使用HTML5處理古書版式

以說文解字注為例

https://shutonggui.cn
Yap, Cheah Shen yapcheahshen@gmail.com
wechat : Sukhanika

關鍵字

說文解字, HTML5 Canvas, 中文直書, 業餘校對,自動標點輸入

前言

我們發展了一個古籍數位化的方法，讓一般人(即使不懂中文者)也可以參與校對的工作，本方法有效地降低校對的時間成本和專業人力的需求。
本文以"清．段玉裁"的《說文解字注》為例，說明這個方法。

背景

由於目前資訊系統的設計對漢文古籍的考慮不周全，古籍的數位化的技術難度很大，漢文研究領域的工作者，普遍缺乏足夠的古籍數位化技術；而擁有技術實力的軟體公司，由於數位古籍的市場規模太小，投入相關研發的意願不高。因此，目前古籍數位化缺乏廉價好用的工具，沒有妥善的工具，製作古籍數位資料庫的成本居高不下。現況是，漢學研究者必須付出高昂的代價取得資料庫的使用授權，或是勉強將就於品質差、功能不足的免費資料庫。

技術規格

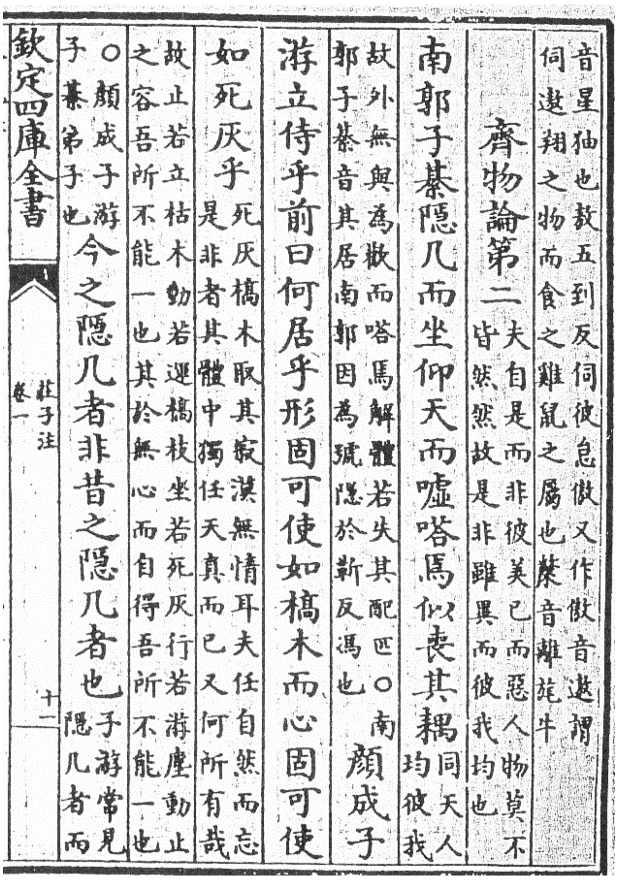
本工作室從1993開始投入佛經的數位化工作。處理佛經的技術和經驗，很大程度適用於漢文古籍，比方說缺字(稍後詳述)、文言文、科判等等。2008 年開始，我們在Accelon 的基礎上，開始設計一個新的古籍處理平台，由於這兩年是數位世界變化相當大的時期，邁入了由行動上網載具和雲端運算為主流的「後PC時代」，本處理平台的規格也經過幾次變動，以及大量的程式碼重寫。以下為最後確立的發展原則及技術規格：

一、使用開放的標準，如HTML5, XML, JSON 。不使用任何專屬格式。
二、以 PC 做為資料庫的製作端。平板設備及電子書閱讀器做為主要消費端。
三、以雲端協作方式，業餘愛好者也可以參與資料庫的製作和改善。和維基模式不同的是，維基是由多人協作創作。而古籍資料庫是由少數的專業人員製作，大部份人引用、加注及糾錯。
四、為符合使用者的習慣，數位化的過程中，我們刻意保留了古書的版式。因為這個決

定，提高了程式開發難度，但得到了意外的收穫。
**由於古書處理涉及的主題較多，本文主要討論如何以HTML5處理古書版式。

**古書的版式

**紙本漢文古籍的總字數，目前無法做精確的統計，若以四全庫書，收錄3000餘種書，共6億字，以平均每種20萬字計，目前已知共十五萬種古籍總字數高達 300億字，這是數以萬計的作者，跨越在數千年的時間，積累出來的創作。也是人類歷史上，重要的文化財產。


