

# TextGrid's Ecosystems of eHumanities Resources and Services

Christoph Ludwig

`ludwig@fh-worms.de`

DEBII – March 20, 2008



Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Overview

## Background: Textual Scholarship

Background:  
Textual  
Scholarship

## The TextGrid Project

The TextGrid  
Project

## The TextGrid Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic Environment

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

## ROA and DEs

# Traditional Critical Editions

- ▶ Attempt to establish an (ideally) error-free “definitive text” or “base text”
- ▶ Describe existing witnesses of the text (old manuscripts, early prints, . . . )
- ▶ Fully cover the genesis of the text (authorial or scribal additions, deletions, comments, etc.)
- ▶ Note (all, most, or selected) variants between witnesses
- ▶ Explain unclear names, words, potentially corrupt passages, historical background, . . .
- ▶ Publish the end result in high-quality print

. . . and often take decades to complete. Intermediate results and information on decision processes are almost always lost.

# The Lone Scholar in his Attic



Étienne Baluze (1630-1718)



Prof. Dr. Gerhard Schmitz



Emil Seckel (1864-1924)

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Today's Requirements (1/2)

- ▶ Publish results – including intermediate results – fast online
- ▶ Publish the end results in several media (online, print)
- ▶ Interact with your peers
- ▶ Work in geographically distributed teams
- ▶ Use software agents to automate repetitive tasks and get more reliable and verifiable results
- ▶ Document progress and your editorial decisions
- ▶ Link text and variants to the digitized witnesses





# Modern Digital Edition Projects

- ▶ Textual Data encoded in XML (TEI)
- ▶ Thousands or tens of thousands very high-resolution images (linked to the textual data)
- ▶ Mostly large project teams, no “lone scholars” anymore
- ▶ Scholars often not very technology savvy
- ▶ (Very) long-term data storage demands
- ▶ Intricate publication requirements
- ▶ Individual tools are similar or identical

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Overview

Background: Textual Scholarship

Background:  
Textual  
Scholarship

The TextGrid Project

The TextGrid  
Project

The TextGrid Ecosystem

The TextGrid  
Ecosystem

TextGrid's Species

TextGrid's Species

Data in TextGrid

Data in TextGrid

TextGrid's Abiotic Environment

TextGrid's Abiotic  
Environment

ROA and DEs

ROA and DEs

Establish an interdisciplinary platform – a community grid  
– for **collaborative research in textual scholarship**  
based on **grid technologies**.

## In particular:

- ▶ Preparation of critical editions
- ▶ Edition and elaboration of dictionaries
- ▶ Corpus linguistics

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Objectives: Tools

Develop a comprehensive tool set for the

- ▶ editing
- ▶ annotating
- ▶ analyzing
- ▶ publishing

of textual data in philologies, linguistics, literary studies  
and related fields.

# Objectives: Grid-Infrastructure

Leverage the grid infrastructure so that:

- ▶ textual data and supporting images from various research projects and content providers such as archives and libraries can fuse into a virtual corpus that can be seamlessly searched and analyzed
- ▶ participating organizations can provide dedicated services – software agents – with well-defined interfaces that can be harnessed together through a user defined workflow to mine or analyze existing textual data or to structure new data both manually and automatically

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Objectives: Grid-Infrastructure

Leverage the grid infrastructure so that:

- ▶ textual data and supporting images from various research projects and content providers such as archives and libraries can fuse into a virtual corpus that can be seamlessly searched and analyzed
- ▶ participating organizations can provide dedicated services – software agents – with well-defined interfaces that can be harnessed together through a user defined workflow to mine or analyze existing textual data or to structure new data both manually and automatically

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# “Non-Objectives”

The following issues were consciously left out of the original architecture:

- ▶ Digital Rights Management (DRM)
- ▶ High-quality printing of critical editions

# Organizational Set-Up

- ▶ Consortium of six academic and two commercial partners
- ▶ Funded by the German Federal Ministry of Education and Research as part of the D-Grid initiative
- ▶ Project started in February 2006 for a duration of three years
- ▶ Deliverables are open source (LGPL)

- ▶ Long-term strategic program to establish a national Grid infrastructure in Germany
- ▶ Provide basic, sustainable resources and services that other eScience projects can build upon
- ▶ Highly collaborative
- ▶ Grew from initially 6 to now 16 “community grids” (sciences, engineering, recently business) plus several horizontal grid projects

