The Text Encoding Initiative: 30 years of accumulated wisdom and its potential for a bright future

Laurent Romary

The Text Encoding Initiative: 30 years of accumulated wisdom and its potential for a bright future

Laurent Romary, Inria

In the beginning

Text archives
Humanities
Standards
SGML

Not intended (immediately) for individual scholars

1. Novembre 1987: Vassar College, Poughkeepsie

A quick historical overview

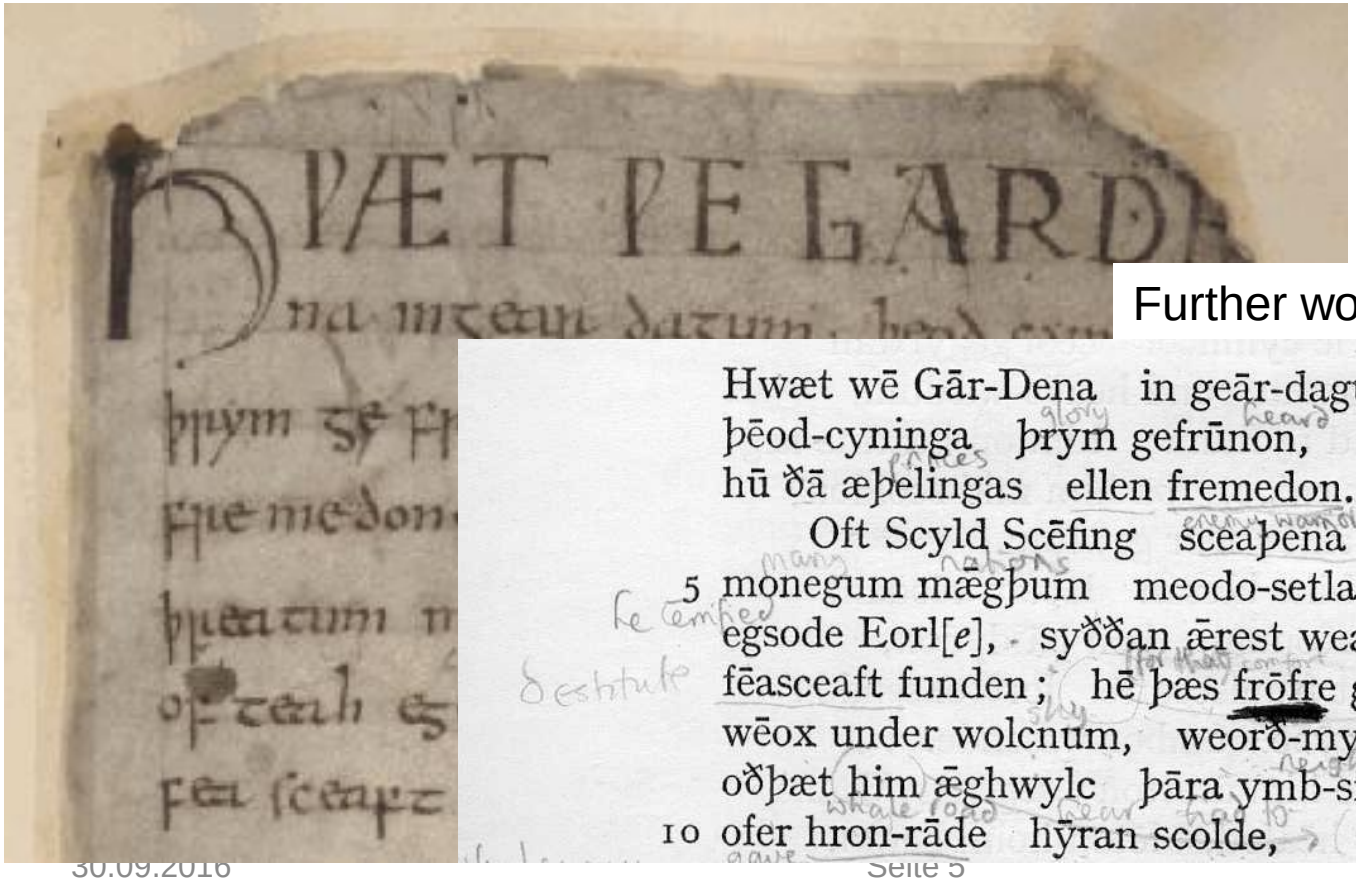
- 1960's — GML (Generalized Markup Language) by IBM
- 1970's & 1980's — ANSI initiates project to develop a Standard text-description language based on GML
- 1983 — SGML became an industry standard
- 1986 — SGML (Standard Generalized Markup Language) becomes an ISO standard: ISO 8879:1986
- 1987 — TEI (Text Encoding Initiative)
- 1990 — HTML 1.0 (HyperText Markup Language)
- 1992 — TEI edition P3 (Michael Sperberg-McQueen and Lou Burnard, eds)
- 1997/1998 — XML 1.0 (eXtensible Markup Language) (Tim Bray, Jean Paoli and Michael Sperberg-McQueen, eds)

TEI for digital scholarly work

- A trend towards digital curatorship
  - Describing digital sources: meta-data
  - Understanding and representing the structure of digital sources: content
  - Enriching (annotations, links), versioning, disseminating
- A wide user community
  - From individual scholars to large digitization projects

The standard scenario?

