漢籍全文數位化工作流程指南

謝筱琳

壹、前言

當火焰燃燒到達華氏 451 度，所有記錄人類智慧的典籍都將灰飛煙滅，但卻燒不毀被壓抑的靈魂……這是一個沒有火災的世界，消防員的工作，是縱火。這是一個所有的書都是「禁書」的世界，消防員的職責，是「焚書」。

這樣的場景，發生於當代科幻大師雷・布萊伯利(Ray Bradbury)1953 年發表的科幻小說——《華氏 451 度(Fahrenheit 451)》。未來的西方世界，實體書籍將因某種因素淪入如同中國古代秦始皇下令焚書的悲慘命運，而愛書人爲了拯救這些人類歷代傳承的知識典籍，自願成爲知識的載具，將經典書籍的內容記憶於腦海，再現於言行，因而每個人都是一本書，如果有人想了解聖經，就來這裡找「聖經」這個人，想拜讀莎士比亞，就去那裡找「莎士比亞」那個人。

「每個人都是一本書」以保存、傳遞知識的這種想法，是 1953 年電腦、網際網路尚未蓬勃發展，布萊伯利針對書籍無法以實體形式存在，又必須另謀存活形式，所構想的解決方案。然而，假設這本書的寫作年代爲二十一世紀的今日，由人作爲知識載具的浪漫劇情，或許直接爲數位化取代，電子化的書籍透過電腦、網際網路，更能客觀地、科學地、久遠地、安全地保存，並且經由網路易於傳遞的性質，知識能夠快速且有效地流通。

圖書文獻的電子化，是將實體(例如紙本)的典籍進行數位化作業，轉換成虛擬的電子形式。目前比較常見的數位化方案有三種，第一是針對書籍的原貌，依照既有的頁數或篇幅，一一拍攝，產出黑白或彩色的數位化影像，稱作「全文影像數位化」；此方案的優點是能夠同時呈現書籍的原文、紙質與原色。第二種「全文數位化」，是將書上的文字進行繕打輸入與校對，產出內文的文字電子檔；此

方案能辨識註解內文、意義模糊不清處，以現行文字代替目前已經不通行的古文難字，並且堤供全文檢索，進而提高典籍之研究價值。上述兩個數位方案，前者重視書籍「形式」的再現，讀者能夠觀看原書的樣貌影像，後者則注重書籍「內容」的呈現與索引，讀者能夠藉此閱讀原書的文字與意義。另外還有第三種方案結合前兩種考量，分別將形式與內容數位化後，共同儲存於資料庫中，再將兩者並置於畫面或建立超連結呈現，以達形式、內容相輔之效。

數位化資料除了有 AAA（Anyone Anywhere Anytime）——任何人隨時隨地均可取得資料之特性，亦有其他四大益處：一、能大量儲存於攜帶方便的光碟片、硬碟、或是磁碟陣列，節省空間，利於保存；二、能夠展現新的資料型式，如超文件(hypertext)、超媒體(multi-media)，令人耳目一新；三、刺激開發研究的新方向，以佛教典籍的全文數位化爲例，藉由數位化後之詞頻統計，可以得知佛典之用詞概況，易於歸納佛典相關之典章、事故；四、易於複製、傳遞、傳播，增進知識之流通【註 1】。

我國最早的電子全文資料庫「漢籍電子文獻資料庫」，爲中央研究院於 1981 年研發製作，收錄二十五史、十三經、清實錄、小說戲曲……等各種歷代古籍叢書之電子全文，是國內現有最周全的經典史料資料庫。行政院文化建設委員會「國家文化資料庫」收集之全國藝文資源中，也有其他古今、地方文學與古文書的全文影像資料。宗教典籍方面，道教有中央研究院中國文哲研究所研究員李豐楙先生帶領的「正統道藏」全文數位化工作；佛教細分經錄與經文，前者有中華電子佛典協會的佛教藏經目錄數位資料庫，以及香光尼眾佛學院之藏經目錄整合查詢系統，後者包括佛光山和中華電子佛典協會製作的《大正》、《阿含》等藏經全文資料庫，其中，佛光山是以精選數部佛典加以新式標點，並附題解及註解，付費使用之方式經營，中華電子佛典協會的「佛典集成電子藏經資料庫」，則按原書全文全套製作，並提供免費瀏覽下載，其經文之質量皆踞漢文佛典數位化之冠。

國科會數位典藏國家型科技計畫自民國 90 年計畫執行起，致力推動國內各項文物、資產的數位化工作，整合台灣漢字全文數位化的「漢籍全文主題小組」

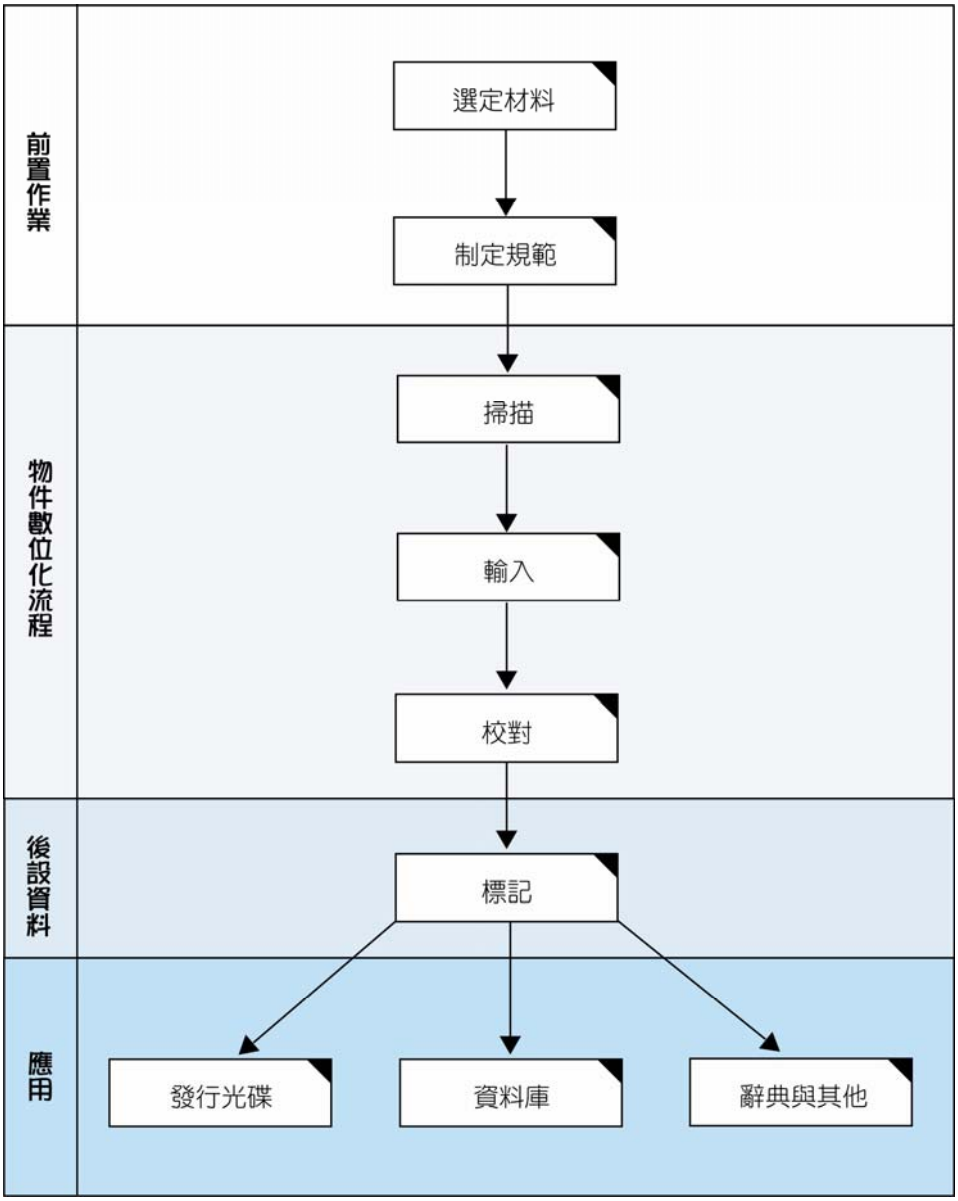
也於 94 年度 7 月成立。計畫成員有中華佛學研究所「佛典數位典藏內容開發之研究與建構——經錄與經文內容標記與知識架構」計畫，主題小組召集人由中華佛學研究所圖書資訊館館長——杜正民先生擔任，積極推動國內全文數位化相關計畫、單位的研究分享與技術切磋，其中，對於全文標記的推廣更是不遺餘力，因爲合適生存於數位時代的數位資訊，必須具備優良成品、加值應用價值高，且能分享流通於世界之特性，故「全文標記」也是本文之撰寫重心。

本「漢籍全文數位化工作流程指南」，係紀錄彙整已執行漢籍全文數位化計畫、單位之工作經驗，參考國內外相關技術、標準，歸納統整一套電子全文數位化之工作流程，期能提供執行中單位之管理者與實際作業人員觀摩、檢視，且讓後續更多文書收藏單位加入數位典藏時，能依此參考依據，更有效的執行數位化工作。

貳、數位化工作流程圖

全文數位化之工作流程可分爲四大部分。第一部分爲數位化之前置作業，即工作前的規劃評估與準備；第二部份爲實際進行文字數位化的工作程序，諸如文本的掃描、影印，文字的繕打輸入以及校對；第三部份標記，是以記錄原有文本的排版資訊與內容資訊，進而增進學術研究與應用價值之重點工作；第四部份應用則爲原始書籍轉爲電子全文後之應用發展，詳細流程可見下表 1。本文亦依此順序介紹說明。

表 1、漢籍全文數位化工作流程圖



叁、前置作業

無論欲進行數位化材料之數量多寡，或是計畫之規模大小，數位化都是一項所費不眥且耗時耗力的工作；因此，事前的籌畫評估與準備，不容小覷。

實際進行文字數位化工作之前的準備作業有兩項，第一爲「選定材料」，此工作涉及對於既有材料之瞭解與整體目標之規劃，第二爲「制訂作業規範」，有助整體作業之執行成效與品質管理。

一、選定材料

(一)數位化材料之選擇

根據數位化目的之不同，數位化材料之選擇標準亦各有異，加上每個計畫經費不一，且多有限，故選擇數位化材料時，應事先考量文物數位化之優先順序，使得人力經費之支出能達到最佳成效。文物數位化之優先順序【註 2】，可依照珍貴性、重要性、成本效益等程度，概分爲以下六項：

1. 典藏品的評等度，如國寶、重要古物等教育部標準、機密程度等等。
2. 典藏品的珍貴度，例如文物具有獨創性、稀有性、時代價值、不可替代性等因素。
3. 典藏品的保存難易度，包括脆弱、無法複製拷貝、有消失之虞等考量。
4. 數位化後之成本效益。
5. 數位化後在研究、教育、經濟等能面的應用價值。
6. 其他。

此外，版本的選擇在全文數位化工作裡，也佔有很大比例的關注與考量。當知識近入書寫時代，印刷術尚未發明之前，書籍多以人工抄錄的方式傳承接遞，故常有抄錯、漏抄之時，即使進入鉛字排版時代，還是會有會錯意、選錯字的情況發生；而且中國以儒家立國，不僅歷代君王，亦含文人雅士、學者志士，重視書籍文獻之考察與典藏，官修、私修之史書叢集不斷，最有名爲清朝乾隆皇帝欽點之「二十四史」，可見史書典籍因年代版本不同有收錄、記載之差異，是協助研究學者進行分析研究之重要線索。此外，數位化的重製作業關係藏品與成品之著作權法，若能請求著作權授權，則可順利進行數位化；若無法取得，或可更換可以取得授權之版本，否則必須重新揀選材料。

(二) 數位化材料清單之建立

選定即將進行數位化之書目後，應建立一份完整詳實的待輸入書目清單。因爲數位化的物件爲書籍，故以書目之最小集合單位冊或卷爲列入清單的基本單位，一冊書即著錄一筆資料，每筆資料所記載的內容應包含以下出版資訊：

1. 書名
2. 作者名
3. 出版地
4. 出版社
5. 出版年限
6. 版本

除上述出版資訊外，亦須記載書目之數量，並且妥善保存此書目資訊（如表 2），作爲之後進行數位化工作之憑據。

表 2、漢籍全文資料庫待輸入書單

	漢籍全文資料庫待輸入書單	
	書　　名	版本
1	七才子詩選	
2	九卿議定物料價值 四卷	(清)工部編，清乾隆元年(1736)刊本
3	二十五史外人物總傳要籍集成	董治安主編；濟南：齊魯書社，2000
4	入告初編 一卷，二編 一卷，三編 一卷	(清)張惟赤撰，清順治(1644-1661)末刊本
5	入幕須知 五種	(清)張廷驤輯，清光緒十八年(1892)浙江書局刊本
6	八旗人口冊	不著編人，清光緒間(1875-1908)排印本
7	十科策略 十卷	(清)劉文安著
8	三流道里表	(清)唐紹祖等纂修，清乾隆年間武英殿刊本
9	三流道里表	不著編人，清同治十一年(1872)江蘇書局重刊本
10	三流道里表 不分卷	(清)查克順等纂，清乾隆四十九年(1784)刊本
11	三省礦防考 二卷	(明)劉應元撰，明隆慶元年(1567)刊本
12	三朝聖諭錄 三卷	(明)楊士奇輯錄，明鈔本；漢籍資料庫有建置《國朝典故》
13	三賢政書 三種	(清)吳元炳輯，清光緒五年(1879)序刊本
14	上諭內閣 一百五十九卷	(清)允祿等輯，清刊本
15	上諭合律鄉約全書 一卷	(清)聖祖諭，(清)陳棨直解，清康熙間(1662-1722)刊本
16	于山奏牘 八卷	(清)于成龍著，清康熙年間(1662-1722)刊本
17	于清端公政書 八卷，外集一卷，續集一卷	(清)于成龍撰，(清)蔡方炳編次，清乾隆二十六年(1761)刊本
18	于肅愍公奏議 十卷	(明)于謙撰，明嘉靖二十年(1541)杭州重刊本
19	于肅愍公集	
20	大明九卿事例案例 不分卷	不著編人，明鈔本
21	大明令 不分卷	(明)太祖撰，鈔本
22	大明律 三十卷	胡瓊集解，胡效才增附，日本蓬左文庫藏明嘉靖刊本(本院圖書館查無此書)

