佛典數位典藏內容開發之研究與建構

數位化工作流程簡介

修訂日期：97.04.15

計畫單位：法鼓佛教研修學院執行/中華電子佛典協會協辦

計畫名稱：台北版電子佛典集成之研究與建構

計畫簡介：

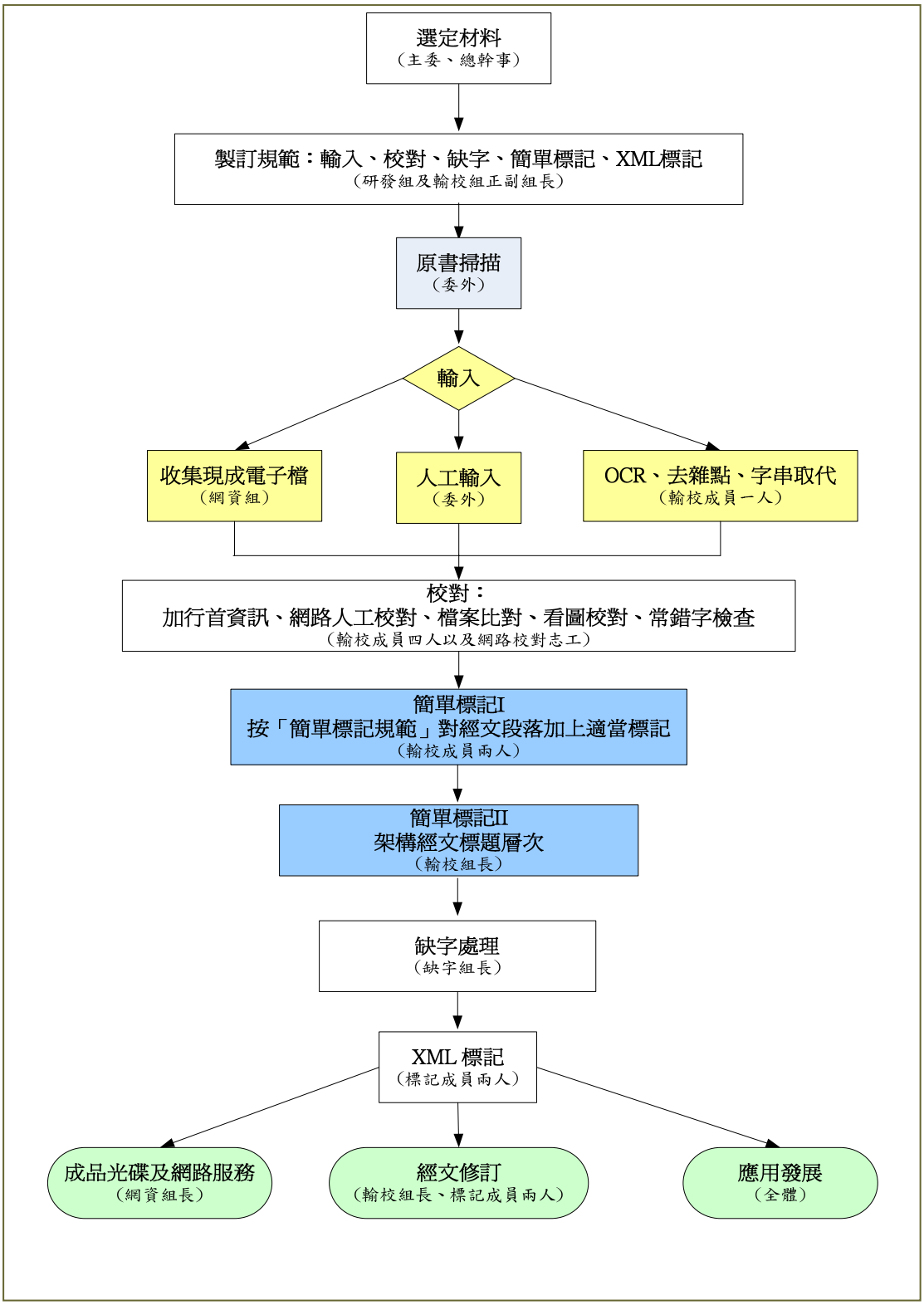
國科會數位典藏國家型科技計畫——「台北版電子佛典集成之研究與建構」（以下簡稱「台北版電子佛典計畫」）以國科會數位典藏內容開發補助專案「佛典數位典藏內容開發之研究與建構--經錄與經文內容標記與知識架構」(以下簡稱「佛典經錄計畫」)爲前期計畫，工作團隊具有建立數位佛典經錄資料庫的經驗與能力，以及全面性整理歷代經錄之研究基礎。在此基礎上，採用合作單位「中華電子佛典協會」(Chinese Buddhist Electronic Text Association，簡稱 CBETA) 累積多年經驗開發出的數位化工作流程，進行佛典數位化工作，以建立一部包羅並超越歷代大藏經內容的電子大藏經爲目標。

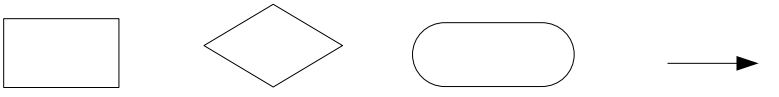
CBETA 於 1998 年 2 月 15 日正式成立，十年間陸續取得日本「大藏出版株式會社」與「株式會社國書刊行會」授權，進行《大正新脩大藏經》（以下簡稱《大正藏》）和《卍新纂續藏經》（以下簡稱《卍續藏》）之數位化工作。第一期《大正藏》數位化計畫由美國的『北美印順導師基金會』贊助，第二期《卍續藏》數位化計畫則由新加坡的居士贊助。十年中 CBETA 工作團隊累積了豐富的經驗，開發出兼具效率和品質之數位化工作流程。然而卻因 CBTEA 第三期計畫未徵得固定的長期贊助者，面臨有人才卻無經費進行後續數位化工作的困境。

因此自 2007 年 9 月開始，「台北版電子佛典計畫」與 CBETA 工作團隊合作，借重 CBETA 工作團隊的多年經驗，進行未收錄於《大正藏》和《卍續藏》之經典的數位化工作。CBETA 工作團隊具有豐富的佛典數位化經驗，「台北版電子佛典計畫」工作團隊則有先前進行「佛典經錄計畫」奠定的基礎，兩者的合作具有相輔相成的作用。結合 CBETA 和「台北版電子佛典計畫」之數位化成果，「台北版電子佛典集成」資料庫將包羅歷代大藏經收錄之中國佛教經典與著述，而「佛典經錄計畫」的成果亦將使「台北版電子佛典集成」資料庫在分類編排上超越歷代大藏經。

今以 CBTEA 開發之電子佛典作業爲例，說明「台北版電子佛典計畫」的數位化工作流程如下：

CBETA 經文數位化工作流程圖



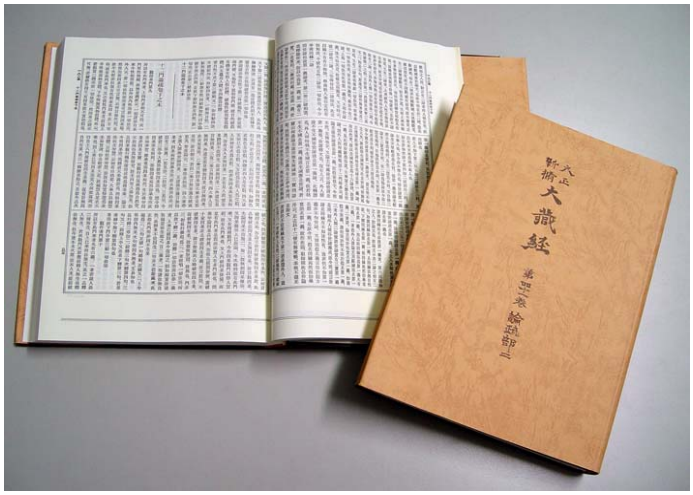


數位化工作流程說明

一、選定材料

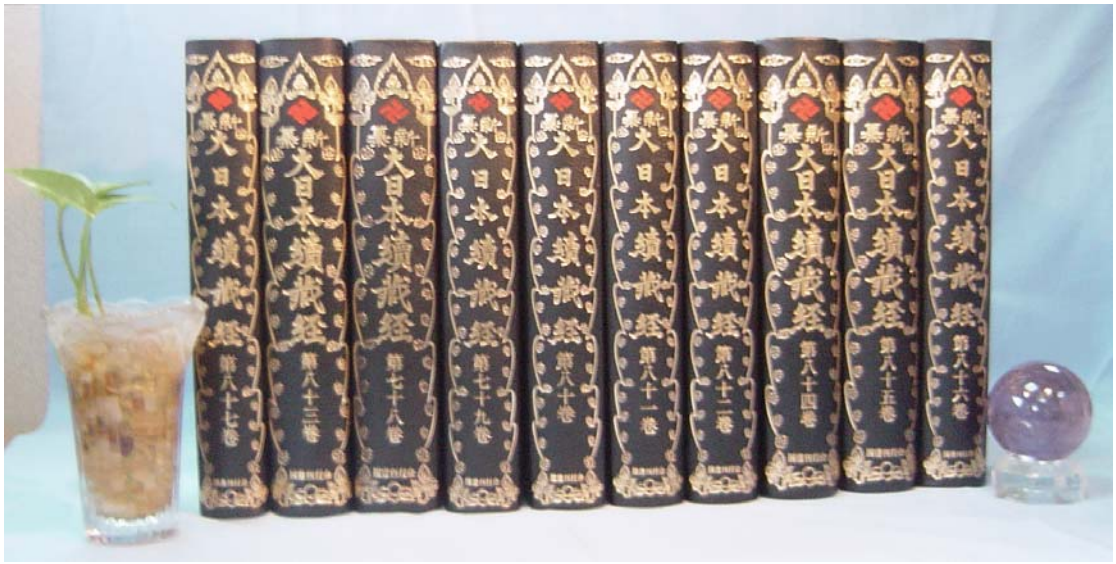
執行者：工作組主委、總幹事

CBETA 以「佛典集成」爲目標，故前期作業以「大藏出版株式會社」授與協會使用之《大正藏》 (圖一)爲底本，擇其中與漢傳佛教較爲相關之第一冊至第五十五冊以及第八十五冊，主要內容有歷代漢譯之〈印度撰述部〉與中國祖師著述之〈中國撰述部〉，共五十六冊，進行藏經電子化工作。數位化工作長達三年，已全數完成。