Digitizing source documents



Further work on documents

Hwæt wē Gār-Dena in geār-dagum
þēod-cyninga þrym gefrūnon,
hū ðā æþelingas ellen fremedon.
Oft Scyld Scēfing sceaþena þrēatum,
monegum mægþum meodo-setla oftēah;
egsode Eorl[e], syððan ærest wearð
fēasceaft funden; hē þæs frōfre gebād:
wēox under wolcnum, weorð-myndum þāh,
oðþæt him æghwylc þāra ymb-sittendra
ofer hron-rāde hȳran scolde,

TEI in a nutshell

- TEI namespace:
  - xmlns="http://www.tei-c.org/ns/1.0"
- TEI documentation:
  - http://www.tei-c.org/release/doc/tei-p5-doc/en/html/
- TEI processor, Roma:
  - http://www.tei-c.org/Roma/
- TEI document model
  - Read: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html
- TEI architecture: modules, classes
- TEI vocabulary: more than 500 elements…
  - Read: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html

TEI –core principles (1)

- The TEI document as a digital surrogate of a physical source
  - A TEI document is always part of a digital library workflow
    - Source – surrogate – enrichment – publication
    - Recorded in the header; encoded in the content
  - Born digital documents may as well encounter a succession of changes/versions
- The TEI document as an autonomous object in a DL workflow
  - Embedded meta-data + content
  - Multiple “hands”: annotation

TEI –core principles (2)

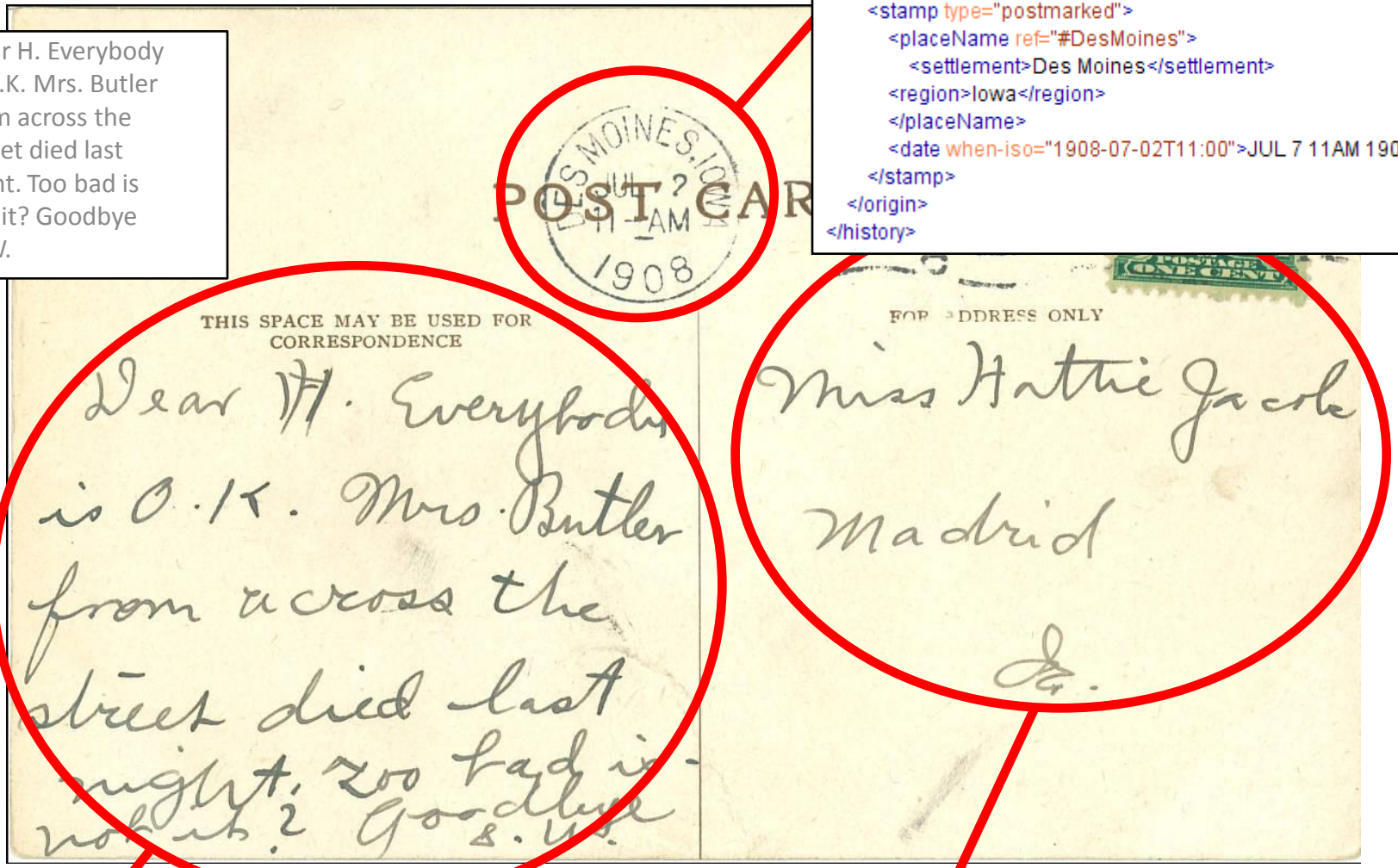
- Favoring the semantics rather than the layout
  - (quasi) No presentational construct
  - Publication requires a transformation stage (XSLT; ePub, pdf, HTML, etc.)
- Document structure
  - Macro-structure: front-body-back
  - Meso-structure: divisions, paragraphs/lists/figures/etc.
  - Micro-structure: in-line annotation mechanisms
    - Dates, names, notes, references, foreign expressions, etc.

