Digitization Procedures Guideline

Text Data

數位化工作流程指南：文字資料

拓展台灣數位典藏計畫

05

拓展台灣數位典藏計畫
Taiwan Digital Archives Expansion Project

數位化工作流程指南：文字資料

Text Data Digitization Procedures Guideline

數位典藏與數位學習國家型科技計畫　拓展台灣數位典藏計畫

出版序

「數位典藏國家型科技計畫」於西元 2002 年開始執行，眾多機構計畫與公開徵選計畫的工作夥伴紛紛加入我們團隊，進行種類繁多而又數量鉅大的數位化工作。第一期五年計畫於西元 2006 年結束，次年即擴大規模，與「數位學習國家型科技計畫」整合為「數位典藏與數位學習國家型科技計畫」（TELDAP, http://teldap.tw/），以「呈現台灣的文化與自然多樣性」為總體目標，持續拓展各類重要資源的數位化工作，並更積極地從事教育、研究與產業的加值工作，希望能更有效地吸引文教與市場力量共同推廣數位化成果，既藉以保存我國寶貴文化資產，也促成數位時代的文化創新。

作為「數位典藏與數位學習國家型科技計畫」的分項計畫，我們的名稱也由第一期「內容發展分項計畫」改為「拓展台灣數位典藏計畫」（http://content.teldap.tw），不僅持續拓展數位內容來源，向更多公私立單位甚或是個人收藏徵集檔案、考古、語言、地理、族群、藝術、民間生活以及動、植、礦物等相關數位化計畫，並且研發如何整合這些自然與人文不同性質數位內容的加值策略，希望製成更多兼具趣味性與啟發性的數位素材，既開放民眾下載便利教育與研究用途，也促成更多廠商與公私典藏者在商業加值方面的彼此合作機會。我們與「數位典藏與數位學習國家型科技計畫」其他分項計畫共同協力，希望能在加速我國數位內容由典藏保存跨入教育、研究與商業加值大目標的過程當中，起到關鍵作用，進而在網路世界更好地呈現台灣的文化與自然多樣性，讓更多國內外民眾體會並珍視我國豐盛美好而又多元互補的歷史文化與自然生態。

在拓展典藏與整合加值各類數位內容的同時，我們都持續針對公私立機關與公開徵選計畫等工作夥伴如何從事各類物件的數位化工作流程及相關技術進行調查與記錄，以這些調查記錄為基礎，本計畫同仁還結合符合國際標準的各項數位化技術與工作流程相關資訊，持續編撰「數位化工作流程叢書」。自西元 2005 年以來，我們即先精選了瓷器、書畫、古籍等單一種類的數位化物件，綜合不同典藏計畫從事這些物件的數位化經驗，並納入國內外相關理論與實務，

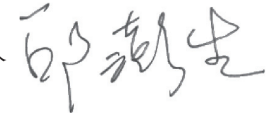
陸續撰寫了 20 冊數位化工作流程指南專書（這些專書都能在「拓展台灣數位典藏」網站的「數位化書籍」欄位以及 Google Books 做全文瀏覽與公開下載）。

西元 2008 年之後，我們不僅持續修訂與擴充這套「數位化工作流程叢書」，並且積極擴大流通管道以供更多博物館、圖書館、機構與個人參考。同時，我們在寫作策略上也做了調整，由兩方面補強這套指南叢書的內容：一是修訂既有的「精選物件」指南，二是新添編撰「共通原則」指南；前者指的是修訂既有的 20 冊工作流程指南，特別是針對數位化新技術與規範的引進、更實用的軟硬體設備以及數位內容保護機制等層面做修訂，這些專書基本上都於 2011 年出版完畢。至於新編的「共通原則」指南，則重點放在導入數位資訊「生命週期」與「品質管理」等關鍵概念，以「跨物件」而非單一精選物件為探究對象，針對這些關鍵的共通原則，做為撰寫此類指南的重點方向。諸如專案規劃、整合性工作流程、圖像管理、影音管理、文字管理、色彩管理、委外製作、數位內容保護機制以及全方位數位博物館建置，我們討論並挑選了這九項共通原則，開始進行調查、研究與撰寫，這類指南也於今年基本完成審查與出版。

在規劃寫作「精選物件」與「共通原則」指南的同時，我們即為這兩類指南設定了一種相輔相成、交互為用的關係。「共通原則」指南著重在分析數位化工作的各項重要主題，引導讀者對數位化的利弊得失做通盤而深入的思考；「精選物件」指南則描述特定物件的數位化實務與技術，便利讀者針對單一物件，選擇最合適、最有效益的數位化工作流程。透過這套「數位化工作流程叢書」的出版，相信可為更多有志投入數位化工作的單位與個人，提供富有整體性思惟並又能循序漸進的一套實用參考書。要特別強調的是：這套叢書的主要內容，仍植基於多年來陸續加入我們團隊的眾多機構與公開徵選計畫，這些工作夥伴多年來累積了眾多寶貴的數位化經驗，這些寶貴經驗讓更多數位內容，可以用更能保障精緻品質以及更能節省成本的方式，來從事數位資源的製作、展示、維護與授權工作，從而豐富我國數位典藏與數位學習事業。在完成出版這套「數

位化工作流程叢書」的同時，我們要特別感謝接受訪問的工作夥伴以及參與寫作的同仁，還有，對於協助我們審查與諮詢這套專書的所有學者專家，我們也在此致上衷心的感謝。最後，也盼望讀者隨時指正與建議，讓我們的工作可以做的更好。

數位典藏與數位學習國家型科技計畫
拓展台灣數位典藏計畫 • 數位內容建置與整合子計畫

計畫主持人  謹誌

西元 2012 年 6 月

編審序

自 2005 年以來「數位典藏與數位學習國家型科技計畫」編撰的《數位化工作流程指南》叢書，不但提供從事數位典藏者實用的參考資料與規範，於國內外數位典藏技術與品質的提升功不可沒；同時，本套叢書也可作為培育新生代的教材，筆者曾多次於課堂使用，學生們的學習效果良好，對於新一代人才培養有相當的助益。

藉由文字，人類得以溝通與表達思想，若能將文字紀錄以適當的方式保存下來，則人類知識運用於生活的記錄，將可長期保存與傳承。因此，自 2002 年起「數位典藏國家型科技計畫」，即致力於「以數位化技術保存重要文化資產」的推動，以確保數位藏品長久的儲存及應用。多年來，所執行的文獻、檔案等文字資料數位化內容已累積相當的成果，為了有效提升數位典藏品質與減低工作執行難度，故編撰此「文字資料」數位化工作流程指南，以供各界參考。

