Mandoku – An Incubator for Premodern Chinese Texts – or How to Get the Text We Want: An Inquiry into the Ideal Workflow

Wittern, Christian
cwittern@gmail.com
Kyoto University

Premodern Chinese texts pose problems that are difficult to accommodate with the current TEI text model, which bases the main hierarchy of a text on its structural content, rather than on a hierarchy that models the pages, lines and character positions. For the TEI, this is a sensible decision and has led to the abolishment of elements like <page> and <line> in the latest release of the Guidelines. For premodern Chinese texts however, especially texts that are transmitted as manuscripts or woodblock printings and have not yet seen a modern edition printed with movable type (let alone as, more recently, computerized typesetting), establishing the structural hierarchy of the text content is, together with the even more daunting question of establishing the proper characters of the text (on which see below), an important part of the research question that motivates the digitization of the text. Requiring an answer to this question before a proper electronic text can be created makes this intractable in the digital medium and glosses over an important leap of faith in the creation of a TEI encoded text. In this paper, I will try to trace some of the implications and propose an approach that allows different models of the text for different stages in the encoding process, thus closer modeling the process of the creation of an electronic text.

To arrive at a text properly encoded according to the TEI Guidelines is not a straightforward process, but in the setup described here requires a detour through at least three stages:

- Draft input of the text without any further markings;
- An incubator phase, in which the text is dealt with as a series of pages (or scrolls), lines and characters;
- The mature text, based on the structural model of a TEI <text>, which is available for further refinement;

Of these steps, the second one is at the center of attention in this paper, which will include the discussion of the following three aspects:

1. A text model according to these requirements;
2. Mandoku, an application that allows manipulating the text;
3. A transform that specifies how a text conforming to this specification can be turned into a TEI encoded text.

1. Different Models for the Same Text

The structural, content based hierarchy of the text has to be established as part of the research process. For this reason, the text at this stage uses the only hierarchy available, that is the one that is based on how the text is physically recorded on the writing surface in the edition used. During the process of working with the text, milestone-like elements are inserted at the starting points of elements of interest, using the incubator as described in the next section. Headings are numbered according to their nesting depth as in a HTML document; this forms the base for their transformation into regular TEI nested s followed by <head> elements.

2. The Incubator: Mandoku

The tool used to manipulate a premodern Chinese text in the incubator phase has been called Mandoku. It makes it possible to display a digital facsimile of one or more editions and a transcribed text of these editions side by side on the same screen. From there, the texts can be proofread, compared and annotated. A special feature is the possibility to associate characters of the transcription with images cut from the text and a database of character properties and variants, which can be maintained while operating on the text. Interactive commands exist also to assist in identifying and record structural and semantic features of the texts.

One of the major obstacles to digitization of premodern Chinese texts is the use of character forms that are much more ideosyncratic than today's standard forms. Since in most cases they cannot be represented, they are exchanged during input for the standard forms. This is a potentially hazardous and error-prone process at best, and completely distorts the text in worse cases. To improve on this situation and to make the substitution process itself available to analysis, Mandoku uses the position of a character in a text as the main means of addressing, allowing the character encoding to become part of the modelling process, thus making it available to study and analysis, which in turn should make the process of encoding more tractable even for premodern texts. The current model is still experimental, but initial results have been encouraging.

Mandoku is work in progress and is developed as part of the Daozang jiyao project at the Institute for Research in Humanities, Kyoto University by Christian Wittern. In this paper, an emphasis will be placed on the different models of a text that are underlying the different stages of preparation of a text and the friction, but also benefits, that arise out of such a situation. The following is a screenshot of the main interface, displaying a facsimile and a transcribed version of the same text side by side.

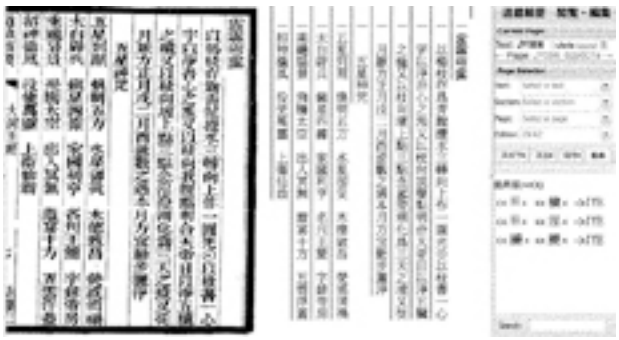

Fig. 1. *Mandoku in action*

3. Transform to TEI <text>

Finally, as a proof of concept, a XSLT script has been developed that performs an algorithmic transformation from the text in the intermediate format to a text as it has to appear as content of the TEI <text> element.

This produces a new version of the text with an inverted hierarchy: The primary hierarchy now is the content hierarchy, whereas the hierarchy of the text bearing medium is demoted to a secondary one, represented by milestones. None of these hierarchies is a priori superior to the other, but in the context of the Daozang jiyao project the purpose of preparing the texts is to make it available for a study of the collection, so the emphasis during the later stages in the life of the text will lie on the content hierarchy. The problem of overlapping hierarchies, which is such a scratching itch for many text projects, poses itself thus in a slightly different incarnation: The different hierarchies occur in two different stages of preparation of the text, which require different viewpoints, but not simultaneous presentation, which makes it easier to accommodate the two in our workflow.

The preparation of a TEI encoded representation of the texts is however not the ultimate goal of the project. The next phase requires analytical interaction with the text for which again the TEI representation might not be the ideal format to work with, so there might be a number of different, purpose-specific derivative formats generated from the TEI texts. They will maintain the required information to refer additional information back to the master files kept in TEI, and to be able to participate in the ongoing evolving of the master text, to which transcriptions of more witnesses will be added, but will otherwise also contain additional commentary, translation and other information that will not belong to the original file. The details of this part of the system are under consideration now and will be the topic for another presentation.

4. Conclusions

The current TEI text model does not allow the direct description of the document as it appears on a text bearing surface without also establishing a content hierarchy. For this reason, a temporary encoding strategy had to be developed, which is TEI conformant to the letter, but not to the spirit of the TEI Guidelines by wrapping all of the text content in one giant <p> (or possibly <ab>) element. Only after the structural hierarchy has been established is it possible to make a transformation to a truly conformant and satisfying TEI document. The slight feeling of uneasiness that this workaround causes might go away once the new <document> element proposed by the TEI working group

on genetic editions has been adapted to the Guidelines and can be used for the phase of work in the incubator, thus making the text fully TEI conformant from the beginning. On the other hand, this project also clearly demonstrates the necessity of being able to represent the document in its own right in a TEI text, even if in the context of this project the documentary part is considered transitory.

References

Esposito, Monica (2009). 'The Daozang Jiyao Project: Mutations of a Canon'. *Daoism: Religion, History and Society*. **1**: 95-153.

TEI Consortium (ed.). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.6.0.* `http://www.tei-c.org/Guidelines/P5/` (accessed 2010-03-09).

Wittern, Christian (2007). 'Patterns of Variation: The textual Sources of the Chinese Buddhist Canon as Seen through the CBETA edition'. *Essays on East Asian Religion and Culture.* Christian Wittern and Shi Lishan (ed.). Kyoto: Editorial committee for the Festschrift in honour of Nishiwaki Tsuneki, pp. 209-232. `http://kanji.zinbun.kyoto-u.ac.jp/~wittern/data/nw-fs/fs-wittern.pdf` (accessed 2010-06-05).

Wittern, Christian (2009). 'Digital Editions of premodern Chinese texts: Methods and Problems – exemplified using the Daozang jiyao'. *Early Chán Manuscripts among the Dūnhuáng Findings – Resources in the Mark-up and Digitization of Historical Texts.* University of Oslo, Sep. 28 to Oct. 3, 2009, preprint PDF available. `http://kanji.zinbun.kyoto-u.ac.jp/~wittern/data/digital-editions-dzjy.pdf` (accessed 2010-03-09).