音星狖也敖五到反伺彼息徼又作儌音邀謂伺遨翔之物而食之雞鼠之屬也斄音離氂牛
齊物論第二夫自是而非彼美己而惡人物莫不皆然然故是非雖異而彼我均也
南郭子綦隱几而坐仰天而噓嗒焉似喪其耦同天人均彼我
故外無與為歡而嗒焉解體若失其配匹○南郭子綦音其居南郭因為號隱於靳反馮也顏成子
游立侍乎前曰何居乎形固可使如槁木而心固可使
如死灰乎死灰槁木取其寂漠無情耳夫任自然而忘是非者其體中獨任天真而已又何所有哉
故止若立枯木動若運槁枝坐若死灰行若遊塵動止之容吾所不能一也其於無心而自得吾所不能一也
○顏成子游子綦弟子也今之隱几者非昔之隱几者也子游常見隱几者而
欽定四庫全書
莊子注 卷一
十二

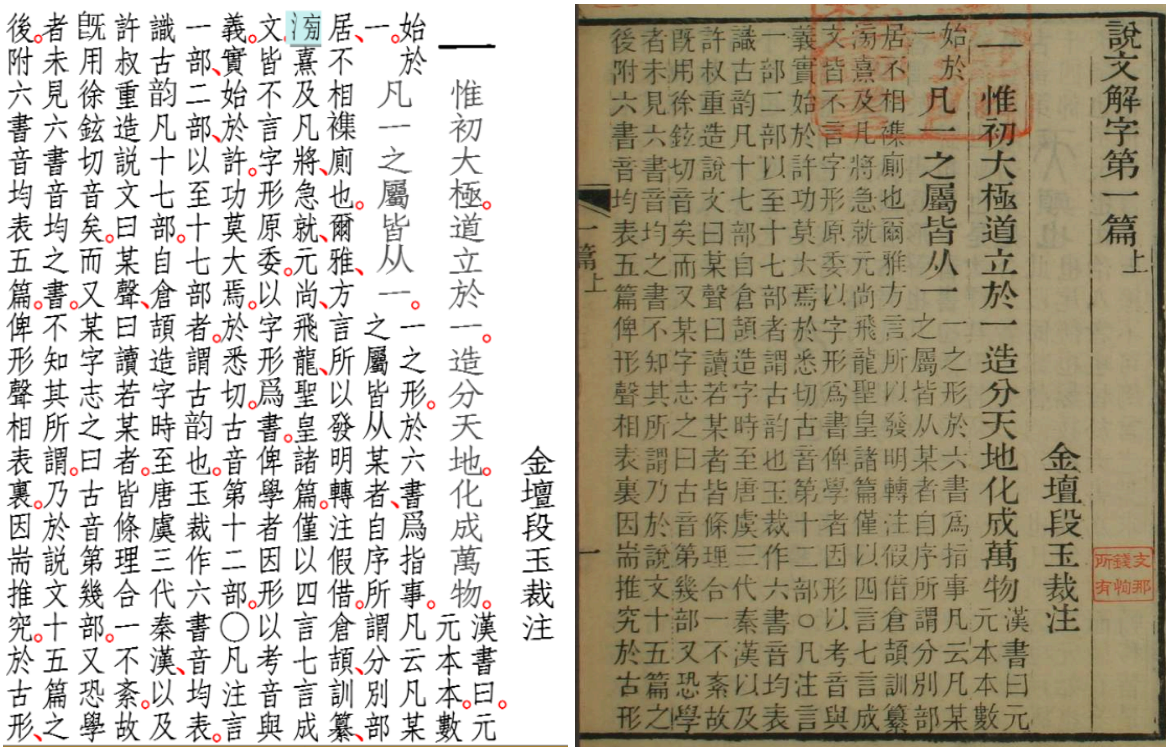
**《四庫全書. 莊子》版面

上圖出處 http://upload.wikimedia.org/wikipedia/commons/e/e6/Eine-Seite-aus-dem-Zhuangzi.jpg

**在電腦排版問世之前，如上圖所示的「直排正文雙行夾注」(vertical text with double interliner annotation) 一直都是紙本漢文古書的主要表現形式。

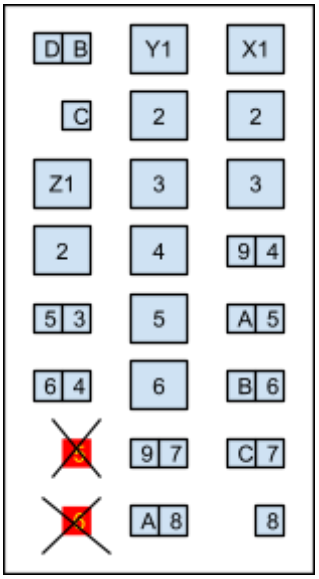
在HTML5 未問世之前，中文直書不容易實現，一般會用**巢狀(nested) HTML表格來模擬，但需大量人工作業而且不利被搜尋。另有像「竹取」的自動豎排套件，但只能在 Firefox 上正確呈現 (Chrome, IE 會有文字轉向的問題) 。