All you can encode…

Examples

- Simple encoded text
  - The Little Riding Hood
- Scholarly paper
  - Towards Higher Ground
- Dictionaries
  - Larousse

Dear H. Everybody is O.K. Mrs. Butler from across the street died last night. Too bad is not it? Goodbye S. W.




```
<history>
  <origin>
    <stamp type="postmarked">
      <placeName ref="#DesMoines">
        <settlement>Des Moines</settlement>
      <region>Iowa</region>
      </placeName>
      <date when-iso="1908-07-02T11:00">JUL 7 11AM 1908</date>
    </stamp>
  </origin>
</history>
```

```
<div type="back" facs="#noble0337b">
  <div type="left">
    <salute>Dear <persName ref="#HJ">H</persName>.</salute>
    <p>Everybody <lb/>is O.K. Mrs. Butler <lb/>from across the <lb/>street died
      last <lb/>night. Too bad is <lb/>not it?</p>
    <signed>Goodbye <lb/><persName>S. W.</persName></signed>
  </div>
```

```
<div type="right">
  <p>
    <address>
      <addrLine>Miss <persName ref="#HJ">Hattie Jacobs</persName></addrLine><lb/>
      <placeName ref="#Madrid"><settlement>Madrid</settlement><lb/>
      <region>Ia</region></placeName>
    </address>.
  </p>
</div>
```

How do you manage this?

TEI as a standardization body (1)

- Consensus building
  - Community based decision process
- Maintenance
  - Two releases per year
- Publication
  - All TEI contents are available under the double CC-BY+BSD 2 clause license.

TEI as a standardization body (2)

- Organization
  - Consortium of institutional and individual members
  - Conference, journal (jTEI)
- The TEI at work
  - Board: administrative aspects
  - Technical council: coordinates the evolution of the TEI guidelines

Standardization work

- Community based workflow
  - Mailing list
  - GitHub – bugs and features
    - Recording all issues and decisions
  - Cf. ODD as a specification platform
- Deliverables
  - Documentation — TEI guidelines (more than 500 elements)
  - Schemas — DTD, RelaxNG, W3C
- Additional resources
  - Tools
    - Online customization: Roma
    - Online processing: OxGarage
    - Stylesheets (included in Oxygen)
  - Examples — TEI by Example

Special Interest Groups (SIGs)

- Computer-Mediated Communication (Michael Beißwenger)
- Correspondence. (Peter Stadler and Joachim Veit)
- Education (TBA)
- Libraries (Stefanie Gehrke and Kevin Hawkins)
- Manuscripts (Dot Porter and Gerrit Brüning)
- Music (Raffaele Viglianti)
- Ontologies (Oyvind Eide and Christian-Emil Ore)
- Scholarly Publishing (Daniel O'Donnell)
- TEI for Linguists (Piotr Bański and Andreas Witt)
- Text and Graphics (John Walsh and Martin de la Iglesia)
- Tools (Serge Heiden)

The TEI guidelines

- Online documentation
  - Prose description organized in chapters
  - Specific documentation for each element
  - Access to all examples from the guidelines
- Schema(s)
  - RelaxNG, W3C (, DTD)
  - Available online from the *Roma* interface
  - Delivered as packages (Ubuntu, Oxygen)
- The TEI guidelines as specifications
  - Documentation and schemas are generated from one single specification file
    - Expressed in a TEI sub-language: ODD (One Document Does it all)

