TEI-Dokumente in TextGrid Repository veröffentlichen und archivieren: neue Features und fluffiger Import Workflow

Calvo Tello, José
calvotello@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0002-1129-5604

Barth, Florian
florian.barth@uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0003-3408-7311

Buddenbohm, Stefan
buddenbohm@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0002-3469-6101

Dogaru, George
george.dogaru@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), Deutschland
ORCID: 0000-0001-9500-9638

Funk, Stefan
funk@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0003-1259-2288

Göbel, Mathias
goebel@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0002-1102-5284

Klammer, Ralf
ralf.klammer@tu-dresden.de
Technische Universität Dresden, Deutschland
ORCID: 0000-0003-4816-6440

Kudella, Christoph
kudella@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0002-9645-7122

Rißler-Pipka, Nanette
Rissler-Pipka@MaxWeberStiftung.de
Max Weber Stiftung, Deutschland
ORCID: 0000-0002-0719-9003

Veentjer, Ubbo
veentjer@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0002-9726-3135

Weimer, Lukas
weimer@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Deutschland
ORCID: 0000-0001-6919-3646

Das TextGrid Repository

TextGrid ist eines der langlebigsten DH-Projekte im deutschsprachigen Raum (Neuroth, Rapp, und Söring 2015). Die beiden Hauptkomponenten, das Repository (TGR) und das Laboratory (TGL), wurden ursprünglich im Rahmen der durch das BMBF geförderten D-Grid-Initiative entwickelt und innerhalb von DARIAH-DE und CLARIAH-DE weiterentwickelt (Calvo Tello u. a. 2023). Ein wesentlicher Bestand des TextGrid Repositories bildet die sogenannte ‘Digitale Bibliothek’ [1], eine umfangreiche Sammlung von Belletristik und Sachliteratur in deutscher Sprache, deren Texte in XML-TEI kodiert wurden und unter einer CC-BY 3.0 Lizenz (Betz 2015) zur Verfügung gestellt werden.

Als XML-spezifisches Repositorium bietet das TextGrid Repository sowohl Projekten als auch Endnutzer*innen eine Reihe von Features. Dazu gehören z.B. die Transformation von TEI in HTML, die Generierung von Inhaltsverzeichnissen, die Vergabe von PIDs für jedes Objekt (wie Dokumente und Bilder) sowie die direkte Anbindung an Werkzeuge und Dienste wie die Voyant Tools oder das CLARIN Language Resource Switchboard. Zudem ermöglicht die sogenannte Regalfunktion die Zusammenstellung personalisierter Sammlungen (Funk und Pempe 2015).

Nach unserem Kenntnisstand ist das TextGrid Repository weltweit das einzige auf TEI spezialisierte Repository, in dem Projekte ihre Dokumente unabhängig vom Herkunftsland, der Gattung oder der Sprache der Dokumente kostenfrei veröffentlichen können (Calvo Tello et al. 2023).

Nachhaltigkeit und Vertrauenswürdigkeit des TextGrid Repository sind durch das CoreTrustSeal ausgewiesen, dessen Wiederbeantragung derzeit in Bearbeitung ist.

Neue Entwicklungen und Korpora

Im Rahmen des NFDI-Konsortiums Text+ (Hinrichs u. a. 2022) wird das TextGrid Repository weiterentwickelt, wobei die Bereitstellung eines neuen, benutzerfreundlichen Import-Workflow (Buddenbohm u. a. 2024) einen der Schwerpunkte bildet. Dank dieser Entwicklung konnten in den letzten Monaten bereits mehrere Korpora und Textsammlungen veröffentlicht werden wie z.B. CoNSSA (Calvo Tello 2021), ELTeC (Rißler-Pipka u. a. 2023; Schöch u. a. 2021) oder textbox (Schöch u. a. 2019). Der Workshop soll es den Teilnehmenden ermöglichen, diesen neuen Import-Workflow kennenzulernen und selbst auszuprobieren.

