The International Journal of Digital Curation
Issue 2, Volume 6 | 2011

TextGrid – Virtual Research Environment for the Humanities

Heike Neuroth & Felix Lohmeier,
State and University Library, Göttingen, Germany

Kathleen Marie Smith,
University of Illinois, USA

Abstract

The TextGrid research group, a consortium of 10 research institutions in Germany, is developing a virtual research environment 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. The TextGrid virtual research environment consists of two main components: the TextGrid Laboratory (TextGridLab), which serves as the entry point to the virtual research environment, and the TextGrid Repository (TextGridRep), 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 re-usability.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The TextGrid research group, a consortium of 10 research institutions in Germany, is developing a virtual research environment 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.[1] Libraries and data centres as well as universities and research institutions are collaborating in a community-driven process that is funded by the German Federal Ministry of Education and Research.[2] As part of the German Grid Initiative (D-Grid), TextGrid maintains a common Grid Resource Centre in Göttingen together with grid projects in physics, medicine, astronomy and climate research.[3,4]

Initially consisting of two academic communities − textual philology and linguistics −in 2006, the TextGrid project was joined by art history, classical philology, and musicology in 2009. These research communities were interested in making use of the capabilities offered by a virtual research environment in order to support their research processes. The TextGrid infrastructure was designed and developed so that new communities can be easily integrated and also take part in the development process (TextGrid, 2008).

The TextGrid virtual research environment consists of two main components: the TextGrid Laboratory (TextGridLab), which serves as the entry point to the virtual research environment, and the TextGrid Repository (TextGridRep), 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 process, preserve and maintain research data, and ensure its long-term usefulness, existing research practices must be supported (Neuroth, Jannidis, Rapp & Lohmeier, 2009). 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 re-usability.

After five years of research and development, TextGrid will release a stable, operational Version 1.0 in July 2011.

The TextGridLab: User Interface and Entry Point

The TextGridLab tools and services scheduled for Version 1.0 include core workbench tools for handling data and metadata, navigation and search functionalities, rights management, and a workflow editor, as well as specialized tools and services to meet the needs of researchers. The core workbench enables collaborative research in a distributed, secure, flexible, and extensible environment. Research groups will be able to work with data stored on local computers or in the grid, and coordinate collaborations in one integrated platform. Specialized tools and services developed jointly in cooperation with the participating academic disciplines mentioned above are:

[1] TextGrid: http://www.textgrid.de/en.html.
[2] Federal Ministry of Education and Research: http://www.bmbf.de/en/.
[3] D-Grid: http://www.d-grid.de/index.php?id=1&L=1.
[4] Grid Resource Centre in Göttingen: http://www.goegrid.de/.

- The *XML editor*, with which users can switch easily between a more technical view with tags and attributes, and a structural view that is oriented towards standard text editing applications. A Unicode Character Table enables the user to search, copy, and insert symbols from the Unicode character set.
- The *Text-Image-Link* editor supports the XML Editor by linking text sequences with image sections in order to create files that contain text elements and topographic descriptions.
- The *Dictionary Search Tool* enables searching in a number of different dictionaries within the TextGrid virtual research environment. The dictionary network *Wörterbuchnetz* at the University of Trier has been integrated into the interface for this purpose.[5]
- A *Text Publisher Web* can be used to present project results and publications on a project website. TextGrid provides standard components for web publishing.

These tools will be available in a stable and operational Version 1.0 with full support and documentation.

Other tools are in development and scheduled for later release during the project term (ending May 2012):

- The *Note Editor* illustrates Music Encoding Initiative (MEI)-encoded scores, and displays them in a simplified format.[6] Unique features of the Note Editor are located in the editorial field, such as the visualization of variants.
- The *Gloss Editor* facilitates: a) the description of specific text structures in gloss comments, b) the synoptical presentation of different gloss comments regarding the same text, c) the synoptical presentation of gloss comments and related digital manuscripts.
- The *Digilib Tool* will be integrated into the virtual research environment to enhance the retrieval and annotation of image data.[7]
- *LEXUS* and *COSMAS*, which are relevant databases for linguistics, will be integrated into TextGrid.[8]
- The *CollateX Tool* will be integrated to compare two or more files encoded in XML (or Text Encoding initiative [TEI]), and annotate any differences in TEI format.[9,10]
- A publishing module called *XML Print* will be integrated into TextGrid to allow printing of scientific texts with complex typesetting layout requirements based on XML data, with a particular view towards the specifications of critical editions and dictionaries.[11]

