Chung-Hwa Buddhist Journal (2012, 25:167-194)
Taipei: Chung-Hwa Institute of Buddhist Studies
中華佛學學報第二十五期　頁 167-194 (民國一百零一年)，臺北：中華佛學研究所
ISSN:1017-7132

Digital Editions of Premodern Chinese Texts: Methods and Problems – Exemplified Using the Daozang Jiyao[1] 道藏輯要

Christian Wittern
Kyoto University

Abstract

Digital editions do have a great potential for new avenues of research, but they also pose vexing research questions that have to be resolved adequately in order to make the resulting edition useful in the long run. One of the many differences between printed editions of texts and digital editions is the open-endedness of the latter, which means that it can be done incrementally and updated without incurring substantial expenses. The medium of digital editions requires the creator to make many assumptions about the texts explicit and record them in a way that can be processed automatically. This is a new concept, which seems foreign to the agenda of a scholar whose ultimate aim is to engage with the text. This article demonstrates that what seems like a detour is actually advancing the understanding of the text and the need objectify a text in this gives access to new dimensions of a text. It then goes on to provide details of a conceptual model for describing a premodern text digitally that has been developed working on a digital edition of the early Qing Daoist collection Daozang jiyao.

Keywords:

Text Encoding, Digital Editions, Character Encoding, XML, Doaist Studies

[1] I would like to thank the anonymous reviewers for this journal for their very helpful suggestions for clarifications and general improvement of the article.

前現代漢語文本的數位版本：方法與問題─以《道藏輯要》為例

維習安
京都大學

摘要

數位版本對於研究的新方向有著極大的潛力，但它們也引起令人困擾的研究問題，而這些問題必須適當地解決，以使得此版本的成果對長期來說是有所助益的。紙本與數位版本的許多差異之一是後者的開放性，也就是能夠不需要實質上的花費而增加或更新。使用數位版本工具需要建立者建構許多有關文本的清晰假設，且能夠以自動執行的方式紀錄；這是一個新的概念，似乎對目標為參與文本的學者之預設立場不同。此篇文章說明那些看起來像是繞道而行，但事實上卻是增進文本理解與客觀化文本需求的情形，依此開啟進入了解文本的新面向；同時並詳盡地提供說明數位化前現代文本的概念模式，而此工作是建立在清初的《道藏輯要》之數位版本上而發展的。

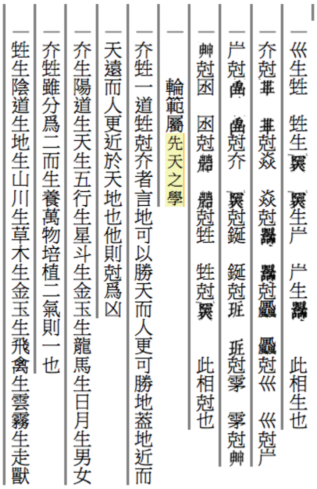
關鍵詞：文本編碼、數位版本、字元編碼、XML、道教研究

Introduction

Text transmitted on traditional written surfaces is immediately available and transparent to the reader, without any additional steps involved. In contrast to this, any text stored digitally, in whatever format, has to be rendered to the screen (or paper) by correctly interpreting (decoding) the values of 0 and 1 that have been used to prepare (encode) the text. Without this correct interpretation, the result of the decoding will be just illegible garbage that does not make any sense whatsoever.

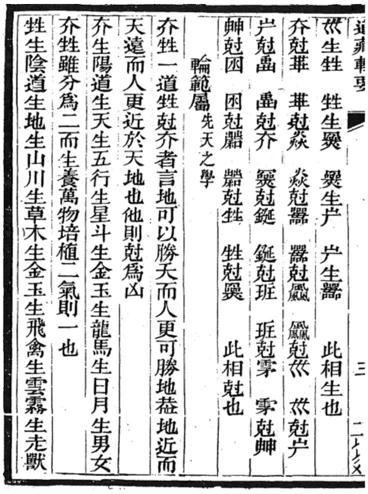
In order to make this decoding successful, the model, according to which the encoding was done, has to be known at the time of the decoding. Even more importantly, as is true for any digital format, the encoding of text into digital format can not be done without a model of the text. The activity of developing and enhancing a model of the text thus becomes a crucial, foundational activity, laying the groundwork for the actual digitization of texts themselves.

The first fundamental decision that has to be made when devising such a model is to whether to treat the text either just as a series of symbols or as a two-dimensional array of spots of different color spread out over a flat surface. Descendants of the first type of model would lead to a transcribed version of a text (an example of a page is shown in Figure 1), while those of the second type of model would be some kind of facsimile representation of the text, these will be called digital facsimile (see Figure 2). None of these representations is intrinsically superior to the other; they do in fact very nicely complement each other.



巛生甡　甡生巽　巽生屵　屵生䨻　此相生也
夰尅丵　丵尅焱　焱尅䨻　䨻尅鬳　鬳尅巛　巛尅屵
屵尅䖵　䖵尅夰　巽尅鋋　鋋尅班　班尅雩　雩尅𠦂
𠦂尅函　函尅鼎　鼎尅甡　甡尅巽　此相尅也
輪範屬先天之學
夰甡一道甡尅夰者言地可以勝天而人更可勝地盇地近而
天遠而人更近於天地也他則尅爲凶
夰生陽道生天生五行生星斗生金玉生龍馬生日月生男女
夰甡雖分爲二而生養萬物培植二氣則一也
甡生陰道生地生山川生草木生金玉生飛禽生雲霧生走獸

Figure 1: An example of a transcribed text



巛生甡　甡生巽　巽生屵　屵生䨻　此相生也
夰尅丵　丵尅焱　焱尅䨻　䨻尅鬳　鬳尅巛　巛尅屵
屵尅䖵　䖵尅夰　巽尅鋋　鋋尅班　班尅雩　雩尅𠦂
𠦂尅函　函尅鼎　鼎尅甡　甡尅巽　此相尅也
輪範屬先天之學
夰甡一道甡尅夰者言地可以勝天而人更可勝地葢地近而
天遠而人更近於天地也他則尅爲凶
夰生陽道生天生五行生星斗生金玉生龍馬生日月生男女
夰甡雖分爲二而生養萬物培植二氣則一也
甡生陰道生地生山川生草木生金玉生飛禽生雲霧生走獸

Figure 2: An example of a digital facsimile

If a text is to be used for information retrieval or any other purpose that requires access to its symbolic content, like for example, text analysis or even the creation of a new version with a different layout, it has to be encoded in a way that somehow represents the symbols used to write the text. This requires a reading of the text and is thus always also an interpretation of the text.

While the transcription of a text as a series of symbols is comparatively straightforward in most alphabetical languages, the logographic languages of East-Asia pose specific problems, since exactly this transcription is not a given, but is open to various interpretations and in fact has to be considered part of the research question. It thus needs a model that allows to make these interpretations transparent instead of hiding them in the transcription process, which takes place before the text even gets to the reader. This paper will discuss models used for such a representation and proposes a new working model specific for premodern Chinese text.

It might be tempting to try to avoid the whole issue of legacy character encoding and try to come up with a completely different way to encode characters. One such attempt is the CHISE project[2], which tries to build a whole ontology of characters and character information. In the model discussed here, the encoding is based on Unicode, but an intermediate layer of dereference is introduced as explained below.

In the practice of transcribing primary sources, there is an additional complication through the fact that there might be more than one witness for a text and therefore a collation and analysis of textual variants in other text witnesses might be required. The model will have to be able to account for this.

One last requirement is that it has to be possible to establish and maintain a normalized version of the text in addition to establishing a copy text faithful to the original.

