IDS
OPEN
ONLINE-ONLY PUBLIKATIONEN
DES LEIBNIZ-INSTITUTS FÜR DEUTSCHE SPRACHE

Standardising language data through the conversion pipeline TEIWorLD

Jennifer Ecker

Abstract The conversion of data into a standard format is a crucial step in many research workflows. Standardisation enables data exchange, reuse, and analysis, which are essential for advancing knowledge in various fields. In this publication, we describe the conversion pipeline TEIWorLD (TEI Workflow for Language Data) that transforms written and spoken language data into standardised formats, specifically I5/TEI P5 XML for written data and ISO/TEI Transcriptions of Spoken Language for spoken data. The pipeline leverages existing tools to convert specific formats into these standards, with an additional transformation step for written data into the archival I5 (short for IDS TEI P5) format used at the Leibniz Institute for the German Language (IDS). We also present two use cases that demonstrate the practical application of standardisation with our conversion pipeline TEIWorLD in language data management on a corpus consisting of more than one format, enabling researchers to efficiently analyse and share their data.

Keywords Keywords conversion, formats, pipeline, use case, TEIWorLD

Table of contents

1. Introduction[1]

The variety of formats used for written and spoken/transcribed data is a significant factor in the heterogeneity of language corpora. Projects often employ different tools to analyse their data, resulting in a variety of formats. While this diversity enables flexibility and adaptability in projects, it also poses challenges for interoperability and data exchange between different projects. However, standardisation can bring significant benefits, facilitating collaboration and data exchange, improving data quality, and reducing the cost of data processing and analysis. For this purpose, we developed the conversion pipeline TEIWorLD,[2] which transforms a variety of different formats for spoken and written language into the standardised formats ISO/TEI Transcriptions of Spoken Language (from here on referred to as TEISpoken) (Schmidt 2011; ISO 2016) and I5.[3] For archiving written data, a component of the pipeline converts TEI P5 XML[4] documents to the format used at IDS, the I5 format (Lüngen/Sperberg-McQueen 2012), which was developed by IDS based on TEI P5.

Tools related to TEIWorLD that handle conversions include Salt'N'Pepper (Zipser/Romary 2010), WebLicht[5] (Hinrichs/Zastrow/Hinrichs 2010), and LAPPS Grid[6] (Ide et al. 2016). Pepper[7] is an open-source framework developed to convert corpora from one linguistic format into another. It leverages the intermediate model Salt, which simplifies the conversion process by reducing the number of direct mappings required between formats. Although Pepper supports many different formats, it currently only supports conversions from TEI P5 XML documents to other formats, but not the other way around. Nevertheless, users have the option to create customised modules to extend its functionality and enable conversions into TEI P5 XML documents. However, TEIWorLD focuses on standardised output formats and does not just provide conversion to and from different formats.

WebLicht (Web-Based Linguistic Chaining Tool) is a web-based platform that provides users with access to a wide range of language processing tools. These tools enable complex processing tasks, such as tokenisation, syntactic analysis, and named entity recognition, to be performed on various languages. With pre-configured pipelines or assembled process chains, users can process their own data. The output of WebLicht is not standardised into XML documents based on TEI P5 for written language data. However, there is an internal processing and XML data exchange format (Text Corpus Format, TCF[8]) that is used as the input and output format for the components. In addition, there are several WebLicht tools that convert spoken data formats into TEISpoken. On the one hand, WebLicht and TEIWorLD differ in the specialisation of the output formats. On the other hand, both can convert written and spoken language data formats.

[1] The development of TEIWorLD is a joint effort with Pia Schwarz and Rebecca Wilm (Leibniz-Institute for the German Language). This publication is based on work within Text+, funded by the Deutsche Forschungsgemeinschaft (DFG, project number 460033370) as part of the German National Research Data Infrastructure (NFDI e. V.).

[2] https://teiwrld.github.io/TEIWorLD/ (last access: 21-5-2025).

[3] https//www.ids-mannheim.de/en/digspra/corpus-linguistics/projects/corpus-development/ids-text-model/ (last access: 21-5-2025).

[4] https://tei-c.org/guidelines/p5/ (last access: 21-5-2025), https://github.com/TEIC/TEI (last access: 21-5-2025).

[5] https://weblicht.sfs.uni-tuebingen.de/weblicht/ (last access: 21-5-2025).

[6] https://www.lappsgrid.org/ (last access: 21-5-2025).

[7] https://corpus-tools.org/pepper/ (last access: 21-5-2025).

[8] It is fully compatible with the Linguistic Annotation Format (LAF) and the Graph-based Format for Linguistic Annotations (GrAF), which were developed by the ISO/TC 37/SC 4 technical committee.

LAPPS (The Language Applications) Grid provides a framework within which users can access, combine, and use many different language processing tools and resources. It is able to orchestrate global access to language resources and processing functions by users in a virtual network of servers and services worldwide. Moreover, the LAPPS Grid enables users to customise their experience by adding their own language resources, services, and entire service grids, tailored to their specific needs. LAPPS Grid primarily utilises the LAPPS Interchange Format (LIF),[9] a JSON-based format for syntactic and semantic interoperability, for data exchange between its services. In contrast to LAPPS, TEIWorLD's main task is format conversion into standardised formats, but it could be upgraded with natural language processing tools that extend the provided metadata in future versions of the tool.

The remainder of this publication is structured as follows: Section 2 provides a detailed description and a schematic representation of the conversion pipeline, including an illustration of the integrated metadata and the components of the pipeline. Section 3 discusses the formats supported by the pipeline, both as input and output. Section 4 presents two use case scenarios, one focused on spoken data and the other on written data, demonstrating the pipeline's value for researchers working with diverse formats. Finally, Section 5 concludes the contribution by summarising the key points and outlining potential future developments. The description of the conversion pipeline TEIWorLD given here represents its current status and functionality. It should be noted that the pipeline is under continuous development and extension. Therefore, additional features and capabilities may be released in future versions.

2. Pipeline description

The conversion pipeline TEIWorLD is based on a modular architecture that reuses existing tools for transforming written and spoken language data into standardised formats. This approach allows for efficient reuse and integration of established solutions, as it reduces redundant development. An overview of the pipeline architecture is illustrated in Figure 1, which provides a visual representation of the components and their interactions. The components of the pipeline have been carefully chosen, including TEICORPO and TEIGarage, which are described in detail in Section 2.2. First, an orchestrator decides whether the input formats are written or spoken language based on their file extension. On the one hand, if the input is a spoken language format, the file or files are transformed with the TEICORPO component. The component outputs the final format, which is TEISpoken. On the other hand, if the input is a written language format, the file or files are transformed with the TEIGarage component in a first step. The component produces TEI P5 XML documents. In a second step, these documents are further processed with the P5ToI5 component. In addition, the user can populate a JSON file with project specific metadata (see Section 2.2) which will be included in the header of the resulting I5 format. The integration of the components ensures a seamless conversion of different input formats into the desired output formats, specifically TEISpoken for spoken language data and I5, with the intermediate TEI P5 XML document, for written language data.

