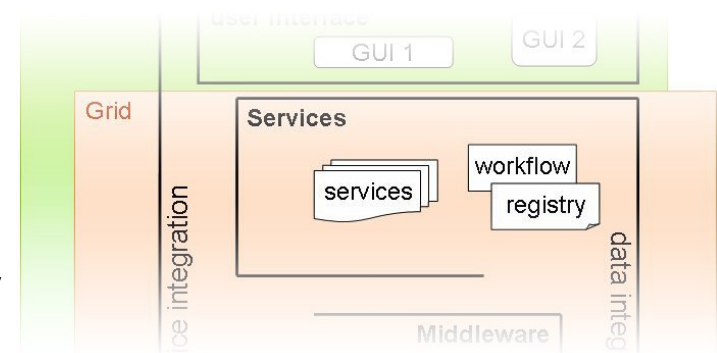
# Aufbau / Service Layer

- Spezialisierte Funktionen implementiert als Web Services
- Drei Arten von Services:
  - **rein Algorithmen-basiert**
  - **mit Zugriff auf statischer Wissensbasis**
  - **mit Zugriff auf (dynamische) Daten des Benutzers**
- TextGrid Registry in Planung
  - **listet verfügbare Services auf**
  - **einfache Suche nach gewünschter Funktionalität**
- Bestimmte (interaktive) Services können aus GUI aufgerufen werden



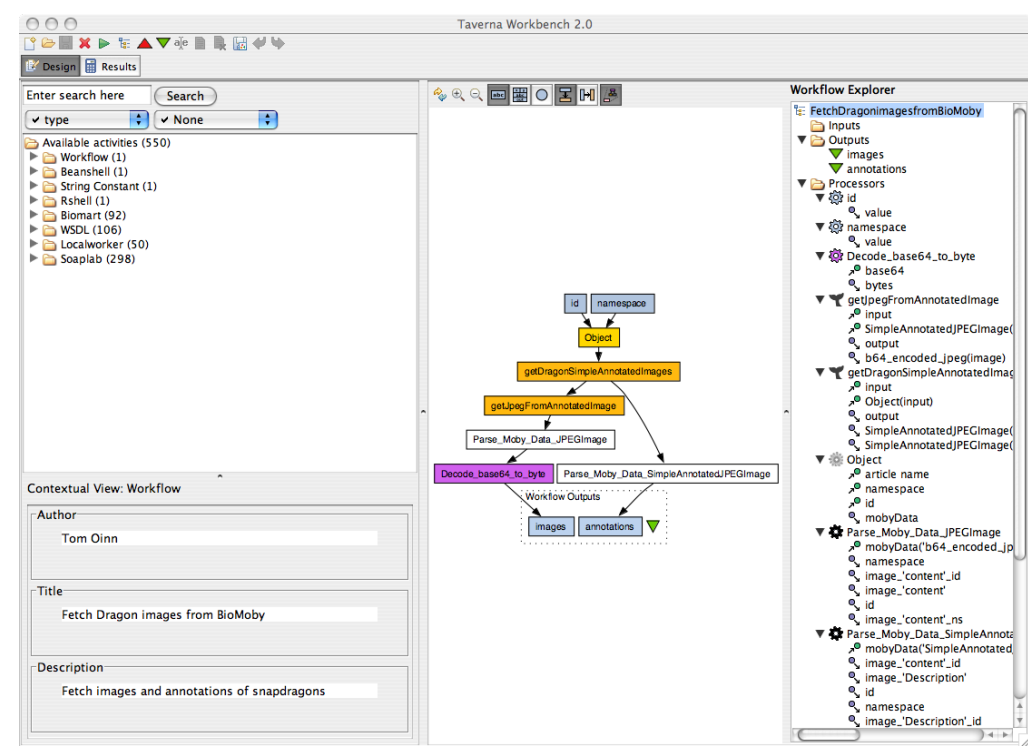
# Aufbau / Service Layer [2]

