1 Data Policies
2 Content

- Collection Development Policy and Data Quality/Reuse
- Data Reuse
- Preservation Policy
  - Aims and Requirements of the Policy
  - Recommendations and List of Preferred Formats
  - Legal and Regulatory Framework
- Important Ethical and Disciplinary Norms
  - Personal Sensitive Data with Disclosure Risk

The Policies of the TextGrid Repository (TextGridRep) for the development of the collection, for data access, quality and reuse as well as preservation is strongly related to the Virtual Research Environment TextGrid (VRE TextGrid) as a whole and strongly influenced by its community driven approach. Hence, it is a result of the development of the virtual research environment in the last 12 years, where researchers using it and experts have mainly contributed to decisions concerning access to data, metadata requirements and options. Aims and requirements of the here described data policies are related to this background, which will be described in the following.
3 Collection Development Policy and Data Quality/Reuse

The TextGrid Repository is especially designed for the publication of digital editions. It is – together with the TextGrid Laboratory (TextGridLab) – part of the TextGrid Virtual Research Environment (VRE). Data are usually published out of the TextGrid Laboratory into the TextGridRep. The TextGridRep is the foundation of many digital edition projects, such as Fontane Notizbücher and Bibliothek der Neologie, that are working with the TextGridLab for editing and transcribing texts and storing images related to an author or topic. The outcome of those editions are then stored (published) in the TextGridRep.

For an overview of research projects using TextGrid as virtual research environment and cooperating with DARIAH-DE see TextGrid – Projects introduce themselves and the picture below:

Cooperations in the context of DARIAH-DE and TextGrid (as Virtual Research Environment)

The long-term research data archive TextGridRep offers in this perspective safe storing, publishing and researching for versatile digital material e.g. XML /TEI formatted text, images and databases.

That’s why our collection development policy focusses mainly in direction of digital editions and related data contributing at the same time to the broader context of the development and set up of a service for digital editions at Goettingen State and University Library. Further (Digital) Humanities related data is very welcome.

To open the TextGridRep and his data to a broader public and other students and researchers outside the research projects using the VRE TextGrid the content of TextGridRep has been enriched in 2011 by the so called TextGrid Digital Library. The TextGrid Digital Library represents an extensive collection of German texts in digital form, ranging from the beginning of the printing press up to the first decades of the 20th century. It contains virtually all the important texts in the canon and numerous other texts relevant to literary history whose copyright has expired. It is therefore of particular interest to German Literature Studies, Philosophy and Cultural Studies as a whole. All these texts have been converted additionally from XML into a valid TEI format in order to make them available for further processing for example in editions and text corpora. They are not only available for reading, but allow also an exact research into the texts.

Fig. 1: Cooperations in the context of DARIAH-DE and TextGrid (as Virtual Research Environment)
The description to make data accessible and reusable is provided by all necessary metadata that is available for every single object of the TextGridRep. The TextGrid metadata schema has been developed during the first two TextGrid project phases. Involved were the designated TextGrid community as TextGrid Repository users as well as metadata experts from SUB and TextGrid. The TextGrid metadata schema was designed to serve three purposes and to fetch all researchers from their level of metadata experience. While developing the metadata schema and repository tools we learned that it is crucial for researchers that do not have a high affinity for computers and IT to be still able to cope with the metadata and tools. Furthermore we learned that the metadata of different projects and/or depositors are very different and heterogeneous so that a single metadata schema with many mandatory fields cannot be served easily for everyone. The complexity level should be as low as possible for those users that simply wants to import and/or publish their data without having a high level of metadata expertise. We also want to serve those researchers that have more complex metadata and more experience in IT digital data. Therefore the TextGrid metadata schema was developed as a chained schema that serves editions, collections, aggregations, works, and items with different mandatory metadata for every level. So any researcher can decide which complexity shall be used and fits best to their requirements (see minimal and mandatory metadata).

This complexity is fully implemented in the TG-search service so that metadata searches can sustain the findability of all documents and furthermore it provides fulltext search, so enhanced and project specific metadata can also be searched and found, such as TEI headers in TEI XML files.