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

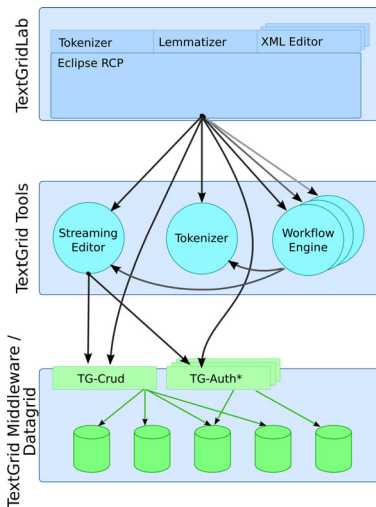
Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Bird's Eye View on the TextGrid Architecture

eHumanities  
Resources and  
Services



Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Overview

Background: Textual Scholarship

Background:  
Textual  
Scholarship

The TextGrid Project

The TextGrid  
Project

**The TextGrid Ecosystem**

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic Environment

**The TextGrid  
Ecosystem**

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

ROA and DEs

# Generic Services in TextGrid

- ▶ Generation and execution of workflows
- ▶ Development of several Eclipse plugins as the central end user interface for grid data storage, grid services and workflow components
- ▶ Handling of metadata and of domain ontologies
- ▶ Finding and retrieving data in the grid

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Services Specific to Textual Scholarship

- ▶ Validating, user-friendly WYSIWYM XML editor
- ▶ Link editor for annotations between XML data and digitized manuscripts
- ▶ Lemmatizer
- ▶ Collator
- ▶ “Streaming editor”
- ▶ Tokenizer
- ▶ Locale-sensitive sorting
- ▶ ...

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

Since TextGrid is based on open standards – UCS / Unicode for character encoding, TEI for Data, SOAP, WSDL, WSRF for web services, Eclipse for GUI etc. – it is easy

- ▶ to integrate new external data providers such as archives
- ▶ to integrate and cross-link networks of historical and domain dictionaries
- ▶ for the community to write new services
- ▶ to provide new GUI components

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

**eScience:** *new form of scientific work building on the global networking of computer resources, knowledge, tools and human beings (Aschenbrenner et al., 2007)*

**eHumanities:** *eScience in the field of humanities*

***Digital Ecosystem:*** self-organizing digital infrastructure, aimed at creating a digital environment for networked organizations (or agents) supporting the cooperation, knowledge sharing and development of open and adaptive technologies and evolutionary domain knowledge rich environments (DEST 2007 Call)

*Digital eco-systems occur through the interactions between both human and computer-based agents, operating in a manner that creates both relationships of cooperation and conflict within the system as well as the overall system itself (DEST 2007 Call for Special Session on eHumanities / Küster and Allen)*

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Evolution and System Breeding (Chuan-Leong Lam)

eHumanities  
Resources and  
Services

Background:  
Textual  
Scholarship

The TextGrid  
Project

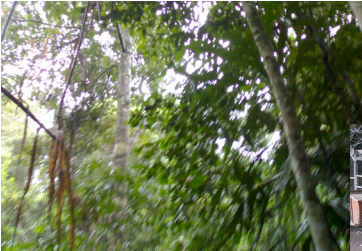
The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs



Source: Private photo



Source: Flickr photo 363283623 „farm animals can all play together“

# The Ecosystem in Stable State

- ▶ Evolved from previous states that left their marks
- ▶ Has both human and digital inhabitants (software agents)
- ▶ Must interact with related ecosystems
- ▶ Has open boundaries – individual species will roam between various digital ecosystems
- ▶ Will be inhabited by different species of many different sizes living in various subsystems

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

1. **End Users:** mostly scholars  
(philologists, linguists, historians, . . . )
2. **Content Providers:** mostly institutions  
(archives, libraries, commercial publishers),  
also scholars publishing in TextGrid
3. **Software Developers:**  
build, maintain & enhance the platform

(1) und (2) don't care about software but about data:

- ▶ For (1), data is object & product of their research.
- ▶ For (2), data is their *raison d'être*.

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

1. **End Users:** mostly scholars  
(philologists, linguists, historians, . . .)
2. **Content Providers:** mostly institutions  
(archives, libraries, commercial publishers),  
also scholars publishing in TextGrid
3. **Software Developers:**  
build, maintain & enhance the platform

(1) und (2) don't care about software but about data:

- ▶ For (1), data is object & product of their research.
- ▶ For (2), data is their *raison d'être*.