- Offenheit des Service Layers soll Initiativen motivieren, evtl. auch auf kommerzieller Basis
- „Service Grid“ - offene Plattform, Streaming Services sind beliebig kombinierbar
- Wie kombinieren? → Workflows



# Aufbau / Service Layer [3]

- Workflows werden im Benutzerinterface erstellt (Workflow Editor) und dann
  - entweder direkt ausgeführt
  - oder dem Workflow Enactor (implementiert als Web Service) übergeben. Dieser führt Workflow automatisch aus
- Workflow Editor ist im Moment noch nicht fertig umgesetzt. Könnte bspw. Auf Taverna basieren



# Aufbau / Service Layer [4]



Milestones/ Reports	Monate	Tool	Tool-Typ	Verantwortlich	DAASI	FH Worms	IDS Mannheim	Saphor	SUB Göttingen	TU Darmstadt	U Trier	U Würzburg	fachi. Betreuung	altes Tool / Kommentar
		<b>PM insges.*</b>			7	2	14	37	10	24	26	3		
M 2.1	1-6	<b>Tokenizer</b>	s	Saphor			x	x	x	x	x		IDS, Trier	
		<b>Workflow-Editor</b>	i	DAASI	x				x	x			TUD	Editor techn. Workflow
M 2.2	7-12	<b>Lemmatisierung</b>	s	IDS Mannheim			x	x			x		IDS, Trier*	* nicht Neuhochdeutsch
		<b>XML-Editor</b>	i	TU Darmstadt					x	x			TUD	
		<b>Rich Client Platform (GUI)</b>	i	SUB Göttingen					x	x			SUB	
R 2.1	12	<b>Dokumentation</b>												
M 2.3	13-24	<b>Recherchetool</b>	i	Saphor			(x)	x		x			IDS, TUD	Query Interface, Text Retrieval
		<b>Streaming-Editor I (XSLT-Komponente, u.a. für Adaptor-Manager)</b>	s	Saphor	x			x					TUD, Worms	
		<b>Metadaten-Annotation</b>	i	SUB Göttingen			x		x	x	x		SUB	
		<b>Datei- / Rechtemanagement</b>	i	SUB Göttingen	x				x					Editor admin. Workflow
		<b>grafischer Link-Editor</b>	i	TU Darmstadt						x		x	TUD, WÜ	Link-Editor
		<b>Bild-Segmentierung</b>	s	TU Darmstadt						x		x	TUD, WÜ	Link-Editor
		<b>Link-Editor Text</b>		Saphor				x		x			TUD	Link-Editor
		<b>Bibliographietool</b>	i	DAASI	x			x			x		Trier	
		<b>Sortieren</b>	s	FH Worms		x	x				x		IDS, Trier	
R 2.2	24	<b>Dokumentation</b>												
M 2.4	25-30	<b>Streaming-Editor II</b>	s	Saphor				x		x			TUD	
		<b>Kollationierung</b>	s	U Trier						x	x	x	Trier, WÜ	
M 2.5	31-36	<b>Text Publisher (Print)</b>	s	FH Worms		x		x						
		<b>Text Publisher (Web)</b>	s	Saphor				x						
		<b>OCR</b>	s	U Würzburg					x		x	x	SUB, Trier, WÜ	
R 2.3	36	<b>Dokumentation</b>												