二、制定規範

爲確保數位化前後環節銜接順暢，產出成果之品質穩定，需制定相關作業規則與檔案格式，以供遵循與評量。

大多數計畫或單位之規範制定，參考同業已訂立之標準，其他則來自自身經驗的累積，不過制定規範還是以滿足個別計畫之最終目標爲最高準則，並非一字不漏、全部採信他人作法。雖然作業規範依計畫目標有所歧異，之間仍有些許共同原則，以下介紹各項數位化作業之工作規範與檔案規則參考。

(一)數位影像檔案規格

提供文字的繕打輸入之底本依據的方式有二種，一爲掃描原書製成數位影像，二爲影印原書製作複本。

掃描原書之圖檔，除可用於繕打輸入與校對用途，還可於之後資料庫建置時，將圖檔與相對應之電子全文連結，成爲全文與影像資料庫。關於數位化影像之規格設定，數位典藏國家型科技計畫區隔數位化檔案規格爲瀏覽級、商務級與典藏級三級：

1. 典藏級圖檔：目的爲永久典藏，影像品質不失真，詳實反映原件狀況。
2. 商務級圖檔：目的爲提供未來之加值應用，如出版、印刷、複製、交換或販售，影像品質須符合印刷之要求。
3. 瀏覽級圖檔：目的爲展示於網路上，影像品質須符合電腦螢幕瀏覽及網路傳輸之要求。

典藏級的影像解析度爲人類眼睛鑑別影像最高值的 300dpi，格式爲適用不同軟體、平台，壓縮不失真，適合作爲原始之 TIFF 檔。商務級的解析度同樣爲 300dpi，影像格式則是高效壓縮使得檔案變小之 JPEG 檔，容易造成影像細節流失。瀏覽級之影像解析度則爲便利傳遞，再利用價值低之 72dpi，影像格式爲 JPEG。三者的色彩模式則都爲 RGB(24bit/pixel)。 (表 3)

表 3、數位典藏國家型科技計畫數位化檔案格式

	瀏覽級	商務級	典藏級
檔案格式	JPEG	JPEG	TIFF
色彩模式	RGB(24bit/pixel)	RGB(24bit/pixel)	RGB(24bit/pixel)
解析度	72dpi	300dpi	300dpi

有鑒於全文數位化之掃描圖檔可能於使用於不同目的，故原始掃描圖檔應設爲高階的典藏級，即 300dpi 的 TIFF 全彩檔，以便日後降階應用。

另外，也有單位使用影印的副本作爲繕打輸入之底本，影印之影像大小則依據原書字體大小與清晰度，決定比例大小。

(二)數位檔案命名原則

一旦執行掃描，產生數位影像之後，便需一一替檔案個別命名，以便利數位資料之管理與檢索。

使用檔案命名字元時，爲確保檔案名稱能夠符合不同作業平台之讀取格式，應注意一般檔案命名事項：

1. 以小寫英文字母與數字做爲檔案命名之編碼組合。
2. 避免使用%、/、?、#、*、-等特殊字元。

除了一般性原則之外，亦需依照數位化物件之媒體類別與不同特性，額外增加能夠突顯物件特性之命名規則。掃描圖書典籍而產生之數位化影像，其檔案名稱包含三種層次，圖書代碼、冊卷號、頁碼。其中頁碼爲檔名，副檔名爲.tif。

例：aaaaaooozzzzzzzzz.tif

aaaaa=圖書代碼；

ooo=冊次號；

zzzzzzzz=頁碼。

1. 第一層：圖書代碼

   長度不固定，計畫單位可自行設定，建議皆爲數字。

2. 第二層：冊卷號

長度固定爲三碼，皆爲數字。

3. 第三層：頁碼

(1) 檔名長度共 8bytes，依原書內容頁碼編頁。

例：第一頁 → 00000001.tif。

(2) 封面頁碼固定爲 c0000001.jpg，倘若爲平裝書加工精裝者以原平裝書之封面爲主。

(3) 原書內文頁碼第一頁前面與內文頁碼不連貫之各頁(即非正文部份)，如序、目次等，可於非正文部份起依序計數，並於頁碼第一位加上英文小寫字母“a”以區別之，如：a0000001.tif，a0000002.tif…

(4) 內文後面多出且與內文頁碼不連貫之各頁，如附錄、圖表、參考資料等，可於非正文部份起依序計數， 並於頁碼第一位加上英文小寫字母“ b ” 以區別之， 如：b0000001.tif，b0000002.tif…

(5) 原文編有頁碼之空白頁或廣告頁，仍依原順序編碼掃入。

(6) 原文未編頁碼且爲多餘之空白頁，則予以跳過不掃。

(7) 內文中之插頁，若未編頁碼，則以接續前頁之編碼後加“_”編入。如：在 86 頁至 87 頁間插頁 2 頁但未編碼， 則以“ 000086_1.tif ” 、“000086_2.tif”編號。

(8) 原文若分左、右版面頁碼者，左版頁碼需以小寫 L 區別，右版頁以小寫 R 區別。

如：

頁左 133→檔名爲 l0000l33.tif

頁右 12→檔名爲 r0000012.tif

(9) 正文若同時有兩組頁碼標示者，例如一組各章節從 1 編頁，一組爲總頁碼者，則掃描取該書冊目次所標示之頁碼爲準【註 3】。

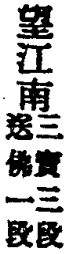
(三)人工輸入規則

執行人工輸入之前，必須建立輸入規則。除了內文的文字，本文以外之符號標誌、圖片、表格、夾注小字、段落、頁碼、欄位、校勘符號，以及空白字元、空白行、圖形、系統缺字……等，都需明確標示著錄格式，如：

1. 頁碼、欄位：每欄開始都要以半形英數先輸入一行 pxxxxn，xxxx 爲四位數，n 爲 a（上欄）或 b（中欄）或 c（下欄）。
2. 序及經卷名：行前不留白。
3. 作者及譯者名：行前留四個全形空白。
4. 正文：行前不留白。
5. 正文夾註小字：以一組半形(　)前後包括。

例  輸入爲：十一月(二段)

6. 雙行夾註小字：同樣以一組半形(　)前後夾住，需注意文字走向。

例  輸入成：望江南(三寶三段送佛一段)

7. 空行：隨文中空行。
8. 空格：按書面空格輸入全形空白字元。

9. 圈點：隨圈點處輸入「。」。

例 

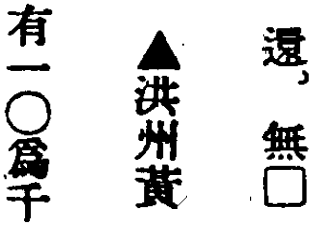
輸入爲：身所居。二自受用土。自受

10. 校勘符號：採兩位數半形阿拉伯數字與中括號

例 

輸入爲：相[01]把成陰陽。

11. 特殊符號：以相似全形符號表示。

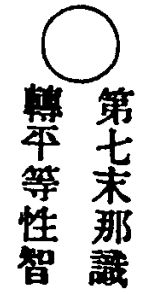
例 

各輸入成：有一〇為千

▲洪州黃

還。無□

12. 圖形：以【圖】表示。

例 

輸入爲：【圖】第七末那識

轉平等性智

13. 缺字：如果可以用組字式或構字式（下一小節將介紹）表示，即示之；若模糊或是難以表達之處，可統一暫以全形●表示【註 4】。

格式化: 項目符號及編號

因爲每種文獻的排版、書寫、或語法等書籍體例各有不同，應根據各書籍體例以及數位化目標，制定適合個別體例之人工輸入規範。不過大致上，關於全文輸入，還是有以下幾基本大原則：

1. 依照書中原文輸入，內文不清楚處，不做模糊判斷，待專業人士進行判斷。
2. 同原書換行位置折行。
3. 由於古文書多無標點符號，輸入時只進行斷句，不加註新式標點符號。

(四)新增缺字系統

漢字發展過程裡，因爲地域、時代或其他因素，衍生了一字多形（例如「眾」與「衆」），無法窮舉之特色，所以現有的電腦交換碼一旦用來處理古籍或佛典、道藏等文獻，缺字問題即層出不窮。缺字的根本及務實的解決之道，應該在現有的編碼方法下，根據漢字的構形規則，針對這些爲數眾多但又不常出現的漢字，提出一套有效的編碼方法【註 5】。

目前國內有兩種缺字組字標準，其一是最爲廣泛使用的缺字系統——中央研究院研發之漢字構形資料庫【註 6】。漢字構形之基本構字單位稱作部件，也就是一個用來構造其它字的形體。如「日」、「京」是「景」的部件，「景」、「頁」是「顥」的部件，而「顥」是「灝」的部件。

部件還有層次，例如「顥」可拆分成「景」與「頁」，「景」又可拆分成「日」和「京」。漢字最常用的拆分方式爲橫連（ ⿰ ）、直連（ ⿱ ）與包含（ ⿴ ），因此，「顥」等於「景」 ⿰ 「頁」，「景」等於「日」 ⿱ 「京」，「圍」等於「口」

 「韋」。另外爲了輸入方便，也造了一些方便符號，表示相同部件之排列方式，如兩個「克」橫連的「兢」等於「  克」，兩個「戈」直連的「戔」等於「  戈」，三個「車」呈三角形狀排列的「轟」等於「  車」，四個「火」呈四角狀排列的「燚」=「  火」。而無從以構字式拆解之字形，則可使使用從缺符號「  」表示（表 4）。

表 4、漢字構形組字規則

符號類別	中文意義	組字符號	使用說明	範例
拆分符號	橫連		當部件的排列順序由左至右	灝、順
	直連		當部件的排列順序由上至下	含、義
	包含		當部件的排列順序由外至內	圍、魁、連
方便符號			二個相同部件直連	炎
			三個相同部件直連	
			二個相同部件橫連	朋、林、孖
			三個相同部件橫連	
			三個相同部件呈三角狀排列	焱、聶、磊
			四個相同部件橫連	
			四個相同部件直連	燚
			四個相同部件成四角狀排列	
其他	起始標示		當拆分有兩種以上時，代替拆分，包夾在所有部件之前面，以及最後	片戶甫
	終止標示			
	缺字標示		代替從缺的部件	

其二爲中華電子佛典協會在進行佛典的數位化工作時，以中央研究院之漢字構字式爲底本，獨家研發出該單位特有之組字式規則。相較於中央研究院之漢字構字式以部件作爲基本構形單位，中華電子佛典協會是以 BIG5(大五碼)系統字作爲組字之基本單位，故無造字問題，使用者無需另外安裝造字程式或圖檔，即可讀取組字式。

組字式採用數學裡的加減乘除四則運算符號來表示，共使用十個符號。這十個符號，其中七個——「*」、「/」、「@」、「-」、「+」、「(」、「)」，用來表示字的左右上下分合關係；問號「?」，表示某字無法用組字方式表示的部分；另外二個半形符號「[」與「]」，表示組字式的起迄【註 7】（表 5）。

表 5、中華電子佛典協會組字式規則

符號	說明	範例
*	表橫向連接	明＝日*月
/	表縱向連接	音＝立/日
@	表包含	因＝口@大　或　閒＝門@月
-	表去掉某部份	青＝請-言
-+	若前後配合，表示去掉某部份，而改以另一部份代替	閒＝間-日+月
?	表字根特別，尚未找到足以表示者	背＝（?*匕）/月
( )	爲運算分隔符號	繞＝組-且+（（土/（土*土））/兀）
[]	爲文字分隔符號	羅[目*侯]羅母耶輸陀羅比丘尼

上述兩種組字規則，前者多爲政府機關單位與中研院院內開發之資料庫所採用，是台灣發展最早之構形系統；後者則爲中華電子佛典協會獨用，它簡化了漢字構形之複雜度，協助繕打人員輕鬆組織缺字。此外，政府也研發一套國家標準中文交換碼方案，並由行政院主計處電子處理資料中心建置「CNS11643 中文標準交換碼全字庫」，以解決個人電腦中文字數不足與自行造字問題，不過，除政府機關與戶政單位採用此系統外，一般民間乏人問津。

(五)通用標記語言

標記（Mark up），是在稿件或文章上加上的記號，以記錄各種不同的資訊。爲避免自創造標記系統影響資料交換之互通性，國際間很早就開始建立通用的國

際標準。1986 年，發明了最早的標記語言 SGML（Standard Generalized Markup Language，標準通用標記語言），它定義了如何描述一組標記標籤（tag）的規則，但由於它相當的複雜，因此應用並不十分普遍。其次是紅遍半邊天的 HTML（Hypertext Markup Language，超文字標記語言），HTML 是 SGML 的一種應用，以其簡單易用的語法隨著網際網路的興起而盛行，世界各地不同語言、文化、電腦作業平台之間，得以藉由 HTML 這個標準的共通語言相互溝通，地球村的資訊交流達到前所未有的迅速與廣度。

然而 HTML 的缺點——正是它的優點——也漸漸的浮現，HTML 不再能滿足網際網路上許多新興的需求。SGML 夠強卻太複雜，HTML 夠簡單卻不夠強大，於是標記語言的專家又爲設計了一套既強大、又不太難、且適用於網際網路的標記語言——XML（eXtensible Markup Language，可延伸性標示語言）。

