TextGrid - Virtuelle Forschungsumgebung, Forschungsdaten und Langzeitarchivierung

Dr. Heike Neuroth, SUB Göttingen

The TextGrid research group, a consortium of 10 research institutions in Germany, is developing a Virtual Research Environment (VRE) for researchers in the arts and humanities that provides services and tools for the analysis of text data and supports the curation of research data by means of grid technology. Libraries and data centres as well as universities and research institutions are collaborating in a community-driven process that is funded by the German Ministry for Education and Research (BMBF). Initially consisting of two academic communities in 2006, textual philology and linguistics, the TextGrid project was joined by art history, classical philology, and musicology in 2009.[1] As part of the German grid initiative D-Grid[2], TextGrid maintains a common grid resource centre in Göttingen[3] together with grid projects in physics, medicine, astronomy, and climate research. The TextGrid VRE consists of two main components: the ***TextGrid Lab(oratory)***, which serves as the entry point to the virtual research environment, and the ***TextGrid Rep(ository)***, which is a long-term humanities data archive ensuring sustainability, interoperability and long-term access to research data.

To support all stages of the research lifecycle, preserve and maintain research data and ensure its long-term usefulness, existing research practices must be supported.

Therefore the TextGridLab provides common functionalities in a sustainable environment to intensify re-use of data, tools and services and the TextGridRep enables researchers to publish and share their data in a way that supports long-term availability and reusability.

TextGridRep is ensuring sustainability, interoperability and longterm access to research data. Researchers can decide how and with whom their data will be shared by using the detailed rights management (based on RBAC). They can publish their findings and research data from the TextGridLab in the repository, and archives and other institutions can ingest enormous amounts of data into the repository via a special interface that uses koLibRI[4], which supports for example automatic metadata validation. On a basic level TextGrid will offer bitstream preservation with redundant storage and tape backup for 10 years (as recommended in the guidelines of the German Research Foundation[5]). Long-term bitstream preservation and higher security levels such as further distributed storage on multiple sites will be available at greater cost.

All data will be addressable via persistent identifiers that TextGrid will allocate by using a reliable handle service that is provided by the local data centre, which is a main developing partner in EPIC.[6] A portal solution will enable rapid searching across public research data via a second instance of the search utility without connection to the rights management. An open REST interface for individual portal solutions will be provided.

---

[1] TextGrid (http://www.textgrid.de/en.html)

[2] The D-Grid Initiative (http://www.d-grid.de/)

[3] GoeGrid (http://www.goegrid.de/)

[4] kopal Library for Retrieval and Ingest (http://kopal.langzeitarchivierung.de/index_koLibRI.php.de)

[5] Proposals for Safeguarding Good Scientific Practice (http://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/self_regulation_98.pdf)

[6] The European Persistent Identifier Consortium (http://www.pidconsortium.eu)

Higher-value long-term preservation services will be provided in 2011 by making use of developments within the WissGrid project,[7] which is also part of the German Grid Initiative and consists of five academic communities from the natural sciences and TextGrid from the humanities. The project is developing a service framework that fulfils more sophisticated long-term preservation needs like a provenance service, metadata extraction, format validation and conversion. Guidelines will be adapted to the specific needs of the humanities and be incorporated in the virtual research environment. The grid storage and all connected resources will be maintained together with those from the other academic disciplines at the common grid resource centre in Göttingen (e.g., 275 terabytes for the humanities). There are plans for the migration of the current repository infrastructure to Fedora and iRODS or dCache/SRM that will probably be implemented in 2011.

[7] WissGrid: Grid for Science (http://www.wissgrid.de/index_en.html)