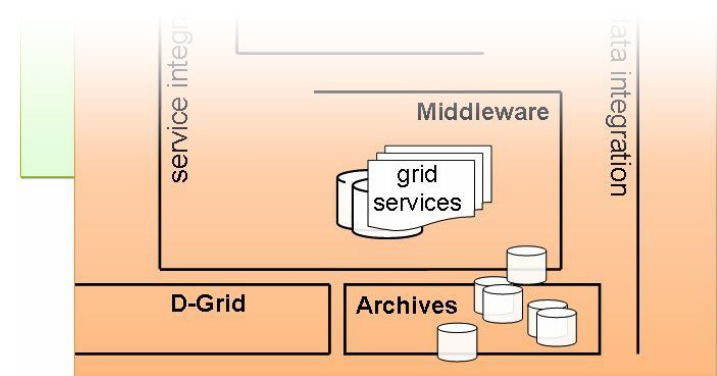
Tool-Typ: „s“ = Streaming Tool, „i“ = interaktives Tool; Grün hinterlegt: gegenüber dem Antrag geänderte Tools

TextGrid Report 2.1

5.2.2007

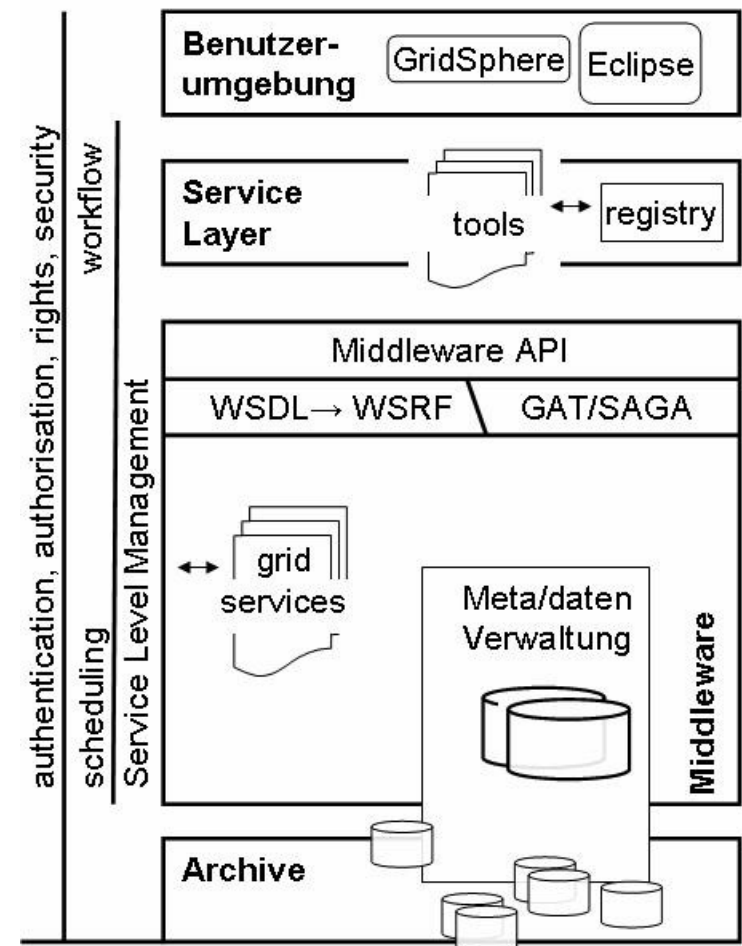
# Aufbau / Middleware

- Kernaufgabe: Virtualisierung der Speicherressourcen
- Grid Services sollen möglichst allgemein und nicht Textwissenschaften-spezifisch sein
  - **Verwendung von TextGrid auch in anderen Geisteswissenschaften**
- Wichtige Aufgaben könnten vom Service-Layer in die Middleware verlagert werden
  - **Steigerung der Verfügbarkeit und Skalierbarkeit**
- Dennoch stehen statt „Computational Grid“ momentan im Vordergrund:
  - **Nachhaltige Datenspeicherung („Data Grid“)**
  - **Offene Dienstplattform („Service Grid“)**



