Current state and future plans for the TLS research platform

Christian Wittern

Kyoto University Institute for Research in Humanities Center for Informatics in East-Asian Studies

2022-05-13

Next steps for the TLS Website

- Next version will be based on TEI Publisher
- Some examples of applictions developed with TEI Publisher
- Better fulltext search (requires big update to text format)
  - Needs some careful considerations
- First experiments with TLS on TEI Publisher

Next steps for the TLS Website

About TEI Publisher

- Developed since 2015, current version 7.1.0
- Abstracted layer of implementation for processing TEI texts
- Latest developments:
  - Backend accessed through standardized API
  - Frontend based on custom Webcomponents
  - Interface for adding annotations

Van Gogh Letter

- Based on TEI Publisher 5.0
- Flexible layout with text, translation, facsimile and notes
- The layout can be configured freely by the user
- Search results include facets
- Original and translation on equal footing

Alfred Escher Briefedition

- Based on TEI Publisher 8.0 (not yet released)
- Additional information about related persons and events
- Context of letter exchanges
- Edited text and diplomatic transcription
  - Text of the letter + registry of persons/places/contexts
  - Line-wise link between text and facsimile
- Timeline

Next steps for the TLS Website

Reminder of the discussion

- Currently, the search function can not search across punctuation
- Every text line (phrase) is treated as an atomic unit
  - this is part of the heritage of the Filemaker database
- To change this, fundamental changes to the text format are required
- A first version of this was proposed in January

Format as discussed in January

```
<div>
<head xml:id="KR5c0057_tls_001-1a.3-h" n="1">第一章</head>
<p xml:id="KR5c0057_tls_p2">
<c/>道可道<c xml:id="KR5c0057_tls_001-1a.3" n=", "/>
非恒道<c xml:id="KR5c0057_tls_001-1a.4" n="。"/>
名可名<c xml:id="KR5c0057_tls_001-1a.5" n=", "/>
非恒名<c xml:id="KR5c0057_tls_001-1a.6" n="。"/>
無名<c n="、"/>天地之始<c xml:id="KR5c0057_tls_001-1a.7" n=", "/>
有名<c n="、"/>萬物之母<c xml:id="KR5c0057_tls_001-1a.8" n=", "/>
故恒無欲<c n=", "/>以觀其妙<c xml:id="KR5c0057_tls_001-1a.9" n="。"/>
恒有欲<c n=", "/>以觀其徼<c xml:id="KR5c0057_tls_001-1a.10" n="。"/>
此兩者同出而異名<c xml:id="KR5c0057_tls_001-1a.11" n="。"/>
同謂之玄<c xml:id="KR5c0057_tls_001-1a.12" n="。"/>
玄之又玄<c xml:id="KR5c0057_tls_001-1a.13" n=", "/>
眾妙之門<c xml:id="KR5c0057_tls_001-1a.14" n="。"/>
</p></div>
```

New format #1

```
<div>
<head xml:id="KR5c0057_tls_001-1a.3-h" n="1">第一章</head>
<p>
<lb xml:id="KR5c0057_tls_001-1a.3" ed="tls"/>道可道<c n=", "/>
<lb xml:id="KR5c0057_tls_001-1a.4" ed="tls"/>非恒道<c n="。"/>
<lb xml:id="KR5c0057_tls_001-1a.5" ed="tls"/>名可名<c n=", "/>
<lb xml:id="KR5c0057_tls_001-1a.6" ed="tls"/>非恒名<c n="。"/>
<lb xml:id="KR5c0057_tls_001-1a.7" ed="tls"/>無名<c n="、"/>天地之始<c n=", "/>
<lb xml:id="KR5c0057_tls_001-1a.8" ed="tls"/>有名<c n="、"/>萬物之母<c n=", "/>
<lb xml:id="KR5c0057_tls_001-1a.9" ed="tls"/>故恒無欲<c n=", "/>以觀其妙<c n="。"/>
<lb xml:id="KR5c0057_tls_001-1a.10" ed="tls"/>恒有欲<c n=", "/>以觀其徼<c n="。"/>
<lb xml:id="KR5c0057_tls_001-1a.11" ed="tls"/>此兩者同出而異名<c n="。"/>
<lb xml:id="KR5c0057_tls_001-1a.12" ed="tls"/>同謂之玄<c n="。"/>
<lb xml:id="KR5c0057_tls_001-1a.13" ed="tls"/>玄之又玄<c n=", "/>
<lb xml:id="KR5c0057_tls_001-1a.14" ed="tls"/>眾妙之門<c n="。"/><lb ed="tls"/></p>
<pb xml:id="KR5c0057_tls_002-1a" ed="tls" n="002-1a"/>
</div>
```

