

The European Literary Text Collection in TextGrid Repository

Rißler-Pipka, Nanette

nanette.rissler-pipka@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH (GWDG)

Calvo Tello, José

calvotello@sub.uni-goettingen.de
Göttingen State and University Library, Germany

Funk, Stefan E.

funk@sub.uni-goettingen.de
Göttingen State and University Library, Germany

Odebrecht, Carolin

carolin.odebrecht@hu-berlin.de
Humboldt-Universität zu Berlin

Schöch, Christof

schoech@uni-trier.de
University of Trier

Veentjer, Ubbo

veentjer@sub.uni-goettingen.de
Göttingen State and University Library, Germany

In this poster, researchers from different projects present the integration of existing TEI-encoded corpora into a repository and analysis infrastructure, as well as the benefits of this integration. The focus is not on resource creation (corpus design or text encoding), but on infrastructure integration, dissemination and re-use of existing resources.

TextGrid Repository

The TextGrid Repository is a long-term archive for humanities research data, with a focus on XML-TEI based publications and images (Neuroth et al., 2015).¹ It offers a way to publish research data in a citable manner, describing it through appropriate metadata for better findability. It is certified with the CoreTrustSeal to reflect trustworthiness and its alignment with the FAIR principles (Wilkinson et al., 2016).² Similar repositories are GAMS or TAPAS, although neither offers free publication to any project.

As of now, most of the texts in the repository are in German, such as the TextGrid Digital Library, a collection of texts in German written in different literary periods.³

ELTeC

The European Literary Text Collection (ELTeC) is a state-of-the-art, open access resource produced in the COST Action ‘Distant Reading for European Literary History’.⁴ ELTeC is a collection of corpora of literary texts (novels published between 1840 and 1920) that are comparable in nature, scope and quality across several European languages, designed as a multilingual benchmark dataset for the development of tools and methods in Computational Literary Studies. (Further characteristics and composition criteria of the corpora have been described by Burnard et al. 2021 and Schöch et al. 2021.)

The latest release of ELTeC has three components:

- ELTeC-core: comparable corpora in 12 languages, each with 100 novels, of the mentioned period.
- ELTeC-plus: 9 corpora in 9 additional languages following the same principles, but containing fewer than 100 novels.
- ELTeC-extensions: additional novels in languages already represented in ELTeC, also covering earlier periods.

Currently, ELTeC contains more than 2000 full-text novels in XML-TEI (Burnard 2014, 2016, Cummings 2019).

Regarding its publication strategy, ELTeC is distributed via multiple platforms: On Github (as a publicly-available collaborative workspace) and on Zenodo (as a permanent archive). ELTeC also serves as a test case for the infrastructure design work conducted in the Computational Literary Studies Infrastructure (CLS INFRA) project.⁵

Integration of Resources within the Text+ Consortium

The landscape of national research data infrastructure in Germany changed recently due to a long-term funding programme (NFDI)⁶ which “aims to create a permanent digital repository of knowledge”⁷ in order to boost research based on data in Germany across all disciplines. Text+⁸ is one of the more than 20 consortia that are part of the NFDI and one of four consortia in the fields of humanities and social sciences (Kett et al., 2022). It is particularly dedicated to text and language based research. The integration of existing community resources like ELTeC is one of the most fundamental objectives in the task area and data domain “Collections” of Text+. The TextGrid repository is one of the Text+ repositories for highly structured (XML-TEI format) resources such as digital editions and literary corpora.

The integration of ELTeC into the TextGrid repository⁹ enhances its value for a broader community by providing multilingual literary corpora inside a well established research infrastructure. Other community resources which fulfill the TextGrid requirements regarding format (XML-TEI) and metadata are welcome. The integration of ELTeC and other non-German literary corpora like CoNNSA¹⁰ (Calvo Tello, 2021) serve as pilots to develop best practices. A Python library for more flexible interaction with TextGrid is under development.¹¹

Advantages of ELTeC in TextGrid

As a domain specific repository, TextGrid enables specific functions for XML-TEI documents, such as their transformation to other formats, like simple HTML, which facilitates the access to traditional scholars.

Regarding the FAIR criteria (Wilkinson et al., 2016), ELTeC's findability is enhanced by persistent identifiers for corpora, works, editions and texts, multiple options for using authority file identifiers for authors, and subject classification by the library classification system Basic Classification (Schulz, 1991). Moreover, each text in the TextGrid repository is findable in OpenAIRE (Tóth-Czifra 2021). The general accessibility of the ELTeC is improved through the TextGrid repository's APIs and services (Funk, 2018). The corpora become more interoperable and reusable, through the possibility of combining them with other corpora in the repository, through the shelf function or the CLARIN Virtual Collection Registry. Finally, single texts or entire corpora can be sent to analytic tools such as Voyant (Sinclair and Rockwell, 2016) or to the CLARIN Switchboard (Zinn, 2016).