[9] http://vocab.lappsgrid.org/schema/lif-schema.json (last access: 21-5-2025).

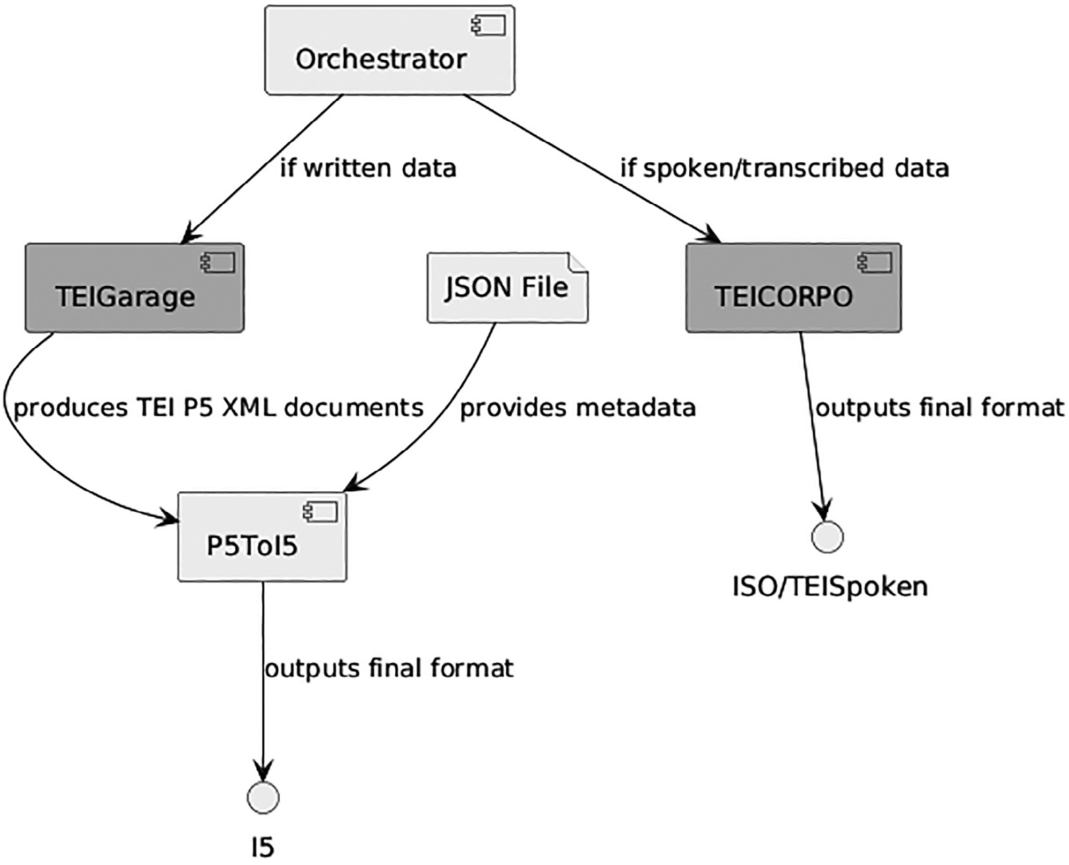


Figure 1: Schematic representation of the components of TEIWorLD

As a member of the NFDI initiative Text+,[10] which focuses on text and language-based research data, we adhere to established standards for metadata. To facilitate the storage of research data in Text+ member institutions' repositories, Text+ has established a defined policy and guidelines for providing metadata. These guidelines ensure a smooth and efficient data submission process,[11] helping researchers to properly document their data and make it easily discoverable, accessible, and reusable. Following these guidelines will help users to properly document their research data, hence making it more discoverable, accessible, and reusable by a wider research community. Even if users do not intend to store their data in a Text+ member institution's repository, the provision of metadata remains a crucial aspect of the pipeline, as it enables effective data management, sharing, and reuse, and facilitates collaboration and reproducibility in research, regardless of the user's plans for data storage. The following subsections go into detail about the metadata retrieved from the JSON file and the key components of the pipeline.

2.1 Metadata

In order to ensure that research data can be easily found and reused, it is essential to provide a comprehensive description of the data using metadata. In Text+, all research data should be described using at least the metadata elements from DataCite.[12] To achieve this, Text+ recommends using the following 20 fields to describe research data, which are either mandatory (M), recommended (R), or optional (O):

[10] https://text-plus.org/en/ (last access: 21-5-2025).

[11] https://text-plus.org/en/daten-dienste/depositing/ (last access: 21-5-2025).

[12] https://datacite-metadata-schema.readthedocs.io/en/4.5/properties/ (last access: 21-5-2025).

- Identifier: Unique identifier for the resource (M),
- Creator: Author or creator of the resource (M),
- Title: Title of the resource (M),
- Publisher: Publisher of the resource (M),
- PublisherYear: Year of publication (M),
- ResourceType: Type of resource (M),
- Subject: Subject area (R),
- Contributor: Contributors to the resource (R),
- Date: Date of creation (R),
- Language: Language of the resource (O),
- AlternateIdentifier: Alternative identifier, if available (O),
- RelatedIdentifier: Identifier of related resources (R),
- Size: Size description (O),
- Format: Format of the resource (O),
- Version: Version number (O),
- Rights: License and copyright information (O),
- Description: Brief description of the resource (R),
- Geolocation: Geographic location of the resource (R),
- FundingsReference: Funding reference or grant number (O),
- RelatedItem: Related resources (O).

For TEIWorLD, we focus on the metadata that can be represented by the I5 elements, but we payed attention that the mandatory metadata fields from DataCite are integrated. The JSON file allows users of TEIWorLD to populate fourteen of the twenty metadata fields, which are then automatically incorporated into the conversion process and stored in the corpus metadata header, enabling a smooth transition from XML documents based on TEI P5 to I5. Example data are integrated into the JSON file to ensure good quality of the user input. Altogether, this ensures that the resulting data is not only accurately converted, but also richly described with relevant metadata, making it more discoverable and reusable. Through the JSON file, the metadata fields creator, title, publisher, publisher year, resource type, subject, contributor, date, language, size, rights, description, geolocation, and fundings reference can be populated. Both, the metadata field identifier and the metadata field version are part of the internal representation of the repository and are automatically filled in. The metadata field format does not concern the corpus metadata and the metadata for the texts are extracted from the respective input files. Related metadata, alternate identifier, and related identifier are not integrated.