標記主要可應用於兩類，一類是關於「排版或顯示格式」的標記，另一類則是關於「資料結構或內容」的標記。例如最爲大眾熟悉之標記語言 HTML（HyperText Markup Language），可能會有如下的用法：

佛教資料電子化技術探討－－以< b >中華電子佛典協會< /b >為例

在這裡，< b >……< /b >表示「中華電子佛典協會」這些字要加粗體字(bold)顯示，這是第一類關於「格式」的標記。而關於「內容」的標記，可能爲如下用法：

史記

< byline type="Author">司馬遷< /byline >

這裡的< byline >……</ byline >標出史記之作者(author)爲司馬遷。這種將「顯示格式」與「內容」分離的做法，能讓電腦「看懂」經文【註 8】。

有了共同的標記語言 XML，標記格式如出一轍，可用同樣的標記語言、標記格式來定義各自不同的標籤名稱，例如要標出一個段落，可以有如下數種不同

的標法：

1. < p >……< /p >
2. < para >……< /para >
3. < 段落 >……< /段落 >

這些都是符合 XML 標準的標記，但在資訊交換上將會造成問題，需要增加一道標記轉換的手續。如果能有共同且統一規格之標籤名稱，這個問題就可以解決。

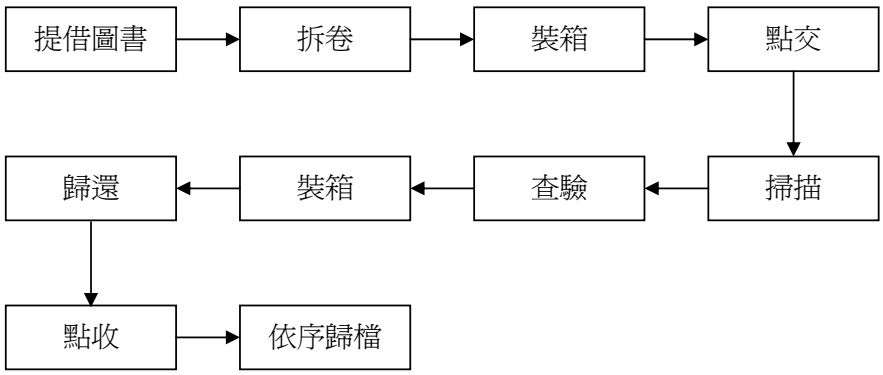
除此之外，早在 SGML 時代就有一個 TEI（Text Encoding Initiative，文件符碼化）專案，研究各種不同西方文獻，整理出一套標籤集（Tag Set），希望獲得各方採用而促進電子文獻的分享交流。由於 TEI 的標籤集是根據文獻所歸納制訂，與 SGML、HTML、或是 XML 相比，忠實反應了文獻的內容與架構，例如完整書目資訊、文獻及其來源之關係和版本、使用語言等，都有特定之標籤，足以滿足文獻標記之需求。本文將於第伍章——後設資料建置，概述 TEI 之基本介紹與操作守則。

肆、數位化物件程序

數位化物件程序指的是書籍在進行數位化過程中的掃描、輸入、與校對工作，此三項作業可以說是全文數位化的主要骨幹，也可以是最底層的基礎。事實上，書籍經過掃描、全文輸入與校對後，已經可以算是數位化過了，只不過只有經過此三項作業所產出之電子全文尚屬生料，若想增加日後之應用層面與研究價值，則需再進入後續之標記作業。

一、掃描

原書圖像之掃描之工作，視數位化藏品數量、計畫經費以及成本效益等考量，可選擇自行購買掃描器自製圖像，或是委託外部掃描廠商承包處理。若選擇自行掃描，可購買具備「自動送紙功能」與「自動編號存檔」之掃描器，節省人力。圖書掃描之作業流程【註 9】如下：



(一)提借圖書

根據書單，提調所選定之書籍，以提供單位內掃描人員或得標廠商影

印及掃描。

(二)拆卷

將原書或原書影本拆卷，裁切騎縫邊，以備送掃。

(三)裝箱

將散裝書籍置入紙箱，並標示書名冊卷號，以資識別。

(四)點交

掃描人員或廠商所屬工作人員提卷，點驗書籍冊、卷名稱與數量正確無誤後，於簽收簿(表 6)上簽名，始得提領。

表 6、圖書提領點收表單

書名	冊數	提卷日期	提卷人簽名	歸還日期	點收人簽名	備註

(五)掃描

1. 掃描之作業流程

(1) 掃描；

(2) 抽樣查看掃描品質——有無線條或歪斜不清者；

(3) 掃描完畢後，檢查有無漏頁；

(4) 按照檔案命名原則編入檔名；

(5) 抽樣檢查頁數正確與否；

(6) 轉檔；

(7) 燒錄；

(8) 燒錄完成後，瀏覽檔案，若有缺漏或無法開啓的檔，加以修改或補齊；

(9) 歸檔；

(10) 清潔掃描器。

2. 掃描之注意事項

(1) 掃描時應按冊或按卷掃描，同一冊(卷)由同一人處理爲原則，一卷完全掃描完畢，始得進行下一冊(卷)，嚴禁不同冊(卷)交疊掃描建檔，或同一冊卷由二人以上分開處理。

(2) 掃描產生之圖檔需先設爲高階影像，即 300dpi 的 TIFF 全彩檔。而交付人工輸入之圖檔，可以再降階轉成 TIFF-g4 黑白格式，畫質清晰且檔案又小。

(六)查驗

影像檔掃描建檔後應建立檔案清單，顯示檔案名稱、大小、筆數等資料，提供品質及數量檢驗。

(七)裝箱、點收、歸還

掃描人員或廠商所屬工作人員歸還書籍時，應恢復點交時狀況予以點收，整理順序後歸檔。

二、輸入

文字的輸入方法有兩種，包括傳統的人工輸入法，以及運用軟體的 OCR（Optical Character Recognition，光學文字辨識系統）辨識法。比較省時省力的方式是採用 OCR 圖檔辨識，此系統能夠自動分析圖檔上的文字並轉爲純文字檔，但是目前的光學辨識系統仍處於辨識鉛字印刷文字之階段，尚無法精確辨識古籍經典中手之抄本或是雕版文字。因此，無法以 OCR 辨識之圖書文獻，則交付傳統的人工輸入，一字一句繕打處理，這種輸入法也是目前國內全文數位化單位最常採用的方式。

此外，還有一種製作文字檔的方式，就是收集網路上的現成電子檔，再加以修改格式成爲符合單位所需文字檔。通常文史或宗教類的經典才可能有現成的電子檔，尤以佛經爲最，據傳民間流傳抄寫佛經能夠累積功德，因而有許多信徒自發地在網路上公布其繕打之電子經文，一方面替自己累積功德，另一方面對於佛法的宣揚有所幫助。目前，僅有製作佛經數位化的中華電子佛典協會，從網路收集不少對於佛法或是佛典有興趣人士所自製的電子經文，作爲校對的參照。以下，將分別介紹人工輸入與 OCR 光學文字辨識系統之作業辦法與須知。

（一）人工輸入

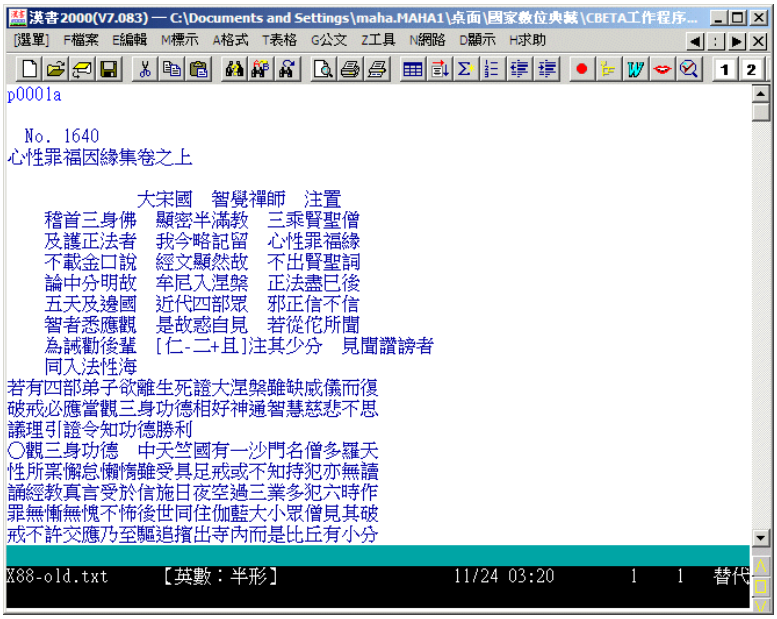
由於人工輸入必須依照原文一字不漏地輸入成電子檔，是費時耗力的大工程，除非預定進行數位化的圖書數量非常少，計畫內人力可以自行應付，否則絕大多數全文數位化單位會將此項作業委外製作，即使在人力成本不斷上揚的今日，委外人工輸入的作法還是全文數位化單位精簡勞務的最佳選擇。

中文人工輸入方面，亞洲這裡除了台灣之外，與台灣較近的市場還有中國大陸以及印度。台灣的人工打字市場，素質頗高，與其溝通協調較具成效，但唯一

美中不足的是人力價格亦高；中國大陸的打字市場，人力價格較低，但是素質參差有別，溝通協調可能繁複，或者需要完整的職前訓練；而印度市場人力成本低，素質又高，在全球數位化浪潮之下，印度已成爲世界全文輸入的最大基地，只不過礙於輸入語言以西方語系或梵語等的限制，依舊令國內數位化單位卻步。因此，國內大部分的全文數位化計畫，多半還是選擇於國內市場交易，選擇一配合優良的廠商，長期合作下去。

輸入時，應以掃描圖檔或影本爲底本，確實遵循先前所制訂之輸入原則輸入。使用的軟體不限，任何能夠進行編碼之文書處理的軟體，如 EmEditor、漢書 98、或是 Windows 內建的記事本皆可。但是繕打之後產生之文字檔案，最好是純文字檔，並且依卷冊號順序命名檔案名稱，以便後續利用時，不受格式之干擾與限制。

以中華佛子佛典協會佛典的人工輸入作業爲例，該計畫以漢書 2000 作爲輸入軟體，而且，爲保留佛教經文之排版特色，以及提供正確的引用註解資訊，格外要求每筆文字輸入除了內文以外，還需仔細記載文字於文內所處之頁數、欄位（下圖一，左上 pooo1a 表示第一頁第一欄）、行數等資訊。



圖一、中華電子佛典協會佛教經文委外人工輸入產出之電子檔

（二）OCR 光學文字辨識系統

OCR（Optical Character Recognition，光學文字辨識系統），是利用掃描器或數位相機等光學輸入設備，獲取印刷文件或手寫於紙上的文字圖片資訊，再以各種模式識別演算法逐一辨識分析文字型態特徵，轉換成電腦可操作的文字編輯，例如美國資訊交換標準碼（American National Standard Code for Information Interchange，簡稱 ASCII code）或是 BIG5(大五碼)，進而轉入資料庫供使用者檢索。

以 OCR 光學辨識而言，中文字的辨識難度遠高於歐美國家的拼音文字。因爲中文字的字數多，且字形架構與字形變化均有其複雜度，故國內的中文 OCR 研究直到最近才邁入實用階段。目前 OCR 的技術開發與研究，在台灣有丹青中英日文文件辨識系統、蒙恬認識王專業系統、全景軟體，在中國大陸則以清華文通和北京漢王最爲著名【註 10】。關於各種 OCR 系統之介紹比較，本計畫去年出版之《報紙期刊全文輸入工作流程參考標準》有詳細的討論與比較，本文不再贅述。

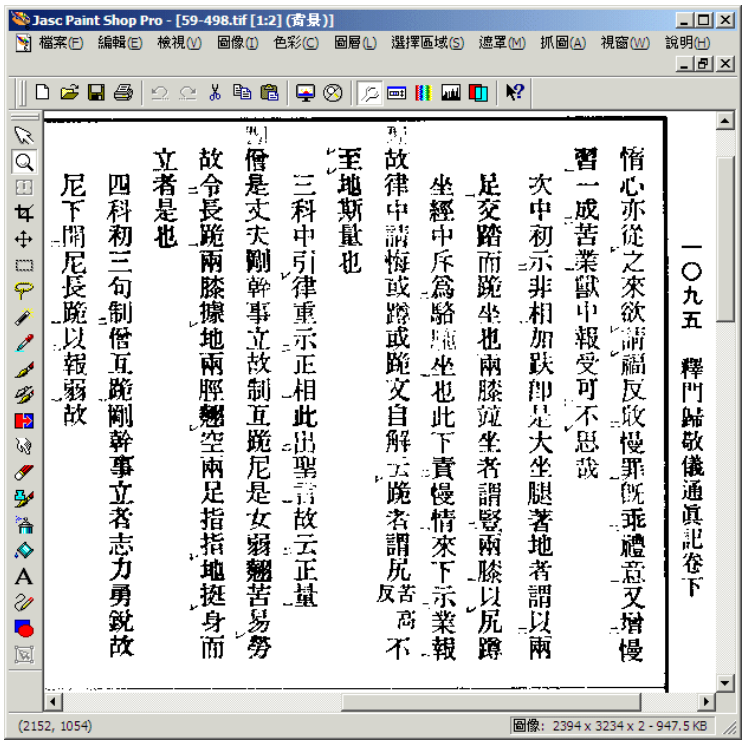
光學辨識易受圖檔清晰程度，影響文字分析結果，故在進行光學辨識前後，有些事前準備與事後修正需特別注意：

