



## 1. TextGrid Data Resources

- ▶ data-centricity of eHumanities
- ▶ approach to the eHumanities data requirements

## 1. TextGrid Data Resources

- ▶ data-centricity of eHumanities
- ▶ approach to the eHumanities data requirements

## 2. Data in the Digital Ecosystem Metaphor

- ▶ resources as abiotic factors in digital ecosystems

## 1. TextGrid Data Resources

- ▶ data-centricity of eHumanities
- ▶ approach to the eHumanities data requirements

## 2. Data in the Digital Ecosystem Metaphor

- ▶ resources as abiotic factors in digital ecosystems

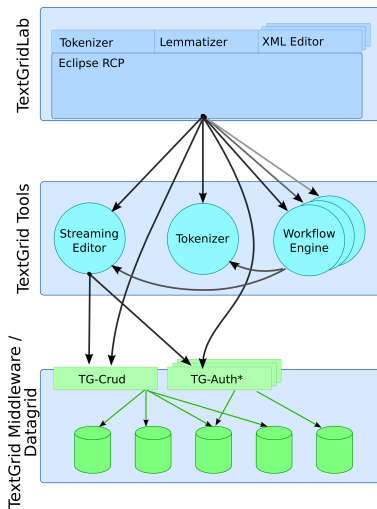
## 3. Resource Oriented Architectures and DEs

- ▶ abiotic factors fit well into ROA paradigm

# The TextGrid Project

- ▶ 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
- ▶ creates a modular, grid-based platform for collaborative textual editing
- ▶ emphasis on critical editions and dictionaries
  - ▶ digitize manuscripts and printed texts
  - ▶ augment texts with information about the writer, linguistic data, references, editorial comments etc.
  - ▶ collate text variants
  - ▶ ...

# Bird's Eye View on the TextGrid Architecture



TextGrid Data  
Resources

Data in the DE  
Metaphor

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. **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 whole purpose.

- ▶ long term stability of
  - ▶ the actual data (accessibility and revisioning)
  - ▶ identifiers
- ▶ access transparent w.r.t. the actual storage location

- ▶ long term stability of
  - ▶ the actual data (accessibility and revisioning)
  - ▶ identifiers
- ▶ 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 17th century uses the word *Natur* (nature).

**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.

## 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.

TextGrid Data  
Resources

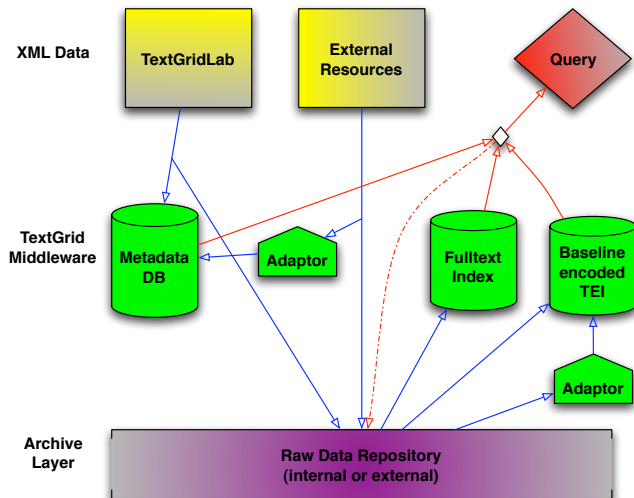
Data in the DE  
Metaphor

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.

# TextGrid Data Storage Architecture



- ▶ High degree of markup variability  
⇒ scholar can choose granularity and expressiveness  
most apt for each project, respectively

- ▶ High degree of markup variability
  - ⇒ scholar can choose 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

*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



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.

## 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.

<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 eHumanities, data is at least as important as software.
- ▶ In the DE metaphor, data belongs to the “abiotic elements” of the environment, does not act for its own benefit or profit.
- ▶ To some degree, the abiotic environment may be “shaped” by the actors, and vice versa.
- ▶ “Abiotics elements” fit well into the ROA paradigm.