

# Modeling dictionary structures

## TEI and TEI Lex-0 for interoperability across dictionaries

Axel Herold<sup>1,2</sup>    The TEI Lex-0 Group

<sup>1</sup>Berlin-Brandenburgische Akademie der Wissenschaften

<sup>2</sup>École Pratique des Hautes Études

September 28<sup>th</sup>, 2018



## The Text Encoding Initiative

### Traditional entry model(s) in TEI

- Lexical modeling

- Typographical view

- Editorial view

- Lexical view

- Problems with vanilla TEI

### TEI Lex-0

- Entries

- Forms and grammatical information

- Usage labels

- Etymology

### Outlook

## Brief overview

- ▶ operating since the (late) 1980s
- ▶ common standard for text representation in the DH for all sorts of printed or written documents based on open X\* standards
  - ▶ prose, verse, drama
  - ▶ dictionaries
  - ▶ list, accounting data
  - ▶ scholarly editions
  - ▶ born-digital communication
  - ▶ ...
- ▶ *lots* of active projects rely on the TEI
- ▶ open collaboration and development through special interest groups
- ▶ <http://www.tei-c.org>, <https://github.com/TEIC>

## Lexical modeling

- ▶ modeling  $\approx$  mapping of objects and their properties (and relations) onto symbolic representations (generally also abstraction)
- ▶ modeling lexical data (esp. in digitization of printed resources) is multi-layered modeling:
  - ▶ printed characters  $\longrightarrow$  codepoints (e.g. Unicode)
  - ▶ spacial relation of characters  $\longrightarrow$  words (tokens)
  - ▶ typographical properties  $\longrightarrow$  (hints as to) functions of words (tokens)
  - ▶ ...
- ▶ every level relies on interpretation and may introduce uncertainty
- ▶ alternative and even incompatible interpretations (and therefore models) are possible

## Lexical modeling in TEI

different “views” on lexical data:

**typographical** “the two-dimensional printed page, including information about line and page breaks and other features of layout”

**editorial** “the one-dimensional sequence of tokens which can be seen as the input to the typesetting process . . . ”

**lexicographic** “... the underlying information represented in a dictionary, without concern for its exact textual form”

(TEI guidelines, chapter 9)

## Lexical modeling in TEI

different “views” on lexical data:

- ▶ in print production: lexical → typographical
- ▶ in (retro-)digitization: typographical → lexical
- ▶ often desirable to retain several views
- ▶ good practice:
  - ▶ keep literal values as character data in elements
  - ▶ add annotations, normalizations as attribute values

## Lexical modeling, example entry

**Flusspat**, von nhd. *Flußspat*, so genannt, weil das mineral als zusatz beim schmelzen verwandt wurde, um die masse *in Fluß* zu bringen. Hierfür holl. *vloeispaath*, engl. *fluor* und *fluor-spar* (vgl. *feltspat*). Zugrunde liegt mlat. *fluor*, eigentlich „das fließen“. Siehe *spat* I.

Falk/Torp (1910)



## Typographical view

`<lb/><p><hi rendition="#b">Flusspat,</hi> von  
nhd. <hi rendition="#i">Flußpat</hi>,  
so genannt, weil das mineral als  
<lb/>zusatz beim schmelzen verwandt wurde,  
um die masse <hi rendition="#i">in</hi>  
<hi rendition="#i">Fluß</hi> zu  
<lb/>bringen. Hierfür holl.  
<hi rendition="#i">vloeispaath</hi>,  
engl. <hi rendition="#i">fluor</hi> und  
<hi rendition="#i">fluor-spar</hi> (vgl.  
<lb/><hi rendition="#g #i">feltspat</hi>).  
Zugrunde liegt mlat.  
<hi rendition="#i">fluor</hi>,  
eigentlich „das fließen“.  
<lb/>Siehe <hi rendition="#g #i">spat</hi>  
I.</p>`



