

# Wake up, standOff!

Piotr Banski, Bertrand Gaiffe, Patrice Lopez, Simon Meoni, Laurent Romary, Thomas Schmidt, Peter Stadler, Andreas Witt

And special thanks to Luca Foppiano and Charles Riondet

# Overview

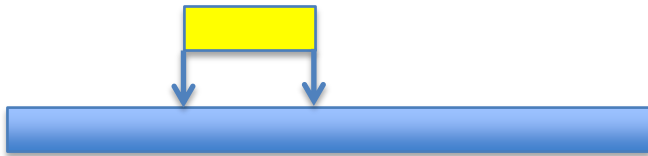
- The way towards a <standOff> element in the TEI architecture
  - Relation to ISO 24624 Transcription of Spoken Language
- Implementation issues
  - Reflecting the open annotation model
  - Open cans of worms (header, annotation body)
- Whither <standOff>?

# The simple picture



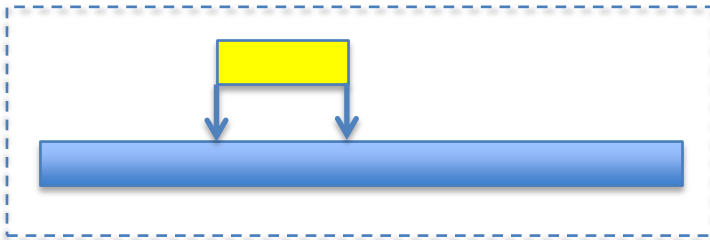
*Inline annotation:*

Intertwined with the source text



*Stand off annotation:*

Source text is referenced from outside



*Embedded stand off annotation:*

Stand off annotations attached to the same document as the source

# Why embedded stand-off annotation?

- Each time the source document is seen as the reference organisational unit
  - Corpus management
  - Transmission workflow
  - Multiple annotation layers
  - Competing annotations
    - E.g. Manual vs. automatic annotation

# Standoff: A long-standing issue

- The idea of standoff annotation is not new in general
  - Thompson & McKelvie, 1997
- Standoff annotation has been a core concept in the TEI guidelines since the beginning
  - Cf. Chapter: Linking, Segmentation, and Alignment
  - Availability of <anchor>, <span>, <interp>, <link>, @ana
- But: not integrated in the TEI architecture
  - Stand-off elements can appear anywhere in a TEI document
  - Usual trade-off between on-site vs. grouping (<back>)
- The NLP community has also developed its own means
  - GraF (Ide & Suderman 2007) , Paula (Zeldes et al. 2009), etc.
- Need for a proper, and inclusive, treatment of standoff annotations in the TEI
  - Better integration, more guidance