Varieties of TEI Conformance

- Pure *TEI-all* subset
  - Most TEI projects
- TEI subset with extensions
  - Cf. TBX in TEI
- Non TEI document with TEI constructs (defined as an ODD)
  - EAG extensions in the EU Cendari project
- Non TEI document defined by means of an ODD document
  - E.g. ISO 24616:2012 Language resources management -- Multilingual information framework

The central role of customization

- Each TEI project starts with the definition of a customisation
  - Module selection
  - Sub-setting elements
  - Reducing possible values or content models
  - Adding, when necessary, new descriptive object
- ODD as the technical platform for customization

Consequences

- Family of formats
  - Comparison of two TEI-based projects through their ODDs
- Support for third-party projects
  - In-house maintenance of customization and documentation
    - E.g. DTAbF at the Berlin Brandenburg Academy of Sciences
  - Even non TEI application!
    - E.g. EAD n ODD
- Does not prevent one from knowing the TEI components…
  - Most projects can live with just a subset of the TEI ontology
    - With the strong possibility to impact on the guidelines themselves
      - E.g. <abstract>

EXPLORING NEW (AND OLD) REALMS

TEI: you're not alone…

- The hidden TEI: scientific information at the European Patent Office
- New components in teh TEI: <standOff>
- Working with others: ISO LMF

SCIENTIFIC INFORMATION?

Characterising scientific documents

- Expert documents describing a specific scientific and technical progress with respect to the state of the art
- Three main domains
  - Scholarly publications
  - Standardisation documents
  - Patents
- Some common characteristics
  - Authorship: the basis of scientific attribution
  - Structure: usually a formal internal organisation
  - Vocabulary: technical terms are essential to convey (or hide) meaning
  - Network of references: relating to the state of the art
  - Certification: workflow, responsibilities, metadata

Authorship

Publications - *The essence of publishing*

- Importance of attribution
- Reflects the context and time of the research (project, affiliation, biography)
- The hidden hand of reviewers

Standards - *Priority to the institution*

- Consensus building => large expert group
- ISO: no authors but project leaders
- W3C: editors

Patents - *A variety of roles*

- Applicant/inventor/representative
- Opponents
- … and *examiners*

Workflow

Publications - *Semi-formal*

- Traditional (vestigial?) concept of peer-review
- From author's initial manuscript to publisher's version
- Evolution in the role of each version (e.g. prior art)

Standards - *Very formal*

- Decision process reflecting membership structure
- ISO: WD, CD, DIS, FDIS, IS
- One single reference document

Patents - *Very formal*

- Review by patent examiners
- Coordination of multiple submissions: national, US, Europe, etc.
- Importance of initial submission date

The European Patent Office

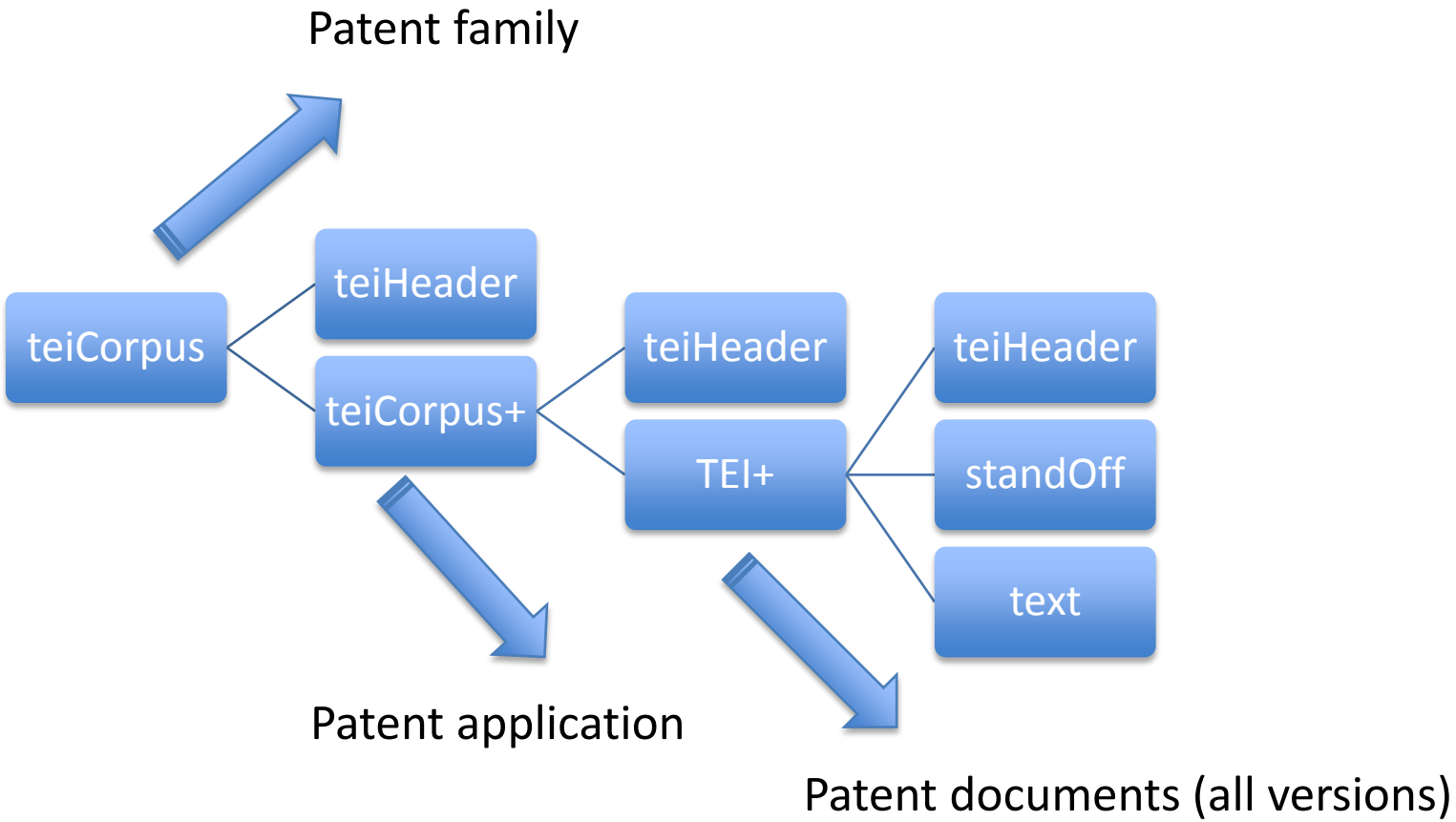
- The European one-stop shop for patent applications
- Examination of each application by experts from the field (examiners)
  - Based on existing patents as well as scholarly publications (aka *Non Patent Literature*)
- Some figures
  - Several thousands of examiners
  - 200 million documents
  - 2 billion annotations…