New format #2

Figure: Change <seg> to <lb>

Evaluation

- The format deviates radically from the current one
  - it completely eliminates <seg>
  - but <seg> is the center node of links to translation and lexicon
- This requires extensive changes to many parts of the database
- Further investigations of the working of TEI Publisher revealed
  - eliminating <seg> is not necessary and not desirable
  - a new format that does continue to use <seg> is a better solution

Latest proposal for new text format

- Minimally necessary changes to current format
- keep <seg> for linking to other parts
- Problem: texts without punctuation do not have <seg>
  - Other ways to link?
- Consider embedding link anchors into the text
  - How is linking done actually?

New text format

```
23        <div>
24 <head xml:id="KR5c0057_tls_001-1a.3-h-h" n="1"><seg
24 xml:id="KR5c0057_tls_001-1a.3-h">第一章</seg></head>
25
26           <p><seg xml:id="KR5c0057_tls_001-1a.3">道可道<c n=", "/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.4">非恒道<c n="。"/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.5">名可名<c n=", "/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.6">非恒名<c n="。"/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.7">無名<c n="、"/>天地之始<c n=", "/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.8">有名<c n="、"/>萬物之母<c n=", "/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.9">故恒無欲<c n=", "/>以觀其妙<c n="。"/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.10">恒有欲<c n=", "/>以觀其徼<c n="。"/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.11">此兩者同出而異名<c n="。"/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.12">同謂之玄<c n="。"/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.13">玄之又玄<c n=", "/></seg><seg
26 xml:id="KR5c0057_tls_001-1a.14">眾妙之門<c n="。"/></seg></p>
27           <pb xml:id="KR5c0057_tls_002-1a" ed="tls" n="002-1a"/><lb/>
28        </div>
```

Figure: Keep <seg>, only add <c>

Next steps for the TLS Website

Linking through annotation file

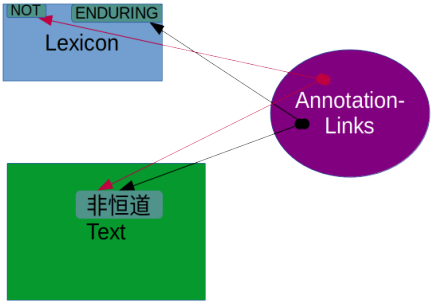

Figure: Text and Lexicon not aware of links

Mechanics of linking (1)

```
<tls:ann xmlns:tls="http://hxwd.org/ns/1.0" concept="NOT"
    concept-id="uuid-2785bda2-fc81-4400-bfbe-53a6bee057b7"
    xml:id="uuid-eb80b06c-f691-4d2d-a753-cd180a59e2c2">
    <link
        target="#KR5c0057_tls_001-1a.4 #uuid-32a30364-5261-438d-8ff8-f493af6e0d17"/>
    <tls:text>
        <tls:srcline title="老子" target="#KR5c0057_tls_001-1a.4" pos="1"
            >非恒道。</tls:srcline>
        <tls:line title="老子 (en)" src="Karlgren 1975">is not the constant
            Way,</tls:line>
    </tls:text>
    <form corresp="#uuid-00e22256-d177-459e-bd67-efa461a8d045" orig="">
        <orth>非</orth>
        <pron xml:lang="zh-Latn-x-pinyin">fēi</pron>
    </form>
    <sense corresp="#uuid-32a30364-5261-438d-8ff8-f493af6e0d17">
        <gramGrp>
            <pos>V</pos>
            <tls:syn-func
                corresp="#uuid-c87f5e8b-6512-404d-84b2-9e99a85aa28e">vt
                N</tls:syn-func>
            <tls:sem-feat
                corresp="#uuid-52f9b87c-5688-4b46-b992-a5fb0bf27fb9"
                >copula</tls:sem-feat>
            <usg type="warring-states-currency">5</usg>
        </gramGrp>
        <def>is not N</def>
    </sense>
```

Mechanics of linking (2)

```
</sense>

<sense xml:id="uuid-32a30364-5261-438d-8ff8-f493af6e0d17">
    <gramGrp>
        <pos>V</pos>
        <tls:syn-func corresp="#uuid-edfd3bad-7083-48fd-bd3b-ad708388fd68"
            >vt+N{PRED}</tls:syn-func>
        <tls:sem-feat corresp="#uuid-52f9b87c-5688-4b46-b992-a5fb0bf27fb9"
            >copula</tls:sem-feat>
        <usg type="warring-states-currency">5</usg>
    </gramGrp>
    <def>is not N</def>
</sense>

<sense xml:id="uuid-0f7d2e76-ffb9-4604-8776-c0a35cd5c953">
```