[5] Wörterbuchnetz: http://www.woerterbuchnetz.de/ (German).
[6] MEI: http://music-encoding.org/.
[7] Digilib: http://digilib.berlios.de/.
[8] LEXUS: http://www.lat-mpi.eu/tools/lexus/manual/ch01s01.html;
COSMAS: http://www.ids-mannheim.de/cosmas2/.
[9] CollateX: http://collatex.sourceforge.net.
[10] TEI: http://www.tei-c.org/index.xml.
[11] XML Print research project: http://www.fh-worms.de/index.php?id=4616&L=1.

- The *Bibliography Tool* can import bibliographical data from existing data inventories, and can capture, process, and administer bibliographies. Users can also export bibliographical data into certain standard formats (e.g., TEI or the Metadata Object Description Schema [MODS]).[12]
- The *OCRopus Service* for automatic character recognition will be enhanced and integrated. This service advances research in the text-based digital humanities by enabling the optical character recognition of Fraktur script (blackletter typeface) in large historically-diverse amounts of text.[13]
- The *Text-Text-Link Editor* is an input assistant for links into XML data, and it connects elements of the Dictionary Search Tool and XML Editor. Users have the opportunity to input links to user-defined elements within TextGrid documents into TEI data.

The advantage of combining the tools and services for text-based research with a data curation system means that users will be able to save their data directly in a safe place (the grid storage), and since the tools and services will be updated continuously, long-term data compatibility and accessibility will be maintained. Rather than having to acquire the technical knowledge necessary for data curation themselves, researchers will be able to build upon services and guidelines for long-term data accessibility and sustainability during the initial planning stages of their projects by using the TextGrid virtual research environment. As described in the JISC Virtual Research Environment Collaborative Landscape Study, there are many potential benefits for the research process, including the flexibility and ease of use offered by using a standard platform that has already been developed and tested, as opposed to developing individual solutions on a case-by-case basis (Carusi & Reimer, 2010). The TextGrid project also aims to provide extensive documentation in both English and German to enable users inside and outside of Germany to use the interface.

Following the philosophy: "release early, release often," TextGrid published its first beta version in January 2009, and has regularly added updates and incorporated new features into re-releases every two to three months for public beta testing.[14] Owing to the use of the Java programming language it is directly available after downloading without installation procedures, and can also be accessed from a Universal Serial Bus (USB) flash drive. The TextGridLab is being developed continuously to reflect the needs of its academic user communities. To faciliate feedback and user participation, there have been specialized TextGrid workshops that address the needs and interests of specific research groups; the most recent workshop in June 2010 focused on the requirements of the Blumenbach-online and Archaeo18 research projects.[15,16] Beta testing has also produced useful feedback. The total number of registered testers for the beta version was about 670 as of June 2011. At this point, there were more than 500 users in Germany, about 110 in other European countries and an additional 60 worldwide. In order to adequately address issues that arise during testing, there is a public bugtracking system based in the JIRA issue-tracking software.[17]

[12] MODS: http://www.loc.gov/standards/mods/registry.php.
[13] OCRopus: http://code.google.com/p/ocropus/.
[14] The latest TextGridLab beta and the changelog is available at: http://www.textgrid.de/beta.html.
[15] Archaeo18: http://www.uni-goettingen.de/en/3240.html?cid=3570.
[16] Blumenbach-online: http://www.blumenbach-online.de/index.php?id=2&L=1.
[17] TextGrid bugtracking system: https://develop.sub.uni-goettingen.de/jira/browse/TG (German).

The TextGridRep: A Grid-Based Repository Infrastructure

The second main component, the TextGridRep, is ensuring sustainability, interoperability, and long-term access to research data by providing 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 (based on role-based access control).[18] 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 time accessability.