然而，因文字類型與載體的不同，文字資料的定義則有所不同。經討論後，本指南「以文字紀錄的文本（text）」為主體，採美國國會圖書館（Library of Congress）文字資料形式的定義，收錄各文本不同形體的文字紀錄，作為本書探討的範圍與內容。同樣，由於文字的結構與記載的載體差異，在進行文字資料的數位化工作時，數位化方式也會有所別。是以，需有一專門的調查與收集，彙整各種文字資料數位作業流程與經驗，就常見的問題提出可行的解決方案，配合參考範例，提供可用的規範與標準。故簡說本書內容如下，以便讀者參照索驥。

本指南工作團隊，首先規劃「文字資料數位化工作流程圖」（見書內圖 2-2），以之作為本書編輯作業的核心。於首章界定「文字資料」的定義與範圍後，即分為「實體物件的數位化」與「後設資料的建立」兩部分，探討文字資料數位化作業。並輔以「前置作業、數位化工作、資料保存、加值應用」四大步驟，分別描述文字資料的數位化流程。

「實體物件的數位化」的工作內容，則分為「輸入」及「文字處理」兩部

分說明。「輸入」一節中，就物件挑選前置作業、輸入的方式、光學辨識及校對等項，闡明輸入作業流程，並提供與「文字資料相關的數位化規格」作為參照標準。於「文字處理」一節中，則以缺字、異體字、標記及斷詞等作業，說明文字數位化獨特的處理方法。

「後設資料的建立」則分為「後設資料規劃」及「資料庫系統建置」兩部份說明。第一節以「後設資料生命週期」（見圖 4-1）為基礎，探討後設資料的需求評估，就不同藏品類型的需求建立後設資料工作表單，推介後設資料欄位與著錄的參考標準。第二節就「資料庫系統建置」說明數位資料的後續儲存管理、資料庫建置，與檢索系統的運用等功能。

文末於「加值應用」一章，就「數位學習」、「商業運用」二大類型，探討數位典藏內容可如何深化於教育、研究、生活與產業發展中，輔以實例說明，以思考數位文字資料的加值應用。

簡言之，本指南依既定的「數位化工作流程圖」流程與步驟，逐一介紹相關作業重點，讀者除了可以依序學習外，亦可就所遭遇的問題，依所需尋求解決方案。文中所列舉的規格標準，以及各單位所提供的寶貴經驗，則可作為參考與借鏡。誠如本書結語所示，希望藉由此數位化工作流程指南的紀錄，彙整文字資料數位典藏的相關技術與經驗，不僅可提供從事文字資料數位化工作之參照，並可拓展數位典藏深層的底蘊價值，成就傳承、傳播文字資料內涵的重要文化意義。

杜正民 敬誌

2012 年 5 月 1 日

目　錄｜CONTENTS

壹、前言

Introduction

一、什麼是文字資料

文字是文明的表徵，將思想資訊紀錄下來，促成知識的累積與傳承，使文明有系統地被留存下來。回顧歷史，在正式文字形成之前，人類大量利用圖案、符號來紀錄概念想法，隨著社會組織逐漸複雜，東西方社會文化漸漸發展出自己的文字，以做為溝通和記事之方法。

為了能閱讀與解釋古代的文獻，文字學研究興起，人們開始進行文字形體的辨識研討，並探究文字體系之演變，以瞭解傳統文化與歷史發展。文字最初是以「形符文字」為發展開端，運用符號、圖形來表達概念或想法，所以不論是在古代歐洲發展的西方文字或是古代中國文字中，都包含象形字符的元素。而在文字發展的過程中，象徵逐漸轉變為符號，並加入讀音結構拼寫的「聲符」。透過形符與聲符之組合產生各種不同結構的文字，例如古蘇美表形文字、埃及人的線性文字、中國商朝漢字…等形符文字，或是蘇美楔形文字、阿茲提克文字、馬雅文字、埃及音段文字、漢語…等形符與聲符混合文字。[1] 在眾多文字發展之中，中國的漢字是世界現存文字中使用時間最長者，也是世界上最古老的三大文字系統之一，從發展之初即不斷演變而沿用至今。在六千多年前新石器時代仰韶文化的陶器上即出現刻寫的符號，至殷商時期的甲骨文已是發展成熟的文字，往後各朝代更發展出各種文字的字體，如金文、篆書、隸書、草書及楷書等。

許慎亦在《說文解字》敘文中寫道：「倉頡之初作書，蓋依類象形，故謂之文。其後形聲相益，即謂之字。字者，言孳乳而浸多也。著於竹帛謂之書。書者如也。以迄五帝三王之世，改易殊體。封於泰山者七十有二代，靡有同焉。」說明字的發展是由「依類象形」的文，到「形聲相益」的字，漢字形體結構發生多次變化，並隨時代發展有不同的紀錄形式。[2] 若從牛津字典對文字的定義來看，其定義「word」為「字」或「單詞」，指的是透過語言或書寫可表達的最小單位（a single unit of language which means something and can be spoken or written）；而「text」指的是書籍雜誌中的正文（the main printed part of a book or magazine, not the

1 海拉 · 哈爾門 著，方奕 譯，《文字的歷史》，台中市：晨星出版，2005 年 4 月，頁 48-49。
2 林尹，《文字學概說》，台北市：正中書局，2007 年 10 月，頁 15。

notes, pictures, etc）、文本（any form of written material）、演講稿或劇本文稿（the written form of a speech, a play, an article, etc）。

故從文字之定義來看，除了直指文字結構本身之外，文字資料亦指當文字成為一種書寫行為時，所記載而成的資料。在文字出現之前，人類即在各種材料上記錄資訊，如岩洞、石材、骨頭、陶土（陶器）、金屬、木頭、樹葉、皮革、織物…等；隨著社會時代之演進，西元前二世紀發明記載文字的重要載體：紙張，至此之後書寫更為便利，促進人類文明的傳播。隨著資訊科技發展進入數位時代，書寫電子化，漸漸形成紙本與電子書同步存在的資訊社會。

二、文字資料與數位化工作

文字紀錄具有重要性，故在解讀文字之外，須將這些文字紀錄保存下來，以永久呈現人類生活與文化。早期文字的典藏只講究完整保存原始載體，但進入數位時代後，數位科技的力量協助我們進行文字資料的數位典藏，透過數位化方式不僅更有效地儲存文字，也加速人類文明的傳播。我國於 2002 年開始執行數位典藏國家型科技計畫，致力於以數位化技術保存重要文化資產、運用後設資料描述文化內容，並建立資料庫與全民分享這些重要資源，以確保數位藏品長期的儲存、維護及使用。數位典藏計畫執行至今已將近十個年頭，除不斷運用資訊處理技術典藏各式各樣的物件，也積極尋求更多創意的加值應用與推廣服務，像目前常見的電子書、漢籍全文資料庫、電子佛典、主題網站或數位典藏有關單位的推展服務皆是很好的應用實例。