Preliminaries and Prerequisites

Before starting to describe the proposed new model, some preliminaries and basic assumptions have to be discussed. This involves a very brief description of the model most widely used for transcribing primary sources, but will also involve a brief discussion of the writing system for Chinese and how its basic properties have been reflected in today's most widely used character encoding, Unicode.

[2] See the CHISE (Character Information Service Environment) project. (*http://www.kanji.zinbun.kyoto-u.ac.jp/projects/chise/*)

The TEI/XML Text Model

Text encoding according to the recommendations of the Text Encoding Initiative (TEI) is today the most widely used format for the creation and processing of texts for research in the Humanities.[3]

In XML, which is the technical basis for the TEI text format, a text is basically seen as a hierarchy of textual content objects, expressed as a hierarchy of XML elements and attributes[4], this is the so-called OHCO (Ordered Hierarchy of Content Objects) view of a text. While this provides a powerful model to deal with many aspects of a text and allows the definition of sophisticated vocabularies, there are a few problems that are hard to solve using this model.

One of these problems is that digital texts do in fact require different hierarchical views, depending on the purpose of the creation and the intended processing of the text. There are several ways the TEI attempts to solve this problem, one of them being considering one of the hierarchies in a document as the primary hierarchy (Guidelines, 20.3 Fragmentation and Reconstitution of Virtual Elements). Textual features that do not nest cleanly into this hierarchy are then arbitrarily split into two (or more) parts. And then introducing additional notions, that can be used for example to virtually join elements together, which have been arbitrarily split within the primary hierarchy.

Another way to overcome this problem is by using elements without text content to indicate points in a text, at which features of the 'other' hierarchy starts. A classic example for this is the use of milestones in TEI. Since the main hierarchy of a TEI document is constructed using elements that describe the semantic content of the document (e.g. <body>, <div>, <p>),[5] elements that hold the content of pages and lines can not exist in the same hierarchy. Pages (and columns and lines; these are all generalized into the concept of 'milestones') are thus only indicated by marking the point in the text flow where a new page begins. This makes it possible to work with both hierarchies at the same time, but there is a tradeoff: It prioritizes one hierarchy, thus making it considerably more difficult to retrieve the content of a page, as opposed to the content of, e.g. a paragraph.

[3] It goes without saying that TEI can be used to encode premodern Chinese texts, which is amply demonstrated for example by the texts produced by the Chinese Buddhist Electronic Text Association (CBETA), whose latest release had to be put on a DVD, since even in compressed form, a CD-ROM could not hold the amount of material anymore. The earliest of these texts are nearly 2000 years old.

[4] See for example Renear&Mylonas&Durand (1996).

[5] Earlier versions of the TEI contained elements <page>, <col> and <line> etc, which could be used to construct a concurrent hierarchy that reflects how the text was laid down on the text bearing surface, but these have been removed in the latest release, P5.

There is also another difficulty of a more practical nature, that is, through what procedure the encoded text is created. If text encoding is seen as a process of gaining insight and enhancing the understanding of a text, this will be a circular process that adds more information in several passes through the text. What this means is that the sophistication of the TEI model, while serving the needs of text encoders well in providing the expressive power to encode the features observed in a text, it puts an enormous burden on text encoders, wishing to employ the system for their texts. This seems to be especially true for premodern Chinese texts, where not only the writing system poses additional difficulties, but there is also usually no indication of paragraph or sentence boundaries, punctuation; the only given is the text as it is divided into 'scrolls', pages and lines. For the purpose of this model then, the main hierarchy in the document is that of the physical representation of the text on the text bearing surface of the witness that is serving as the source for digitization. As the encoding of the text progresses, markers of the points of change in the content hierarchy are inserted, thus gradually bringing this other hierarchy into existence. In some ways this is thus an inversion of the relationship between these hierarchies as they exist in the TEI model. The following discussion will be targeted at requirements of Chinese text and no claims are made about usefulness in other areas.

The model described in this paper is not intended as a replacement for the TEI text model, but rather as a heuristic, methodological model that allows the creation of a sophisticated text, most likely as the childhood of a text that will prepare it to spend its adult life in a TEI environment.

Writing System

The main difficulty with encoding Chinese texts lies in the writing system. Over thousands of years, the script used to write Chinese texts has evolved and has seen many changes in conventions, styles and character usage. The result is thus a rich and deep cultural heritage, which engraves in the writing system memories of a people that values history and memory in a way few others do, resulting in a writing system that contains an open ended, unknown number of distinct characters[6]. Since the beginning of the 20th century, there have been attempts at dealing with this problem from a practical side, by limiting the use of characters in daily life and thus making it possible for the first time to enable more than a tiny elite to acquire enough knowledge of the writing system to participate in a modern society based on the written word, be it application forms, contracts, newspapers or novels.

[6] The largest dictionary known to this writer is the *Zhonghua zihai*, which contains 85000 characters, but the difficulty here is not really the number of distinct characters, but the question what has to be seen as a character as opposed to a mere variant of another character. We will return to this question.

The last incarnation of the Unicode character set provides almost 75000 Chinese characters[7]. In this case also the definition of what has to be considered a separate character changed significantly during the process of defining these, which has been going on more than 20 years[8].

Although there are now assigned code-points for all characters in daily use and even most rare characters that appear in historical sources, there are still problems with the character encoding that are intrinsic to the way it is defined and evolved over the years of its development: unwanted unification and unwanted separation of characters[9].

- **Unwanted unification:** Especially in the early phase of the development, when there was only insufficient space set aside and processing memory limited, efforts were made to unify similarly looking character shapes into one code-point value. This makes it impossible to refer to just one of the character shapes as opposed to the other character shapes also defined with a given code-point in a universal way.[10]
- **Unwanted separation:** On the other hand, there are certain code-points that encode characters of a slightly shape separately; the most famous being 說 (U+8AAC) and 説 (U+8AAA); the character shapes in many fonts do indeed look identical for characters in this group, thus making it extremely difficult to consistently only using one of them and avoiding the unwanted other pairs.[11] Another reason for this is the 'code separation rule' which meant that characters already encoded separately in one of the character encodings that formed the source of Unicode, these had to be treated separately.
- **Inconsistencies, duplications, wrong assignments:**[12] do also exist, but these are not by design and much less disrupting.

While these are annoying problems when dealing with Unicode, it is clear that the advantages of using a universal encoding for all texts far outnumber the problems mentioned here. The strategy adopted here is thus not the development or use of a

[7] With the release of Unicode 6.1 the total count of CJK characters is 74617.

[8] Development of Unicode started with a document (*http://www.unicode.org/history/unicode88.pdf*) by Joe Becker of Xerox corporation, published in August 1988.

[9] It would be more precise to talk about glyphs here, but what it comes down to in digital text is code-points.

[10] In practice, this can be done by specifying one specific font to be used to represent a character. Modern font technology also allows fonts to contain several character shapes for one codepoint and allow a rendering program to select them as needed. There is however no standardized way to do so across applications.

[11] In practice, the only way to deal with this is to preprocess a document with a table that changes the unwanted member of such a pair into the desired one.

[12] See Kawabata (2006) for some examples.

different encoding system, but rather a strategy to deal with these problems within and on top of Unicode. This will be achieved through a character database and the definition of additional private characters where necessary.

The Process of Encoding a Character

It might be useful here to look a bit more carefully into what exactly happens in the process of encoding a character, that is transcribing a character from a source text to its digital equivalent. In an encoded character set, each character that has been assigned to a code-point can be seen as a kind of platonic, ideal character that stands for any number of real-world, existing character shapes (glyphs), as we see them on a text bearing surface. However, it is impossible to design such an encoded character set in a way that each platonic character is only represented once, since it is in many cases impossible to unambiguously assign one specific glyph shape to only one character, since it is not only the shape, but also meaning and sound that contribute to this assignment and all of these might be dependent on the specifics of area and era as additional conditionals. In the case of the Unicode/ISO 10646 character set, this has led to a development where more and more glyphs that had already been represented as members of the set of glyphs represented by a given character, are now also encoded separately. The result is thus that a given glyph can be logically represented in several sets.