於是我們決定用 HTML5 提供的Canvas，重新撰寫一個豎排的文字呈現元件，不但充份支持豎排，還可以呈現夾注，不佔字位的標點符號，效果見下圖：



左：HTML5 排版效果。右：原書掃瞄圖檔

**中文直書的排列規則



**CSS 沒有描述直排雙行夾注的機制，因此我們必須撰寫排版引擎(layout engine)，它可以全自動地處理換行，以及空白處的平均分佈。上圖的例子中有X,Y,Z三個段落，每一段落有「正文」及「夾注」兩部份。文字的排入順序是從上到下，從右到左。必須注意的是，當一頁結束時，Z段的5,6兩個字元，必須折到左邊。

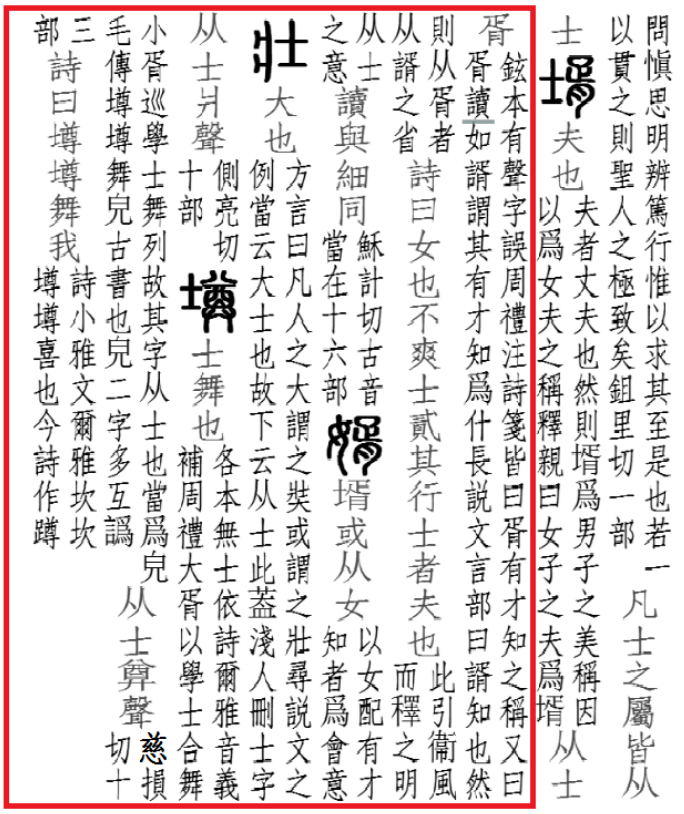
**我們設計了自動折行(line wrapping)的機制，除了夾注必須標註(tagging)之外，其他都是自動完成的，XML 如下：

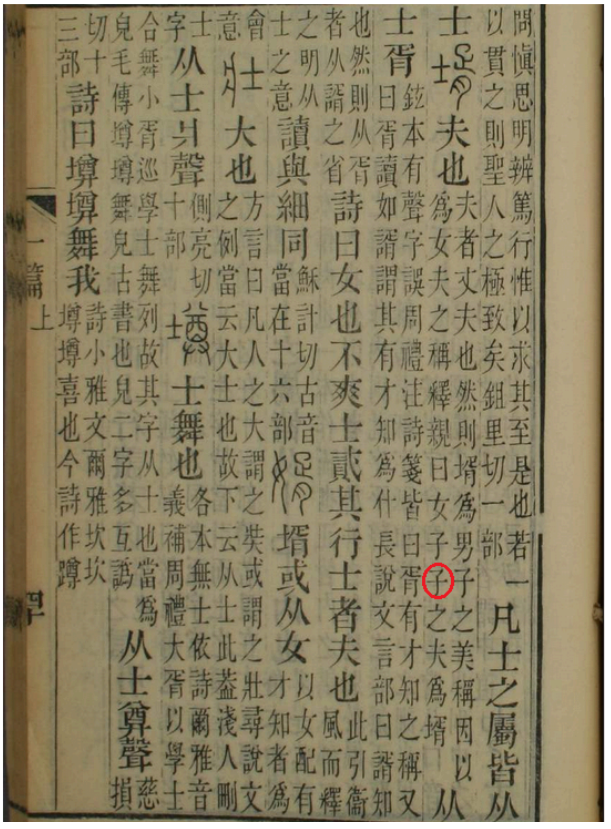
```
X123<anno>456789ABC</anno>Y123456<anno>789ABCD</anno>Z12<anno>3456</anno>
```