文字紀錄為最普遍用來呈現知識的媒介，以數位化的形式將文字進行典藏亦具有重要的意義。然而，由於文字的結構不同、記載的載體不同，故在進行文字資料的數位化工作時，需依照文字資料之特性採用不同的數位化方式，以完整地將原始文字轉化為數位紀錄。有感於此，本指南將進行文字資料數位化作業探討，並整理歷年數位典藏計畫在執行文字資料數位化方面之成果，以期能做為執行文字資料數位化單位之參考。

本指南所指的文字資料，不特別鑽研文字的形音義，而是著重文字紀錄，

即「text」的意涵，也就是在各種形體上的文字紀錄。同美國國會圖書館（Library of Congress）在“Structural Metadata Dictionary for LC Digital Objects” 中定義文字資料形式為：「可透過肉眼辨識的印刷或手寫資料，包括圖書、小冊子、手稿、樂譜（Printed or handwritten material accessible to the naked eye. This includes books, pamphlets, broadsides, manuscripts, and musical scores.）[3]」，本指南將依此定義說明文字資料如何成為數位內容。

文字資料一般是透過人工打字輸入、掃描、拍照等形式將其轉換為數位內容，並在輸入過程中進行缺字、異體字、編碼與斷詞等文字處理，同時選定資料描述標準並對數位檔案的內容及屬性進行詮釋，以利數位化文字資料在資訊系統中的呈現與檢索查詢利用。因此，本指南將依文字資料之特性與數位化工作程序，依序說明文字資料之型式、數位化方式、後設資料運用、資料儲存與檢索，期望能呈現出文字資料數位化的執行流程與基本概念，並思考在數位化之後，有哪些加值再利用的面向，以讓數位化的優質內容能達到更多的傳承與推廣。

三、章節說明

第壹章、前言

從文字資料的起源、重要性和現今的數位化趨勢，縱談《數位化工作流程指南：文字資料》的編撰背景，並定義本書欲處理的「文字資料」範圍及撰寫目的。

第貳章、文字資料的數位化

此章節試圖從文字資料的載體與文字的類型分析，輔以目前國內數位典藏機構的相關技術與經驗為實例說明。並彙整出一個文字資料數位化工作流程簡圖，可從圖片簡要說明數位化工作的各個主要流程步驟。

---

3 Library of Congress. (2008). *Structural metadata dictionary for LC digital objects*. Retrieved November 30, 2011, from http://memory.loc.gov/ammem/techdocs/repository/attdefs.html

第參章、數位化工作

著重說明數位化工作流程中的「物件數位化工作」階段，主要闡述「輸入」與「文字處理」兩大數位化程序。針對文字資料各類的輸入方式（包括人工輸入、手繪、掃描、拍攝、光學文字辨識等）和數位化工作流程中文字處理的解決因應方法（包括各類編碼標準介紹、缺字、異體字等文字處理方式），除了說明其數位處理的方式並列舉相關的機構經驗以茲參考。

第肆章、後設資料規劃與資料庫系統建置

後設資料的建置可增進管理數位資料，是連結實體物件與數位檔案之間的橋樑，亦有利於建置資料庫的檢索系統。此章節以「後設資料生命週期」為基礎，由後設資料的需求評估為始，針對不同藏品類型的需求列舉常見的後設資料工作需求表單、介紹目前後設資料欄位建立與著錄的幾項參考標準，並延伸討論數位資料的後續儲存管理概念與資料庫檢索系統的運用。

第伍章、加值應用

將珍貴的文化資產進行數位化保存、管理與檢索應用後，更重要的是能透過各種加值應用的方法，使這些資產擁有最佳的附加價值。本章節將加值的應用模式分為「數位學習」與「商業運用」兩種類型，每個類型皆輔以實例說明，希冀透過本章節可讓讀者了解不同的加值應用形式，以便在數位化過程中，能選擇最適切的加值方式將數位資料做最佳的應用。

第陸章、結語

希望藉由此數位化工作流程指南的紀錄，可以彙整目前國內從事文字資料數位典藏的相關技術與經驗。不僅是協助從事文字資料的數位化工作，並且拓展數位典藏深層的底蘊價值，更是傳承、傳播文字資料內涵的重要文化意義。

貳、文字資料的數位化

Digitization for Text Data

文字資料的數位化，意指將史料、手稿、書籍、期刊報紙等紙本資料轉成數位電子檔，並利用電腦作業軟體系統加以處理與管理。這些文字資料儲存為容易調用、可再編輯、搜尋、檢索、瀏覽、再利用等格式。其文字資料數位化後的好處更是便於保存、管理與再利用，對於資料流通交換、描述事件等用途皆大有益處。[4]

一、文字資料的內容

關於文字資料數位化的工作，就數位典藏現有的計畫為例，以下依「文字資料的載體」和「文字的類型」兩大類分別加以說明。

（一）文字資料的載體

文字的書寫並非一開始都記錄於紙張，在紙張尚未問世以前，上古時期的人們會以結繩來記事。後來中國的文字書寫開始藉由雕刻的方式在龜甲、獸骨、金石、竹簡或絲帛等物上做記錄，接著才有了紙張、甚至膠片、磁帶等數位載體。

因此文字資料若以載體來區分，從早期開始有文字歷史記錄算起，可涵括書寫於甲骨、青銅器、竹簡、帛書等器物上的所有文字資料，以及最普遍的書籍、期刊、報紙類的載體，至近現代科技時代儲存文字的數位電子檔案等，每一階段的文字資料載體都有其特有的記錄方式。「數位典藏與數位學習國家型科技計畫」已執行多年的數位典藏工作，亦有不少單位針對文字資料的數位典藏有了豐富的經驗與成果。下文將依各個計畫執行不同載體的文字資料數位化工作，做一簡要的介紹。

1. 甲骨

古代人們利用龜甲、獸骨刻畫各種符號、紀錄其生活文化，在早期考古出土材料流通不便、取得不易的情況下，以此類載體所記錄的文字

4 陳昭珍，〈文字資料的數位化〉，數位典藏學程—人文領域，檢索：2011 年 12 月，http://humanities.lis.ntu.edu.tw/md/20070314.pdf。

