Complex documents and XStandoff

Maik Stührenberg
Bielefeld University

Stand-off Annotation and GrAF Workshop
18.12.2013, CCeH, Cologne University
Complex documents

Two types of complex documents:

1. Multimodal documents with both textually and visually encoded information
2. Already annotated documents containing a markup hierarchy on their own
Use the mode dial (1) to select the 'A' or 'P' mode
Press the shutter (2) half-way down, to focus and to release the built-in flash (4)
Alternatively, use an external flash connected to the hotshoe (3)
Press the shutter(1) full-way down to take the picture
Example: Instruction manual

Use the mode dial (1) to select the 'A' or 'P' mode
Press the shutter (2) half-way down, to focus and to release the built-in flash (4)
Alternatively, use an external flash connected to the hotshoe (3)
Press the shutter (1) full-way down to take the picture

Numbers are used to reference parts of the camera in the image and in the textual instructions – there have to be better ways...
Another situation in which team A is not able to finish its move: Y tries to pass the ball through the small gap to X (instead of passing it to G) while Z is unintentionally obstructing Y's way.

But before the ball reaches X, D intercepts and passes the ball to A.
Example: Soccer match analysis

Another situation in which team A is not able to finish its move: Y tries to pass the ball through a small gap to X (instead of passing it to G) while Z is unintentionally obstructing Y's way.

But before the ball reaches X, D intercepts and passes the ball to A.

Where is the player named G on the graphic?
Example: Soccer match analysis

Another situation in which team A cannot finish its move: Y tries to pass the ball through the small gap to X (instead of passing it to G) while Z is unintentionally obstructing Y's way.

But before the ball reaches X, D intercepts and passes the ball to A.
Multimodal documents

Example: Soccer match analysis

Another situation in which team A is not able to finish its ball to X (instead of passing it to G) while Z is unintentionally obstructing Y's way.

But before the ball reaches X, D intercepts and passes the ball to A.

In which kind of relationship are the textual and the graphical representations?
Pre-annotated primary data

• Most NLP tools use raw text as input data instead of already annotated documents
• Even if tools accept documents of a given markup language, the addition of annotation layers may fail because of overlapping markup
• However, using XPath 2.0 or XPointer expressions it is possible to traverse pre-annotated markup
XStandoff

- Standoff meta annotation
- XSD 1.1 schema
- Multiple primary data files
- Segmentation mechanism for textual primary data, multimedia primary data, spatial primary data and pre-annotated primary
- Level (concept)/Layer (serialization) distinction
- Multiple annotation levels – no restriction about markup inventory
- ISOcat attributes for imported markup layers
Segmentation in XStandoff

- **Textual primary data (char or byte positions)**
  
  <segment xml:id="seg1" primaryData="txt" type="char" start="0" end="3"/>

- **Multimedia primary (time code)**
  
  <segment xml:id="seg2" primaryData="video" start="310079" end="310302"/>

- **Spatial primary data (shapes and coordinates)**
  
  <segment xml:id="seg3" primaryData="img" type="spatial" shape="poly" coords="30,6 60,6 60,32 30,32"/>

- **Annotated primary data (XPath 2.0 expressions)**
  
  <segment xml:id="seg4" primaryData="pd1" target="xhtml:html/xhtml:body/substring(xhtml:div[1],4,5)"/>
Use the **mode dial** to select the 'A' or 'P' mode.
Press the **shutter** half-way down, to focus and to release the **built-in flash** ...
Working with XStandoff Instances

Creation:
• Converting inline XML instances with the XStandoff Toolkit
• Using XML::Loy::XStandoff

Processing:
• Visualization with XStandoff Toolkit
• Query with XQuery scripts
Differences to GrAF

• Textual primary data may be included in the XStandoff instance
• Overlapping segments are possible without creating helper segments
• Graph structure not explicitly modeled as part of the serialization
• No restriction about annotation layers – no feature structures (of course you’re allowed to use them if you want to)
• Metadata may be included in XStandoff instances or standalone