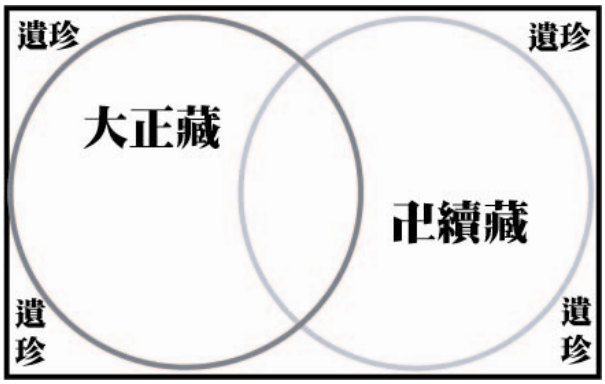
圖一、《大正新脩大藏經》

繼《大正藏》後進行數位化之藏經爲《卍續藏》(圖二)，已於 2007 年全數完成上線，現正進行《嘉興大藏經》(以下簡稱《嘉興藏》) 之數位化工作。未來將持續搜尋其他漢文佛典之遺珍，納入數位化工作，以達「佛典集成」之效。



圖二、《卍續藏》

選定《大正藏》乃因其爲國際上佛學研究之權威版本，現成電子檔與相關資源較多；《卍續藏》有極爲豐富的中國祖師大德著述，深具價值；加以《大正》與《卍續》兩藏皆爲鉛字排版，較適合輸入作業的進行；若集兩藏，重要的漢文佛典幾乎囊括大部份(圖三)，此乃 CBETA 選定材料之優先原則。



圖三、《大正藏》與《卍續藏》之關係圖

而如同圖三所示，在《大正藏》和《卍續藏》之外，仍有漢文佛典遺珍分散於其它藏經中，《嘉興藏》便是其中一部。儘管《嘉興藏》並非鉛字排版，但其收錄大量未收錄於前兩藏之明清漢文佛典，因此成爲「台北版電子佛典計畫」首選之數位化材料。

二、制定規範

執行者：工作組研發組正、副組長與輸校組正、副組長

爲確保數位化前後環節銜接順暢，各項流程需制定作業規範以利工作遵循。這些規範來自經驗累積，且以最終目標——「XML 標記」爲考量。本計畫針對幾項數位化重要作業：輸入、校對、缺字、簡單標記、XML 標記等，皆制定詳盡之作業規範。

(一)輸入

輸入規範包括對本文、本文以外之符號標誌，以及圖片、表格等等狀況提出規定，例如一般本文、夾注小字、段落，本文以外之頁碼、欄位、校勘符號，或是空白字元、空白行、表格、圖形、缺字……等。

(二)校對

計畫採用「檔案比對」程式進行校驗，因此校對規範著重於比對前之格式化準備，以及程式之使用方式與程序。

(三)缺字

經文中常可見非現行使用之古漢字或異體字、符號等，爲一般 BIG5(大五碼)系統無法辨識，故需建立一套缺字處理辦法，例如組字式規範，及以缺字資料表記錄缺字。

(四)簡單標記

簡單標記規範經文之經號、經名、作者、標題、段落…等之文字屬性。以簡單符號記錄，較 XML 標記容易上手。

(五)XML 標記

該計畫使用 XML 做爲佛典電子檔的標記語言，並採用國際規範 TEI(Text Encoding and Interchange)做爲基礎標籤集，再依實務標記作業經驗，修訂或新增標籤，建立適用於漢文電子佛典的標籤集。

三、原書掃描

執行者：早期自製，現委外執行。

掃描需將藏經原書或原書之影本拆卷，裁切騎縫邊，以散裝方式進行掃描。掃描要點如下：

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，就奇數頁與偶數頁檢查有無漏頁。
4. 編頁碼——先編奇數頁後編偶數頁，然後合併。
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。

早期使用具備「自動送紙功能」與「自動編號存檔」之掃描器，可一次自動掃存五十頁，程式能依冊、號編名存檔。後再以圖形處理軟體快速瀏覽圖檔以檢查掃描狀況。現因人成本效益考量，委託外部廠商執行，成本約每頁一・五元。

掃描產生之圖檔(圖四)需先設爲較高階影像：解析度 300dpi，色彩模式灰階或黑白，以供日後依不同目的降階應用。而該計畫之圖檔用途，一供「OCR 辨識」使用，二備爲「看圖校對」查看，故再將圖檔由 300dpi 灰階 轉成 Tif-g4 黑白格式，檔案既小，畫質又清晰。

No. 1

❶長阿含經序

❷長安釋僧肇❸述

夫宗極絕於稱謂。賢聖以之沖默。玄旨非言不傳。釋迦所以致教。是以如來出世。大敘有三。約身口則防之以禁律。明善惡則導之以契經。演幽微則❹辨之以法相。然則三藏之作也。本於殊應。會之有宗。則異途同趣矣。禁律律藏也。四分十誦。法相阿毗曇藏也。四分五誦。契經四阿含藏也。增一阿含四分八誦。中阿含四分五誦。雜阿含四分十誦。此長阿含四分四誦。合三十經。以爲一部。阿含秦言法歸。法歸者。蓋是萬善之淵府。總持之林苑。其爲典也。淵博弘富。❺韞而彌廣。明宣禍福賢愚之迹。剖判眞僞異❻旨之原。歷記古今成敗之數。墟域二儀品物之倫。道無不由。法無不在。譬彼巨海百川所歸。故以法歸爲名。開❼析修途。所記長遠。故以長爲目。翫茲典者。長迷頓曉。邪正難❽辨。顯如晝夜。報應冥昧。照若影響。劫數雖遼。近猶朝夕。六合雖曠。現若目前。斯可謂朗大明於幽室。惠五目於衆瞽。不閾戶牖。而智無不周矣。大秦天王。滌除玄覽。高韻獨邁。恬智交養。道世俱濟。每懼徵言翳於殊俗。以右將軍使者司隸校尉晉公姚爽。質直清柔。玄心超詣。尊尚大法。妙悟自然。上特留懷。每任以法事。以弘始十二年歲次上章❾閹茂。請罽賓三藏沙門佛陀耶舍。出律藏❿一分四十五卷。十四年訖。十五年歲次昭陽⓫赤奮若。出此長阿含訖。涼州沙門佛念⓬譯。秦國道士道含筆受。時集京夏名勝沙門。於第校定。恭承法言。敬受無差。竊華崇朴。務存聖旨。余以嘉遇猥參聽次。雖無翼善之功。而預親承之末。故略記時事。以示來賢焉

佛說⓭長阿含經卷第一

⓬後秦弘始年佛陀耶舍共竺佛念譯

(一)第一分初⓮大本經第一

如是我聞。一時佛在⓯舍衛國⓰祇樹⓱給孤獨園。與⓲大比丘衆千二百五十人俱。時諸比丘於乞食後集⓳花林堂。各共議言。諸賢比丘。唯無上尊爲最奇特。神通遠達。威力弘大。乃知過去無數諸佛。入於涅槃。斷諸結使。⓴消滅戲論。又知彼佛劫數多少。名號姓字。所生種族。其飮食壽命脩短。所更苦樂。又知彼佛有如是戒。有如是法。有如是慧。有如是解。有如是住。云何諸賢。如來爲善別法性。知如是事。爲諸天來語。乃知此事。爾時世尊在閑靜處。天耳清淨。聞諸比丘作如是議。即從座起。詣花林堂。就座而坐。爾時世尊知而故問。謂諸比丘。汝等集此。何所語議。時諸比丘具以事答。爾時世尊告諸比丘。善哉善哉。汝等以平等信。出家修道。諸所應行。凡有二業。一曰賢聖講法。二曰賢聖默然。汝等所論。正應如是。如來神通。威力弘大。盡知過去無數劫事。以能善解法性。故知。亦以諸天來語。故知。佛時頌曰

比丘集法堂　講說賢聖論
如來處靜室　天耳盡聞知
佛日光普照　分別法界義
亦知過去事　三佛般泥洹
名號姓種族　受生分亦知
隨彼之處所　淨眼皆記之
諸天大威力　容貌甚端嚴
亦來啓告我　三佛般泥洹
記生名號姓　哀愍音盡知
無上天人尊　記於過去佛

又告諸比丘。汝等欲聞如來識宿命智知於過去諸佛因緣不。我當說之。時諸比丘白㉑言世尊。今正是時。願樂欲聞。善哉世尊。以時講說。當奉行之。佛告諸比丘。諦聽諦聽。善思念之。吾當爲汝分別解說。時諸比丘受教而聽

佛告諸比丘。過去九十一㉒劫時世有佛名㉓毘婆尸如來至眞。出現于世。復次比丘。過去三十一劫有佛名㉔尸棄如來至眞。出現於世。復次比丘。卽彼三十一劫中有佛名㉕毘舍婆如來至眞。出現於世。復次比丘。此㉖賢劫中有佛名㉗拘樓孫。又名㉘拘那含。又名㉙迦葉。我今亦於賢劫中成最正覺。佛時頌曰

過九十一劫　有毘婆尸佛
次三十一劫　有佛名尸棄
卽於彼劫中　毘舍如來出

❶此序宋元明三本ニ依テ載ス ❷[長安]ー㊥ ❸[述]ー㊥ ❹辨ニ辯㊥* ❺韞ニ温㊥㊦ ❻旨ニ濟㊥㊦ ❼析ニ斤㊥ ❽闇ニ掩㊥㊦ ❾一ニ二㊥ ❿[赤]ー㊥㊦ ⓫長阿含經Dīrgha-āgama～Dīgha-nikāya ⓬後秦弘始年ニ姚秦三藏法師㊂ ⓭大本經ニ大本緣經～D. 14.Mahāpadhāna-suttanta. ⓮舍衛～Sāvatthī. ⓯祇樹～Jetavana. ⓰花林窟～Kareri-kuṭikā. ⓱消ニ錯㊂* ⓲泥洹ニ涅槃㊂ ⓳鞞ニ壓㊥ ⓴(佛)十言㊂ ㉑劫～Kappa. ㉒毘婆尸～Vipassin. ㉓尸棄～Sikhin. ㉔毘舍婆～Vessabhū. ㉕賢劫～Bhadda-kappa. ㉖拘樓孫ニ拘留孫㊂～Kakusandha. ㉗拘那含～Koṇāgamana. ㉘迦葉～Kassapa.