The (simplified) patent life-cycle

- Patent application in one or several patent offices
  - USPTO, Japan, EPO (directly or initiated in a specific country)
  - First application: reference date for the patent (“coming into force”)
  - Form a “Patent family”
- Examination process for one application
  - Search report, communications, decision, appeal, opposition
  - Patent documents may be revised at each stage
- Necessity to have a single model for dealing with all stages and versions

- The TEI appeared to be the optimal choice

The Patent Document Model



The situation so far

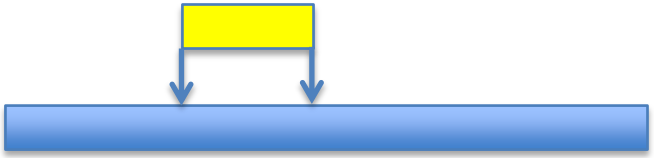
- Complete implementation in the back-office system
  - Integration of several so-far dispersed data-bases
  - First large-scale implementation of <standOff> (be patient!)
- Quite a few customisations – maintained in a reference ODD specification
  - Re-use of TEI attributes at various places
    - @type, @cert, @sortKey
  - Bibliographic references to patents
  - Complex classification mechanism (<classCodeGroup>)
- All in all a large scale demonstration of the TEI possibilities
- Next steps
  - All scholarly publications (NPL)
  - All official communications

WAKE UP STAND-OFF!

The simple picture

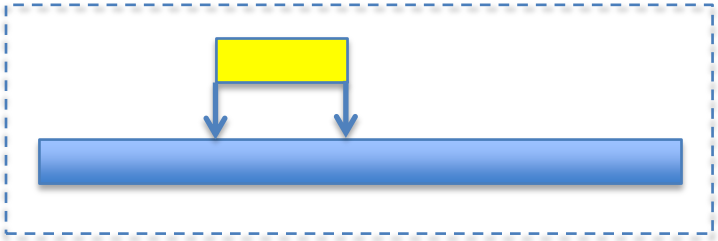
Inline annotation:

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?

- In line (!) with the TEI philosophy
- 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>, , <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