In such a situation, the process of assigning a character code to a given glyph has to look for the set of glyphs that as a whole most closely resemble the given glyph, or, to put it differently, to look for the most specific representation of a given character. If that can not be found, there are in principle two choices:

- To add this glyph (G) to an existing set, encoded by an existing character code (C) and thus in fact extending the set to accommodate this new glyph.
- To add a new character code (N) to the system, with this glyph as the most representative of the set of glyphs represented by this character code.

The first option makes the assumption G has been recognized as in principle belonging to the set of glyphs represented by C, which assumes knowledge of G and of the set of allowed representatives for C. Since the set of allowed representatives for C is an open set, which is not defined exhaustively in the relevant standards, but only by giving a sample of such representatives, this decision has to be made case by case and can not be generalized[13]. The second option does not require any knowledge of the character beyond

[13] Text encoding is in this respect more of an art than an exact science in that many decisions depend on the encoder. This can and should be made less arbitrary than this sounds by recognizing this fact and define a policy as to what exactly should the set of represented glyphs be. The first step to this could be for example to use a specific reference font and define what kinds of deviation from the glyphs used in this font are allowable. Such

this glyph and is the only one available if nothing more is known about this character. The downside is of course that this new character is not integrated into the network of implicit knowledge that is already in the system, through system level character properties and/or a database. It would therefore be wise to provide also a way to add such information together with the character.

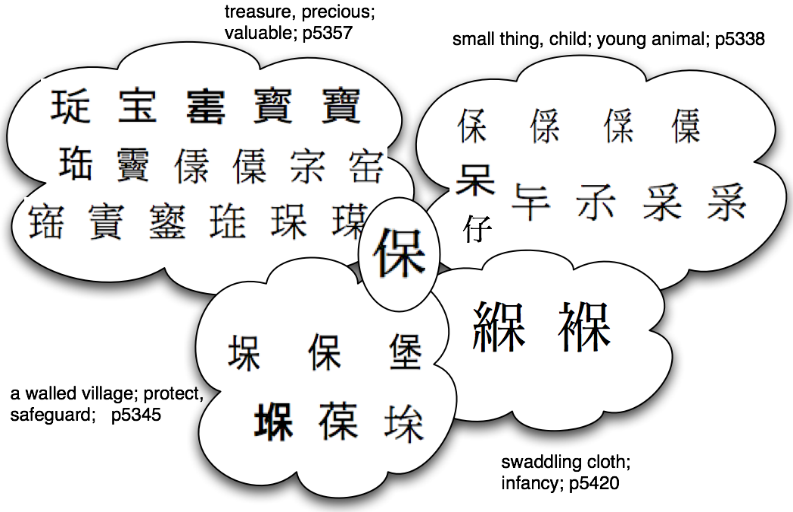


Figure 3: The semantic fields around the character 保 according to the HYDCD

Given this situation, information about the relationship between the characters in the character set has to be maintained. Different types of such relations have to be distinguished.

On the one hand, characters can be seen as mere variants of each other, serving essentially as a replacement for each other. More often, however, such a relationship covers only part of the semantic field of a given character, which makes it necessary to allow for a character to belong to different groups of variant characters, depending on which aspect of its meaning is called upon[14]. In other cases, the relationship might be due to a phonetic replacement or even error. Dictionaries and commentaries have for a long time collected such information, which has to be taken into account. This type of relationship could be called a generic relationship, which is true for all characters in this set, thus it is a relationship (to use a technical term) on the level of the class of characters, not the instances.

definitions should go into the project documentation.

[14] The historic dimension of the development of the writing system towards more specific characters is also playing a role here; what had been written with the same character in earlier texts might be delegated to different characters later on.

On the other hand, out of all the possible relationships that exist on a class level, or sometimes even in addition to these, for every instance of a character that is not identical with the character in modern usage, the corresponding modern character form needs to be established. While this might not seem necessary for a pure diplomatic transcription of a text, it is necessary to do proper searches and other text analytic tasks. Without this the value of a transcribed version is not much more than a digital facsimile.

Between these two types of relationships, the one completely generic and the other completely tied to the specific instance, it might well be useful to generalize from the instance-specific relationships to relationships that are relevant for the whole text, text corpus or text collection, thus forming a third type of relationship (of which could exist a number of sub-types depending on the scope).

A New Model for Encoding Chinese Primary Sources

In this paper, a new model is presented, together with a description of an implementation that acts on the model. The model again is described in two parts that are complementing each other, that is (1) a representation of the text and (2) a database of characters.

Representation of the Text

With respect to the character encoding, the main problem for premodern Chinese texts is that there is a friction between the modern usage, as reflected in the encoding systems available for digital texts, and the characters as they are used in a source text. In order to learn more about the writing system, and better understand the development of character forms and usages, one ideally should not have to rely on modern encoding systems for premodern texts, since they tend to hide exactly the differences that are the object of such a study, but if we are to transcribe the texts digitally, there is in fact hardly another way then to use such a modern encoding system. The only realistic way out is to give up on using character encoding as the only trace of the characters from the written source. This is however not easily achieved, since due to the way text encoding is done at the moment, the character encoding is a given, on which the layer of markup is built. Although there is some support, for example in TEI P5 to reach down into the encoding layer and introduce additional characters through markup, this mechanism is not flexible enough for cases, where the research questions involve investigation of the writing system itself.

The reason character encoding is performed is that this opens the way to computationally simply deal with the symbols encoded and abstract from the idiosyncrasics of the actual written characters. In alphabetical languages, this is very seldom problematic and even for logographic languages, this is only problematic where fundamental questions about the characters themselves need to be answered. On the other side, if character encoding does not provide the stable framework on which the following interpretative layers can be built, something else has to take its place.

The fundamental difference with respect to character encoding in the model proposed here is that first and foremost the location of a position in the text is recorded. Only in a second step is this position than associated with an encoded character that might provisionally serve to represent it.[15]

The model proposed here takes one representative edition of a text as a reference edition for digital encoding. This text is seen for the purpose of this model as a sequence of pages (or scrolls or other writing surfaces), which contain a sequence of lines, and the lines again containing a sequence of characters. While there is a provisional transcription into encoded characters, these encoded characters are considered to be preliminary and serve mainly as placeholders to mark slots for the positions in the text they fill. The characters used might be replaced by others or further annotated and linked to. The encoding is considered to be mainly positional (that is, identifying a character at a specific position in a text), rather than mainly symbolic (i.e. identifying the symbol that will be used for all such characters in this text).

In addition to the transcribed text of the reference edition, there are additional layers of text, that might contain characters as they are found on other witnesses of the text, or for example a regularized form that reflects modern usage. These layers are considered to be linked positionally through the sequential numbers of the pages, lines and characters (See Figure 5). The number of layers is unlimited, but for practical purposes they are assigned to different categories:

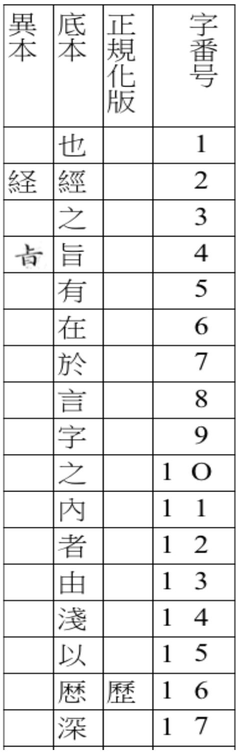
- The new edition to be created
- The reference edition
- Editions used for collation
- Other editions[16]

By convention any character position left empty will be filled by the character in the reference edition, which has to be present for all characters. In addition to these transcribed layers, a digital facsimile of the reference edition is linked to each page. If necessary, a cutout from this digital can be linked to the characters on this position, thus providing a connection between these two different representations of the text. The model also allows for the possibility of linking a digital facsimile of other editions (with possible different page arrangement) to the reference edition.[17]

[15] This idea is of course not new, it has been used implicitly in previous work, for example Yasuoka (2005).