圖四、原書掃描之圖檔

四、輸入

對於大量佛典經文的輸入，應針對不同內容，選擇採用人工輸入或是掃描圖檔辨識的方法來產生文字檔。

該計畫之輸入方法有三種，分別爲收集現成電子檔、人工輸入，以及 OCR 圖檔辨識。決策方式爲：如一佛典已有現成電子檔，則該電子檔可供日後檔案比對使用；無電子檔又難以透過 OCR 辨識之文字，如手抄本與刻版經文，則採用人工輸入。

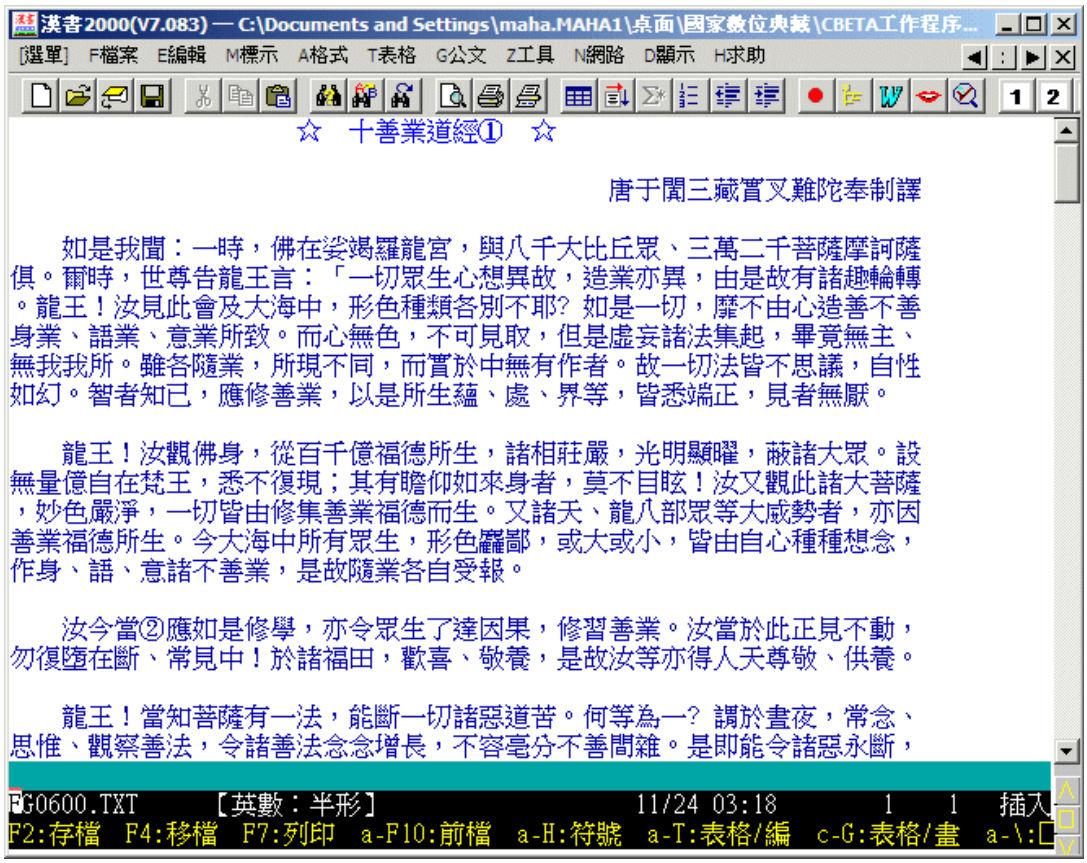
不論使用何種輸入方式，一部經文至少需產生兩份電子檔。

(一)收集現成電子檔：

執行者：工作組網資組

早在計畫實行前，網路上已流傳許多對佛典有興趣之志工團體的輸入電子檔，或是其他佛教機構、學術單位研發之電子佛經。

現成電子檔之收集大都以流通較廣的經文爲主，這些電子佛經(圖五)通常不符合計畫之規定格式(如需加註頁、欄資訊)；故收集得來之檔案在檔案比對前，還需經過格式化之後續處理。



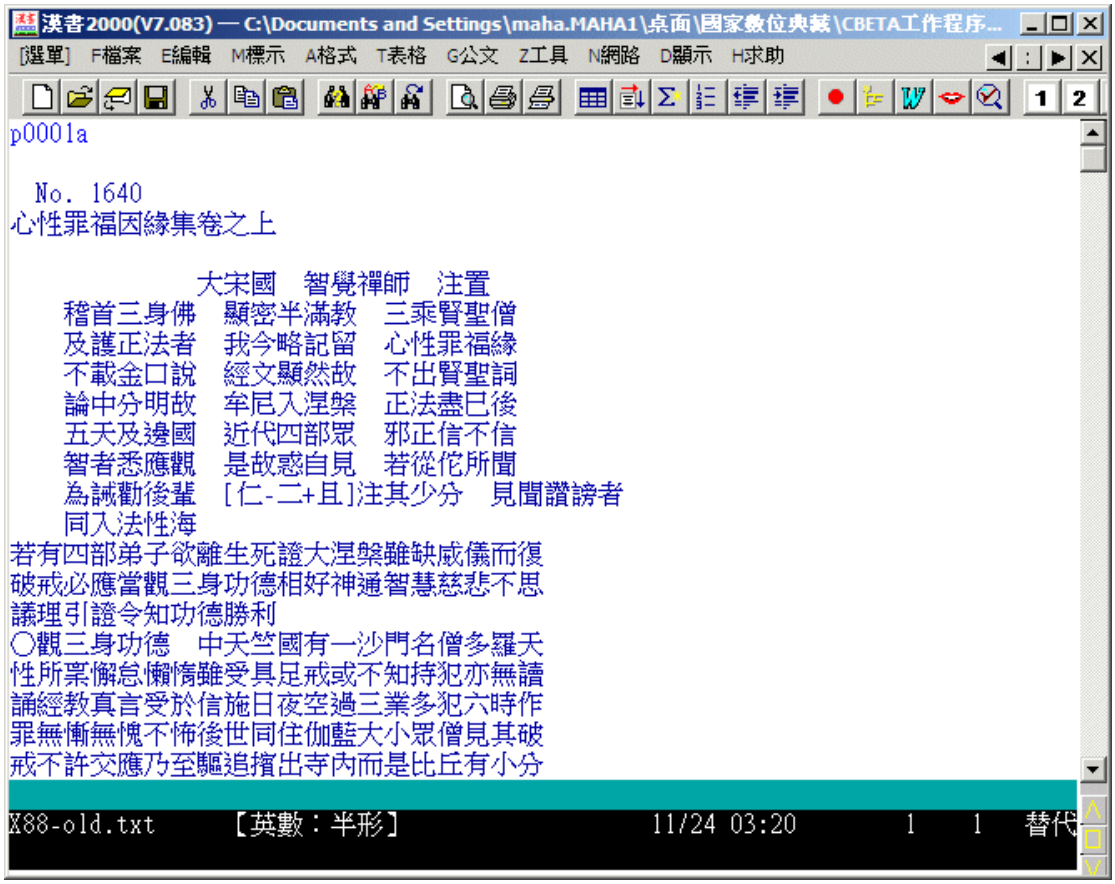
圖五、《大正藏》之現成電子經文

(二)人工輸入：

執行者：委外執行

無法使用 OCR 辨識軟體辨識之佛經，委外交由專業承包公司進行人工繕打。

委外之前，必須事先制定輸入規範，將之交與廠商人員比照辦理。人工輸入產生之純文字電子檔，需包含頁、欄(圖六)資訊，以及依冊號順序命名之檔案名稱。人工輸入成本約每千字五十元。