或圖像，多以拓片形式將其影像複製以利方便傳播。

以中央研究院歷史語言研究所建置的「甲骨文數位典藏資料庫」[5] 為例，是處理典藏於史語所的甲骨文拓片，約有四萬餘件，大部分都以掃描方式建立數位影像檔。然而因文字、圖像都刻畫於甲骨片上，所以拓片所得的資料多少已有些不清楚之處或有破損的問題。因此基於人力、物力考量，目前該資料庫所建置的內容以較為完整、清楚的拓本為數位掃描的底本。將甲骨文拓片數位化並建置資料庫後，可透過圖文對照功能，檢索甲骨文的基本數據和文字資訊，並且也能與拓片影像進行對照。唯該資料庫呈現仍有漢字缺字的問題，亦是一項待克服的文字處理難題。

2. 青銅器

青銅器拓片（或稱金文拓片）保留了刻或鑄於青銅器上的銘文、紋飾及器形資料，此類珍藏拓片的數位化工作模式大多亦是和甲骨文拓片一樣採取掃描方式進行。此外，使用數位機背攝影拍攝是較不易損毀原拓的方式，且掃描時間較平台掃描短，不需經過切割掃描及人工接圖，也更利於編排管理檔案。[6]

目前已開放檢索的相關資料庫有中央研究院歷史語言所所建置的「青銅器拓片數位典藏」[7] 資料庫，以及國家圖書館「金石拓片資料庫」[8] 等，可依全形拓片和銘文拓片檢索相關資料。

---

5 甲骨文數位典藏資料庫，檢索：2011 年 12 月，http://ndweb.iis.sinica.edu.tw/rub_public/System/Bone/home2.htm。

6 數位典藏與數位學習國家科技型計畫，〈中央研究院歷史語言研究史語所藏青銅器拓片數位化工作流程簡介〉，拓展台灣數位典藏計畫，檢索：2011 年 11 月，http://content.teldap.tw/index/?p=1102。

7 青銅器拓片數位典藏資料庫，檢索：2011 年 11 月，http://rub.ihp.sinica.edu.tw/~bronze/index.htm。

8 金石拓片資料庫，檢索：2012 年 2 月，http://rarebook.ncl.edu.tw/gold/。

3. 簡牘

簡為竹片、木片，牘為木板，是古代在紙張尚未普及前用來書寫文字的載體。目前發現的簡牘年代主要是以戰國、秦漢、三國至西晉的為主，中央研究院歷史語言研究所收藏了不少漢代邊塞地區出土的簡牘文書，是國內一批價值非凡的文化資產。有鑑於數位化的應用在漢簡研究上的助益，這些重要的簡牘透過紅外線攝影儀、掃描器、電腦多方數位化工作，重新釋讀藏品簡牘，並進行影像釋文資料的數位化，建置了「歷史語言研究所藏漢代簡牘資料庫」[9]。此資料庫提供了漢簡基本資料、釋文以及彩色照、紅外線照、反體照等影像瀏覽，透過「歷史語言研究所藏居延漢簡遺址查詢系統」，可整合資料庫與地理資訊系統的時空資訊，增進漢簡研究的便利性。

4. 手稿

文字書寫的紀錄在進入紙張時代後，成就了許多的重要文獻、藝術作品、歷史檔案等。許多文書檔案的手稿正是數位典藏工作的重點之一，舉凡文學作家、音樂家的創作手稿、建築家的手繪稿，甚至不少古籍經典的手抄本等，都有具有相當價值的文化歷史意義。在文字資料的數位化處理方面，大致以掃描文本為影像檔案為主，若需建置全文資料時，則採用打字輸入的方式。但全文打字輸入往往需要不少經費與時間，若手稿原件內容多含古字、異體字，甚至可能有較潦草難辨的文字，則輸入工作的人員素質也必須具有相關的學術或經歷背景才行。[10]

例如台灣大學「台灣文獻文物典藏數位化計畫」將《淡新檔案》、《伊能嘉矩手稿》等台灣珍貴檔案、善本古籍的全文與影像數位化，並製作後設資料、資料內容全文輸入、判讀與斷句等。其中《伊能嘉矩手稿》

9 漢代簡牘數位典藏資料庫，檢索：2012 年 2 月，http://rub.ihp.sinica.edu.tw/~woodslip/index.htm。

10 洪淑芬，《文獻典藏數位化的實務與技術》，台北：數位典藏國家型科技計畫 訓練推廣分項計畫，2004 年 2 月，頁 21。

為日本台灣研究學者伊能嘉矩的珍貴手稿、書信、手繪圖、照片、田野調查筆記等台灣日治時期史料，這些都以掃描的方式建置為影像數位檔。《淡新檔案》除了進行掃描外，亦採全文輸入打字，並且進一步印刷排版成冊出版，網路上也提供全文及影像資料的檢索利用。但是《淡新檔案》的原件使用大量俗體字、異體字，甚至有罕用字缺字的情形，這些文字處理問題也都透過中研院「漢字構形資料庫」工作小組一同研究配合解決資料庫呈現的形式。本書後面章節也將針對此類文字處理方法再做進一步介紹。

5. 書籍 / 期刊 / 報紙等紙質印刷類

有別於針對早期貴重的手稿、文獻的數位化處理，對於書籍、期刊雜誌、報紙的紙質印刷類數位化執行工作又有所不同。這些從年代久遠的珍籍資料到近年的成冊文獻、雜誌書籍的數位化工作，除了需考量製作成本價格、機器選擇外，不同文件類別（文件型或單件文物、雜誌書籍等）掃描為數位影像的規格需求也有所差異。若欲全文建檔尚需採人工輸字或以光學文字辨識（Optical Character Recognition, OCR）轉換等數位化工作處理方式。[11]

以法鼓佛教學院「台北版電子佛典計畫」[12] 經驗為例，其資料庫的建置包羅歷代中國不少佛教經典著述，工程可謂浩大艱辛。其中中華電子佛典協會（CBETA）開發的電子佛典作業，從選定材料、制訂各式作業規範都有一定的流程守則。大量佛典經文的輸入，則針對不同內容，分別採用收集現成電子檔、人工輸入，以及 OCR 圖檔辨識等方式來產生文字檔，再加上後續的校對、標記、缺字處理等都是建置書籍全文資料庫

11 洪淑芬，〈紙質文獻類的雜誌書籍之數位化〉，《佛教圖書館館刊》，第 45 期，2007 年 6 月，頁 19-25。

12 數位典藏與數位學習國家科技型計畫，〈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〉，拓展台灣數位典藏計畫，檢索：2011 年 11 月，http://content.teldap.tw/index/?p=1035。