2.2 Components

As a first step in the development process of TEIWorLD, we collected and tested already existing tools that perform format conversions into standardised formats (TEI). Since we did not want to start from scratch and build all the tools ourselves, and since reuse is beneficial, we first integrated the selected tools. Thus, the conversion pipeline contains existing tools for transforming written and spoken language data in one place. For written language data, the pipeline transforms the input into the I5 format, and for spoken language data, it transforms the input into TEISpoken. So far, the pipeline incorporates TEIGarage[13] (Parkoła et al. 2024) and TEICORPO[14] (Parisse/Etienne/Liégeois 2020). We are currently evaluating the feasibility of integrating TEIDrop, a component of EXMARaLDA.[15] The components already included are described in the following subsections.

2.2.1 Components for spoken language formats

In this section, we describe one tool for spoken language conversion (TEICORPO), which is integrated into TEIWorLD. TEICORPO facilitates the process of working with spoken language data, especially when integrating data from diverse transcription formats into the TEISpoken standard. It is designed to facilitate the conversion and utilisation of spoken language corpora. Furthermore, it enables the conversion of transcriptions created with alignment software such as CLAN, Transcriber, Praat, or ELAN into the TEISpoken format, which acts as a lossless pivot format outlined in its specification for spoken language use (ISO 2016). In many cases, backwards conversion is also possible, although there may be limitations depending on the target format. This helps maintain the integrity and accessibility of linguistic data throughout the research process.

2.2.2 Components for written language formats

TEIGarage is a versatile tool based on OxGarage[16] with a web service[17] available. It is designed to facilitate the transformation of diverse written language formats into a range of output formats, including TEI P5 XML documents, which is a crucial step in our pipeline. The conversion process in TEIGarage ensures that each input document is transformed into a corresponding output document. By using the already existing tool TEIGarage, we can ensure that various input formats are standardised and converted into a consistent output format. The web service aspect of TEIGarage allows for easy accessibility and automation of the conversion process, making it a core part of our transformation pipeline. Furthermore, TEIGarage's ability to handle multiple input formats and produce TEI P5 XML output documents enables us to process a wide range of written language data, which is then further processed by our custom-built conversion component P5ToI5 to produce I5-compliant output. Some conversions with TEIGarage lead to invalid documents, that are well formed, but do not conform to the underlying document type definition (see Figure 2 for an example of the output from TEIGarage with a validation error). In this example, the dateline element is not allowed inside the p element as indicated by the red rectangles on the right and the red underline. Consequently, the subsequent conversion to I5 improves the overall accuracy of our pipeline. The P5ToI5 component ignores these mistakes and uses XSL transformations to transfer the needed data from elements of the P5 XML documents into the corresponding elements in the I5 format.

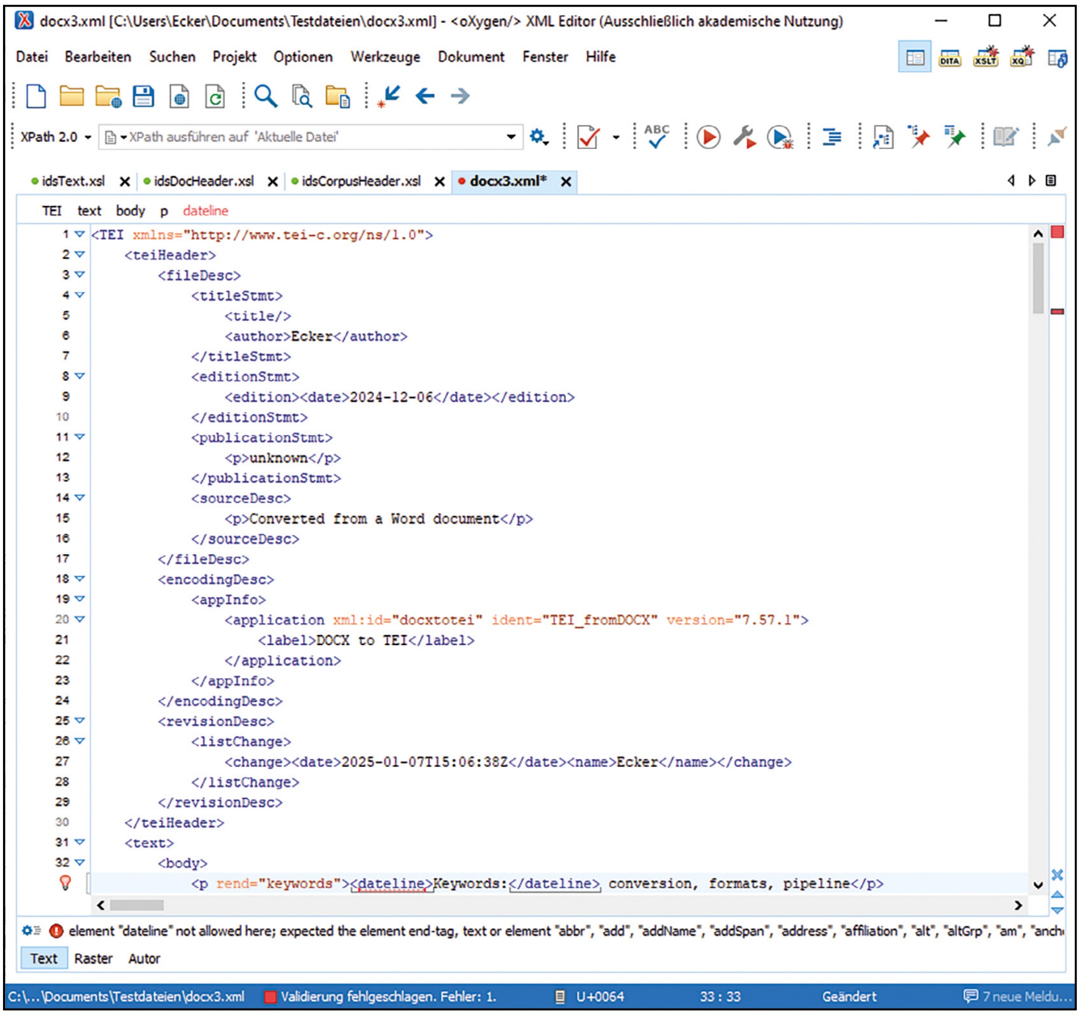
[13] https://github.com/TEIC/TEIGarage (last access: 21-5-2025).