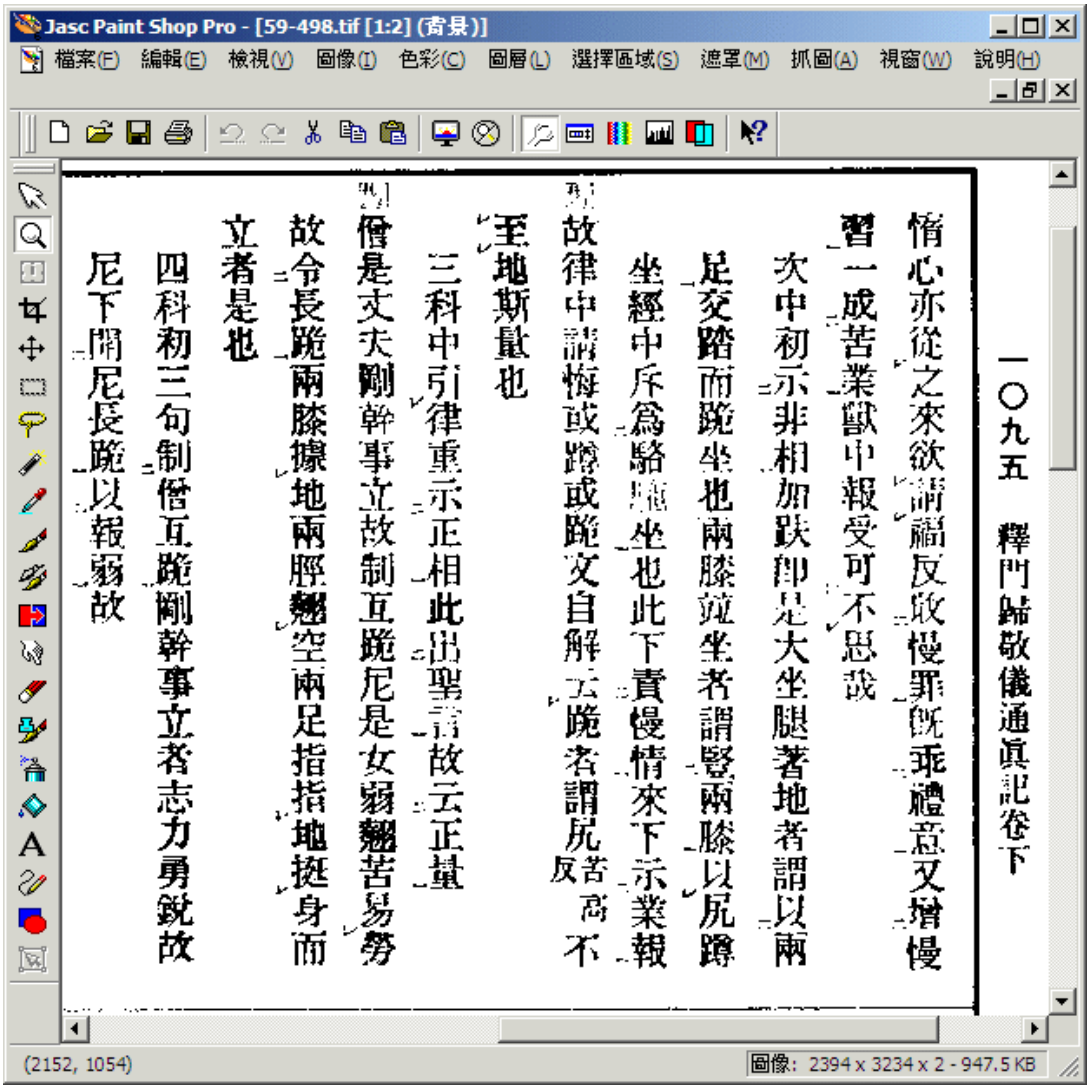
圖六、委外人工輸入產出之電子檔

(三)OCR 圖檔辨識：

執行者：工作組輸校組成員一人

1.去除雜點

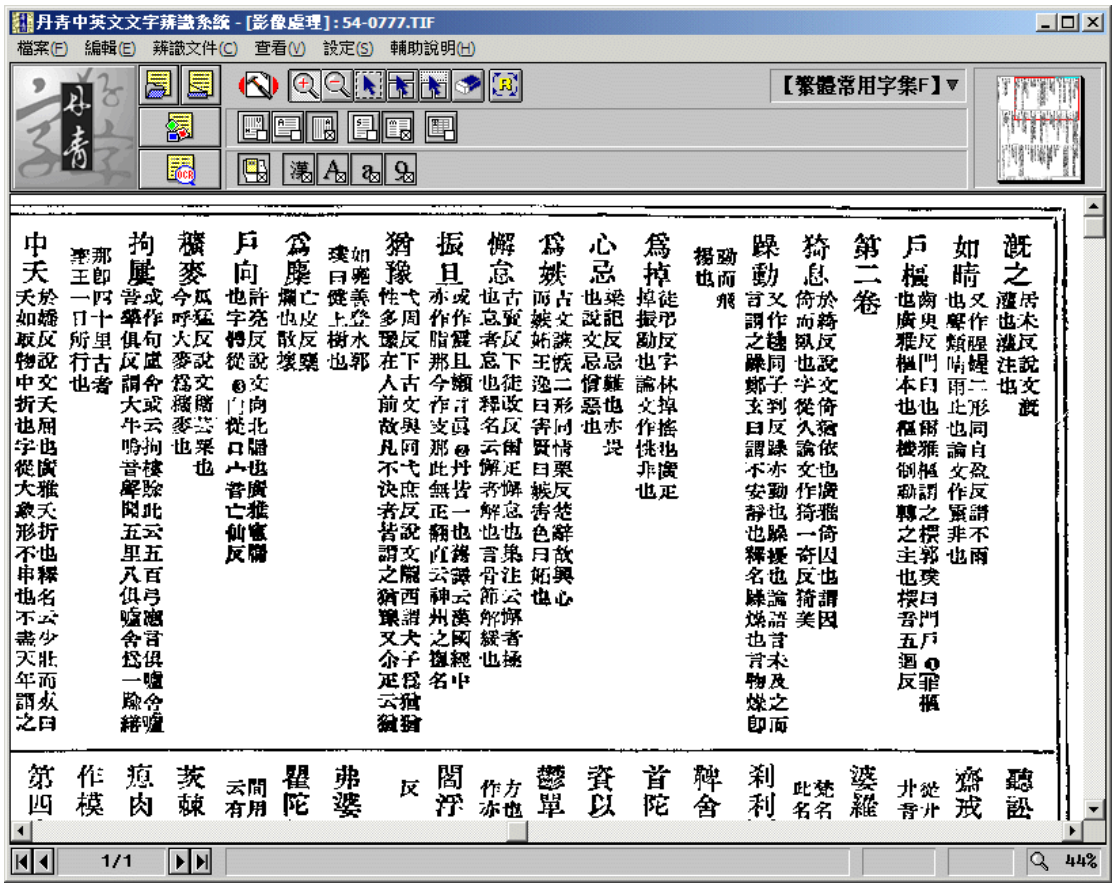
經文書上常有異於文字之讀音符號與注釋標記(圖七)，嚴重影響 OCR 辨識之判讀結果；故掃描後之經文圖檔，須先以程式去除雜點，產生一新 TIF 圖檔。



圖七、含讀音符號與雜點之原始掃描圖檔

2. OCR 圖檔辨識

將去除雜點後之新圖檔，匯入丹青公司特別爲該協會量身訂作之 OCR 程式(圖八)進行辨識，產出一份經文之「純文字檔」。

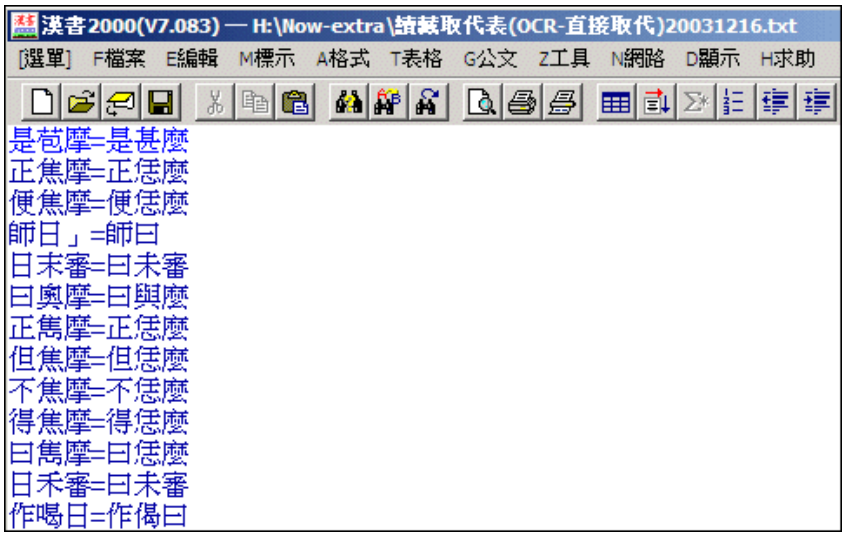


圖八、丹青 OCR 操作介面

該程式與一般辨識程式不同處在於「丹青 for CBETA」可判讀經文特有之雙排小字。

3.字串取代

使用「常錯字串取代程式」，以正確字串快速批次取代 OCR 後可能之常錯字串(圖九)，免除逐字校對之不便，約可提升純文字檔文字精確度至 90%。



圖九、OCR 常錯字串取代表

※進行至此，輸入步驟可能產生三種皆未格式化(未加行首資訊)之電子檔：

1). 網路收集之現成電子檔。

2). 委外人工繕打(包含頁欄資訊)，正確率約爲 97%之電子檔。

3). OCR 辨識後，正確率 90%之電子檔。

五 、校對

執行者：工作組輸校成員四人與網路校對志工

校對程序包括「加行首資訊」、「網路人工校對」、「檔案比對」、「看圖校對」、「常錯字檢查」五項。前二項爲第三項「檔案比對」之前置作業，須先妥善執行，後續之比對工作才能順利完成。

(一)加行首資訊

加行首資訊屬於格式化作業。行首資訊用於記錄每行電子經文在紙本經書上之相對位置，此舉不僅幫助後續之標記處理，也嘉惠學術引用之便。

將含有「頁欄資訊」之未格式化經文純文字檔匯入「加行首資訊程式」，執行後稍加編輯即可產生包括冊數、經號、頁、欄、行等資訊之新純文字檔。內容格式如下：

例： T10n0279_p0070a04 ‖ 菩薩在家 當願眾生 知家性空
T10n0279_p0070a05 ‖ 免其逼迫 孝事父母 當願眾生
T10n0279_p0070a06 ‖ 善事於佛 護養一切 妻子集會