所需費心的工作。

同為紙質印刷類的期刊、報紙文獻數位化方式也多以掃描影像、重新打字輸入、光學文字辨識等方式數位化。但是早期報紙除了以原件類型收藏外，也有彙集製作成微縮軟片（microfilm）及拍攝成單張黑白底片的形式。原件的保存狀況即攸關數位化的品質，因此紙質與印刷品質、破損狀況、缺頁及裝訂方式，有時尚需經過專業的修補才能進行數位化。這些期刊、報紙的數位化執行作業，也多以影像掃描、人工輸入、光學文字辨識等方式進行文字數位化。以微縮軟片載體形式保存的期刊報紙，將列於下一小節做介紹。

6. 微縮軟片（Microfilm）

微縮資料是利用微縮攝影機把比較大的檔案、文獻、圖書等資料，以數倍甚至數百倍的縮小比例，攝製於微縮軟片，經過化學或物理加工製成的微縮品，所用化學藥劑極微細小。因其特性，早期又譯做微粒資料、微影資料或微縮技術、微縮資料等。[13] 從 1930 年代以來，微縮軟片即成為承載人類知識的重要媒介，許多重要的文獻資料皆以微縮軟片複製本保存，甚至原件受到毀損或遺失，僅剩微縮軟片型態的資料依舊保存存在。因此大量的期刊報紙常選擇以此型態做為複本保存時，一旦要將其數位化，就得針對不同的微縮軟片形式（例如，微縮捲片、條狀微縮軟片、夾檔、微縮單片等）進行不同的數位化作業。

以「北平世界日報內容數位化開發計畫」[14] 為例，該計畫在取得「世

13 林彥宏、程婉如、張思瑩，《微縮資料數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2011 年 6 月，頁 12。

14 數位典藏國家型科技計畫，〈北平世界日報內容數位化開發計畫之數位化工作流程圖文說明〉，《國家數位典藏通訊電子報》，檢索：2011 年 11 月，http://www2.ndap.org.tw/newsletter06/news/read_news.php?nid=544。

界日報」原報紙內容全文的微縮捲片後，利用微縮閱讀機讀取微縮捲片全文，並列印全文影像以製作新聞原版報紙，再選擇需要全文處理的資料後，進行新聞資料的打字輸入成電子檔，並針對缺字情況做資料處理、資料校對與匯入，完成網站資料庫的建置。

「國家圖書館古籍文獻典藏數位化計畫」[15] 則因館藏善本古籍已攝製微捲（黑白），為加速進行數位化故決定由古籍微縮資料轉製成黑白影像，並利用微縮影片掃描機將微縮片的影像放大與原書尺寸相同再進行掃描，轉製成影像檔案格式。完成掃描的檔案還需去除黑邊影像、污點，並檢查影像是否歪斜、校對頁碼與原始物件的對應等品質，以利後續的文字資料等處理、輸入。

7. 數位資源

有別於利用器物、紙張等載體記錄文字資料，在邁入所謂的數位化時代之際，大量的資訊以數位化的形式產生並儲存於各式的數位載體。一般來說，數位資源（digital resources）的形式分為兩種，即「數位化資料」（Digitized Materials）以及「原生數位化資料」（Born-Digital）兩類。前者是將文件或是其他媒體利用掃描、數位攝影等方式轉換為電子形式，間接產生數位資源；後者則是指數位化資料原本就以電子形式創造。有時數位資源也稱做數位檔案、電子檔案。[16]

此處所欲討論的數位資源，即是原生數位化資料，以各式數位檔案、媒體為載體儲存記錄文字的部分。舉凡，電子公文、電子期刊、電子報等電子格式的各種資料皆是。此類的文字資料本身就是數位檔案，所以實體數位化的程序即可省略，直接檢視資料本身的數位化品質，並進行數位化後製、數位檔儲存、資料著錄、檢索系統建置等。

---

15 程婉如，〈微縮資料委外轉製影像專訪報告—以國家圖書館、世新大學為例〉，拓展台灣數位典藏計畫，檢索：2011 年 11 月，http://content.teldap.tw/index/blog/?p=305。

16 詹雅蘭，《OAIS 參考模式應用在國家檔案永久典藏機制之探討》，台北市：台灣師範大學圖書資訊學研究所，2004 年 6 月，頁 6。

例如國內的聯合報報系，已將報紙編排方式數位化，每日的新聞資料文稿皆儲存於資料庫中，這些數位檔案資料便可有效建立更多的檢索運用。其在 2000 年成立了聯合新聞網，是一多元化的數位媒體網站，包括其下的聯合知識庫（會員制），完整收錄聯合報系所發行的海內外各類報紙，已有一千多萬則以上新聞資料、新聞照片等，並建有全文檢索、專卷查詢、影像圖庫等多項查詢功能。不僅是將歷年來的傳統資料數位化外，並直接有效處理數位內容的一個多元資訊平台線上資料庫。[17]

（二）文字資料的類型

文字資料的數位化工作，可依據文字資料的載體類別來決定數位化處理方式，以及這些文字資料如何管理與應用；此外，依文字的類型與結構不同，也有相異的數位化方式。以目前國家型數位典藏計畫為例，除了漢字為大宗的文字資料外，亦有不少少數民族或其他文字類型的文字資料存在。在數位化的程序裡，文字的處理正因不同民族文字特性得採用不同的軟硬體設備，以達有效地數位化。尤其是欲透過資料庫做資料檢索與加值應用時，在影像處理、全文建置甚至後設資料的著錄等流程，都需建立一套相關作業規範。以下將依據不同的文字類型所使用的數位化處理方式做介紹，並佐以簡介幾個相關的數位典藏計畫經驗為參考。

1. 漢字

文字資料的數位內容，在目前的執行數位典藏計畫當中，以漢字的文字資料為大宗，尤其是處理漢籍文獻等方面者為最。透過掃描或數位攝影可將書籍期刊等文獻數位化，但其中的文字內容處理則需進一步的軟體協助管理建置。數位時代下，電腦處理漢字的方式是一字一碼，然而不同區域使用的編碼方式也不盡相同。以台灣區域使用漢字的情況

17 聯合新聞網，檢索：2011 年 12 月，http://udn.com/NEWS/main.html。

來說，有使用 Big5 碼，也有使用 Unicode 碼；大陸地區則是除了使用 Unicode 碼外，還有使用 GBK 碼；其他地區如韓國、越南、日本等地也各自有所屬的編碼系統。[18]

