A closer look at HXWD.org and how the TLS is implemented there

Christian Wittern

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

2021-07-09



A closer look at HXWD.org and how the TLS is implemented there

Introduction

The HXWD interface

The road to HXWD

HXWD.org behind the scenes

More interactions with HXWD



Background

- ► Task
 - Porting the TLS from Filemaker to a new format that allows collaborative editing
 - Improve the quality of the source texts
 - Stay within the available budget and timeframe
- (...) Finally, I decided to take up the challenge ...
- My simple plan:
 - Attach TLS to the Kanseki Repository
 - After two failed attempts, I had to give up.
- ► HXWD.org is the third attempt



Bird's eye view of the TLS

- ▶ (1) Textual layer: the source texts
- ▶ (2) Interpretional layer:
 - Ontology of
 - concepts (or synonym groups)
 - syntactic functions and semantic features
 - rhetoric devices
 - ... and eventually: character taxonomies
 - Words
 - Phonological reconstructions
 - Bibliography
 - **.**..



Introduction

0000

- ▶ (since tls.org was already taken by Tokyo Little Satellite...)
- ▶ The new application for the TLS was implemented at hxwd.org (漢學文典)
- In the bird's eye view, this became:
- ► (3) The web application



A closer look at HXWD.org and how the TLS is implemented there

The HXWD interface



- don't overwhelm the user
- exploration reveals functionality

- don't overwhelm the user
- exploration reveals functionality
- consistent use of UI elements



- don't overwhelm the user
- exploration reveals functionality
- consistent use of UI elements
- simple and elegant style



- don't overwhelm the user
- exploration reveals functionality
- consistent use of UI elements
- simple and elegant style
- simple



- don't overwhelm the user
- exploration reveals functionality
- consistent use of UI elements
- simple and elegant style
- simple
- and even simpler



Technical details

- **XML** based standards for text and interpretation
 - ► TFI wherever feasable
- eXist-db as back-end engine (recently upgraded to 5.3.0)
 - app development using eXist-db's template processing
- XQuery 3.0 as server development language
- Bootstrap 4 for the front-end interface elements
- ▶ Javascript (Jquery and other libraries) for interface programming



User groups

- Different views for different user groups:
 - guest (not logged in users)
 - test (a generic account with limited view and no editing rights)
 - tls-user (basic editing: translation, attribution)
 - currently, most users are here
 - tls-editor (more elaborate/destructive edits)
 - tls-admin (additional administrative rights)
 - dba (database administrator)
- (a proper user management system still needs to be implemented)



Web interface Demo

- ► Text search
 - other searches, options
- Concept browsing
- Character taxonomies



Web interface Demo

- Text search
 - other searches, options
- Concept browsing
- Character taxonomies
- ▶ Sorry, the demo has been moved to the end of this presentation

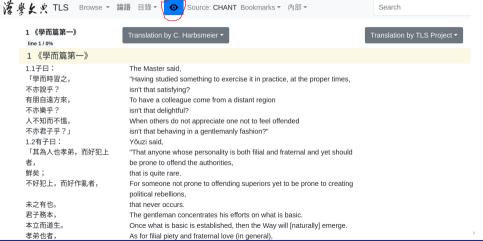


Web interface Demo

- Text search
 - other searches, options
- Concept browsing
- Character taxonomies
- ▶ Sorry, the demo has been moved to the end of this presentation
- But we will have some screen shots now



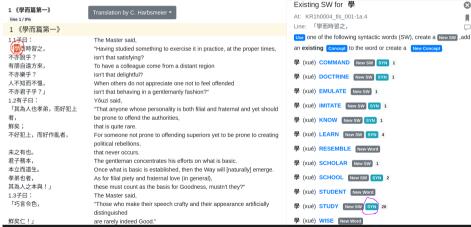
First chapter of the Analects with translation