For reuse of the data (mainly XML/TEI transcribed texts and images of digitized manuscripts) the mandatory metadata fields are sufficient as the community evaluates the quality of data usually by accessing the encoded texts and digital editions – a procedure inherent to text-based scientific disciplines such as Editorial Philology or other disciplines of the Humanities when dealing with textual sources.

Additionally technical metadata is extracted for every TextGrid object during the publishing process, and is then also publicly available. It is stored as an extra file beneath the data and metadata files. You can get the object’s technical metadata from textgridrep.de on object site (see Tech. Metadata (XML)) or directly from the TG-crud via https://textgridlab.org/1.0/tgcrud-public/rest/textgrid:2v117.0/tech.

Within the framework of CLARIAH-DE, the further TextGridRep collection development is expanded. The works especially include measures to increase the interoperability of the collections stored in the TextGridRep with other data sets as well as measures to increase the reuse of the data, for instance by establishing connectivity to further tools for digital text analysis. For example, standardized approaches for converting the collections of the TextGrid Repository stored in the in TEI format to the DTA basic format (DTABf) used by The German Text Archive. This way, it would be possible to collections stored in the TextGridRep to be merged with the collections of The German Text Archive and thus simultaneously with the other CLARIN collections, for example for the purpose of text analysis. In terms of text analysis further necessary steps for the integration of tools like Switchboard and Weblicht are currently evaluated for implementation. In addition, an evaluation of the DTA basic format as potential standard for digital editions is ongoing. In this context also the use of the DRAIAH-DE Data Modelling environment is considered and evaluated for the integration of heterogeneous text data.

Further adjustments and developments will be considered by observing and analysing best practice of the designated community and taking into account their feedback.
4 Data Reuse

As stated in its mission statement the TextGrid Repository promotes free and open access to research data.

In terms of reuse the following definition of DARIAH-DE for research data follows the understanding of the designated community of the repository. The definition takes into account the special characteristics and the resulting heterogeneity of scientific research data in the Humanities:

"[…] all those sources / materials and results collected, written, described and / or evaluated in the context of a research and research question in the field of human and cultural sciences, and in machine-readable form for the purpose of archiving, citation and for further processing."

For data stored in the TextGrid Repository reuse is mainly about digital editions, digital text corpora and related data that are further processed for text-based scientific research. Especially machine readable XML/TEI encoded texts allows the research community to analyse them with further tools such as XML databases, XML analyzers, annotation tools, or by tools scripted by themselves.

Additionally the collection of texts of the TextGrid Digital Library are available open access and for free. The canon includes 94,461 works from 693 authors from the 20th century to the present with 106,832 XML/TEI files.

In terms of analysing and visualising research into texts, the voyant tools provide researchers with web-based, diverse explorative and analytical approaches to any text or text collection allowing to discover and visualize specific properties and structures. Besides a number of different single Voyant Tools, DARIAH-DE offers with Digivoy a tool which connects the TextGrid Digital Library directly with the Voyant Tools. They allow e.g. a visualisation of occurrences and relations of figures in literary works.

A good and illustrating example for reuse of data from the TextGridRep for further analysis with external tools is the project Digital Literary Network analysis (dlina) dealing with network analysis of dramatic texts. The working group was looking into hundreds of dramatic texts ranging from Greek tragedies to 20th century plays and worked on larger German, French, English, and Russian corpora. The TextGridRep was the source for the German Drama Corpus. Finally, the results of statistical analysis have been visualised to demonstrate literary networks and relations between the different works, which can be accessed by zooming in the Drama Networks Superposter.

A further reuse scenario of data from the TextGridRep is related to the TextGridLab as further component of the virtual research environment. A researcher may make use of an already existing XML encoded edition of a text and may add either further text, his annotation or further XML markup to the text. In this case a complete copy of the text with a new identifier (TextGrid URI) will be created, that can be finally published again in the TextGrid Repository. The metadata will automatically refereee the new text or new edited text to its primary source (relation metadata field: refers to). The community may develop different digital editions with different XML schemas developed and adapted to their specific research needs and later research generations may be able to analyse developments and approaches in text-based research related to certain topics, authors, genres or disciplines.

To sum up the reuse of data from the TextGridRep at a more abstract level, it can be described according to the Research Data Lifecycle as described here by DARIAH-DE and visualised in the following schema:

![Fig. 2: The DARIAH-DE research data lifecycle](image-url)

The lifecycle and its processes can be further specified for the TextGrid repository in terms of a publication process. The publication process is a central part of scientific research in the humanities and for the TextGrid Repository and VRE. Publication is the basis for the dissemination of research results and data to allow all researchers to referee (cite) to each other’s work. New research can follow up on the basis of already existing scientific results and sources. The TextGrid Repository covers the fundamental processes of the DARIAH-DE research data lifecycle in its publication lifecycle within its framework of the virtual research environment as a whole.
Fig. 3: Publication workflow in the TextGrid VRE

If we apply the following mapping from the DARIAH-DE research data lifecycle (Fig. 2) to the TextGrid publication workflow (Fig. 3), it maps perfectly:

1. Source definition – is applied before importing research data into the TextGrid VRE (SOURCES) and implies the evaluation and appraisal of the source data (so 6. is already included).
2. Processing – if imported (Ingest) in the VRE data can be processed using tools, it can be annotated, linked (VIRTUAL RESEARCH ENVIRONMENT), and then, if finished...
3. Data production – …new research data has been produced and can be published (Publication).
4. Conclusion – published data can then be visualized in various places, such as the TextGrid Repository, project specific internet portals (Fontane Notizbücher, Bibliothek der Neologie), or printed books, all based on the data published in the TextGridRep (ACCESS & CURATION).
5. Verbalization – Articles and other project related works (such as talks) will be done and published...
6. Peer review – …and finally the research data will be used as new sources and research data for other publications (SOURCES).
5 Preservation Policy

The Preservation Policy of the TextGrid Repository is in line with the open access strategy of the University of Göttingen and its research data policy. It represents a clear commitment to open access of research data in promoting and making data of the designated community of the repository as widely accessible and usable as possible. Here it follows clearly the mission statement of the repository in supporting the use of publications and data without any access restriction.

The TextGrid Repository commits itself to undertake all necessary efforts to enable sustainable open access to digital research data. Within its organisational and technical infrastructure the repository ensures continuous access to its resources by following an active preservation policy. The Humanities Data Centre and its responsible institutions SUB and GWDG assumes responsibility and provides all necessary resources for long-term preservation and accessibility of data stored in the repository.

It is especially designed for the most suitable formats for long-term preservation in the field of editorial publication which represents the current state of the art and common practice within the designated community. The TextGrid Repository explicitly and actively recommends especially supported formats for long-term preservation (such as XML/TEI formatted text and TIFF images) and offers respectively additional services as well as support and consultation. Further recommendations are also given for other data types. The designated community consists of scholars in the Humanities and represents mainly text-based research and editorial philology. A variety of disciplines are working with similar formats, mainly XML, TEI and TIFF. An overview of all used formats is available at the TextGrid Repository. The represented research disciplines can be classified as follows:

- Editorial Philology
- German Philology
- Slavic Studies
- Jewish Studies
- Ancient American Studies
- Theology
- Philosophy
- Ethnology
- Historical Science
- Legal History
- Cultural History
- Art History
- Musicology

Due to this interdisciplinarity the TextGrid metadata schema was developed as a chained system starting from a minimal set of mandatory metadata to ensure the reuse and evaluation of all data at a basic level. Depending on individual research needs the minimal mandatory metadata description can be expanded to a more complex schema. In addition, due to the xml encoding of textual data of the repository specific metadata can be researched through a full text search using TG-search. To evaluate data by accessing the (xml)-encoded texts is a usual procedure in the humanities and inherent to disciplines dealing with (digital) text based research. Text data can be enriched by expanding or enhancing the xml schema and adding of new marked up elements. This data can then be republished as a new related source for specific research needs. The repository supports format standards that ensure usability, access to data and its preservation for the designated community (see CollectionDevelopmentPolicyandDataQuality/Reuse). Within an ongoing collection development the repository stays in touch with the needs and the state of the art of the designated community and undertakes necessary steps including format changes or adding of new formats. Due to its commitment to open access and open science the repository supports in this context open formats in the sense of free file formats, that can be used by anyone at no monetary cost and whose specifications are visible and maintained by a standards organisation relevant for the designated community.