[16] This category includes for example other electronic transcriptions of the text that are linked to the reference edition to improve the proofreading, but are not in themselves witnesses of the text.

[17] This can become rather complex and may in practice be difficult to realize if there are big differences in the arrangement of text in different sources.



異本	底本	正規化版	字番号
	也		1
経	經		2
	之		3
占	旨		4
	有		5
	在		6
	於		7
	言		8
	字		9
	之		1 O
	內		1 1
	者		1 2
	由		1 3
	淺		1 4
	以		1 5
	歴	歷	1 6
	深		1 7

Figure 4：Representation of the different editions

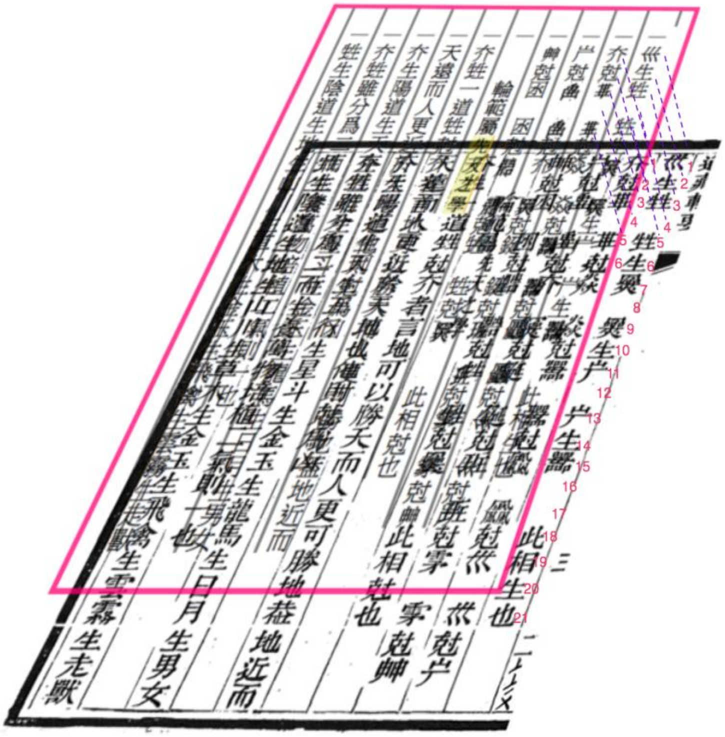


Figure 5: Attempt to visualize the connection between two layers.

The provisional encoding is by no means the only or final encoding that should be used, its main purpose is simply to occupy the position and show a representative that might stand for the character used at that position. Closer examination of this and other similar characters might bring up other possible candidates.

The transcription of the text is not seen just as a precondition for dealing computationally with the text, but is in itself a means to acquire better understanding of the writing system used to write the text and ultimately the content of the text. To gain an increasingly detailed understanding of the text, a kind of hermeneutical circle has to be performed, consisting of several steps to be performed in sequence.

- For every character that seems doubtful, unintelligible or a non-standard representation, the word intended by this character needs to be established.
- This can be done by
    - Looking at the context of the occurrence of this character and compare it with other, similar contexts
    - Looking at characters that are similar, either in visual, phonetical or semantic respects
- The result of this research gets registered into the database and thus provides context for future lookups.

- Information about context and registered variants becomes only available as the processing of the text progresses, therefore several loops of this activity have to be performed.
- Like a hermeneutical circle, this activity is in principle open ended and holds the potential for ever new discoveries and observations.

Through the performing of several loops of proofreading and digesting of different representations of characters, a new understanding of the text and the conventions and idiosyncrasy used to write it is gained.

Quite separate from these layers of textual representation there is an interpretative layer that might be thought to hover over the positional layer; in this layer connections or disconnections between similar or different characters are established and investigations of characters and their contexts is conducted.

Character Database

The model developed here relies on a database of characters. In this database, relations between characters, their occurrences within the text and among groups of characters are registered.

The groupings of the characters can be organized according to different properties of the characters, thus allowing the researcher to built sets of characters similar in its phonetic, semantic or visual properties. Since the relation to the occurrence of the character in the text is maintained, these relations are never thought to be abstract and generic, but are specific to the text under investigation.

Information in the database is held in two parts. One is holding generalized relations, as they are recorded in dictionaries, here the table of variant characters of the *Hanyu dacidian* 漢語大詞典 (HYDZD) and the *Dictionary of Variant Characters* compiled by the Taiwanese Ministry of Eductions are used, these are the most comprehensive tables of this kind. This serves as a backdrop for a specific database, which records the relations as they are observed in the text. This information is thus specific to the text it was developed with and the records of the database are always tied to the context the information was abstracted from. Nevertheless, as the number of texts processed with this system increases, and information held for these texts in the databases is aggregated, it is hoped that more general information on the Chinese writing system and its development can be gained, which are not available at the moment.[18]

[18] It should be noted here, that the development of encoded character sets by necessity predates the creation of textual material using these character sets. This precludes then of course any statistical base that might be used as a guidance in developing such encoded character sets. The results of work using systems such as the one developed here could serve as a guidance for the future development of such character sets.

The database connects the specific instance of the character, which is registered not with a character code, but with the location of the character within the text, with a generic identifier that is, an encoded representation of the character, if such a representation is available in the encoded character set. If no such representation is available, a private character will be created in order to allow computational processing and representation of this character. In such cases, structural information about the character, as well as an image cut from the digital facsimile is added to the record for this character.

If a suitable representation can be found within the almost 75000 character codes registered in Unicode, there might still be slight differences in appearance that can't be accounted for using the standard glyphs present in the operating system of the used computer. In such cases, and whenever a doubt about this character arises, an image cut from the facsimile representation of the text will be added to the record. The database can thus also be seen as connecting the digital facsimile representation and the transcribed representation of the text.

The Daozang Jiyao and its Editing Environment

The Daozang Jiyao

After the Daoist Canon of the Ming period (正統道藏 Zhengtong Daozang, 1445), the Daozang jiyao (Essentials of the Daoist Canon) is the most important collection of Daoist texts. It is by far the largest anthology of premodern Daoist texts and an indispensable source for research on Daoism in the Ming and Qing period (fourteenth to late nineteenth century). Although the collection is chiefly derived from the Ming Canon, it contains more than 100 texts that are not included there and thus is undoubtedly the most valuable collection of Daoist literature of the late imperial period. It features texts on neidan or inner alchemy, cosmology, philosophy, ritual, precepts, commentaries on Buddhist, Confucian and Daoist classics, hagiographic, topographic, epigraphic and literary works, and much else.

At the Institute for Research in Humanities in Kyoto, a research project on the DZJY is being conducted. This was started by the late Monica Esposito with the help of Mugitani Kunio and Christian Wittern, with the aim to investigate the origin of the collection, but also create a new critical electronic edition and develop the tools for exploring all aspects of its content[19].

[19] More on the history of the Daozang jiyao and the projects sponsored by the Chiang-ching Kuo Foundation (CCK) and the Japanese Society for the Promotion of Science (JSPS) can be found at *http://www.daozangjiyao.org*. Due to the untimely passing away of Dr. Monica Esposito in March 2011, the project has seen a reassessment and will be continued under the leadership of Lai Chi-tim and in close collaboration with the Centre of Daoist Studies at the Chinese University in Hong Kong.

