Open source library improving networking, interoperability, standardization and flexibility of software for spoken language corpora

https://github.com/zumult-org/zumultapi

BACKGROUND

Web applications for access to spoken language corpora: the same functionality, but totally different technical basis strongly tied to the data format and user needs they were designed for (cf. Batinić et al. 2021)

Problem: Reuse of software components for new corpora types and new usage scenarios

Proposed solution: Object-oriented modeling and three-layer client-server architecture for corpus platforms

Use of Standards

PROJECT

ZuMult stands for German “Zugänge zu multimodalen Korpora gesprochener Sprache” (Access to multimodal corpora of spoken language, Fandrych et al. 2022) – funded by the DFG (LiS-program) https://zumult.org

Implementation / Use of Standards

ISO 24624:2016

Language resource management — Transcription of spoken language

Media:
- PCM-WAV/MP3 (Audio)
- MPEG-4 (Video)

Transcriptions and annotations:
- ISO 24624:2016 based on XML and Text Encoding Initiative (TEI), interoperable with the formats of the established transcription editors EXMARaLDA, FOLKER, ELAN, Praat and Transcriber
- Event und speaker metadata:
- XML, CMDI

Backend:
- Java EE Framework, REST API
- Search functionality realized with MTAS
- Open source Lucene-based search engine for querying text with multilevel annotations + CQP Query Language

Client/web application prototypes:
- jQuery, XSLT, HTML5, Bootstrap
- Data visualization with XSL, SVG, WebVTT

References:
