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

製作日期：2005/11/30

計畫單位：中華電子佛典協會執行／中華佛學研究所協辦

計畫名稱：佛典數位典藏內容開發之研究與建構

計畫簡介：

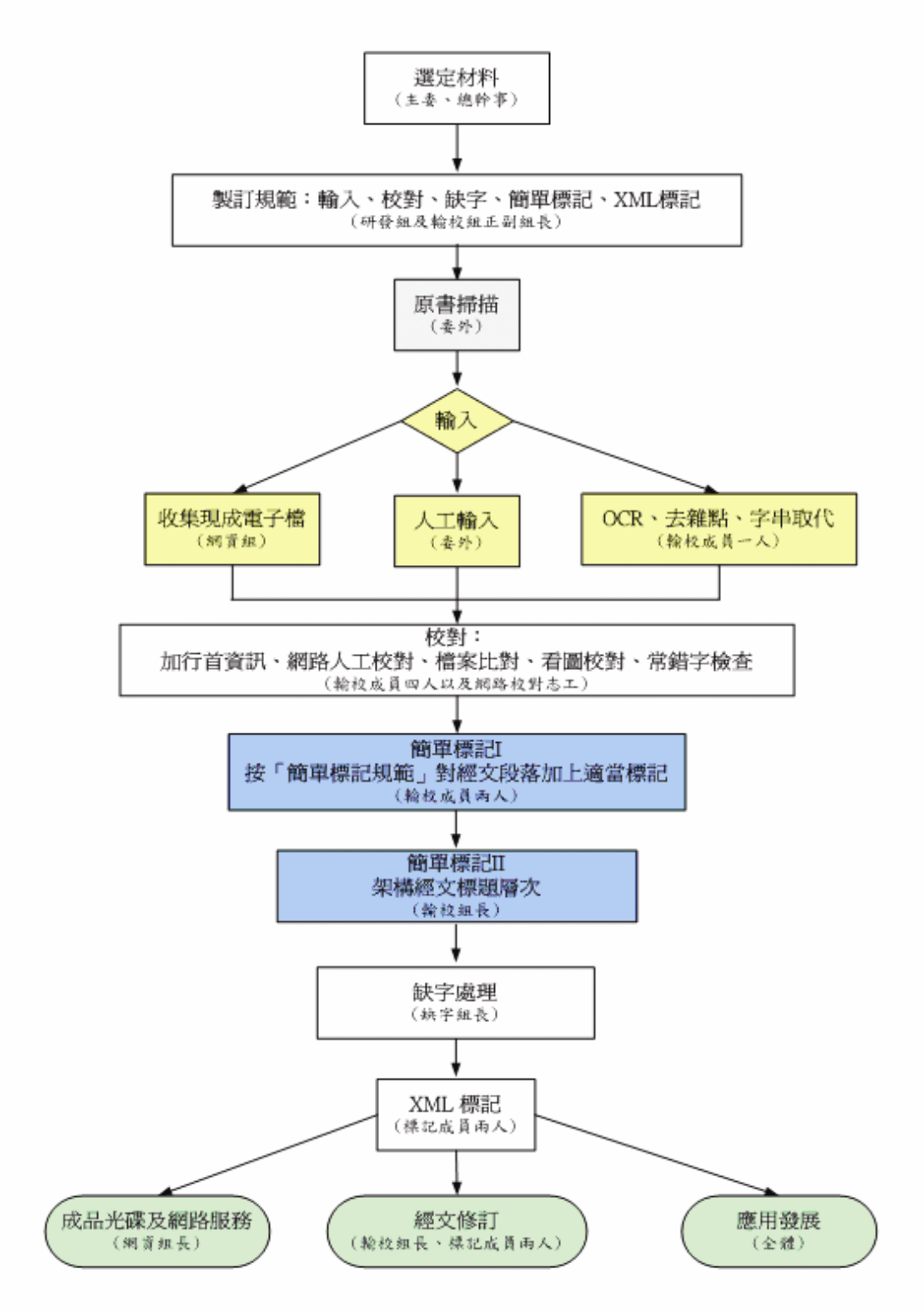
在 1998 年之前，經由眾人努力，網路上集結了不少佛教經典，也因此帶動佛典電子檔輸入熱潮。大家的目標主要著重於網路應用，比如將這些電子檔置放於 FTP 供人免費下傳，或是透過 GOPHER、WWW 方便使用者瀏覽，最近更在 WWW 上提供檢索查詢功能。另一發展是將電子檔包裝設計成電子書，使經文呈現更加精緻。所有努力，莫不希望透過網路，使佛典普及，讓更多人同霑法益，並利用電腦拓展佛典的應用範圍及閱讀方式。

有計畫之經典輸入始於網路上電子佛典討論版（BudaTech）的朋友草擬了電子版大藏經輸入計畫；後在蕭鎮國先生提供二十五冊 CCCII《大正藏》電子稿，並於 1997 年 1 月 6 日在台灣大學佛學研究中心（以下簡稱台大佛研中心）成立 25T 小組後，著手進行大規模的藏經電子化作業。台大佛研中心由釋恆清法師籌募所需經費，「北美印順導師基金會」與「中華佛學研究所」全力支持贊助，於 1998 年 2 月 15 日假法鼓山安和分院舉辦籌備會議，並於當日正式成立「中華電子佛典協會」（Chinese Buddhist Electronic Text Association，簡稱 CBETA）。1998 年 9 月 30 日，CBETA 更與日本「大藏出版株式會社」正式簽約，授權使用《大正新脩大藏經》，並同意其發行電子版之網路版與光碟版，使 CBETA 得以學術界通行的《大正新脩大藏經》爲底本，完成第一冊到第五十五冊，以及第八十五冊之佛經電子化作業。

CBETA 下分六組：研發組、網資組、輸校組、標記組、缺字組、財務組。研發組負責數位化過程之研究開發及製訂標記規範；網資組負責網站維護、開發讀經器、撰寫作業工具軟體；輸校組、標記組、缺字組處理所有「文字」與「標記」相關業務；財務組則管理協會之財務運作。

協會宗旨爲：一、收集所有漢文佛典，以「佛典集成」爲目標；二、研發佛典電子化技術，提昇佛典交流與應用；三、利用電子媒體之特性，以利佛典保存與流通；四、期望讓任何想要閱藏的人都有機會如願以償。