Christian Wittern Kvoto University The HXWD interface 00000000000

1 《學而篇第一》 line 1/0%		Translation by C. Harbsmeier ▼	Translation by TLS Project ▼
1 《學而篇第一》			
1.1子曰:		The Master said,	
	子 (zǐ)	TEACHER	n[post-N] our master; person in authority [Note that the [post-N] can never in fact be made explicit without breaking the rules of the grammar! CH]
「學而時習之,		"Having studied something to exercise	e it in practice, at the proper times,
Rhet:	TRICOLON-2+1+CRESCENDO		
	學 (xué)	STUDY	vt[oN] act devote oneself to study; be devoted to study; engage in intellectual work; work to improve oneself morally/intellectually
	而 (ér)	AND	padV1.postV2 sequence e.g. 拜而受之 "bow and accept it" and then, and thereupon; (often coordinates two verbal predicates with the same subject)
	時 (shí)	TIMELY	nadV appropriate at the appropriate time; when the need arises Note 非時
	時 (shí)	OFTEN	nadV from time to time> sometimes; often; periodically; several times; at regular intervals; all the time $\bigcirc \bullet \times$
	之 (zhī)	PRONOUN	npro.postVt nonreferential an object, things; a place; something [cognate object; indefinite dummy object pronouns without reference: things, people; someone 以約失之者"Those who, through keeping to the main thing, get things wrong"]

Concepts related to 學 xué







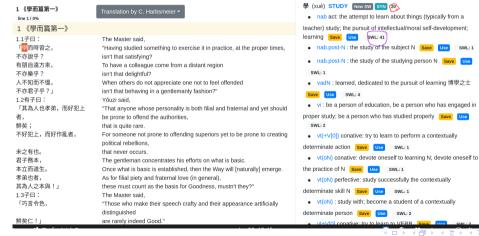
```
學 (xué) SCHOLAR New SW 1
學 (xué) SCHOOL New SW SYN 2
學 (xué) STUDENT New Word
學 (xué) STUDY New SW SYN 20
1. The dominant word is xué 學 (ant. jiào 教 "train teach")which refers
primarily to studying or training under another person, and
secondarily to the learning by heart texts. Very often, the word retains
a tinge of immitation.
2. Shī 師 and cóng 從 refer to deciding to study under someone and
treating him as one's teacher.
3. Kǎo 考 refers to an investigation of a subject in a "scientific" spirit.
4. Jiū 究 and ván 研 refers to an in-depth study of a subject, typically
involving a fair amount of reflection.
5. Zhì 治 and gōng 攻 refer to the specialised study, typically of a
certain text
6. Xí 習 refers to rehearsal of what one has learnt, through repetition
of text and/or through enactment in practice.
學 (xué) WISE New Word
```

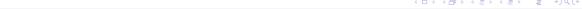


Christian Wittern Kvoto University

Christian Wittern

Syntactic words defined for STUDY/學 xué









Christian Wittern Kyoto University

A closer look at HXWD.org and how the TLS is implemented there

The road to HXWD



Porting details (1): low hanging fruits

- Criteria for prioritation of texts
 - Complete text
 - ► Translation available
 - Reliable source text available
 - e.g. either Kanripo or CHANT
- Core features
 - Concepts, Syntactic functions, Semantic features
- Rhetorical devices (added, but not yet fully functional)



Porting details (2): todos and donots

- ► TODO:
 - Phonetic representation ("Guangyun" file)
 - Bibliography
- Other stuff
 - Word composition
 - ▶ .



Porting procedure

- Content of Filemaker database written to ad-hoc XML representation
 - Final export by Jens Østergaard Petersen on July 29, 2016
- Data modeling and transformation
 - For texts, translations and annotations:
 - intermediate combined version in Kanripo format (ManDoku format, hereafter MD)
 - comparison and manual proofreading of TLS text and new source text
 - conversion to separate format (Chinese text and translation)
 - development of and conversion to HXWD annotation format (attributions)
 - Concepts, Syntactical functions and semantic features (later also rhetorical devices)
 - Customized and highly formalized files, close reflecting the display screens in Filemaker
- ▶ Until Feb 2017: Conversion of the whole database to MD format
- ► From Mar 2017: Re-modeling into XML, mostly based on TEI (Text Encoding Initiative) recommendations
- From Oct 2018: Development based on eXist-db
- ▶ 2019-01-05: Register of HXWD.org domain and test version of online version