Weitere Projekte werden gerade in Vorbereitung ihrer Publikation im TextGrid Repository betreut, darunter u.a. Korpora mit Bibelübersetzungen in ca. 100 Sprachen, französische und deutsche Romane, deutsche Lyrik, Folklore aus der Kaukasusregion und spanische Fabeln. Diese Betreuung erfolgt in mehreren Schritten, in denen die Forschenden Beispiele aus den Textsammlungen schicken und eine Fachperson Vorschläge und Feedback zur Verbesserung der Qualität der Metadaten gibt. Typische Beispiele sind das Fehlen wichtiger Metadaten in den TEI-Dokumenten, die für das gesamte Korpus oder die Sammlung zutreffend sind. Zum Beispiel hat ein Korpus aus französischen Romanen, die von Autorinnen geschrieben wurden, wahrscheinlich keine expliziten Metadaten zur Gattung, Sprache oder zum Geschlecht der Autor*in in den TEI-Dokumenten, da diese Informationen für das Projekt selbstverständlich sind. Diese Metadaten sind jedoch notwendig, wenn diese Dokumente neben anderen Dokumenten in einem Repository abgelegt werden. Ein weiteres Beispiel für mangelnde Qualität in den Metadaten ist der Verweis auf die Sprache durch einen nicht-standardisierten Freitext; zu bevorzugen wäre ein Standard wie z.B. durch einen ISO-Code. Das Hinzufügen vieler dieser Metadaten zu den TEI-Dokumenten kann automatisiert werden. Im Workshop wird es auch Zeit geben, um konkrete Beispiele gemeinsam zu diskutieren. Dadurch wird die Qualität der Metadaten und ihr FAIR-Status (Wilkinson u. a. 2016) erhöht, was die Auffindbarkeit der Forschungsdaten unabhängig von ihrer Veröffentlichung im TextGrid Repository verbessert.

Ein wesentlicher Aspekt dieser neuen Entwicklungen basiert auf einer stärkeren Berücksichtigung der FAIR-Prinzipien. So wurden z.B. neue Metadatenoptionen im Zusammenhang mit Linked Open Data Ressourcen ermöglicht, wie die Zuordnung zu Gattungen über die GND (Kett u. a. 2022) oder zu Fachklassifikationen über die Basisklassifikation (Calvo Tello u. a. 2023; Balakrishnan und Voß 2022; Schulz 1991). Die Integration und Ergänzung der GND-Entitäten wird derzeit mit der GND-Agentur in Text+ (Buddenbohm und Fischer 2023) abgestimmt.

Darüber hinaus wurden Features entwickelt, die bei der Veröffentlichung der Daten im TextGrid Repository projektspezifisch ausgestaltet werden können. Dazu zählt die Gestaltung einer individuellen Landing-Page als auch projektspezifische Metadaten für die Facettierung (z.B. Geschlecht der Autor*in), XSLT-Stylesheets, Projekt-Icons etc. (Calvo Tello u. a. 2023). Das Repository bietet zudem auch neue programmatische Zugänge zu den Texten und Metadaten, z.B. durch die neue Python-Bibliothek ‘tgclients’ (Hynek u. a. 2024).

Im TextGrid Repository publizierte Projekte profitieren zudem von einer Anbindung an die Dienste von Text+ wie z.B. Federated Content Search (Körner u. a. 2023), Registry (Gradl u. a. 2024) und Langzeitarchivierung (Dogaru 2023).

Wenn Interesse besteht, können im Workshop weitere Aspekte angesprochen werden, wie z.B. die Publikation von abgeleiteten Daten oder abgeleiteten Textformaten im TextGrid Repository (Calvo Tello und Rißler-Pipka 2023) oder die Einbindung von weiteren Ressourcen wie z.B. die Natural-Language-Processing-Pipeline MONAPipe (Barth u. a. 2023).