## Editorial view

`<p><hi rendition="#b">Flusspat,</hi> von  
nhd. <hi rendition="#i">Flußspat</hi>,  
so genannt, weil das mineral als  
zusatz beim schmelzen verwandt wurde,  
um die masse <hi rendition="#i">in</hi>  
<hi rendition="#i">Fluß</hi> zu bringen.  
Hierfür holl.  
<hi rendition="#i">vloeispaath</hi>,  
engl. <hi rendition="#i">fluor</hi> und  
<hi rendition="#i">fluor-spar</hi> (vgl.  
<hi rendition="#g #i">feltspat</hi>).  
Zugrunde liegt mlat.  
<hi rendition="#i">fluor</hi>,  
eigentlich „das fließen“. Siehe  
<hi rendition="#g #i">spat</hi> I.</p>`

## Lexical view

```
<entry>
  <form><orth>Flusspat,</orth></form>
  <etym>von <lang>nhd.</lang>
  <mentioned>Flußpat</mentioned>, so
  genannt, weil das mineral als zusatz
  beim schmelzen verwandt wurde, um die
  masse <mentioned>in Fluß</mentioned>
  zu bringen. Hierfür <lang>holl.</lang>
  <mentioned>vloeispaath</mentioned>,
  <lang>engl.</lang>
  <mentioned>fluor</mentioned> und
  <mentioned>fluor-spar</mentioned>
  (vgl. <ref>feltspat</ref>).
  Zugrunde liegt <lang>mlat.</lang>
  <!-- ... --> </etym>
</entry>
```

## Lexical view, alternative

```
<entry type="main">
  <form type="headword">
    <orth>Flusspat,</orth>
    <gramGrp><pos value="NN"/>
  </gramGrp></form>
  <etym>von <lang>nhd.</lang>
  <mentioned
    xml:lang="de">Flußspat</mentioned>,
    so genannt, weil das mineral als
    zusatz beim schmelzen verwandt wurde,
    um die masse <mentioned
      xml:lang="de">in Fluß</mentioned>
      zu bringen. Hierfür <lang>holl.</lang>
    <!-- ... --> </etym>
</entry>
```

## Problems with vanilla TEI

(specifically for dictionary modeling)

- ▶ sometimes several almost similar models
- ▶ often several ways to encode the same abstract model
- ▶ sometimes abstract models have no direct equivalent in TEI
- ▶ difficult to encode alternative models
- ▶ often open or semi-fixed vocabulary for annotations

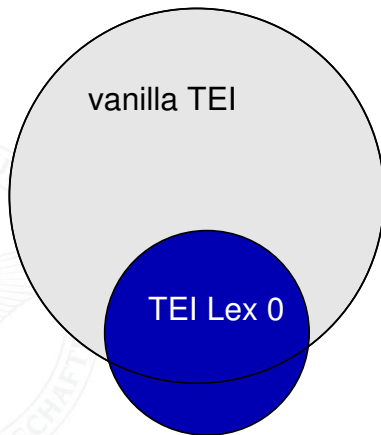
Some of these problems will be discussed in this talk.

## Background

- ▶ work started in 2016, supported by ENeL, Dariah, research institutes
- ▶ international group with strong ties to the TEI
- ▶ general use-case: mapping typographic structures onto lexical structures
- ▶ not a chapter 9 replacement, rather a proposed baseline by way of a TEI customization
- ▶ aims at interoperability by
  - ▶ restricting some alternatives
  - ▶ streamlining some content models
  - ▶ closing and fixing vocabulary
  - ▶ ...
- ▶ some changes even went upstream already

## Relation between vanilla TEI and TEI Lex-0

“not a chapter 9 replacement, rather a proposed baseline”  
... for lexicographic information



## Important side note: What's a TEI name, anyway?

```
<form>  
  <orth>tree</orth>  
</form>
```

```
<form>  
  <orth>tree</orth>  
  <form>  
    <orth>trees</orth>  
    <gramGrp><number>plur.</number></gramGrp>  
  </form>  
  <gramGrp><pos>noun</pos></gramGrp>  
</form>
```

- ▶ “form” vs. “container for form-related information”
- ▶ more pronounced typing of elements is needed

## Entries

(“sometimes several almost similar models”)