漢字編碼使資訊可藉由電腦相互溝通，但是電腦處理漢字資料仍有許多的問題待解決，尤其漢字的演進主要經歷了甲骨文、金文、篆文、隸書、楷書等字體的嬗變，亦產生了不少字型結構的差異。加上歷代的文獻中常出現當時的俗字，這些都是處理電子古籍中普遍面對的現象，亦是「字」不夠用的原因之一。這些字大多是屬於字的「異體」，因此處理缺字問題時，不能不兼顧到「異體字」的問題。[19]

中央研究院建有「漢籍電子文獻」[20] 資料庫，是目前處理漢字資料的重要單位之一，其所處理的文字內容、文獻是以 Unicode 編碼系統為主要方式。有關漢字構形以及電腦字型檔的處理問題，則會借重中央研究院資訊所文獻處理實驗室的專業技術，加以研發與處理更多的缺字或異體字等。

2. 藏文、梵文及其他少數民族文字

除了漢字的文字資料以外，亦有數位典藏計畫的執行內容是有關於其他少數民族的文書或是藏文、梵文等佛典的數位化工作。這些文字資料的數位化主要以先掃描為影像圖檔為主，再針對掃描圖檔進行文字識別處理。

數位時代下，佛教文獻因大量數位化的成果，越來越多人仰賴網路、電子資料庫來進行佛學檢索與相關研究。佛學歷史悠久，舉凡梵文、巴利語、藏文、漢語等各類語言、文字、版本繁多，透過網路的整合力量，

18 羅凡聶，《文字學數位內容加值應用之研究》，台北縣：花木蘭文化出版社，2010 年 9 月，頁 7。
19 謝清俊，〈漢字的字形與編碼〉，文獻處理實驗室，檢索：2011 年 11 月，http://cdp.sinica.edu.tw/paper/1996/19961004_1.htm。
20 漢籍電子文獻瀚典全文檢索系統，檢索：2011 年 11 月，http://hanji.sinica.edu.tw/。

數位內容資訊愈加豐富，文字處理的技術也不斷地在精進。

以梵文為例—梵語，是一種古印度所使用的語言，梵文就是相對使用的文字，印度教經典《吠陀經》就是梵文寫成。梵文通常由一套符號式的字母（sign alphabet）組成的天城字體（Devanågarª script）來書寫。[21] 唯梵語是一項很古老的語言，其語言和梵文字體之間，也會經歷語言史的分化、疊合、演變，而有極大的變化，加上梵文佛典的網路資源尚屬短缺，解讀、轉寫和校對的工作仍不太精細，許多的檔案整合仍有其困難度，全文資料庫的檢索建置也還有很長的一段距離。

再者以藏文字為例，是一種拼音文字。如圖 2-1，藏文是透過基本字符和基本字符縱向疊加而成的字符串，構成一個完整藏文詞素，基本單位是由藏文中的「音節分割符（tsheg bar）」來確定。一個藏文詞由一個或多個音節構成。[22] 但因其拼寫具有縱、橫向的拼寫性，有些字母會出現變形，相似字符也多，所以藏文辨識的難度也相對增加。有關藏文的識別研究，目前多集中在印刷體識別技術中，礙於大量相似字的區別難度，手寫識別方法確實仍有研究空間，是非常必要的文字處理技術。

21 蔡耀明，〈網路上的梵文與梵文佛典資源〉，《佛學數位資源之應用與趨勢研討會論文集》，2005 年 09 月，檢索：2011 年 12 月，http://buddhism.lib.ntu.edu.tw/BDLM/seminar/book0.htm。

22 劉芳、歐珠，〈藏文文字識別系統中的數字圖像預處理方法研究〉，《中國少數民族語言文字信息處理研究與發展》，北京市：民族出版社，2010 年 6 月，頁 257。

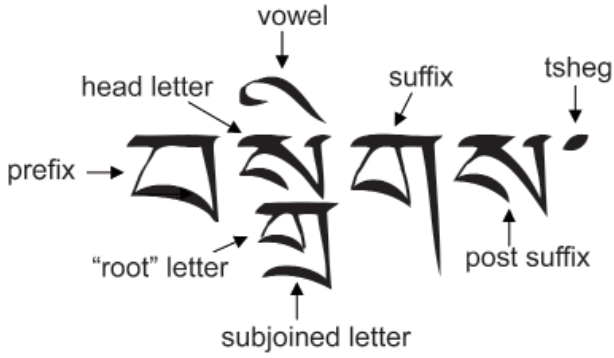


圖 2-1：藏文字的各組成構建
資料來源：《計算機工程與應用》，歐珠、普次仁、大羅桑朗杰、趙棟才、劉芳、邊巴旺堆等 [23]

在少數民族文字方面，中央研究院歷史語言所建置的「中國西南少數民族資料庫」[24]，其資料庫處理了不少西南少數民族的文書資料。以納西民族的「東巴文」為例，因其專門被納西族祭司「東巴」用於書寫宗教經書和宗教活動等方面，所以稱為「東巴文」。東巴文是一種原始的圖畫象形文字，是一種比甲骨文更原始的文字，在書寫過程中也比較隨意，相對地在文字處理方面也增顯難度。除了透過影像掃描的方式數位化外，為求資料的正確性與清晰度，還是得請相關研究專家親自再次手繪圖像。

3. 台灣白話字

白話字（Peh-o ē -j ī , POJ）是一種以羅馬字母所拼寫而成的閩南語正字法。由於原先是由 19 世紀的基督教長老教會於福建廈門所創設並推行的拼音文字，因此又被稱為「教會白話字」或是「教會羅馬字」（Church

23 歐珠、普次仁、大羅桑朗杰、趙棟才、劉芳、邊巴旺堆，<印刷體藏文文字識別技術研究>，《計算機工程與應用》，2009 年，第 45 卷第 24 期，頁 166。

24 中國西南少數民族資料庫，檢索：2011 年 11 月，http://ndweb.iis.sinica.edu.tw/race_public/index.htm。

Romanization）。[25] 台灣因基督教的傳入，故將白話字的書寫方式也帶進台灣。「白話字」稱呼的起源，一開始是為了區別不同的漢語書寫方式。[26] 因閩南人日常所使用的白話既非高深漢文，亦非母語以外的官話，而是平日口語的書寫，所以將其稱為「白話字」。

「白話字」在傳入台灣後，使用者還延伸至客家及原住民族的語文書寫。由於書寫系統的歷史悠久且普遍被使用，已被通稱為「台灣羅馬字」。尤其是 1980 年代末期後，已不少透過漢羅並用（漢字、羅馬字）的書寫方式，呈現台灣各族群的文學作品或論述。