T：大正藏　　10：冊數　　n0279：經號

p007：頁　　a04：a 欄(第一欄)第 4 行　　‖：分隔符號

經此步驟，所有純文字電子經文皆已格式化成 CBETA 所需格式，即可進行下階段之數位化工作。

(二)網路人工校對

OCR 產出之電子經文純文字檔經字串取代後，正確率僅達 90%。若將之與另一電子檔(如人工輸入檔)比對，勢必差異數量龐大，需動用大量人力方能完成校對程序。

CBETA 有一「網路校對」機制，即於網路上徵集志工約九百人，投入線上一人一頁分工校對行列。線上校對程序爲：

1. 上 CBETA 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. 回傳 CBETA。

看圖校對程式係該協會之程式設計師開發設計，校對者可同時閱覽純文字檔與其相對之圖檔，達成看圖替代翻書之快速校閱。

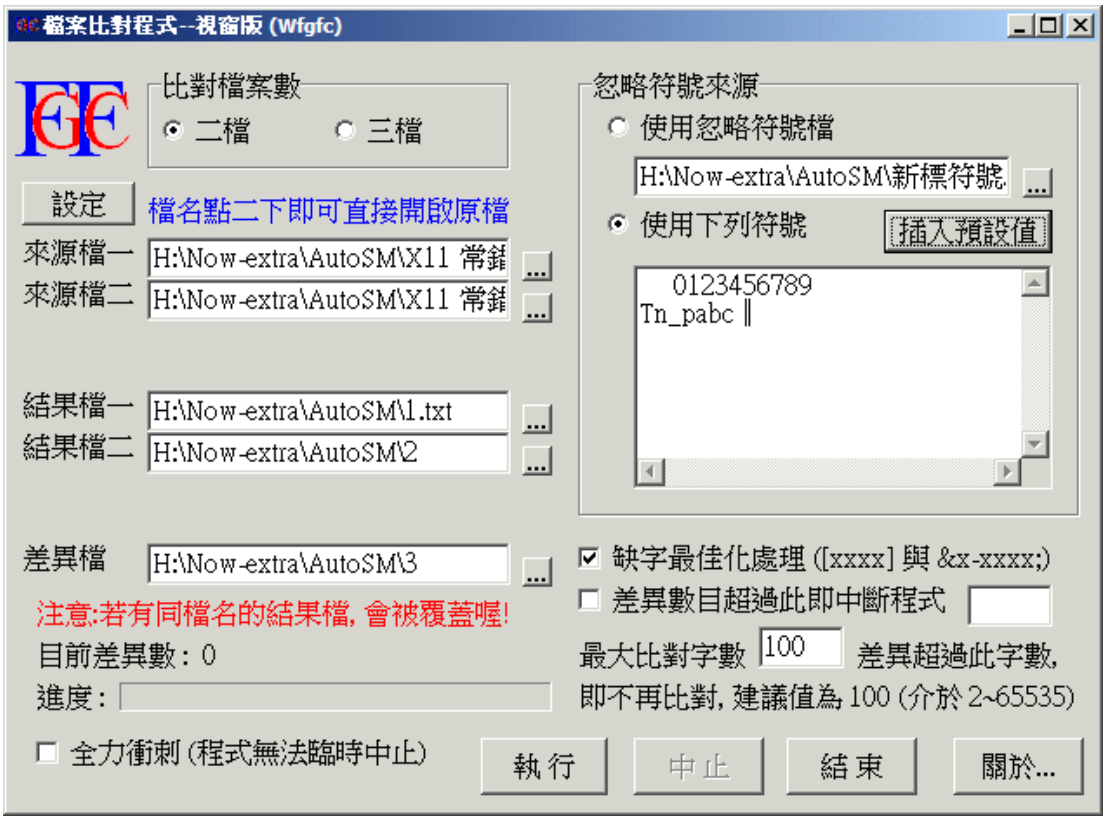
網路校對後之 OCR 經文，正確率可提升爲 98%。

(三)檔案比對

傳統人工校對，即使四校或十校，總有無法避免的死角。該計畫利用電腦檔案比對，即同一份經文內容，由兩個版本予以輸入，然後以檔案比對程式找出兩者差異，再以看圖校對方式進行訂正，產生一份超越一般人工校對水準之經文檔。

首先，收集兩份同一經文但輸入來源不同之純文字電子檔。若有一頁一頁的小檔，可利用「檔案合併程式」，將兩檔各自所含小檔之純文字檔案合併成大檔，以利文書編輯處理及後續比對作業的進行。

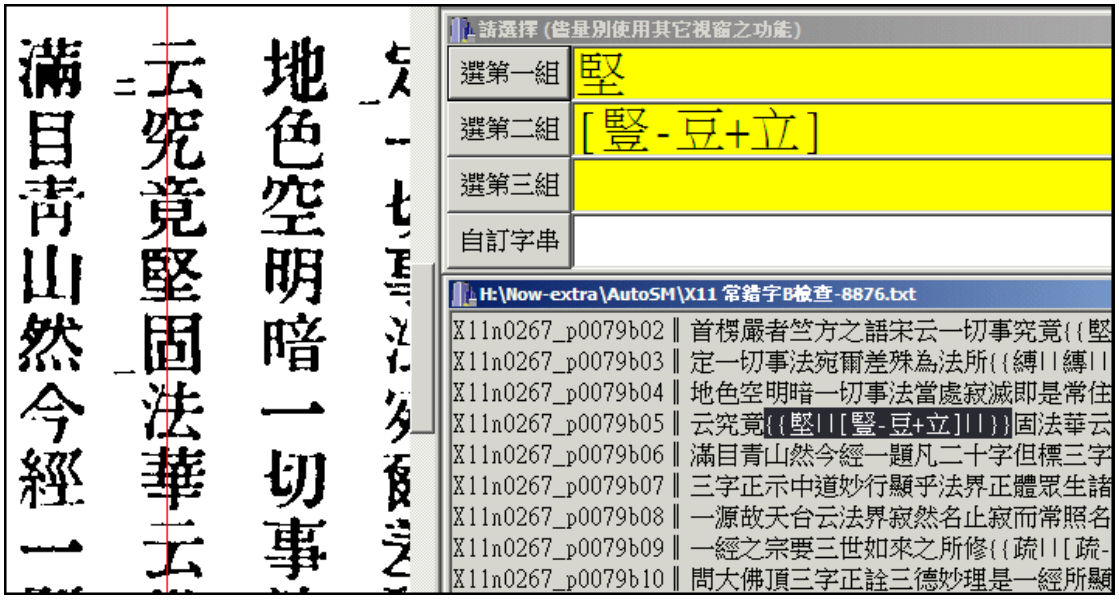
將合併成大檔之兩檔匯入「檔案比對程式」(圖十)，執行第一次兩檔比對。比對後產生一個主要差異檔。以《大正藏》而言，平均每冊約產生兩萬個差異。



圖十、檔案比對程式

(四)看圖校對

比對後之差異檔，交由兩位熟識經文之經驗人員各自利用 SeeCheck「看圖校對程式」(圖十一)，以差異檔比照原書掃描圖檔予以訂正。

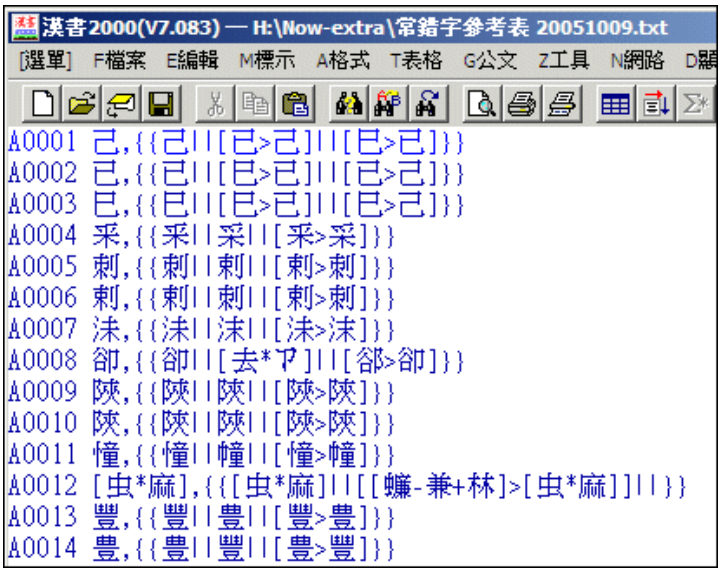


圖十一、看圖校對程式介面

該兩位人員訂正後交回的兩份校對完成檔，再以程式執行第二次檔案比對，比對後會產生一數量較小之差異檔。將此差異檔交由一位人員進行最後把關，方法也是以差異檔比照原書掃描圖檔看圖校對。

(五)常錯字檢查

校對最後的工作重點是對於任何值得疑慮的字元，我們將之列入「常錯字參考表」(圖十二)，並透過程式對檔案進行取代，形成差異以利用看圖方式來校對。這個概念是我們對看圖校對程式的充分應用，可以發揮事半功倍的效果。