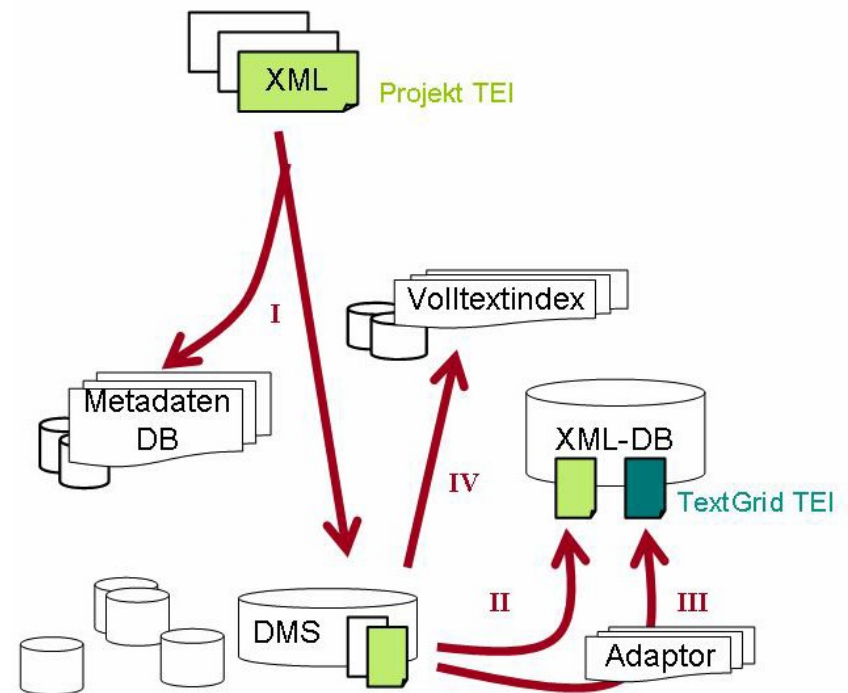
# Aufbau / Middleware [2]

- Baut auf *Globus<sup>®</sup> Toolkit* auf
  - große Community (guter Support)
  - spezielle Komponenten (z. B. GridShib) werden vornehmlich hierfür entwickelt
- Abstraktionsschicht zwischen Service Layer und Middleware
  - Austauschbarkeit der Middleware möglich durch Einsatz von GAT bzw. SAGA
  - Vermittlung zwischen WSRF und der auf einfachen Web Services beruhenden TextGrid API

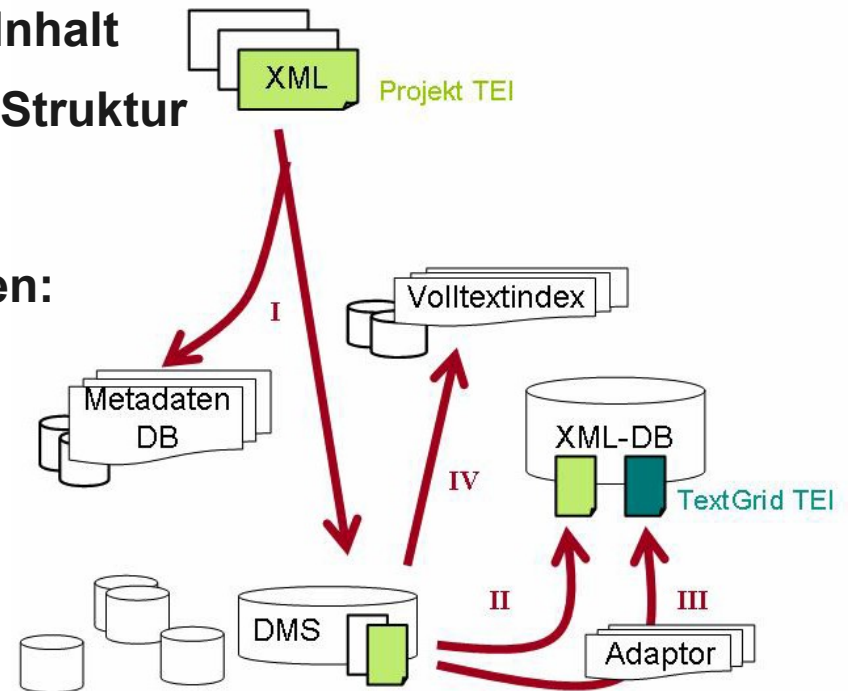


- Archiv-Schicht ist Teil der Middleware, wird dennoch gesondert geführt
- TextGrid speichert Daten in vielfacher Form