The following sections provide an overview of the main aspects of the preservation policy.
5.1 Aims and Requirements of the Policy

The preservation policy and its implementation aims:

- to operate for the community as a trusted digital repository for data in the humanities and related disciplines with a special focus on digital editions and relevant data for text-based scientific research
- to guarantee long-term preservation and open accessibility of the stored research data
- to keep data long-term searchable and citable
- to ensure the authenticity and integrity of the data and provide reliable data to researchers, and
- to keep the repository standards in compliance with the state of the art of the designated community including its ethical and legal standards following applicable law

For this purpose the TextGrid Repository strives to ensure the following requirements by its organisational and technical infrastructure:

- The SUB and GWDG as two well recognised institutions with the respective relevant expertise declare their responsibility for the long-term operation of the repository through common founding of the Humanities Data Centre (HDC) as operator of the TextGrid Repository and to take care of providing all necessary resources (technical, financial and in terms of knowledge and expertise of stuff members) – in addition to public project funding of associated projects and independently whenever necessary. See in this context also the founding manifesto of the HDC.
- All phases of the TextGrid Repository’s publication and preservation workflows are based on the Open Archiving Information System (see TextGrid and the Open Archival Information System (OAIS)),
- at bitstream preservation level the repository ensures data preservation in unchanged form for every item,
- for TEI/XML and TIFF as most suitable long-term preservation formats according to the state of the art of the designated community and respective technical standards the repository offers support and will also undertake format changes in the future if necessary,
- further recommendations in terms of preferable long-term-preservation-formats are given,
- the data is accompanied by appropriate metadata standards for the professional cataloguing of the data and to enable use and reuse for research purposes,
- appropriate ingest procedures ensure that data are checked and validated according to community standards (such as mandatory metadata fields and generated additional administrative and technical metadata),
- the integrity and authenticity of data is regularly checked through a technical based routine,
- the repository has implemented periodical local and distributed backups (located in dedicated computing centers with strict access control) allowing to reinstall the repository data from backup and to recover data in case of technical failures,
- security issues are covered through security and disaster plans including responsible persons and actions to undertake,
- documentation, data, metadata, and all related information are regularly maintained suitable to long-term archival storage,
- all involved entities and stuff members agree to regularly observe and evaluate if changes are to be considered necessary due to changing scientific practice or technical developments and how they are to be implemented (To see ongoing evaluations and planned actions that will be implemented see the section "Collection Development Policy and Data Quality/Reuse", and the wiki page "Digital Object Management"),
- also on an organisational and strategic level SUB and GWDG ensures that the repository stays closely related to its designated community and ongoing innovative developments through associated projects and engagement in new developments and initiatives at a national and international level.
5.2 Recommendations and List of Preferred Formats

In a long-term perspective not all file formats will ensure long-term usability, access to data and its preservation. Therefore, the TextGrid Repository recommends certain file formats according to the current state of the art and common practice within the designated community.

The TextGrid Repository as part of the Humanities Data Centre and of DARIAH-DE supports long-term preservation for the following formats, which are widely used by the designated community and represents the major part of the stored data, as proposed in the the nestor criteria Catalogue of Criteria for Trusted Digital Repositories:

> „Open, disclosed and frequently used formats are preferred as archive file formats, the assumption being that these will have a longer life, and there are more likely to be techniques and tools for converting or emulating them, given that they are supported by a wide circle of users."

A list of those proposed file formats is listed here below (see p. 26f.), the preferred formats for TextGrid are mentioned first:

- for structured text: XML (http://www.w3.org/XML/) preferably TEI/XML
- for unformatted text: ASCII/Unicode

Further recommendations from the nestor criteria catalogue:

- for video files: MPEG 4 File Format, ISO/IEC 14496 (http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm) All levels of the publication workflow of the TextGrid VRE and for the repository (as illustrated and described above) support and consultation are given.