CBETA 經文數位化工作流程圖



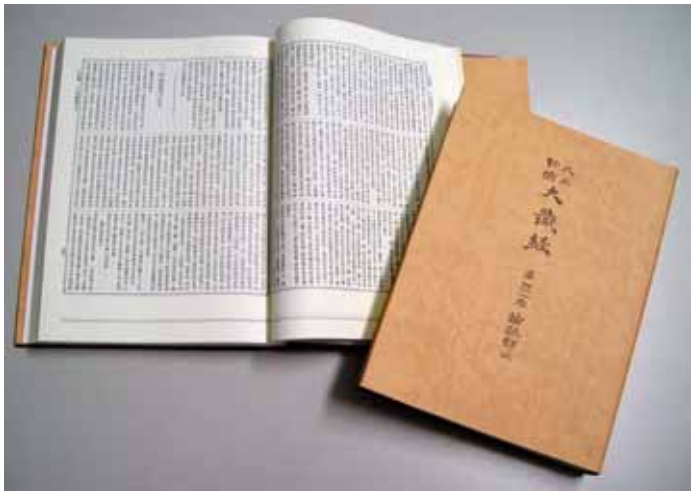
製作日期：2005/11/30

製作單位：內容發展分項計畫

數位化工作流程說明

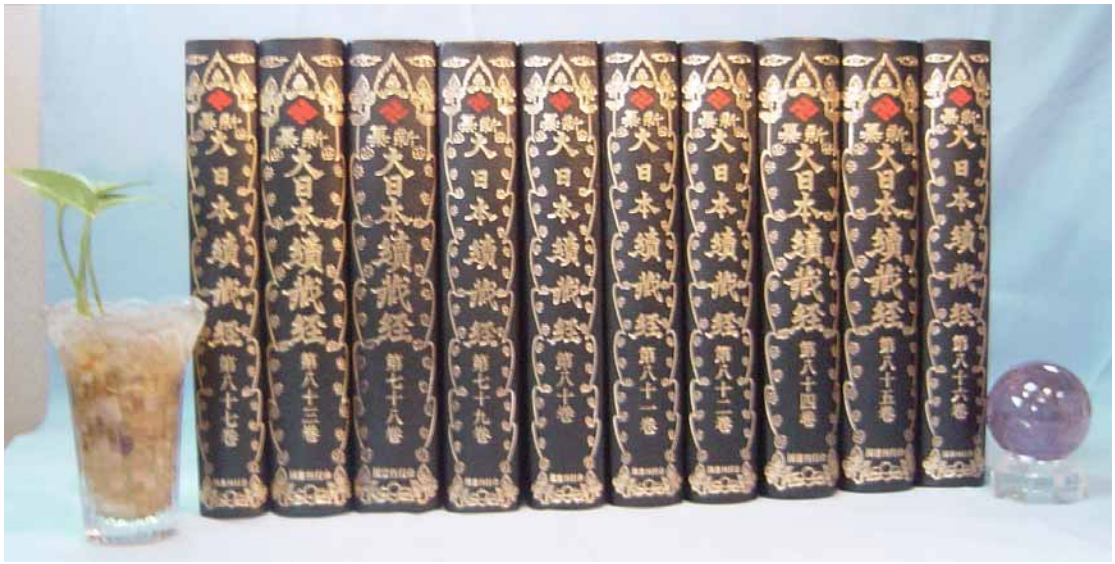
一、選定材料（執行單位：主委、總幹事）

計畫以「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《大正新脩大藏經》（以下簡稱《大正藏》）爲底本（圖一），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，目前已全數完成。



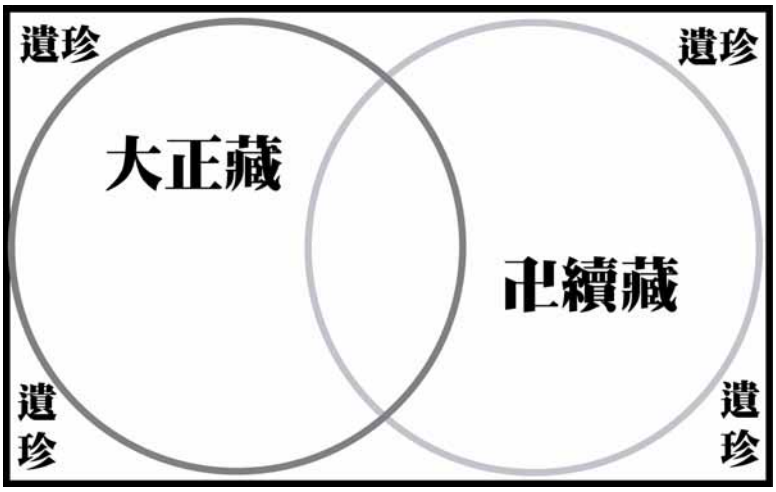
圖一、《大正新脩大藏經》

現正進行數位化之藏經爲《卍續藏》（圖二），預計 2006 年底全數完成上線。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），此乃本計畫選定材料之優先原則。



圖三、《大正藏》與《卍續藏》之關係圖

二、制定規範（執行單位：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。此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（執行單位：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。現因人員及成本效益考量，委託外部廠商執行，成本約每頁１.５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。而計畫之圖檔用途，提供「OCR 辨識」使用，並備爲「看圖校對」查看，故再將圖檔由 300dpi 灰階 轉成 TIFF-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齊之原。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。不閱戶牖。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。每懼微言翳於殊俗。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。❾特留懷。每任以法事。以弘始十二年歲次上章❿閹茂。請罽賓三藏沙門佛陀耶舍。出律藏❷二分四十五卷。十四年訖。十五年歲次昭陽⓫赤奮若。出此長阿含訖。涼州沙門佛念爲譯。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。蠲華崇朴。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⓬長阿含經卷第一

⓭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花林窟。與大比丘衆千二百五十人俱。時諸比丘於乞食後。集⓲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。⓳消滅戲論。又知彼佛劫數多少。名號姓字。所生種族。其所飮食。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。卽從座起。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⓴泥洹
記生名號姓　哀㉑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㉒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。過去九十一㉓劫時世有佛名㉔毘婆尸如來至眞。出現于世。復次比丘。過去三十一劫有佛名㉕尸棄如來至眞。出現於世。復次比丘。卽彼三十一劫中有佛名㉖毘舍婆如來至眞。出現於世。復次比丘。此㉗賢劫中有佛名㉘拘樓孫。又名㉙拘那含。又名㉚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㊀ ❸[述]ー㊀ ❹辨ニ辯㊀* ❺韞ニ温㊀元 ❻齊ニ濟㊀元 ❼析ニ斤㊀ ❽闢ニ掩㊀元 ❾一ニ二㊀ ❿[赤]ー㊀元 ⓫長阿含經Dīrgha-āgama～Dīgha-nikāya ⓬後秦弘始年ニ姚秦三藏法師㊁ 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消ニ銷㊁* ⓲泥洹ニ涅槃㊁ ⓳戲ニ懇㊀ ⓴〈佛〉十言㊁ ㉑劫～Kappa. ㉒毘婆尸～Vipassin. ㉓尸棄～Sikhin. ㉔毘舍婆～Vessabhū. ㉕賢劫～Bhaddakappa. ㉖拘樓孫ニ拘留孫㊁～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。

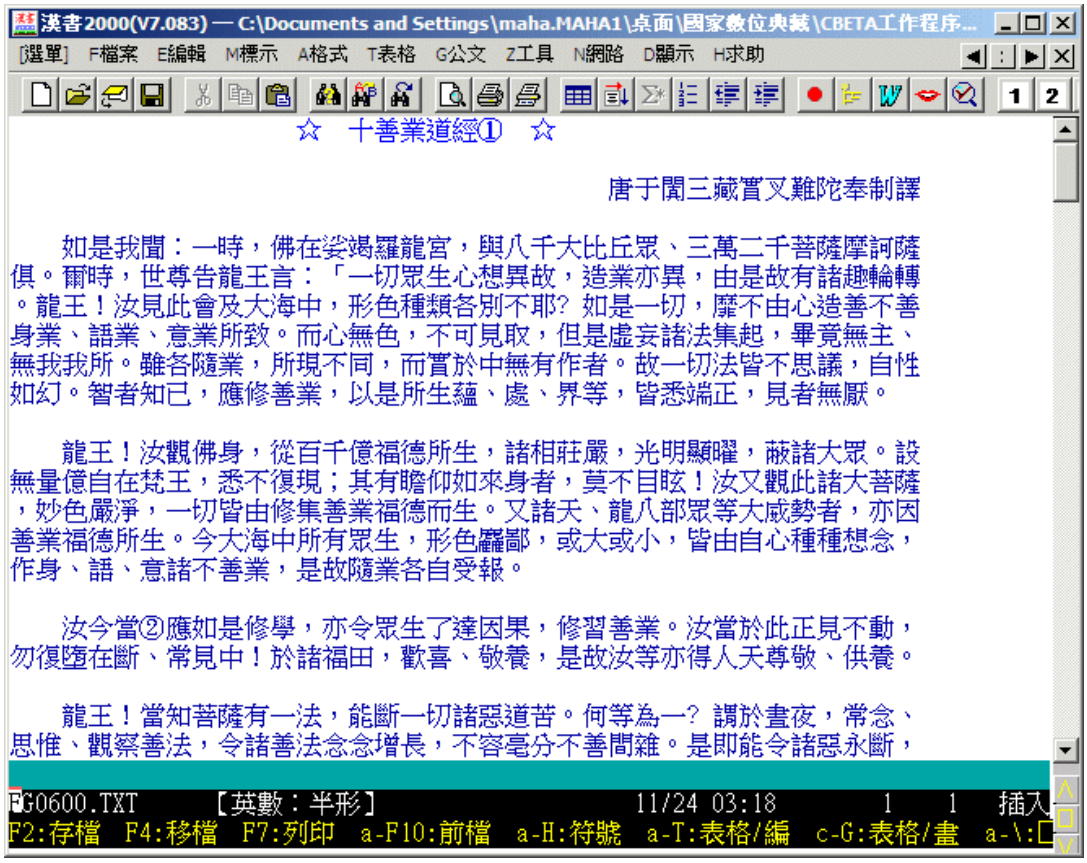
該計畫之輸入方法有三種，分別爲收集現成電子檔、人工輸入，以及 OCR 圖檔辨識。決策方式爲：如一佛典已有現成電子檔，則該電子檔可供日後檔案比對使用；無電子檔又難以透過 OCR 辨識之文字，如手抄本與刻版經文，則採用人工輸入。

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：（執行單位：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。

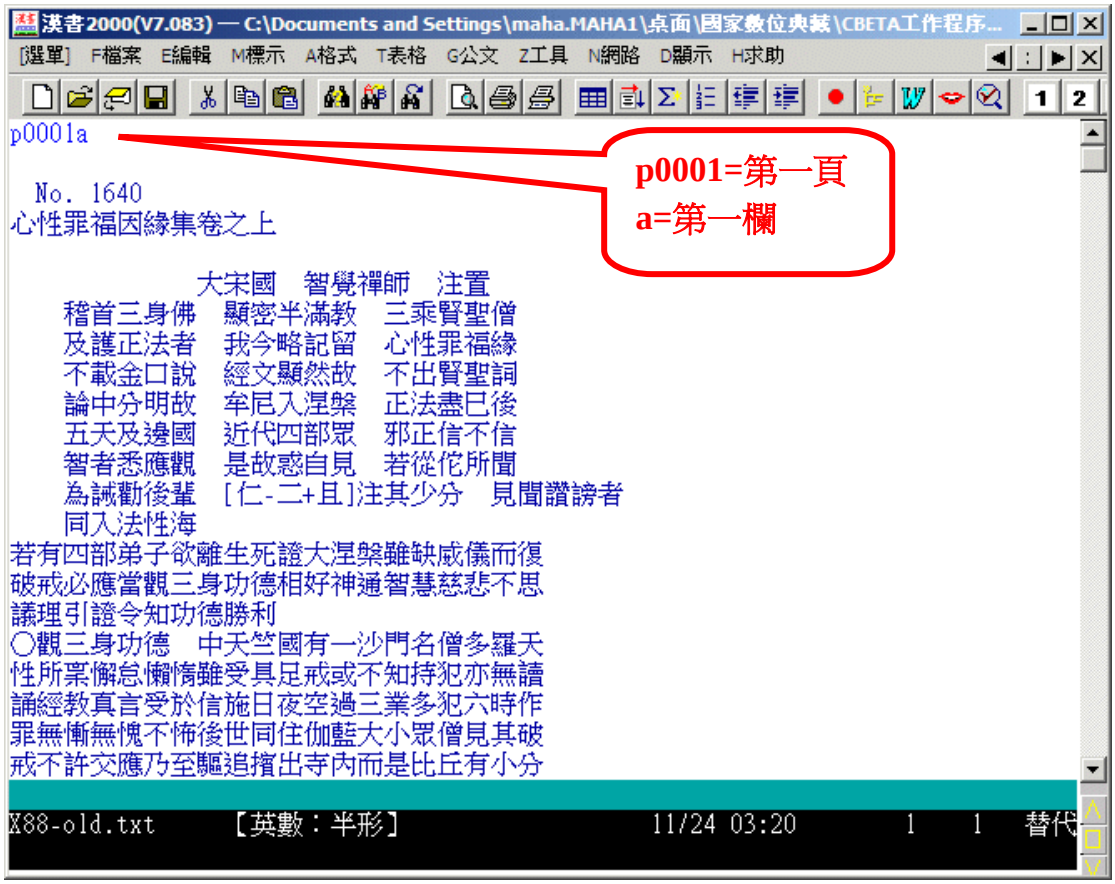


圖五、《大正藏》之現成電子經文

（二）人工輸入：（執行單位：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、欄資訊（圖六），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。

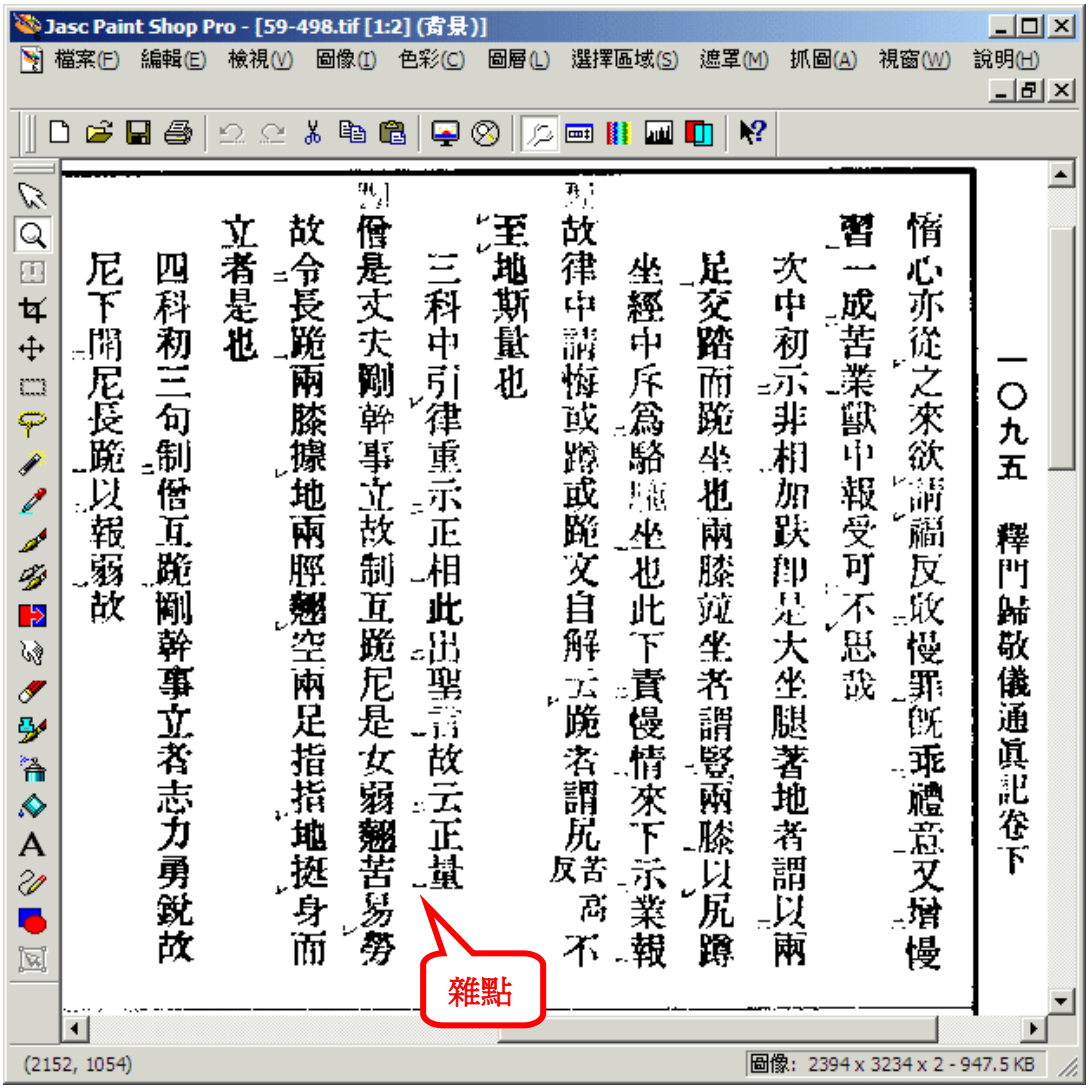


圖六、委外人工輸入產出之電子檔

（三）OCR 圖檔辨識：（執行單位：輸校組成員一人）

1.去除雜點

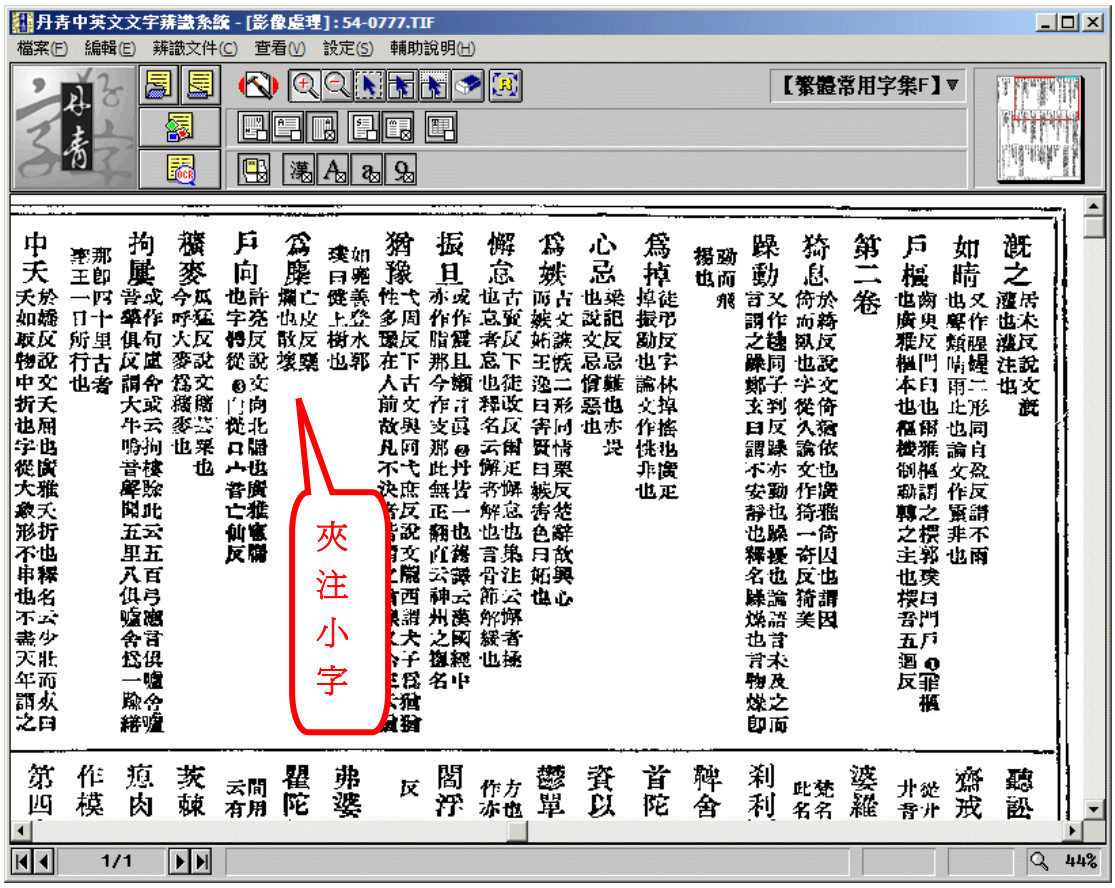
經文書上常有異於文字之讀音符號與注釋標記（圖七），嚴重影響 OCR 辨識之判讀結果；故掃描後之經文圖檔，須先以程式去除雜點，產生一新 TIFF 圖檔。



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 程式進行辨識（圖八），產出一份經文之「純文字檔」。

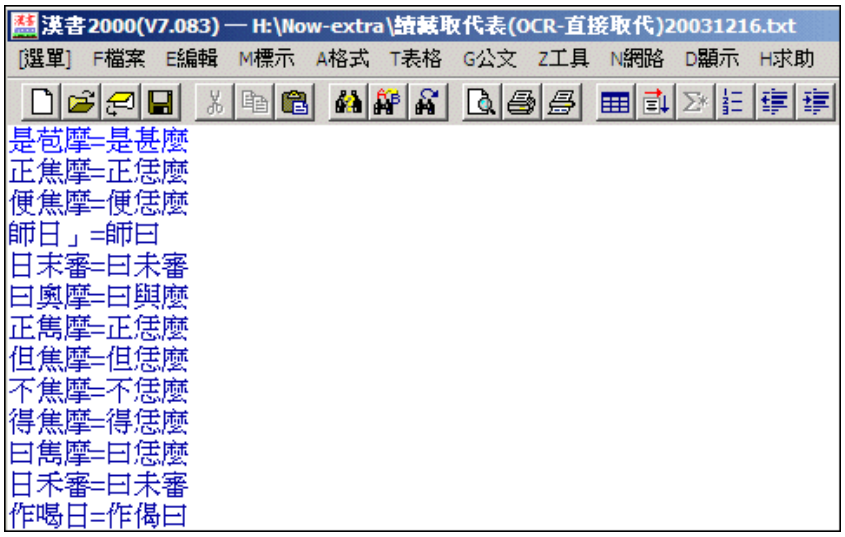


圖八、丹青 OCR 操作介面

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（圖八）。

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（執行單位：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。

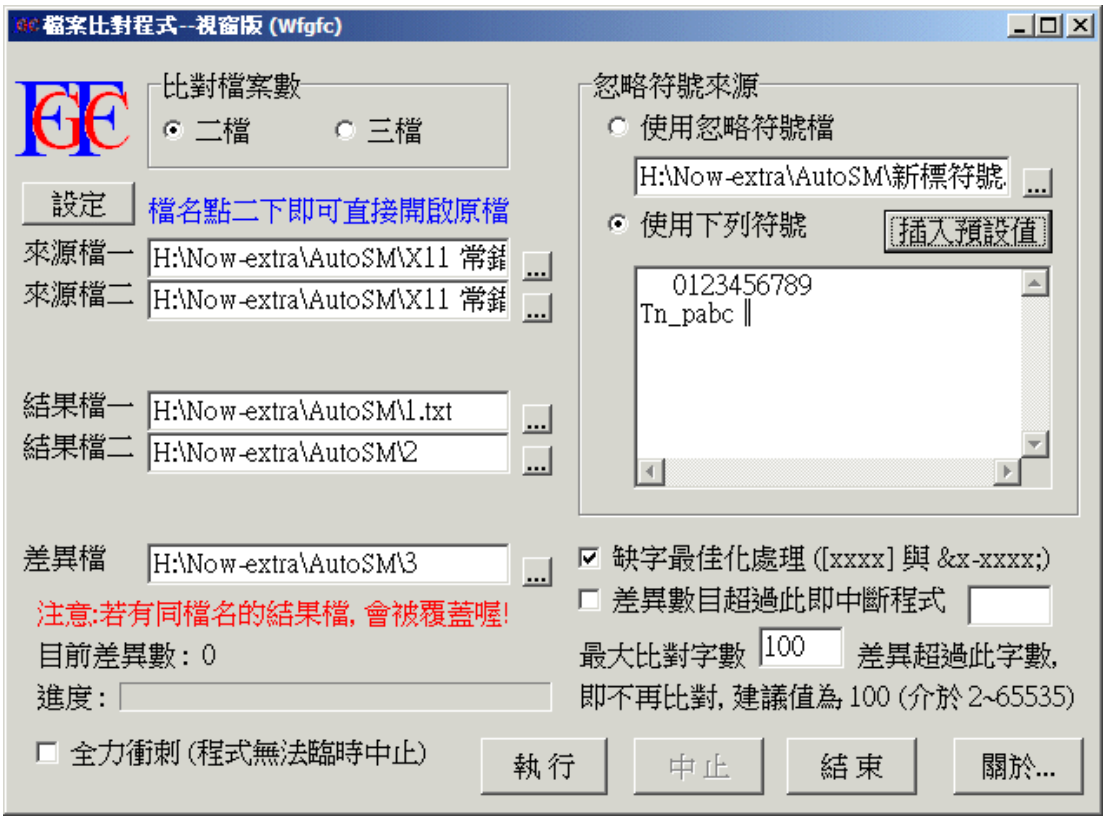
網路校對後之 OCR 經文，正確率可提升爲 98%。

（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。

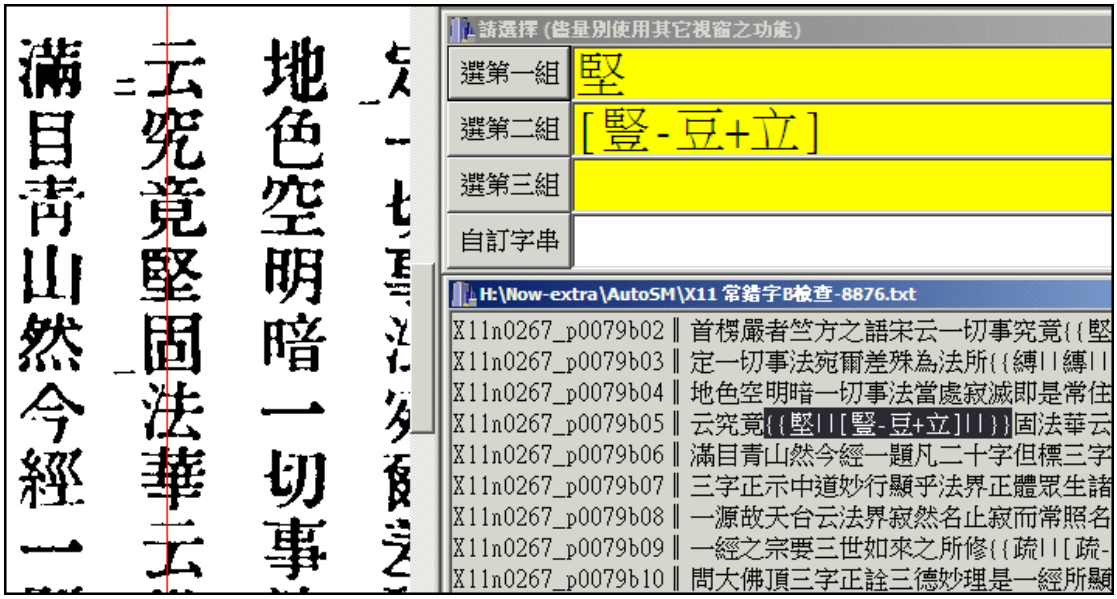
將合併成大檔之兩檔匯入「檔案比對程式」（圖十），執行第一次兩檔比對。比對後產生一個主要差異檔。以《大正藏》而言，平均每冊約產生兩萬個差異。



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。

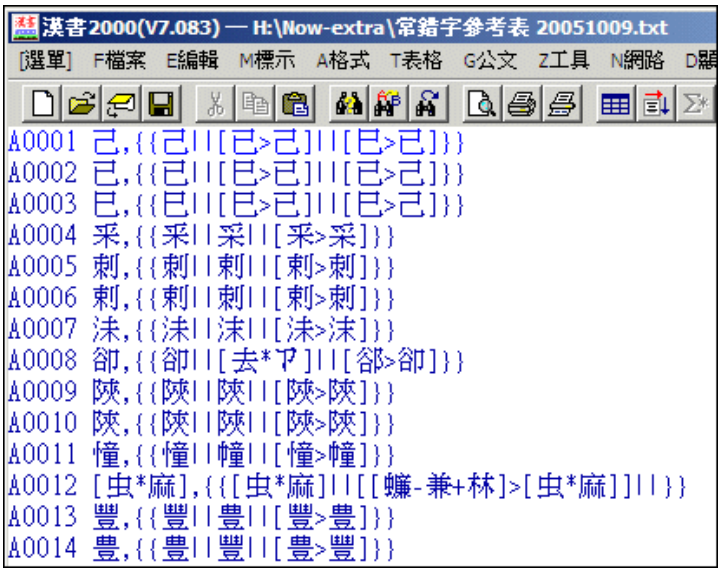


圖十一、看圖校對程式介面

此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。



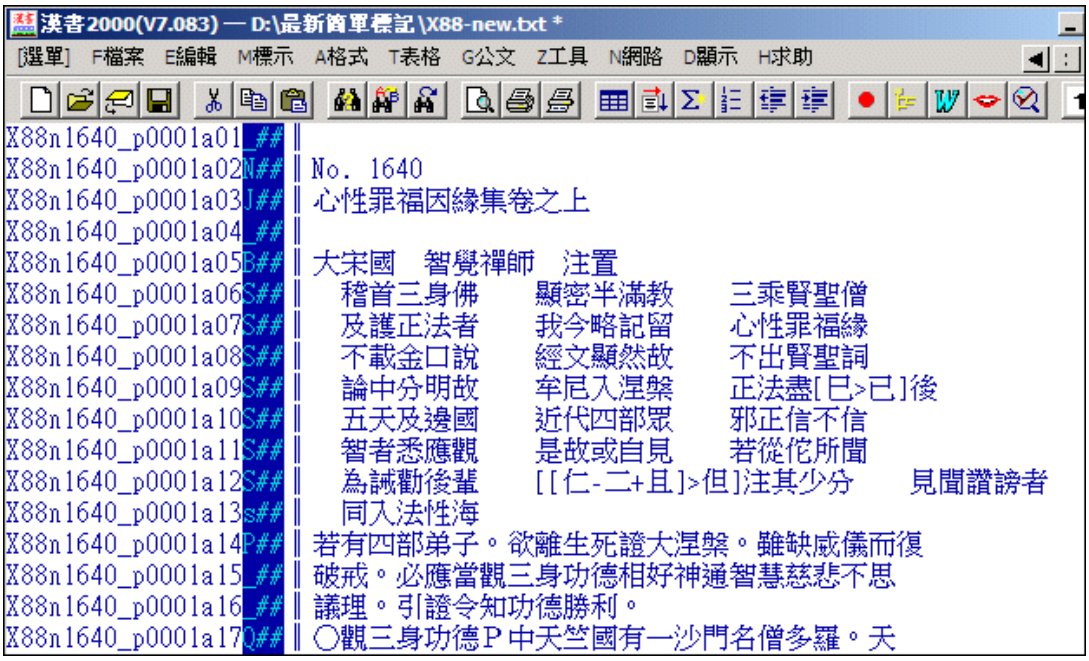
圖十二、常錯字參考表

六、標記

（一）簡單標記 I：（執行單位：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：（執行單位：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（執行單位：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。

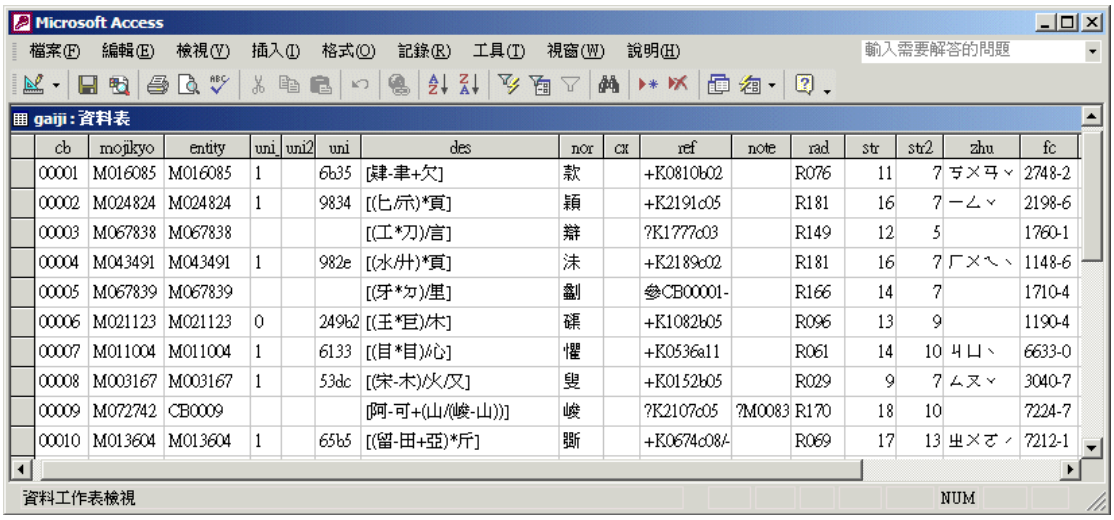
該組字法含「*」、「/」、「@」、「-」、「+」、「?」六個半形基本符號，及「(…)」、「[…]」兩組半形分隔符號。

舉例說明如下：

表 1、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

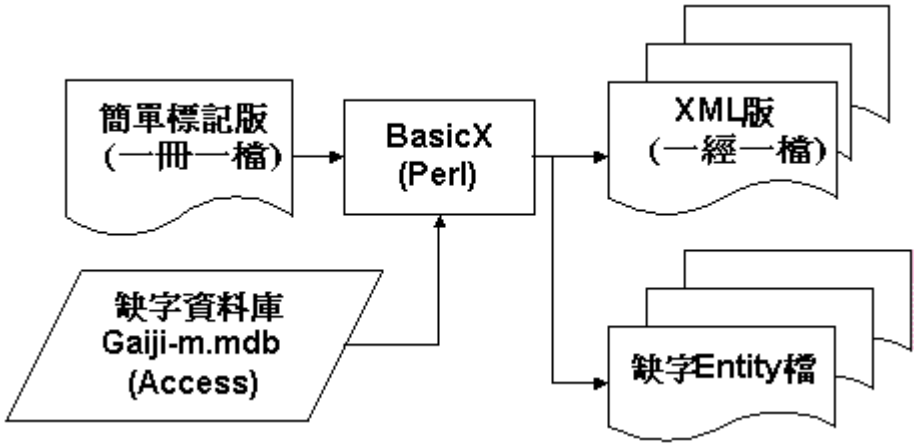
記錄缺字後，並將缺字相關資訊，包括注音、筆畫、部首、通用字、Unicode…等建構成漢文佛典缺字資料庫（圖十五）。



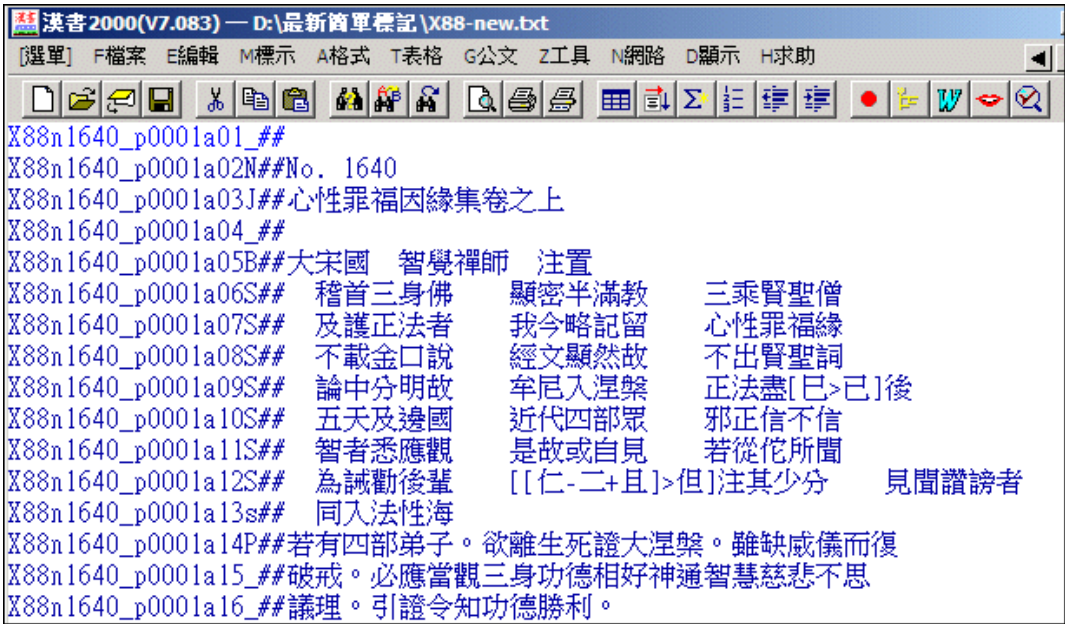
圖十五、缺字資料庫畫面

八、XML 標記（執行單位：標記成員兩人）

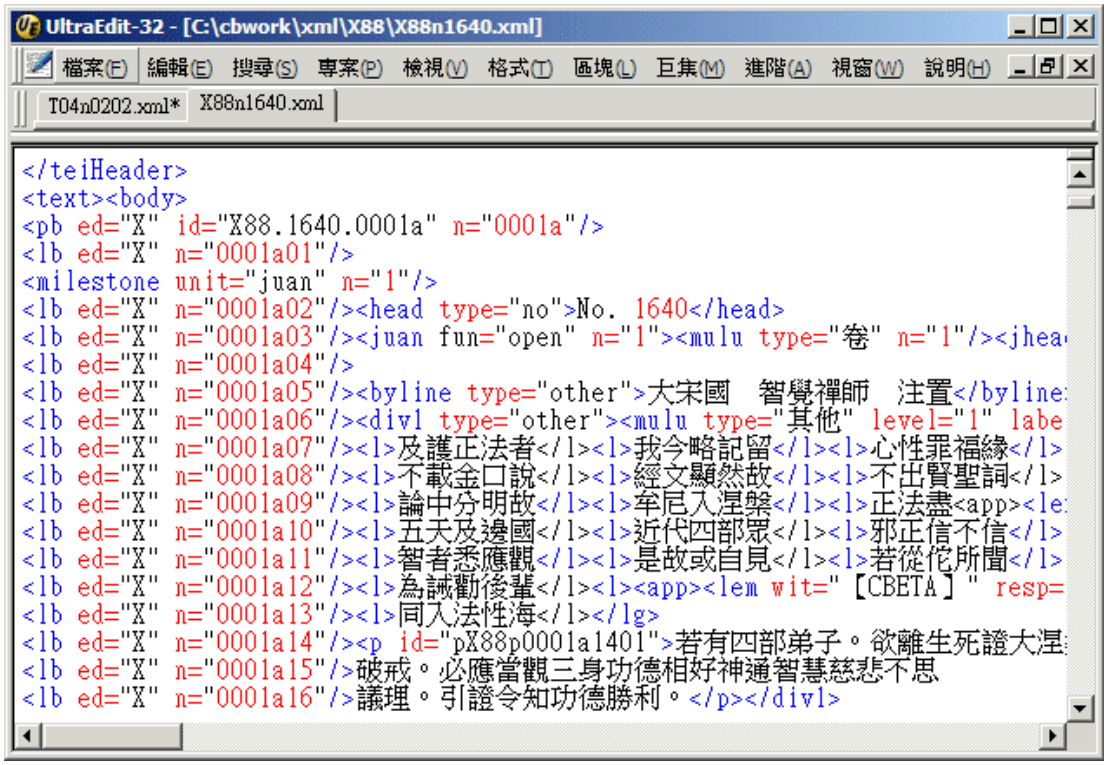
經簡單標記、缺字處理後之經文，以程式將簡單標記經文轉爲 XML TEI 標記經文（圖十六至圖十八）。