- ▶ in TEI, several models for entries (and entry-like entities):
  - entry** “contains a single structured entry ...”
  - entryFree** “contains a single unstructured entry ...”
  - superEntry** “groups a sequence of entries ...”
  - hom** “groups information relating to one homograph within an entry.”
  - re** “contains a dictionary entry for a lexical item related to the headword ...”
- ▶ all models with slightly different content models
- ▶ in TEI Lex-0: only **entry**, but recursively nestable



## Entries

---

Leder	nn	leather
leder n		of leather; leathern, leathery, tough
ab leder n		wipe with chamois skin
Ober leder	nn	upper leather of shoe
Unter leder	nn	sole leather

---

Keller (1978)



## Entries

---

Leder	nn	leather
leder n		of leather; leathern, leathery, tough
ab leder n		wipe with chamois skin
Ober leder	nn	upper leather of shoe
Unter leder	nn	sole leather

---

Keller (1978)

```

<entry type="word-family">
  <entry type="word">
    <form type="lemma" xml:lang="de">
      <orth>Leder</orth></form>
      <gramGrp<pos>nn.</pos></gramGrp>
      <sense<def xml:lang="en">leather</def></sense>
    </entry>
    <entry type="word"> <!-- ledern [vb.] --> </entry>
    <entry type="word"> <!-- abledern [vb.] --> </entry>
    <!-- ... -->
  </entry>

```

## Forms and grammatical information

(“often several ways to encode the same abstract model”)

- ▶ `gramGrp` may appear on `entry`, `form`, `sense`, ... in vanilla TEI
- ▶ TEI Lex-0 restricts this and relies on enheritance as expressed in the XML structure:
  - ▶ entry-level grammatical information on `entry`
  - ▶ sense-level grammatical information on `sense`
  - ▶ form specific grammatical information on `form`
  - ▶ (with narrow exceptions)
- ▶ mandatory `form/@type`, e. g. `lemma`, `inflected`, `paradigm`, `variant` in Lex-0

## Forms and grammatical information

**grunt** vb. ME. *grunte gronte*  
 OE. *grunnettan*; ident. w. G.  
*grunzen*, DAN. *grynte*, SW. *grynta*  
 A more primit. stem appears in  
 OE. *grunian* 'grunt'. The  $\sqrt{\text{grun}}$   
 is imitation of sound; cp. LAT  
*grunnire*.

Kluge/Lutz (1898)

```

<entry>
  <form type="lemma"><orth>grunt</orth></form>
  <gramGrp><pos>vb.</pos></gramGrp>
  <etym>
    <!-- ... -->
  </etym>
</entry>

```

## Forms and grammatical information

```

<entry>
  <form type="lemma">
    <orth>aid</orth>
    <pron>e&#305;d</pron>
  </form>
  <entry>
    <gramGrp<del>pos</del>noun</pos<del>/gramGrp>
    <!-- ... -->
  </entry>
  <entry>
    <gramGrp<del>pos</del>verb</pos<del>/gramGrp>
  </entry>
</entry>

```

**aid** /eɪd/ *noun* **1.** help, especially money, food or other gifts given to people living in difficult conditions ○ *aid to the earthquake zone* ○ *an aid worker* (NOTE: This meaning of **aid** has no plural.) □ **in aid of** in order to help ○ *We give money in aid of the Red Cross.* ○ *They are collecting money in aid of refugees.* **2.** something which helps you to do something ○ *kitchen aids* ■ **verb** **1.** to help something to happen **2.** to help someone

## Forms and grammatical information

**ACHTER**

Woordsoort: vz., bw.

Modern lemma: achter

voorz. en bijw. Dnl. *aftr, after* (Ps. 57, 5; 62, 9; Gl. Lips. in Ps. 81, 13; 118, 8), in samens. 1, 11); *nhd. after*; *ags. æfter* (ETTM. 39); *eng. after*; *osaks. aftr, after* (SCHMELLER 4); *ofri. mnl. ave, af, goth. af, hd. ab*, hetwelk verwijdering en scheiding uitdrukt. De lettergreep *-ter*, wordt (BOPP, *Vergl. Gramm.* § 291). De afleiding verklaart de beteekenis: wat *achter* is, is v

- ☐ I. Als voorzetsel.
- ☐ II. Als bijwoord.
- ☐ III. In samenstellingen.