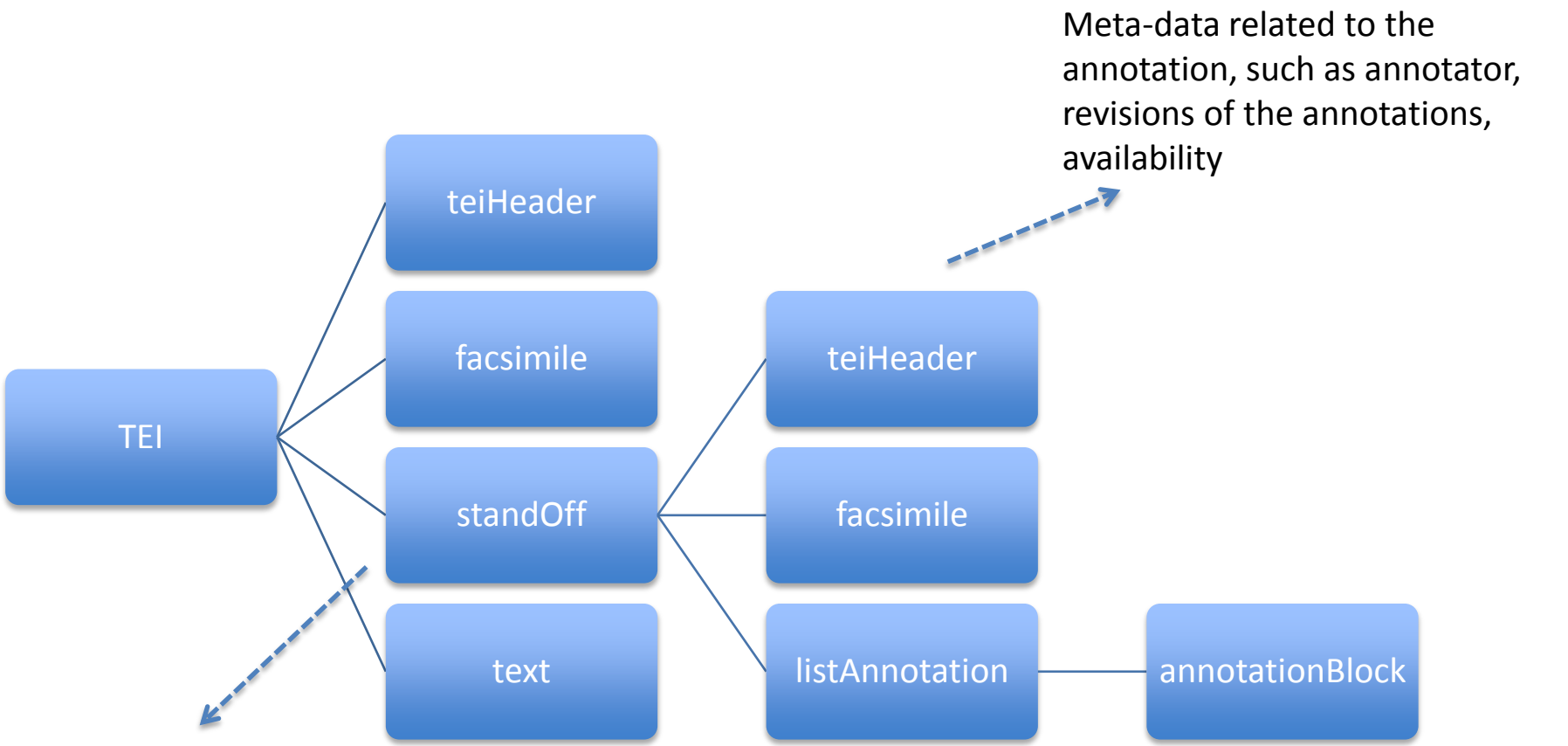
Embedded standoff: Basic concept

- 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

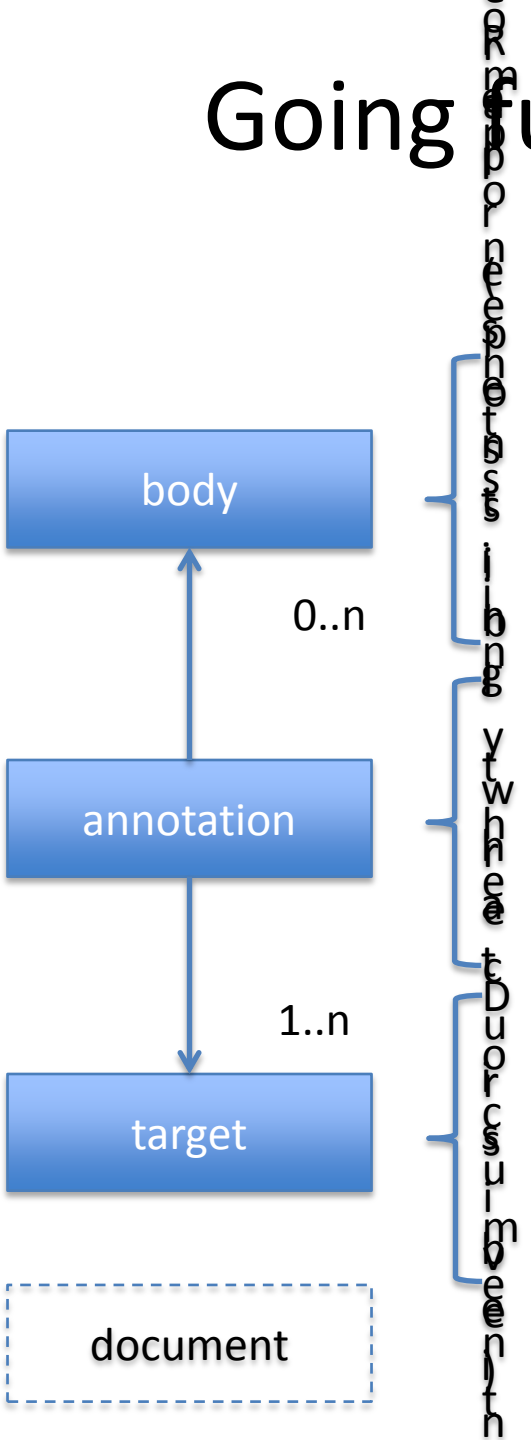
- Encoding interlinear annotation as stand-off markup
  - In <standOff>

```
<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=“”>

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

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

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

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

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…

JOINING EFFORTS WITH OTHERS: TEI AND LMF

A divided landscape

- The TEI print dictionary chapter
  - Available since more than 20 years
    - See http://www.tei-c.org/Vault/Vault-GL.html
  - Used in a wide variety of dictionary projects
    - 6 entries just in http://www.tei-c.org/Activities/Projects/
    - Disseminated at quick pace within the COS E-NEL network (credits: Toma Tasovac)
- ISO 24613:2008 Language resource management - Lexical markup framework (LMF)
  - Shorter life span
  - Mostly implemented in NLP related activities
- Is it worth reconciling the 2?
  - Yes: for the sake of combining a well-defined model with a rich XML infrastructure
  - A need for the TEI to have a terser model
    - Curation, interchange, tools, automatic generation of TEI constructs
- Is it just possible.
  - Yes: and now!

The need for a revision

- Main assets of ISO 24613 LMF
  - Comprehensive core model + series of annexes for additional modules
  - Perfectible XML serialisation…
- Going towards a multi-part standard
  - Simplifying the editorial process (drafting, decision making, revising; various tempi)
  - Reflecting the needs of specific communities (modules, serialisation)

Overview of the current plans

- Resolution 2016-04.2 (WG 4) Multi-part development of LMF
  - Part 1: Core model
  - Part 2: MRDs
  - Part 3: Diachrony-Etymology
  - Part 4: TEI serialisation
  - Part 5: LBX serialisation

(Part 4) A TEI serialisation for LMF

- Objective