**相較於巢狀表格的方案，這種精簡的 XML 節省了大量的標註人力。

業餘校對

古籍的校對非常困難，我們採用的文本數據，Kanjibase 的說文解字注XML，是一份製作精良、錯誤率很低的電子檔。日本人一向以態度嚴謹而著稱，這份文件相信當初是費了不少力氣製作的。縱然如此，經由豎排，我們的工讀生還是很輕易地發現不少錯誤，由下圖可知，從第五行開始，排版就亂掉了，於是問題點一定出在上一行，也就是第四行漏打了一個「子」字。誤：「女子之夫為壻」正：「女子子之夫為壻」。除了奪文(漏字)，衍文(贅字)也無所循形。


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之夫爲壻　从士胥聲
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說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衞風而釋之明从士之意　讀與細同　蘇計切古音當在十六部
壯　大也　方言曰凡人之大謂之奘或謂之壯尋說文之例當云大士也故下云从士此蓋淺人刪士字从士爿聲　側亮切十部
壵　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壵壵舞兒古書也兒二字多互譌　从士舜聲　慈損切十三部　詩曰壵壵舞我　詩小雅文爾雅坎坎壵壵喜也今詩作蹲



*古書的標點(句讀)

**保留原書版式製成的數位文件，還有一個很大的好處，就是可以重現古書的標點，對中文古書來說，標點並不是由作者加上的，換句話說，標點並不是預先寫好的文字，而是在閱讀的過程中，隨看隨作的「內容標記」(contextual markup)，因此，中文古書中的標點，含有很大的主觀的成份(subjectivity)，從這個角度來看，同樣內容的古書，由不同的人閱讀過加上的標點，就產生了獨一無二的加值 (unique value-added) 。

**我們將內容和標點分個兩個圖層(layer)，技術上它們是獨立並重疊(seperate and overlapped) 的HTML5 Canvas tag。在內容的圖層上，繪製出和原書版式相對位置一致的文字，因此，從原書的某個座標點(coordinate point)，很容易對映到電子版的某個字，反之亦然(vice versa)。而在標點的圖層，有兩種輸入來源，其一是讓現代讀者(主要是文學院的學生)在電子版面用滑鼠直接加上標點，老師根據這些標點，可以評量出學生理解古書內容的水平。這部份我們已有雛型(prototype)。

**另一種輸入的方式，則是利用影像識別的技術(visual pattern recognition)，找出古人的標點 (通常是用紅色的硃砂cinnabar做成的墨水)，並從座標點換算回電子版的位置(character position)，達成全自動化的標點輸入。這是未來我們會努力的方向。

**下圖是一個珍貴的刊本(printed version)，十三世紀元代(13 century Yuan dynasty)。上面有明顯的紅色標點。標點的年代已不可考(unknown)。

http://catalog.digitalarchives.tw/item/00/07/e8/d8.html

左傳 元代刊本

結語

以CBETA(中華電子佛典協會)為代表的佛經的數位化成果，免費開放供所有人使用，近幾年來對佛教學術界帶來重大的影響，佛學專家節省了大量翻閱資料的時間，論文的數量和品質都有顯著的提升。我們相信在佛學界發生的事，不久的將來，也會在漢學界發生。

我們對原書版式協助校對的效果尤其感到振奮，從此，單調枯燥的校對工作變得有趣，基本上不必經過專門訓練，一般志工都可以承擔校對的工作。而中文的專家可以省下大量的時間和精力，專注於更有創造力的工作。

古籍漂流在漫漫歷史長河，終於有了一個擺脫物質束縛的契機，以能量的形式無遠弗屆，分身無數億。目前，一個8GB的隨身儲存(USB Disk/ SD Card)只需不到10美元，就可以容納所有的漢文古籍，這也許就是佛經中所形容的「芥子納須彌」吧。

和所有的數位資料庫一樣，它的製作和維護(非常艱難，但是完成之後，從技術的角度而言，每個單位的持有和分享成本趨近於零。因此，古籍的數位資料庫，先天上就和公共建設一樣，有強烈的「前人種樹、後人乘涼」的特質，而相關的計畫也常常具有濃厚的社

會主義色彩。

古籍數位化一者帶不來選票，政府無心於此；二者帶不來鈔票，廠商興趣缺缺。身為一個非營利機構，在人力和財力不充裕的情況下，我們懷著繼往開來的心情，承擔起這個具有歷史意義的任務，致力於打造優良的數位化工具。希望讓漢文典籍得以在數位世界保存及發揚，讓古人的智慧，繼續嘉惠後世。

感謝

一、早稻田大學提供說文解字注的掃瞄圖檔 (scanned transcript from Waseda University, Japan)
http://archive.wul.waseda.ac.jp/kosho/ho04/ho04_00026/ho04_00026_0001/
二、説文解字注データ (swjz database)
http://kanji-database.sourceforge.net/dict/swjz/
三、漢字構形資料庫(The CDPHanzi Database)，由台灣中央研究院(Taiwan Acedemia Sinica) 資訊科學研究所(Institute of Information Science) 文獻處理實驗室(Chinese Document Processing)提供。http://cdp.sinica.edu.tw/