Preserving the Original Layout of Ancient Chinese Texts Using HTML5 : Using Shuowen Jiezi as an example

Author: Yap, Cheah Shen (Ksana Forge) yapcheahshen@gmail.com
Translated by Ian McIlwaine

Keywords
Shuowen Jiezi, HTML5 Canvas, Vertical Page Layout, Amateur Proof-reading

Abstract

We have developed a software platform for converting the Chinese Classics into a digital format which the original layout is preserved. This method of conversion allows those that are not experts in classical Chinese literature, even those that do not understand Chinese, to partake in the proof-reading process.

Since this kind of application involves complicated manipulation of 2D graphics, in the past, we have no choice but writing a native desktop program, with the introduction of CANVAS tag to the HTML 5 standards, we are able to move our code from desktop to browser without any plug-in or JAVA applet. The cost of cross-platform development and maintenance is greatly reduced.

In order to clarify the various elements involved in the conversion and proof-reading process, we will use Duan Yu Cai's compilation of the Shuowen Jiezi (ref 1) as an example throughout this paper.

Background

Due to a lack of financial and political incentive, current information systems have been developed without consideration towards classical Chinese texts. For this reason, converting these texts into a digital format is very difficult, and experts across the field of Chinese literature continue to lack the necessary technology to complete such conversions.

As the market for this type of digitization is very small and specialized, the probability that large software companies will invest resources to address the problem in the future is also quite low. Without effective and affordable tools, the cost of compiling a digital archive will remain high. Currently, scholars of Sinology have to pay a high price to gain the authorized usage of such databases, or tolerate inferior free databases lacking in functionality.

System Specifications

Using Accelon as a foundation, we began designing the platform in 2008. With great paradigmatic changes in the digital world over the past two years and the rise of the post PC era, mainly in the form of Internet mobile devices and cloud computing, the structure of this platform has undergone several changes, as well as a large rewriting of source code.

The system maintains an open source format using HTML5, XML and JSON (Java Script Object Notation), and contains no proprietary data formats. The database data producer is designed mainly for PCs, while the target data consumer is tablet devices and e-book readers.

Cloud computing is also incorporated, and allows amateurs and enthusiasts to participate in database construction and improvement. Differing from the Wikipedia Model, we are able to limit database editors to a small number of experts, while the majority of users have access to the functions of tagging quotes, adding annotations, and finding typos.

Finally, to aid user editing and review, we intentionally maintain the original page layout of the ancient texts during the digitization process. Although this decision has increased the difficulty of development, it has also led to some unexpected benefits that will be discussed later on.

As the number of topics involved in the processing of ancient Chinese texts is very large, this paper will focus on how to use HTML5 to handle the format of texts ancient texts.

The Layout of Ancient Chinese Texts

Before the introduction of modern publishing, the majority of Chinese books were printed using woodblock, and some using moveable type printing methods. In these books, vertical text with double interlinear annotations (ref 2) was the main visual format. Due to the immense size, there is no way to make a precise calculation of the total number of Chinese characters contained in paper copies of Classical Chinese texts. For example, the Imperial Siku Quanshu collection alone contains over 3,000 titles, each averaging 200,000 characters in length, for a total of over 600 million characters. Currently, there are more than 150,000 known titles with an estimated total of over 30 billion characters. This is the product of tens of thousands of writers spanning thousands of years, and is also an important relic of human history and culture.

Prior to the introduction of HTML5, these texts were very difficult to display digitally in their original layout. The most common solution was to use nested HTML tables to imitate the original layout, but this required a lot of labor-intensive work. The HTML tagging required to display the texts in their original format would be 2~3 times as large as the texts themselves. Given the size of the collections, this would be very impractical. Moreover, these tables were not compatible with search engines, as the order of Chinese vertical texts read from right to left, but search engines would read the HTML tables from left to right. Vertical text plugins like Taketori (ref 3) existed, but could only be viewed properly in Firefox. (Chrome and IE would exhibit character alignment problems).

For this reason, we decided to use the new CANVAS element provided on HTML5 to rewrite a new text display component that could not only adequately support vertical script, but could also display interlinear annotations and ruby punctuation, which are symbols added to the right of characters by the reader, not the author.

Layout Rules for Chinese Vertical Texts

CSS does not define instructions for vertical text with double interlinear annotations. For this reason, we had to develop a layout engine that could automatically handle line wrapping and compact spacing. Figure 1 contains three parts (X1,Y1, Z1), each with the ‘main text’, shown as large squares, and ‘interlinear annotations’, shown as small squares. The order of the characters reads from top to bottom, and from right to left.