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. TextGrid is not a closed system but rather an open platform that enables scholars to adapt the environment to their needs. Owing to the modular structure it is easy to integrate external web services (e.g. the dictionary network "Wörterbuchnetz" at the University of Trier (see footnote 5), and the layered architecture also allows access to the services via other graphical user interfaces in addition to the Eclipse-based TextGridLab software currently provided by TextGrid (Figure 1).[19]

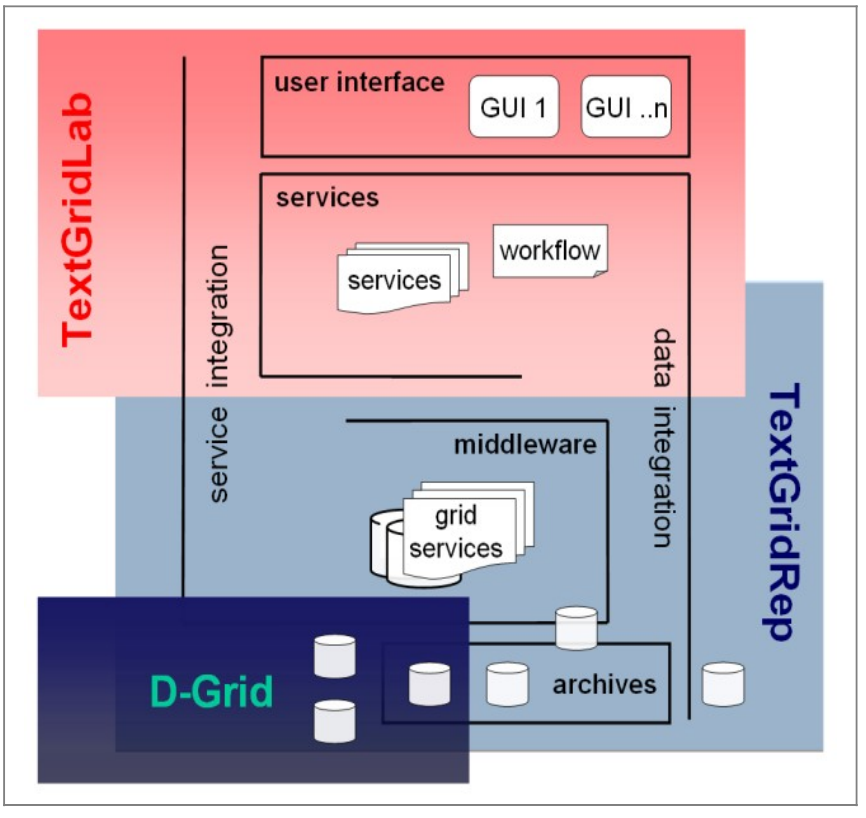

Figure 1. The TextGrid architecture. GUI: Graphical User Interface.

[18] Role-based access control: http://csrc.nist.gov/groups/SNS/rbac/.
[19] Eclipse Rich Client Platform: http://wiki.eclipse.org/index.php/Rich_Client_Platform.

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).[20] Long-term bitstream preservation and higher security levels such as further distributed storage on multiple sites will be available at greater cost in the future. When researchers publish their research data via the TextGridLab in the repository, the metadata provided will be automatically validated. 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, GWDG, which is a main developing partner in the European Persistent Identifier Consortium as well as the computer centre for the Max Planck Society.[21,22,23] As part of the publishing process the data will be frozen and moved to a static storage cluster used for long-term preservation. A portal solution will enable rapid searching across public research data via a second search index without connection to the rights management service, *TG-auth**. An open Representational State Transfer interface for individual portal solutions will be provided, so research groups may provide specific elaborated access to their research collections. Archives and other institutions can ingest enormous amounts of data into the repository via a special interface that uses the kopal Library for Retrieval and Ingest software, which supports automatic metadata validation, for example (Figure 2).[24]