[14] https://github.com/christopheparisse/teicorpo (last access: 21-5-2025).

[15] https://exmaralda.org/en/release-version/ (last access: 21-5-2025).

[16] https://github.com/TEIC/oxgarage (last access: 21-5-2025).

[17] https://teigarage.tei-c.org/(last access: 21-5-2025).



docx3.xml [C:\Users\Ecker\Documents\Testdateien\docx3.xml] - <oXygen/> XML Editor (Ausschließlich akademische Nutzung)
Datei Bearbeiten Suchen Projekt Optionen Werkzeuge Dokument Fenster Hilfe
XPath 2.0
XPath ausführen auf 'Aktuelle Datei'
idsText.xsl
idsDocHeader.xsl
idsCorpusHeader.xsl
docx3.xml*
TEI text body p dateline

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title/>
                <author>Ecker</author>
            </titleStmt>
            <editionStmt>
                <edition><date>2024-12-06</date></edition>
            </editionStmt>
            <publicationStmt>
                <p>unknown</p>
            </publicationStmt>
            <sourceDesc>
                <p>Converted from a Word document</p>
            </sourceDesc>
        </fileDesc>
        <encodingDesc>
            <appInfo>
                <application xml:id="docxtotei" ident="TEI_fromDOCX" version="7.57.1">
                    <label>DOCX to TEI</label>
                </application>
            </appInfo>
        </encodingDesc>
        <revisionDesc>
            <listChange>
                <change><date>2025-01-07T15:06:38Z</date><name>Ecker</name></change>
            </listChange>
        </revisionDesc>
    </teiHeader>
    <text>
        <body>
            <p rend="keywords"><dateline>Keywords:</dateline> conversion, formats, pipeline</p>
```

element "dateline" not allowed here; expected the element end-tag, text or element "abbr", "add", "addName", "addSpan", "address", "affiliation", "alt", "altGrp", "am", "anch

C:\...\Documents\Testdateien\docx3.xml | Validierung fehlgeschlagen. Fehler: 1. | U+0064 | 33 : 33 | Geändert | 7 neue Meldu...

Figure 2: Validation error in Oxygen

So far, we utilise the web service in TEIWorLD, but to guarantee long-term sustainability and flexibility, we plan to integrate the TEIGarage web service into our internal infrastructure, allowing for more efficient and reliable processing of large volumes of written language data. This will enable us to better monitor and maintain the conversion process, ensuring that our pipeline remains stable and efficient.

2.2.3 P5ToI5

After the transformation into P5 XML documents, an additional step in the pipeline is the conversion from P5 XML to I5 with the new P5ToI5 component integrated into TEIWorLD. It is based on Java with included XSL transformations and it integrates the metadata from the JSON file. This step is applied to written data that has been transformed into TEI P5 XML documents by the pipeline using TEIGarage. In contrast to TEI P5 XML documents, I5 represents many texts in one XML file and the P5ToI5 component transforms all documents into one I5 document. As outlined in Section 2.1, the corpus metadata is embedded in the header of the corpus. For the I5 files, it is crucial to assign unique identifiers to both the corpus as a whole and each individual text within it. These identifiers are automatically generated from different metadata fields extracted from the JSON file. The corpus identifier, in particular, is derived from a combination of the title, date, and description fields. Specifically, the first letter of the title is extracted, the last two digits of the year are retrieved from the date, and a hash code is generated from the description. To avoid duplicating existing corpus identifiers in the IDS archive,[18] careful naming conventions are essential. To illustrate the identifier generation process, consider a corpus with the title "Letter Corpus", dated 2022, and a description "This corpus represents a collection of letters from a German soldier writing to his mother". The resulting identifier for the corpus sigle would be generated from the first letter "L", the last two digits "22", and a five-digit part of the beginning of the hash code generated from the description (e.g. L22e298b). While the current identifier generation process is designed to minimise collisions, there is still a small risk of duplicate identifiers being generated, which would require manual intervention to resolve. As there are already existing identifiers for specific corpora in the IDS repository, this naming convention ensures that none of the occupied names is selected for new corpora. For this approach of generating identifiers to work for new corpora in the I5 format made by the pipeline, the identifiers of the corpora that end up in the IDS repository must be checked in a second step. As filler sentences like "This is a sample corpus." might lead to the same hash code, this needs to be double-checked. The document identifier on the other hand is generated with information from the creator and the title. The identifier consists of three capital letters derived from the first letter of the title and the initials of the creator's last name and first name. In addition, the generated corpus identifier is put in front of the document identifier separated by a slash (e.g. L22e298b/LMM) to serve as the document sigle. Finally, the different text sigles are numbered consecutively with a dot between the number and the document sigle (e.g. L22e298b/LMM.00001).

Three different XLS transformations are initiated by the Java-based component P5ToI5 to convert P5 XML to I5 XML. They are loaded and applied using the TransformerFactory class in Java. The first transformation is responsible for generating the header of the I5 corpus, which includes the metadata described in Section 2.1. The second transformation generates the header of each document within the corpus, while the third transformation transforms the text content for each input file. The XSL transformations are parameterised, allowing the pipeline to pass in values such as the different identifiers and metadata extracted from the JSON file. This enables the transformations to generate unique identifiers and incorporate metadata into the I5 XML documents. Finally, the outputs from the three transformations are combined to form the final I5 document. The resulting I5 document contains all the necessary metadata and text content, and is formatted according to the I5 DTD. The combination of the outputs from the three XSL transformations is done by concatenating the strings generated by each transformation, resulting in a single string that represents the complete I5 document. This final document is then written to a file, which can be used for archiving or further processing. The use of XSL transformations provides a flexible and efficient way to convert P5 XML documents to I5 XML documents, and allows the pipeline to generate high-quality I5 XML documents that meet the requirements of the IDS repository.