  - Preventing re-inventing element that already exist
  - Eliciting constraints on the TEI model
- Method
  - Covering core model and a selected number of extensions
    - Remaining in the scope of the Print dictionary chapter
    - Extending scope if we feel there is a need from the potential TEI applications (e.g. syntax)
  - Sub-setting the TEI guidelines
    - Associating a definite TEI construct for each component of the LMF Meta-model
    - Adding constraints when necessary
      - (e.g. @xml:lang mandatory on <entry>?)
  - Complementing the TEI
    - Defining new constructs (or elements?) if necessary
      - We are not bound to the existing chapter, even if we have to abide to the Birnbaum principle

Gathering mapping proposals

Component	TEI construct
Lexical Entry	<entry>…</entry>
Form	<form>…</form>
Lemma	<form type="lemma">…</form>
Word Form	<form type="inflected">…</form>
Syntactic Behaviour	??
??	<etym>

Data category	TEI construct
/PartOfSpeech/	<pos>
/Gender/	<gen>
…	

How far should we go here?

Once upon a time, the clergyman…

```
<entry xml:lang="en">
  <form type="lemma">
    <orth>clergyman</orth>
    <gramGrp>
      <pos>commonNoun</pos>
    </gramGrp>
  </form>
  <form type="inflected">
    <orth>clergyman</orth>
    <gramGrp>
      <number>singular</number>
    </gramGrp>
   </form>
   <form type="inflected">
     <orth>clergymen</orth>
     <gramGrp>
       <number>plural</number>
     </gramGrp>
   </form>
</entry>
```

(Part 3) The case of etymology

- A flat model in the current TEI chapter
  - No sense of etymon: <mentioned>
  - No sense of etymological process
    - Typed and recursive <etym>
  - No grouping of etymon related information
    - Usage, grammatical constraints, source, date, language, etc.
- A need for revision
- Pushing a fully fledged model

Before-after example

Old school

```
<etym>
  <lang>Ahd.</lang>
<mentioned>âband</mentioned>,
<lang>mhd.</lang>
<mentioned>âbent</mentioned>;
<bibl>zur Etym. s. Kluge Mitzka 18. Aufl.
unter ,,Abend'', ferner Schwäb. Wb. 1,
11ff. Schweizdt. Wb. 1,34ff.</bibl>
</etym>
```

Structured