The genesis of this collection is still hardly explored. According to the most common account, often presented even in recent articles and primarily based on Zhao Zongcheng (1995)'s hypothesis[20], it is believed that there are at least three different editions of the Daozang jiyao:

- by 彭定求 Peng Dingqiu (1645-1719) compiled around 1700 and containing 200 titles from the Ming Canon;
- by 蔣元庭 Jiang Yuanting (予蒲 Yupu, 1755-1819), who reportedly added 79 texts not contained in the Ming Canon (Weng Dujian, 1935) during the Jiaqing era (1796-1820);
- by 賀龍驤 He Longxiang and 彭瀚然 Peng Hanran published in 1906 at the 二仙菴 Erxian'an of Chengdu (Sichuan) under the name of Chongkan Daozang jiyao 重刊道藏輯要 (New Edition of the Essentials of the Daoist Canon), and (according to this hypothesis) containing a total of 319 titles.

However as early in 1955, 吉岡義豊 Yoshioka Yoshitoyo in his work entitled *Dōkyō kyōten shiron* 道教教典史論 (*Historical Studies on Daoist Scriptures*) cast doubt on the belief and affirmed that there were only two editions of the Daozang jiyao (number 2 and number 3).

One avenue that might provide new light in this controversy is the establishing of a stemma of existing textual witnesses. This should provide an answer to this question. However, a close reading and comparing of the existing witnesses is required, as well as a method to computationally compare these versions and calculate the respective closeness of individual witnesses.

Editing Environment

The editing environment has been realized as a Web application that can be used from any compatible browser, anywhere on the Internet. One of the reasons for choosing this platform was to be able to allow collaborative editing in a distributed environment, another was the hope to use this interface either directly, or at least most of it for a web-based publication of the texts.

Mapping to a Relational Database

A relational database management system (in this case, PostgreSQL 8.3) has been used to hold the data, while the user interface was developed with the Python-based web application framework Django (post 1.0 SVN version) and the Javascript framework ExtJS. In Django terms, there are two applications, 'textcoll' for holding the textual

[20] Zhao (1995).

content and 'chardb' for the character database; these two are glued together with a front-end called 'md'. One of the difficult tasks at the outset was to model the text collection, which has been done in the following tables[21]:

Tablename	Kind of Information	Relations
Work	Title of the work, date and other information	
Edition	Information about the edition, editor, publication details	Work
TextPage	Page number, graphical image of the page, serial number of the first character, number of characters	Edition, TextChar
TextLine	Line number, serial number of the first character, number of characters	TextPage, TextChar
TextChar	Serial number of character, associated extra information[22], Unicode value of the character, serial number of previous and next character	TextLine, Edition, TextChar, Interpunction

As can be seen, there is in principle a hierarchical relationship from the Work through Edition, TextPage and TextLine down to the TextChar table, which holds all the information related to the character at this position. It goes without saying that this incurs a tremendous overhead for the storage and processing of a simple text, but it should be kept in mind that this is the equivalent to a raster electron microscope, which tries to study the atomic units of a text, so there has to be some effort for isolating and handling these atomic units. There are some anomalies in the hierarchy, which are for the convenience of processing, which are that through the serial numbers of the first character on pages and lines the TextPage and TextLine tables are linked also to the TextChar table, which also has some internal links to the previous and following character position. Any character that spans more than one position in the grid, as well as talismans, outlines of movements in rituals or similar material that falls out of this simplistic model for the layout of a text is treated as a graphic outside of the textual flow.

In addition to these tables representing the text and allowing the modeling of its digital representation, there are a few other tables necessary for holding information about the text structure and content, as follows:

[21] Only tables and information relevant to this discussion are shown, implementation details are ignored to keep the table simple.

[22] Information about interpunction or other extra characters attached to this character is held here. This does include the possibility to add additional information, for example in the case of space characters that are used honorifically before names.

Tablename	**Kind of Information**	**Relations**
Attribute	key, value, note	TextChar (start), TextChar (end), Mark
Mark	tag, name, gloss, scope, note, color	
Interpunction	position, category	

The Mark table provides the tags that can be associated with locations in the text, whereas the Attribute table does provide the actual connection between an instance of a mark and a specific text location, given its start and end TextChar. Interpunction, except for space that is already present in the source text, is held in a separate table, linked to the text from the TextChar; besides the character used to represent the interpunction, the position relative to the character[23] and a category[24] is recorded.

Here is a table of the tables in the Chardb, the part of the application that maintains the character database:

Tablename	**Kind of Information**	**Relations**
Char	unicode codepoint, character, types	external link to TextChar
Unihan	key, value	Char
CharGroup	members, type	Char through Variant
Variant	type, character, note	Char, CharGroup
Pinyin	pinyin reading	Char
IDS	IDS (Ideographic Descriptor Sequence)[25]	Char

Groups of characters are built by linking the characters through the Variant table to a CharGroup and declaring thereby membership to that group. Additional properties can be set on Variant and CharGroup. The modeling of semantic is currently done through the definition in the Unihan table; the sound is modeled through the Pinyin table. This is provisional and is awaiting a more thorough solution.

User Interface

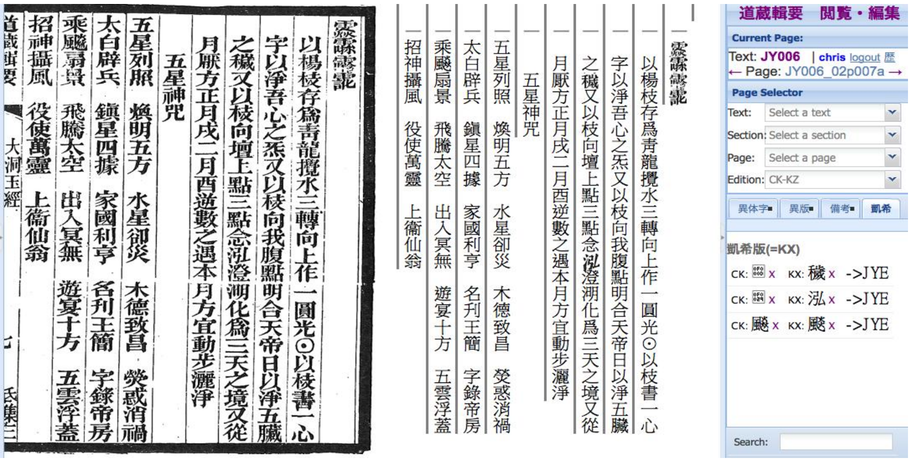
The user interface is accessed by opening the URL. It requires an account in the web application. Upon login, the user will be presented with the last page visited before leaving the system, like in Figure 6. The initially visible screen space is divided into three parts, at the right is a page as digital facsimile of the text, in the center pane is a transcribed version of this same page, while the left pane holds some administrative functions: There is information about the current page, the user (including a logout button and a possibility to

[23] This is given as one of eight compass positions with the character in question at the center, numbered clockwise and starting in the 'East', that is, after the character.

[24] At the moment, the categories are phrase-end, sentence-end and phrase/sentence-start

[25] The IDS is a sequence of operators and character parts that together describe how a character is composed.

look at a change log), in the second part is a panel for navigating the text collection and finally the bottom left has a multifunction panel for showing additional information and perform other tasks on this text page.