Figure: 非 in concept NOT

Linking without annotation file

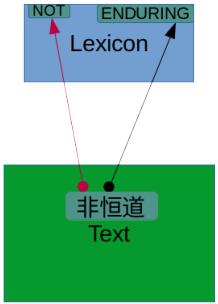

Figure: Link anchors embedded in text

New text format with embedded anchors

```
        <div>
<head xml:id="KR5c0057_tls_001-1a.3-h-h" n="1"><seg
xml:id="KR5c0057_tls_001-1a.3-h">第一章</seg></head>

            <p><seg xml:id="KR5c0057_tls_001-1a.3">道可道<c n=", "/></seg><seg
xml:id="KR5c0057_tls_001-1a.4"><anchor
xml:id="uuid-eb80b06c-f691-4d2d-a753-cd180a59e2c2"
ref="uuid-32a30364-5261-438d-8ff8-f493af6e0d17" resp="#CH"
modified="2019-11-05T23:05:39.385+09:00" len="1" type="SWL"/>非<anchor
xml:id="uuid-3a45c08d-5c2f-423b-82eb-b733838ccacb"
ref="uuid-4e5c1744-84c8-4b48-bb71-674294718581" resp="#CH"
modified="2019-11-05T23:04:00.227+09:00" len="1" type="SWL"/>恒道<c n="。"/></seg><seg
xml:id="KR5c0057_tls_001-1a.5">名可名<c n=", "/></seg><seg
xml:id="KR5c0057_tls_001-1a.6">非恒名<c n="。"/></seg><seg
xml:id="KR5c0057_tls_001-1a.7">無名<c n="、"/>天地之<anchor
xml:id="uuid-b56779f1-db75-4ce9-931e-43a50aa2b2db"
ref="uuid-0bbd0bfa-0082-4c14-96e2-86c6437036f5" resp="#CH"
modified="2019-03-26T13:08:54.554+09:00" len="1" type="SWL"/>始<c n=", "/></seg><seg
xml:id="KR5c0057_tls_001-1a.8">有名<c n="、"/>萬物之<anchor
xml:id="uuid-61a5c49f-f63a-471b-a1f7-29edbc2aa633"
ref="uuid-e23a5301-ee04-40f6-b693-8d6ebbc34a7e" resp="#CH"
modified="2019-03-26T13:08:54.554+09:00" len="1" type="SWL"/>母<c n=", "/></seg><seg
xml:id="KR5c0057_tls_001-1a.9">故恒無欲<c n=", "/>以觀其妙<c n="。"/></seg><seg
```

Text with link anchors

```
30            <p>
31               <seg xml:id="KR5c0057_tls_001-1a.3">道可道<c n=", "/></seg>
32               <seg xml:id="KR5c0057_tls_001-1a.4"><anchor
33                  xml:id="uuid-eb80b06c-f691-4d2d-a753-cd180a59e2c2"
34                  ref="uuid-32a30364-5261-438d-8ff8-f493af6e0d17" resp="#CH"
35                  modified="2019-11-05T23:05:39.385+09:00" len="1" type="SWL"/>
非<anchor
36                  xml:id="uuid-3a45c08d-5c2f-423b-82eb-b733838ccacb"
37                  ref="uuid-4e5c1744-84c8-4b48-bb71-674294718581" resp="#CH"
38                  modified="2019-11-05T23:04:00.227+09:00" len="1" type="SWL"/>恒道<c
n="。"/></seg>
39               <seg xml:id="KR5c0057_tls_001-1a.5">名可名<c n=", "/></seg>
40               <seg xml:id="KR5c0057_tls_001-1a.6">非恒名<c n="。"/></seg>
41               <seg xml:id="KR5c0057_tls_001-1a.7">無名<c n="、"/>天地之<anchor
42                  xml:id="uuid-b56779f1-db75-4ce9-931e-43a50aa2b2db"
43                  ref="uuid-0bbd0bfa-0082-4c14-96e2-86c6437036f5" resp="#CH"
44                  modified="2019-03-26T13:08:54.554+09:00" len="1" type="SWL"/>始<c
n=", "/></seg>
```

Figure: Anchor for 非 in Laozi 1

Next steps for the TLS Website

What is currently working

- new text format can be displayed
  - translations not yet
- anchors not yet visible
- adding of new anchors (annotation)
  - ordering of list for existing definitions?
- other annotations:
  - Persons, places, organizations etc
- text critical markup
- different display modes
- search needs more work