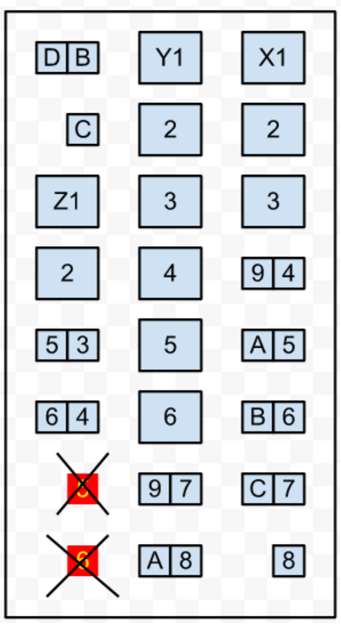


Figure 1: the flow of vertical text with interlinear annotaion

In the example Z1 above, it is important to notice the compact spacing of the annotations, and how squares 5 and 6 have been moved to the left hand side. This is done to mimic the spacing rules used by the original authors, thereby maintaining the original layout in the digitized version of texts. Other than adding XML tags to delineate the main text from annotations, the rendering process is completely automatic, and saves a large amount of work relative to using nested HTML tables.

Implementation of the Layout Engine

The source code of layout engine consists of some 2000 lines of javascript code, it is available at http://dev.predragon.org/cgi-bin/ksanaedit.cgi/dir?ci=tip **,** the source code is too long to go line by line, let me explain the key work-flow and concepts. The engine is made up of three parts, namely typesetter, layouter and ksanaview .

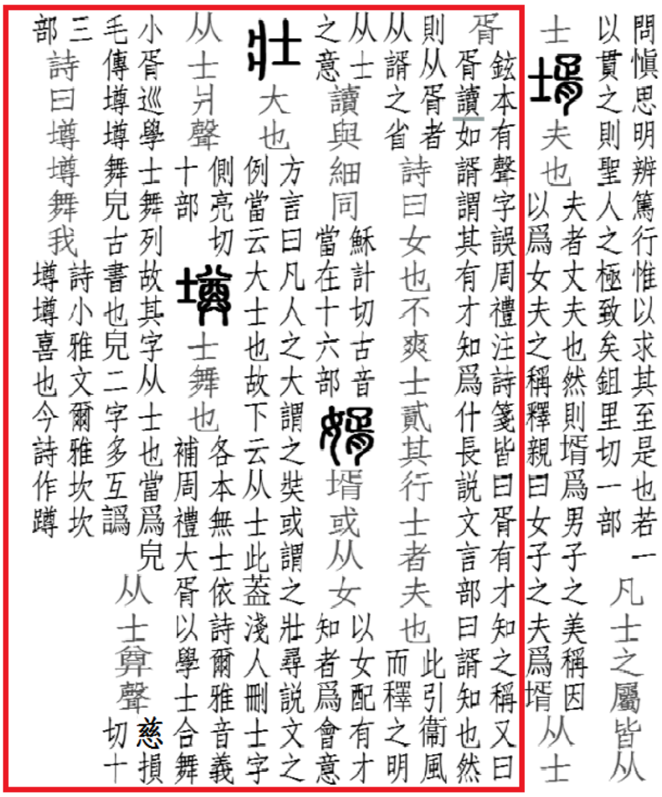
Typesetter takes XML as input, splits into an array of tokens. There are 2 types of tokens, text token and tag token, tag tokens are not visible, they serve as page layout instructions to adjust the font size, thus the minimum bounding rectangle of each text token can be decided.

Layouter tries to fit the tokens into a page, from top to bottom, right to left, for normal text, the text-wrapping rule is same as normal text editor, for the interlinear annotation, when a character hit the bottom of the page, layouter will first check if it is right side of annotation (e.g. X4~X8), if so, the next line of annotation will start from the bottom of previous normal characters, not a new line from the top of the page.

Ksanaview takes the output of layouter and renders the tokens on HTML5 canvas, draw a caret, and handles mouse and keyboard input. Typesetter and layouter is platform independent, which means by modifying only Ksanaview, it is possible to generate SVG format or fallback to a less feature-rich “DIV+CSS absolute positioning” solution when HTML5 canvas is not available.

In contrast with our similar engine written in Object Pascal, the Javascript version is about 2 times fewer lines of code, and it is much easier to maintain, extend and deploy. we think the dynamic types (the JSON object) is ideal for this kind of job.