3. Formats

The conversion pipeline TEIWorLD is designed to handle a wide range of input formats and produce standardised output formats, enabling data exchange and integration across different research projects and systems. The diversity of formats used in research presents both opportunities and challenges, as it allows researchers to choose the best format for their specific needs, but also creates difficulties when sharing and combining data from

[18] www.ids-mannheim.de/en/digspra/corpus-linguistics/projects/corpus-development/archive/ (last access: 21-5-2025).

different sources. In the following subsections, we will provide a brief overview of the various input and output formats that the pipeline supports. Figure 3 shows a graphical overview of the already supported formats.

To date, the pipeline transforms the following spoken input formats into TEISpoken: EAF, TextGrid, CHAT, TRS, and QDPX, which have their individual format that is in all cases except for one based on XML, with TEICORPO. For written data, the pipeline transforms plain text and DOC/DOCX files into TEI P5 XML documents with TEIGarage, which is then further transformed into I5 with P5ToI5. We are currently expanding our format support, with additional formats being integrated. The following section provides an overview of the input formats already supported by TEIWorLD and the output formats that it generates.

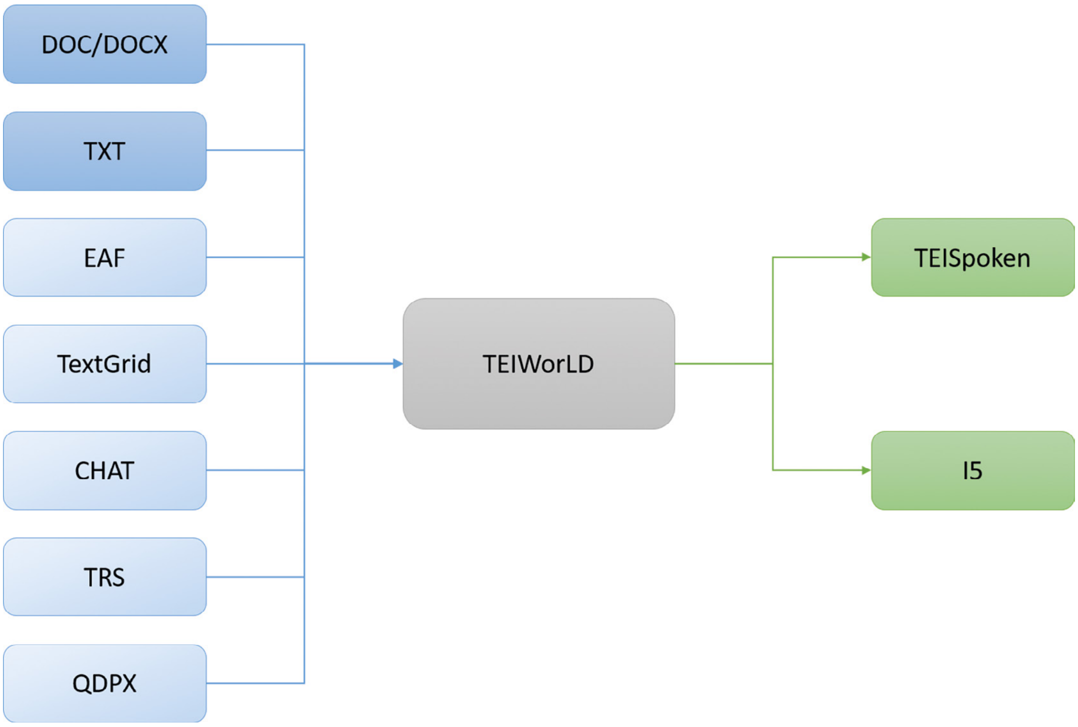


Figure 3: Supported input formats and output formats of TEIWorLD

3.1 Input formats

Formats that are used by researchers to store and manage their research data can be diverse, ranging from simple text files to complex databases, and from standardised mark-up languages such as XML and TEI to proprietary formats specific to particular software or tools. This diversity of formats brings both benefits and drawbacks, as it allows researchers to choose the best format for their specific needs, but also creates challenges when it comes to sharing, exchanging, and integrating data across different projects and research groups. In the following, the input formats integrated into the pipeline and the software that produces them are introduced.

ELAN[19] (EUDICO Linguistic Annotator) (Sloetjes 2013), developed by the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, is an annotation tool designed for annotating audio and video recordings. It allows the creation of annotations whose content is represented as Unicode text, with the resulting annotation documents being stored

[19] https://archive.mpi.nl/tla/elan (last access: 21-5-2025).

in an XML format, the ELAN Annotation Format or EUDICO Annotation Format (EAF), with the file extension *.eaf. EAF is a format for serialising objects that are part of the Abstract Corpus Model (Brugman/Wittenburg 2001). A schema, defined in an XML Schema Definition file, constrains XML elements and attributes of an EAF document and the annotations are contained in tier objects.

TextGrid is a file format with the file extension *.TextGrid used in the Praat[20] (Boersma 2001, 2013) open-source software for phonetic analysis and speech processing. Praat is designed to analyse, synthesise, and manipulate speech and other sounds and for creating graphics. TextGrid works as an annotation file that aligns linguistic information with audio data, allowing researchers to label and segment speech into tiers (e. g. phonemes, words, phrases) for a detailed analysis. The format is suitable for various linguistic and phonetic studies, and can be used for tasks such as speech transcription, prosody analysis, and phonetic alignment.

The CHAT (Codes for the Human Analysis of Transcripts) transcription and coding format (MacWhinney 2000) was developed as the standard transcription system for CHILDES (Child Language Data Exchange System), now a component of the TalkBank[21] system. With the help of the CLAN (Computerized Language ANalysis) program, data in the CHAT format with the file extension *.cha can be analysed. CHAT provides a standardised format for annotating spoken language interactions and produce computerised transcripts.

TRS with the file extension *.trs is an XML-based format of the Transcriber[22] tool (Barras et al. 2001) for creating and annotating speech transcripts for audio or video recordings. Transcriber allows manual segmentation and the transcription of long-duration broadcast news recordings, with the annotation of speech turns, topics, and acoustic conditions. Multilingual transcriptions with Unicode support are also allowed by the TRS format.

QDPX (the REFI-QDA-Project-Format) is a file format used by MAXQDA,[23] a software for qualitative and mixed methods data analysis. The QDPX format with the file extension *.qdpx stores all project-related data, including coded segments, annotations, and metadata, in a single file, serving as the project's central container. This format allows for systematic management, coding, and analysis of qualitative data, and is particularly useful for collaborative research. The QDPX file is designed to be used within MAXQDA, ensuring compatibility and consistency. However, it can be unzipped to obtain the transcription of recordings in the plain text format.

Plain text[24] files with the file extension *.txt are files that contain only text, without any formatting, special encoding or graphical elements. They store data as a simple sequence of characters and hence are universally compatible across different platforms and programs. Their compactness and human-readability make them a convenient and accessible format.