霦霦霧靇

以楊枝存爲青龍攬水三轉向上作一圓光⊙以枝書一心

字以淨吾心之炁又以枝向我腹點明合天帝日以淨五臟

之穢又以枝向壇上點三點念泓澄湖化爲三天之境又從

月厭方正月戌二月酉逆數之遇本月方宜動步灑淨

五星神咒

五星列照　煥明五方　水星卻災　木德致昌　熒惑消禍

太白辟兵　鎭星四據　家國利亨　名刋王簡　字錄帝房

乘飈扇景　飛騰太空　出入冥無　遊宴十方　五雲浮蓋

招神攝風　役使萬靈　上衞仙翁

道藏輯要　閲覧・編集

Current Page:

Text: JY006 | chris logout 歷

← Page: JY006_02p007a →

Page Selector

Text: Select a text

Section: Select a section

Page: Select a page

Edition: CK-KZ

凱希版(=KX)

CK: x KX: 穢 x ->JYE

CK: x KX: 泓 x ->JYE

CK: 飈 x KX: 飇 x ->JYE

Search:

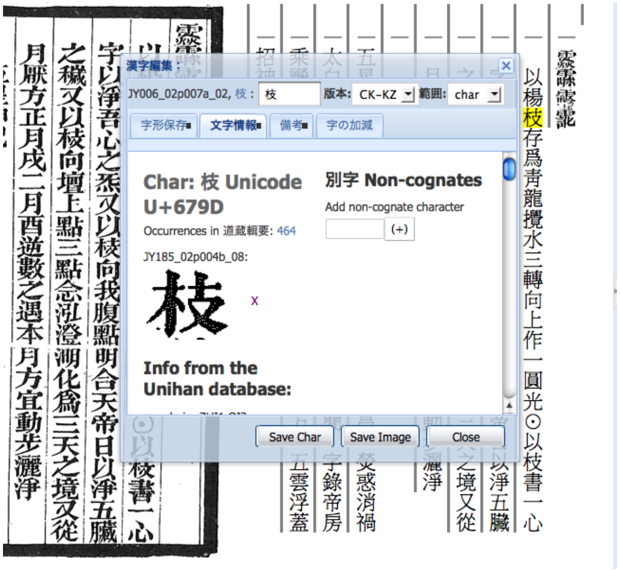
Figure 6: The web application interface for establishing the source text

The main functions for interacting with the text however are not visible here. Most editing actions are performed by clicking or selecting text and through the dialog boxes that pop up following such an action. Figure 7 shows an example of this popup window, in this case the fourth character position in the second line has been clicked, as a visual feedback to remember which character position is the target of the actions taken in this dialog, the character in this position is highlighted. The new window that opened gives in the top line the TextLine of this position, the character and then a number of input boxes. The first input box has the current character for the edition [CK-KZ][26]which is given in the second box. By providing a different character and selecting a different edition, the user can associate a new reading for another witness of the text, or give a different character to be used in the JYE edition. If the correction or replacement is occurring several times, the scope for this action can be set in the third selection box to be either valid for the current character, for the whole page, or even for the remaining part of the text[27]. Below this line, there are four tabs for further action or inspection; by default it

[26] The conventions for identifying the edition here is constructed as follows: Currently, there are two edition groups, indicated by CK and YP. The actual edition from within the group is then indicated in the second part of the sigle, in this case it is the Kaozheng reprint of the Chongkan edition CK-KZ. An exception to this scheme is the new regularized edition created here, which will be indicated as JYE.

[27] This is mainly to make the editorial process more efficient, under the assumption that only text not yet seen will be touched.

opens to the second tab, which provides a glimpse into the information in the character database for the character at this position. Among other things, the number of occurrences of the character here are given (464) and images of the character as it has been cut from the text. The main part gives additional information about the character, including pronunciation and definition according to the Unihan database[28]. More important however, for the present context, is the ability to maintain character relations here. The information about character variants, that is hold for the character 枝 is shown in Figure 8. In this case, the *Hanyu da cidian* 漢語大字典[29], on which the initial information is based, has assigned this character to five different groups of characters. For all characters in this group, the Unicode code-point, number of occurrences in the DZJY, as well as definition and pronunciation is given. Characters can be added to groups or deleted from groups, or new groups created as necessary, thus allowing to model this information exactly as is needed for this text collection. In addition to that, to assist the user in distinguishing characters that might be mistaken for each other, it is also possible to register characters to the system which are not cognates of the current character.



漢字編集：

JY006_02p007a_02, 枝：枝 版本: CK-KZ 範囲: char

Char: 枝 Unicode U+679D

Occurrences in 道藏輯要: 464

JY185_02p004b_08:

Info from the Unihan database:

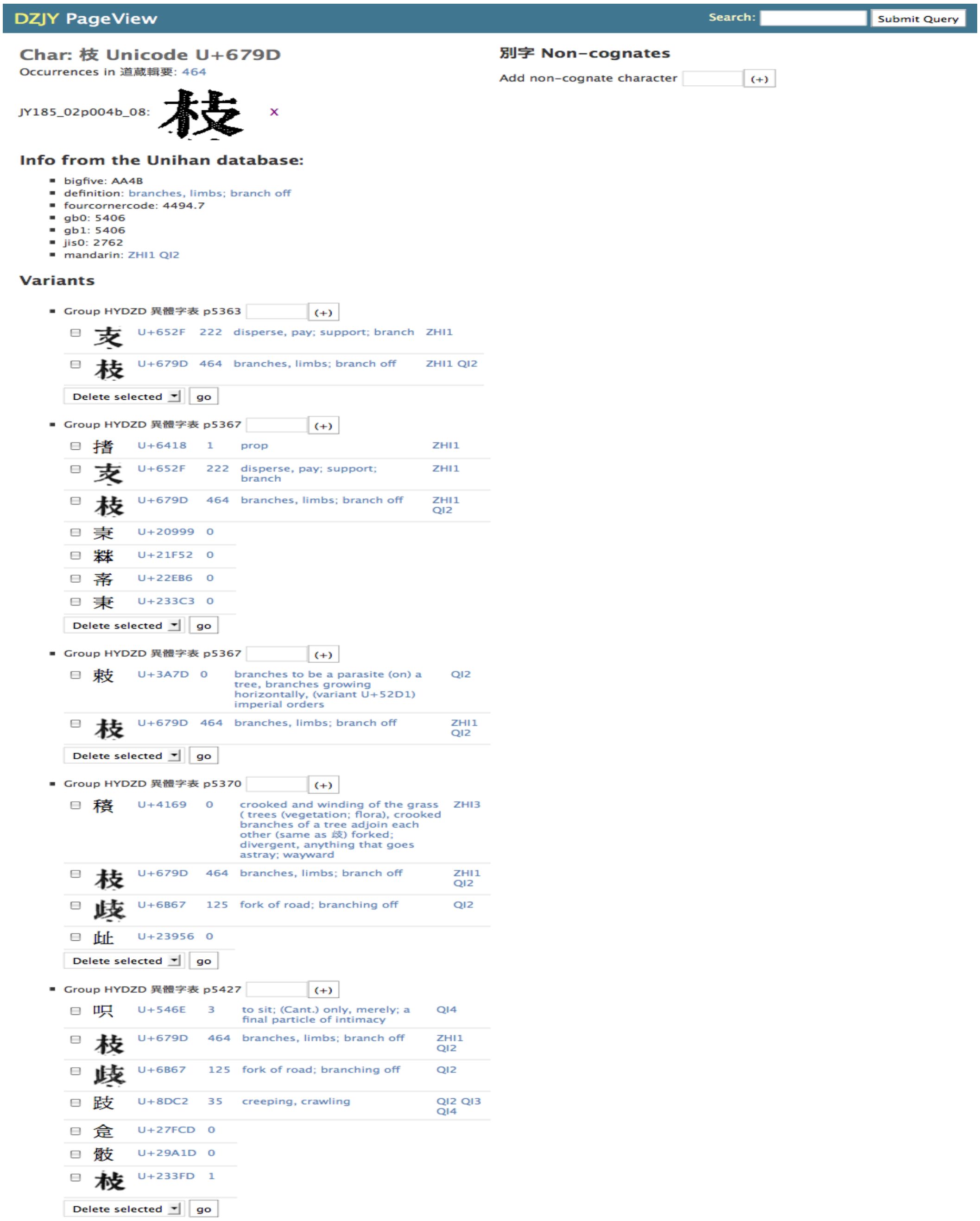
別字 Non-cognates

Add non-cognate character

Figure 7: The dialog box that opens when a character position in the transcribed text is clicked

[28] This is a database of basic character properties, maintained by the Unicode Consortium.