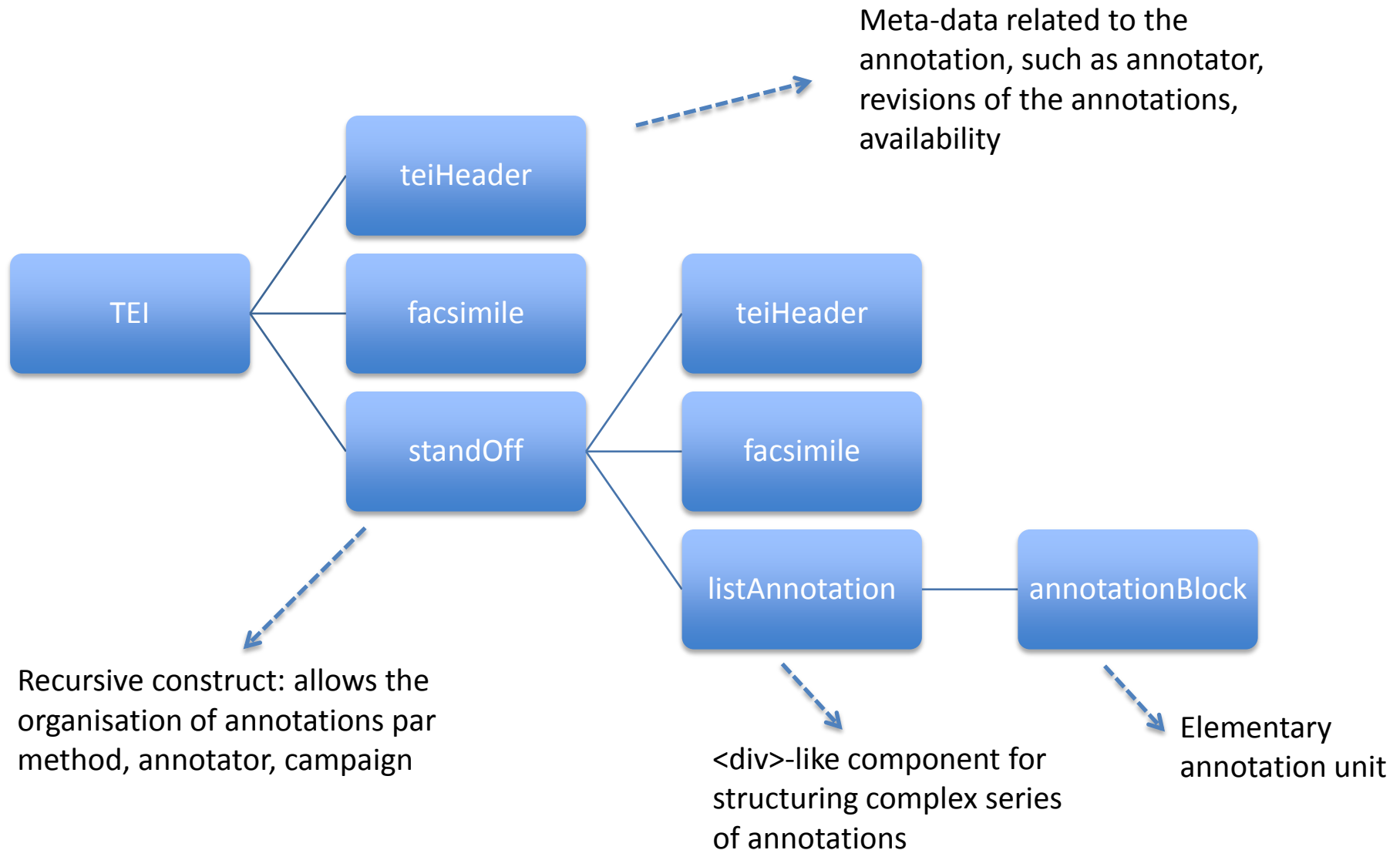
# Embedded standoff: Basic concept

- Building up an autonomous document containing primary source and additional annotations
  - Annotations are conveyed with their specific meta-data
  - Annotations have their specific place in the TEI document architecture
  - Standoff annotations may be recursively organized
  - Standoff annotations may point to textual as well as facsimile content
  - Well-defined elementary annotation units
  - Coherence with existing models (Open Annotation, ISO TC 37) should be ensured
- Typical use-cases
  - Annotated corpora
    - Treebanks
  - Text mining
    - Named entity recognition, keyword/terms extraction
  - Human annotations on a document
    - critical editions, patent examination, peer review...
- Strong relation with interlinear annotation

# Timeline

- 2011: Paper by Thomas Schmidt in jTEI (<https://jtei.revues.org/142>)
- August 2012: new tickets by Javier Pose (EPO)
- January 2014: Workshop in Berlin
  - Draft of a first proposal
  - Setting-up a github environment
- 2012-2016: ISO 24624 project (Editor: Thomas Schmidt)
  - Need for a annotation grouping component (<annotationBlock>)
- May 2015: Council meeting in Ann Arbor
  - Several updates to the proposal
  - Stabilisation of element names
- March 2016: TEI release 6.0.0
  - New element <annotationBlock> for interlinear annotation
- August 2016: publication of ISO 24624 Transcription of Spoken Language

# Annotations in TEI: <standOff>





# Application: interlinear annotation

- Encoding interlinear annotation as inline content (in <text>)

```
<annotationBlock who="#SPK0" start="#T9" end="#T12" xml:id="au1">
  <u xml:id="u1">
    <seg xml:id="seg45" type="utterance" subtype="declarative">
      <w xml:id="w43">Nee</w> <pc xml:id="pc3">,</pc> <w xml:id="w44">hab</w> <w
xml:id="w45">kein</w> <w xml:id="w46">Führerschein</w>
    </seg>
  </u>
  <spanGrp type="en">
    <span from="#T9" to="#T12">No, I don't have a driver's license.</span>
  </spanGrp>
  <spanGrp type="pos">
    <span from="#w43" to="#w43">NE</span>
    <span from="#pc3" to="#pc3">$,</span>
    <span from="#w44" to="#w44">VAIMP</span>
    <span from="#w45" to="#w45">PIAT</span>
    <span from="#w46" to="#w46">NN</span>
  </spanGrp>
</annotationBlock>
```

