Digital Humanities 2012 – Poster Submission
16-22 July 2012, Hamburg

Title: TextGrid Repository – Supporting the Data Curation Needs of Humanities Researchers

Authors: Felix Lohmeier (Saxon State and University Library Dresden), Kathleen M. Smith, Sibylle Söring and Ubbo Veentjer (State and University Library, Göttingen, Germany)

Abstract:
This poster will show how the TextGrid Repository assists researchers in the curation of their data and how they can make the data available with TextGrid tools to foster scientific re-use. The increasing importance of digitally-aided research methods has caused exponential growth in the creation of research data. New research methods and collaborative ways of using data sets require sophisticated research infrastructures to support researchers in the Digital Humanities and to enable the re-use of existing data. Data curation must be included in the planning stage as a fundamental requirement for all projects dealing with sustainable data. Therefore, this poster will present an overview of the technical infrastructure and the applicability of the TextGrid Repository for humanities researchers.

The TextGrid Virtual Research Environment (VRE), funded by the German Federal Ministry of Education and Research,[1] provides tools, data and services in one integrated interface and supports the long-term archiving and management of research data. It provides a platform for researchers in the Arts and Humanities to curate their data that reflects universally-recognized best practices and standards. TextGrid consists of two main components: the TextGrid Laboratory (TextGridLab), the entry point to the VRE, and the TextGrid Repository (TextGridRep), a long-term humanities data archive. To preserve and maintain research data and ensure its long-term viability, current research practices in all stages of the research lifecycle must be supported. Therefore, the TextGridLab provides common functionalities in a sustainable environment to facilitate the re-use of data, services, and tools, and the TextGridRep enables researchers to publish and share their data in a way that supports long-term availability and re-usability. Rather than acquiring the technical knowledge necessary for data curation themselves, researchers can make use of services and guidelines for long-term data accessibility and sustainability during the initial planning stages of their projects through the TextGrid VRE.

After five years of research and development, TextGrid released a stable, operational version 1.0 in July 2011 and will release a version 2.0 in May 2012.[2] The value of this project is demonstrated by the fact that there are already eight long-term research groups actively using the TextGrid virtual research environment for the creation of scholarly editions, for the analysis of humanities research data, as a basis for the development of project-specific tools for specialized analysis and visualization, and for long-term digital archiving and facilitating world-wide access to research data for the scientific community. (In addition to these established research projects, as of February 2012 there were concrete requests from 18 additional research groups about how TextGrid can be integrated

[1] The initial funding phase by the German Federal Ministry of Education and Research (BMBF) lasted from February 2006 to May 2009 (BMBF reference number 07TG01A-H). The second funding phase covered the period from 1 June 2009 to 31 May 2012 (BMBF reference number: 01UG0901A).

[2] TextGrid v1.0: http://www.textgrid.de/1-0.html

into their projects.) These projects deal with large amounts of humanities data that require specialized tools to reflect individualized requirements. To name a few examples:

- The project Blumenbach-Online (State and University Library, Göttingen) is producing an online resource providing access to the writings and collections of the German physician and anthropologist Johann Friedrich Blumenbach (1752-1840), in addition to secondary literature resources.[3]
- The project Hybrid-Edition von Theodor Fontanes Notizbüchern (University of Göttingen) is creating a critical annotated edition of 67 notebooks from the writer Theodor Fontane.[4]
- The Virtuelles Skriptorium St. Matthias (University of Trier / City Library Trier / Technical University Darmstadt) is developing an virtual reconstruction of the medieval manuscript collection of St. Matthias.[5]
- The Deutsches Literaturarchiv Marbach is creating an online edition of the letters of Ernst Kantorowicz.[6]

These projects create significant amounts of data during the research process that require curation. This poster will show how the TextGridRep assists researchers in the curation of their data and in ensuring persistent access to data with TextGrid tools to support scientific re-use.

The first section of the poster will give an overview of the technical functionalities and infrastructure of the TextGridRep, which has been fully operational since July 2011. The TextGridRep provides a repository infrastructure based on grid technology. Researchers can decide how and with whom their data will be shared by using the detailed rights management module. Findings and research data can be published directly from the TextGridLab in the repository via a publishing process that guides researchers in preparing the data for long-term accessibility. The middleware consists of various components for handling files in the data grid, rights management in a role-based access control-enabled database, metadata in an XML database, and relations in a Resource Description Framework (RDF) triple store. On a basic level, TextGrid will offer bitstream preservation with redundant grid storage and tape backup for 10 years (as recommended in the guidelines of the German Research Foundation).[7] TextGrid developed its own metadata schema, especially suitable for digital editions, that supports different layers in its object model (item, work, edition, and collection).