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。

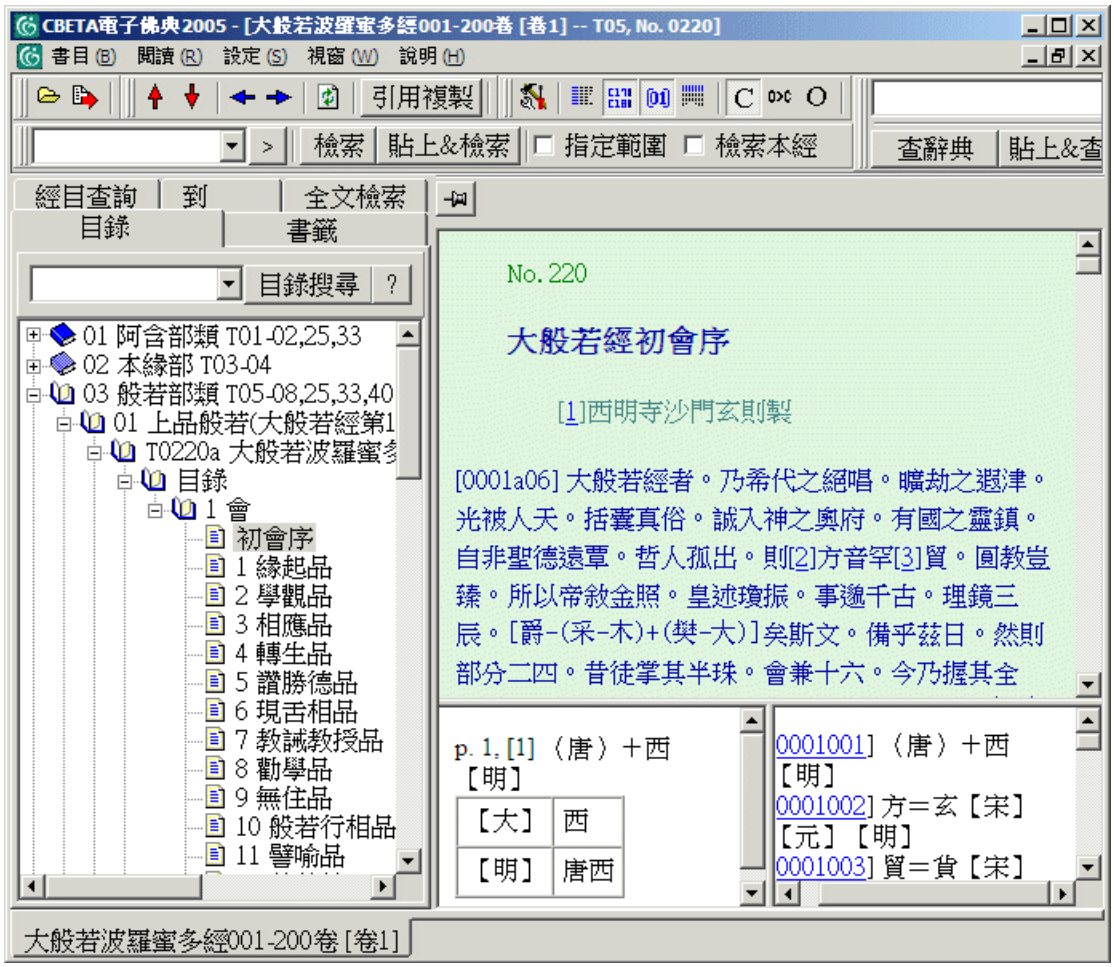
九、應用服務

（一）成品光碟與網路服務（執行單位：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

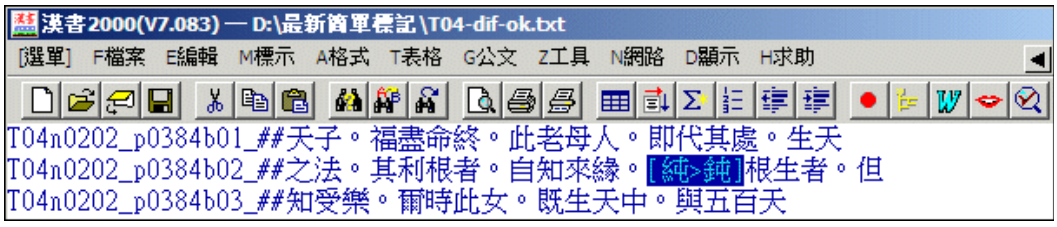
圖二十、CBETA 每年發行之光碟



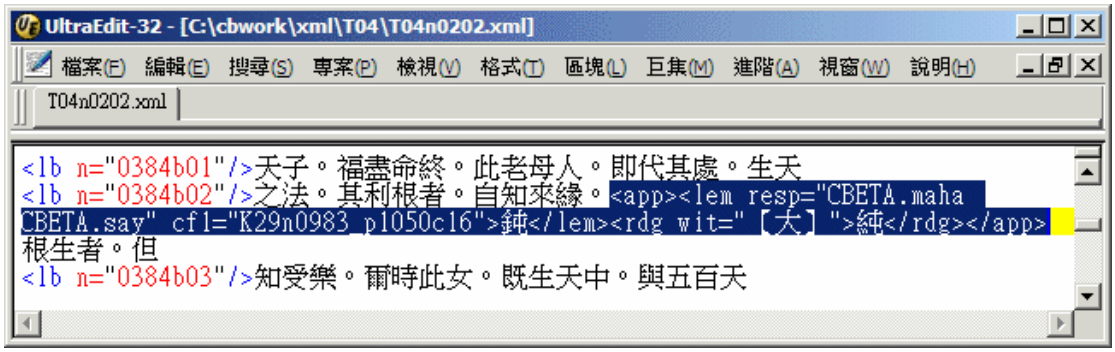
圖二十一、優異檢索及閱覽功能的 CBReader

（二）經文修訂（執行單位：輪校組長、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。



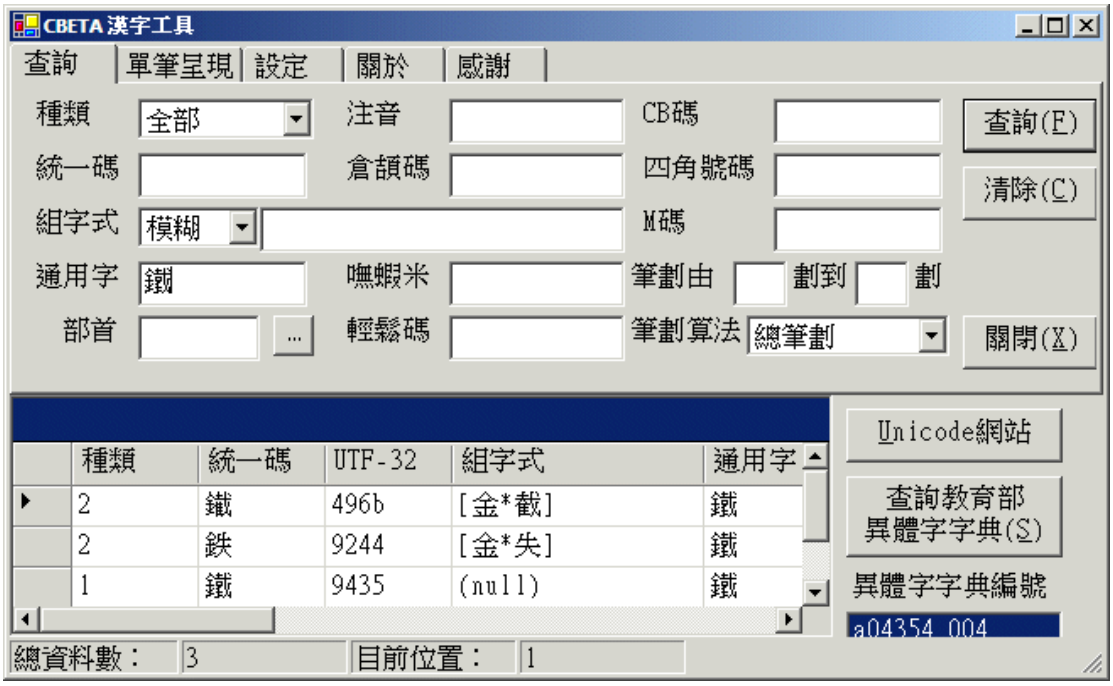
圖二十二、簡單標記版修訂



圖二十三、XML 版修訂

（三）應用發展（執行單位：全體）

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：數位典藏國家型科技計畫　內容發展分項計畫

中華電子佛典協會

中華佛學研究所

※ 文字撰寫：中華電子佛典協會　輸校組組長 吳寶原

數位典藏國家型科技計畫　內容發展分項計畫

漢籍全文主題小組助理 謝筱琳

※ 圖片拍攝：數位典藏國家型科技計畫　內容發展分項計畫

漢籍全文主題小組助理 謝筱琳、林淑惠

※ 圖片提供：中華電子佛典協會

※ 圖文編輯：中華電子佛典協會　輸校組組長 吳寶原

數位典藏國家型科技計畫　內容發展分項計畫

漢籍全文主題小組助理 謝筱琳

致謝：

感謝「佛典數位典藏內容開發之研究與建構」計畫主持人　杜正民老師、中華電子佛典協會輸校組組長吳寶原先生撥冗指導及提供實地拍攝與簡介編寫。並感謝中華電子佛典協會其餘相關人員之協助。