中國佛教寺廟志數位典藏系統之建置

洪振洲
法鼓佛教學院副教授

法鼓佛學學報第12期　頁145-187（民國102年），新北市：法鼓佛教學院
Dharma Drum Journal of Buddhist Studies, no. 12, pp. 145-187 (2013)
New Taipei City: Dharma Drum Buddhist College
ISSN: 1996-8000

摘要

佛教於東漢時始傳入中國本土，歷經兩千多年的發展，已經成爲中華文化中不可或缺的一部分。於此長達千年的時間內，佛教歷經多次的興衰與各分支宗派的開展與合流，期間所出現的高僧大德、於各地所建立的佛寺塔院與相關的佛教活動，可說是不可勝數。對於如此豐富的文化活動，官方正史的記載並不夠詳實，導致研究者多半仍需借助佛教典籍中紀錄來還原當時的眞實情形。於 16 世紀至 20 世紀之間，由於佛法的興盛加上大量居士的護持，促使私版藏經的刊刻成爲一件可實現的理想。在此一條件的協助之下，產生了大量佛教寺廟史志型態的書籍，也就是一般通稱的佛寺志。佛寺志的內容有別於一般的論述性文體，而是將各地蒐集 相關於該佛寺的相關著作，統整編輯而成的合集作品。雖然這些佛寺志並非原創作品，但反而更能代表各時代關於該佛寺的文獻史料紀錄，也因此佛寺志反而成爲研究晚期中國佛教發展的重要參考資料來源。而這些巨量的佛寺志書籍，於近代匯集成兩部叢書，分別是於 1980–1985 年間由杜潔祥主編之《中國佛寺史志彙刊》，此部叢書共有 110 冊，前後共分三輯，由台北明文書局與丹青圖書發行。而另一部是：2005 年由張智主編，杭州廣陵書社出版的《中國佛寺志叢刊》，內容共 130 冊。爲使此兩部具有高度宗教與歷史研究價值之典籍能廣泛爲世人所知與利用，並達到永久典藏之目的，於西元 2007 年，法鼓佛教學院圖書資訊館啓動「中國佛教寺廟志數位典藏」的專案計畫。專案的主要目的，就是要將此兩部叢書的內容，進行高品質之全文數位化處理，並藉由數位媒體的便利性，公開爲世人所用。本文之目的即以爲介紹專案數位化之過程、處理方式、介面之功能與未來可能的應用與發展。讓相關人士於使用此一數位典藏內容時，更能清楚其各方向的面貌，或於其他數位專案能吸取本專案之經驗，作爲開發時之參考。

目次

關鍵詞
中國佛寺志叢刊、中國佛寺史志彙刊、數位典藏、文本標記、TEI

* 收稿日期：2013/01/17，通過審核日期：2013/04/23。
佛寺數位典藏專案悉由中華佛學研究所與杭州徑山寺（徑山寺與徑山集兩志的數位化）資助之下得以完成，僅以此感謝。此外，也感謝歷年來參與本專案的相關同仁的辛勞付出。

一、前言

佛教始東漢時傳入中國本土，歷經兩千多年的發展，已經成爲中華文化中不可或缺的一部分。於此長達千年的時間內，佛教歷經多次的興衰與各分支宗派的開展與合流，期間所出現的高僧大德、於各地所建立的佛寺塔院與相關的佛教活動，可說是不可勝數。對於如此豐富的文化活動，官方正史的記載並不夠詳實，導致研究者多半仍需借助佛教典籍中紀錄來還原當時的眞實情形。在歷代集結的大藏經中，有許多記載佛教歷史的典籍、也有以弘揚各代高僧行誼所編撰的僧傳，這些都是了解佛教歷史十分重要的資料，而另一重要的參考資料來源，就是於明、清兩代，民間大量開始編撰的佛寺志。

於 16 世紀至 20 世紀之間，由於佛法的興盛加上大量居士的護持，促使私版藏經的刊刻成爲一件可實現的理想。大部分的佛寺志，就是在這樣條件下所產生。佛寺志的編撰目的，主要是以地域資源爲敘述的核心，加以描述於該區域發生的佛教活動。大部分的佛寺志內容是以單一佛寺爲主要的描述對象，但也不乏以一個區域，或一座山頭爲範圍所編撰的佛寺志。佛寺志的內容，主要紀錄佛寺的歷史沿革、附屬地域空間發展，與撰文之時，佛寺發展情況等事。而曾於該寺院住錫之歷代祖師、文人雅士之軼事，也通常會成爲該寺志的紀錄對象。因此佛寺志的內容除用於佛教史研究之外，也見於歷史、經濟、甚至醫學的研究之中，其重要性不可謂之不大。以文體而言，佛寺志的內容有別於一般的論述性文體，或依年代條列式的記事體裁，而採用將不同文類的文本（傳記、詩、散文、地圖、銘文、序跋、圖像等），匯集而成的作品。與其說是一個單一的創作，不如說是當時將所有各地蒐集而來的相關著作，統整編輯而成的合集作品。作者本身的貢獻多在整理與編輯，但有時也會有一些補充的資料。

近代對於佛寺志主要的匯集作品，主要爲兩部於 20 世紀所完成的木刻版編輯叢書。第一部爲：1980 年至 1985 年由杜潔祥

主編之《中國佛寺史志彙刊》，此部叢書共有 110 冊，前後共分三輯，分別由台北明文書局（前兩輯）與丹青圖書（第三輯）發行。而另一部是：2005 年由張智主編，杭州廣陵書社出版的《中國佛寺志叢刊》，內容共 130 冊。110 冊的《中國佛寺史志彙刊》，內容共包含 100 部寺志，而《中國佛寺志叢刊》的 130 冊內容中，收錄 197 部寺志。兩叢書所收錄的所有寺志當中，共有 60 部寺志是完全相同或僅有小部分差異。因此扣除重複的部分之後，我們可認爲內容共有 237 部具參考價值的寺志。[1] 由於此二冊書籍篇幅極爲龐大，導致發行量較少、典藏地也不多，造成研究者參閱上的阻礙，實爲可惜。因此法鼓佛教學院圖書資訊館於西元 2007 年開始，正式啓動「中國佛教寺廟志數位典藏」[2] 的專案計畫（以下簡稱佛寺志數位典藏）。專案的主要目的就是要將此兩套叢書所錄之 237 個寺志，進行高品質之全文數位化處理。其目的就是要藉由數位媒體的便利性，爲使此兩部具有高度宗教與歷史研究價值之典籍能廣泛爲世人所知所用，並達到永久典藏之目的。

除了將兩部寺志叢書以國際標準完成數位化作業，提供高品質的數位檔案之外，專案也提供了一個以網頁爲基礎的閱讀介面。在介面中，除了提供寺志的全文、掃描的木刻版影像圖之外，也辨識出寺志內所有的人名、地名、時間，並提供詳細的參考資料與中國年與西元年的對照表。以便讓對於中國歷史細節並非十分熟悉的讀者，也能領略通達寺志之內容。在後續章節中，我們將以針對專案數位化之過程，處理方式，介面之功能，程式

1 此兩部叢書的內容多有重複，但也有其相異之處。其書目與內容之詳細比較，可參閱：Marcus Bingenheimer（馬德偉），〈中國佛寺志初探及書目研究〉，《漢語佛學評論》2，頁377-408。

2 於筆者撰寫本文之時，佛寺志專案將正式邁入第六個年頭。目前，所有寺志的影像檔已經掃描完成，並且公開於專案網站之上，供各界下載，目前已有17部寺志全文已有高品質的標記成果。專案後續將陸續完成其他寺志的標記作業。 詳情請參閱：法鼓佛教學院，「中國佛教寺廟志數位典藏」，2012/12/22，http:// buddhistinformatics.ddbc.edu.tw/fosizhi/。

架構與未來可能的應用與發展，作一個深入淺出的完整介紹。

二、佛寺志內容數位化程序

數位典藏建置之主要目的，就是將重點歷史文物的全貌以數位方式保存下來。但對於不同型態之文物，典藏策略自然有所不同。而此專案的典藏標的——《中國佛寺史志彙刊》與《中國佛寺志叢刊》，乃是屬於具有豐富內容的書冊。對於此類別的物件，以數位化全文與掃描圖片並陳，做爲典藏標的，自然是最佳的保存方式。

(一) 佛寺志專案數位化工作重點

在實際進行專案的數位化工作之前，首先必須先針對專案的範圍與產出目標，加以討論與規劃，以防專案進行時，出現偏離主題之情形。以本專案而言，數位化範圍是兩部叢書中，不重複的237個寺志。而數位產出目標，則是數位化文字內容與保留圖文編排，並提供使用者於網路上直接閱讀的可能性。

爲使數位化之全文，也具有長久保存之可能性。因此本專案針對數位化之文字內容，有別於一般以僅記載文字內容的數位化方式，而採用TEI (Text Encoding Initiative)[3] 標準的XML (Extensible Markup Language) 語言，針對文章內容進行標記，用以表達文字之外的其他重要訊息。採用了XML / TEI之後的文本範例如下圖一，我們可以發現到，除文獻原有的文字之外，數位內容中多了許多的標記。這些都是符合XML / TEI這個專門用來處理文字資料電子格式化的國際標準的標記。我們使用這個標準