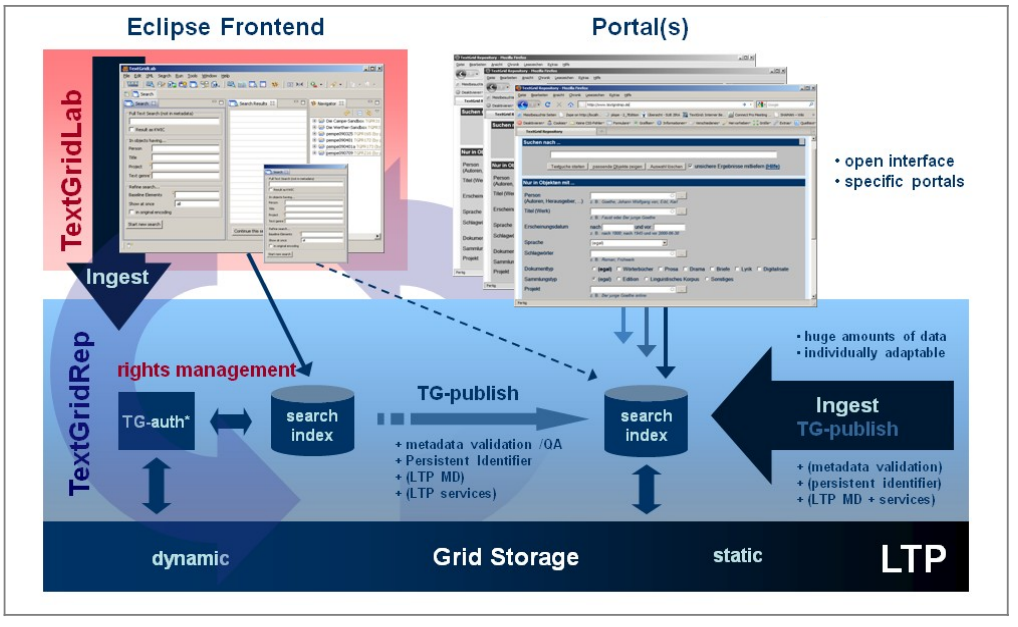

Figure 2. Functionalities of the TextGrid Repository. It provides capabilities for Long-Term Preservation (LTP) and LTP-Metadata (LTP-MD).

[20] Proposals for Safeguarding Good Scientific Practice: http://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/self_regulation_98.pdf.
[21] Local computer centre GWDG: http://www.gwdg.de/index.php?L=1.
[22] The European Persistent Identifier Consortium: http://www.pidconsortium.eu.
[23] The Max Planck Scoiety: http://www.mpg.de/en.
[24] kopal Library for Retrieval and Ingest: http://kopal.langzeitarchivierung.de/index_koLibRI.php.en.

Higher-value long-term preservation services will be provided in 2011 by making use of developments within the WissGrid project, which is also part of D-Grid, and consists of five academic communities from the natural sciences, and TextGrid from the humanities.[25] The project is developing a service framework that fulfils more sophisticated long-term preservation needs such as a provenance service, metadata extraction, format validation, and conversion. Guidelines will be adapted to the specific needs of the humanities and incorporated into the virtual research environment. The grid storage for the humanities 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 Flexible Extensible Digital Object Repository Architecture (Fedora) and integrated Rule-oriented Data System (iRODS) or dCache/Storage Resource Management (SRM) that will be implemented after the release of Version 1.0 in 2011.[26,27,28] Fedora and iRODS, and dCache/SRM, are widely-used repository and storage software systems with large developer communites. By using more standardized components the TextGrid Repository will be more sustainable.

Our expectation is that diversity and decentralization will increase with the number of repository systems (Pempe & Aschenbrenner, 2010). TextGrid was planned as an open repository environment and incorporates an open storage interface and federation patterns that facilitate cooperation and interconnectability (Aschenbrenner, 2010). On a national level, TextGrid is working on a federated repository infrastructure for the humanities together with the eSciDoc open source e-research environment.[29]

Sustainability: The TextGrid Organizational Model

TextGrid is part of D-Grid, which works to provide basic, sustainable resources and services as a foundation for other e-Science projects. The D-Grid initiative began as a collaboration with 6 academic community grid projects and has grown to include over 20 grid projects from a variety of fields (the natural sciences, engineering, business, etc.), along with several horizontal projects to deliver generic services for all community grid projects, for example, a project to facilitate use of short-lived credentials (SLCs) in portal-based grids (Gap-SLC Project).[30]