[29] Hanyu da zidian weiyuanhui (1986-1989).



Char: 枝 Unicode U+679D
Occurrences in 道藏輯要: 464

JY185_02p004b_08: 枝 ×

別字 Non-cognates

Add non-cognate character (+)

Info from the Unihan database:

- bigfive: AA4B
- definition: branches, limbs; branch off
- fourcornercode: 4494.7
- gb0: 5406
- gb1: 5406
- jis0: 2762
- mandarin: ZHI1 QI2

Variants

- Group HYDZD 異體字表 p5363 (+)

⊟	支	U+652F	222	disperse, pay; support; branch	ZHI1
⊟	枝	U+679D	464	branches, limbs; branch off	ZHI1 QI2

Delete selected go

- Group HYDZD 異體字表 p5367 (+)

⊟	搘	U+6418	1	prop	ZHI1
⊟	支	U+652F	222	disperse, pay; support; branch	ZHI1
⊟	枝	U+679D	464	branches, limbs; branch off	ZHI1 QI2
⊟	𠦙	U+20999	0		
⊟	𡽒	U+21F52	0		
⊟	𢺶	U+22EB6	0		
⊟	𣏃	U+233C3	0		

Delete selected go

- Group HYDZD 異體字表 p5367 (+)

⊟	㩽	U+3A7D	0	branches to be a parasite (on) a tree, branches growing horizontally, (variant U+52D1) imperial orders	QI2
⊟	枝	U+679D	464	branches, limbs; branch off	ZHI1 QI2

Delete selected go

- Group HYDZD 異體字表 p5370 (+)

⊟	䅩	U+4169	0	crooked and winding of the grass (trees (vegetation; flora), crooked branches of a tree adjoin each other (same as 歧) forked; divergent, anything that goes astray; wayward	ZHI3
⊟	枝	U+679D	464	branches, limbs; branch off	ZHI1 QI2
⊟	歧	U+6B67	125	fork of road; branching off	QI2
⊟	𣥖	U+23956	0		

Delete selected go

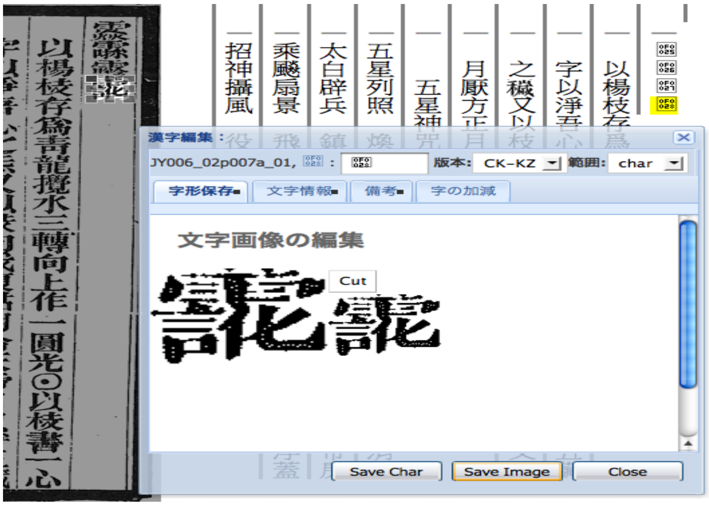
- Group HYDZD 異體字表 p5427 (+)

⊟	呮	U+546E	3	to sit; (Cant.) only, merely; a final particle of intimacy	QI4
⊟	枝	U+679D	464	branches, limbs; branch off	ZHI1 QI2
⊟	歧	U+6B67	125	fork of road; branching off	QI2
⊟	跂	U+8DC2	35	creeping, crawling	QI2 QI3 QI4
⊟	𧿍	U+27FCD	0		
⊟	𩨝	U+29A1D	0		
⊟	𣏽	U+233FD	1		

Delete selected go

Figure 8: Information about 枝 held in the character database

The first tab on this window allows the user to cut an image from the digital facsimile and associate it with the current position in the transcribed text. In addition, this image is also associated with the corresponding character in the character database.



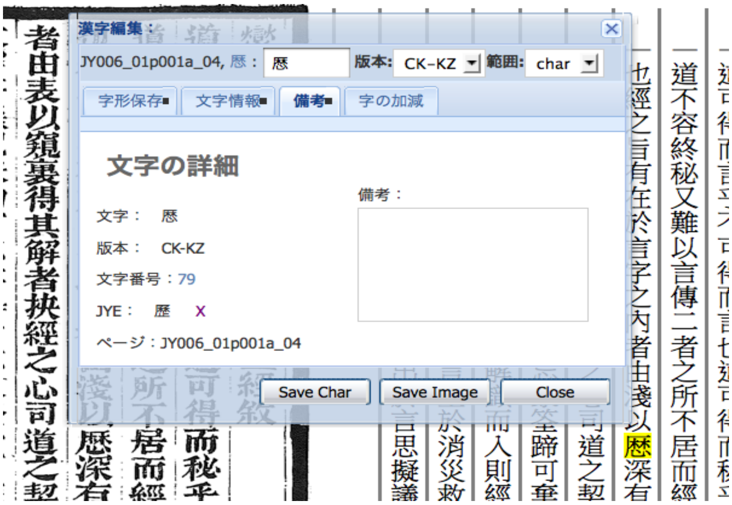
漢字編集：

JY006_02p007a_01, :

版本: CK-KZ 範囲: char

Figure 9: Cutting a character from the text

The next tab on this window allow the user to see all information associated with a character, as shown in Figure 10. Here, a regularized version of the character has been registered for the JYE edition. It is also possible to add further notes to the character into the textbox to the right. The last tab (not shown), allows for adding or deleting of larger chunks of text.