- **Datenhaltung (DMS)**
- **Metadaten DB**
  - Titel, Autor, Bearbeiter, Zeitraum
- **Volltextindex**
- **Strukturdatenbank (XML-DB)**
  - Ermöglicht das Suchen in XML-Daten mit Abfragen über Xpath o. Ä.

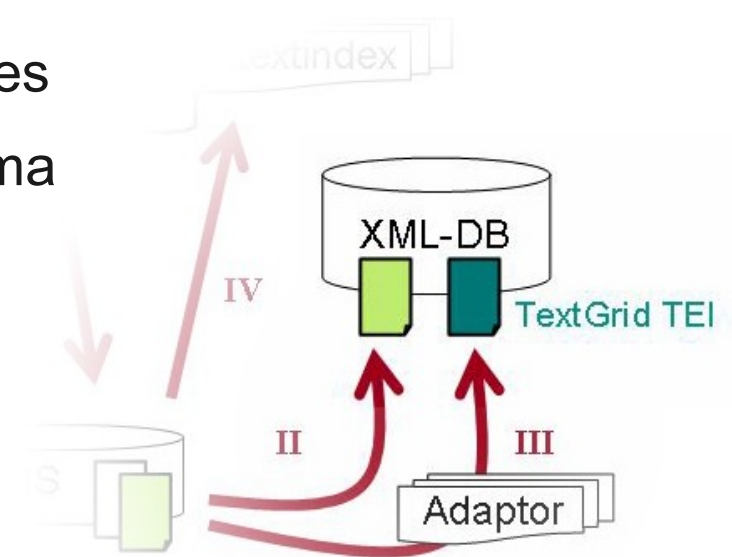


- Ist das nicht zu viel des Guten? - Nein!
  - **Metadaten-DB** ist immer nötig, besonders für nicht-XML-Dateien
  - **Volltextindex** durchsucht den gesamten Inhalt von Textdaten ohne Beachtung der XML-Struktur
  - **Strukturdatenbank (XML-DB)** ermöglicht gezielte Einschränkung von Suchanfragen:
    - z.B. auf gesprochenen Text (Drama)
    - z.B. auf Überschriften
    - z.B. Wörterbuch-Stichworte



# Aufbau / Archive [3]

- Problem bei Suche in Strukturdatenbank: Jedes Projekt soll sein eigenes Auszeichnungsschema verwenden dürfen
  - **Fremde Benutzer benötigen Kenntnis des Schemas, um Suchanfrage zu formulieren**
  - **Abhilfe: TextGrid definiert ein Basisschema**
  - **Projektmitglied erstellt XSLT-Stylesheets (Adapter), mit denen von projektspezifischen Schemata auf das Basisschema abgebildet wird**
- In der Strukturdatenbank werden Daten sowohl im Originalschema als auch im Basisschema abgespeichert
  - **Erhöhte Datenmenge zu Gunsten besserer Suchperformance**