, = .2.4C

Christian Wittern Kyoto University

A closer look at HXWD.org and how the TLS is implemented there

HXWD.org behind the scenes

Introduction

The HXWD interface

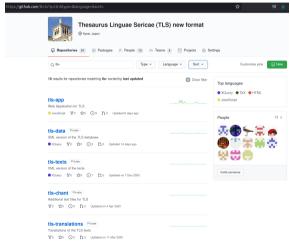
The road to HXWD

HXWD.org behind the scenes

More interactions with HXWE



TLS source files are on GitHub



HXWD.org behind the scenes

000000000000



TLS Repositories

- Public repositories
 - tls-app: Application source code
 - tls-texts: Chinese source texts ("Core texts")
 - these are texts that were part of the original TLS or have been curated in the new format

HXWD.org behind the scenes

000000000000

- tls-data: Interpretational layer (excluding proprietary translations)
- Private repositories
 - tls-translations: proprietary translations
 - tls-chant: Texts provided by the CHinese ANcient Texts project in Hongkong ("More texts")
- In the database, tls-translations is part of the tls-data app; tls-chant is part of tls-texts
- ▶ There is also a user area for the registered users, this will not be pushed to GitHub

Christian Wittern Kvoto University

Views of the database

- ► The public view of the database is generated by the code in tls-app
 - it provides a customized, focussed view, usually combining data items from many different sections of the database

HXWD.org behind the scenes

0000000000000

- All user interactions, including editing, translating, attributing are chanelled here
- eXist-db provides also other interfaces to the database:
 - Dashboard / eXide: Browser based edit/view of database content
 - ► XML-RPC: used by oXygen XML editor to interface with eXist-db
 - RESTXQ: execute XQuery through a REST interface



Rhet-dev.xml in oXygen XML editor

```
rhetorical-devices.xml [oxygen:/eXist-db localhostSeXist-db hxwd.org/db/apps/tls-data/core/rhetorical-devices.xml] - <oXygen/> XML Editor (Academic
File Edit Find Project Options Tools Document Window Help
n = = n c ♥ Q a = × ← → [
                                                  - o. D. O. & O. = D > - m ×
XPath 2.0 - D . Execute XPath on "Current File
                       西東米 *text-and-trans.xg × *rhetorical-devices.xml* × *text-pron.xgl × *transform1.xsl × *add-swl-id.xg × *find-att-wo-id.xg × *anye.html × *text.xsl × *anpendix3.xsl × *xml2mc
مه Data Source Explorer
                               TEI text body div div list
                                    1vkTEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:mods="http://www.loc.gov/mods/v3" xmlns:xlink="http://www.w3.org/1999/xlink">
_LeXist-db krx.hxwd.org
                                                <fileDesc>
A eXist-db hxwd.org
                                                    <titleStmt>
· fl db
                                                        <title>Rhetorical Devices in TLS</title>
 . Anns
                                                    </titleStmt>
   - Adashboard
                                                    <publicationStmt>
  - [] doc
                                                         Published electronically
   - 🗎 eXide
                                                    </publicationStmt>
   - [] existdb-backup
                                                    <sourceDesc>
   - B fundoes
                                                         Generated from the Filemaker export of 2016-07-29 
   - B markdown
                                   12
                                                    </sourceDesc>
   - fl monex
                                   13
                                                </fileDesc>
   - [] packagemanager
                                   14
                                           </teiHeader>
   - | packageservice
                                   15
                                           -tayta
   - 1 shared-resources
                                   16
                                                <body>
   - El tis-ann
                                   17
   . B tleudata
                                   18
                                                    <div type="rhet-dev" xml:id="uuid-acbd1441-4ala-4ccd-839a-cf5fe55de62e">
    - [] bibliography
                                   19
                                                         <head>HENDTADYS</head>
    - fl concepts
                                                        <div type="definition">
    + Fl core
                                   21
                                                             PLEONASM in two consecutive words.
                                                             重点法 The expression of what can be said with one word or expression by the
       Arhetorical devices xml
                                   22
                                                                 adjacent use of two near-synonymous words or expressions.
       semantic features vml
                                   23
                                   24
                                                        </div>
       evotactic-functions yml
                                   25 1
                                                         <div type="pointers">
       taychar yml
                                   26
                                                            t type="taxonymy">
    - 🗎 guangyun
                                   27 -
    - F medical
                                                                     <ref target="#uuid-bf228c09-612c-4fdc-b9e7-87e07cbee113">HENDIADYS-MINIMAL</ref>
                                   28
    + fil modules
                                   29
                                                                 </item>
    - A notes
                                   30 =
                                                                 <item>
    » B schams
                                   31
                                                                     <ref target="#uuid-76558a96-3337-4b96-b692-9bcafa840250">HENDIADYS-MINOR</ref>
A. Data Source Explorer 💢 Project
                                   32
                                                                 </item>
2 Outline
                                                                 citems
```