Amateur Proof-Reading


問愼思明辨篤行惟以求其至是也若一以貫之則聖人之極致矣鉏里切一部
凡士之屬皆从士
壻夫也
夫者丈夫也然則壻爲男子之美稱因以爲女夫之稱釋親曰女子子之夫爲壻
从士胥
鉉本有聲字誤周禮注詩箋皆曰胥有才知之稱又曰胥讀如諝謂其有才知爲什長説文言部曰諝知也然則从胥者从諝之省
詩曰女也不爽士貳其行士者夫也
此引衛風而釋之明从士之意
讀與細同
穌計切古音當在十六部
婿壻或从女
以女配有才知者爲會意
壯大也
方言曰凡人之大謂之奘或謂之壯尋説文之例當云大士也故下云从士此葢淺人刪士字
从士爿聲
側亮切十部
墫士舞也
各本無士依詩爾雅音義補周禮大胥以學士合舞小胥巡學士舞列故其字从士也當爲兒毛傳墫墫舞兒古書也兒二字多互譌
从士尊聲
慈損切十三部
詩曰墫墫舞我
詩小雅文爾雅坎坎墫墫喜也今詩作蹲

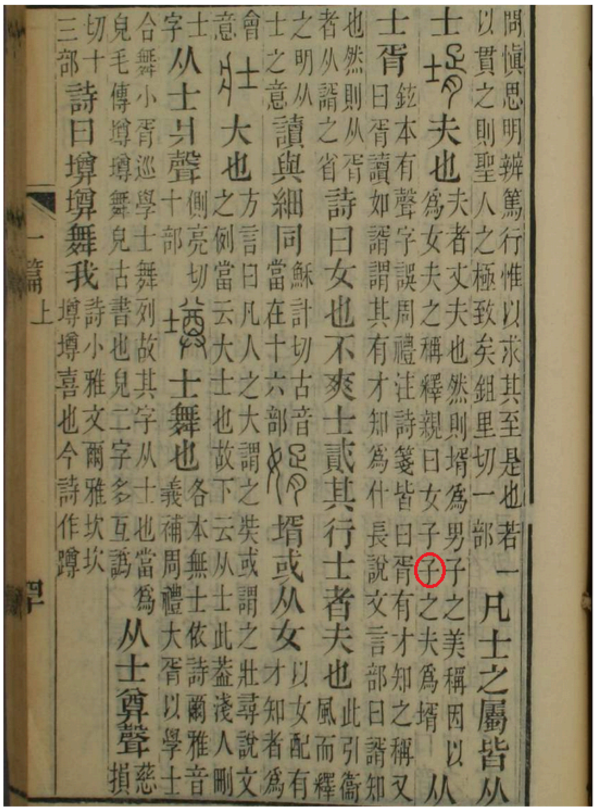

問愼思明辨篤行惟以求其至是也若一以貫之則聖人之極致矣鉏里切一部
凡士之屬皆从士
壻夫也
夫者丈夫也然則壻爲男子之美稱因以爲女夫之稱釋親曰女子子之夫爲壻
从士胥
鉉本有聲字誤周禮注詩箋皆曰胥有才知之稱又曰胥讀如諝謂其有才知爲什長説文言部曰諝知也然則从胥者从諝之省
詩曰女也不爽士貳其行士者夫也
此引衛風而釋之明从士之意
讀與細同
穌計切古音當在十六部
婿壻或从女
以女配有才知者爲會意
壯大也
方言曰凡人之大謂之奘或謂之壯尋説文之例當云大士也故下云从士此葢淺人刪士字
从士爿聲
側亮切十部
墫士舞也
各本無士依詩爾雅音義補周禮大胥以學士合舞小胥巡學士舞列故其字从士也當爲兒毛傳墫墫舞兒古書也兒二字多互譌
从士尊聲
慈損切十三部
詩曰墫墫舞我
詩小雅文爾雅坎坎墫墫喜也今詩作蹲

Figure 2: comparison of HTML5 version and original manuscript
Left: HTML5 version Right: The original scanned manuscript

The proof-reading of classic Chinese literature is very troublesome. The source texts we are using, raw data from Kanjibase’s Shuowen Jiezi website, are well elaborated electronic files with a very low error rate. The Japanese are well known for their meticulous attention to detail, and it is apparent that a lot of effort was put into the development of these files. Despite this, after realigning the text into a vertical layout, our intern student very easily pointed out several flaws within the Japanese source text files. When comparing the scan below to the original text, it is obvious that starting from the fifth line, the layout differs from the original text. One can deduce that a problem must have occurred on the previous line, and as you can see, it is on the fourth line that the character ‘子’ has been omitted and「女子子之夫為壻」is mistakenly represented as「女子之夫為壻」. Aside from missing characters, extra characters added through the process of digitization will also become apparent through changes in layout.