&lt;entry&gt;

&lt;form type="lemma"&gt;&lt;orth&gt;ACHTER&lt;/orth&gt;&lt;/form&gt;

&lt;gramGrp&gt;&lt;pos&gt;voorz. en bijw&lt;/pos&gt;&lt;/gramGrp&gt;

&lt;etym&gt; &lt;!-- --&gt; &lt;/etym&gt;

&lt;sense n="I"&gt;

&lt;gramGrp&gt;&lt;pos&gt;Als voorzetsel.&lt;/pos&gt;&lt;/gramGrp&gt;

&lt;!-- ... --&gt;&lt;/sense&gt;

&lt;sense n="II"&gt;

&lt;gramGrp&gt;&lt;pos&gt;Als bijwoord.&lt;/pos&gt;&lt;/gramGrp&gt;

&lt;!-- ... --&gt;&lt;/sense&gt; &lt;!-- ... --&gt;&lt;/entry&gt;

## Forms and grammatical information, exception

```
<entry>
  <form type="lemma">
    <orth>go</orth>
    <gramGrp>
      <gram type="tense">present</gram>
    </gramGrp>
  </form>
  <!-- ... -->
</entry>
```

We really want inheritance to work!

## Forms and grammatical information, inflected forms

```
<entry>
  <form type="lemma">
    <orth>go</orth>
  </form>
  <form type="inflected">
    <orth>went</orth>
    <gramGrp>
      <gram type="tense">past</gram>
    </gramGrp>
  </form>
  <!-- ... -->
</entry>
```

We really want inheritance to work!



## Usage labels

(“often open or semi-fixed vocabulary for annotations”)

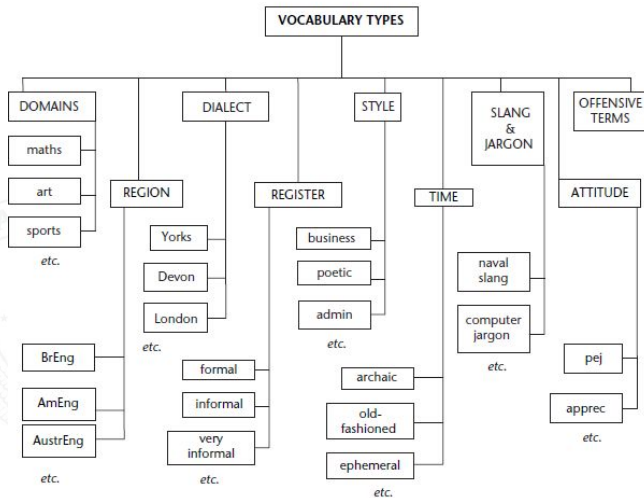
- ▶ usg covers multiple dimensions:
  - ▶ geo(graphic)
  - ▶ time
  - ▶ dom(ain)
  - ▶ register, style
  - ▶ plev (preference level)
  - ▶ lang(uage)
  - ▶ gram(matical)
  - ▶ syn(onym), hyp(ernym)
  - ▶ colloc(ation), comp(lement), obj(ect), subj(ect), verb
  - ▶ hint
- ▶ many dimensions have their own (better) model in TEI
- ▶ TEI Lex-0 tries to streamline this vocabulary (ongoing discussion)