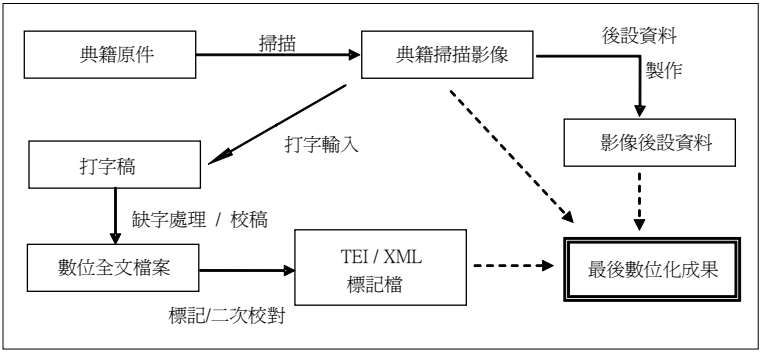
3 更多有關TEI的細節，請參閱：馬德偉，《TEI使用指南——運用TEI處理中文文獻》（台北：數位典藏與數位學習國家型科技計畫-拓展臺灣數位典藏計畫，2009）。

的規範，來記錄文件的分段層次架構、文體架構、引用參照、校勘附註、內容修訂資訊等重要訊息，使文件本身具有完整的自我解釋能力，也因此可應付未來可能的處理需求。

```
<p><placeName key="CN0320504T08AA"> 寒山寺 </placeName> 位於
<placeName key="CN0320000E03AA"> 江蘇 </placeName><placeName
key="CN0320504A05AA"> 姑蘇 </placeName><placeName
key="CN0320504A07AA"> 楓橋 </placeName> 下。<placeName
key="CN0320504A07AA"> 楓橋 </placeName> 在 <placeName
key="CN0320503A03AA"> 閶門 </placeName> 西七里地，與 <placeName
key="CN0320503Z03AA"> 長邑 </placeName> 合治爲水陸孔道，自古有名，南
北客經由，未有不憩此橋而題詠者。<date notBefore-iso="1604-01-31" notAfter-
iso="1605-02-17" key="j23069392307322"> 明萬曆三十二年 </date>（一六○四）築
<placeName key="CN0320504A07AA"> 楓橋 </placeName> 隄；<date notBefore-
iso="1731-02-07" notAfter-iso="1732-01-26" key="j23533322353685"> 清雍正九
年 </date>（一七三一）修築 <placeName key="CN0320503Z03AA"> 長洲縣 </
placeName> 運河塘。<placeName key="CN0320504A07AA"> 楓橋 </placeName>
舊作「<placeName key="CN0320504A07AA"> 封橋 </placeName>」，<persName
key="A018277"> 王 <roleName> 郇公 </roleName></persName> 居。
```

圖一、佛寺志專案 XML 標記文件片段

本專案除以標記紀錄文章原有樣貌之外，我們也考慮如何減低讀者對於此艱澀難讀的古籍的閱讀困難。因此我們決定在製作數位化文本的過程中，也爲這些文本加入新式標點符號，並且對於寺志中所有的人名、地名、時間也加以識別，並於原有數位文獻中加上相關標記（如圖一中所出現的 persName、placeName 標記）。我們這些人時地實體的背景資料，則統一記錄於資料庫之中，並由程式提供適當處理，以便於文獻閱讀的環境中，提供相關的資訊，讓讀者在瀏覽古文的同時，可以即時得知人時地相關的背景資料，減低閱讀的困難，進而達到佛教文獻的加值與廣泛流通的目的。這些標記資料除了提供與實體相關背景資料的呈現功能外，也讓我們有機會可以針對人時地的資料，進行統計分析。在後續章節中，我們將以一些簡單例子，說明這些標記文獻的簡單分析結果。爲使對於文獻數位化流程有全面性理解，茲將詳細的文獻數位化流程，以圖二表示。



圖二、佛寺志數位典藏專案數位化流程示意圖

依圖二所示，本專案之典藏目標以文字與原書掃描影像並存爲主。因此古籍清單確定後，將會將古籍內容加以掃描。掃描結果將成爲本專案成果之一部分，並將製作影像的後設資料，以便提供影像完整的描述資料。此外，文獻經過掃描之後，將繼續送交打字，以取得數位文字檔。數位文字檔將經過多次的校對，以確保其品質。由於古籍文獻常含有許多現在社會少用的難字與罕見字，我們也必針對這些字，進行特別處理程序。經此程序後，才能製作出一個內容文字錯誤率低的全文文字檔。而如上所述，文件內部豐富的描述資料，都是在製作 TEI / XML 標記的過程中加入的。此步驟最花時間，但也最爲重要。待標記完成後，後續其他的相關應用，如製作網頁閱讀介面、轉換爲 PDF 格式、電子書或其他更多的應用方式，都將以此標記完成之文獻作爲主要材料。因此此階段的重要性是不可忽視的。在後續各小節中，我們將針對各數位化工作環節的細節加以描述，讓讀者可更明瞭相關細節，並通曉數位化成果之使用方式。

(二) 文獻掃描

由於古文獻具有稀有性、易毀性的特性，以掃描的方式將重

要文獻數位化，不僅可以讓資料可以易於保存，也是讓文獻流通的十分有效方法之一。因此本專案計畫中《中國佛寺志叢刊》與《中國佛寺史志彙刊》兩套寺志叢書的掃描圖檔，是專案的重點成果之一。我們將兩套書籍內容分頁掃描成圖檔，以便於保存及傳播。但由於文獻本身也是影印檔案，所以此次掃描亦受限於原稿的影響，部分圖文無法十分清晰。圖三爲掃描結果的範例圖檔。

圖三、原書掃描成果圖檔範例

《佛寺志》文獻爲單色印刷，大多爲文字，少部分附圖亦爲木刻畫，沒有中間灰色的層次，所以全部可以歸類爲「線條稿」，也就是僅有黑白兩色的出版品。掃描線條稿時，多半將每一個畫素（pixel）以 1 個位元（bit）來表示黑與白的分隔，[4] 但爲謹愼起見，本次掃描都以「灰階」（gray）處理。改以灰階掃描後，每畫素改用 8 個位元表示，檔案大小隨之變大 8 倍，但仍在可接受範

4 由於線條稿本身僅有黑與白的分別，因此使用1bit儲存（0或1）單一像素的內容便已足夠。

圍。而其優點是，顏色的層次擴充至 256 個層次，能保留圖檔的較多資訊。如此一來，若處理過程中，發現掃描檔案因爲墨色太淡，或底紙顏色太深，使掃描內容無法完整反應內容時，此時我們可以針對這些不滿意之掃描結果進行個別調整。本專案的掃描圖檔，採用 tiff 檔案格式作爲主要影像儲存格式，[5] 再視需要轉爲其他格式。此外，考量到未來可能的印刷出版需求，因此，所有的影像解析度皆設爲 400 ppi（pixels per inch）。如此一來，這些圖檔在不放大的情形下，也能以標準輸出的設定 200 lpi（line per inch）來進行印刷工作而不失眞。

掃描完成的文獻圖檔，雖然已經可以充分表達文獻的原貌，但爲使這些「數位物件」成爲學者可信任的參考文獻，我們必須將進行掃瞄工作時的詳細背景資料也完整的紀錄下來。在本專案中，圖檔的後設資料分爲兩個層次來處理。首先，在進行掃描之時，掃描設備自動爲每頁的掃描圖檔，加入 EXIF (Exchangeable Image File Format) [6] 紀錄，讓使用者可以明瞭該圖檔取得的掃描條件。由於 EXIF 對於圖片檔案的載體資訊與內容資訊並沒有紀錄，因此我們使用 MIX (Metadata for Images in XML Standard，XML 標準影像詮譯資料) [7] 的標準加以描述，以補 EXIF 資訊的

---

5 對於掃描圖檔的儲存格式部分，一般而言多半採用jpeg2000或tiff 格式作爲主要的影像壓縮格式。jpeg2000爲一種非破壞性壓縮檔案格式，方便數位儲存及傳播。tiff檔格式爲印刷業界採用已久的高階影像格式，但其壓縮比較小，因此產生的檔案較大，需要較大的空間儲存。由於本專案採用的掃描機不支援jpeg2000之影響壓縮格式，因此我們選擇以tiff 作爲主要儲存格式。

6 所謂的EXIF後設資料的內容，主要是紀錄數位裝置取得影像時的相關設定紀錄，其內容包含如光圈、快門、顏色校正、像素、掃描時間、座標等等相關訊息，詳細內容可參閱：Japan Electronics and Information Technology Industries Association. 2002. Exchangeable Image File Format for Digial Still Cameras: Exif Version 2.2, 2013/05/01, http://www.exif.org/Exif2-2.PDF。

7 MIX, 2012/06/06, http://www.loc.gov/standards/mix/.

不足。MIX 使用 XML 作爲載體，主要用於紀錄數位影像的語意資訊，包括圖檔檔名、圖檔所佔磁碟空間、圖檔格式、驗證圖檔完整性的 MD5[8]、圖片與元件的長寬尺寸、顏色數，圖檔掃描後處理、處理前圖檔資訊等等，茲將以 MIX 完成的後設資料內容，摘錄基本資訊部分如圖四所示。[9]

```
<mix:BasicDigitalObjectInformation>
 <mix:ObjectIdentifier>
  <mix:objectIdentifierType>Filename</mix:objectIdentifierType>
 <mix:objectIdentifierValue>luoyangqielanji_p0001.jpg</mix:objectIdentifierValue>
 </mix:ObjectIdentifier>
 <mix:fileSize>32394</mix:fileSize>
 <mix:FormatDesignation>
  <mix:formatName>image/jpg</mix:formatName>
 </mix:FormatDesignation>
 <mix:byteOrder use="system">little endian</mix:byteOrder>
 <mix:Compression/>
 <mix:Fixity>
```

圖四、佛寺原書掃描後圖檔後設資料範例部分摘錄

(三) 文字輸入與校對

文字輸入作業是典藏數位全文專案中，最基本但卻是非常重要的環節，其執行過程也需花費大量的人力與金錢成本。因此在決定要自行輸入文字之前，先確定是否已經有其他單位已經完成的可用數位物件以及這些物件的可取得性。最後，才考慮是否由自己進行文字輸入作業。

8 詳細內容可參閱：Rivest, R. 1992, The MD5 Message-Digest Algorithm. United States: RFC Editor, 2013/05/01, http://www.ietf.org/rfc/rfc1321.txt。

9 完整的圖檔後設資料，可於專案網頁之上，數位資料下載的內容中取得。

文字的輸入有兩個最主要的方式：一、以外包方式，交由專業打字行協助文字輸入。二、使用電腦光學辨識（Optical Character Recognition，簡稱 OCR），來進行文字內容自動辨認。[10] 由於 OCR 的效果取決於文稿本身的清楚程度，因此大部分情況下，外包打字仍是文字輸入的第一選擇。一般而言，以專業打字行輸入的錯誤率，約可以規範至千分之四或千分之五左右，但錯誤仍是免不了的，也因此後續校對作業將更形重要。

最有效率的校對方式，是請兩家不同的專業打字公司，針對相同的文稿，進行兩次打字輸入作業。取得兩份的打字稿後，利用電腦軟體互相比對，這可以快速找出大部分的輸入錯誤，並使錯誤率顯著降低。但是繕打兩份文字稿，將使成本上升爲兩倍，因此並非所有專案都能負擔。以此專案來說，考慮到預算額度的關係，我們僅繕打一份打字稿。但爲降低錯誤率，我們在收到打字稿後，邀請對於寺志有內容有興趣的義工，進行紙本的閱讀與電子檔內容的比對，這是我們進行的第一次校對。此後，於全文標記作業時，因工作人員需逐字閱讀，瞭解文義，此時也同時進行第二次校對。加上打字公司於打字交稿之前，會自行進行一次校對工作。因此一份文稿從製作到完成，共會經過三次校稿，如此可使錯誤率降至可接受之範圍。

（四）特字處裡程序

處理古代典籍文字稿的一大挑戰，就是文本中的特字數位化的工作。「特字」在此指的並非聲韻學定義下的「特字」，而是指：由於現代十分罕用，以致讀音及意義較不爲人知，不易以常見輸入法輸入的特殊字。換言之，「特字」乃是我們爲處理古代典籍，

10 當然也可以選擇自行打字輸入的方式，但因爲自行打字成本較高，因此唯有在處理絕對不可外流的珍貴手稿，或是文稿以專業語言寫成，找不到適當的打字廠商時，才會考慮此選項。

而對難以輸入的文字所專門使用的一種稱呼。從長期以來處理特字的經驗中，我們發現一個文字被判定爲特字，其原因大致包括：

1. 就外觀上難以判斷眞正讀音，例如：「匋」。
2. 由於該字使用範圍相當窄，以致無論讀音是否易於猜測、字形是否單純，時下輸入法仍多不收入該字，例如：「豶」。[11]
3. 字形複雜難以拆解。

特字處理並非新的議題，但截至今日，雖然數位化的範圍日趨廣大，但此問題仍然沒有一個完美的解決方案。特字處理最大的困難，仍在於「忠於文本」理念難於實現。原因在於，字形的差異無法以「是此字形」與「不是此字形」二分，而是一個連續體。從與電腦使用的字型最相似的木刻版印刷字體到相差最遠的手抄本，到底哪一些字是與電腦編碼字形之間是「一樣而可直接採用的」，判斷時因標準的訂定而異。簡言之，若以極嚴苛的眼光來看，手抄本每個字都可說是異體字，就算是再龐大字庫，也不一定能表示所有的字型。因此，以多年的特字處理經驗，我們決定以字庫容量極大，且國際間較認可的統一碼（Unicode）編碼字庫作爲主要的處理標準，並且以盡量不作特字處理爲主要原則，來減低工作負擔。

對此，我們採取先把特字分成具有統一碼編碼的字（也就是說，已經收錄在統一碼字庫內），及尚無統一碼編碼的字兩大類，再於此兩大類之下依情況做適當處理。對於已經有統一碼編的字，大部分情況下這些字都不需額外處理，直接採用與字形寫法原文相同的 unicode 編碼字即可。但統一碼編碼中，也收錄許多今日漢語中不常出現且不易於辨識之罕用字。在文獻中，直接使用這些 unicode 編碼字，當然可以最直接達到「忠於文本」理念，但也使內容難以閱讀。因此若這些罕用字能找到於現代漢語

11《說文解字》：「豚屬」專指一種特定種類的豬。

中相對應的常見用字時，則我們在數位文本之中，將同時紀錄原字與現代常見的通用字。[12] 例如：㝎（= 定），灋（= 法），銕（= 鐵）。[13] 對於尚無統一碼編碼的字，情況較爲複雜，分成以下三種情況討論：

一、該字與某一統一碼編碼的字僅有些許差別，例如：少一撇、多一點的字。這可能是因刻版慣例，書寫習慣產生的差異。此時，若該特字搭配前後文時，易於辨識者。皆直接以該統一碼編碼字取代，不另作紀錄。例如：在《徑山集》中，第 13 頁「中本龍湫，化爲寶所，國一禪師開山於天寶之初，特爲偉異。天作地藏，待斯人而後發；道成名震，召歸長安，代宗爲之執弟子禮。」的「發」字（原文作  ），就是此志中如此處理的常見例子。

二、該字不易辨識，但可以找到一個於統一碼編碼中，意義完全相同的常見用字，則以該同義字取代文本中原先無法輸入之罕用字，加以紀錄此處進行了統一標準化的動作。例如以「  」取代 [叠 * 毛] 。[14]

三、該字不易辨識，且找不到相對應的常見用字。這是最難處理的字。對於這些字，我們除給予特殊紀錄與唯一編號之外，我們保留該字的圖片，並爲其建立一個「組字式」。所謂的組字式乃是由中華電子佛典協會 (Chinese Buddhist Electronic Text Association，簡稱 CBETA) [15] 組織於其數位專案中，處理特字時

12 在數位文本中，如何針對同一個字，保留其兩種不同的表現方式，同時又不干擾文獻本義的處理方式，將在後續「標記處理」的章節中詳述。

13 詳見：佛寺志專案_特字處理，2013/05/03，http://wiki.ddbc.edu.tw/pages/佛寺志專案_特字處理。

14 在此表示叠爲字左半部，而毛爲字右半部的特字。

15 CBETA，「中華電子佛典協會」網站，2013/05/03，http://www.cbeta.org/。

所提出的方式。其方式是將組成文字的部件稱爲「字根」[16]，並以簡易符號來表示各字根間的位置關係。使用組字的最大優點，就是可以在文獻中很直觀的表示原本無法展現的文字，而不需要經過繁複的造字程序，使用者也不需安裝額外程式。但其主要缺點就是組字式並非只有一種拆解規則，因此也造成使用者在搜尋上的困難。下表一列出 CBETA 組字式的規則簡述。

表一、CBETA 組字式規則簡述

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 閒 = 門 @ 月
-	表去掉某部分	青 = 請 - 言
-+	若前後配合，表示去掉某部分，而改以另一部分代替	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

16 因爲有時會遇到特殊情況，故我們也不排除採用全形注音、標點及英文符號做爲組字用字根。

進行缺字處理的過程之中，我們也查閱了許多值得參考的網路資源，包含有：教育部異體字字典[17]、Unihan Database[18]、CBETA 漢文字辭資訊網[19]、漢典[20]等網站。值得一提的是，檢索的過程中，我們發現中文字的部首分類標準不一，且有時各相關資源的分類令人難以猜測。例如：「艱」（Unihan Database 網頁部首歸在「隹」部下，而非「艹」或「舟」部）。此外，我們也意外發現，當遇到一個未知的字，以部首和筆劃爲線索來做搜索時，並不如想像中那麼容易搜索，例如：「垂」字，在不同線上字典中，其部首歸類及筆劃計算。教育部異體字字典認爲「垂」字乃土部、部外筆劃六劃；漢典將「垂」分爲土部（簡體）及士部（繁體），並認爲不論繁簡部外筆劃皆爲五劃；但 Unihan Database 網頁將兩個字都歸爲土部五劃。又如：「灪」與「灩」，同樣在 Unihan Database，卻只因字形繁簡之別，前者被歸至「鳥」部、後者則歸至「水」部。因此在查找過程中發現不少困難。

因此，爲正確處理正字，文字學、聲韻學、訓詁學、文字學之中的六書造字法則等相關素養，都可以加快檢索的效率。此外，倘處理人員除國學專業外，倘能具備佛學基礎知識，則將更有助於特字處理作業的效率。

（五）標記作業

TEI 標記編碼，主要是針對文字資料要轉換成電子形式儲存時，提供一套明確化的標準，這也是一個可被擴充到多種不同軟

17 教育部，「教育部異體字辭典」網站，2012/06/20，http://dict.variants.moe.edu.tw/。

18 Unicode Consortium, Unihan Database, 2012/06/20, http://unicode.org/charts/unihanrsindex.html.

19 CBETA，「CBETA字辭資訊網」網站，2012/07/04，http://dict.cbeta.org/word/search.php。

20 龍維基，「漢典」網站，2012/06/20，http://www.zdic.net/。

體語言的編碼架構。這些語言的共通之處是以元素與屬性來定義文件，並有完整規則來規範元素與屬性在文件中的用法。第一版指引手冊於 1994 年 5 月出版後，不論是在軟體語言，或是在全球資訊網的發展上，該指引即在數位圖書館的發展中漸具影響力。爲使佛寺志的數位典藏的數位資料達到國際標準之要求，此專案的數位內容都是採用此一國際化標準的電子格式所標記完成。

TEI 標記作業，是佛寺志數位典藏的最核心作業，也同時是困難的部分。標記的目的，是要利用鑲嵌在全文之中的 XML 標籤，盡量保留文章的型態與內容意義的資訊，以供後人使用。因此，執行過程之中，工作人員除了要熟知 TEI 標籤的使用時機與使用方式外，對於內容也必須完全的了解，始能做出正確的判斷。在 TEI 的標記規則指引之中，定義了幾百個元素的使用時機與方式。但在實際作業時，對於相同的內容，根據標記者的目的與偏好，還是有許多不同的標記方式可以達到類似的效果。因此，爲了保持標記的品質與一致性，在專案開始之初，我們訂定了此專案在標記文獻內容時，相對應的 TEI 標記使用方式，並於專案執行過程中，視情況進行不斷的修正。這些自行定義的標記標準，實爲此專案除了內容產出之外的另一個重要心血結晶，以下爲這些自訂規範的介紹。

1. TEI 標記檔首資訊

TEI 的標記規範中，明確要求所有的標記必須含有 <teiHeader> 的部分，用以紀錄此標記內容的背景資訊。除作品題名之外，此區塊也紀錄了專案相關人員及其貢獻、發行單位、授權方式以及原書書目資料內容。我們以專案內容的其中一個寺志——《徑山寺志》的標頭作爲範例，詳細內容如圖五所示。我們可以清楚發現到，標頭中記載了文獻的原始發行紀錄，數位化的工作項目與人員，數位版的發行授權等內容。

```
<teiHeader>
    <fileDesc>
      <titleStmt>
        <title type="main" xml:lang="zh"> 名山古刹－《中國佛寺史志》數位典藏 </title>
        <title type="main" xml:lang="en">Digital Archive of Chinese Buddhist Temple
          Gazetteers</title>
        <title type="subordinate" xml:lang="zh"> 徑山志 </title>
        <title type="subordinate" xml:lang="en">Jing shan zhi</title>
        <author>Dharma Drum Buddhist College, Library and Information Center, Digital
          Archives Section 法鼓佛教學院 圖書資訊館 數位典藏組 </author>
        <sponsor>Jingshan Temple 徑山寺 </sponsor>
        <principal>Marcus Bingenheimer 馬德偉 ， 洪振洲 </principal>
        <respStmt>
          <resp>Programming and Interface 程式及介面撰寫 </resp>
          <name> 花金地 </name><name> 李志賢 </name>
        </respStmt>
        <respStmt>
          <resp>Authority Files 規範資料庫架設、維護 </resp>
          <name> 洪振洲 </name><name> 葛賢敏 </name>
        </respStmt>
        <respStmt>
          <resp>Encoding 標記 </resp><name> 王秀雯 </name>
        </respStmt>
        <respStmt>
          <resp>Archive creation</resp><name> 周邦信 </name><name>Simon Wiles</name>
        </respStmt>
      </titleStmt>
      <publicationStmt>
        <publisher> 法鼓佛教學院 </publisher>
        <address><addrLine>20842 新北市金山區西勢湖 2-6 號 </addrLine>
                  <addrLine>da@ddbc.edu.tw</addrLine></address>
        <date>2011-2012</date>
        <availability>
          <p>This document is published under a CC Attribution-Share Alike License</p>
        </availability>
      </publicationStmt>
```

```
<sourceDesc>
  <bibl><title type="main"> 中國佛寺史志彙刊 </title>
        <title type="subordinate"> 徑山志 </title>
        <editor> 杜潔祥 </editor>
        <pubPlace> 臺灣 · 臺北 </pubPlace>
        <publisher> 宗青圖書出版公司 </publisher>
        <date>1994</date>
        <biblScope type="vol"> 第一輯第三十一冊 </biblScope>
        <biblScope type="pp">1 -1077</biblScope>
        <idno type="CallNo">DDBC:017954</idno>
  </bibl>
 </sourceDesc>
</fileDesc>
</teiHeader>
```

圖五、以徑山寺志爲例的 TEI 標記檔案之 teiHeader 部分

2. 內文標記

爲了使標記規則一致，在本數位專案開始之初，我們便決定專案標記的範圍與重點主要涵蓋以下四個重點：

（1）對於文章的架構（標題、分段）與樣貌（分頁、頁碼、目錄、圖像頁）必需進行標記紀錄。

（2）文本中的人名、地名、時間皆加以標記及並額外建置資料庫，以紀錄各條目詳細內容。

（3）製作數位化的文本過程中，加入新式標點符號。

（4）進行特字處理。

標記目標決定後，我們也必須決定使用 TEI 標記範圍與使用方式，以避免產生標記的混亂。此部分的細節，我們整理如下表二。[21]

21 此表僅爲標記方式之簡明摘錄，若要進一步瞭解細節，可參閱：法鼓佛教學院，「《中國佛寺史志》標記作業」網頁，2012/05/03，http://wiki.ddbc.edu.tw/pages/。

表二、佛寺志數位典藏 TEI 標記使用規範

標記內容	實際文章情況	標記方式
標題	段落標題	<head> 標題 </head>
	標題與作者並列	<head> 標題 </head><byline><persName> 作者 </persName></byline>
	有附帶小字副標題的標題	<head> 標題 <seg rend="font-size:small"> 副標題 </seg></head>
縮排	縮排的文段	內文部分
表格	以表格形式呈現、包含在直行橫列中的文字內容。 <table> 標記整體表格。 <row> 標記表格中的一列。 <cell> 標記表格中的一格。	<table> <row> <cell> 儲存格 1</cell> <cell> 儲存格 2</cell> </row> </table>
詩詞歌賦偈	韻文區塊	以 <lg> 包含形式上視爲一組的詩行，裡面包涵許多 <l> 元素，各標記詩文一行。而 <caesura/> 標記韻律詩行可能被截斷的位置
頁碼	文章分頁標示	<pb facs="1B009P003.jpg" n="0000a"/>

人名	確定人物之人名	<persName key=" 人名辨識編號 "> 某高僧大德 </persName>
	無法辨識之人名	<persName key="unknown"> 未知的人名 </persName>
	具有稱謂之人名	<persName>人名<roleName> 稱謂</roleName><persName>
	具有尊稱、封號之人名	<persName> 人名 <roleName type="honorific"> 尊稱、封號、謚號 </roleName></persName>
	混雜地名之人名	<persName><placeName> 天童 </placeName> 圓悟 <roleName> 大師 </roleName></persName>
地名	確定地點之地名	<placeName key=" 地名辨識編號 "> 地名 </placeName>
	無法辨識之地名	<placeName key="unknown"> 未知的地名 </placeName>
時間	無法確定的時間區段（僅知道開始不晚於某時間，結束點不早於某時間）	<date key="j ＋ (起始時間碼) ＋ (結尾時間碼)" notBefore-iso=" 起始西元年 - 月 - 日 "notAfter-iso=" 結尾西元年 - 月 - 日 "> 區段性的某時間 </date>
	確定的時間區段	<date key="j ＋ (開始時間碼) ＋ (結束時間碼)" from-iso=" 開始西元年 - 月 - 日 "to-iso=" 結束西元年月日 "> 事件持續進行的時間 </date>
	精確的單日	<date key="j ＋ (當日時間碼) ＋ (當日時間碼)" when-iso=" 西元年 - 月 - 日 "> 歷史上的某一天 </date>
	無法辨識的時間	<date key="unknown"> 未知的時間 </date>

代名詞	用來表示之前描述過的人名或地名實體	<ref key=" 人名辨識編號 "> 人名代名詞 </ref> <ref key=" 地名辨識編號 "> 地名代名詞 </ref>
底本小字註解	文本中使用小字表示的夾注內容	<note rend="font-size:small"> 小字註 </seg>
新增附註	標記人員額外加入的參考附註	<note resp="ddbc.da"> 註解文字 </note>
空格	文章中的空白格	<space quantity="1" unit="eng_chars"/> （一個英文半形空格） <space quantity="1" unit="chi_chars"/> （一個中文全形空格）
印章	原始文本中的印章	<stamp> 印章上的文字 </stamp>
上標字	文章中的上標字	<seg rend="vertical-align:super"> 上標字 </seg>
特字	具有統一碼，但不易辨識形義的特字	<choice><orig> 文本原字 (unicode)</orig><reg resp="ddbc.da"> 通用字 </reg></choice>
	不具統一碼，但有一通用字之特字	<reg> 通用字 </reg>
	不具統一碼，也無對應之通用字之特字	<glyph xml:id=" 特字參考編號 "> <glyphName>Non Unicode Character</ glyphName> <mapping type="cbeta">CBETA 組字式 </mapping> </glyph> <g ref=" 特字參考編號 ">

(六)規範資料庫

在專案標記過程中，如遇到人名、地名或時間之實體時，除依上述表二以相對應標記進行處理之外，我們也針對這些實體進行基本背景與相關資料的查找。如此一來，不僅可以增加此數位專案的內容豐富度，也同時可以提高文本內容的可讀性。而這些背景資料，實際上不僅可以用於此專案，於其他類似的數位專案，甚或佛學資料閱讀的過程中，這些資料都有其重要的參考價值。有鑑於此，法鼓佛教學院建立了一個佛學名相規範資料庫，[22] 以便將這些資料有系統的整理與分享，達到避免各相關專案的重工與資源共享之目的。此一規範資料庫其下又分成三大子資料庫，分別是：人名規範資料庫、地名規範資料庫與具備中日韓與西元年對照的時間規範資料庫。

人名規範資料庫主要提供與佛典相關的人名規範資料的存放與查詢，而此資料庫的設計目的，主要是針對古籍文本大量重複的人名，提供足以分辨的基本資料，如：別名、生卒年、朝代、籍貫、說明等欄位來輔助判斷，以協助使用者正確地找出所需要的人物。此外，本資料庫使用「唯一碼」(Authority ID)，作爲數位化佛教文獻人物的唯一識別方式。人名唯一碼是由一組不重複的七位文數字組合而成，使用者僅需要引用此唯一碼，便可以避免掉文獻上同名同姓所造成的識別上困擾。而針對這筆資料的紀錄，我們的工作人員可能會探查許多可能的資料來源，例如其他重要參考文本，或也可能來自其他網路資源。但不論是哪種來源，我們都會詳實記錄下來，以供後世使用者理解與判斷之用。人名規範資料庫目前約有兩萬兩千餘筆資料，主要的來源爲《梁高僧傳》、《唐高僧傳》、《宋高僧傳》、《明高僧傳》與此專案數位化之兩部中國佛寺志。

22 法鼓佛教學院，「佛學規範資料庫」，2012/07/04，http://authority.ddbc.edu.tw。

地名規範資料庫主要提供佛典相關地名規範資料查詢，由於佛教的產生與活動地點多以亞洲爲主，因此地名規範資料庫的內容也依佛教活動爲範圍。目前約有五萬五千餘筆資料，約有三萬筆中國歷史行政地名（行政疆域）是由中研院所提供，其餘則由高僧傳、佛寺志等佛教文獻所產生。每一筆地名資料中都提供有一規範碼（Authority ID），可以用作「唯一」識別之用。資料庫的資料內容包括地名之別名、群組、經緯度、現在行政區、朝代、註解等資料。值得一提的是，資料庫上提供地名的十進位地理經緯度座標資訊（Longitude, Latitude），可用於 GoogleMap 或 GoogleEarth 等圖形化的地理資訊系統軟體（GIS, Geographic Information System）上面，用以明確辨識出該地名的地理位置。透過座標資料的紀錄，相關的佛學數位專案便可依此產生視覺化的效果，不再局限於文字上的描述。

在中國的佛教古籍經典中，有關於「時間」的記載都是以中國年爲主，所謂中國年大致包括了朝代、帝號、年號（如漢桓帝建和二年），或使用干支來表達時間（如太和七年癸丑歲九月二十二日）。這種中國年的時間表示法，對非具有歷史專業背景的人而言，他們很難把文本描述的時間與歷史事件相結合起來。而藉由建置完備的時間規範資料庫，便可將此困擾減至最低。而時間規範資料庫的內容，主要是提供中國曆法資料查詢。此外，也提供了相對應的日本與韓國時間曆法之查找。在時間規範資料庫中，最小單位是「日」，也就是說，每一天有一個獨一無二的規範碼。爲與國際標準接軌，因此採用具有國際標準的 Julian Day Number 作爲規範碼，而其他曆法也就被當作是每一日的一種解釋。在介面中，我們也提供了「並行年代」查詢的功能，使用者將可同時查到某一日在中日韓曆法中的完整解釋。

(七)永久性資料保存

數位典藏的一個重要任務，就是必須確保數位資料的內容可以被永久保存。但實際上，數位物件的存活時間，卻不如我們想像中的雋永。以光碟或是硬碟的儲存媒體的年限來看，儲存時間約在數十年間，比起書頁或實體建築這種動輒數百年以上的存活時間，數位媒體的保存年限實在不可謂之長。而影響數位檔案內容是否能永久保存的另一要素，就是檔案格式問題。由於軟體的功能日新月異，爲了搭配更多、更新穎的功能，檔案格式經常隨著軟體版本的推陳出新而改變。因此幾年前製作的檔案，雖然承載之媒體沒有受損，但常常發生當使用者想以今日的軟體開啓時，卻已經無法正確開啓的現象。這些就是因爲原始檔案格式沒有更新，也被沒有被現代軟體所支援的結果。因此透過資料的數位化，想賦予原始紙本的文獻內容有更長壽命的論點，實際上並不完全正確。其實，數位形式帶給資料的好處，是高度的可散佈性，以及極低成本的重新利用性。因此，數位資料維護者必須一直重複的轉換數位資料爲最新的格式，才能將資料長久保存。但綜觀以往的經驗，專案有其資源的限制，興趣的更迭，與結束的一天。只靠單一團隊並不容易長久維護一份資料。因此，將專案完成之數位資料的製作過程的相關資料，資料定義格式封裝在一起，並且開放讓更多人使用，讓資料在被新的使用者需要時，能依當時之需求，進行資料格式傳換之作業。如此結合眾人之力，才容易達到數位資料永久保存的目的。

因此，本專案除將文獻數位化之外，也將資料供國際上互通使用，並且可以爲後世所應用，而不是只存在一個時間，一個地點的目標，那作專案的重點考量。爲了達到這個目標，其所採用的數位化方式、規格及永續性就相對的非常重要。除了典藏標的物製作時必須以最高品質爲要求的方式來製作方式來執行之外，也必須考量永久保存資料內容的完整性。也就是說，除了典藏標

的物之外，一些執行過程中的輔助資料，最好也能夠一併留存，才能提供足以爲後世應用所需要的數位典藏檔案資料。因此本專案所提供的資料下載內容中，包含以下重要內容：

1. 兩部佛寺志文獻的原書掃描圖檔。
2. 以 TEI 標準完成的佛寺志文獻標記檔案。
3. 以 RELAX NG[23] 格式撰寫的專案標記文件之驗證綱要檔。
4. 用來產生專案標記文件之驗證綱要檔之 ROMA 設定檔。[24]
5. METS[25] 格式的完整後設資料檔案（內含以 MIX 格式完整紀錄的原書圖檔後設資料與以 TEI 格式紀錄的文獻內容的後設資料）。

由這些資料，使用者將可以掌握此數位典藏內容的所有面相，並在需要的情況下，具有重構整體數位典藏系統的能力。

(八)數位化成果統計

如前所述，兩部佛寺志一共包含 237 個不重複的寺志。其

23 RELAX NG爲撰寫XML驗證用的綱目資料（schema）的主流語言之一，其特點是簡單易學且相容於xml的基本格是。詳細資料可參見：RELAX NG home page, 2012/06/18, http://relaxng.org/。

24 本專案使用之TEI標記規範，實際上已經過客製化之調整。因此，爲產生符合本專案使用之TEI標記驗證檔，我們使用TEI協會提供的ROMA線上服務，產生客製化之驗證檔。而操作過程中之詳細設定，皆紀錄在此ROMA設定檔之中。使用者以此檔案，重新調整內容，以微調產生所需的TEI標記驗證檔。詳細內容請參見：Roma: Generating Customizations for the TEI, 2012/03/05, http://www.tei-c.org/Roma/。

25 Metadata Encoding & Transmission Schema（簡稱METS）是眾多常用的後設資料規範之一。但METS本身並非是爲了某一種特殊物件而設計出來的後設資料規範，而相反地，METS是用來將各種不同的schema包裝在一起的容器型後設資料。藉由METS的規範，我們可以將多種不同的後設資料整合在同一個檔案之中，而不會喪失細節。詳細內容請參見：Metadata Encoding& Transmission Schema, 2012/07/02, http://www.loc.gov/standards/mets/。

中，根據目前的數位化狀況，我們以分成：僅完成圖檔掃描、已完成圖檔掃描與全文打字與初略標記、圖檔掃描與全文詳細標記都已完成等三個類別。[26] 其中，完成詳細標記的全文共有 15 部寺志，我們將其列表如下：

表三、15 部已標記完成之寺志清單

編號	寺志名	版本紀錄
g008	《重修普陀山志》	明萬曆三十五年（C.E. 1607）太監張隨刊本
g009	《普陀洛迦新志》	民國十三年（C.E. 1924）排印本
g010	《明州阿育王山志》	明萬曆四十年（C.E. 1612）刻本 清乾隆二十二年（C.E. 1757）正續合刊本
g011	《明州阿育王山續志》	清乾隆二十二年（C.E. 1757）正續合刊本
g017	《玉岑山慧因高麗華嚴教寺志》	清光緒七年（C.E. 1881） 錢塘丁氏重刊本
g032	《徑山志》	明天啓四年（C.E. 1624）原刊本
g043	《寒山寺志》	清宣統三年（C.E. 1911）初稿纂成 民國十一年（C.E. 1922）吳縣潘氏刊本
g049	《峨眉山志》	民國二十三年（C.E. 1934）排印本
g062	《福建泉州開元寺志》	民國十六年（C.E. 1927）重刻本
g077	《九華山志》	民國二十七年（C.E. 1938）排印本

26 初略標記與詳細標記的差別在於，詳細標記內容增加了新式標點標逗，與人、時、地等實體的辨識。

g081	《清涼山志》	明萬曆二十四年秋（C.E. 1596）纂輯成書 民國二十二年（C.E. 1933）排印本
g084	《雞足山志》	清康熙三十一年（C.E. 1692）刊本 中央研究院傅斯年圖書館藏
g086	《黃檗山寺志》	民國十一年（C.E. 1922）萬福禪寺據道光四年（C.E. 1824）刊本排印； 中央研究院傅斯年圖書館藏
g089	《天台山方外志》	明萬曆三十一年（C.E. 1603）撰成本志清光緒二十年（C.E. 1894）重刊佛隴眞覺寺藏板中央圖書館臺灣分館藏
y109	《徑山集》	明萬曆七年（C.E. 1579）刻本

針對上列已完成的 15 部寺志的內容，我們進行了簡單的統計計算，獲得許多有趣的結果，茲表列如下表四。其中我們發現到，目前 15 寺志中，總計約 300 萬字，而新增標點符號 30 萬個，也就是平均 10 個字，就需要一個標點。而特字數量約 9700 字，雖僅佔全文比例的千分之三，但處理起來卻複雜度極高。專案在這 15 個志中，我們也辨識了約 9000 位的人物，其中女性只有 186 位，大概是百分之二。由此也可看出自古以來女性在佛教界的歷史中，確實較不被重視。

表四、15 部已標記完成寺志之標記內容統計

總字數	2,997,003 字	男性人名數目	8,746 位
新增標點符號數	290,472 個	女性人名數目	186 位
特字處理字數	9,776 字	其他人名數目	219 位
修訂字數	1,073 字	地名出現總數	43,777 個

附加註解數	947 處	不重複之地名總數	7,093 處
時間標記總出現次數	9,699 個	中國內地名	6,700 處
人名總出現次數	41,254 個	不在中國的地名	393 處
不重複之人名總數	9,151 位		

此外，我們統計出在兩志之中最常被提及的人名與地名前十名，如下表五，括號中表示出現的次數。

表五、15 部已標記完成寺志之最常出現的人名與地名前十名

出現次數排名	人 名	地 名
1	文殊菩薩（532）	普陀山（980）
2	釋迦牟尼佛（470）	天臺山（705）
3	普賢菩薩（411）	五臺山（680）
4	觀世音菩薩（378）	九華山（673）
5	智顗（355）	峨嵋山（571）
6	宗杲（355）	阿育王寺（512）
7	寒山（307）	普濟寺（404）
8	摩訶迦葉（200）	徑山寺（361）
9	蘇軾（182）	雞足山（350）
10	地藏菩薩（169）	徑山（346）

由於人物被提及的次數與重要性之間，必有其正向的關聯性。由於佛寺志多半內容仍與佛教有關，因此前十位常被提及的人物中，有多數爲神佛。其他常被提及之人物，多半是十分重要的宗師。例如：智顗（C.E. 539-598）：世稱智者大師，天台大師，是中國佛教天台宗四祖，天台宗的實際創始人。智者大師德高望

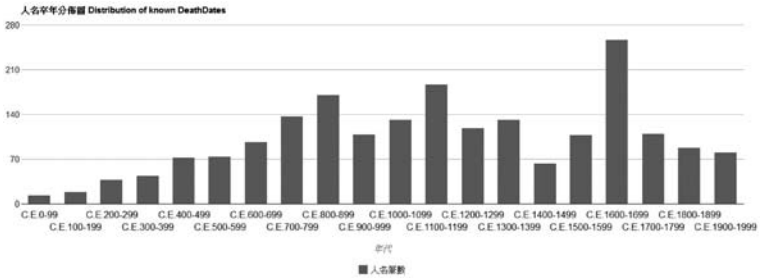
重，與皇室的關係良好。當時凡是講院所傳，多天台智者之教。也因此許多寺志當中，多半會提到智顗教理。而宗杲（C.E. 1089-1163），他不僅是中興徑山寺的祖師，中國禪宗文化的發展亦因宗杲的出現，而進入禪宗的成熟期。而寒山子（C.E.691-830）常於國清寺出入，傳說其爲文殊菩薩之化身，與豐干（彌陀化身）、拾得（普賢化身）號稱「國清三隱」，其相關事蹟話語也常出現於多部寺志之中。而摩訶迦葉，又名大迦葉、迦葉波、迦攝波。意爲飲光，爲釋迦牟尼十大弟子之一，迦葉除常出現於佛教教理的宣說過程之中，也常被人塑像於寺廟之中，也因此常見於寺志記載。除這些有名的佛教人物之外，蘇軾（C.E. 1036 -1101）也進入榜中，是一個令人意外的驚喜。雖然蘇軾一生並未出家，但蘇軾首創「十方住持制」，以取代舊有的「自傳制」，破除門戶之見，選賢與能的制度讓徑山祖席獨著功不可沒。此外，蘇軾亦嘗多次遊覽佛教聖地，作詩留文，因此許多寺院皆有與其相關的記載。而常提到的地名前十名當中，多半都是寺志的討論對象，因此成爲前十位常被提及之地名，也就不令人意外了。

此外，我們進一步針對人物的卒年紀錄，以每百年爲間隔，統計文獻中各年代人物的數量，其結果如下表六。

表六、15 部已標記完成寺志中所提及之人物卒年分佈統計表

年 代	人 數	年 代	人 數
西元一世紀（C.E. 0-99）	14	西元十一世紀（C.E. 1000-10 99）	131
西元二世紀（C.E. 100-199）	19	西元十二世紀（C.E. 1100-1199）	186
西元三世紀（C.E. 200-299）	38	西元十三世紀（C.E. 1200-1299）	119
西元四世紀（C.E. 300-399）	45	西元十四世紀（C.E. 1300-1399）	131

西元五世紀 （C.E. 400-499）	73	西元十五世紀 （C.E. 1400-1499）	62
西元六世紀 （C.E. 500-599）	75	西元十六世紀 （C.E. 1500-1599）	106
西元七世紀 （C.E. 600-699）	97	西元十七世紀 （C.E. 1600-1699）	256
西元八世紀 （C.E. 700-799）	135	西元十八世紀 （C.E. 1700-1799）	108
西元九世紀 （C.E. 800-899）	168	西元十九世紀 （C.E. 1800-1899）	87
西元十世紀 （C.E. 900-999）	108	西元二十世紀 （C.E. 1900-1999）	81



圖六、15 部已標記完成寺志之中提及之人物卒年分佈統計圖

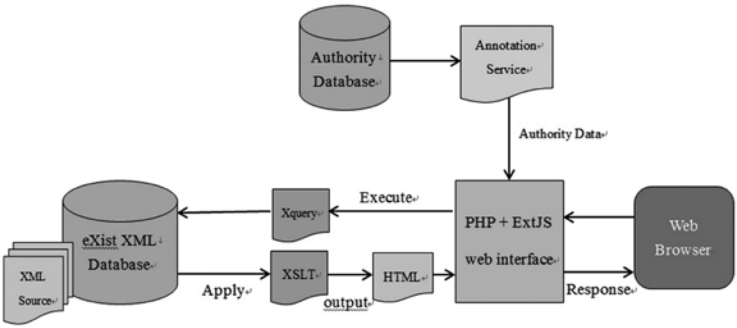
由以上圖六與表六得知，人物卒年主要有三個集中的趨勢，分別爲西元九世紀（C.E. 800-900）、西元十二世紀（C.E. 1100-1199）、與西元十七世紀（C.E. 1600-1699）。也就是說在這 15 個《佛寺志》內的人物主要以唐、宋時期與明末清初的人物最多。因爲佛寺志撰寫的年代多半於明、清，因此內容紀錄許多明清時期的佛教人物，並不令人意外。唐朝時期，中國佛教大興，佛教人才輩出，而南宋時許多佛寺受到政府的冊封，而香火鼎盛，也吸引許多人才，例如徑山寺，也是在南宋被封爲江南五大禪院之首。這也造成我們系統中唐、宋時期較多的現象。

三、佛寺志數位典藏系統建置與操作介面

(一)系統介面技術

爲使佛寺志數位典藏的使用者能有最好的使用體驗，本專案建立了一個專屬的佛寺志數位內容閱覽網頁系統（http://buddhistinformatics.ddbc.edu.tw/fosizhi/）。

此系統使用了多種先進的資訊技術，但本專案所有使用技術皆爲免費公開的技術，這意味著此專案的系統架構可以容易的被複製到其他專案中使用，而且能使系統建置成本降到最低。圖七爲本數位典藏系統之架構圖，如圖所示，本系統所使用的 eXist[27] 資料庫作爲本專案使用之 XML 文獻的主要儲存機制，eXist 資料庫是專門爲儲存 XML 文件所設計的資料庫，可以支援許多前端語言需求的高階 XML 操作功能。其主要優點爲實作了一般交易資料庫中最重要的交易紀錄機制，以及提供 XML 文件樹狀結構與其他特殊需求的索引機制。也因此 eXist 提供了本專案一個可信賴且高效能的資料儲存空間。



圖七、徑山志數位典藏專案線後端技術架構

27 詳見：eXist Database, 2012/07/04, http://exist-db.org/。

爲打造高效能且功能豐富的前端網頁介面，我們選擇以 PHP 程式語言搭配 ExtJS[28] 函式庫作爲程式的平台。PHP 程式語言負責系統後端資料存取的服務，而 ExtJS 的函式庫則協助打造豐富的網頁前端互動介面功能。在實際執行過程中，當使用者進行資料瀏覽時，位於伺服器端的 PHP 網頁程式將會被啓動，讓使用者的瀏覽器，載入使用 ExtJS 的函式庫爲主體所寫成的使用者操作介面。此時，當使用者由介面要求閱讀任何的寺志內容時，使用者操作介面將會對後端 PHP 程式送出更新資料的需求。而 PHP 程式將視使用者之需求，執行一個預先撰寫的 XQuery[29] 腳本，以便向 eXist 資料庫取用需要的資料。並將回應結果的 XML 內容繼續應用一段 XSLT[30] 的程式碼，將回應的 XML 轉譯成 HTML，並根據前端需求來呈現不同的格式給使用者。此外，爲了減少使用者在使用過程中的等待時間，此系統大量的使用了 Ajax 技術來降低網頁前端與後端伺服器溝通的負擔，讓使用者可以即時得到結果而無需忍受網頁重複載入所帶來的大量等待時間。

---

28 ExtJS爲web介面專用的javascript framework，ExtJS整合了CSS樣式文件，各種網頁元件（按鈕、下拉式選單、文字框等）都有現成的樣式，甚至不需要美工進行特別的美化就可以直接使用，功能涵蓋了一個Web 2.0網站所需要的幾乎所有的功能，非常完備。詳見：ExtJS, 2012/07/01, http://www.sencha.com/products/extjs/。

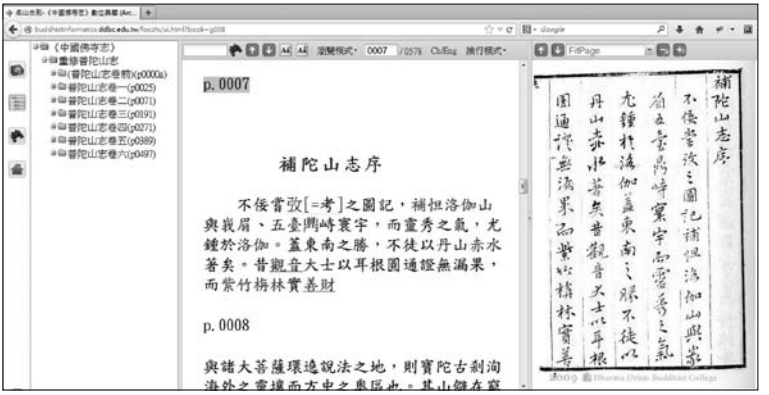
29 XQuery是由W3C所定義的XML資料庫使用的標準查詢語言，其目的就是用來將所需資料由XML資料庫中取出。參閱：XQuery 1.0: An XML Query Language, 2012/07/04, http://www.w3.org/TR/xquery/。

30 XSLT是用來定義如何將XML的原始資料內容轉換爲其他格式的語言，詳細內容可以參見：XSL Transformations (XSLT), 2012/05/04, http://www.w3.org/TR/xslt20/。

(二)系統介面操作

圖八、佛寺志數位典藏專案首頁

爲迎合現今使用者對於網路瀏覽與數位資料使用的習慣，佛寺志數位典藏系統以網頁作爲主要的呈現媒體，本專案首頁如圖八所示。點選專案網頁下方的任一個寺志，將可以開啓佛寺志數位典藏系統的線上閱讀介面。系統閱讀介面的基本樣貌如圖九所示。畫面基本上包含三大部分，最左邊以樹狀結構的方式，顯現此寺志的文章結構。而中間區塊則是文章的主要內容，右邊區塊是相對應目前閱讀內容的圖檔。基本上，文章內容與圖檔之間具有連動的效果，也就是說當使用者捲動文本視窗的內容時，右方的圖檔區塊也會隨之重新載入相對應的圖檔。



圖九、佛寺志數位典藏專案線上閱讀介面

而畫面中也設計了一些小功能讓使用者可以有更好的閱讀體驗，例如使用者可以選擇關閉左方樹狀結構，以取得更多的本文閱讀空間，或按下畫面上方的「閱覽模式」按鈕，選擇「僅閱讀全文」、「檢視雙描圖模式」或是「圖文對照模式」以得到最佳的閱讀空間配置。此外，畫面上面的一些小按鈕，則可以讓使用者進行對於文字、圖片顯示的大小進行微調。對於內容的切換，系統提供了向上卷頁、向下卷頁的按鈕。當然，使用者也可以直接給予頁碼，讓系統直接捲動到該頁，進行閱讀。

除上計基本閱讀功能之外，文獻閱讀介面也會利用不同的顏色來表示寺志內所有式別出來的的人名、地名、時間實體，並且針對每一個實體，提供一個彈跳式的解釋框。當使用者用滑鼠點擊文本中實體出現的位置時，此一彈跳框就會出現，並提供這些實體的適當解釋。例如當使用者於瀏覽的介面上點選已標記的人名時，即會跳出如圖十中的「 人名規範資料」的小視窗，顯示這個人物相關資料，諸如：生卒年、籍貫、性別、註解，及這個人物在其他古籍出現的位置。

圖十、人名規範資料彈跳式視窗

而這個彈跳式的小視窗，並非僅對於人名，實際上對於內文中的人名、地名與時間的實體，我們在介面上都提供了詳細資料的彈跳視窗。而這樣的設計，並非專門爲此一專案而設計的功能，而是來自於第二章第六節中所提到的佛學名相規範資料庫所提供的註解服務（Annotation Service）[31]。這一註解服務的目的，就是要爲所有使用此一規範資料庫內容的數位專案，提供一個方便且快速取得參考資料的服務。更詳細的說，在佛寺志數位典藏的系統中，我們在把 TEI 的文獻標記資料，轉換爲 HTML 的內容之時，也將人名、地名與時間的實體標記，轉換爲符合該註解服務的要求格式。如此，我們便可以在介面中提供此彈跳式資訊視窗的服務。而此一彈跳視窗的運作原理是透過簡單的 http 呼叫，與規範資料庫中預先設置的服務程式來溝通，而達到資料交換的目的。此一架構的最大優點是，數位專案的開發者無須負責參考資料的維護、也不用花費空間來儲存，資料的提供與維護都是由另一專案——規範資料庫來負責，因此佛寺志數位典藏便可以專心在文獻內容的發展，而不必擔心其他參考資源的內容更新問題。可說是專案間整合的一個良好應用示範。

同上所述，當使用者點選文獻中已標記的地名時，如圖十一所示之「地名規範資料」小視窗就會跳出來，以顯示這個地名相關資料，諸如：別名、行政區、座標、註解等資訊。

31 參閱李志賢、洪振洲，〈法鼓佛教學院權威資料註解服務〉，2009電腦與網路科技在教育上的應用研討會，新竹，中華大學，2009年11月26-27日。

圖十一、地名規範資料彈跳式視窗

若使用者點選的是已標記的時間，則如下圖十二的「時間規範資料」資訊視窗就會跳出。這個視窗顯示這個時間相對應的西元年，以及相對應的日本曆法跟韓國曆法時間，如下圖十二所示，以便讓使用者能快速掌握時空訊息。

圖十二、時間規範資料彈跳式視窗

除豐富的內容閱讀介面之外，系統也提供了快速精確的搜尋功能。使用者可以在上方搜尋框中搜入關鍵字，啓動全文搜索的

功能。以下圖十三爲例，在搜尋欄輸入搜尋字串「宗杲」，並執行查詢後，就可查到文本內所有出現「宗杲」字串的位置。所有搜尋的結果會在畫面左方顯示，而按下左方的搜尋結果後，中間的文獻視窗將自動捲動至所指定的搜尋結果位置，並將關鍵字標記，讓使用者能一目了然。

圖十三、佛寺志數位典藏線上介面之搜尋「宗杲」之結果

四、佛寺志數位典藏未來發展

明清時代於中國本土各地大量編撰的寺廟佛寺志，是研究近代晚期中國佛教的寶典。雖然內容珍貴，但由於出版品的內容龐雜，不易整理，一直到二十世紀晚期，兩套以此爲主題的叢書才由台灣與大陸的編者，不約而同地相繼出版上市。這兩部叢書，分別是：1980年至1985年由杜潔祥主編之《中國佛寺史志彙刊》，110冊，分三輯上市，分別由台北明文書局與丹青圖書發行。而另一部是：2005年由張智主編，杭州廣陵書社出版的《中國佛寺志叢刊》，內容共130冊。兩套叢書由於數量龐大，發行量與典藏地皆有限，爲使此兩部具有高度宗教與歷史研究價值之典籍能廣泛爲世人所知與利用，並達到永久典藏之目的，於西元2007年，法鼓佛教學院圖書資訊館正式啓動「中國佛教寺廟志數

位典藏」的專案計畫。專案的主要目的就是要將此兩套叢書中所收錄的237個寺志，進行高品質之全文數位化處理，並藉由數位媒體的便利性，公開爲世人所用。

在約略六年的執行過程之中，我們完成全部237個寺志內容的影像掃描。此外，有關數位全文的部分，目前已經有29個寺志已可提供的全文的數位檔案，而其中的15個寺志的全文，更是已經完成細節的標記（包含人名、地名、時間的辨別），新式標點符號修訂，缺字處理等品質強化的步驟，可說是目前世界上品質最高的佛寺志數位內容，這些修訂完成之寺志，也將於今年以紙本的方式公開發行，以供研究者參考與收藏之用。而這些內容都開放在本專案網頁，供研究者免費下載使用。此外，我們也利用了最新的網頁撰寫技術，完成了具有豐富互動功能的網頁閱讀介面，讓使用者在無須安裝任何系統的情形之下，就可以使用佛寺志數位典藏的豐富內容。

而目前，佛寺志的數位計畫並未結束，未來將持續朝幾個方面進行。首先要務當然是持續進行全文數位化的工作。但根據前幾年的經驗，我們察覺到，若要將每個寺志的內容，皆進行細節的標記（包含人名、地名、時間的辨別），則需花費相當大的時間成本。因此我們將調整專案執行方向，優先執行基本的全文數位化動作，預計於未來兩年，先將所有237個寺志的全文能於網路上提供爲目標，之後再持續進行各志的細節的標記。而第二個目標是將標記完成的佛寺志，加入中華佛典協會所發行的CBETA光碟之中。由於CBETA可說是目前最知名的佛學參考資源，將佛寺志的數位成果加入CBETA光碟之後，將可有助於佛寺志數位成果的流通。

引用文獻

中日文專書、論文或網路資源等

洪振洲、李志賢（2009）。〈法鼓佛教學院權威資料註解服務〉。2009 電腦與網路科技在教育上的應用研討會論文集。新竹：中華大學資訊管理學系。

馬德偉（2009）。《TEI 使用指南——運用 TEI 處理中文文獻》。台北：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。

CBETA。「CBETA 中華電子佛典協會」網站。2013/05/03，http://www.cbeta.org/。

CBETA。「CBETA 字辭資訊網」網站。2012/07/04，http://dict.cbeta.org/word/search.php。

法鼓佛教學院。「佛學規範資料庫」網站。2012/07/04，http://authority.ddbc.edu.tw。

法鼓佛教學院。「佛寺志專案_特字處理」網頁。2013/05/03，http://wiki.ddbc.edu.tw/pages/ 佛寺志專案 _ 特字處理。

法鼓佛教學院。「《中國佛寺史志》標記作業」網頁。2012/05/03，http://wiki.ddbc.edu.tw/pages/。

法鼓佛教學院。「中國佛教寺廟志數位典藏」。2012/12/22，http://buddhistinformatics.ddbc.edu.tw/fosizhi/。

教育部。「教育部異體字辭典」網站。2012/06/20，http://dict.variants.moe.edu.tw/。

龍維基。「漢典」網站。2012/06/20，http://www.zdic.net/。

西文專書、論文或網路資源等

eXist Database. 2012/07/04, http://exist-db.org/.

Japan Electronics and Information Technology Industries Association. 2002.Exchangeable Image File Format for Digial Still Cameras: Exif Version 2.2. 2013/05/01, http://www.exif.org/Exif2-2.PDF.

Library of Congress. Metadata Encoding & Transmission Schema. 2012/07/02, http://www.loc.gov/standards/mets/.

Library of Congress. Metadata for Images in XML Standard (MIX). 2012/06/06, http://www.loc.gov/standards/mix/.

Rivest, R. 1992. The MD5 Message-Digest Algorithm. United States: RFC Editor. 2013/05/01, http://www.ietf.org/rfc/rfc1321.txt.

Sencha Inc. ExtJS. 2012/07/01, http://www.sencha.com/products/extjs/.

Text Encoding Initiative. Roma: Generating Customizations for the TEI. 2012/03/05, http://www.tei-c.org/Roma/.

Unicode Consortium. Unihan Database. 2012/06/20, http://unicode.org/charts/unihanrsindex.html.

W3C. XQuery 1.0: An XML Query Language. 2012/07/04, http://www.w3.org/TR/xquery/.

W3C. XSL Transformations (XSLT). 2012/05/04, http://www.w3.org/TR/xslt20/.

The Construction of Digital Archive of Chinese Buddhist Temple Gazetteers

Jen-jou Hung
Associate Professor
Dharma Drum Buddhist College

Abstract:

Buddhism was introduced to mainland China in the Eastern Han Dynasty (東漢), 206 BC – 220 AD. After two thousand years of development, it has become an integral part of Chinese culture. The long history of Buddhism in China has witnessed the activities and travels of a large number of eminent monks, the building of an enormous amount of magnificent temples and the undertaking of uncountable large scale enterprises of various sorts. Unfortunately, such rich cultural activities were not documented in full detail in official historical writings. Most researchers still need to resort to the Buddhist texts themselves in order to be able to restore the actual historical situation at any given time, albeit with a certain degree of approximation due to the complex and at times problematic nature of the sources in question.

Between the 16th and the 20th centuries, with the modern rise of Buddhism and the increased availability of the support of large numbers of believers, compiling and publishing Buddhist scriptures becomes an achievable dream. Thus, a large number of Chronicles recording the history of local Buddhist temples comes into being under such circumstances. Gazetteers are a distinct genre of Chinese historiography. Instead of a single descriptive work of a region, city or temple, these gazetteers are usually compilations including a variety of texts from different authors, which makes them especially representative historical documents and records about the temples

in each era of modern China. As a result, the Buddhist gazetteers become an important source of reference for studying the late developments of Chinese Buddhism. These exceedingly large amounts of gazetteers were assembled into two series of books in the course of the 20th century: the Zhongguo Fosi Shizhi Huikan (中國佛寺史志彙刊), in 110 volumes, compiled by Du Jiexiang (杜潔祥) in 1980-1985, and the Zhongguo fosizhi congkan (中國佛寺志彙刊), in 130 volumes, compiled by Zhang Zhi and others (張智等) in 2006.

In order (a) to make these religious and historical primary sources that were included in these two book series widely known and used, and (b) to achieve the purpose of their permanent preservation, from year 2007, the Library and Information Center at Dharma Drum Buddhist College, Taiwan, started the "Digital Archive of Chinese Buddhist Temple Gazetteers" project. The main goal of the project is to create a high-quality digital full-text archive of these two book series and to make them freely available to the public with the help of digital media. The project has been running for 5 years and now bears its fruitful results. In this paper, we briefly describe the digitization process behind the creation of this archive, the main web interface and future possible developments and applications of the project. We believe that this overview will provide to be very helpful to users who want to have a clear and complete picture of this digital archive. Besides, other digital projects will also benefit from our experience in building their own archives.

Keywords:

Chinese Buddhism; Temple Gazetteers; Digital Archive; Content Markup; TEI