Neuer *fluffiger* Import Workflow

Das Metadatenmodell des TextGrid Repositories wurde vor dem theoretischen Hintergrund des konzeptuellen Modells ‘Functional Requirements for Bibliographic Records’ (FRBR) entwickelt. Hierbei wurden die vier Ebenen des FRBR-Modells - Werk, Expression, Manifestation und Exemplar (Gantert 2016; Taylor 2007) - auf drei Ebenen vereinfacht: Werk, Aggregation und Item (Neuroth, Rapp, und Söring 2015). Die Aggregation kann jedoch von unterschiedlicher Art sein, z.B. Edition und Sammlung. Jeder dieser Objekttypen wird zusätzlich in zwei Dokumente unterteilt: ein Dokument für die Daten und ein Dokument für die Metadaten. In der Praxis führt dies dazu, dass für jedes TEI-Dokument, das ein Projekt veröffentlichen will, bis zu fünf zusätzliche TextGrid Repository-Metadaten-Dokumente erzeugt werden müssen. Bisher wurde von den Forschenden erwartet, dass sie mit diesem Modell und den verschiedenen Ebenen von Dokumenten im TextGrid Repository umgehen können. Die Erfahrungen der letzten Jahre haben jedoch immer wieder gezeigt, dass dies die Veröffentlichung von Projekten behindert hat.

Eine mögliche Lösung wäre, das Metadatenmodell des TextGrid Repositories zu vereinfachen. Dies würde jedoch bedeuten, die gesamte Architektur des Repositoriums und der bereits publizierten Projekte zu ändern. Das Ziel der NFDI-Konsortien ist jedoch nicht die Neugestaltung bestehender Infrastrukturen, sondern die Integration bestehender Ressourcen.

Deshalb wird im neuen Import-Workflows davon ausgegangen, dass beinahe alle Metadaten, die für die Dokumente im TextGrid Repository benötigt werden, bereits in den ursprünglichen TEI-Dokumenten vorhanden sind. Einige der fehlenden Metadaten können dem gesamten

Korpus pauschal zugeordnet werden, wie z.B. Genre, Geschlecht der Autorinnen und Sprache für das oben erwähnte fiktive Beispiel eines Korpus französischer Frauenromane. Mit dem neuen *fluffigen* Import erhalten die Forschenden zuerst ein von tg-model [2] vorbereitetes YAML-Dokument. Dieses Dokument enthält Vorschläge für xPaths, die die Stellen in den XML-Dokumenten angeben, wo die benötigten Metadaten zu finden sind. Die Nutzenden müssen dann nur noch diese xPaths anpassen oder die Metadaten als pauschale Werte für das gesamte Korpus eingeben. Wenn die Nutzenden fertig sind, erzeugt tg-model für jedes Input-Dokument die fünf benötigten TextGrid Repository-Dokumente. Die Nutzenden können diese dann über tgadmin [3] hochladen und in einem geschlossenen Bereich des TextGrid-Portals überprüfen, ob die Daten korrekt sind. Falls nicht, können die TEI-Dokumente, xPaths oder Werte erneut angepasst werden und der iterative Prozess kann so lange durchgeführt werden, bis die Nutzenden mit der Präsentation der Daten zufrieden sind.

Der neue Import-Workflow kann am einfachsten auf dem Text+-Jupyterhub durchgeführt werden. Für den neuen Import-Workflow wurden neue Tools entwickelt (tg-model, textgrid-import-ui [4]) und bestehende weiterentwickelt (tgadmin, tgclients [5]). Die benötigten Bibliotheken können entweder direkt über eine Kommandozeile verwendet werden (von Nutzenden mit Programmierkentnissen), oder mittels eines Jupyter Notebooks, das diese Tools über eine Web-Oberfläche bedienbar macht. Eine Installation des Notebooks und der dazugehörigen Software auf dem eigenen Gerät ist möglich, aber unnötig, weil auf dem Text +-Jupyterhub alles Nötige für die Verwendung bereitliegt.