One central goal of the remaining project time is to develop an organizational model to ensure the long-term sustainability of TextGrid after its current funding source ends in May 2012. Even if costs are reduced to a minimum by economies of scale, it will still be necessary to fund the storage and sophisticated data curation services when the infrastructure is in operational mode. By the end of the project funding period, our objective is to establish TextGrid as a community-driven virtual research environment and to have a sustainable business model based on data curation within the TextGridRep that is integrated with other subject-specific virtual research environments (such as climate research and astronomy) within the framework of

[25] WissGrid: Grid for Science: http://www.wissgrid.de/index_en.html.
[26] Fedora: http://www.fedora-commons.org/about.
[27] iRODS: https://www.irods.org/index.php/IRODS:Data_Grids,_Digital_Libraries,_Persistent_Archives,_and_Real-time_Data_Systems.
[28] dCache: http://www.dcache.org/index.shtml.
[29] eSciDoc: http://www.escidoc.org.
[30] Gap-SLC: http://gap-slc.awi.de/ (German).

WissGrid. A joint venture of the academic community grids that participated in the initial D-Grid call for projects, WissGrid consists of high-energy physics, astronomy, medicine, climate research, and the arts and humanities. It aims to develop an operational model for academic grid users, create blueprints for new academic community grids (e.g. social sciences), and foster long-term storage for research data. By using the storage-backend and long-term-preservation services from WissGrid, TextGrid can take advantage of robust, secure technology based on accepted standards. Grid-based infrastructures are already widely accepted in the natural sciences for data storage and processing, and are being adopted in the arts and humanities now (Carusi & Reimer, 2010). There is a strong commitment of the high-performance computer centres in Germany to the grid technology (Neuroth, Kerzel & Gentzsch, 2007). One other important aspect of sustainability is, naturally, cooperation with new research projects which also provide new sources of funding. TextGrid is, for example, currently developing collaborations with the research projects Blumenbach-Online and Archaeo18.

- The project Blumenbach-Online will produce 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.
- The Archaeo18 research group is working to compile multiple manuscripts of student notes of the lectures of the scholar and archaeologist Christian Gottlob Heyne (1729-1812) over a 40-year period (approximately 1769-1810) in order to reconstruct Heyne's lectures.

These research collaborations are important for developing specific user case studies for the TextGrid project, and their interest in TextGrid is perhaps the best indicator of TextGrid's feasibility in the long term.

Cooperation

In addition to the cooperation within the D-Grid and the integration of existing tools and services mentioned above, TextGrid is cooperating with other research infrastructures on international and European levels:

- The Coalition of Humanities and Arts Infrastructures and Networks is exploring the use of digital technologies in arts and humanities research in Europe.[31]
- The vision of the European Strategy Forum on Research Infrastructures project Digital Research Infrastructure for the Arts and Humanities (DARIAH) is to facilitate long-term access to, and use of, all European arts and humanities digital research data.[32] Therefore it will enhance and support digitally-enabled research across the humanities and arts by developing and maintaining an infrastructure in support of information and communication technology-based research practices.
- Project WisNetGrid ('Networks of Knowledge on the Grid') is working to link the information sources and administer knowledge.[33]

[31] Coalition of Humanities and Arts Infrastructures and Networks: http://www.arts-humanities.net/chain.
[32] Digital Research Infrastructure for the Arts and Humanities: http://www.dariah.eu.
[33] WisNetGrid: http://www.wisnetgrid.org/

- The TEXTvre project is working to implement TextGrid's experience into the community of practice in the UK.[34]

In its ambition to support all stages of the research lifecycle and build upon existing research practices, TextGrid works together with other research groups to provide interoperability. In Germany, TextGrid cooperates, for instance, with the *Forschungsnetzwerk und Datenbanksystem* humanities research and publication platform at the University of Trier, and with the International Tustep User Group (ITUG), which is an organization that supports the training and continuing education of Tustep users.[35,36]

Conclusions

TextGrid is the first large multi-year project in Germany dealing with the development of a research infrastructure and virtual research environment for the arts and humanities. As such, it is still a research project and a process: bringing arts and humanities researchers together with information technology specialists in order to develop a common language and culture will take time and commitment from those involved. To this end, some of the instruments implemented, such as the programming sprints, workshops with external researchers, and so on, have proven to be very useful.