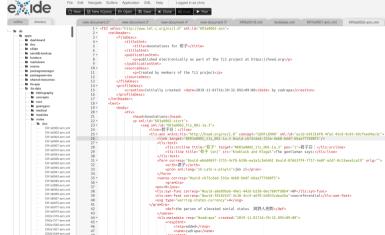
HXWD.org behind the scenes

0000000000000



Christian Wittern Kyoto University

Attributions for Xunzi seen in eXide





Beginning of Lunyu 1, TEI XML format

```
- 🗎 chant
                                 1 VKTEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="KR1h0004">
    - manifests
                                       <teiHeader>
    - | modules
                                 3 -
                                          <fileDesc>
                                             <titleStmt>
    + 🗎 tis
                                                <title>論語</title>
      KR1a0001 vml
                                             </titleStmt>
      KR1b0001.xml
                                             <editionStmt>
      R v
                                                    <edition>
      KR1c0066.xml
                                 9
                                                        <idno type="kanripo">KR1h0004</idno>
      KR1d0001.xml
                                                    </edition>
      KR1d0052.xml
                                11
                                                </editionStmt>
      KR1e0001.xml
                                123
                                             <publicationStmt>
      KR1e0008 vml
                                13
                                                Published electronically
      KR1f0001.xml
                                14
                                             </publicationStmt>
      KR160014 vml
                                15
                                             <sourceDesc>
      KR1h0001.xml
                                                Generated from the TLS textfiles in Kanripo format on 2019-03-20
                                16
      KR1h0004.xml
                                             </sourceDesc>
      KR1i0007.xml
                                18
                                          </fileDesc>
      KB2a0001 vml
                                19
                                       </teiHeader>
      <text>
      KR2e0001.xml
                                21 🔻
                                          <body>
      22 0
                                             -div-
                                23 9
      <seg xml:id="KR1h0004_tls_001-1a.3-h">1 《學而篇第一》</seg>
                                24
      KR2a0017.xml
                                25
                                                </head>
      KR2m0054.xml
      KR2n0001 vml
                                26 =
                                                >
                                27 
      KR200025.xml
                                28
      KB2n0015 vml
                                                   <seg xml:id="KR1h0004 tls 001-1a.3">1.1子曰: <lb/>
                                29 🕶
      KR2n0018.xml
      KR2n0019 xm
                                31 3
                                                   <seq xml:id="KR1h0004 tls 001-1a,4">「學而時習之, <lb/>

    L Data Source Explorer

                    Project
                                32
Outline
                      anx
                                33 ♥
                                                   <seg xml:id="KR1h0004_tls_001-1a.5">不亦説乎? <lb/>
                     -Q #
                                34
                                                        </sea>
TEI *KR1h0004*
                                 35 1
                                                   <seq xml:id="KR1h0004 tls 001-1a.6">有朋自遠方來、<lb/>
- P telHeader
                                 36
e text
                                                   <seq xml:id="KR1h0004 tls 001-1a.7">不亦樂乎? <1b/>
                                37 9
· 🗎 body
                                38
                                                        </seq>
```

