CoRDI 2025 Conference on Research Data Infrastructure

Humanities and Social Sciences

Making the Repository Programmable

The TextGrid Repository as a multi-layered Research Environment

Lukas Weimer*[1] https://orcid.org/0000-0001-6919-3646, José Calvo Tello[1] https://orcid.org/0000-0002-1129-5604, Stefan Buddenbohm[1] https://orcid.org/0000-0002-3469-6101, Daniel Kurzawe[1] https://orcid.org/0000-0001-5027-7313, and Ubbo Veentjer[1] https://orcid.org/0000-0002-9726-3135

[1] University of Göttingen, Göttingen State and University Library, Germany

*Correspondence: Lukas Weimer, lukas.weimer@sub.uni-goettingen.de

Abstract

The TextGrid Repository (TGR) is a dedicated research-data repository for the humanities and cultural studies that specializes in XML/TEI–encoded texts. Developed in the DFG-funded TextGrid project from 2006 to 2015, TGR was a pioneering infrastructure, as it embraced the TEI format—a de facto standard in digital humanities and an essential foundation for computational philology. Initially, TGR offered basic storage, download, archiving, and structured metadata for literary texts. Over time, it has evolved into a sophisticated research environment that transcends conventional archival functions.

To support advanced scholarly workflows, TGR integrates tools for automated text analysis and annotation—such as Voyant Tools [1], the Language Resource Switchboard [2], and the Annotation Sandbox [3]—with direct export capabilities, thereby lowering technical barriers and streamlining complex analyses. Its incorporation into the NFDI consortium Text+ ushers in a new era of modernization, component upgrades, and enhanced user engagement, opening TGR to emerging generations of researchers.

Contemporary literary and linguistic scholars demand virtual research environments that differ markedly from those of earlier years. Text-editing projects now emphasize rich presentation layers and custom transformations for reading and highlighting annotated data. Computational literary studies require straightforward access to plain text, programmatic interfaces, and libraries. Library-driven initiatives prioritize authority data integration. Some digital humanities inquiries hinge on author attributes—such as gender—while corpus linguistics projects center on detailed linguistic annotations.

TGR addresses these diverse requirements by unifying multiple services and access modalities. From the end user's vantage point, data can be retrieved via direct reading links, faceted search in the portal's graphical interface, persistent identifiers (PIDs), or programmable interfaces—including the Python client library tg_client [5]. Prospective data publishers receive expert guidance on metadata quality. In the Text+ context, TGR now offers new services—Notebook Actions [6], which provide a graphical import interface in Jupyter Notebooks, and tg_model [7], which generates the metadata documents required for data ingestion—alongside established tools (tg-crud [8] and tg_admin [9]) that handle repository maintenance and document management. Collectively, these enhancements simplify and accelerate data import and publication workflows.

A clear indicator of TGR's transformation is the surge in new projects over recent years, which has greatly enriched the repository's content. Whereas TGR once catered primarily to German studies, it now houses materials in over one hundred languages and multiple script systems (including Coptic, Cyrillic, Arabic, Hebrew, Amharic, Chinese, Japanese, Korean, and Armenian), reflecting the needs of a broad spectrum of disciplines.

In our presentation, we will demonstrate key new functionalities and illustrate how TGR's current multi-layered research environment departs from its original archival role. Special emphasis will be placed on the latest automated processes, which not only facilitate but actively promote computer–assisted analyses, all while ensuring the highest standards of metadata quality.

Keywords: TextGrid Repository, Digital Humanities, XML-TEI, Research Data Infrastructure, Automated Text Analysis

Resources

The material and resources (e.g. data, model, codes, documentation, videos) which underly, support or are closely related to your contribution deposited on a repository, please include a title, the respective DOI(s) and brief description. E.g.

- Software.
    - https://pypi.org/project/monapipe/ MONAPipe stands for "Modes of Narration and Attribution Pipeline". It provides natural-language-processing tools for German, implemented in Python/spaCy.
    - https://github.com/DARIAH-DE/eXanore JWT enabled implementation of annotatorjs.org Storage API as eXist-db library used by the DARIAH-DE Annotation Store. View, share and export your annotations with the AnnotationViewer.
    - https://gitlab.gwdg.de/textplus/textplus-io/nb-actions/ The notebook actions used in the import workflow.
    - https://gitlab.gwdg.de/textplus/textplus-io/tg_model To generate the TextGrid metadata files, the tool tg-model was developed, which can be used as a Python library and on the command line.
    - https://gitlab.gwdg.de/dariah-de/textgridrep/tgadmin A command line tool for managing your projects in the TextGrid repository without TextGridLab.
- Services.
    - https://textgridrep.de/?lang=en The TextGrid Repository.
    - https://dariah-de.pages.gwdg.de/textgridrep/textgrid-python-clients/docs/# TextGrid Python clients documentation.
    - https://textgridlab.org/doc/services/submodules/tg-crud/tgcrud-webapp/docs/ TG crud documentation.
- Resource.
    - https://textgridlab.org/doc/services/index.html The TextGrid Repository documentation.
- Website.
    - https://text-plus.org/en/ The Text+ portal. Text+ is the NFDI consortium for text- and language-based research data.

Author contributions

All authors of this submission have been involved in Writing – original draft (ID: 43ebbd94-98b4-42f1-866b-c930cef228ca) and Writing – review and editing (ID: d3aead86-f2a2-47f7-bb99-79de6421164d).

Competing interests

The authors declare that they have no competing interests.

Funding

The development of TGR has been made possible by various generous grants by the DFG (German Research Foundation/Deutsche Forschungsgemeinschaft) and the BMBF (Federal Ministry of Education and Research/Bundesministerium für Bildung und Forschung).

References

[1] Sinclair, Stéfan and Geoffrey Rockwell, 2016. Voyant Tools. Web. http://voyant-tools.org/.(26.04.2025)

[2] Claus Zinn, The Language Resource Switchboard. Computational Linguistics 44(4), pages 631-639, December 2018.

[3] https://github.com/DARIAH-DE/eXanore (26.04.2025)

[4] Buddenbohm, S., Calvo Tello, J., Funk, S. E., Klammer, R., Rißler-Pipka, N., Steckel, A., Veentjer, U., Weimer, L., & Dogaru, G. (2025). Fluffy Import: Preserving Humanities Research Data with the TextGrid Repository. Preprint. In TRANSFORMATIONS - A DARIAH Journal (Workflows: Digital Methods for Reproducible Research Practices in the Arts and Humanities). Zenodo. https://doi.org/10.5281/zenodo.15222895.

[5] https://dariah-de.pages.gwdg.de/textgridrep/textgrid-python-clients/docs/# (26.04.2025)

[6] https://gitlab.gwdg.de/textplus/textplus-io/nb-actions/ (26.04.2025)

[7] https://gitlab.gwdg.de/textplus/textplus-io/tg_model (26.04.2025)

[8] https://textgridlab.org/doc/services/submodules/tg-crud/tgcrud-webapp/docs/ (26.04.2025)

[9] https://gitlab.gwdg.de/dariah-de/textgridrep/tgadmin (26.04.2025)