## Usage labels

Criterion	Type of marking	Unmarked centre	Marked periphery	Examples of labels
Time	diachronic	contemporary language	archaism – neologism	<i>arch., dated, old use</i>
Place	diatopic	standard language	regionalism, dialect word	<i>AmE, Scot., dial.</i>
Nationality	diaintegrative	native word	foreign word	<i>Lat., Fr.</i>
Medium	diamedial	neutral	spoken – written	<i>colloq., spoken</i>
Socio-cultural	diastratic	neutral	sociolects	<i>pop., slang, vulgar</i>
Formality	diaphasic	neutral	formal – informal	<i>fml, infml</i>
Text type	diatextual	neutral	poetic, literary, journalese	<i>poet., lit.</i>
Technicality	diatechnical	general language	technical language	<i>Geogr., Mil., Biol., Mus.</i>
Frequency	diafrequential	common	rare	<i>rare, occas.</i>
Attitude	diaevaluative	neutral	connoted	<i>derog., iron., euphem.</i>
Normativity	dianormative	correct	incorrect	<i>non-standard</i>

Svensén (2009) after Hausmann (1989)

## Usage labels



Atkins/Rundell (2008)

## Etymology

(“sometimes abstract models have no direct equivalent in TEI”)

etymological prose ...

- ▶ often is exactly this: prose  
(i. e. not necessarily rigidly structured)
- ▶ outlines complex linguistic entities
- ▶ outlines complex (historical) relations among them

→ Can we formalize this more deeply than vanilla TEI?

## Etymology

**grunt** vb. ME. *grunte gronte*  
OE. *grunnetan*; ident. W. G.  
*grunzen*, DAN. *grynte*, SW. *grynta*  
A more primit. stem appears in  
OE. *grunian* 'grunt'. The  $\sqrt{\text{grun}}$   
is imitation of sound; cp. LAT  
*grunnire*.

- ▶ non-etymological information is covered in TEI
- ▶ etymological information not so much ...
- ▶ essentially, we need a device to model complex *mentioned* forms (such as etymons) and their relations
- ▶ need a model of the temporality of etymological processes

## Etymology

**grunt** vb. ME. *grunte gronte*  
 OE. *grunnetan*; ident. w. G.  
*grunzen*, DAN. *grynte*, SW. *grynta*  
 A more primit. stem appears in  
 OE. *grunian* 'grunt'. The  $\sqrt{\text{grun}}$   
 is imitation of sound; cp. LAT  
*grunnire*.

```
<entry>
  <form type="lemma"><orth>grunt</orth></form>
  <gramGrp><pos>vb.</pos></gramGrp>
  <etym>
    <!-- ... -->
  </etym>
</entry>
```

## Etymology

**grunt** vb. ME. *grunte gronte*  
 OE. *grunnetan*; ident. w. G.  
*grunzen*, DAN. *grynte*, SW. *grynta*  
 A more primit. stem appears in  
 OE. *grunian* 'grunt'. The  $\sqrt{\text{grun}}$   
 is imitation of sound; cp. LAT  
*grunmire*.

&lt;etym&gt;

&lt;lang rendition="sc"&gt;me.&lt;/lang&gt;

&lt;mentioned&gt;grunte&lt;/mentioned&gt;

&lt;mentioned&gt;gronte&lt;/mentioned&gt;

&lt;lang rendition="sc"&gt;oe.&lt;/lang&gt;

&lt;mentioned&gt;grunnetan&lt;/mentioned&gt;;

&lt;!-- ... --&gt;

&lt;/etym&gt;

Problem in vanilla TEI: associate lang and mentioned

## Etymology

**grunt** vb. ME. *grunte gronte*  
OE. *grunnetan*; ident. w. G.  
*grunzen*, DAN. *grynte*, SW. *grynta*  
A more primit. stem appears in  
OE. *grunian* 'grunt'. The  $\sqrt{\text{grun}}$   
is imitation of sound; cp. LAT  
*grunnire*.

```
<cit type="etymon">  
  <lang rendition="sc">me.</lang>  
  <form type="lemma" xml:lang="enm">  
    <orth>grunte</orth>  
    <orth>gronte</orth>  
  </form>  
</cit>
```

Remeber: What's a TEI name, anyway ...



