Annotating complex documents with XStandoff

Maik Stührenberg
Bielefeld University, Germany

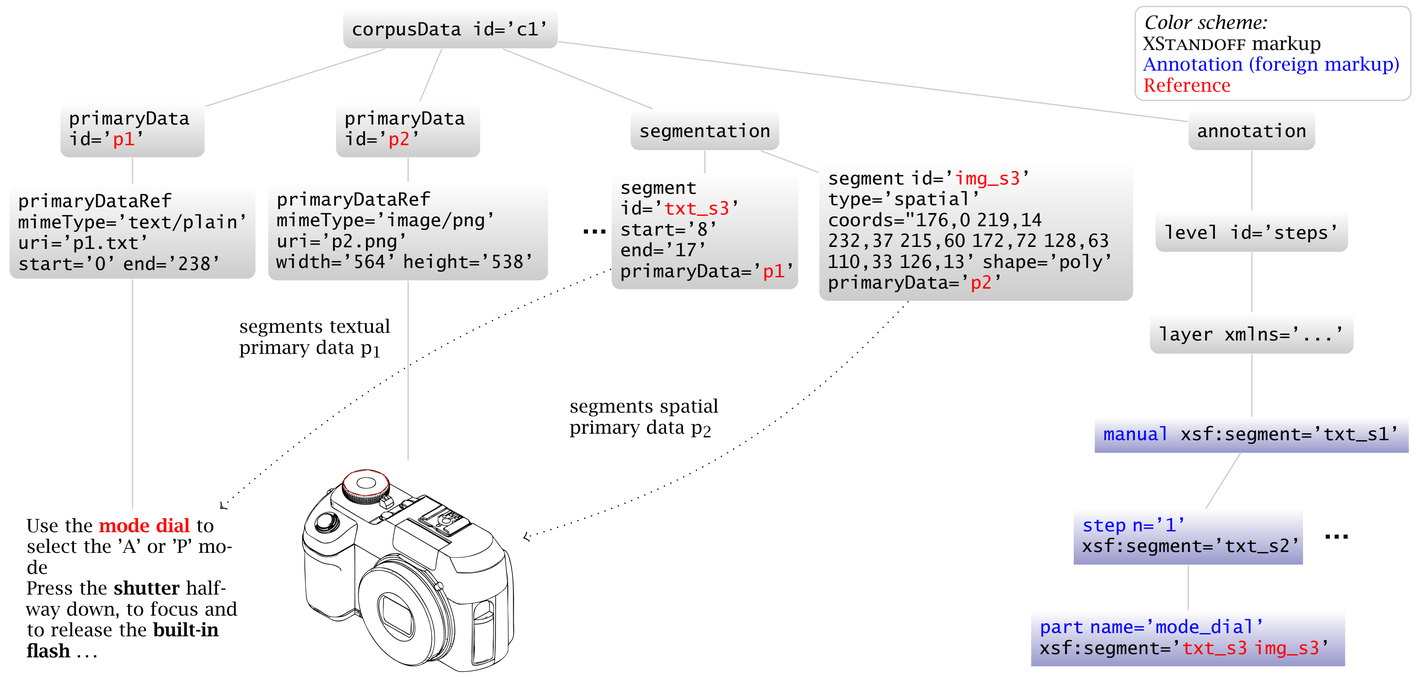
VolkswagenStiftung

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, Sperberg-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

Segmentation 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
  ```
  <segment xml:id="seg1" primaryData="txt" type="char" start="0" end="3"/>
  ```
- video and audio primary is segmented by start and end values using timecodes (or frame numbers)
  ```
  <segment xml:id="seg2" primaryData="video" start="310079" end="310302"/>
  ```
- spatial primary data is segmented by using primitive shapes in conjunction with coordinates in space (Stührenberg 2013)
  ```
  <segment xml:id="seg3" primaryData="img" type="spatial" shape="poly" coords="30,6 60,6 60,32 30,32"/>
  ```
- already annotated primary data is segmented by using XPath 2.0 expressions or XPointer's `xpointer()` scheme
  ```
  <segment xml:id="seg4" primaryData="pd1" target="xhtml:html/xhtml:body/substring(xhtml:div[1],4,5)"/>
  ```

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-aggreement

Serialization

```
<xsf:corpusData xsfVersion="2.0" xmlns:xsf="http://www.xstandoff.net/2009/xstandoff/1.1">
 <xsf:primaryData xml:id="p1">
  <xsf:primaryDataRef uri="camera.txt" mimeType="text/plain" encoding="utf-8" start="0" end="238"/>
 </xsf:primaryData>
 <xsf:primaryData xml:id="p2">
  <xsf:primaryDataRef uri="camera.png" mimeType="image/png" width="564" height="538"/>
 </xsf:primaryData>
 <xsf:segmentation>
  <!-- [...] -->
  <xsf:segment xml:id="txt_s3" primaryData="p1" start="8" end="17"/>
  <xsf:segment xml:id="img_s3" primaryData="p2" coords="176,0 219,14 232,37 215,60 172,72 128,63 110,33 126,13" shape="poly"/>
  <!-- [...] -->
 </xsf:segmentation>
 <xsf:annotation>
  <xsf:level xml:id="camera_manual">
   <xsf:layer xmlns="http://www.xstandoff.net/example/manual">
    <manual xsf:segment="txt_s1">
     <step n="1" xsf:segment="txt_s2">
      <part name="mode_dial" xsf:segment="txt_s3 img_s3"/>
     </step>
     <!-- [...] -->
    </manual>
   </xsf:layer>
  </xsf:level>
 </xsf:annotation>
</xsf:corpusData>
```

References

Sperberg-McQueen, C. M. and Claus Huitfeldt (1999). "GODDAG: A Data Structure for Overlapping Hierarchies". In: *Proceedings of ACH-ALLC1999. Joint International Conference of the Association for Computers and the Humanities (ACH) and the Association for Literary & Linguistic Computing (ALLC)*.

Stührenberg, Maik (2013). "What, when, where? Spatial and temporal annotations with XStandoff". In: *Proceedings of Balisage: The Markup Conference*. Vol. 10. Balisage Series on Markup Technologies. Montréal.

Stührenberg, Maik and Daniela Goecke (2008). "SGF – An integrated model for multiple annotations and its application in a linguistic domain". In: *Proceedings of Balisage: The Markup Conference*. Vol. 1. Balisage Series on Markup Technologies. Montréal.

Stührenberg, Maik and Daniel Jettka (2009). "A toolkit for multi-dimensional markup: The development of SGF to XStandoff". In: *Proceedings of Balisage: The Markup Conference*. Vol. 3. Balisage Series on Markup Technologies. Montréal.

Witt, Andreas (2004). "Multiple Hierarchies: New Aspects of an Old Solution". In: *Proceedings of Extreme Markup Languages*. Montréal.

→ Visit `http://xstandoff.net` for further information