A DOCX[25] file is a Microsoft Word document in Open XML format, introduced with Word 2007 as a successor of DOC files. It is based on XML and uses ZIP compression to store multiple XML files and assets, reducing file size. DOCX files can contain formatted text, tables, graphics, diagrams, hyperlinks, comments, and macros.

[20] www.fon.hum.uva.nl/praat/ (last access: 21-5-2025).

[21] https://talkbank.org/ (last access: 21-5-2025).

[22] https://trans.sourceforge.net/en/usermanUS.php (last access: 21-5-2025).

[23] www.maxqda.com/ (last access: 21-5-2025).

[24] https://datatracker.ietf.org/doc/html/rfc2046#section-3 (last access: 21-5-2025).

[25] https://docs.fileformat.com/word-processing/docx/ (last access: 21-5-2025).

3.2 Output formats

A crucial aspect of the conversion pipeline TEIWorLD is the ability to produce standardised output formats, enabling seamless integration and compatibility of data from diverse sources. The following explores the key output formats utilised in the pipeline, including TEISpoken and I5 (with the intermediate XML document based on TEI P5). These formats play a vital role in facilitating data exchange, searchability, and representation within research systems.

Spoken language transcription is a crucial aspect of humanities research, but the lack of a standardised approach has resulted in a fragmented landscape of different transcription conventions, file formats, and tools. This makes it challenging for researchers to work with data from multiple sources, as they must navigate a complex array of incompatible formats and conventions. To address this issue, Schmidt (2011) has developed a set of guidelines for spoken language transcription, known as TEISpoken. This standard aims to provide a common framework for transcribing spoken language based on the TEI Guidelines (TEI Consortium 2025), enabling researchers to work with data from different sources in a more efficient and compatible way.

Similarly, in the domain of written language, standardisation is also an important aspect that allows data exchange and collaboration. For written language, TEI P5 XML has a comparable significance as TEISpoken for spoken language transcriptions. TEI P5 was developed as a toolkit by the Text Encoding Initiative (TEI) to provide a standardised framework for encoding digital texts. One of the key benefits of TEI P5 based XML is its ability to facilitate seamless data exchange between individuals and research groups using different software, systems, and applications.[26] By providing a comprehensive inventory of features commonly used in computer-based text processing, TEI P5 also serves as a valuable resource for developers designing new systems and creating new materials, even when data interchange is not the primary goal. The IDS stores its written language corpora in the IDS text model, utilising the I5 format. I5 is a TEI-based document grammar derived from TEI P5 through customisation using the TEI P5-specific ODD mechanism. Through the I5 format, the IDS corpora are searchable and displayed within the IDS's COSMAS II[27] research system. For the IDS's KorAP[28] research system, the I5 format is converted into the KorAP-XML format, which presents primary data, metadata, and annotation layers for a document in separate XML files. The I5 format serves as the primary format for ensuring representation in the research systems and functions as the primary ingest format for the German Reference Corpus (Kupietz et al. 2018). This allows data from external sources to be seamlessly integrated into the corpus.

4. Use case scenarios

We identify two distinct needs of researchers potentially using TEIWorLD. On the one hand, a researcher aims to standardise their data formats, but the data will not be stored in the IDS repository. They use the tool's output for their own research and store the data on their local servers or other external repositories. On the other hand, another researcher seeks to integrate their data into the IDS repository for collaborative analysis and sharing purposes. They utilise the pipeline to transform their data into standardised formats, which

[26] https://tei-c.org/release/doc/tei-p5-doc/de/html/AB.html (last access: 21-5-2025).

[27] www2.ids-mannheim.de/cosmas2/ (last access: 21-5-2025).

[28] https://korap.ids-mannheim.de/ (last access: 21-5-2025).

are archived in TEISpoken for spoken data or I5 for written data. In order to demonstrate the practical application and versatility of the conversion pipeline TEIWorLD, two use cases are presented in the next subsections: one for written data and one for spoken data. These examples illustrate how the pipeline transforms diverse data formats into standardised structures, enabling seamless integration into the IDS repository. The spoken data use case focuses on transcriptions of recordings from tasks like picture-naming and interviews, while the written data use case highlights the transformation of plain text and DOCX files into the archival I5 format. Both scenarios underline TEIWorLD's ability to handle different data types and in the written data case, it ensures their compatibility with state-of-the-art linguistic research tools such as KorAP.

4.1 Multi-format data integration for documentation of endangered languages

A linguist documenting an endangered language creates recordings in two different fieldwork settings. In one setting, the researcher conducts a picture-naming task with a single speaker and annotates the resulting audio files using Praat, which generates a TextGrid file. In the other setting, the researcher conducts a qualitative analysis in MAXQDA and produces QDPX files with transcription data. The pipeline processes the different formats as follows:

- For Praat annotations: The TextGrid files, which capture the picture-naming task by a single speaker, are directly processed with TEICORPO into TEISpoken preserving the alignment between the text and the audio annotations.
- For MAXQDA transcriptions: The QDPX file, which contains interview transcriptions with multiple speakers, is first unzipped to obtain the contained plain text file with the transcription of an interview. This is further transformed via the TEICORPO tool to transform it into the TEISpoken format.