# Standoff interlinear annotation

- Encoding interlinear annotation as stand-off markup

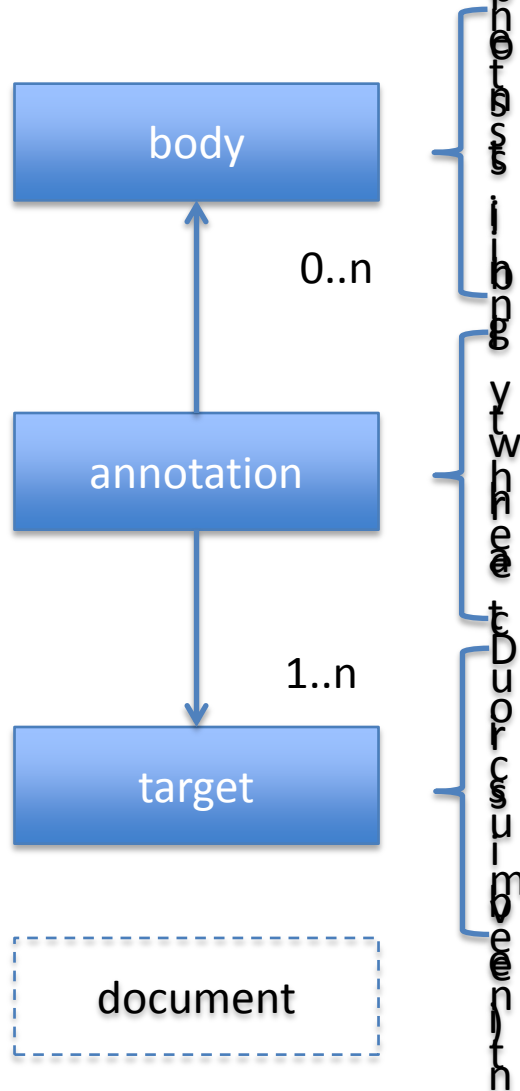
- In <standOff>

```
<annotationBlock inst="#u1">
  <spanGrp xmlns="http://www.tei-c.org/ns/1.0" type="en">
    <span from="#T9" to="#T12">No, I don't have a driver's license.</span>
  </spanGrp>
  <spanGrp xmlns="http://www.tei-c.org/ns/1.0" type="pos">
    <span from="#w43" to="#w43">NE</span>
    <span from="#pc3" to="#pc3">$,</span>
    <span from="#w44" to="#w44">VAIMP</span>
    <span from="#w45" to="#w45">PIAT</span>
    <span from="#w46" to="#w46">NN</span>
  </spanGrp>
</annotationBlock>
```

- In <body>

```
<u xml:id="u1" who="#SPK0" start="#T9" end="#T12">
  <seg xml:id="seg45" type="utterance" subtype="declarative">
    <w xml:id="w43">Nee</w><pc xml:id="pc3">,</pc>
    <w xml:id="w44">hab</w> <w xml:id="w45">kein</w> <w
xml:id="w46">Führerschein</w>
  </seg></u>
```