HXWD.org behind the scenes

0000000000000

Christian Wittern Kyoto University

Network of information items

- XML data allows for a great degree of freedom in information modeling
- ▶ different parts of the TLS can be modeled differently
 - as appropriate for the items at hand:
 - texts as hierarchy of text objects
 - concept descriptions as structured collection of information items (which can themselves be hierarchical, and linked to other items)
 - ▶ attributions as nexus points (a link between a text location and a syntactic word SW)

HXWD.org behind the scenes

00000000000000

- these elements are connected through ID and IDREF
 - similar to cross references in printed books: "see page 27" is an IDREF, the page with the number 27 is a referrable location (ID)
 - in HXWD, this is implemented as oxml:id and mostly occurresp (or occurre



HXWD.org implements a stateless client/server architecture

- ► Client requests a page from server
- ▶ Server assembles the page, sents it to the client and forgets about it
- ▶ Client requests the "attributation floater" for 莊周
 - this is done via XHR: the page is not updated, but only modified
- Client clicks on "Use" to attribute
 - ▶ a XHR is sent to the server, the server saves the attribution
 - if the action succeeds, the interface displays a short acknowledgement
 - at this point, the page on display differs from the state in the server
 - to bring this in sync, the client issues another XHR to get the latest state from the server.



HXWD.org implements a stateless client/server architecture

- ► Client requests a page from server
- ▶ Server assembles the page, sents it to the client and forgets about it
- ▶ Client requests the "attributation floater" for 莊周
 - ▶ this is done via XHR: the page is not updated, but only modified
- Client clicks on "Use" to attribute
 - ▶ a XHR is sent to the server, the server saves the attribution
 - ▶ if the action succeeds, the interface displays a short acknowledgement
 - ▶ at this point, the page on display differs from the state in the server
 - to bring this in sync, the client issues another XHR to get the latest state from the server.

HXWD.org behind the scenes

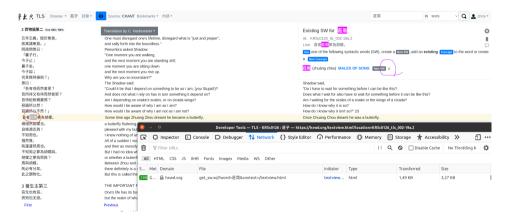
0000000000000

▶ We will look at an example in the next slides



Christian Wittern Kyoto University

Looking at 莊周 attributions

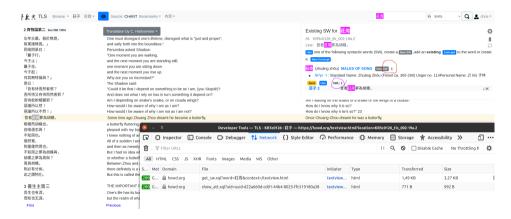


HXWD.org behind the scenes

000000000000



Searching for text locations



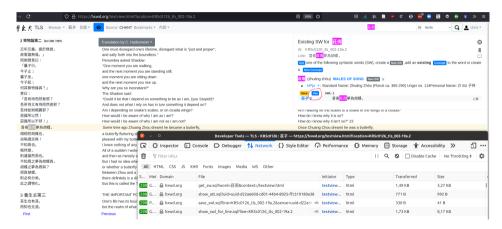
HXWD.org behind the scenes

000000000000



Kvoto University

Making a new attribution



HXWD.org behind the scenes

0000000000000



Pros and cons of the client/server approach

- ► Pro
 - User interfaces can be designed to make specific tasks very convenient
- Con
 - Every interaction has to be specifically designed, most of the times in several steps

HXWD.org behind the scenes

000000000000

- Only interactions foreseen and planned for can occur
- Other issues:
 - network issues can hamper the communication, with unpredictable results
 - if in doubt, reload the page to be sure to get the latest state from the server
 - do not use the back button of the browser:
 - the back button displays a locally cached version of the page received from the server earlier, without any modification that occurred later through user interaction!