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Data Requirements

- ▶ long term stability of
  - ▶ the actual data (accessibility and revisioning)
  - ▶ identifiers
- ▶ virtual corpus, access transparent w.r.t. the actual storage location
- ▶ easy & transparent interfacing with content providers not fully integrated in TextGrid
- ▶ content providers keep control over their data (incl. IP), make final authorization decisions

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Data Requirements

- ▶ long term stability of
  - ▶ the actual data (accessibility and revisioning)
  - ▶ identifiers
- ▶ virtual corpus, access transparent w.r.t. the actual storage location
- ▶ easy & transparent interfacing with content providers not fully integrated in TextGrid
- ▶ content providers keep control over their data (incl. IP), make final authorization decisions

**Typical Cross-Corpus Query:** Give the context of all instances where any author of the 18th century uses the word *Trost* (consolation).

**Therefore:** Text retrieval engine must be able to

- ▶ tell when a text was originally written (metadata)
- ▶ differentiate between text from the author and additions by, e. g., the editor (text markup)

## Data Formats (2/3)

⇒ cross-corpus queries require  
standardized metadata & markup schemata.

### **Naïve approach:** Require

- ▶ Dublin Core metadata
- ▶ markup according to the *Text Encoding Initiative's* (TEI) guidelines P5.

### **But:**

- ▶ External repositories already have their own metadata format.
- ▶ Markup schemata (necessarily) vary from edition to edition, dictionary to dictionary, even if they follow TEI P5.

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

## Data Formats (2/3)

⇒ cross-corpus queries require  
standardized metadata & markup schemata.

### **Naïve approach:** Require

- ▶ Dublin Core metadata
- ▶ markup according to the *Text Encoding Initiative's* (TEI) guidelines P5.

### **But:**

- ▶ External repositories already have their own metadata format.
- ▶ Markup schemata (necessarily) vary from edition to edition, dictionary to dictionary, even if they follow TEI P5.

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

## TextGrid approach:

- ▶ Define *baseline encodings* (BE) for the most important genres.
- ▶ BEs are reduced subsets of TEI P5; the tags have very specific semantics and constraints.
- ▶ Users provide *adaptors* from the project specific markup to the BE. (In general, adaptors perform lossy transformations!)
- ▶ TextGrid stores data twice, once in project specific markup, once in BE.

⇒ Both cross-corpus and project specific queries possible.

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

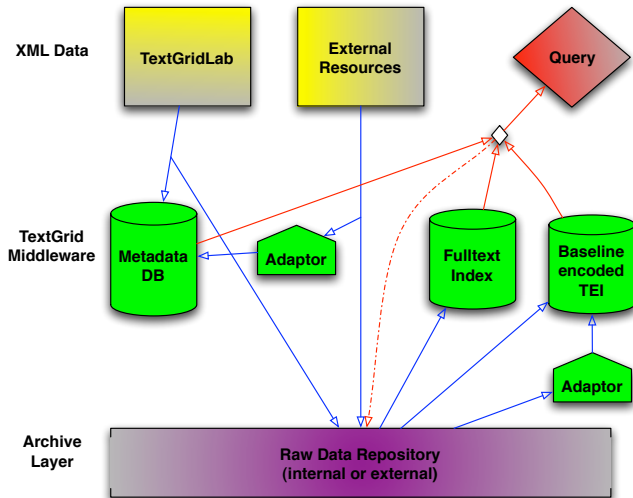
TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# TextGrid Data Storage Architecture



- ▶ High degree of markup variability  
⇒ scholar can choose markup granularity and expressiveness most apt for each project, respectively
- ▶ BE preserves common ground required for querying and retrieval of documents  
⇒ allows the eHumanities DE to thrive

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

- ▶ High degree of markup variability  
⇒ scholar can choose markup granularity and expressiveness most apt for each project, respectively
- ▶ BE preserves common ground required for querying and retrieval of documents  
⇒ allows the eHumanities DE to thrive

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

Yet another description of DEs:

*The digital ecosystem is defined as an open, loosely coupled, domain clustered, demand-driven, self-organising collaborative environment, where each species is proactive and responsive for its own benefit or profit.  
[DEST 2008 CFP]*

- ▶ Textual data is neither proactive nor responsive, so it cannot be a DE species.
- ▶ Textual Data can only be an *abiotic* part of the eHumanities DE

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Abiotic Ecosystem Elements

- ▶ one of the key factors determining how an ecosystem develops
- ▶ soil structure, climate, etc. typically long-term stable
- ▶ minor changes can disrupt the ecosystem



Source: private photos

⇒ Scholars “shape” the environment of the eHumanities DE by linking and editing resources and generating new ones.