```
<etym type="inheritance">
  <cit type="etymon" xml:lang="goh">
    <oRef>âband</oRef>
    <lang>Ahd.</lang>
  </cit>
  <etym type="inheritance">
    <cit type="etymon" xml:lang="gmh">
      <oRef>âbent</oRef>
      <lang>mhd.</lang>
    </cit>
  </etym>
  <bibl>zur Etym. s. Kluge Mitzka 18.
Aufl. unter ,,Abend'', ferner Schwäb. Wb.
1, 11ff. Schweizdt. Wb. 1,34ff.</bibl>
</etym>
```

An interesting moment

- Time to complement and consolidate the existing practices
  - TEI as reference framework
  - ISO as a precise standardisation background
- Various ongoing projets and groups
  - Clarin Standards committee, DARIAH WG Lexical Resources, TEI LingSIG
  - COST E-NEL WG2, EU Parthenos
- Joining efforts
  - Towards a single information space
    - Basecamp, GitHub, Blog
    - Exchanging information
  - Increasing participation as experts
    - ISO-TEI in particular

WHITHER TEI?

The TEI is doing well – the hidden TEI

- Antonio Zampolli price by ADHO
  - Reflects that the TEI is pervading all fields in the (digital?) humanities
- TEI has become a natural component of a humanities project based on textual sources
  - Many small editions are flourishing everywhere
  - Now recommended or requested by funding organisations
  - Numerous training events (cf. DiXiT)
- Taken up by larger organisations
  - Academies, Dictionary projects, EPO… especially in Europe

Consolidating our conceptual model

- TEI as a rich space of elementary constructs
  - Attributes (classes), “entities”, bibliographical and dictionary entries, etc.
- Multifarious document types for various communities
  - From scholarly editions to dictionaries, including computer mediated communication, scientific information, etc.
  - More precise guidelines for specific applications
    - Collaboration with ISO (standards), DARIAH (recommendations)
    - Reducing syntactic freedom in specific application domains, not in TEI as a whole
  - Complementing our stock: onomasiological constructs, standOff
- Strong conceptual basis with pure ODD
  - For TEI and non TEI based application
  - Starting point for offering support to other dissemination formats (JSON, LOD) – Interfacing the trends
  - XML is likely to remain central for a long time for sustainable back-office content

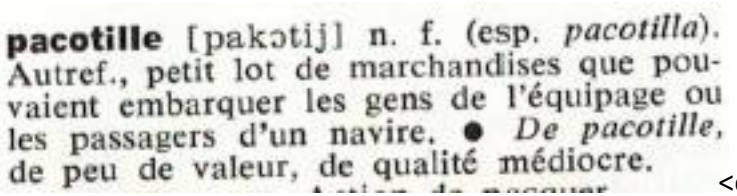
Focusing, enlarging?

- Enlarging our expert basis
  - Stronger role for SIGs
  - Close coordination with council
  - Bringing in more technical experts from outside
- Institutional partnership
  - Archives, Clarin, DARIAH, MEI, Europeana
    - Further enforcement of the TEI guidelines
    - Sharing our technical platform
      - E.g. EAD maintenance
  - Thinking together the sustainability of TEI material
    - Repositories (Tapas)
    - The TEI already offers a strong basis for sustainability
- Need for a stable communication framework
  - Lively conference and journal (jTEI)
  - Investing in the web site and the wiki

MERCI !

Automatic dictionary structure recognition

PhD theses by Mohamed Khemakhem (Inria, projet H2020 Parthenos)



pacotille [pakɔtij] n. f. (esp. *pacotilla*). Autref., petit lot de marchandises que pouvaient embarquer les gens de l'équipage ou les passagers d'un navire. ● *De pacotille*, de peu de valeur, de qualité médiocre.

CRF (Conditional Random Fields) based data mining

```
<entry>
        <form type="lemma">
          <orth>pacotille</orth>
          <pron><pc>[</pc>pakɔtij<pc>]</pc></pron>
          <gramGrp><pos>n.</pos><gen>f.</gen></gramGrp>
        </form>
        <etym><pc>(</pc><lang norm="es">esp.</lang>
          <oRef>pacotilla</oRef><pc>)</pc></etym>
        <sense>
          <usg type="time">Autref.</usg> <pc>,</pc>
          <def>petit lot de marchandises que pouvaient embarquer les gens de
l'équipage ou les passagers d'un navire</def>
          <pc>.</pc>
        </sense>
        <re>
          <form type="compound">
            <orth>De pacotille</orth>
          </form><pc>,</pc>
          <sense><def>de peu de valeur, de qualité médiocre.</def></sense>
        </re>
      </entry>
```

Fine grained recognition of the various component of an entry in a legacy dictionary

Using the TEI as reference output format (coordination with ENEL recommendations)

Perspectives: Creating step by step a large-scale network (diachronic and synchronic) of our lexical patrimony