圖十二、常錯字參考表

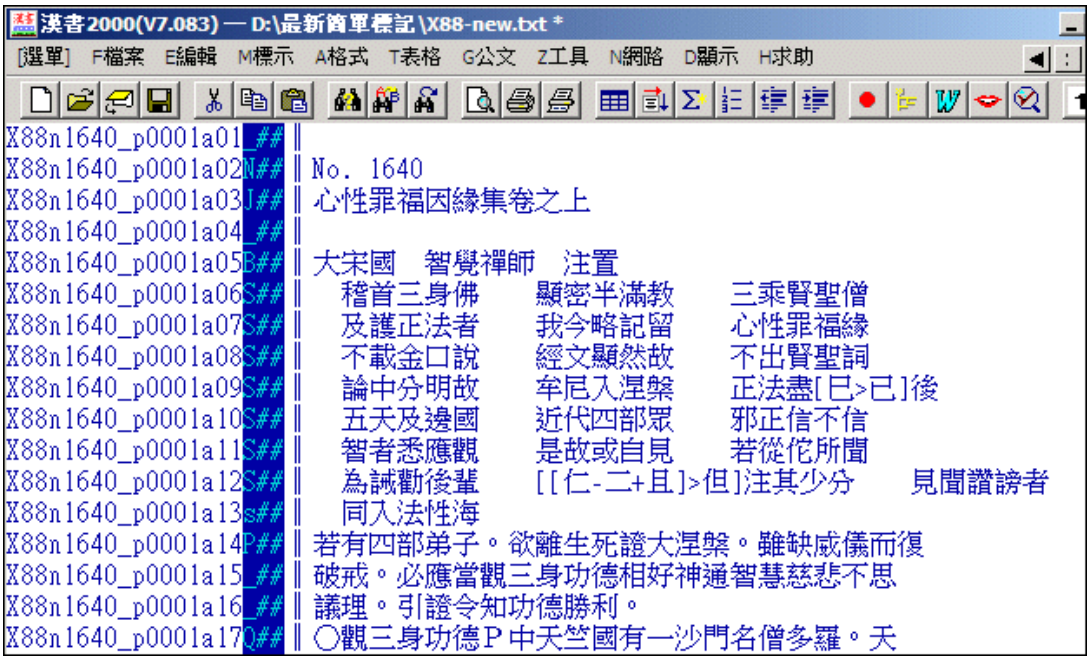
六、標記

(一)簡單標記 I

執行者：工作組輸校組成員兩人

標記，是針對已完成校對文件之進一步編輯作業。在進入正式 XML 標記之前，輸校組需對經文段落加上適當標記，成爲「簡單標記版」的經文電子檔。

這一套簡單標記，目的是把經文當中 「經號 N」、「經卷 Jj」、「品 D」、「著者 A」、「譯者 Y」、「序 X」、「偈頌 Ss」、「咒語 Z」、「附文 W」、「段落 P」、「其它標題 Q」、「行中小段落 P」…等，以簡單符號標示區分出來，方便電腦認識經文各段落之不同屬性，並能加以進一步運用。簡單標記主要是在行首資訊後的三欄「_##」標記欄位置中標示出來(圖十三)，或標記於經文中的「行首」、「行中」、「行尾」。

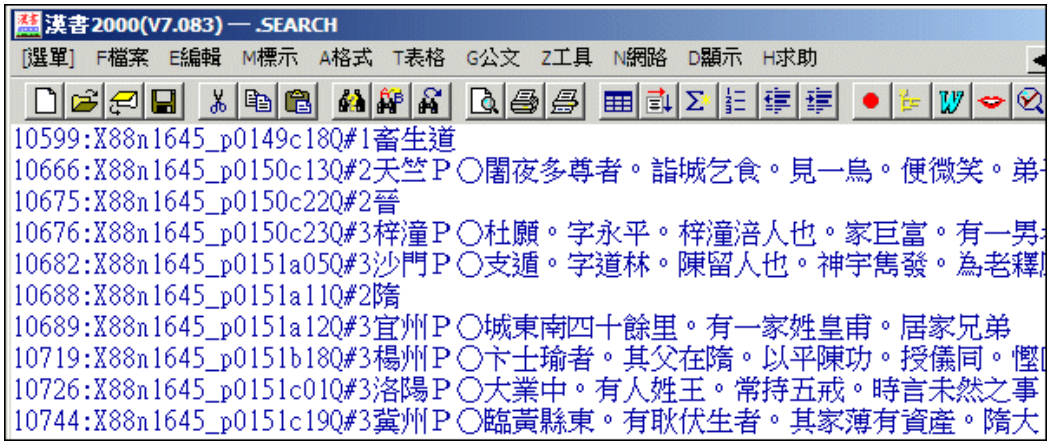


圖十三、第一次簡單標記產出之純文字檔

(二)簡單標記 II

執行者：工作組輸校組組長

第二階段簡單標記之重點工作爲「架構經文標題層次」(圖十四)。此自訂標記可讓電腦認識整篇經文之章節架構，如：



圖十四、經文之標題層次架構

七、缺字處理

執行者：工作組缺字組長

CBETA 以「BIG5(大五碼)」加上「組字式」作爲記錄缺字的基礎。

使用一般組字式來表達佛典缺字的方法，是考量使用者能在純文字環境下閱讀，不需另外安裝造字檔或圖檔而設計的，這種方式提供了閱覽、散播上的便利性，也不會佔用使用者對造字檔自行運用的空間。

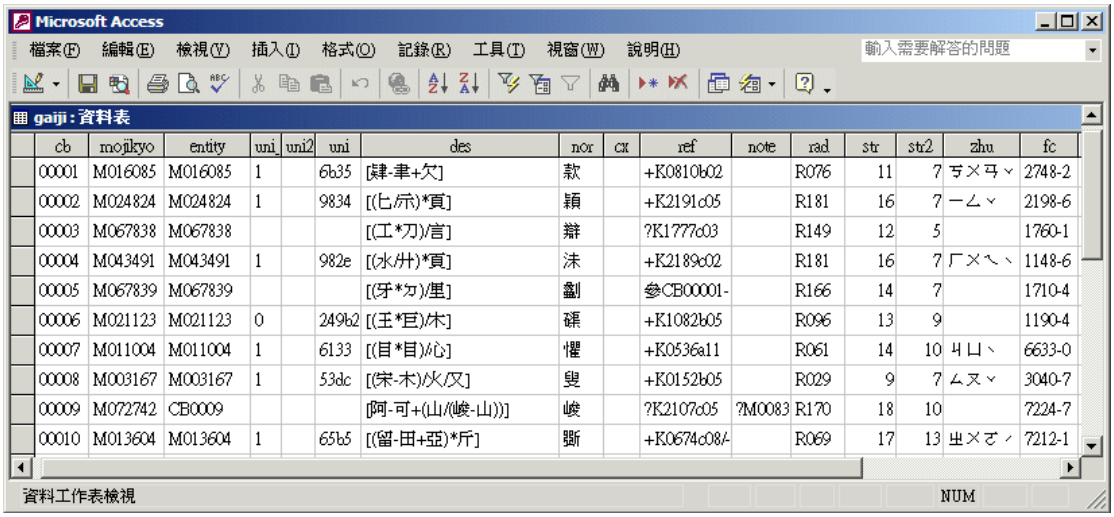
該組字法含「*」、「/」、「@」、「-」、「+」、「?」六個半形基本符號，及「(…)」、「[…]」兩組半形分隔符號。

