ANNOTATING COMPLEX DOCUMENTS WITH XSTANDOFF
Maik Stührenberg
Bielefeld University, Germany

Complex documents

- Documents often encode information in more than one way, typically both textual and visual (multimodal documents, e.g. instruction manuals). These multiple encodings bear information that may or may not be related to each other
- Already annotated documents may prevent the creation of additional markup layers due to possible overlaps and other validation issues

XSTANDOFF in brief

- XSTANDOFF is a meta-markup language for multiple annotation hierarchies (Stührenberg and Goecke 2008)
- The formal model is that of a generalized ordered-descendant directed acyclic graph (GODDAG, Sperber-McQueen and Huitfeldt 1999) supporting discontinuous annotation elements
- The serialization format is defined by an XSD 1.1 schema including assertions for rule-based validation, and makes use of XML’s inherent hierarchy (parent child relationship) and integrity features (xs:ID/xs:IDREFS type attributes)
- corpusData elements can be nested recursively, allowing even for corpora to be stored in a single file (including cross-references between corpus items)
- The primaryData element is optional, supporting annotations over real-time instantiated segments (e.g. in case of describing sensor data such as eyetrackers)
- Differentiation between annotation level (concept) and layer (serialization, see Witt 2004) is supported
- Elements of imported annotation layers can have optional meta, update, and delete child elements in addition to ISOCat attributes according to ISO 12620:2009
- Creation and visualization of XSTANDOFF instances is done by converting inline annotations via the XSTANDOFF-Toolkit introduced in Stührenberg and Jettka 2009

Segments and annotation

Segments (or markables) are a finite set of regions (spans) over the primary data that are used as anchors for one or more annotations. The segmentation mechanisms supported by XSTANDOFF are related to the primary data type:
- textual primary data is segmented by start and end values using character or byte positions
- video and audio primary data is segmented by start and end values using timecodes (or frame numbers)
- spatial primary data is segmented by using primitive shapes in conjunction with coordinates in space (Stührenberg 2013)
- already annotated primary data is segmented by using XPath 2.0 expressions or XPointer’s xpointer() scheme

Annotation layers bear the actual information added to the primary data. XSTANDOFF makes only little restrictions about the serialization of annotation layers:
- element and attribute names remain the same (including element hierarchy) but are slightly converted into a standoff serialization, i.e. no textual element content
- elements are linked to corresponding segments via XML’s inherent integrity feature (xs:ID/xs:IDREFS type attributes added to the annotation layer’s elements)

Application scenarios

- Text-image corpora, including discourse analysis of relations between different information encodings regardless of annotation software formats
- Annotation and query of multiple annotated corpora, including comparison of markup languages, annotation tools, and inter-annotator-agreement

Serialization

References