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

Other digital ecosystems have “abiotic factors” as well:

- ▶ **geo spatial information:** in, e. g., DE with localization based services
- ▶ **meteorological information:** weather data going back 100 years and more precious to meteorologists

**Typically:** The abiotic factors are very stable resources and vice versa.

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Overview

Background: Textual Scholarship

Background:  
Textual  
Scholarship

The TextGrid Project

The TextGrid  
Project

The TextGrid Ecosystem

The TextGrid  
Ecosystem

TextGrid's Species

TextGrid's Species

Data in TextGrid

Data in TextGrid

TextGrid's Abiotic Environment

TextGrid's Abiotic  
Environment

ROA and DEs

ROA and DEs

## Resource Oriented Architecture:

- ▶ resource
- ▶ resource name (URI)
- ▶ resource representation
- ▶ resource links
- ▶ statelessness
- ▶ addressability
- ▶ uniform (HTTP) interface
- ▶ connectedness

([Richardson, Ruby, 2007], expanding on [Fielding, 2000])

## Examples (more or less strictly RESTful):

- ▶ Google's GData API
- ▶ Amazon's Simple Storage Server
- ▶ many social networks like BibSonomy, del.icio.us, etc.

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# RESTful TextGrid Data

<b>ROA</b>	<b>TextGrid Data</b>
statelessness	text documents, no application state
stable URIs	requires stable PSI mechanism (locational transparency!)
multiple representations	project & baseline encoding
stable links	crossreferences in markup

The “abiotic” elements of DEs typically fit well into the ROA paradigm!

In TextGrid, resource registries are geared toward use by human end users.

⇒ Free text descriptions, definitions etc. often more comprehensible and more useful than (over-)formal specifications.

⇒ Couple registry entries (both class descriptions and concrete instances) with Web / Wiki pages  
(cf. *subject indicator* in the Topic Maps Data Model)

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Federation of eHumanities Resource Registries

TextGrid plans for cooperation and integration:

- ▶ Eclipse RCP allows addition of GUI plug-ins.
- ▶ Most TextGrid services may be called by external parties and vice versa.
- ▶ Various levels of integration of archives and data repositories

⇒ Resources must be discoverable across projects and institutions!

**But:**

- ▶ Everyone wants to keep control over their services and (meta-)data.
- ▶ Centralized registry will soon become stale.

⇒ Employ P2P-model for resource registry federation!

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Federation of eHumanities Resource Registries

TextGrid plans for cooperation and integration:

- ▶ Eclipse RCP allows addition of GUI plug-ins.
- ▶ Most TextGrid services may be called by external parties and vice versa.
- ▶ Various levels of integration of archives and data repositories

⇒ Resources must be discoverable across projects and institutions!

**But:**

- ▶ Everyone wants to keep control over their services and (meta-)data.
- ▶ Centralized registry will soon become stale.

⇒ Employ P2P-model for resource registry federation!

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# Theses on Grids, SOAs and ROAs

- ▶ GT4 technology stack is geared towards science, it is only a limited match for humanities requirements
- ▶ Grid will move from a technology stack to a more flexible, but domain-specific model of a digital ecosystem of services and resources
- ▶ The eHumanities DE will be built along combined SOA and ROA principles and centre around standards-compliant, lightweight, federated semantic registries of services and resources (in hype-speak Web 3.0)
- ▶ ... and will evolve in the next decades in manners utterly unforeseen now, but certainly something completely different

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

# A Few Ecosystem Research Challenges

## Organizational interoperability:

- ▶ Ensure longterm viability of TextGrid, especially also for the commercial partners
- ▶ Develop long-term data strategy
- ▶ Expand trust and security infrastructure
- ▶ Establish loosely-coupled confederation between partners of very different size, technology prowess and interests beyond the TextGrid project proper

# A Few Ecosystem Research Challenges

## Semantic interoperability:

- ▶ Federate diverse dictionaries as valuable semantic resources and informal ontologies
- ▶ Agree on integration mechanisms with library metadata
- ▶ Use Topic Maps for the semantic description and location of services

Background:  
Textual  
Scholarship

The TextGrid  
Project

The TextGrid  
Ecosystem

TextGrid's Species

Data in TextGrid

TextGrid's Abiotic  
Environment

ROA and DEs

Every habitat influences its species / *The medium is the message* (McLuhan) → Possible bias for research that can easily be done with the tools and contents available