The described workflow is pictured in Figure 4 on the left as a use case diagram. By converting both data types into TEISpoken, the pipeline ensures that diverse datasets, such as structured annotations from a picture-naming task and conversational data from interviews, are standardised into a unified format. Additionally, the TEISpoken format enhances data interoperability, supports integration with other linguistic resources, and facilitates long-term preservation and sharing of the collected data.

4.2 Standardising and preparing written corpora for repository integration and advanced linguistic research

A corpus linguist builds a custom corpus consisting of written data in plain text and DOCX formats. The goal is to standardise the formats and prepare the corpus for integration into the IDS repository for long-term storage. Later, the data can be made searchable within KorAP using annotations. The conversion pipeline TEIWorLD supports the initial steps of this process:

- Conversion to TEI P5 XML documents: Both, the plain text and the DOCX are subsequently converted into the TEI P5 XML format using the TEIGarage component. In this process, metadata such as author, edition, date of publication, or any other bibliographic information is extracted from the source documents and is put in TEI P5 XML encoding. This ensures that both the content and metadata are standardised and preserved in a widely accepted format.

- Conversion to I5: In this step, the text and metadata for the individual texts from the TEI P5 XML documents are integrated into the I5 format with the P5ToI5 component resulting in one file. Additional metadata at the corpus-level that are provided by the corpus linguist in a JSON file are also integrated into the I5 format as overarching metadata for the entire collection.

The described workflow is pictured in Figure 4 on the right as a use case diagram. Once the corpus is in the I5 format, it can be converted into KorAP-XML using, for example, IDS's KorAP-XML-TEI tool.[29] When integrated into KorAP, the platform displays linguistic annotations of tokens at several layers, such as part-of-speech tags, lemma annotations, and other linguistic features including dependency annotations on the sentence level, making the corpus more usable in advanced linguistic analysis and enabling sophisticated queries. In this use case, the conversion pipeline TEIWorLD standardises written data in plain text and DOCX format and enriches it both with text level and corpus-level metadata. The resulting I5 format prepares the corpus for integration into the IDS repository, while subsequent annotation in KorAP unlocks advanced research capabilities.

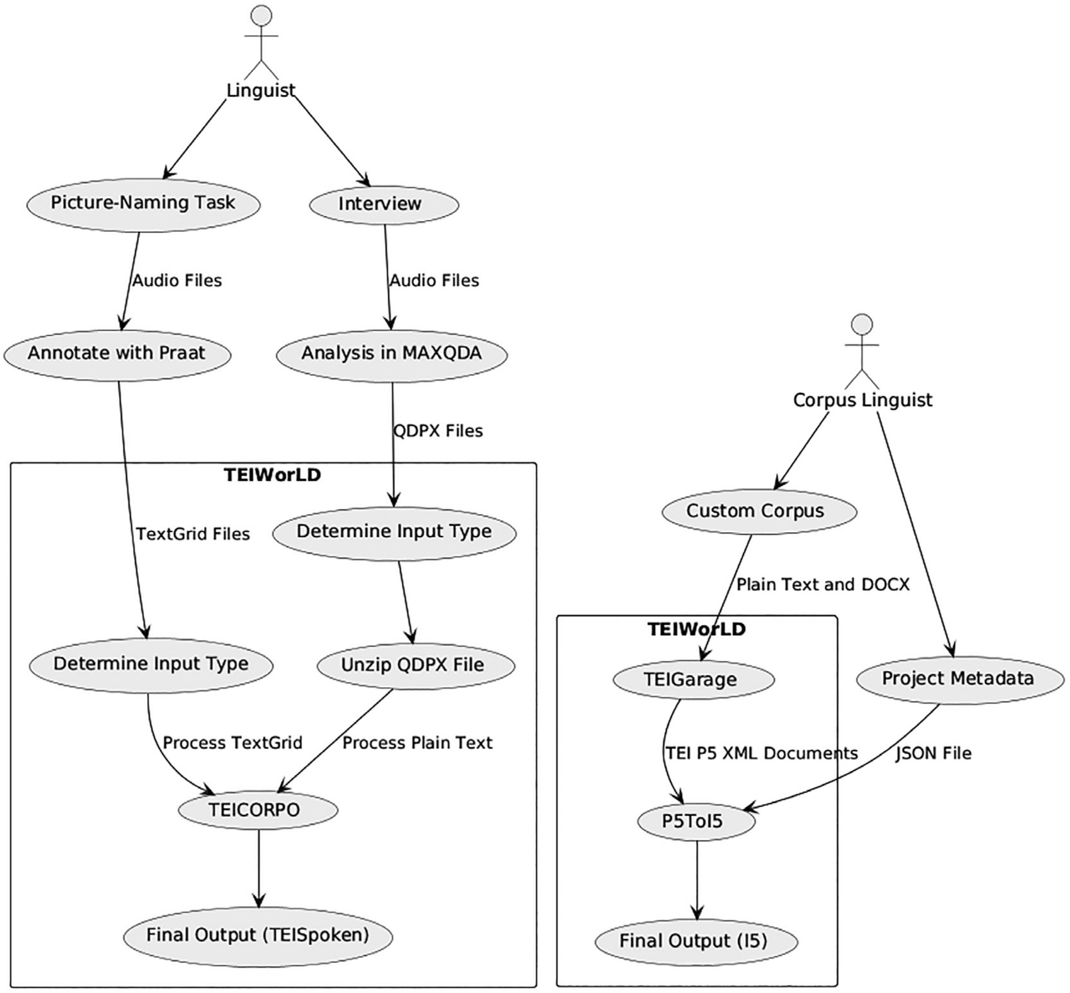


Figure 4: Use case diagrams of the workflows of both described use cases (on the left for spoken language and on the right for written language)

[29] https://github.com/KorAP/KorAP-XML-TEI (last access: 21-5-2025).

5. Conclusion and summary

This publication has shown how our conversion pipeline TEIWorLD for standardising language data is able to enhance research data management. It simplifies the work process in the sense that a researcher does not have to try different tools, but can use one that includes various options for the input format for spoken/transcribed data as well as written data. The implementation of such a pipeline not only simplifies the research process but also enhances the overall efficiency and productivity of researchers. With a single, comprehensive tool at their disposal, researchers can focus on higher-level tasks, such as data analysis and interpretation, rather than devoting excessive time and resources to data preparation and conversion. Originally designed to ease format integration into the IDS repository, TEIWorLD can be used by other researchers for their own projects to standardise their data. We provide our users with the standardised formats TEISpoken and I5, which are derived from TEI P5. Additional usage is the archiving of data in our repository.

Up to now, the TEIWorLD integrates existing components to transform written and spoken language data into standardised formats (TEIGarage and TEICORPO) and one additional component P5ToI5. The key points of the components are:

- TEICORPO: A tool that converts spoken language transcriptions into the TEISpoken format, which is a lossless pivot format. It supports conversion from formats used by CLAN, Transcriber, Praat, and ELAN.
- TEIGarage: A versatile tool that transforms diverse written language formats into TEI P5 XML documents with metadata from the documents. It has a web service available and can handle multiple input formats.
- P5ToI5: A new tool that transforms the output of TEIGarage into the standardised I5 format and enriches it with metadata provided by the user. So far, this tool works for plain text files and DOC/DOCX files that are converted into P5. There will be other text format conversions available soon.

By means of the use cases, we showed that our conversion pipeline TEIWorLD offers a robust solution for standardising and preparing both written and spoken language data with respect to its potential integration into the IDS repository and subsequent linguistic analysis. In the written data use case, a corpus linguist uses the pipeline to convert plain text and DOCX files into I5. This process integrates the text-level metadata extracted from the documents with the corpus-level metadata provided by the user so that the data is structured and ready to be integrated into the repository. The I5 format will allow not only long-term preservation of the corpus itself but also its later optional transformation into the KorAP-XML format. This will provide opportunities for automatic annotation and advanced linguistic analysis using the KorAP platform. In the spoken data use case, the pipeline handles the transcriptions of recordings from a variety of formats, such as TextGrid resulting from a picture-naming task and QDPX files from MAXQDA-supported interview transcriptions. These formats are converted into TEISpoken that standardises the spoken language data while preserving its attributes, such as speaker contributions, time alignment, and contextual metadata. By enabling the transformation of different written and spoken data types into standardised formats, TEIWorLD facilitates interoperability, enhances accessibility, and ensures that corpora are preserved and optimised for modern linguistic research. TEIWorLD can also be used for the researcher's own purpose so that the data does not have to be archived.