「台灣教會公報（1895-1969）白話字文獻數位典藏計畫」的數位化處理的主要工作，大致是先將文獻掃描為影像檔儲存，也選取出部分白話字文獻重新打字，並且同步進行打字後的文章作「漢羅台語」的翻譯，讓使用者更加容易閱讀與查詢。其所建置的「台灣白話字文獻資料館」，即是將《台灣教會公報》內容以打字、翻譯、校對、上稿等工作外，亦將「北部台灣教會公報」《芥菜子》和早期白話字出版品加以掃描影像數位化，並將部分內容數位化以及漢羅翻譯，是一個文字與影像內容豐富的文獻資料庫。

4. 其他

以台灣的數位典藏計畫而言，目前處理文字資料的文字類型仍以漢字為最，其他如日文文獻、英文、西班牙文等其他文字類型，因處理方式和漢字文字資料數位化方式相類似，暫不再贅述。本指南後面章節也將針對幾個普遍遇到的文字處理問題，列舉目前的數位化方式以及未來文字資料可能發展的方向再加以討論述之。

---

25 國立台灣師範大學台灣文化及語言文學研究所，〈台灣白話字發展簡介〉，台灣白話字文獻館，檢索：2011 年 11 月，http://www.tcll.ntnu.edu.tw/pojbh/script/about-2.htm。

26「白話字」之稱，乃為區別其他種不同的漢語書寫方式：第一種是文言的漢詩、漢文等傳統的書寫方式，過去稱為「孔子字」；第二種則為中國北京話的白話文書寫方式，稱「唐人字」。檢索同註 25。

二、 文字資料的數位化工作流程

透過上一節文字資料的內容簡要介紹，大致了解文字資料的載體與類型在數位典藏工作上主要採取的數位化方式以及可能面臨的問題等。並針對文字資料的數位化工作，試圖歸納出整個工作流程的主要步驟、程序，繪製一個流程圖重點標示。

（一）文字資料的數位化工作流程圖

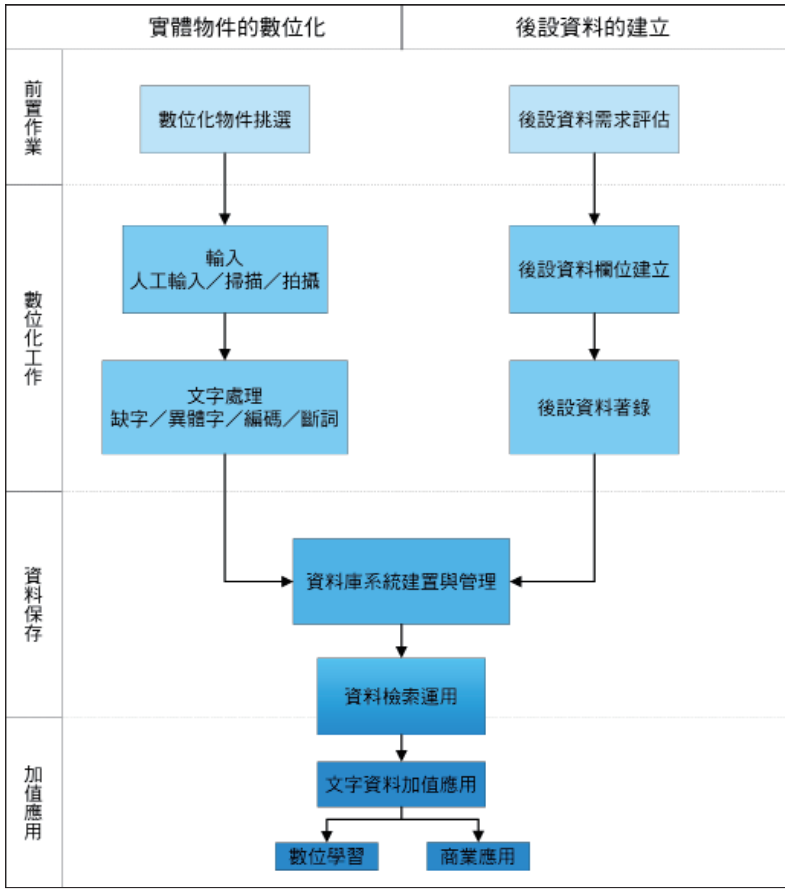


圖 2-2：「文字資料」數位化工作流程圖
資料來源：拓展台灣數位典藏計畫繪製

（二）數位化工作流程圖簡說

文字資料的數位化工作流程，在初步判別所屬資料的內容類型後，需規劃考量欲數位化的方式，並評估後設資料的需求建置，以及資料儲存管理甚至後續的運用檢索等層面。如上圖 2-2 所示，主要的流程步驟有四：前置作業、數位化工作、資料保存、加值應用等程序。

1. 前置作業：

釐清文字資料的載體與類型是考量數位化處理方式的第一步。此外，主要的前置工作還包括了瞭解藏品狀況及物件的清查、清冊整理，並挑選欲數位化的物件，進行工作規劃與選擇。同時包括研擬後設資料需求評估與內容分析，後設資料欄位調整、訂定著錄規範等。

2. 數位化工作：

文字資料的數位化流程在此階段屬於重點工作。需依照文字資料的載體與類型選擇數位化「輸入」的方式，例如人工輸入打字、掃描、攝影等，建置初步的數位檔。再者，進行特殊文字資料處理工作，還需針對缺字、異體字、編碼、斷詞等問題尋求適切的軟硬體解決之道。數位檔完成後需搭配後設資料的完整欄位，其資料的詳細著錄有助於後續管理與檢索運用等層面。

3. 資料保存：

受到資訊科技發展的影響，所有數位資料的保存皆面臨相同的儲存管理問題。如何延續數位資料生命週期的正常運作，除了硬體設備的因應外，舉凡不同作業系統的相容性、儲存媒體的時效性、數位檔的標準格式，或是資料是否適用等問題，這些都是資料保存亦需有相對應的規劃。

4. 加值應用：

文字資料數位化，在網路時代除了讓大眾更易於獲取資訊、進行研究外，也開啟更多面向的可能。如何應用這些龐大的文字資料數位內容，以更多不同的面貌、多元化方式的呈現，亦是文字資料數位化關心的範