舉例說明如下：

表一、CBETA 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

記錄缺字後，並將缺字相關資訊，包括注音、筆畫、部首、通用字、Unicode…等建構成漢文佛典缺字資料庫(圖十五)。

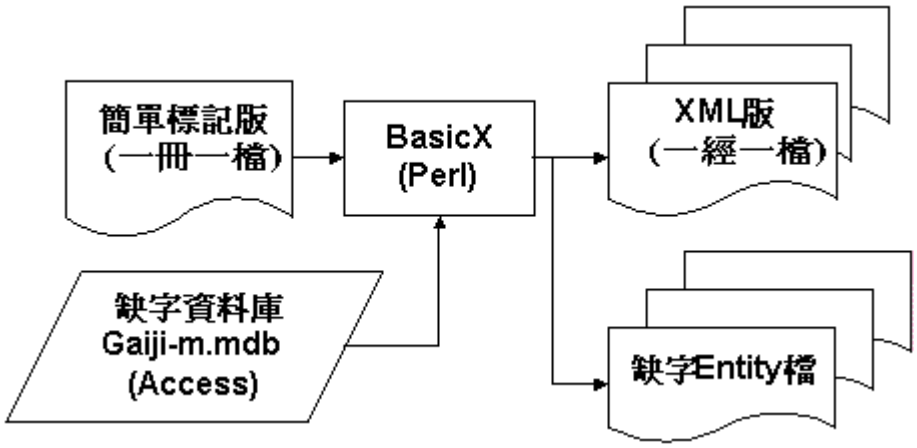


圖十五、缺字資料庫畫面

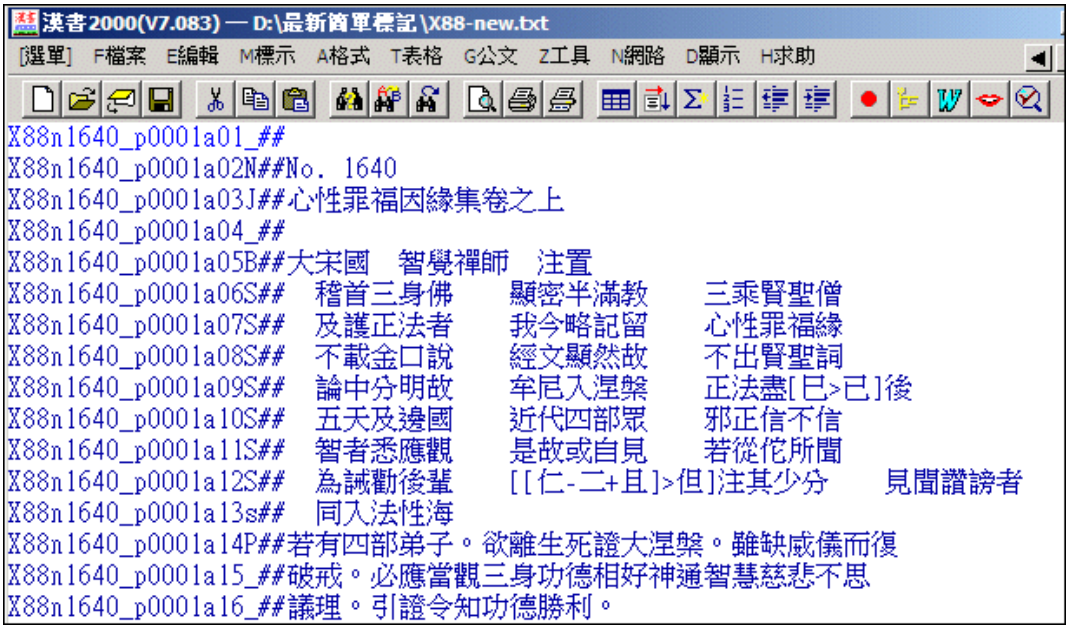
八、XML 標記

執行者：工作組標記成員兩人

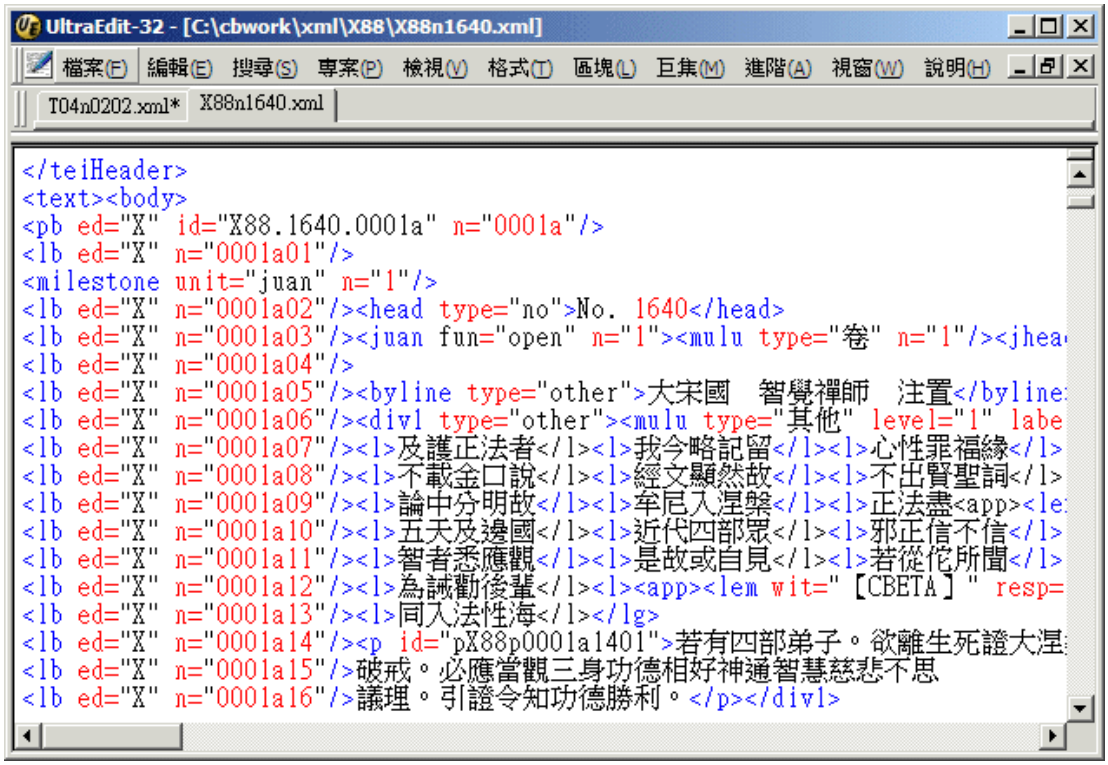
經簡單標記、缺字處理後之經文，以程式(圖十六)將簡單標記經文(圖十七)轉爲 XML TEI 標記經文(圖十八)。



圖十六、簡單標記轉換爲 XML 標記之程式流程圖



圖十七、簡單標記經文



圖十八、XML TEI 標記經文

之後仍需做語法檢查及人工編輯，最後以程式將 XML 版輸出與簡單標記版相互比對。

九、應用服務

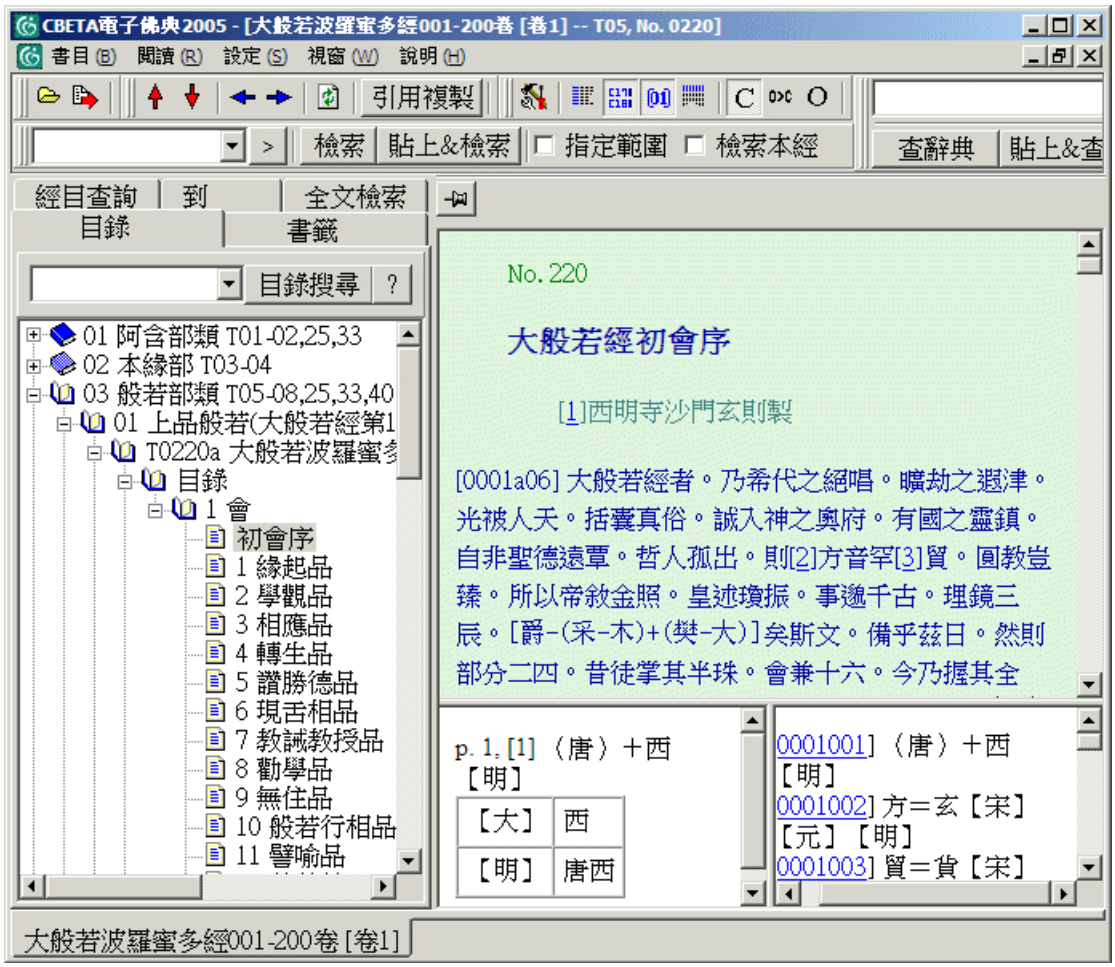
(一)成品光碟與網路服務

執行者：工作組網資組長

該計畫利用標記完成之經文，轉換成普及網路版放置網路上供大眾免費瀏覽、檢索與下載(圖十九)；此外，CBETA 每年發行一萬份電子佛典光碟(圖二十)，光碟含有優異檢索及閱覽功能的 CBReader(圖二十一)，提供免費索取，與大眾結緣。

圖十九、CBETA 網站

圖二十、CBETA 每年發行之光碟

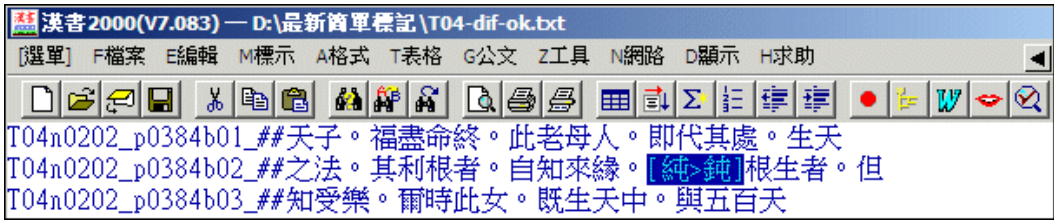


圖二十一、優異檢索及閱覽功能的 CBReader

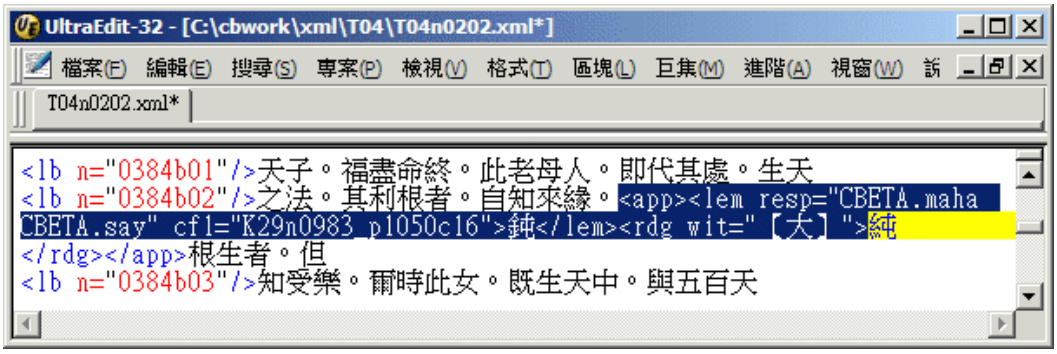
(二)經文修訂

執行者：工作組輸校組長、標記成員兩人

儘管經文已上線、壓光碟，仍需不斷查證相關資料以確認讀者及內部作業發現之經文用字問題，並執行經文資料庫之修訂，包括簡單標記版(圖二十二)及 XML 版(圖二十三)，兩者必須同步修訂；期望透過修訂，提升經文資料庫之品質。