Kvoto University

Pros and cons of the client/server approach

- Pro
 - User interfaces can be designed to make specific tasks very convenient
- Con
 - Every interaction has to be specifically designed, most of the times in several steps

HXWD.org behind the scenes

000000000000

- Only interactions foreseen and planned for can occur
- Other issues:
 - network issues can hamper the communication, with unpredictable results
 - if in doubt, reload the page to be sure to get the latest state from the server
 - do not use the back button of the browser:
 - the back button displays a locally cached version of the page received from the server earlier, without any modification that occurred later through user interaction!
- ▶ It is far from ideal, but on the Web anything else is difficult



A closer look at HXWD.org and how the TLS is implemented there

Introduction

The HXWD interface

The road to HXWD

HXWD.org behind the scenes

More interactions with HXWD

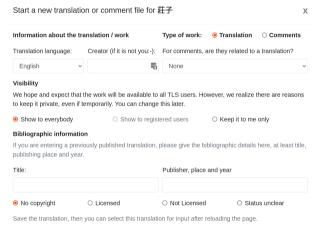


Create a new translation/comment area

- All regular users can start a new translation or comment file
 - (these are the files that populate the columns in the middle and right area of the screen)
 - to activate the dialog, click on the translation header and select "New translation/comments"
- translation files are specific to the text they accompany
- ► The following slide shows the input screen
 - Be careful when using it: Everybody can create files
 - but only the admin can delete them
- Translation files marked as "Keep it to me only" will be saved to your user area and not be visible to other users
- The actual translation can than be entered line by line. Moving to the next line with <tab> saves the line to the server



Add a new translation dialog





Christian Wittern Kvoto University

Stars, logs etc.

- Texts of specific interest can be rated with stars
 - This will add them to the "Starred texts" tab for easy access and allows to prioritize them in searches
- Interactions that change data are usually marked with userid and modification date
 - in some cases earlier versions of a changed state (for example a translation line) is kept in a log file
- Visits to text locations in the textview screen are logged. This allows you to return to the most recent place in a text easily, e.g. from the "Starred texts" tab
- Fixed locations of interest can be bookmarked (from the attribution floater)



Sometimes the server freezes :-(

- Some interactions cause a lot of internal changes to keep the data consistent
- In such cases, the server might stop to respond to new requests
- Other unknown and untraceable circumstances can also lead to freezes
- At noon and midnight JST the server makes a full backup
 - to ensure consistency of the backed up data, this also requires all interactions to stop
 - this will usually take 3-5 minutes



- Some interactions cause a lot of internal changes to keep the data consistent
- In such cases, the server might stop to respond to new requests
- Other unknown and untraceable circumstances can also lead to freezes
- At noon and midnight JST the server makes a full backup
 - to ensure consistency of the backed up data, this also requires all interactions to stop
 - this will usually take 3-5 minutes
- Most of the time, the server will unfreeze itself in due time



- ► A lot work needs to be done, specifically
 - Marking and annotating of text segments
 - ► Integrating Zen Knowledgebase (zenkb)
 - Metadata for texts (esp. CHANT texts)
 - Adding/Linking of additional texts
 - User management
 - **.**..



Web interface Demo

► (now, finally)



Web interface Demo

- ► (now, finally)
- Concept browsing
- Text search
 - other searches, options
- Character taxonomies



- (now, finally)
- Concept browsing
- ► Text search
 - other searches, options
- Character taxonomies
- And of course: Questions?



Web interface Demo

- ► (now, finally)
- Concept browsing
- Text search
 - other searches, options
- Character taxonomies
- And of course: Questions?
- ► Comments?