漢字編集：

JY006_01p001a_04, 歴 : 歴

版本: CK-KZ 範囲: char

文字の詳細

文字： 歴

版本： CK-KZ

文字番号：79

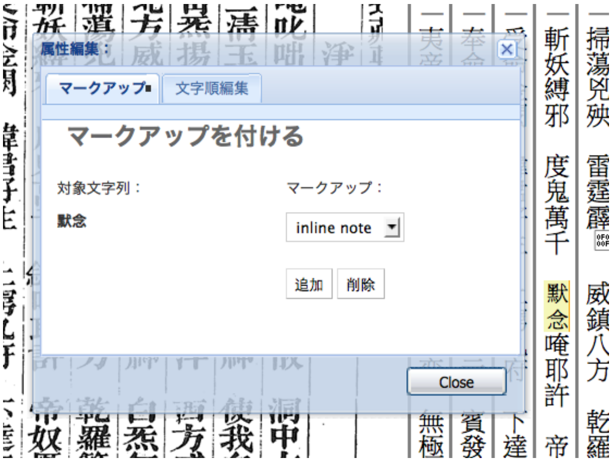
JYE： 歴 X

ページ：JY006_01p001a_04

備考：

Figure 10: Detailed information about this text location

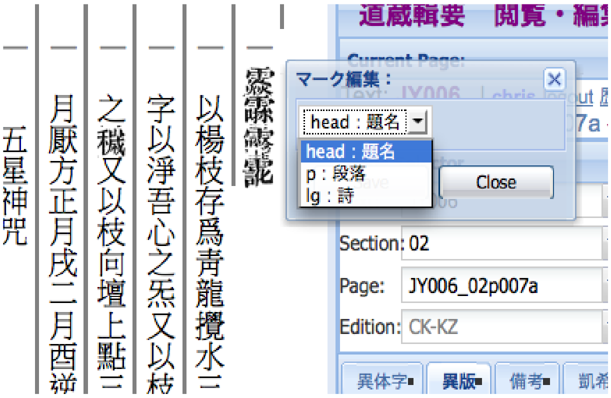
Another way to interact with the text is to select a string of characters. The action following a selection can be configured to either copy the selected string to the search box, or to apply markup to the selection, as shown in Figure 11. Currently, this is mostly used to record characters that have been printed smaller as inline notes, but this will also be used for titles, personal names and other items of interest in the text. To record structural elements in the text, like paragraphs, verse lines or section headings, yet another dialog can be used that pops up when clicked on the horizontal bars at the top of a text line (see Figure 12); this assumes however that the features is starting at the beginning of the line.



対象文字列：黙念

マークアップ：inline note

Figure 11



head：題名

Section: 02

Page: JY006_02p007a

Edition: CK-KZ

Figure 12: Applying markup to a line

Context

The discussion here stands in the context of practical experience and theoretical considerations with digital text in Chinese. Some ideas have been pursued and have been discussed in earlier presentations and articles. In particular, in the last several years, I was developing an ontological model[30] for understanding text from a perspective quite different from the one taken here. The model presented here is meant to complement this from a different perspective, filling some of the gaps in the earlier model.

The work here can also be seen as a continuation of an earlier line of thought, which was concerned with a 'scholarly workbench'; the last incarnation of which was a Filemaker-based application called KanDoku that supported annotation, translation and markup of digital texts. When I tried to implement support for more flexible handling of character representation and variant readings for different text witnesses, I quickly ran into the limitations inherent in that platform. The present work should be seen as aiming in a similar direction, except that this time and attempt has been made to start with a firm foundation. It is planned however, to gradually add more of the possibilities of that earlier KanDoku. Another difference of the present work, to KanDoku is that the latter took as its input a completed TEI P5 compatible digital version of a text, while the former will attempt to produce such a thing as its output (among other things), in fact one of its design goals is to improve the workflow of creating high quality digital edition of text, but hopefully its usefulness will extend beyond that and allow the user to gain new insights into the text itself.

In the Daozang jiyao project, the work was initially done by editing TEI conformant XML files with the XML editor oXygen. This was considered cumbersome and time consuming by the researchers involved, so this editing application has been developed to provide a more convenient interface for performing specific tasks on the text easier than could be done otherwise. It should be noted however, that such a specialization also involves an enormeous limitation to what can be done while editing the text, there will therefore be many cases where such a solution can not be applied. It is planned to add a routine to export the texts edited using this interface into TEI conformant XML documents.

As it stands at the moment it is very much work in progress and much of the necessary functionality, for example to visualize textual context in a way that takes into account the several different layers of characters that might be available at a given point in the text is still missing. The results that have been achieved so far in the context of

[30] In English, this is presented most detailed in Wittern (2007), but more references can be found here *http://kanji.zinbun.kyoto-u.ac.jp/~wittern/publications articles/index.html*.

work on the Daozang jiyao seem to suggest that the work is going in the right direction and will indeed be able to open up new avenues for digital texts.

It will be interesting to see how well this approach could also be applied to earlier stages of the development of the Chinese writing system, such as bronze inscriptions or texts on bamboo slips. The model presented here implicitly assumes a regular grid for the layout of a text, so that model would require some extension, but it will have to be actually tried with such a text to see in what way such extensions should be implemented.

Abbreviations

CBETA	Chinese Buddhist Electronic Text Association 中華電子佛典協會 (see *http://www.cbeta.org*)
CJK	Chinese, Japanese and Korean Characters
CK	Chongkan 重刊 (reprint) edition of the DZJY, Sichuan 1906ff.
CK-KZ	Facsimile edtion of CK published by the Kaozheng publishing company
DZJY	Daozang Jiyao 道臧輯要
HYDCD	*Hanyu Dacidian* 漢語大詞典
IDS	Ideographic Definition Sequences
JYE	New Electronic edition of the DZJY
TEI	Text Encoding Initiative (see *http://www.tei-c.org*)
UCS	Universal Character Set, also known as Unicode (see *http://www.unicode.org*)
XML	eXtensible Markup Language (see *http://www.w3.org/XML/*)
YP	DZJY original edition by Jiang Yupu 蔣予蒲 (1755-1819)

References

Becker, Joe. 1988. *Unicode 88*. (*http://www.unicode.org/history/unicode88.pdf*. Accessed 2012-03-23).

Hanyu da zidian weiyuanhui 漢語大字典委員會, eds. 1986-1989. .*Hanyu da Zidian* 漢語大字典. 8 vols. Wuhan: Hubei cishu chubanshe and Sichuan cishu chubanshe.

Kawabata, Taichi 川幡太一. 2005. *Possible Multiple-encoded Ideographs in the UCS*. (*http://www.cse.cuhk.edu.hk/~irg/irg/irg25/IRGN1155_Possible_Duplicates.pdf*. Accessed 2012-03-23).

Kawabata, Taichi 川幡太一. 2006. *IDS による UCS 漢字の 「同一性」の判定手法 (Methods to Assert 'Sameness' of a Character in UCS Kanji Through IDS.) 東洋学へのコンヒュ゚ータ利用第17回研究セミナー*. Kyoto: Institute of Research in Humanities.

Leng, Yulong 冷玉龍 and Wei, Yixin 韋一心, eds. 1994. *Zhonghua Zihai* 中華字海. Beijing: Zhongua shuju.

Luo, Zhufeng 羅竹風, ed. 1987-1994. *Hanyu Dacidian* 漢語大詞典. Shanghai: Dictionary Publishing House.

Morioka, Tomohiko. 守岡知彦 The CHISE (Character Information Service Environment) project. (*http://www.kanji.zinbun.kyoto-u.ac.jp/projects/chise/*. Accessed 2012-03-23).

Renear, Alan; Mylonas, Elli; Durand, David. 1996. Refining Our Notion of What Text Really is: the Problem of Overlapping Hierarchies. *Research in Humanities Computing*. Ed. Ide, Nancy and Hockey, Susan. Oxford: Oxford University Press.

Wittern, Christian. 2007. Digital Text, Meaning and the World: Preliminary Considerations for a Knowledgebase of Oriental Studies. *Higashi Ajia ni Okeru Reigi to Keibatsu* 東アシ゛アにおけろ儀礼と刑罰 (*Ritual and Punishment in East Asia*). Ed. Tomiya, Itaru 冨谷至. Kyoto: Institute for Research in Humanities. 41-58.

Yasuoka, Koichi 安岡孝一. *Text-Searchable Image and Its Applications* (*http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/publications/2005-01-22.pdf*. Accessed 2012-03-23).

Yoshioka, Yoshitoyo 吉岡義豊. 1955. *Dōkyō Kyōten Shiron* 道教教典史論 (*Historical Studies on Daoist Scriptures*) Tokyo: Gogatsu shobo.

Zhao, Zongcheng 趙宗誠. 1995. Daozang Jiyao de Bianzuan yu Zengbu 道藏輯要的編纂与增補 (The Compilation of the Daozang Jiyao and its Enlarged Editions) *Sichuan Wenwu* 四川文物 2:27-31.