1. 降階轉檔

   因爲內文與底色反差較大的圖檔較利於 OCR 光學辨識，因此在進行光學辨識之前，必須先將全彩的圖檔轉成黑白或是灰階，抽離多餘的色彩干擾，有助辨識的正確度。

2. 去除雜點

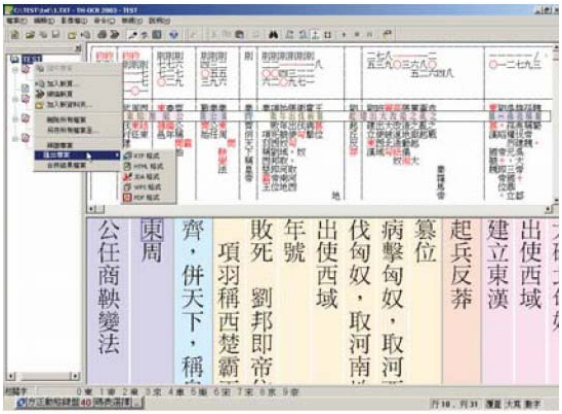
   有些書籍之原文本身有非文字之讀音符號或注釋標記(圖二行間打勾與畫橫槓處)，需以影像處理程式去除雜點，再生一個新的清晰圖檔，才能匯入 OCR 進行光學辨識。



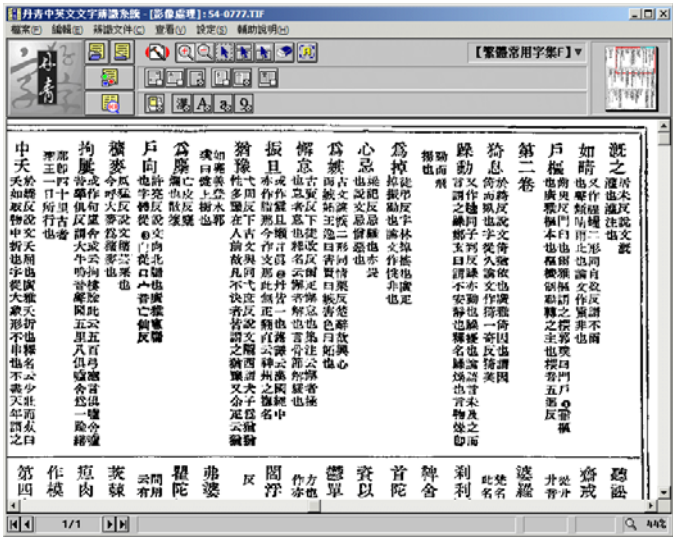
圖二、含讀音符號與雜點之原始掃描圖檔

3. 匯入 OCR 光學辨識系統

將降轉圖檔匯入選用之光學辨識系統（圖三、圖四），執行文字辨識功能，產生純文字檔。



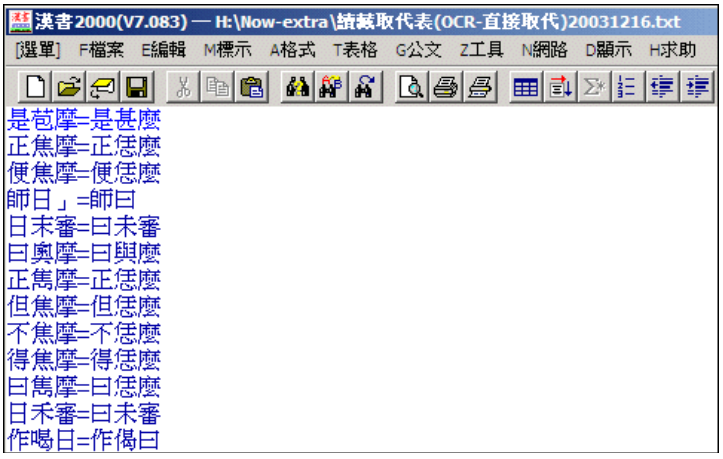
圖三、清華辨識系統介面



圖四、丹青辨識系統介面

4. 字串取代

由於中文字形繁複，相似字形繁多，OCR 無法百分之百分析出正確的內文字形，也會出現錯字。根據錯字出現的前後文，能夠判斷其正確用字，若能再依此整理出一份常錯字串表（圖五），便能以此正確字串快速批次取代 OCR 之常錯字串，減輕逐字校對之不便，並且提升 OCR 產生之文字檔的文字精確度約至 90%。



圖五、常錯字串取代表

三、校對

根據研究與統計，人工打字每人每天約可產出四萬八千字，一般錯誤率爲千分之四至千分之五，換言之，人工輸入的正確率約爲 96%；而 OCR 光學文字辨識系統，依原文清晰度程度，正確率從 90%(內文含有較多符號、異體字、缺字的古籍)到 96%(鉛字印刷工整的近代圖書)不等。由於文字的正確率關係數位化成品之品質，錯誤率盡可能降得越低越好。目前全文數位化單位多將錯誤率的目標訂爲萬分之一，也就是每一萬字只允許出現一個錯字，而達成此理想狀況的唯一方法就是投入心力校對。

現有校對的主要方案是人工校對，而義工校對與檔案比對則是拜網路、科技之賜而新起的發明與工具。將人工校對與後起的校對方案互相搭配執行，可以改善、提升文字的正確率，進而產出更接近原文的高品質電子全文。

然而，提升文字的正確率並非僅止於輸入錯誤的字，中國文字變異以及電腦系統缺字引發之各種缺字、異體字、避諱字等問題，都會於校對過程裡一一浮現，因而關於缺字、異體字與避諱字的解決方法，也是此章節的重點之一。

（一）人工校對

1.一校

逐字逐頁的人工校對是最爲傳統的方案，雖然人力成本、時間、經費的支出都高，但是技術門檻最低，仍爲多數單位執行校對的首選。

通常進行第一次校對的人員爲廠商或是計畫裡負責輸入的人員，輸入完成即進行逐字校對並即時訂正。

2.二校

一校文字檔送回後，計畫人員應立即執行第二次校對。二校的工作方式有別於一校，是以隨機抽樣方式查驗，若達到合格要求，即燒製光碟進行備份；未達要求者則退回廠商或掃描人員處進行修改，並安排二次查驗。目前國內之全文數位化單位，多以錯誤率於萬分之一以下爲最終目標。

（二）義工校對

因爲某些數位化的文本具有特殊的傳播性質，或是計畫單位在執行、推廣數位化工作時佔有特殊優勢，動用大規模無償人力的校對志工團，應運而生。前者如中華電子佛典協會，受惠於民間傳說抄寫或是閱讀佛經有助累積功德之文本特性，因而衍生出「網路校對」機制，即於網路上徵集志工約九百人，投入一人一頁分工的線上校對工作，其程序爲：

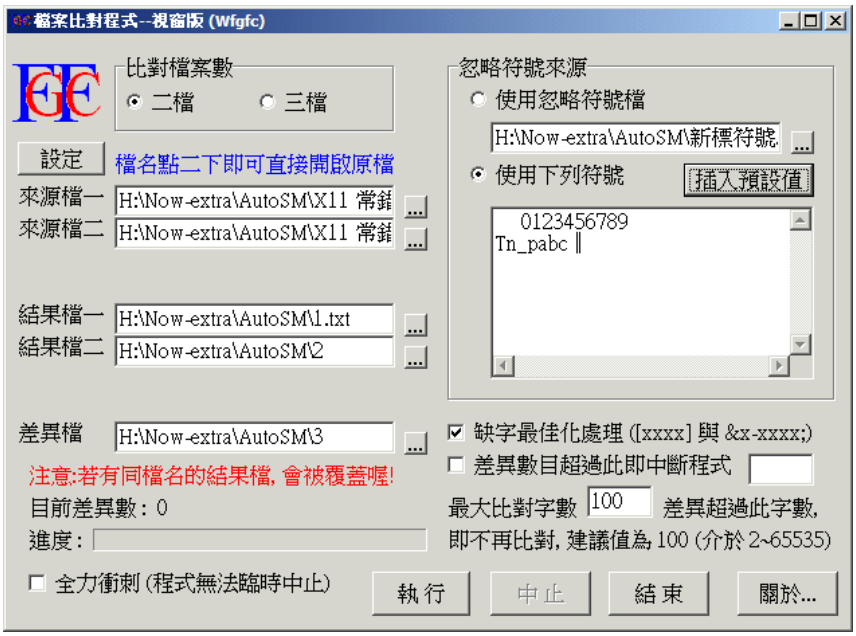
1. 上該單位網站（**http://www.cbeta.org/index.htm**）申請登記。
2. 提領經文之純文字檔與圖檔。
3. 利用看圖校對程式對純文字檔進行逐字校對。
4. 校對完畢即可回傳。

後者爲北京的一個全文數位化計畫，由上行單位直接向下發函分派多所大學之相關領域的教授進行點校。募集大量學者協助數位化工作，一方面可加速工作效率，另一方面也強化數位化成品的品質。

至於義工校對的成效爲何？根據中華電子佛典協會的統計，一般透過 OCR 辨識產生的佛典文字檔，正確率只有 90%，但經過網路義工校對之後，正確率可達 98%。

（三）檔案比對

傳統人工校對，即使四校或十校等重複校驗，還是可能有未發現的漏網錯字。有鑑於此，中華電子佛典協會自行設計了檔案比對程式（如圖），即將相同內容但輸入方式不同（例如人工輸入和 OCR 光學文字辨識）的兩個文字檔，匯入此程式檔案，交由其依照字形、文字編碼對應出兩者的差異處，並予以另存成一個紀錄差異的檔案，節省人工校對必須逐字尋錯的時間，校對人員可直接針對錯誤之處再行查驗訂正。其他計畫單位若欲參照檔案比對之校對方式，或可委請資訊部門協助設計類似程式，以便加快校對速率。



圖六、檔案比對程式畫面

(四)異體字處理程序

異體字是讀音、意義與正體字相同，但寫法不同的漢字，例如「體」的異體字，從教育部異體字字典(http://140.111.1.40/)查詢，可查出「体」、「骵」、「骵」、「躰」……等字(圖七)。異體字的形成時間、區域和原因不盡相同，但在很多的

應用底下，異體字應該被視爲相同的字。不過部份異體字的使用與上下文內容有關，因此異體字處理相當困難【註 11】。

圖七、教育部異體字字典畫面

雖然異體字判讀不易，難取捨，不過現在全文數位化單位多還是以「保留原文」的原則進行輸入。異體字的處理原則，以製作漢籍電子文獻資料庫著名的中央研究院漢籍工作室，他們對於異體字的處理辦法爲：

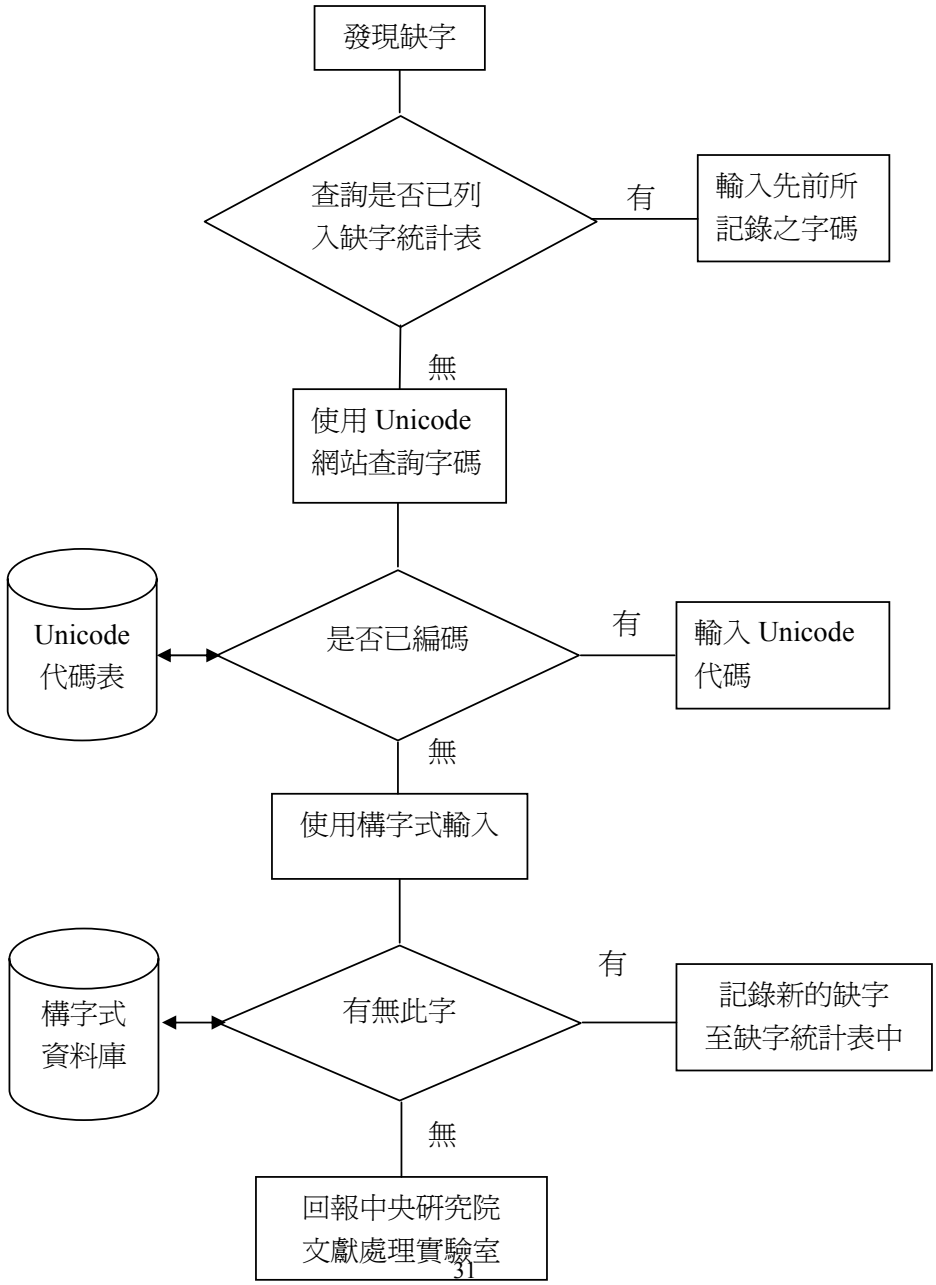
1. 若遇意義、用法相當等同於其正體字的純異體字，則輸入的時候以其正體字表示。
2. 若判讀前後文仍無法確認異體字之義與用是否合於正字，則保留原文樣式示之。
3. 還有一些異寫字，也就是音、義、用完全相同，與異體字間的差異主要是部件的異寫的字，如「福」與「福」皆由部件「示」構成，只不過「示」在「福」中異寫成「礻」。因爲使用者瀏覽的方便性考量，只需找一個通用的參考字形輸入，另外再標示部件的異寫現象即可。但若無法判讀異寫字的用意，則同樣用保留原書字形。