Da die Jupyter-Technologie viele Vorteile bietet, entstehen im Text+-Kontext weitere Notebook-basierte Anwendungen, unter anderem auf Basis der Software, die für den Import-Workflow entwickelt wurde. Der Text+-Jupyterhub wird außerdem in den NFDI-Basisdienst Jupyter4NFDI [6] integriert, was einen breiteren Zugang zu diesen Anwendungen und den darunterliegenden Tools ermöglicht.

Zielgruppe und Pläne

Der Workshop richtet sich an zwei verschiedene Gruppen: Zum einen an Forschende, die bereits über zu veröffentlichende Dokumente verfügen. Diese Dokumente sollten vorzugsweise bereits in XML-TEI kodiert sein, aber wir können auch über die Konvertierung von anderen Formaten (“plain text”, Word-Dokumente) beraten. Zum anderen richtet sich der Workshop auch an Forschende, die keine Dokumente für ihre Publikation haben, aber am Import interessiert sind. Für diese Teilnehmenden werden kleine Sammlungen zum Ausprobieren vorbereitet. Obwohl in TextGrid Repository hauptsächlich literarische Sammlungen veröffentlicht wurden, können Projekte auch Editionen publizieren, vor allem als Fallback-Lösung.

Im Vorfeld des Workshops werden die Teilnehmenden gebeten, grundlegende Informationen zu ihren Daten zu geben (ob sie ein Korpus haben, Anzahl der Dokumente, ob die Daten in TEI vorliegen, ob Bilder vorhanden sind) und dass sie sich erfolgreich in TGR angemeldet haben (entweder über einen DARIAH-Account oder über die eigene Institution).

Der geplante Workshop wird zwei halbe Tage dauern. Am ersten Tag ist eine Vorstellung des Teams, der Teilnehmenden und ihrer Daten vorgesehen. Die Hauptmerkmale des TGR werden erklärt und einen Überblick über den Import-Workflow mit grundlegenden Metadaten und Zeit zum Ausprobieren gegeben. Am zweiten Tag werden weitere Optionen zur Verbesserung und Erweiterung der Metadaten und projektspezifische Optionen (eigene XSLT-Transformation, Landing Page, Icon) vorgestellt und die Daten für eine abschließende Veröffentlichung überprüft. Weitere TGR-Dienste und Optionen (abgeleitete Textformate, Zugang zur tg-model API, Python Library) werden ebenfalls vorgestellt. Insgesamt werden die Teilnehmenden ihre Projekte und Daten vorstellen, wir werden das TextGrid Repository, seine alten und neuen Features vorstellen und ein erstes Publizieren eines Testkorpus gemeinsam durchführen. Anschließend haben die Teilnehmenden die Möglichkeit, ihre Daten in das TextGrid Repository hochzuladen. Verschiedene Experten und Entwickler der verschiedenen TextGrid-Komponenten werden anwesend sein, so dass die Teilnehmenden eine gute Beratung für die verschiedenen Schritte erhalten können. Das Hauptziel des Workshops ist es, dass die Projekte eine Version ihres Datensatzes in einem geschlossenen Bereich des Repositories hochgeladen haben.

Fußnoten

1. https://textgrid.de/digitale-bibliothek
2. Ein speziell für den Import entwickeltes Python-Tool, das die Erstellung der erforderlichen TextGrid-spezifischen Metadaten-Dateien erleichtert.
3. Ein Python-Tool zur Erstellung und Verwaltung von Projekten in TextGrid Repository.
4. Eine Web-Anwendung zur Durchführung des Import-Workflows, die über ein Jupyter Notebook bedienbar ist.
5. Ein Python-Tool für die Interaktion mit der TGR-API, worauf auch tgadmin basiert.
6. S. base4nfdi.de/projects/jupyter4nfdi und nfdi-jupyter.de .