The ELTeC in the TextGrid repository will benefit from future developments as part of the Text+ portfolio. For example, a recently added feature is the possibility for projects to edit flexibly their own branded page. We are developing new workflows to facilitate bulk import of already existing corpora. New features are being developed relating to the authority files identifiers (by Wikidata, VIAF, GND) of authors and works enabling Linked Open Data functions, such as complex queries relating both the metadata and the fulltext, or the extraction of information from third-party resources and its integration into TextGrid. These and further developments will be applicable to all texts in the TextGrid repository, including the ELTeC.

Notes

1. TextGrid Repository: <https://textgridrep.de/?lang=en>.
2. TextGrid Repository – CoreTrustSeal Certification: <https://www.coretrustseal.org/wp-content/uploads/2020/05/TextGrid-Repository.pdf>
3. TextGrid Digital Library: <https://textgrid.de/en/digitale-bibliothek>.
4. See: <https://www.distant-reading.net/>. A good starting point for more information about ELTeC is <https://www.distant-reading.net/eltec/>.
5. <https://clsinfra.io/>
6. For an overall explanation, see: https://www.dfg.de/en/research_funding/programmes/nfdi/index.html.
7. <https://www.nfdi.de/association/?lang=en>
8. For a description of the consortium, see: <https://www.text-plus.org/en/home/>.
9. For the integration of ELTeC in TextGrid, see: <https://textgridrep.org/project/TGPR-99d098e9-b60f-98fd-cda3-6448e07e619d>.
10. For the integration of CoNLSA, see: <https://textgridrep.org/project/TGPR-8b44ca41-6fa1-9b49-67b7-6374d97e29eb>.
11. <https://gitlab.gwdg.de/dariah-de/textgridrep/textgrid-python-clients>

Bibliography

- Calvo Tello, J.** (2021). *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. (Digital Humanities Research 4). Bielefeld: transcript <https://www.transcript-verlag.de/978-3-8376-5925-2/the-novel-in-the-spanish-silver-age/?c=331025282>.
- Burnard, L.** (2014). *What Is the Text Encoding Initiative? : How to Add Intelligent Markup to Digital Resources*. Marseille: OpenEdition Press <http://books.openedition.org/oep/426> (accessed 20 March 2015).
- Burnard, L.** (2016). *ODD Chaining for Beginners*. TEI Consortium <https://teic.github.io/PDF/howtoChain.pdf>.
- Burnard, L., Schöch, C. and Odebrecht, C.** (2021). In search of comity: TEI for distant reading. *Journal of the Text Encoding Initiative*(Issue 14). Text Encoding Initiative Consortium doi: 10.4000/jtei.3500 . <https://journals.openedition.org/jtei/3500> (accessed 10 September 2021).
- Cummings, J.** (2019). A world of difference: Myths and misconceptions about the TEI. *Digital Scholarship in the Humanities*, 34(Supplement_1). Oxford Academic: i58–79 doi: 10.1093/llc/fqy071 .
- Funk, S. E.** (2018). *Elektronisches Publizieren von Digitalen Forschungsdaten am Beispiel des TextGrid Repositoriums – Umsetzung von Digitalen Publikationsworkflows für die eHumanities*. Köln <http://dx.doi.org/10.20375/0000-000B-D269-2>.
- Kett, J., Kudella, C., Rapp, A., Stein, R. and Trippel, T.** (2022). Text+ und die GND – Community-Hub und Wissensgraph. *Zeitschrift Für Bibliothekswesen Und Bibliographie*, 69(1–2): 37–47 doi: 10.3196/1864295020691262 .
- Neuroth, H., Rapp, A. and Söring, S.** (eds). (2015). *Text-Grid: Von der Community - für die Community: eine virtuelle Forschungsumgebung für die Geisteswissenschaften*. Glückstadt: Hülsbusch.
- Odebrecht, C., Burnard, L. and Schöch, C.** European Literary Text Collection (ELTeC) COST Action Distant Reading for European Literary History (CA16204) <https://doi.org/10.5281/zenodo.4662444>.
- Schöch, C., Erjavec, T., Patras, R. and Santos, D.** (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*(1). Liverpool University Press: 25 doi: 10.3828/mlo.v0i0.364 .
- Sinclair, S. and Rockwell, G.** (2016). Voyant Tools <http://voyant-tools.org/>.
- Schulz, U.** (1991). Die niederländische Basisklassifikation : eine Alternative für die ‘Sachgruppen’ im Fremddatenangebot der Deutschen Bibliothek. *Bibliotheksdienst*, 25: 1196–219.
- Tóth-Czifra, E.** Connecting Arts and Humanities to the European open data commons: the OpenAIRE-DARIAH Research Community Gateway Billet *DARIAH Open* <https://dariahopen.hypotheses.org/995> (accessed 2 November 2022).
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al.** (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3 doi: 10.1038/sdata.2016.18 . <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/> (accessed 20 May 2020).
- Zinn, C.** (2016). The CLARIN Language Resource Switchboard. Aix-en-Provence https://www.clarin.eu/sites/default/files/zinn-CLARIN2016_paper_26.pdf.