（五）缺字處理程序

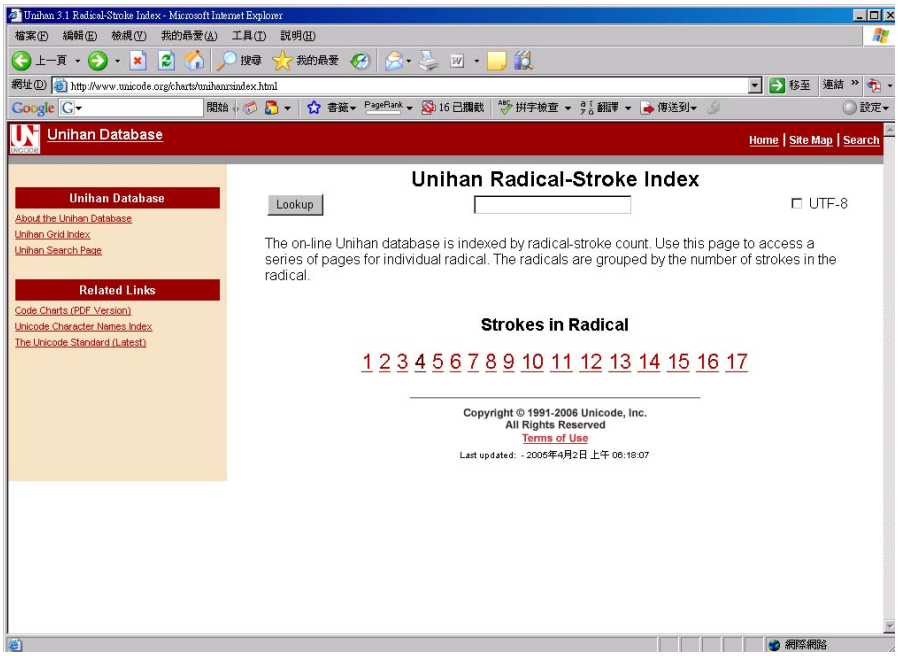
台灣早期的全文數位化計畫多採用 BIG5（大五碼），因爲當時國際通用的編碼在漢字的編輯上不甚完備，國人因而自造編碼系統彌補缺字，但自從字集更爲強大的 Unicode（標準萬國碼）出現之後，不僅新的計畫直接採用 Unicode，歷

史悠久的常設單位也循序將原先的 BIG5 轉換爲 Unicode，而其中有些在 BIG5 裡依然無法顯示的缺字已爲 Unicode 編入，能夠減少缺字減。

國內對於缺字的處理對策，普遍爲從早期沿用至今的「缺字即造字」原則，若輸入時發現未列於 Unicode 的漢字，則交予中央研究院專責處理電腦缺字的文獻實驗室，由其漢字構形資料庫進行造字，產出字形圖檔，並以自立的編碼模式予以編碼，參考流程【註 12】如下：



於 Unicode 網站（http://www.unicode.org/charts/unihanrsindex.html）查詢缺字（圖八），先以部首筆畫進入，之後再以扣除部首之外的筆畫查詢，即可找出所遇缺字是否已編入 Unicode。而有關中研院資訊所文獻處理實驗室研發之漢字構形資料庫的相關介紹與使用說明，可上其網站（http://www.sinica.edu.tw/~cdp/）瀏覽、下載。



圖八、Unihan 資料庫

使用漢字構形資料庫造字，它會將缺字新增成一個圖檔於全文中顯現，可以即時緩解電腦缺字問題；但是瀏覽端必須安裝造字程式始能閱讀圖像，而且只要處理古籍，缺字必定層出不窮，加上造字管理不易，實非治本之計。

現在全文數位化處理缺字問題的態度，趨向不再自行自行造字，也不等待 Unicode 新編缺字，而是改從標記著手，也就是在內文缺字出現處，以標籤與文字描述該缺字的之用與義，再以某種可以純文字表示之組字或者構字方式組合標

示其形。如此一來，可節省造字工時，使用者可輕鬆閱覽，研究者更可循標記得到有關該缺字的原始與詮釋資訊。

（六）套用新式標點

我們現在所使用的新式標點符號，是民國初年五四運動之後，胡適與馬裕藻等人以古代舊式標點符號（句讀號）爲基礎，並吸收西方標點符號所制定的；在此之前，漢文古籍多只以句、讀（。與、）表達語氣的停頓與終了，其餘的疑問、驚嘆語氣無從標明，後人重拾讀本，時常只能望文興嘆，不得其義。

因此，全文數位化除了參照原文依樣畫葫蘆，製作出相同的電子全文外，亦可於電子全文上標示新式標點，協助讀者以熟悉的斷句、語氣來解讀該本，增加資料之研究與傳播的價值。

唯此工作需動用中文句讀相關的學者、研究人員，針對個別文獻深入分析後標以新式標點，並於其後校對，耗費大量人力與時間成本。

文獻處理至此，在經過電腦輸入、校對、以及異體字、缺字等處理過後，已稱得上是一份完整的電子全文，而大部分單位的全文數位化工作，也僅止於此。但是在這強調數位化、資源流通、加值應用的時代，僅止於輸入、校對的電子全文不僅無法滿足應用加值的需求，還因輸入與校對的方法各異，不易互享資料，形成資源浪費。爲了達成資源共享，知識流通，並促進學識研究，電子全文因於校對完成之後，再行以符合國際標準的標記語言進行標誌作業，以提升其研究、應用與分享之價值。

現有通用全球且適用於全文的國際標記語言，爲專爲博物館、圖書館之語文資料設計的 TEI (Text Encoding Initiative)，因其標誌範圍涵蓋結構面與內容面，

又有一個類似圖書版權頁的標頭能夠概述資料，也是一種用來詮釋資料的後設資料，因此本文將標記列入後設資料工作範疇內，有關後設資料之著錄與應用，則於下一章節介紹。

伍、後設資料建置

一、文獻語料的專屬後設資料——TEI

TEI（http://www.tei-c.org/）是一國際性與跨學科性的標準，協助圖書館、博物館、出版者和個別學者以電子文本形式呈現各種文獻和語言學相關的文本，以達到線上教學與利用之便。TEI 利用標準通用標誌語言(SGML)展現電子形式的文本，可不受軟硬體、平台的限制，達到資料交換、再利用之目的。TEI P4 版本(2002)，已可使用可擴充標誌語言（XML）編碼，而 2006 年釋出的最新版本 TEI P5 也新增了缺字的子集，並且補完整體標記系統。

TEI 的文件結構可分成兩大部分：標目(Header)和文本(Text)。由於標目類似圖書文獻的版權頁，除記載原始文本的來源、出處、作者、出版資訊等基本書目資料，也記錄標記者的姓名、身分、標記年份、以及用途，一般也常作爲文本之後設資料使用；文本部分則可標誌內文的層級架構、使用語言，甚至詮釋內容、註釋特殊字句、記錄缺字等。TEI 訂立很多標籤作爲標記時使用的元素，諸如<作者>、<引用>、<新增>、<刪減>、<異體字>、<名稱>、<段落>、<行>、<篇章>……等等。

二、TEI 的核心元素

因爲 TEI 不斷改進創新，現在流通的版本共有 TEI P4、TEI P5 以及 TEI Lite 三種版本。TEI P5 爲 TEI P4 的補完版本，而 TEI Lite 是爲 TEI P4 之選錄輕量版，

內含簡單的 TEI 編碼架構，標籤抽取自大量元素中的易用選集，可滿足 TEI 社群中九成使用者的九成需求，不過較不適用於複雜的文本。有關 TEI 的核心元素，以 TEI P4 的核心元素示之（表 7）。

表 7、TEI (P4)核心元素一覽表

(紅字爲必要元素；綠字爲屬性)

元素名稱	定義	範例
類型 (type)	header 所加進的檔案類型	語料庫(corpus)
建立者(creator)	指出 TEI Header 的建立者	
階段 (status)	說明 header 是新的或是已經改版過	
建立日期(date.created)	指出 header 第一版的建立日期	
更新日期 (date.updated)	指出現在版本的建立日期	
檔案描述 (fileDesc)	描述電子檔案(computer file)本身的完整書目資訊。從這些描述裡，文件的使用者可以得到適當的參考文獻，或當這些檔案由圖書館或檔案館收藏時，館員可以根據此描述建立目錄。這裡的「電子檔案」是指由 header 所描述的整批文件或檔案，而不管是否分別儲存在數個作業系統下。這個類別也可以描述電子檔案的來源資訊。	
標題敘述 (titleStmt)	一組有關作品的標題和負責智識內容者的資訊。	
題名 (title)	一件作品的題名，作品可以是文章、期刊、書籍或叢書；標題包含了別名(alternative titles)或副檔名(subtitle)	Two stories by Edgar Allen Poe: electronic version
層級(level)	標題的書目層級，可以指出是屬於文章、期刊、書籍、叢書或未出版文獻的題名	
類型(type)	題名的類型，依據一些合適的分類標準來分類題名	
作者 (author)	書目索引裡的作者名稱，包含作品作者的名稱，可以是個人的或團體的名稱。在任何書目資料裡是對負責者的主要敘述	Poe, Edgar Allen (1809-1849)
贊助單位(sponsor)	指出贊助機構或組織的名稱	

		主辦單位 (funder)		爲文件或計畫出資的個人、學術機構或組織的名稱	Wellcome Institute for the History of Medicine
		主要建立者 (principal)		負責建立電子檔案的主要研究人員名稱	Dominik Wujastyk
		負責者敘述 (respStmt)		提供負責文件內容、版本、紀錄或叢書負責人的敘述。通常作爲當作者或編輯者等元素不足以描述或沒有描述時的補充說明元素	
		負責內容(resp)		以短語的方式描述負責者智識上的工作內容	由---編輯(compiled by)
		姓名(name)			James D. Benson
			類型(type)	以短語的方式對物件類型命名	
	版本敘述 (editionStmt)			一組有關文件某版本的資訊	
		版次(edition)		描述某一文件某一版本的特殊性	第二版草稿，較前版大爲擴展、改版和修正
		負責者敘述 (respStmt)		提供負責文件內容、版本、紀錄或叢書負責人的敘述。通常作爲當作者或編輯者等元素不足以描述或沒有描述時的補充說明元素	
		名稱(name)			George Brown
			類型(type)	以短語的方式對物件類型命名	
		負責內容(resp)		以短語的方式描述負責者的工作內容	由---全新註釋
	大小 (extent)			描述電子檔案儲存在某些媒介裡的約略大小，須以合適的單位表示	(1) 4532 bytes (2) 3200 句
	出版描述 (publicationStmt)			一組有關電子或其他文件出版或發行的資訊	
		出版單位 (publisher)		負責出版或發行書目項目的組織名稱	牛津大學出版社
		發行者/單位 (distributor)		負責文件發行的個人或其他代理人的名稱	Oxford Text Archive
		權威人士 (authority)		非出版者或發行者，但負責使電子檔案可通行的人或機構名稱	James D. Benson
			出版地點 (pubPlace)	一個書目項目出版的地點名稱	牛津
			地址 (address)	提供出版者、組織或個人的郵件或其他地址	21 High Street, Wilmslow, Cheshire M24 3DF
			識別碼	用於識別一個書目項目的標準式或非標準式的號碼	0-19-254705-4

			(idno)			
				類型(type)	識別碼的類型，例如 ISBN 或其他標準序號	<idno type='ISBN'>
			取用權 (availability)	提供一份文件的取用權的資訊，包括使用或發行限制、著作權限等。		James D. Benson
			日期(date)			1989
				曆法(calendar)	指出時間表示的系統或曆法	
				格式(value)	以標準的格式表示日期，通常以 yyyy-mm-dd 表示	
				精確度 (certainty)	描述日期的精確程度	
	序號敘述 (seriesStmt)			有關序號的一組資訊，通常在出版上使用		
		題名 (title)		一件作品的題名，作品可以是文章、期刊、書籍或叢書；標題包含了別名(alternative titles)或副檔名(subtitle)		Machine-Readable Texts for the Study of Indian Literature
			層級(level)	標題的書目層級，可以指出是屬於文章、期刊、書籍、叢書或未出版文獻的題名		<title level="S">
			類型(type)	題名的類型，依據一些合適的分類標準來分類題名		
		識別碼(idno)		用於識別一個書目項目的標準式或非標準式的號碼		1.2
			類型(type)	識別碼的類型，例如 ISBN 或其他標準序號		<idno type="vol">
		負責者敘述 (respStmt)		提供負責文件內容、版本、紀錄或叢書負責人的敘述。通常作爲當作者或編輯者等元素不足以描述或沒有描述時的補充說明元素		
		負責內容(resp)				由---所編
		姓名(name)				Jan Gonda
			類型(type)	以短語的方式對物件類型命名		
	附註敘述 (notesStmt)			收集有關文件資訊的補充說明以增加書目描述的其他部分		
		附註 (note)		包含附註或註釋(annotation)		歷史評註由 Mark Cohen 提供
			類型(type)	描述附註的類型		
			註解者	說明負責註釋的人員，例如：作者、編		