# Wie kommen Daten ins TextGrid?

- Mehrere Möglichkeiten für die Datenanbindung

- **Metadatenintegration**

- Es liegen nicht notwendigerweise assoziierte Datenobjekte vor

- **Archiveinbindung**

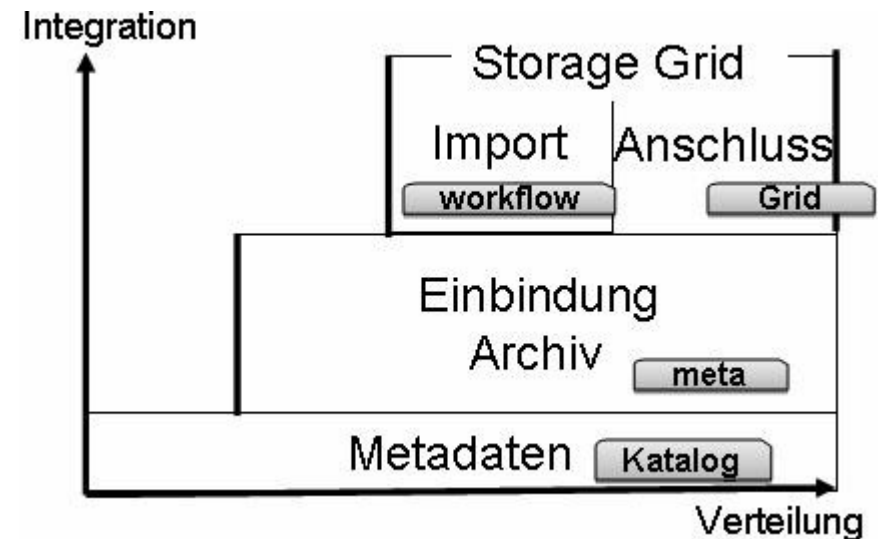
- Metadatenintegration + Transfer der Datenobjekte bei Bedarf

- **Datenimport**

- Datenobjekte werden in TextGrid integriert

- **Grid-Anschluss**

- Betreiben der TextGrid-Software (bisher in Göttingen, Mannheim, Würzburg)



- Accounting/Logging
  - **Schützen laufender Projekte vor unbefugtem Zugriff**
  - **Rollenverteilung bei Projekten: Leiter, Bearbeiter, Beobachter**
  - **Logging, um Missbrauch zu erkennen**
- Bisher kein Billing für Nutzung urheberrechtlich geschützter Dokumente
  - **später jedoch denkbar**
- Authentifizierung/Autorisierung per Shibboleth
  - **Einmalige Anmeldung an der Heimatorganisation, danach Zugriff auf alle zugewiesenen Ressourcen**

# Demonstration von TextGridLab

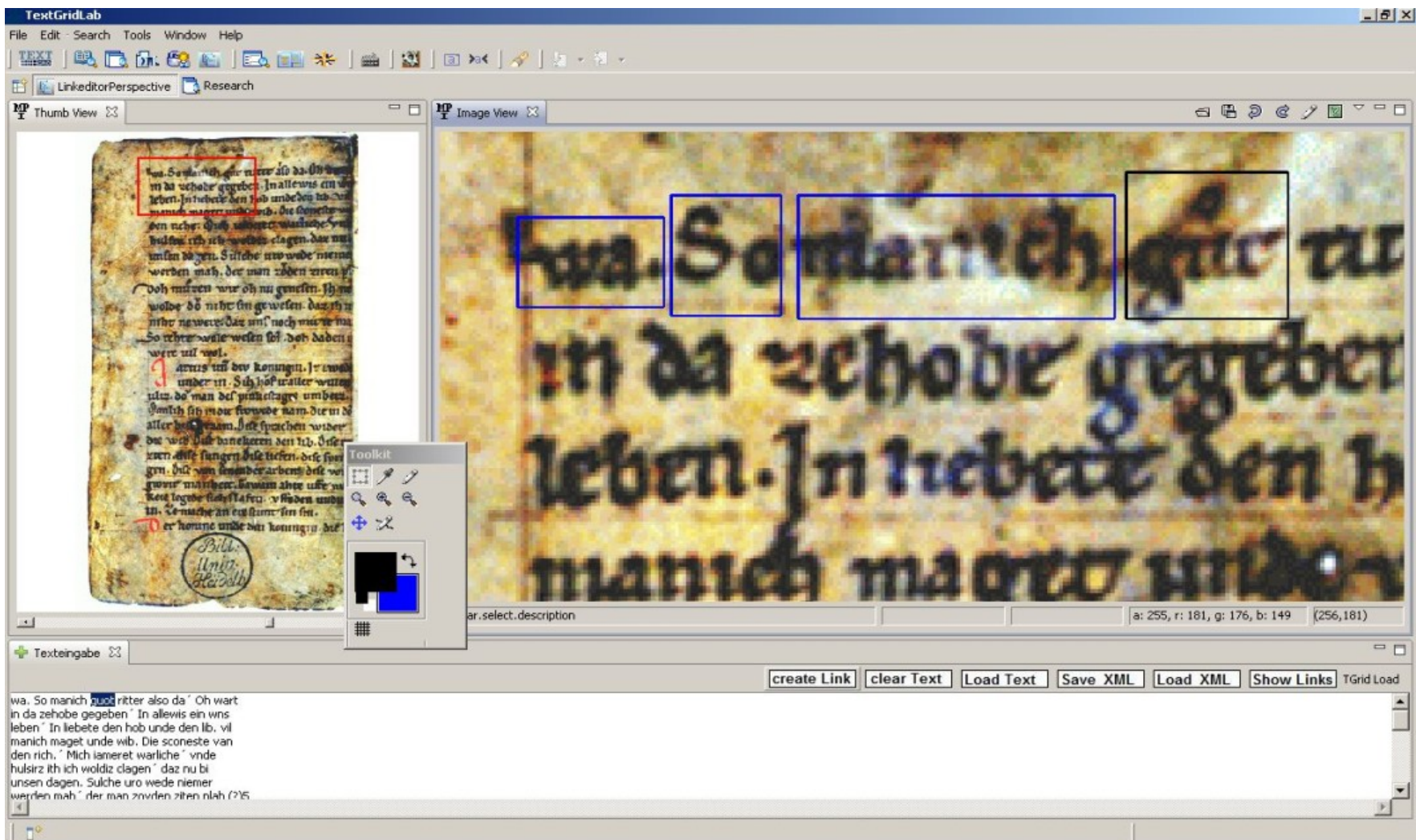


A screenshot of the TextGridLab web application interface. The browser window title is "TextGridLab" and the address bar shows "http://chann-4:~/Desktop/textgridlab". The menu bar includes "File", "Edit", "XML", "Search", "Tools", "Window", and "Help". The main content area features the TextGrid logo and the text "Welcome to TextGridLab". Below this, there are several interactive buttons: "Login" (with a person icon), "Project &amp; User Management" (with a cat icon), "Search" (with a magnifying glass icon), "Image Editor" (with a document and image icon), "XML Editor" (with an XML document icon), and "Workflow Editor" (with a flowchart icon). A note at the bottom of the main content area reads "For Navigator, Dictionaries and other tools see the Tools menu." At the bottom center, there is a "Help" button with a question mark icon. The status bar at the bottom right shows "AuthNDialogue" and a small gear icon.