圖二十二、簡單標記版修訂

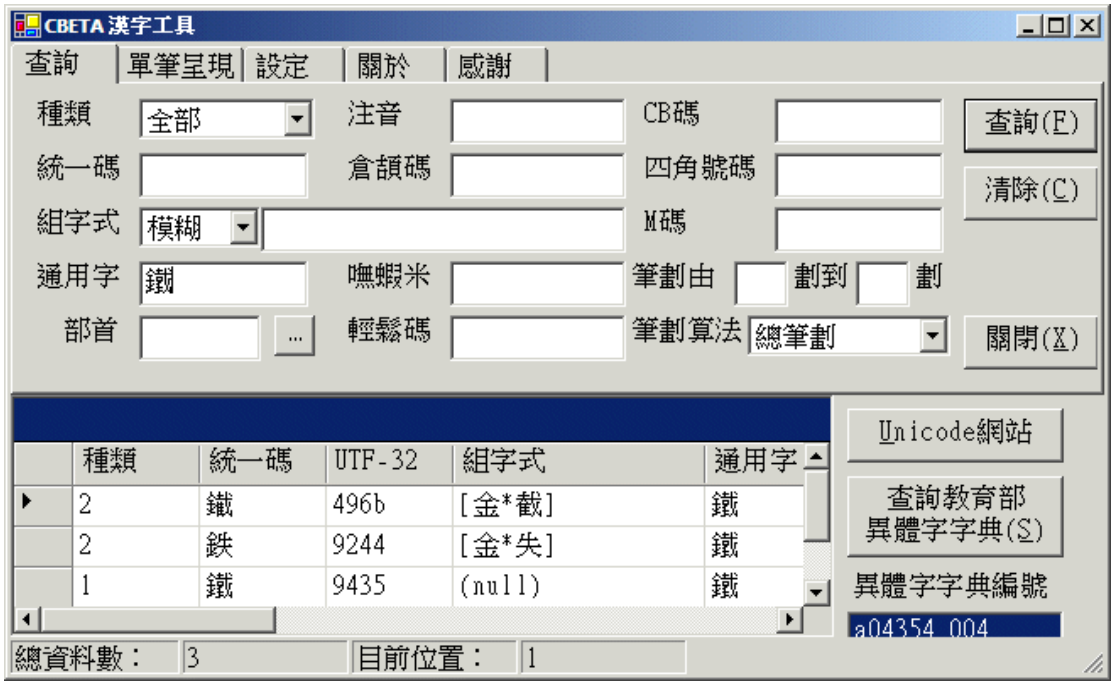


圖二十三、XML 版修訂

(三)應用發展

執行者：全體工作人員

除生產預定經文典籍外，CBETA 也亟欲推廣與經文資料庫相關之應用及技術，例如漢字工具(圖二十四)、新式標點、通用詞庫、相關字(辭)典、藏經目錄資料庫、各版藏經經文對照資料庫…等。



圖二十四、漢字工具

※ 製作單位：數位典藏與數位學習國家型科技計畫

拓展臺灣數位典藏計畫　數位內容建置與整合子計畫

中華電子佛典協會

法鼓佛教研修學院

※ 文字修訂：法鼓佛教研修學院「台北版電子佛典計畫」陳以儒　修訂

拓展臺灣數位典藏計畫　數位內容建置與整合子計畫

— 文獻與檔案主題小組助理　陳美智　修訂

※ 圖片拍攝：法鼓佛教研修學院「台北版電子佛典計畫」陳以儒　修訂

※ 圖片提供：法鼓佛教研修學院

※ 圖文編輯：法鼓佛教研修學院「台北版電子佛典計畫」陳以儒　修訂

致謝：

感謝「台北版電子佛典集成之研究與建構」計畫共同主持人杜正民老師、法鼓佛教研修學院陳以儒先生，撥冗指導及提供實地拍攝與簡介修訂。並感謝法鼓佛教研修學院其餘相關人員之協助。

數位典藏國家型科技計畫內容發展分項計畫

數位典藏工作流程調查表

單位：國立臺北藝術大學　共同科　　數位化物件名稱：漢文大藏經經文

子計畫名稱：台北版電子佛典集成之研究與建構　　分項計畫名稱：＿＿＿＿＿＿＿＿＿＿

主持人（負責人）(E-mail、Tel)：郭敏芳（釋惠敏） huimin2525@gmail.com　02-2498-0707#2271

聯絡人（E-mail、Tel)：陳以儒 sraddhabala@gmail.com　02-2498-0707#2254

程序	工作內容	操作人員（數量、專業能力之要求）	硬體（名稱、版本、價格）	軟體（名稱、版本、價格等）	依循標準（技術規範、成品規格、品質要求…等）	耗時	總結（困難、缺失、特色…等）	成本估算	備註
1	●選定材料	主委、總幹事			以「佛典集成」爲目標		配合現藏目錄整理以得知待補足典籍		
2	●製訂規範： 輸入、校對、缺字、簡單標記、XML 標記	研發組及輸校組正副組長	PC	MS Office、漢書 2000、UltraEdit	以繁體中文 BIG5 爲作業基礎。大體保持原書用字及版面格式，並方便程式進行文字處理。		規範不是一開始就齊備的，必須從工作經驗中不斷累積、修正。不可過於拘泥書版格式，須配合電子化特性及考量作業方便。		
3	●原書掃瞄	委外或 輸校人員一人	PC、Scaner	掃瞄器附帶軟體	300dpi 灰階 or 黑白。最後轉成 Tif-g4 黑白格式做爲作業運用材料。		有了掃瞄圖，可以少買幾套書。另個重點是，後續作業依靠掃瞄圖的機會很多，包括一般查閱以及程式運用。	1.5 元/頁	
4-1	●輸入一： 收集現成電子檔	網資組	PC		主要是由各友好單位及個人提供，少部份是上網搜尋取得。不同編碼或編輯格式皆可，後續由程式統一轉化處理。		缺字處理方法不一致。根據的輸入底本不一定是我們所要使用的底本。因此事後的消化整理要花一些功夫。		
4-2	●輸入二： 人工輸入	委外	PC		按輸入規範作業。若同時輸入兩份電子檔，必須分找不同輸入單位，以免互拷檔案。		按輸入材料狀況決定人工輸入 or OCR。爲配合檔案比對至少必須產生兩份電子檔。最常遇到的困難是原文不清或缺字太多。另外，雕版藏經用字異體化嚴重，若不予以規範勢必窒礙難行。	50 元/千字	
4-3	●輸入三： OCR、去雜點、字串取代	輸校成員一人	PC	漢書 2000、丹青 OCR、自行研發的各種工具軟體	善用 OCR 軟體，不做線上逐字校對，利用「取出表」快速進行字串取代。				

5	●校對： 加行首資訊、網路人工校對、檔案比對、看圖校對、常錯字檢查	輸校成員四人以及網路校對志工	PC	漢書 2000、自行研發的各種工具軟體	同一經文由兩個人同時執行校對，校對完畢再予以比對除錯，以求得更高精確度。		以檔案比對爲主，人工校對(網路志工)爲輔。校對理想標準爲錯誤率 1/10000 以下。		
6	●簡單標記 I	輸校成員兩人	PC	漢書 2000、PERL、自行研發的各種工具軟體	按「簡單標記規範」對經文段落加上第一階段標記		加上第一階段自訂標記，讓電腦認識經文各個段落的基本不同屬性。		
7	●簡單標記 II	輸校組長	PC		架構經文標題層次，以及加入諸如「問答」、「原文解釋」、「辭書」等特殊標記。		決定經文在瀏覽時的樹狀目錄，以及深化標記內涵。		
8	●缺字處理	缺字組、網路查詢志工	PC	漢書 2000、MS Access、Paint Shop pro、IrfanView	處理新增缺字及維護缺字資料庫。		建立缺字相關資訊，包括注音、筆畫、部首、通用字、Unicode 等，並吸取 BIG5 系統字資料，以期建立完整的文字資料庫。		
9	●XML 標記	研發組、標記成員兩人	PC	UltraEdit、PERL、WinCvs、 SP、Python、MS Office	Big5, CP950, Unicode, XML, TEI. 利用程式將 簡單標記 轉爲 XML 標記。		以符合國際標準的 XML 語言建立經文資料庫。舉凡經文排版呈現、目錄架構、檢索….，都可做出有效運用。		
10	●成品光碟及網路服務	網資組長	PC	Borland C++ Builder、UltraEdit、MS Office	提供 CBReader 讀經器，以及 normal、app、xml 等各種版本經文，並爲使用者解決使用上的問題。		每年發行一萬份光碟免費與大眾結緣。網站提供經文檢索及經文下載服務。		
11	●經文修訂	輸校組長、標記成員兩人	PC	UltraEdit、MS Office、自行研發的各種工具軟體	查證相關資料以確定讀者及內部作業所發現的經文用字問題，並執行經文資料庫修訂。		透過不斷修訂，經文資料庫的品質越來越好。		
12	●應用發展、推廣	全體	PC		漢字庫、新式標點、通用詞庫、辭書、藏經目錄資料庫、各版藏經經文對照….		除了生產預定經文典籍外，與經文資料庫相關的應用軟體也是非常重要的。		

註：若程序多於七個，請複製本表使用

調查人：陳以儒　　調查地點：法鼓佛教研修學院、中華電子佛典協會　　調查日期：2007/12