			(resp)	輯者、翻譯者等	
			地點(place)	指示附註出現在來源檔案的位置	
			下錨處(anchored)	說明是否複製的文件爲附註顯示了正確的參照位置	
			標的結尾處(targetEnd)	如果附註沒有包含在文件裡，則說明附註加接範圍的終點	
	來源描述 (sourceDesc)			提供有關電子文件來源的複製文件的書目描述	
		書目資料(bibl)		包含書目資料的粗略描述，其中的次類不一定要明顯標記(tagged)	The first folio of Shakespeare, prepared by Charlton Hinman (The Norton Facsinile, 1968)
		書目結構(biblStruct)		包含結構化的書目引用，其中只會出現有關書目的次元素並以特定的順序出現	
		完整書目(biblFull)		包含書目資料的完整結構，其中會出現TEI 檔案描述的所有組成成分	
		條列書目(listBibl)		將書目引用以條列方式表示	
		腳本描述(scriptStmt)		包含對口語檔案的詳細腳本的引述。使用在電子檔案的來源文件是口語檔案時。	
		記錄描述(recordingStmt)		描述口語檔案轉寫時的紀錄。使用在電子檔案的來源文件是口語檔案時。	
		紀錄(recording)		描述口語檔案來源的錄音或錄影事件，影音來源可以從大眾傳播上取得。	U-matic recording made by college audio-visual department satff, available as PAL-standatd VHS transfer or sound-only casssette
			類型(type)	說明錄音/影的種類	<recording type='video'>
			時長(dur)	說明錄音/影的時長	<recording dur="30 mins">
		設備(equipment)		提供錄音/影設備或媒體的詳細敘述，這些聲音或影像的紀錄是作爲口語檔案的來源。	數位錄音自 FM 廣播

		廣播節目 (broadcast)	描述作爲口語檔案來源的廣播節目	主題：Interview on foreign policy 製作單位：BBC Radio 5 主持人：Robin Day 受訪者： Margaret Thatcher 節目名稱：The World Tonight 附註：First broadcast on 27 Nov 1989
編碼描述 (encodingDesc)			描述電子檔案和其來源之間的關係。這個類別詳細描述了在轉寫過程中，文件如何標準化、編碼者如何解決原始文件內歧義的問題、應用了何種層級的編碼或分析方法	
	計畫描述 (projectDesc)		詳細描述電子檔案編碼的目的，以及電子檔案集結過程的相關資訊	Texts collected for use in the Claremont Shakespeare Clinic
	取樣宣告(samplingDecl)		以散文的方式描述建立語料庫時，文本取樣的原理和方法	文件取樣是從開頭算起兩千字
	編輯宣告(editorialDecl)		描述當爲文件編碼時，所使用的編輯原則與實作	
		修正(correction)	說明在何種情況下以及如何修正文件	拼字錯誤檢查是藉由 WordPerfect spelling checker 來執行
		程度(status)	指出應用在文件上的修改程度	<correction status="unknown">
		方法 (methond)	用於指出文件內標明更動的方式	<correction metnod="silent">
		標準 (normalization)	指出轉成電子檔案的原始文件內，施行標準化的範圍	藉由韋氏第九版 Collegiate 字典將字轉成標準美式拼字 (Modern American spelling)
		來源 (source)	指出任何施行標準化的權威檔	<normalization source="w9">
		方法 (methond)	用於指出文件內標明標準化的方式	<normalization method="silent">
		引號(quotation)	在編輯時，原始檔內引號的應用	所有開引號由參考實

				體(entity reference) ODQ 表示；所有閉引號由參考實體 CDQ 表示
		引號(marks)	指出引號在文件內是否被保留作爲內容的一部份。	<quotation marks="all">
		形式(form)	說明引號在文件內指示功能的運作方式	<quotation form="std">
	連字號(hyphenation)		摘要敘述原始	
		行尾(eol)	說明文件裡行尾的連字號是否被保留	
	斷詞(segmentation)		描述文件斷詞的原則，例如是依句子、聲調單位或字素圖層等	
	標準值(stdVals)		當使用標準化的日期或數字表示時，指出使用的格式(format)	
	詮釋(interpretation)		描述除了轉譯以外，任何加在文件上的分析或詮釋資訊的內容	第四部份的言談分析是以手寫方式加入，還未被檢查
標籤宣告(tagsDecl)			詳細描述應用在 SGML 文件裡標籤的	
	翻譯(rendition)		提供有關一個或多個元素欲轉成樣式的資訊	
	標籤使用方式(tagUsage)		文件內特定元素的使用資訊	只用來加標籤在複製文件裡的斜體字
		(gi)	標籤所標示的元素名稱(一般辨識名稱)	<tagUsage gi="p">
		(occurs)	文件內元素的出現次數	<tagUsage occurs="28">
		識別(ident)	在全球識別屬性(global id attribute)擁有區辨值的文件內，元素的出現次數	<tagUsage ident="321">
		翻譯(render)	指出「翻譯<rendition>」元素的識別，而翻譯元素是定義元素是如何被翻譯的	<tagUsage render="style1">
參照宣告(refsDecl)			說明如何爲這份文件建立正式的參照(canonical references)	
	檔案類型(doctype)		說明在參照宣告內的檔案類型	<refsDecl doctype="TEI.2">
	階梯式(step)		指出由階梯式方法定義的正式參照的一個構件	
		參照單位(refunit)	在正式參照中，給予這步驟所識別出的單位(書、章、詩篇 canto、詩節 verse)命名	<step refunit="chapter" >
		長度(length)	指出參照構件的固定長度	<step length="3" >
		定界(delim)	提供跟隨在參照構件後的定界線(delimiting string)	<step delim=":" />

			起點(from)	指出在正式參照裡，藉由此步驟參照的起點	<step from="DESCENDANT" (1 DIV2 N %2)" />
			終點(to)	指出在正式參照裡，藉由此步驟參照的終點	<step to="DITTO"/>
		里程碑式(state)		指出由里程碑式方法定義的正式參照的一個構件	
			版本(ed)	指出里程碑方式應用在何種版本上	<state ed="first"/>
			單位(unit)	指出在這里程碑上什麼部份被改變了	<state unit="page"/>
			長度(length)	指出參照構件的固定長度	<state length="2"/>
			定界(delim)	提供跟隨在參照構件後的定界線(delimiting string)	<state delim="."/>
	類別宣告(classDecl)			以一或多個分類法定義檔案內的分類碼	
		分類法(taxonomy)		定義類別來分類文件，可以不明顯地藉由書目索引，或明顯地採結構化分類	
		類別(category)		包含個別的描述類別，可能在使用者定義的分類法內，套合在 superordinate 類別裡。	
		類別描述(catDesc)		描述文件分類裡的一些類別，可以以短文的形式或藉由使用在 TEI 正式檔案描述(textDesc)的狀況參數(situational parameters)描述	報紙報導(Press Reportage)
	特徵系統宣告(fsdDecl)			識別出特徵系統宣告，其宣告包含對特徵結構的特定類型的定義。	
		類型(type)		指出記錄在 FSD 內特徵結構的類型。這將會是至少一個特徵結構裡的類型屬性值。	<fsdDecl type='myA1'>
		特徵系統宣告(fsd)		指出包含特徵系統宣告的外部實體。在檔案的 DTD 次集合的實體宣告必須和系統內具有檔案的實體名稱相關連。	<fsdDecl fsd='myFeatures'/>
	韻文宣告(metDecl)			當韻文的型態是以結構化的元素屬性表示出來時，此元素記錄被運用以顯示韻文型態(metrical pattern)的符號。	
		類型(type)		指出符號是否表達了抽象的韻律形式(metrical form)，真正的韻律展現(prosodic realization)，或者韻律架構(rhyme scheme)，或一些相關的組合。	<metDecl type="METREAL">
		型態(pattern)		指出規則性的表示方法來定義對符號的合法值。	<metDecl pattern="((1|0)+\ |?/?)*">

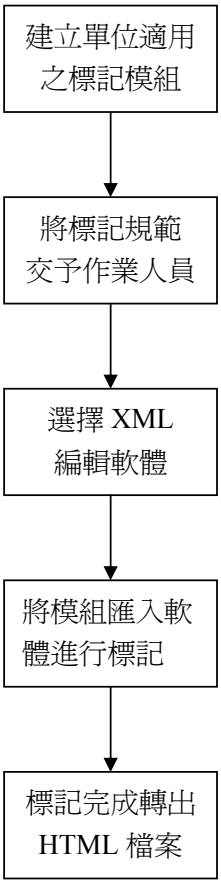
		象徵(symbol)		記錄在韻文符號內，特定字串的重要性，可以明顯表示或藉由在同一個 metNotation 內的象徵元素	韻律顯著(metrical prominence)
			內含值(value)	指出被紀錄的字元或字串	<symbol value="1">
			終點(terminal)	指出象徵符號是由其他符號定義(terminal=N)或以描述法(prose)定義(terminal=Y)。	<symbol terminal="y">
	文件變體編碼(variantEncoding)			宣告變體文件的編碼方法	
		方法(method)		指出變體裝置的編碼方法	
		地點(location)		指出裝置(apparatus)是隨檔案運作出現或在檔案運作外圍出現	
文件描述 (profileDesc)				提供一份文件非書目部分的詳細描述，特別是所使用的語言和次要語言，文件建立的情況、參與者以及其背景。	
	建立(creation)			有關一份文件建立的資訊	<date value="1992-08">1992 年 8 月</date> <rs type="city">新墨西哥州，Taos 城</rs>
	使用的語言(langUsage)			包含一組有關描述文件的主要語言、次要語言、登錄者、方言等的資訊	
		語言(language)		描述文件內單一的語言或次要語言	加拿大商用英語(Canadian business English)
			書寫系統宣告(wsd)	爲包含書寫系統宣告的實體，用來顯示文件上的語言	<language wsd="wsd.en">
			使用法(usage)	指出文件內使用某語言的冊數所佔的約略百分比	<language usage="20">
	文件類別(textClass)			一組描述文件性質或標題的資訊，可以藉由標準化的分類架構來描述，例如 thesaurus 等	
		關鍵詞(keywords)		含有關鍵詞或短語的列表，用來指出一份文件的主題或性質。	
			架構(scheme)	定義關鍵詞時所依據的控制詞彙	<keywords scheme="lcsh">
		分類碼(classCode)		依據一些標準分類系統爲文件訂定分類碼	005.756
			架構	指出使用的分類系統或分類法	<classCode

			(scheme)		scheme="ddc19">
		類別參照(catRef)		一些分類學上所定義的一個或多個類別	
			目標(target)	指出有關的類別	<catRef target="b12 b15">
			架構 (scheme)	指出定義類型集所依據的分類架構	<catRef scheme="brown"/>
	文件描述(textDesc)			提供依據狀況參數(situational parameters)表示的文件描述	
	背景描述(settingDesc)			描述語言互動(language interaction)發生時的背景，可以是散文式描述或利用一系列元素來描述	
	筆跡(handlist)			包含描述來源筆跡的元素列表	
改版描述 (revisionDesc)				允許編碼者提供在電子檔案發展過程中，檔案變動的歷史。改版歷史對版本控制(version control)和解決文件歷史的問題都很重要。	
	變更(change)			摘要描述一份多位研究者所共有的電子文件的變更或改版的內容	
		日期(date)		以任何形式表示的日期	5/25/91:
			曆法 (calendar)	指出時間表示的系統或曆法	
			格式(value)	以標準的格式表示日期，通常以 yyyy-mm-dd 表示	
			精確度 (certainty)	描述日期的精確程度	
		負責者敘述 (respStmt)		提供負責文件內容、版本、紀錄或叢書負責人的敘述。通常作爲當作者或編輯者等元素不足以描述或沒有描述時的補充說明元素	<name>EMB</name> <resp>ed.</resp>
		項目(item)		包含一列表的一個組成部分	檔案格式更新

(發表於 2001 年 6 月，URL: http://www.tei-c.org/P4X/元素名稱、定義與範例是由技術發展分項計畫後設資料工作組翻譯)

三、TEI 標記實務

（一）標記作業程序參考

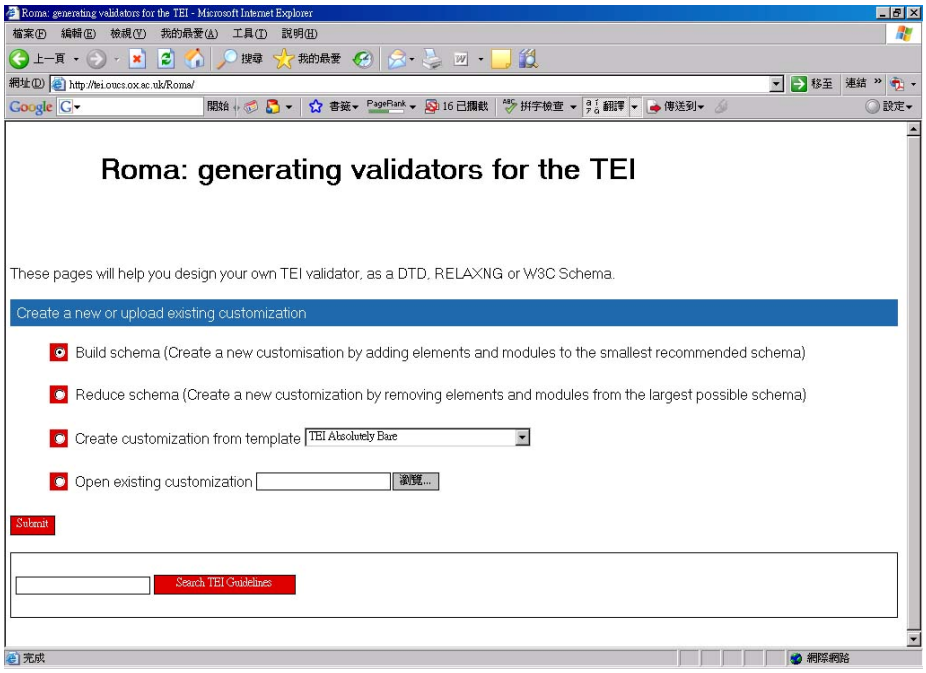


（二）建立適合計畫的標記模組

TEI 標記語言是以上百個描述元素（標籤）組成，由於標籤數量眾多且規則繁雜，要熟悉並學會使用所有標籤實爲難事，況且多數計畫單位的文本不需使用所有的標籤，反而是依照不同文本特性選用不同標籤，集合成能夠標記計畫文本

