Character taxonomies, text editing and other recent developments in the TLS collaborative research environment

Christian Wittern

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

2022-10-28

Recent developments on HXWD

HXWD Roadmap

Text editing in HXWD

Text import

Character taxonomies

Wikidata (WD) integration

Reasons for not moving to TEI Publisher (TP)

- Very different approach to Web development
  - TP: Web components, etc.
  - HXWD uses a template based environment
- TP is mainly concerned with (read-only) publication
  - HXWD attempts to support ongoing research and text preparation
- In TP annotations are inserted into the texts
  - We want to maintain (most) annotaions external to the texts

But there are some nice features in TP!

- Different views of a text (by structural division, or printed page)
- Number of panels displayed can vary
- Facsimile can be shown next to the transcribed text
- Flexible catalog and listing of texts
- User upload of texts
- Download in many formats
- => We will try to do most of this (and more!) in HXWD

Some features on our wishlist (not in TP)

- User management
  - Self-application for account with subsequent vetting
- Bibliography
  - Adding and editing of references
  - Linking references to other parts of the TLS
- Please join the discussion in GitHub Issues

Recent developments on HXWD

HXWD Roadmap

Text editing in HXWD

Text import

Character taxonomies

Wikidata (WD) integration

Text editing

- Annotations in HXWD require text segmented into phrases
- So far, the texts had to be punctuated to be made available in HXWD
- We now support punctuation within HXWD
- Text can be imported without punctuation
  - the process of text establishment should be entirely supported within HXWD

Process of punctuation

- Unpunctuated texts come into HXWD aribtrarily segmented into textual units of variable size
- These segments are locked and not available for annotation
  - They do show up in search results!
- Locked segments can be loaded into a dialog for punctuation
  - These segments will then be split at punctuation marks into smaller segments
  - Where necessary a segment can also be merged with the following one
- The lock will now be removed for these segments, thus making them available for annotation
  - so, even before the whole text is available, isolated segments can be annotated

More on the editing process

- All members of tls-edit have editing rights
- Other users can be given the right to edit specific texts
- Deletion of unwanted segments is also possible
- Text critical annotation will be made available soon

Recent developments on HXWD

HXWD Roadmap

Text editing in HXWD

Text import

Character taxonomies

Wikidata (WD) integration

Import from the Kanseki Repository (KR)

- Most of these texts are not punctuated, therefore were not directly available
- We now can import them in the current state and edit within HXWD
- Since HXWD allows only one text per work, the “master” version will be used
    - In the future, other versions could be added via a text-critical apparatus
- Structural features of the text can be partially interfered from layout
- Commentaries and intralinear notes will be handled
    - There is a great variety of how these things are expressed, so at the moment the development of heuristic procedures proceeds one text at a time

Import from CBETA

- The Chinese Buddhist Electronic Text Association (CBETA) is making all their texts available on GitHub (`https://github.com/cbeta-org/xml-p5`)
- These texts are of high quality and contain text critical information on many other versions
- The XML format is based on TEI, but with modfifications
- A import routine for HXWD has been developed to take advantage of this and supports a direct conversion to the format required by HXWD

Assesment of imported texts

- Since the import workflows outlined above is largely an automated process, its results need to be evaluated before they can be used
- Routines for analysing structure and content and detect errors have been developed
- However, because of the great variety of text types and textual features, they have also to be checked for textual integrity manually
- HXWD is using now a traffic light code for communicating the state of a text to the user
  - Imported texts are that pass basic tests are marked "red"
  - Texts that have been evaluated as stable and completely segmented are marked "yellow"
  - Texts that have been completely established text-critically will be marked "green"

Requests for texts

- Texts in the Kanseki Repository can be requested for inclusion
  - A button for requests is shown on the title search page
- If the texts are originally from CBETA, their XML version will be used to retain maximum information
- Texts from the Daozang will be taken from KR, but those of the Daozang jiyao texts where XML versions are available, we will use those
- Texts from other sources (ctext.org, zh.wikisource.org) etc might be available for requesting at a later state
- At some point, we want to allow users to upload their own texts as well

Recent developments on HXWD

HXWD Roadmap

Text editing in HXWD

Text import

Character taxonomies

Wikidata (WD) integration

Background

- During the transition between the Filemaker TLS and HXWD, Christoph Harbsmeier was working on the project of establishing how different semantic fields of a character relate to each other
- We called these tree structures character taxonomies
- He created these structures using `org-mode` in `Emacs`
- They were based on the then current CONCEPTS and annotaions
- Character taxonomies for the most frequent characters in Classical Chinese are listed on HXWD

Editing of Character taxonomies

- Since the TLS is constantly changing, the existing character taxonomies are outdated
- Replicating the editing capabilities of `Emacs` on a web page proved difficult
- We now have at least some rudimentary way of updating and editing the character taxonomies
- This feature is only available to tls-editor members
- However, all users can enjoy the results:-)

Prototypical annotaions

- The character taxonomies did occasionally include example sentences to explain certain fine points
- In the new format, this is frowned upon (although not entirely impossible)
- As a remedy, we introduced a way to mark certain SWL annotaions as prototypical
- These will be shown exclusively or more prominently in certain contexts, such as the character taxonomies
- At the moment, they are simply distinguished by a different background color

Recent developments on HXWD

HXWD Roadmap

Text editing in HXWD

Text import

Character taxonomies

Wikidata (WD) integration

Overview

- Inspired by Fudie Zhao's presentation two weeks ago, I took another look at connecting HXWD to WD
- At some places on HXWD pages are now little blue-green icons labelled "WD"
- These can be used for a very broad and general search on the Wikidata website
- This is still explorative and experimental

Use case: Text catalog disambiguation

- Some texts are available in HXWD and KR with different identification numbers
- This leads to unwanted text duplication
- By associating the texts to the corresponding items in Wikidata the catalog can be disambiguated
- This is now implented on the title search page and on the text information display "SOURCE" in the top section of the screen on a displayed text.

Implementation details: Qitems

- All entities in Wikidata are addressed by an identification number beginning with Q, called Qitem here
- For every WD item registered to establish a link between HXWD and WD, a Qitem file is created on HXWD
- Initially, these files serve only to link between this sites
- At a later stage, links to additional Qitems and properties on WD can be established to discover new information
- The coming months will show the usefulness of the linking to WD
- If this turns out to be a hassle, it can be easily removed again

Next steps

- The developments reported today mark an important point in the development of HXWD
- Part of the original vision of merging the Kanseki Repository with the TLS seems realizable now
- Experience will show whether the server can cope with the load of all 10000 texts from KR
- Important fields for further investigation are intertextuality, specifically
  - link between texts and commentaries
  - link between texts and text reuse (quotations)
  - semantic linking through usage of key terms
- New specialized fields, eg. mathematical texts
- And don't forget: please join the discussion on GitHub Issues