TEITOK API

Maarten Janssen

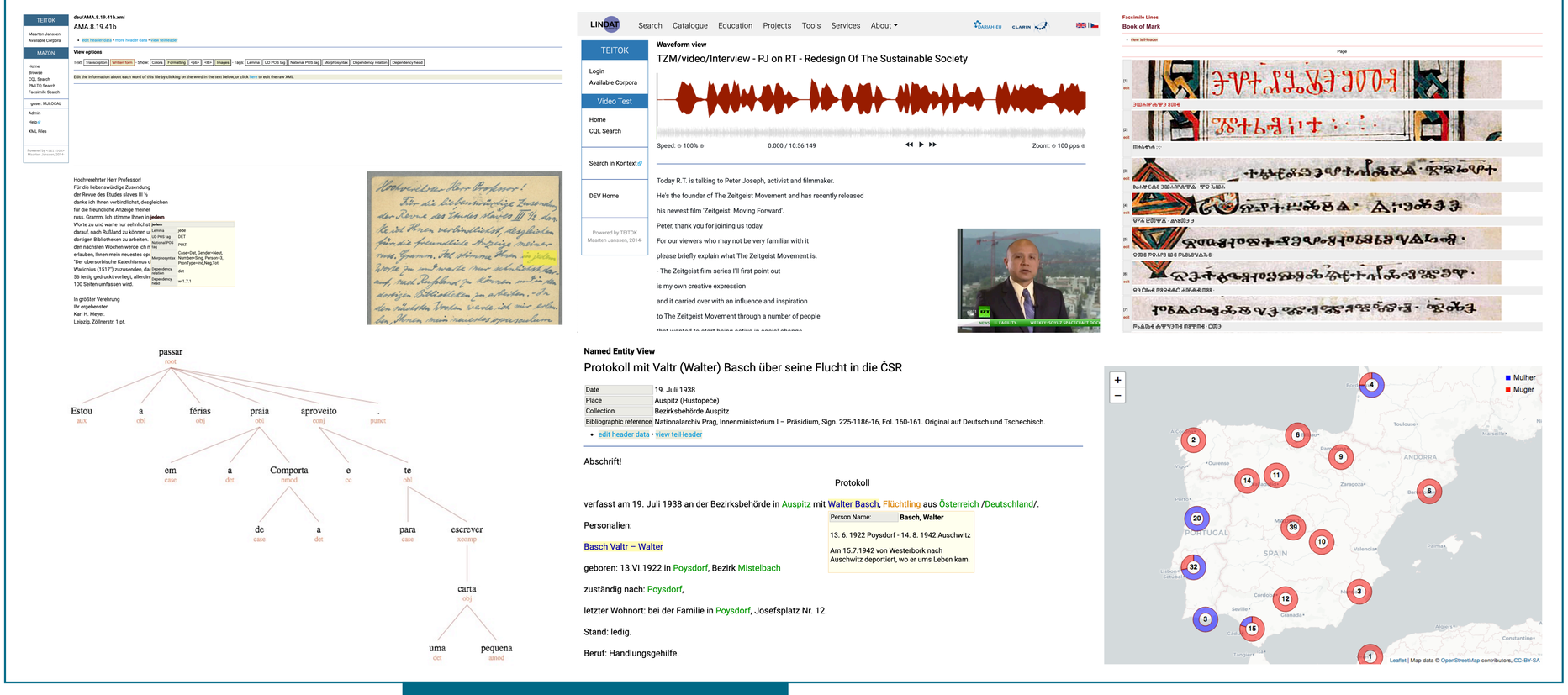
ÚFAL, Faculty of Mathematics and Physics, Charles University

<TEI:TOK>

TEITOK is an online platform for creating heavily annotated TEI/XML based corpora, visualising them online and making them searchable. Documents can contain stylistic and linguistic annotations, facsimile alignment, multimedia alignment, semantic markup, rich metadata, etc.

TEITOK has built-in editing functions to edit individual tokens, selections of tokens from a query, XML regions, run NLP processes, etc. These make it possible to edit XML document without having to look at the code or know any programming.

Richly Annotated TEI/XML documents



API

Partial Export

Export part of a TEI/XML document in a field-specific standard - such as CoNLL-U for NLP - in order to annotate using external tools, and import the resulting annotations back into the TEI/XML file.

Annotation tools

Manual Annotation with INCePTION, CoNLL-U Editor, CATMA, Brat, etc.

Automatic Annotion with SpaCy, Flair, Stanza, UDPipe, etc.

Custom Annotation with Python or R scripts.

Search Export

Export content from the corpus from a search query. The results can be used in external tools for statistics or visualisation.

Processing tools

Visualization with Stylo, Voyant, etc.

Statistical processing with Python or R scripts, QuitaUp, etc.

Start Processes

Start NLP processing steps such as tokenisation, lemmatisation, tagging, parsing, etc.

Document import

Import content in a number of different formats. All documents will be converted to TEI/XML before being added to the corpus using predefined conversion scripts.

External sources

Add spoken files in ELAN or EXMARaLDA

Add HTR files in PagesXML or ALTO

Add annotated files in VCF, Brat or CAS

Convert locally to TEI/XML

Script Repository

Scripts to import and export data from and to TEITOK using the API are often small. We are keeping a repository of scripts for specific purposes that people can use directly or use as a template for similar tasks. Examples:

- Tag Macedonian texts using CLASSLA
- Create a corpus from a folder of RTF files
- Provide ISO code for textual date metadata

API Advantages

- Programmable TEITOK corpora - both reading and writing
- Provides freedom to use any external tools in combination with TEITOK
- Increases reusability and fair-ness
- Allows backing up all corpus data
- Allows adding large volumes of data without going through the interface
- Allows modifying large collections of (meta)data
- All without impeding the GUI interface actions
- While keeping any complex TEI encoding

Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ / H2020 101004984 — CLS INFRA