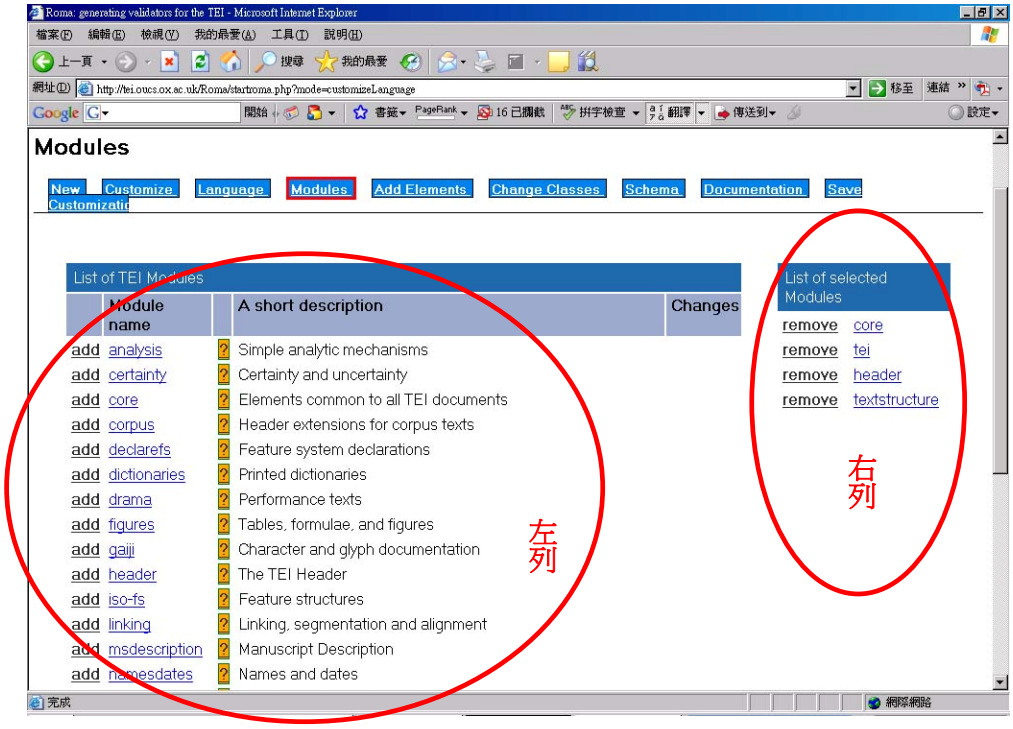
的標籤集合。

爲了方便計畫單位建立或者參考他人所選用的標籤，網路上有自發研究的工程師設計了一個線上系統 Roma(http://tei.oucs.ox.ac.uk/Roma/，圖九)，提供所有人士免費登入，於其中創造、修正，或是分享自己的 TEI 標籤集合。



圖九、Roma 首頁

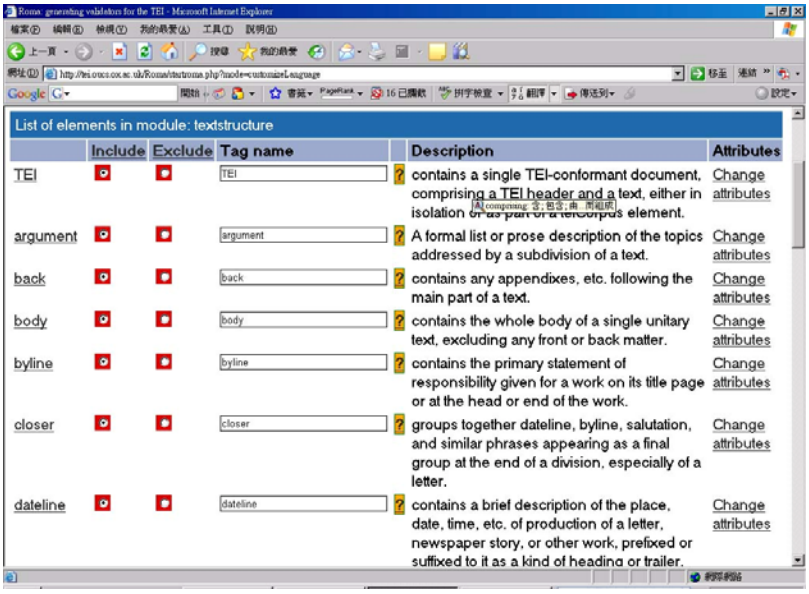
在這個系統裡，點選最上列的 module 選項，可看見標籤依據其功能與用途，歸納條列成核心、文件結構、表頭，以及適合詩歌、語言分析、散文、戲劇等各類標籤集合（圖十左列），如 drama(戲劇)、figures(圖表)、gaiji(缺字／外字)、corpus(語料)……等，並賦予這些標籤集合一個專有名詞稱做 TEI 模塊(TEI Module)。



圖十、Roma Modules 選擇介面

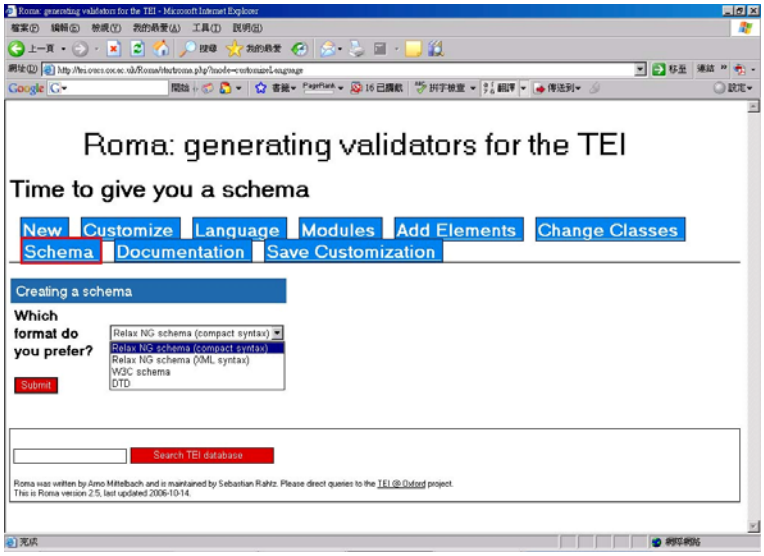
使用者根據手邊文本的特性與複雜度，只要點選左列模塊名稱旁的 add（增加）字樣，就能新增模塊至右列的已選模塊裡，若要刪去，點選 romove(移除)字樣就能刪除，這些選擇的模塊會組成 TEI 標記的子集，又稱作 TEI 模組(TEI Schema)。爲確保所有使用者創造之模組符合 TEI 標準結構，不失資料於國際、館際交換流通之通用標準特質，此系統還特別將 core、tei、header、teistructure 四個主要構成 TEI 結構之模塊強制加入右列清單，使用者無法刪除。

TEI 除了模塊的組合自由外，使用者還可針對模塊裡的標籤進行屬性的修改，只要點選模塊名稱本身，例如 teistructure，就能進入該標籤集合頁面進行標籤的新增與刪除，及其屬性的定義與修正（圖十一）。

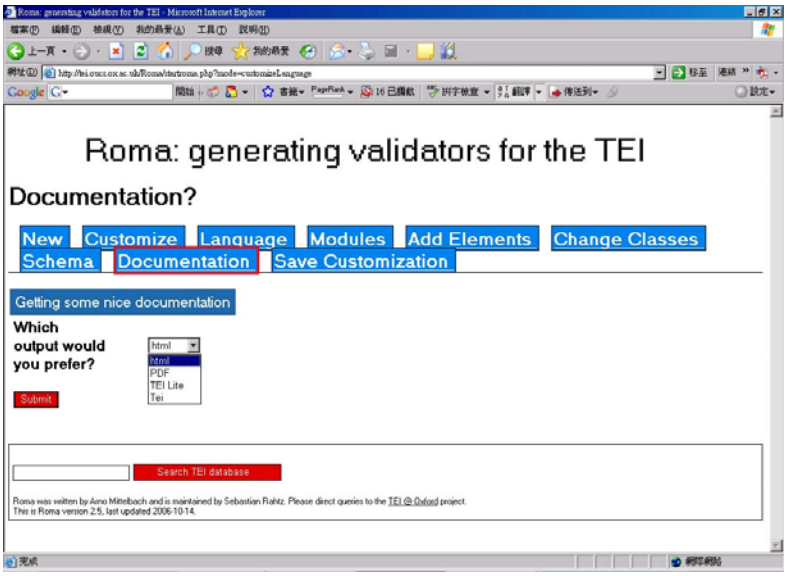


圖十一、teistructure module 修改畫面

待模塊選擇完畢，且內含標籤的屬性都定義完畢後，點選畫面最上方工具列的 schema 選項(圖十二)，可以進入選擇轉出模組格式的畫面，系統會以選擇之格式輸出 TEI 模組的 DTD(Document Type Definition，文件格式定義)；而點選 documentary 選項(圖十三)，則能產出文件定義之文字說明檔。



圖十二、建立 TEI Schema 畫面



圖十三、建立 TEI Documentation 畫面

(三) 標記人員

標記分爲層級標記與內容標記，前者標記內文的結構如題名、作者、段落、行句……等，這些屬於比較制式化、不需專業學識判斷的基本標記，可交由輸入、繕打的一般人員於輸入文字時一併處理。

而後者的內容標記，因爲牽涉到內容的判讀與辨識，例如缺字、異體字、訛字、或是特殊註釋等，則需由專業人士（例如計畫成員或經過訓練的人員）使用參考工具如字典、辭典等工具書進行考究，才能正確無誤標誌。

（四）使用軟體

1. UltraEdit

UltraEdit(http://www.ultraedit.com/)是一種純文字編輯器，可以使用它來編輯 XML 檔，並在 UltraEdit 裏設定呼叫 XML Parser，隨時做 XML 的語法檢查，

方便尋找、修正錯誤。其特色與功能如下：

(1) 提供超強的文字檔編輯、預視、列印功能。

(2) 提供直接編輯十六進位碼功能。

(3) 可編輯 HTML 檔案，以彩色顯示 HTML 標記，方便網頁編輯。

(4) 新版在畫面左邊提供快速檢視欄框，無論切換編輯視窗或檔案都很容易。

(5) 新增的 project（計畫）功能，可以把數個文字檔合成一個「計畫」，方便一次編輯數個彼此有關連的文件。

2.Oxygen

Oxygen (http://www.oxygenxml.com/index.html)是一種 XML 編輯器。其特色是能夠匯入自訂模組，並在定義的模組內編輯、檢查語法，錯誤語法會一一條列於下方視窗內，點擊條目上方的編輯畫面就會顯現相對的錯誤處，方便修訂。此外，也能靈活轉換成 HTML、PDF 與 PostScript。

(五)轉出 HTML 格式

標記完成後並且語法檢查無誤之文字檔，可轉出 HTML 格式儲存，並搭配適宜之 Style Sheets，即可於網頁上呈現出型式與內容兼具的電子全文。

陸、資料庫與其他應用

圖書文獻製成電子全文後，舊有的厚重形式變化爲虛擬無形的數位檔案，文本內容擴展爲可無限延伸的超文本，因爲形式的跳脫與內容的提升，數位化釋放了圖書文獻受制於圖書館藏的舊規，進而啓發知識應用的無限可能。就電子全文現有的應用發展來說，有建置全文資料庫、製作電子全文光碟、以及開發週邊產品如字典、辭典或是百科全書。

一、建置資料庫

全文資料庫是全文數位化單位對於電子文本的基本應用，用意類似實體圖書館的數位化，意即將館藏文獻都搬到網路上，增加資源的流通性，並透過容易與其他媒體結盟的電子形式，提升資源的可用性。完備的全文資料庫應具備下列功能：

(一)全文檢索功能

一般圖書館或是資料庫的文字檢索功能止於書目、作者名以及關鍵字的搜尋，這樣的功能邏輯來自於以書找文，將以文找書的狀況排除在外。全文檢索能夠提供使用者在只知部分內文的情況下，依舊查出其作者出處，並且並列具有相似內文的文獻資料，使用者可用以比較、分析或統計。

(二)層級檢索功能

多數圖書文獻都有經、史、子、集或是宗、冊、卷、件等層級關係，而此層級資訊不會呈現於一般的書目檢索或是全文檢索結果裡，因而建立層級檢索功能，一方面還原圖書的知識架構，一方面也協助使用者認識文獻間的層級關連。

(三)權威控制功能

權威控制的方式主要用於建立人名、地名、機關名及主題等標目，提供使用者檢索同質異名，例如馬英九、小馬哥，陳水扁、阿扁，或是相同款目的所有資料，以建立檔案資料的聚集及一致，控制並提高檢索精準率。