Adding Reader Punctuation to Ancient Texts

Unlike modern literature, punctuation in ancient Chinese texts is not added by the author. Instead, it is a kind of contextual markup added by readers while studying an ancient work, and is largely subjective. Readers place punctuation marks (equivalent to commas and periods in English) to break a text down into smaller fragments, and through this process, their understanding and interpretation of the text is revealed.

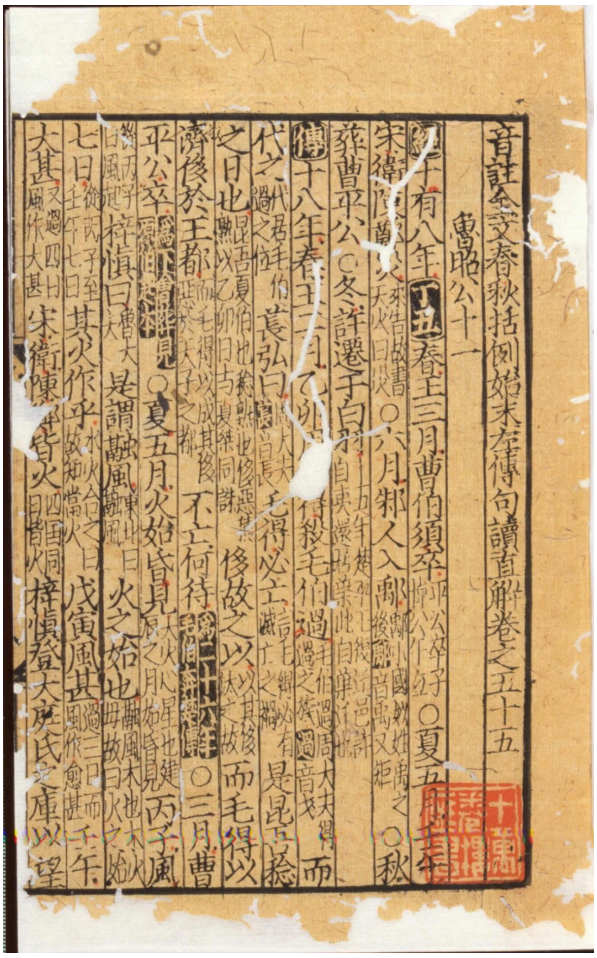
A large amount of value-added content can be created by collecting statistics on the way a large group of readers punctuate the same texts in unique ways. For one thing, punctuation added by experts can easily be shared, compared and used as a reference by others. Moreover, statistical analysis could be used to pinpoint confusing areas within texts where students are more likely to face problems in interpretation, as many students would punctuate those areas incorrectly.

To provide a platform to add and store punctuation data without changing the main text, we created separate layers for the ‘main text’ and ‘punctuation’. This was achieved by creating independent and overlapping HTML5 Canvas tags. Similar to the idea of an overhead transparency, the ‘punctuation’ layer lies on top of the ‘main text’ layer, and punctuation can be added by many different users without affecting the ‘main text’.

On the punctuation layer, there are two sources of input. One comes from current readers of the texts (mainly university literature students) directly adding punctuation to electronic texts with a mouse. Based on this punctuation, professors are able to evaluate a student’s level of understanding with respect to the content in ancient texts, a method which we have already prototyped successfully.

The second method utilizes visual pattern recognition techniques to locate punctuation in ancient paper texts (often using ink made from red cinnabar), and convert them to electronic copies. As characters in the electronic copy of a text are displayed at positions consistent with the transcripts' original scanned image, this punctuation can be mapped to a coordinate in the electronic text, and then associated with the sequential position of the character at that position. Through the character's sequential position, the character can be determined. The result is a completely automatized method of punctuation input. This is an area in which we hope to expand upon in the future, and is also another reason why it is important to maintain a text’s original format in its digital conversion.

Figure 3 is a precious printed version of a classic scripture from the thirteenth century Yuan dynasty. It contains obvious red punctuation.