# Going further: mapping the Open Annotation model



<bibl>, <person>, <place>, <fs>, <note>,  
<body>, MAF, SynAF

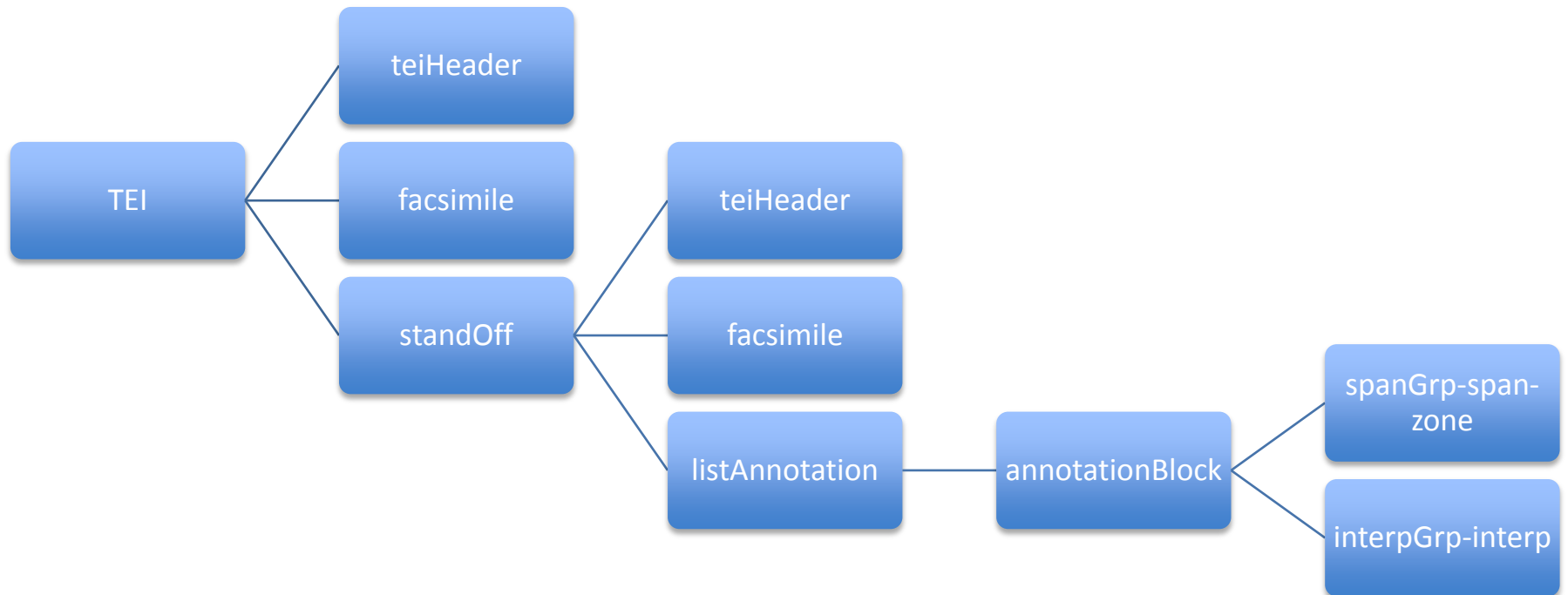
<interp type="" inst="" ana="">

<span type="" from="" to="">

<zone type="" corresp="#\_theSurface"  
ulx="1253" uly="802" lrx="22" lry="29"/>

Any TEI object (with @xml:id) or <surface>

# Going deeper into <standOff>



# Systematizing the use of `<span>` and `<interp>` in `<annotationBlock>`

- `<span>`
  - Close semantic to the notion of *target* in the OA model
  - Identifies a markable within the full-text of the document
  - Requires a precise guidance concerning pointing options
  - Kind-of business as usual
- `<interp>`
  - Extended usage
  - `@type`: provides the type of the annotation
    - Cf. `@type` on the parent `standOff` element
  - `@resp`: the entity who is responsible for this annotation
  - `@inst`: lists the components (span or surface) to be annotated
  - `@ana`: points to annotation content (body in OA speak)

# Prototypical example

Dates in a named entity recognition context

```
<annotationBlock>  
  <date xml:id="E4N1" from="1944-08-17" to="1944-08-25">  
    17 - 25 août 1944</date>  
  <interp ana="#E4N1" inst="#d1e173"/>  
  <span xml:id="d1e173" from="#E4T6" to="#E4T10" />  
</annotationBlock>
```

Great advantage on readiness and programmatic treatment

# Example from the ANR Termith project

```
<annotationBlock>
```

```
  <fs>
```

```
    <f name="lemma"><string>corpus</string></f>
```

```
    <f name="pos"><symbol value="NOM"/></f>
```

```
  </fs>
```

```
  <interp/>
```

```
  <span target="#t1"/>
```

```
</annotationBlock>
```

# Can we make the model more implicit?

```
<annotationBlock inst="#t1">  
  <fs>  
    <f name="lemma"><string>corpus</string></f>  
    <f name="pos"><symbol value="NOM"/></f>  
  </fs>  
</annotationBlock>
```

- Closer to the speech transcription version
- Risks:
  - Loosing the link with the OA model (hindrance to automation)
  - Allowing all types of possible (creative) encodings



# Issues (many)

- Which header do we need?
  - Standoff annotation usually requires very restricted meta-data
  - If we adopt the TEI header, we need to make it more flexible...
    - Should we have a convergence with biblFull (where profileDesc is missed, BTW, SF:533, deeply ambered)
  - Stand-off annotations may be generated by humans and machines
    - how to put <author> (editionStmt) and <appInfo> (encodingDesc) at the same place?
- How do we provide guidance concerning annotations?
  - Mapping the OA model to precise TEI constructs?
  - Allowing a wide variety of possible vocabularies depending on the use case?
    - TBX entries, MathML, full-text annotation (<body>?)
  - Aligning with the various ISO standards: MAF, SynAF and SemAF series

# Leaving dust under the carpet for today: pointing mechanisms

1. Offset based mechanism: *string-range(...)*
  - not stable in case the original text is modified. The annotation needs to be rebuilt
2. word tokenisation `<p><w>.</w><w>,</w></p>`
  - may generate an insane amount of data
3. `<span xml:id="s1" to="#a1"/> + <anchor xml:id="a1"/>`

Example:

`<p>....`

`<span to="#a2"/><span to="#a1"/>le petit chat<anchor xml:id="a1"/> est mort <anchor xml:id="a2"/>...`

`</p>`

- what about the purity of the source text?

# Next steps

- Finalising the content model of <annotationBlock>
  - Completely open model?
  - Constrained with specific model classes? (OA)
  - Alternation between the two (or more) options
- Gathering reference example from existing implementations
  - Istex, Termith, EPO, IDS
- Finalising the graft in the guidelines
  - Section in chapter 16 Linking, Segmentation, and Alignment?
- Don't give up the fight...

**MERCI !**