## Etymology

**grunt** vb. ME. *grunte gronte*  
OE. *grunnetan*; ident. w. G.  
*grunzen*, DAN. *grynte*, SW. *grynta*  
A more primit. stem appears in  
OE. *grunian* 'grunt'. The  $\sqrt{\text{grun}}$   
is imitation of sound; cp. LAT  
*grunnire*.

```
<cit type="etymon">  
  <lang rendition="sc">oe.</lang>  
  <form type="lemma" xml:lang="ang">  
    <orth>grunian</orth>  
  </form>  
  <def>'grunt'</def>  
</cit>
```

## Etymology

**grunt** vb. ME. *grunte gronte*  
OE. *grunnetan*; ident. w. G.  
*grunzen*, DAN. *grynte*, SW. *grynta*  
A more primit. stem appears in  
OE. *grunian* 'grunt'. The  $\sqrt{\text{grun}}$   
is imitation of sound; cp. LAT  
*grunnire*.

```
<cit type="cognate">  
  <lang rendition="sc">dan.</lang>  
  <form type="lemma" xml:lang="da">  
    <orth>grynte</orth>  
  </form>  
</cit>
```

## Etymology

**grunt** vb. ME. *grunte gronte*  
OE. *grunnetan*; ident. w. G.  
*grunzen*, DAN. *grynte*, SW. *grynta*  
A more primit. stem appears in  
OE. *grunian* 'grunt'. The  $\sqrt{\text{grun}}$   
is imitation of sound; cp. LAT  
*grunnire*.

grunt



grunte, gronte



grunnetan

Problem in vanilla TEI: make (chained) relations explicit

## Etymology

**grunt** vb. ME. *grunte gronte*  
 OE. *grunnetan*; ident. w. G.  
*grunzen*, DAN. *grynte*, SW. *grynta*  
 A more primit. stem appears in  
 OE. *grunian* 'grunt'. The  $\sqrt{\text{grun}}$   
 is imitation of sound; cp. LAT  
*grunnire*.

```
<etym type="inheritance">
  <cit type="etymon">
    <lang rendition="sc">me.</lang>
    <form type="lemma" xml:lang="enm">...</form>
  </cit>
  <cit type="etymon">
    <lang rendition="sc">oe.</lang>
    <form type="lemma" xml:lang="ang">...</form>
  </cit>
</etym>
```

## Etymology

complex descriptions of linguistic signs via

`cit[@type="etymon"]`, `cit[@type="cognate"]`:

- ▶ in a way, `cit[@type="..."]` is very close to entry
- ▶ may contain
  - ▶ `lang` (not in vanilla TEI)
  - ▶ `date`
  - ▶ `form`
  - ▶ `def/gloss` (even sense?)
  - ▶ `usg`
  - ▶ `xr`
  - ▶ `gramGrp`
  - ▶ `ref`
  - ▶ `bibl`

## Etymology

complex relations among cits via  
`etym[@type="..."]`:

- ▶ types may include borrowing, inheritance, compounding, derivation, metaphor, ...
- ▶ typing may be expensive, therefore optional
- ▶ etyms contain mostly cits, maybe segs (for unmarked stretches of prose)
- ▶ conflicting etymologies can be siblings (and may get indications of responsibility)
- ▶ upcoming paper by Bowers/Herold/Romary

## The Text Encoding Initiative

### Traditional entry model(s) in TEI

- Lexical modeling

- Typographical view

- Editorial view

- Lexical view

- Problems with vanilla TEI

### TEI Lex-0

- Entries

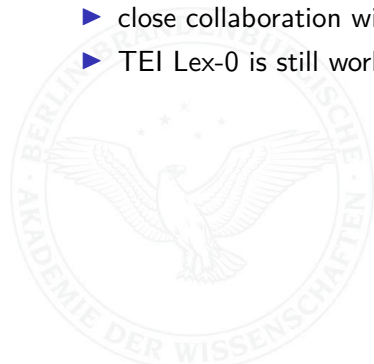
- Forms and grammatical information

- Usage labels

- Etymology

### Outlook

- ▶ more areas of work not covered here, e. g. for cross-references, bilingual dictionaries, ...
- ▶ frequent group meetings
- ▶ open collaboration on GitHub (take a look!)
- ▶ close collaboration with the TEI consortium
- ▶ TEI Lex-0 is still work in progress





# Thank you for listening!