Bibliographie

Balakrishnan, Uma, und Jakob Voß. 2022. „Automatische Anreicherung der Sacherschließung des Verbundkatalogs K10plus mittels coli-rich“. In *#FreiräumeSchaffen*. Leipzig: Bibliothek und Information Deutschland. https://bid2022.abstractserver.com/program/#/details/presentations/27.

Barth, Florian, Yannic Bracke, José Calvo Tello, George Dogaru, Tillmann Dönicke, Keli Du, Stefan

E. Funk, Philippe Genet, Mathias Göbel, Lennart Keller, Daniel Kurzawe, Ubbo Veentjer, & Lukas Weimer. (2023). MONAPipe: Modular Natural Language Processing Pipeline for Digital Humanities (1.0). Text+ Plenary 2023: Connecting People an Data, SUB Göttingen. Zenodo. https://doi.org/10.5281/zenodo.8424925

Betz, Katrin. 2015. „Ein virtuelles Bücherregal: Die Digitale Bibliothek im TextGrid Repository". In *TextGrid: Von der Community - für die Community*, herausgegeben von Heike Neuroth, Andrea Rapp, und Sibylle Söring, 229–39. Glückstadt: Verlag Werner Hülsbusch. https://publications.goettingen-research-online.de/handle/2/14388 .

Buddenbohm, Stefan, José Calvo Tello, George Dogaru, Stefan Funk, Ralf Klammer, Alex Steckel, und Lukas Weimer. 2024. „Preserving Humanities Research Data: Data Depositing in the TextGrid Repository Aka The Fluffy Import". Gehalten auf der DARIAH Annual Event - Workflows, Lisbon, Juni 18. https://zenodo.org/records/11279675 .

Buddenbohm, Stefan, und Barbara K. Fischer. 2023. „Das G in GND: ein Erfahrungsaustausch von GND-Agenturen". *Text+ Blog* (blog). 3. Mai 2023. https://textplus.hypotheses.org/5071 .

Calvo Tello, José. 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 .

Calvo Tello, José, Florian Barth, Stefan Funk, Daniel Kurzawe, Nanette Rißler-Pipka, und Ubbo Veentjer. 2023. „Between corpora, tools and standards data: TextGrid Repository for Hispanic Studies". In *Hispanistik in neuen Umwelten: Digitalisierung, Schnittstellen, Reinskriptionen*, 253–56. Graz. https://doi.org/10.25364/513.2023.1 .

Calvo Tello, José, Stefan E. Funk, Daniel Kurzawe, und Ubbo Ventjeer. 2023. „Inhaltserschließung für Forschungsdaten: TextGrid Repository, Normdaten und Basisklassifikation". In *FORGE23*. Tübingen. https://doi.org/10.5281/zenodo.8341605 .

Calvo Tello, José, und Nanette Rißler-Pipka. 2023. „¿Qué hacer con textos que no se pueden publicar? Datos derivados, criterios FAIR y TEI". *Journal of the Text Encoding Initiative*, Nr. 16 (Mai). https://doi.org/10.4000/jtei.4720 .

Dogaru, George. 2023. „Das Text+-Langzeitarchiv: Eine generische Lösung für den nachhaltigen Erhalt von Daten in den Geisteswissenschaften". Juni 22. https://doi.org/10.5281/zenodo.8108792 .

Florian Barth, Stefan Buddenbohm, José Calvo Tello, George Dogaru, Stefan E. Funk, Mathias Göbel, Ralf Klammer, Ubbo Veentjer (8. Mai 2024). Fluffy Workflow: Neue Tools für den Datenimport ins TextGridRep. Text + Blog. Abgerufen am 19. Juli 2024 von https://doi.org/10.58079/11nmp

Funk, Stefan E., und Wolfgang Pempe. 2015. „Vom Konzept zur Umsetzung — Einblicke in die Entstehung des TextGrid Repository". In *TextGrid: Von der Community - für die Community*, herausgegeben von Heike Neuroth, Andrea Rapp, und Sibylle Söring, 191–200. Glückstadt: Verlag Werner Hülsbusch. https://publications.goettingen-research-online.de/handle/2/14388 .