音註全文春秋括例始末左傳句讀直解卷之五十五

魯昭公十一

經 十有八年 丁丑 春王三月曹伯須卒 平公卒子悼公午立 ○夏五月壬午

宋衛陳鄭災 天火曰災 來告故書 ○六月邾人入鄅 鄅小國妘姓 [illegible] 鄅音禹又矩 ○秋

葬曹平公 ○冬許遷于白羽 [illegible] 十五年楚遷許 [illegible] 自夷還 [illegible] 此自葉遷于析 [illegible]

傳 十八年春王二月乙卯周毛得殺毛伯過而 [illegible] 周大夫 得

代之 代君毛伯過之位 萇弘曰 周大夫 萇音長 毛得必亡 言毛得必有滅亡之禍 是昆吾稔

之日也 昆吾夏伯也 稔熟也 侈惡積熟 以乙卯日亡 與桀同誅 侈故之以 以其侈 [illegible] 之故 而毛得以

濟侈於王都 而毛得以成其侈 [illegible] 於天子之都 不亡何待 為二十六年毛伯奔楚傳 ○三月曹

平公卒 為下會葬見 [illegible] ○夏五月火始昏見 大火心星也 建辰之月始昏見 丙子風

[illegible] 丙子 [illegible] 梓慎曰 魯大夫 是謂融風 東北曰融風 火之始也 融風木也 木火母 故曰火之始

七日 從丙子至壬午七日 其火作乎 水火合之日 故知當火 戊寅風甚 過三日而 [illegible] 壬午

大甚 又過四日 風作大甚 宋衛陳鄭皆火 四國同日皆火 梓慎登大庭氏之庫以望

十萬 [illegible]

Figure 3: Manuscript of Zuozuan (左傳)

Conclusion

The CBETA (the Chinese Buddhist Electronic Text Association) provides a strong illustration of the impact free open-source data and technology can have on a field of study, as well as society. Using CBETA's free database, Buddhist scholars have saved a large amount of time once wasted on data retrieval and the repeated entry of text. Accordingly, the quantity and quality of papers published by Buddhist scholars has shown a significant increase. We believe that the same results will soon be realized in the field of Sinology.

We are especially excited about the results of using the vertical text layout to aid in the proof-reading of ancient texts. From this, the monotonous job of proof reading becomes interesting and, for the most part, requires no special training, allowing average volunteers to undertake the job. Chinese experts can save large amounts of time and energy, giving them more time to focus on more productive work.

As with all digital databases, creation and maintenance are very difficult tasks, but once finished, from a technical point of view, the cost of owning and sharing the database between organizations is next to nothing. For this reason, the Classics digital database is similar in nature to public infrastructure. There is a Chinese proverb that states “the ancestor plants the tree, and the successor enjoys the shade.” Indeed, the tools we have developed are the water, the sunlight and the fertilizer, but a strong sense of social welfare is still required for the first seed to be planted.

The digitization of the Chinese classics will not bring electoral votes, so the government is unlikely to become deeply involved. It does not pull in great revenues, so commercial interest is lacking. As a non-profit organization restricted by conditions of insufficient manpower and resources, we maintain the attitude that we must not let the wisdom of the past slip through the cracks of this modern society. For this reason, we have undertaken this duty, which is rich in historical meaning, with a sense of obligation, and have endeavored to create a high quality platform. We hope that through these actions we can maintain and promote these collections in the digital world, thereby allowing the wisdom contained in the Classics to be passed on to future generations.

Floating along the river of time, the classics finally have an opportunity for liberation from the bonds of their material form, morphing into countless incarnations of energy, limited by neither time nor space. An 8GB USB disk costs less than £5 and can hold all the wisdom contained within the Chinese Classics. Perhaps this depicts the ancient Buddhist saying "The world can be contained in a mustard seed."

Acknowledgements

1. The original scanned manuscript is provided by Library of Waseda University (早稲田大学) , Japan. http://archive.wul.waseda.ac.jp/kosho/ho04/ho04_00026/ho04_00026_0001/
2. The Shuowen Jiezi Chu Data (説文解字注データ) http://kanji-database.sourceforge.net/dict/swjz/
3. The CDPHanzi Database(漢字構形資料庫) provided byTaiwan Acedemia Sinica (台灣中央研究院) Institute of Information Science (資訊科學研究所)Chinese Document Processing (文獻處理實驗室) http://cdp.sinica.edu.tw/

References

1.Introduction of Shuowen Jiezi:
http://www.ihp.sinica.edu.tw/~asiamajor/pdf/2008a/12%20Bottero%20v21.pdf
2.Vertical Layout http://unicode.org/notes/tn22/RobustVerticalLayout.pdf
3.Taketori http://taketori.org/js.html