# Geplante Tools

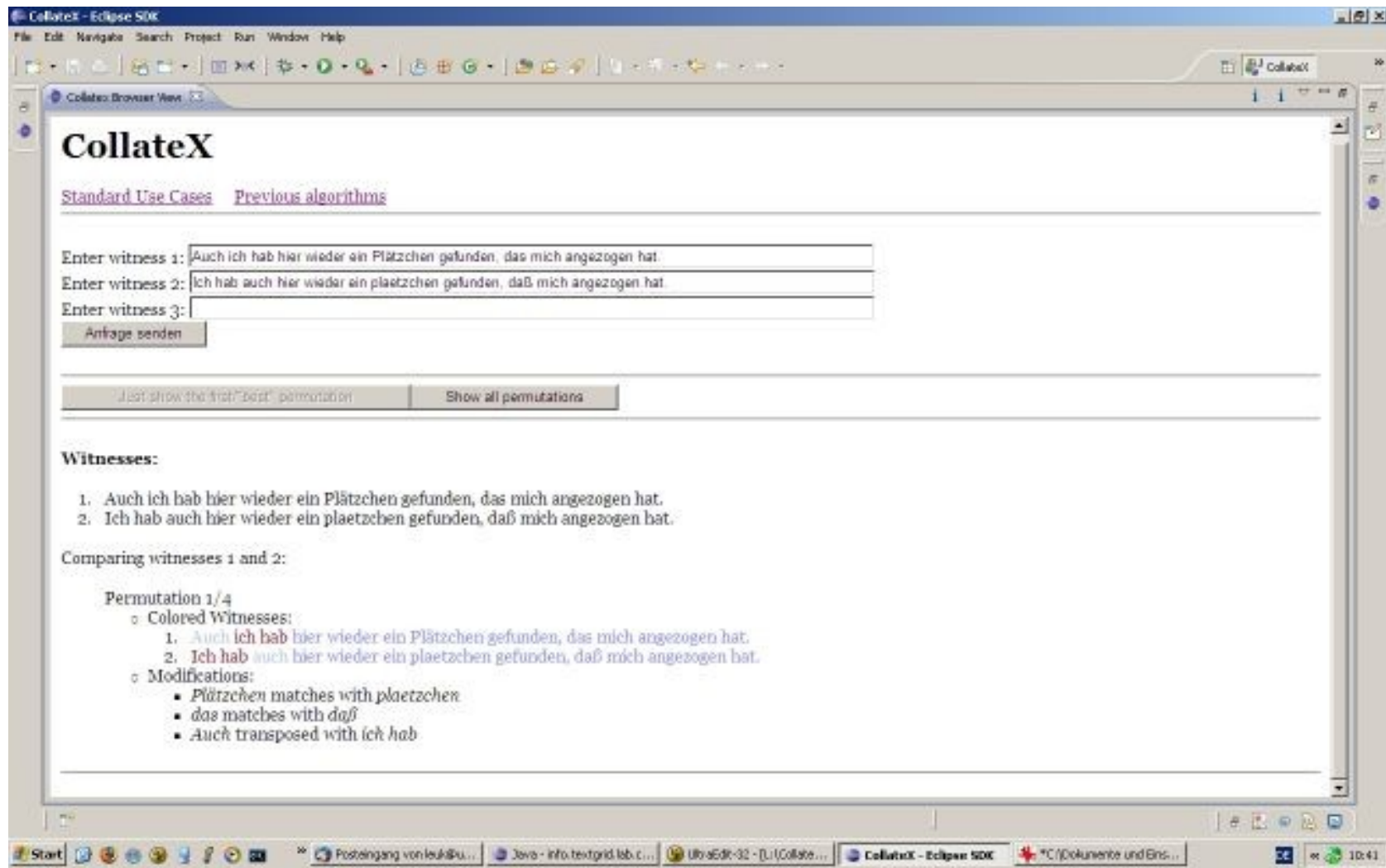


- Image Link Editor



# Geplante Tools [2]

- Kollationierer



# Zusammenfassung



- TextGrid lässt dem Benutzer viele Freiheiten bei der Gestaltung (Annotation) der Daten
- Durch offenen Service Layer können weitere Funktionen implementiert werden. TextGrid ist auf solches Engagement angewiesen
- TextGrid setzt sich die einfache Handhabung zum Ziel – sowohl bei der Forschungsarbeit und der Projektadministration als auch bei der Programmierung von Zusatzfunktionen
- TextGridLab ist teilweise noch unvollständig und fehlerbehaftet
- Stabilität und Notwendigkeit von Skalierung des Grids wird sich in der Praxis erweisen, wenn Zahl der Nutzer steigt

Vielen Dank!

Bilder: Entnommen aus TextGrid-Reports

Vielen Dank an Andreas Aschenbrenner für die Betreuung