At all levels of the publication workflow and lifecycle of the TextGrid VRE and for the repository (as illustrated and described above in the section DataReuse) support and consultation are guaranteed and given by staff members of the SUB and DARIAH-DE dealing with the TextGrid VRE via:

- consultation of research projects
- user meetings and Workshops
- online-Tutorials
- email for support
- user’s mailing list

All cooperating projects and research projects using the TextGrid VRE are entitled to receive at the beginning an initial consultation for starting using the TextGrid VRE and to be aware of relevant issues for data publication into the TextGrid Repository. Consultation and support is usually used by our designated community and highly recommended for the following issues:

- data ingest and data publication
- ingest of large amount of data
- data description with relevant metadata (mandatory and optional)
- data quality and data reuse
- digital text data: standards, file formats
- creating, editing and publishing digital editions
- use of the repository in general
- discovered bugs or needed adjustments for own research projects
5.3 Legal and Regulatory Framework

The TextGrid Repository manages ethical standards by a legal and regulatory framework as stated by the TextGrid Terms of Use. The depositor of data has to agree to these Terms and will therefore have to be clear about the following questions before publishing his data:

- Are scientific and ethical norms as well as applicable law considered?
- Are personal data contained in the data, and if so: is privacy protection ensured?
- Does the depositor hold all rights to publish the data?
- Is the data suitable to be published in an open access and publicly available repository?
- Which license to end-users shall be granted as access is always free and publicly available?

DARIAH-DE takes legal precautions to ensure that users of deposited data of the TextGridRep do not violate the law with regard to legal and ethical criteria related to personal privacy, copyright issues, computer fraud, abuse and dissemination of unlawful offending material, etc. Hence, the relationship between the repository and the data depositor is organised by the TextGrid Terms of Use as a legally-binding agreement covering several relevant areas important for a trustful digital repository, such as:

- All necessary rights and obligations of both parties are stated and confirmed.
- The repository has all necessary rights and permissions to undertake all necessary operations to ensure long-term preservation, accessibility and security of data.
- The depositor agrees to accept all ethical and scientific standards set by the repository and its designated community as stated in the terms of use. This is of legal and ethical nature.
- The depositor is aware of the TextGrid Repository being an open access archive without restrictions of access. In this context, the depositor is informed by the terms of use that the repository is not suitable to publish personal data which needs a restricted access according to applicable law and that he/she has to take care of legal and ethical criteria related to personal privacy.
- As far as sensitive data with disclosure risk are concerned the depositor agrees to not publish those data in the repository without permissions of the data subject whose rights have to be protected according to applicable law.
- Prohibitions to publish and disseminate harmful unlawful offending content are formulated and legal implications as removing of data, exclusion from publishing access to the repository or further legal implications are stated.
- The depositor agrees to follow the community standards of good scientific practice as stated in the recommendations of the German Research Association and to be subject to possible non-legal and legal implications if not.

Furthermore, most of the data stored in the TextGridRep is provided with a license, which determines the rights of use. In the case that no licence is provided, German copyright applies. Every user who stores data in the TextGridRep grants DARIAH-DE the right – unlimited in time, non-exclusive, and free of charge – to copy and store the data on its own servers and to make them accessible in electronic form via international data networks “a non-exclusive right of use”.

Users of deposited data in an open access repository therefore do not have to state explicitly, that they accept the terms of use as the depositor of data assures that

“he/she alone is entitled to dispose of the copyright to use the data, collections and metadata and that he/she has not made any dispositions contrary to the rights of these Terms of Use. In particular, he/she is responsible for ensuring that the permanent publication does not violate any third-party rights or copyright laws.” (TextGrid Terms of Use §8)

See for more details the TextGrid Terms of Use and the following section for important ethical and disciplinary norms.
6 Important Ethical and Disciplinary Norms

The technical infrastructure of the TextGrid Repository runs on a well-supported operating system. The hardware, software and used technologies are appropriate to serve nationally and internationally research, teaching and learning by providing long term preservation, further processing, openly sharing and dissemination of digital research data according to ethical and scientific standards of the international research community. The Designated Community of the TextGrid Repository follows the good scientific practice as recommended by the German Research Association, which also does the University of Göttingen. In terms of practice this means, as highlighted by the University on the respective website (https://www.uni-goettingen.de/en/604506.html) as well as in related documents and listed here:

- "The conduct of science rests on basic principles valid in all countries and in all scientific disciplines. The first among these is honesty towards oneself and towards others. Honesty is both an ethical principle and the basis for the rules, the details of which differ by discipline, of professional conduct in science, i.e. of good scientific practice" (Memorandum of the German Research Foundation DFG, 2013:67)
- Cooperation in scientific working groups must allow the findings, made in specialized division of labour, to be communicated, subjected to reciprocal criticism and integrated into a common level of knowledge and experience.
- Experiments, numerical calculations or analysis have to be reproducible and therefore all important steps must be recorded.
- Primary data as the basis for publications shall be securely stored as the publication itself.
- "appropriate methods are used and all results are consistently doubted by oneself,
- academic qualification work is actually based on personal contribution,
- preliminary academic work should be adequately considered and correctly cited,
- the authors listed in a publication have actually contributed substantially to the creation of the work,
- one's own research data can be checked and used by others within the framework of standards customary in the respective field,
- scientists and scholars who teach and instruct meet their responsibility for communicating these principles and ensure adequate supervision." (http:
  s://www.uni-goettingen.de/en/604506.html).

For more details see the relevant publications about Safeguarding Good Scientific Practice of the German Research Foundation as well as the guidelines of the University:

- Memorandum on Safeguarding Good Scientific Practice by the Commission on Professional Self Regulation in Science (German Research Foundation, the English version starts at page 61)
- DFG Leitlinien zur Sicherung guter wissenschaftlicher Praxis (2019)
- Rules Governing the Safeguarding of Good Scientific Practice (2016)
- Research Data Policy (2016)
6.1 Personal Sensitive Data with Disclosure Risk

The TextGrid Repository recommends to be very careful in dealing with sensitive data throughout the whole research lifecycle of data collecting, handling and publishing. Disclosure risk is not only an issue for data allowing directly discovering of personal or sensitive data, but also indirectly by combination with other data.

Sensitive data can be defined as data that for ethical or legal reasons has to be protected against disclosure risk. Safeguarding of sensitive data may also be related to personal privacy or proprietary issues. Due to its open access commitment the TextGrid Repository excludes publication of sensitive data and data with disclosure risk which needs a restricted access or are not allowed to be published according to applicable law. The depositor and author has to take care of legal and ethical criteria related to personal privacy – according to applicable law. Authors and depositors have to be aware of new regulations as consequence of the EU data protection directive.

DARIAH-DE as service provider for the TextGrid Repository follows the Code of Conduct for DARIAH Services and will respect the relevant legal frameworks for the protection of personal data, especially the EU data protection directive. The way DARIAH-DE deals with private data itself is furthermore described in the DARIAH-DE Privacy Policy.

The depositor of data is informed by the TextGrid Terms of Use that he/she is responsible for ensuring that the permanent publication does not violate any third-party rights or copyright laws. This applies to all supplied text, image, sound, or other data formats or forms. Even if only individual data objects are subject to restrictive rights, publication is not possible. For personal data a fully informed consent is necessary for publication and applicable law is to be considered by author and depositor. This has to be presented in writing to DARIAH-DE.

If permissions are not given for open access publication, non sensitive data are to be seperated from sensitive data with disclosure risk for publishing. Sensitive data may will have to be published outside in a secure environment – not in an open acess repository as the TextGridRep. Measures and legal implications in case of non-compliance are stated in the Terms of Use. In case of dealing with sensitive data and data with disclosure risk legal advice is recommended.

More information and recommendations are available online:

- General Data Protection Regulation (European Law, EUR-Lex)
- Data Protection in the EU (European Commision)
- Law topic data protection (European Commission)
- Legal grounds for processing data (European Comission)
- Legal grounds processing sensitive data (European Comission)
- Portal for licence information on research data (DARIAH-DE)
- Guide on legal issues for the humanities (DARIAH-DE Working paper by Paul Klimpel and John H. Weitzman, language: german)
- Data licences for research data in the humanities (DARIAH-DE Working Paper by Beer, N. et al., language: german)
- Legal framework for research data (DataJus Project)