Future developments aim to extend the pipeline's support for additional data formats that can be handled by the tools TEICORPO and TEIGarage. On another note, we are currently evaluating whether to integrate TEIDrop for written data, and at a later stage for potentially other tools, into the pipeline to handle the formats that cannot be transformed by TEICORPO. Another task is to extend the P5ToI5 component to handle other input format conversions from TEI P5 XML documents, because the TEI P5 XML output of TEIGarage results in some small variations depending on the input format. So far, the metadata from the JSON file is only integrated into the I5 format, but it is planned to merge the metadata with TEISpoken. Furthermore, we are exploring the integration of tools for automatic transcription, which would significantly streamline the processing of spoken language data. The user would be able to insert video or audio files into the conversion pipeline that are automatically transcribed and then transformed into the TEISpoken format. The KorAP-XML-TEI tool is currently only available in the Perl programming language. There are plans to integrate a Python version into TEIWorLD. Users will then be able to choose into which format they wish to convert their data.

References

Barras, Claude/Geoffrois, Edouard/Wu, Zhibiao/Liberman, Mark (2001): Transcriber: Development and use of a tool for assisting speech corpora production. In: Speech Communication 33, 1–2, pp. 5–22.

Boersma, Paul (2001): Praat, a system for doing phonetics by computer. In: Glot International 5, 9, pp. 341–345.

Boersma, Paul (2013): The use of Praat in corpus research. In: Durand/Gut/Kristoffersen (eds.), pp. 342–360.

Brugman, Hennie/Wittenburg, Peter (2001): The application of annotation models for the construction of databases and tools. Overview and analysis of MPI work since 1994. In: Bird, Steven/Liberman, Mark/Buneman, Peter (eds.): Proceedings of the IRCS Workshop on Linguistic Databases: 11–13 December, 2001. Philadelphia: University of Pennsylvania.

Durand, Jacques/Gut, Ulrike/Kristoffersen, Gjert (eds.) (2013): The Oxford handbook of corpus phonology. (= Oxford Handbooks). Oxford: Oxford University Press.

Hinrichs, Marie/Zastrow, Thomas/Hinrichs, Erhard (2010): WebLicht: Web-based LRT services in a distributed eScience infrastructure. In: Calzolari, Nicoletta/Choukri, Khalid/Maegaard, Bente/Mariani, Joseph/Odijk, Jan/Piperidis, Stelios/Rosner, Mike/Tapias, Daniel (eds.): Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10). Malta: European Language Resources Association (ELRA), pp. 489–493.

Ide, Nancy/Pustejovsky, James/Cieri, Christopher/Nyberg, Eric/DiPersio, Denise/Shi, Chunqi/Suderman, Keith/Verhagen, Marc/Wang, Di/Wright, Jonathan (2016): The Language Application Grid. In: Murakami, Yohei/Lin, Donghui (eds.): Worldwide language service infrastructure. Proceedings of the 2nd International Workshop, WLSI 2015, Kyoto, Japan, January 22–23, 2015. Revised Selected Papers. (= Lecture Notes in Computer Science). Cham: Springer, pp. 51–70.

ISO (2016): Language resource management – transcription of spoken language. www.iso.org/standard/37338.html (last access: 21-05-2025).

Kupietz, Marc/Lüngen, Harald/Kamocki, Paweł/Witt, Andreas (2018): The German Reference Corpus DeReKo: New developments – new opportunities. In: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga,Takenobu (eds.): Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC '18). Miyazaki: European Language Resources Association (ELRA), pp. 4353–4360.

Lüngen, Harald/Sperberg-McQueen, Christopher M. (2012): A TEI P5 document grammar for the IDS text model. In: Journal of the Text Encoding Initiative 3, pp. 1–18.

MacWhinney, Brian (2000): The CHILDES Project: Tools for analyzing talk. Vol. 1: Transcription, format and programs. 3rd edition. Mahwah, NJ: Erlbaum.

Parisse, Christophe/Etienne, Carole/Liégeois, Loïc (2020): TEICORPO: A conversion tool for spoken language transcription with a pivot file in TEI. In: Journal of the Text Encoding Initiative: Selected Papers from the 2018 TEI Conference 13, pp. 1–38.

Parkoła, Tomasz/Stanisławczyk, Mariusz/Cummings, James/Burnard, Lou/Rahtz, Sebastian/Werla, Marcin/Mittelbach, Arno/Cayless, Hugh/Viglianti, Raffaele/Röwenstrunk, Daniel/Stadler, Peter/Ferger, Anne (2024): TEIGarage. https://zenodo.org/doi/10.5281/zenodo.11516696 (last access: 21-05-2025).

Schmidt, Thomas (2011): A TEI-based approach to standardising spoken language transcription. In: Journal of the Text Encoding Initiative 1, pp. 1–22.

Sloetjes, Han (2013): ELAN: Multimedia annotation application. In: Durand/Gut/Kristoffersen (eds.), pp. 305–320.

TEI Consortium (2025): TEI P5: Guidelines for electronic text encoding and interchange. https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf (last access: 21-05-2025).

Zipser, Florian/Romary, Laurent (2010): A model oriented approach to the mapping of annotation formats using standards. In: Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010, May 2010, La Valette, Malta. http://hal.archives-ouvertes.fr/inria-00527799/en/ (last access: 21-05-2025).

Bibliographic information

Information on the citation of this publication:

Ecker, Jennifer (2025): Standardising language data through the conversion pipeline TEIWorLD. (= IDS*open* 15). Mannheim: IDS-Verlag.

DOI https://doi.org/10.21248/idsopen.15.2026.54

Contact information

Jennifer Ecker
Leibniz-Institut für Deutsche Sprache
R 5, 6–10
68161 Mannheim
Germany
E-Mail: ecker@ids-mannheim.de

Imprint

Bibliographic information of the German National Library
The German National Library lists this publication in the German National Bibliography; detailed bibliographic data is available online at
http://dnb.dnb.de.

IDS-Verlag · Leibniz-Institut für Deutsche Sprache
R 5, 6–13 · 68161 Mannheim
www.ids-mannheim.de

IDS IDS-Verlag

Leibniz
Gemeinschaft

Publication series: IDS*open*: Online-only Publikationen des Leibniz-Instituts für Deutsche Sprache
Series editors: Norman Fiedler, Katrin Hein, Siegwalt Lindenfelser, Beata Trawiński
Copy editor: Melanie Kraus
Typesetter: Joachim Hohwieler

ISBN: 978-3-948831-78-3 (PDF)

ISSN: 2749-9855