Gantert, Klaus. 2016. *Bibliothekarisches Grundwissen*. *Bibliothekarisches Grundwissen*. Berlin, Boston: De Gruyter Saur. https://www.degruyter.com/view/title/302969 .

Gradl, Tobias, Harald Lordick, Christoph Kudella, und Daniela Schulz. "Towards a Registry for Digital Resources – The Text+ Registry for Editions." Datenbank-Spektrum (2024): . https://doi.org/10.1007/s13222-024-00479-0.

Hinrichs, Erhard, Peter Leinen, Alexander Geyken, Andreas Speer, und Regine Stein. 2022. „Text+: Language- and Text-Based Research Data Infrastructure". Zenodo. https://doi.org/10.5281/zenodo.6452002. https://zenodo.org/record/6452002 .

Hynek, Stefan, Ubbo Veentjer, José Calvo Tello, Florian Barth, Stefan Funk, Mathias Goebel, Daniel Kurzawe, und Lukas Weimer. 2024. „TextGrid Python Clients: Making the Repository Programmable". In . https://doi.org/10.5281/zenodo.10706157 .

Kett, Jürgen, Christoph Kudella, Andrea Rapp, Regine Stein, und Thorsten Trippel. 2022. „Text+ Und Die GND – Community-Hub Und Wissensgraph". *Zeitschrift Für Bibliothekswesen Und Bibliographie* 69 (1–2): 37–47. https://doi.org/10.3196/1864295020691262 .

Körner, Erik, Thomas Eckart, Axel Herold, Frank Wiegand, Frank Michaelis, Matthias Bremm, Louis Cotgrove, Thorsten Trippel, und Felix Rau. 2023. „Federated Content Search for Lexical Resources (LexFCS): Specification". https://doi.org/10.5281/zenodo.7986303 .

Neuroth, Heike, Andrea Rapp, und Sibylle Söring, Hrsg. 2015. *TextGrid: Von der Community - für die Community: eine virtuelle Forschungsumgebung für die Geisteswissenschaften*. Glückstadt: Verlag Werner Hülsbusch. https://publications.goettingen-research-online.de/handle/2/14388 .

Rißler-Pipka, Nanette, José Calvo Tello, Stefan E. Funk, Carolin Odebrecht, Christof Schöch, und Ubbo Veentjer. 2023. „The European Literary Text Collection in TextGrid Repository". In *Collaboration as Opportunity*, herausgegeben von Walter Scholger, Georg Vogeler, Toma Tasovac, Anne Baillot, Elisabeth Raunig, Martina Scholger, Elisabeth Steiner, und Patrick Helling. Graz: ADHO. https://doi.org/10.5281/ZENODO.8107707 .

Schöch, Christof, José Calvo Tello, Ulrike Henny-Krahmer, und Stefanie Popp. 2019. „The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in XML-TEI". *Journal of the Text Encoding Initiative*. https://journals.openedition.org/jtei/2085 .

Schöch, Christof, Tomaz Erjavec, Roxana Patras, und Diana Santos. 2021. „Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives".

Modern Languages Open, Nr. 1 (Dezember), 25. https://doi.org/10.3828/mlo.v0i0.364 .

Schulz, Ursula. 1991. „Die niederländische Basisklassifikation: eine Alternative für die ‚Sachgruppen' im Fremddatenangebot der Deutschen Bibliothek". *Bibliotheksdienst* 25:1196–1219.

Taylor, Arlene G., Hrsg. 2007. *Understanding FRBR: what it is and how it will affect our retrieval tools*. Westport, Conn: Libraries Unlimited.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3 (März). https://doi.org/10.1038/sdata.2016.18 .