When researchers publish their research data via the TextGridLab in the repository, the metadata provided will be automatically validated. The system validates against the TextGrid object model and checks if obligatory metadata fields like rights owner and license are well defined. In the next step,

---

[3] Blumenbach-Online: http://www.blumenbach-online.de

[4] Hybrid-Edition von Theodor Fontanes Notizbüchern: http://www.uni-goettingen.de/de/303691.html

[5] Virtuelles Skriptorium St. Matthias: http://kompetenzzentrum.uni-trier.de/projekte/kernprojekte/virtuelles-skriptorium

[6] Edition Ernst H. Kantorowicz: http://www.dla-marbach.de/dla/entwicklung/projekte/edition_ernst_h_kantorowicz/index.html

[7] *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft"; Denkschrift = Proposals for safeguarding good scientific practice. Deutsche Forschungsgemeinschaft*. (Weinheim: Wiley-VCH, 1998). ‚Recommendation 7' (Pp. 55-56) : http://www.forschungsdaten.org/informationen-fur-ihr-forschungvorhaben/informationen-fur-ihr-forschungsvorhaben/#DFG

persistent identifiers are allocated by using a reliable handle service that is provided by the data centre in Göttingen, GWDG, which is a main developing partner in the European Persistent Identifier Consortium as well as the computer centre for the Max Planck Society. As part of the publishing process, the data will be frozen and moved to a storage cluster used for long-term preservation. If researchers want to update their data, they can copy it to their workspace, correct or further annotate the data, and publish the data as a new revision that is linked to the old revision. Both revisions will be available but the newer one will be more prominent in search results. The grid storage for the humanities and all connected resources are maintained together with those from the other academic disciplines at the common Grid Resource Centre in Göttingen (which has allotted 275 terabytes for the humanities).

The second section of the poster will show how researchers can make their data available with the TextGrid Repository. There are currently three different ways for research groups to enable access to their data in the repository.

1) All published data is available via the TextGridRep portal, which is already in place. It enables rapid searching with both simple and advanced search capabilities, in addition to the option of browsing repository content, across public research data with fulltext and metadata indexes.[8] Complex editions can be browsed according to the TextGrid object model (see above) and predefined XSLT stylesheets provide HTML representations of TEI documents. Links between texts and images, such as those created in the TextGridLab interface, will be displayed in a synoptical view in a future version of the portal.

2) Research groups who create a digital edition often want to present their data in their own portal with specific graphics, labels, and predefined browse and search options. Therefore, an open REST interface for individual portal solutions is provided so that research groups may provide specific elaborated access to their research collections with common technologies like Javascript, CSS, HTML.

3) Research groups who want to provide complex customized visualizations and complex project-specific search queries for their digital editions often use their own database for their project that is not connected to any long-term archiving solutions. We are developing a straightforward and easy way to sync the data stored in the TextGridRep (for long-term access) with a project-specific XML database (for the project-specific representation of the digital edition). A prototype is already in place that enables users to publish data from the TextGridRep to any eXist database with drag & drop functionality. Users can continuously test the representation of their XML data (e.g., TEI) in their own environment while they are still working on the digital edition with TextGrid. This allows research projects to annotate their data and develop the representation with XSLT and XQuery scripts at the same time. They can also easily publish new revisions of their data through TextGrid in the TextGridRep as well as in their own environments. TextGrid will provide a TextGrid-specific XQuery-module for the eXist XML database via the newly announced eXist AppRepository.[9] Users who install this module will be able to enhance their eXist environment to interface with the TextGrid metadata schema. To provide an out-of-the-box solution for the representation of a digital edition, TextGrid

[8] TextGridRep Portal: http://www.textgridrep.de

[9] eXist AppRepository: http://atomic.exist-db.org/blogs/eXist/AppRepository

collaborates with the TELOTA working group at the Berlin-Brandenburg Academy of Science and Humanities to use their publishing framework Scalable Architecture for Digital Editions (SADE).[10]

XML technologies like XQuery and XSLT support the representation of digital editions using a common standard that promotes long-term reusability of and reliable access to research data. Therefore TextGrid facilitates the publication of digital editions in ways that are both easy to use and encourage the use of established best practices.

[10] Scalable Architecture for Digital Editions (SADE): http://www.bbaw.de/telota/sade