(四)影像連結功能

於全文資料外附上原始圖檔的連結，將電子全文資料庫加值爲電子全文影像資料庫，能使使用者同時閱讀原書內文以及觀看原書的編排格式，就像瀏覽原書一般。

二、製作成品光碟

光碟具有保存與流通兩種功能，尤其對於需要傳播宣揚的資訊來說，成本低、攜帶方便、讀取容易的光碟，絕對是網路之外不可或缺的傳播工具。以國內的全文數位化計畫來說，中華電子佛典協會每年固定壓制最新版本的電子佛典集成光碟（圖十四），廣發大眾。

圖十四、中華電子佛典協會電子佛典集成光碟

三、發展相關百科、字(辭)典

在全文數位化過程中，校對時記錄之缺字、異體字，標記時詮釋、註解的詞句語法，以及權威控制的詞條、字彙，都可以加值成提供資訊與知識的線上工具書。

柒、數位資料保護

資訊時代，透過無所不侵的網際網路，數位資源唾手可得，國家珍貴文物與個人私人財產很容易爲不法利益盜用。全文資料雖未像數位影像容易被不法商人侵權使用，牟取暴利，但是引用他人文字卻未標明出處，仍會觸犯剽竊、抄襲等法律規範。爲防制電子全文著作權或智慧財產權之受損，其數位資料之保護方式，有以下幾種爲部分電子全文資料庫使用之方案：

一、 限制使用端 IP

有些圖書管或是博物館會採用付費或是限制使用者 IP 的方式，管理館藏文獻的資源流通。選擇以此方式來限制資源流向的機關單位，通常是其館藏具有特殊、珍貴性質，或是牽涉著作權、授權等法律問題。

二、 創用 CC 授權條款

Creative Commons(創意公用授權條款，簡稱創用授權)是來自美國的一種著作權授權條款，透過開放授權的方式，釋出一部分的著作利用權限（通常是非營利利用的部分），僅包留一部分的著作利用權限（通常是營利利用的部分），讓著作可以在一定的範圍內被公眾在授權的範圍內利用。

至於實際授權的方式，作者可以在「是否要求標示作者名稱」、「是否允許商業利用」、「是否允許創作衍生著作」、「衍生著作是否應以原授權條件提供授權(share like)」等選項，依其自由意願選擇適合的方案釋出著作，並且透過各種簡

易的圖示（圖十七）的搭配使用，使利用人得輕易辨識該著作授權方案的種類，以利著作之自由流通【註 13】。

圖十七、創用 CC 核心授權條款

Creative Commons 與一般網站宣稱 All Rights Reserved 之不同，在於 All Rights Reserved 保留所有權利，除經合法授權或合理使用，不然就有侵害著作權

之問題，而 Creative Commons 則是透過授權條款釋出的部分權利，使任何人都可依據授權條款進行利用。

三、 無償使用

相較於使用者付費與限制使用者 IP 之數位資料保護措施，也有一些機構所藏文獻基於知識共享、服務大眾之心態，免費提供資訊於網路傳播，以達無遠弗屆之宣揚、告示目的。使用無償的全文資訊雖無需負擔費用，並且可以自由瀏覽使用，但在引用或是引述的時候，還是需要注意標明出處，以示尊重。

捌、設備與成本分析

數位化本是一項所費不眥的浩大工程，除了有形的、可以計算的設備、軟體費用，隱形的、容易忽略的場地租用、水電費用與設備維修，還有無形的、難以計算的勞務費用如人力訓練費用，以及全文數位化特別需要文史相關研究者標點、標記的所付出之知識成本與時間。

相較於其他物件的數位化工作，在全文數位化過程裡，無形知識的付出占所有工作的大部分，資料有限，經濟成本難以考量，故本書僅就勞務費與設備費，以及部分委外支出，提供選擇方案與可能花費，作爲數位化成本之基本參考。

一、 設備選擇考量

此部分，我們針對全文數位化工作所需之相關器材設備，進行選購說明。主要設備包括影像掃描器以及電腦的軟硬體設備。

（一）掃描器的選擇

市面上掃描器分有桌上型平台掃描器、桌上型自動進紙式掃描器、桌上型無邊縫掃描器、以及滾筒掃描器。如果欲進行全文數位化的單位想自行掃描，建議使用自動進紙式掃描器，並搭配自動編號存檔之功能，能夠有效節省掃描時間，簡化掃描工作。

（二）電腦硬體的選擇

市面上的電腦可分成兩類，一種是針對商務或一般文書處理的個人電腦（PC），另一種則是爲繪圖出版等作業所使用的麥金塔系列（MAC）。由於全文

數位化工作大多處理的是輸入、校對以及標記等文書處理的工作，所以選購一般個人電腦即可。

而在個人電腦的主機選購上，雖然多數工作僅止於文書處理，但由於可能需要同時作業多個視窗，或是瀏覽掃描圖檔，在部分設備的選擇上，還是需要特別挑選，例如：

1. 隨機存取記憶體（Random Access Memory，RAM）的容量關係軟體的順暢執行。一般來說，目前的個人電腦至少要有 512MB 或 1GB 以上的 RAM。
2. 顯示卡能夠呈現色彩層次的細膩度，越需要影像處理的工作，其電腦的顯示卡就需挑選效能越高的。目前市售的顯示卡，有附較佳繪圖功能的顯示卡，價格大約都會在 3000 元以上。若爲一般的文書處理，可採買較爲低階的顯示卡，或是直接使用主機板內建的顯示卡即可。
3. 儲存檔案所使用的硬碟，容量越大越好，並且需配合妥善的儲存規劃。此外，也可購置外接式硬碟，異地備援。

（三）電腦軟體的選擇

電腦軟體方面，由於各項作業所需軟體不一，以下就不同作業流程，分項介紹。

1. 數位影像處理：有關色彩校正、色彩管理之工作，可選用 Adobe Photoshop 系列軟體(http://www.adobe.com/tw/products/photoshop/)。
2. 繕打輸入：一般來說，只要可以編輯字碼之軟體都可使用，而推薦使用之軟體爲國人開發之「漢書」，其好處爲記憶體佔用少、作業速度快、可讀取大容量的檔案，能與 WORD 連結。
3. OCR：欲知 OCR 光學辨識系統之介紹，可參照數典計畫內容發展分項計畫出版之《報紙期刊全文輸入工作流程參考標準》，內含多種廠牌之辨識系統比較，其中能夠同時辨識中英日三種語言的「丹青文件辨識系

統」，是多數已執行全文數位化之單位建議使用之軟體。

4. 標記軟體：標記所使用之軟體，和輸入時使用的軟體大同小異，原因多半是現在的文書處理軟體多可同時處理純文字以及 XML 標記，故推薦使用之軟體同樣爲漢書與 UltraEdit。此外，Oxygen XML Editor 亦爲標記時可使用之軟體，它能夠匯入單位所需之標記基模(schema)，快速檢查、追蹤編輯之 XML 標記是否符合規定。
5. 其他：由於全文數位化的主要工作是進行大量繁複的輸入、校對以及標記，所需人力成本極高，爲簡化勞務加速工時效率，有些計畫或單位會自行研發有助改善流程之相關系統軟體，例如看圖校對、檔案管理、檔案比對、文書處理、標記轉換……等，這些軟體除可請計畫內資訊部門協助開發，還可以委託外部廠商製作，只不過軟體委外時，最好要求設計師將系統軟體以開放碼處理，以便日後修改有方。

二、 成本分析

全文數位化工作之成本包括：材料費、勞務費以及經費。材料費爲數位化工作所需使用之耗材費用；勞務費爲工作人員的薪資；經費則爲機器設備與軟體之費用，包含折舊費用，以及場地的修繕、租賃、水電、雜支等費用。計算數位化費用時，應依上述羅列項目，一一統計核算。而有關掃描、繕打之市價行情，長期進行大宗漢籍文書全文數位化的中華電子佛典協會，提供下列參考價格：

表 8、文書掃描與文字繕打市價

作業項目	價格
掃描	委外掃成 300dpi 黑白影像檔，價格爲 NT1.5 元/頁(A4)
輸入	台灣：NT50 元/1000 字
	中國大陸：NT25 元/1000 字 NT15000 元/冊

玖、結語

過去，台灣在漢籍全文的數位化領域，不論質量，皆處於獨步全球的地位；然而現今，中國大陸也逐漸進行相同的全文數位化工作，產量方面的表現尤其驚人，並非台灣所能匹敵。有鑑於此，國內長期從事全文數位化的單位、機構，在質與量難以兼得的強況下，莫不選擇投入大量時間精力與智慧腦力，以製作出高水準、高品質的全文資料庫爲計畫目標，而非汲汲營營趕製產量。此數位化工作流程指南，亦抱持高品質高水準之觀念宣導，在全文數位化程序上，著重文件標記之詮釋資料的介紹。如前面後設資料之章節所述，文件標記是提升電子全文研究價值與應用廣度的必備利器，而現有文件標記系統裡，又以本文所介紹之 TEI 最爲人文學者以及圖書、博物館界所推崇，希望所有已經或是即將進行全文數位化工作之計畫，都能將之納入標準作業程序，有利數位化成品日後之再利用。

數位典藏國家型科技計畫—內容發展分項計畫，在漢籍全文主題小組召集人杜正民先生的協助下，自成立以來，一直致力推動台灣漢籍全文數位化之研究發展的成果共享，現已開設兩梯次的 TEI 實作工作坊，用以訓練培養台灣的標記好手；而在未來的規劃裡，除希望持續推廣 TEI 的在地化與中文化，也期盼開設缺字處理相關課程或講座，以及其他與全文數位化有關研究與技術的國內（外）研討會，使台灣內的漢籍全文數位化環境，更臻完善，更具競爭力。

最後，本數位化工作流程之撰寫，有賴於所有全文數位化先鋒之經驗累積與分享貢獻，在此，特別感謝漢籍全文主題小組召集人，同時身爲法鼓山中華佛學研究院圖書資訊館館長—杜正民先生—的撥冗指導，以及中華電子佛典協會吳寶原組長、中央研究院歷史與語言研究所漢籍工作室聯絡人李芳瑩小姐，屢屢於忙碌工作之中不厭其煩的提供資料與協助。並且誠摯希望這本蘊含台灣首屈一指的全文數位化計畫經驗的參考手冊，能夠吸引更多計畫單位或人員進入全文數位化的範疇，並協助順利完成工作。

註釋

1. 黃鴻珠，〈「觀前顧後」資料電子化的要訣之一〉，《佛教圖書館館訊》，第十五期 1998 年 9 月。
2. 陳秀華，《書畫數位化工作流程參考標準》，2005 年，頁 4-5。
3. 數位典藏國家型科技計畫，10-4 地方文獻影像編碼原則，《技術彙編》，2004 年。
4. 中華電子佛典協會，〈續藏輸入規則及範例〉、〈大正藏內文格式〉。
5. 莊德明，〈漢字缺字處理與梵巴藏字母的輸入〉，《佛教圖書館館訊》，第十四期， 1998 年 6 月。
6. 數位典藏國家型科技計畫，〈第三部份第二章漢字部件及組字規則〉，《技術彙編》，2002 年。
7. 中華電子佛典協會，http://www.cbeta.org/data-format/rare-rule.htm。
8. 周邦信，〈標記語言的應用〉，《佛教圖書館館訊》，第二十四期，2000 年 12 月。
9. 數位典藏國家型科技計畫，〈1-9 檔案數位化影像製作文件處理工作場所規範：以外交檔案爲例〉，《技術彙編》，2002 年。
10. 程婉如，《報紙期刊全文輸入數位化標準作業流程參考》，2005 年，頁 18-20。
11. 袁國華，〈建立 UNICODE 漢字異體字表與異體字辭典相關研究〉，《國家數位典藏通訊》，第三卷第八期，2004 年。
12. 國史館台灣文獻館難字處理程序。
13. 創用 CC — Creative Commons Taiwan，http://creativecommons.org.tw/。

拾、參考資料

中央研究院文獻處理實驗室，《漢字構形資料庫使用手冊》，2002 年。

林彥宏，《文書檔案數位化工作流程參考標準》，2005 年 12 月。

香光尼眾佛學院圖書館，《佛教圖書館館訊》，第二十四期，2000 年 12 月。

香光尼眾佛學院圖書館，《佛教圖書館館訊》，第十八/十九，1999 年 9 月。

香光尼眾佛學院圖書館，《佛教圖書館館訊》，第十五，1998 年 9 月。

香光尼眾佛學院圖書館，《佛教圖書館館訊》，第十四期，1998 年 6 月。

香光尼眾佛學院圖書館，《佛教圖書館館訊》，第三十二期，2002 年 12 月。

香光尼眾佛學院圖書館，《佛教圖書館館訊》，第三十五/三十六期，2003 年 12 月。

香光尼眾佛學院圖書館，《佛教圖書館館訊》，第四十期，2004 年 12 月。

國立臺灣大學典藏數位化計畫，《台灣文獻文物典藏數位化計畫》，2006 年 11 月 20 日，<http://140.112.113.4/project/default.asp>。

莊德明，〈漢字資訊化的困境及因應：談如何建立漢字知識庫〉，。

陳秀華，《書畫數位化工作流程參考標準》，2005 年 12 月。

程婉如，《報紙期刊全文輸入數位化標準作業流程參考》，2005 年 12 月。

數位典藏國家型科技計畫，《技術彙編》，2002 年。

數位典藏國家型科技計畫，《國家數位典藏通訊》，第三卷第八期，2004 年。

數位典藏國家型科技計畫內容發展分項計畫，《數位典藏叢書——數位化工作流程：漢籍全文主題小組》，2006 年 1 月。