Since researchers in the arts and humanities do not in general have a high level of expertise dealing with data exchange and re-use, not to mention data management and data curation, in this area there has been slower progress. One way of improving researcher involvment and encouraging interdisciplinary collaboration is through the use of hands-on practical education (Rings et al., 2010).

With TextGrid, our cooperative and collaborative method means that we are working together in new types of collaboration: a very large team in virtual and geographically-separated situations, above and beyond institutional borders.

Another challenge is to build awareness in funding agencies in the arts and humanities that these virtual research environments and research infrastructures will need to be supported financially on a long-term as well as a short-term basis, similar to systems such as the Large Hadron Collider at the European Organization for Nuclear Research in other fields.[37]

Acknowledgements

The joint research project TextGrid is part of D-Grid, and is funded by the German Federal Ministry of Education and Research for the period starting June 1, 2009 to May 31, 2012 (reference number: 01UG0901A).

[34] TEXTvre: http://textvre.cerch.kcl.ac.uk.
[35] Forschungsnetzwerk und Datenbanksystem: http://fud.uni-trier.de (German); http://www.interedition.eu/wiki/index.php/AssociatedProjects#FuD_-_Forschungsnetzwerk_und_Datenbanksystem (English langauge description).
[36] International Tustep User Group: http://www.itug.de (German); Tübinger System von Textverarbeitungsprogrammen: http://www.tustep.uni-tuebingen.de/tdv.html.
[37] Large Hadron Collider at the European Organization for Nuclear Research: http://public.web.cern.ch/public/.

References

Aschenbrenner, A. (2010). *Reference framework for distributed repositories – towards an open repository environment.* (Doctoral Dissertation, University of Göttingen, Germany, 2010). Retreived June 5, 2011, from http://webdoc.sub.gwdg.de/diss/2010/aschenbrenner/.

Carusi, A., & Reimer, T. (January 2010). *Virtual research environment collaborative landscape study - A JISC funded project*. Retrieved October 21, 2010, from http://www.jisc.ac.uk/media/documents/publications/vrelandscapereport.pdf.

Neuroth, H., Kerzel, M. & Gentzsch, W. (Ed.) (2007). *German grid initiative*. University of Göttingen, Germany. Retrieved October 31, 2010, from http://resolver.sub.uni-goettingen.de/purl/?webdoc-1521-2.

Neuroth, H., Jannidis, F., Rapp, A., & Lohmeier, F. (2009). Virtuelle Forschungsumgebungen für e-Humanities. Maßnahmen zur optimalen Unterstützung von Forschungsprozessen in den Geisteswissenschaften. *Bibliothek. Forschung und Praxis*, *33*, 161-169.

Pempe, W., & Aschenbrenner, A. (2010). *Development of a federated repository infrastructure for the arts and humanities in Germany* (Report 1.3.1). TextGrid; German Federal Ministry of Education and Research. (German; report available upon request at info@textgrid.de).

Rings, T., Aschenbrenner, A., Grabowski, J., Kalman T., Lauer, G., Meyer, J., Quadt, A., et al. (2010). An interdisciplinary practical course on the application of grid computing. In *Proceedings of the 1st Annual IEEE Engineering Education Conference – The Future of Global Learning in Engineering Education (EDUCON 2010)*, Madrid, 2010.

TextGrid. (2008). *TextGrid—Vernetzte Forschungsumgebung in den eHumanities: Nachtrag zu der am 27.06.2008 eingereichten Vorhabenbeschreibung*. Project Proposal. Retrieved October 31, 2010, from http://www.textgrid.de/fileadmin/TextGrid/div/090804_Nachtrag_oeffentlich.pdf