疇。不僅是數位資料庫的檢索運用，在學術教育的數位學習方面、商業加值的層面，甚至是公共社會的各種創新服務型態，都是加值應用階段可執行的重要方向。

此四大步驟流程主要是概括式的統整程序，每一環節都相輔相成，至於數位化過程的細節問題，仍須依照不同的物件選擇相應的處理方法。

參、數位化工作

Digitizaion

本章節著重說明數位化工作流程中的「物件數位化工作」階段，以「輸入」和「文字處理」兩大數位化程序為主。除了簡述前置作業的基本物件挑選作業外，將介紹文字資料所常見的輸入方式（包括人工輸入、手繪、掃描、拍攝、光學文字辨識等）和數位化工作流程中文字處理的解決因應方法（包括各類編碼標準介紹、缺字、異體字、標記等文字處理方式）。以下介紹目前數位典藏相關計畫在文字資料數位化的執行現況，並列舉幾個特殊文字處理案例做為參考。

一、數位化物件挑選

數位化物件挑選是數位化工作流程「前置作業」中重要的程序，包括對資料的瞭解、清查與製作清單和相關表格、訂定數位化規格、作業標準等項目。在文字資料物件的挑選工作，主要著重於下列幾個步驟：

（一）數位化物件的瞭解與盤點

文字資料來源以古籍為多數，多屬特殊、珍貴的歷史資料，在初步清點時應有一定程度的瞭解。資深管理員的協助，或負責人對於資料的特殊性應有所註記，並加以指導相關工作人員進行盤點。例如，有受損者還需進行哪些的修補程序等。對於資料內容的保存狀況、文件內涵等內容，必要時還得再加以註記，以利數位化工作的進行及資料管理。[27]

（二）物件選定與建立清冊

前述資料盤點是規劃數位化工作的基礎流程，從中選定即將進行數位化的資料、物件之後，應建立一份更為詳實完整的資料清冊，作為之後進行數位化工作的憑據與管理。

---

27 洪淑芬，《文獻典藏數位化的實務與技術》，台北：數位典藏國家型科技計畫 訓練推廣分項計畫，2004 年 2 月，頁 8-10。

（三）訂定規範

為確保數位化工作流程的每個環節銜接順利，產出的成果品質良好，訂定相關作業標準規範、檔案格式以及後設資料建立評估與內容分析等工作，都是前置作業中亦需建立的遵循事項。目前各類數位檔案格式及後設資料規格等細節，後面的章節將有更進一步說明。

二、輸入

在挑選完數位化物件的前置作業後，即進入數位化工作的第一步：輸入。文字資料輸入方式之選擇主要考量到文字資料本身的類型與狀況，並且因數位化之目的與用途不同，而採用不同的技術與標準。文字資料的輸入有許多方式可以選擇，包括以人工透過電腦輸入文字、手繪文字、將文字資料拍攝成圖片，以及掃描成圖檔或微縮片等。而在文字資料輸入電腦之後需進行校對工作，以確保資料之正確性。

本小節除說明輸入作業外，亦簡述如何將數位化資料（Digitized Materials）進行光學文字辨識（Optical Character Recognition，簡稱 OCR），以及說明數位化建議規格與校對等輸入相關工作如下：

（一）輸入方式

1. 人工輸入

人工輸入是一種最基本的文字數位化方法，主要是將文字資料之原件內容運用人工方式重新打字鍵入電腦，或是將過去已經掃描或拍攝之影像檔案之內容重新輸入成電子檔。雖然現今科技技術已能將掃描檔案透過 OCR 技術將檔案分析處理而獲取純文字檔，但許多年代較久遠的文獻或是手稿等無法透過電腦進行文字辨識之檔案資料則仍需依賴人工逐字輸入。

因人工輸入工作僅需基本電腦打字輸入技能且費時耗力，故大多執行數位化工作單位皆採用委外製作方式進行；然而若遇到以古字、

變體字為主的文字資料，則還是建議交由專業人員執行建檔。人工輸入看似簡單，但由於文字資料相當多元，故不論是委外或是由專業人員進行人工輸入，皆需事先製作文字輸入規範，以做為輸入時之參考。

文字輸入規範主要是將著錄格式明確標示出來，包括內文的文字、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，且因為每種文獻的排版、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，制訂適合個別體例之人工輸入規範。[28]

舉例來說，中華電子佛典協會進行佛典數位內容建構時，即運用人工繕打來輸入無法 OCR 辨識的佛經，在進行〈大藏經〉數位化時，即訂定〈續藏輸入規則及範例〉，說明佛經中空行、空格、圈點、單行夾註小字、雙行夾註小字、校勘符號、特殊符號……等輸入原則。

若以報紙類文字資料人工輸入為例，「世新大學北平世界日報內容數位化開發計畫」[29] 在執行數位化過程時，首先規範數位化的範圍，主要收錄國內要聞、世界要聞、各省新聞、經濟界、教育界、世界所聞及地方新聞；另由於報紙資料排版有既定的規則，故在人工輸入資料時依照流水號（no）、標題（ti）、副標題（ts）、日期（da）、版次（ed）、版名（cl）及全文（ct）鍵入，並以「則」為單位，作為網站上檢索之用途。

在文字輸入作業中除了依文字資料特性不同調整輸入原則外，另一個工作要點在於文字辨識。在文字輸入時若遇無法辨別之文字應交由專業人士判斷，若遇缺字則需進行造字。一般在遇到中文缺字狀況時，可利用中央研究院文獻處理實驗室發展的「漢字構形資料庫」解

28 王雅萍、謝筱琳，《漢籍全文數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2011 年 6 月，頁 21-23。

29 郭哲裕、羅玉容、汪怡慧，〈世新大學北平世界日報內容數位化開發計畫〉，拓展台灣數位典藏計畫，檢索：2012 年 2 月，http://content.teldap.tw/index/?p=85。

決缺字與異體字問題，或是下載行政院主計處電子處理資料中心建置之「CNS11643 中文標準交換碼全字庫」[30] 應用工具。有關缺字及異體字之處理將於本章第三節說明。

2. 手繪文字

並非所有文字皆可透過人工電腦打字的方式輸入，當文字無法以打字輸入時，為了能將文字內容於電腦中呈現，故委請專業人員以人工手繪文字後再掃描文字成圖檔輸入電腦。如中央研究院歷史語言研究所執行「民族學調查標本、照片與檔案數位化工作」時，由於許多文書內容為少數民族特有文字，無法以人工鍵盤輸入，故將特定需要著錄的文字（例如文書標題），委請編目譯解者以手繪的方式繪製文字，再掃描成數個圖檔，上傳至資料庫供文書翻譯頁面使用。

以納西族東巴經後設資料著錄及翻譯者繪製的納西族東巴經原文影像為例，圖 3-1 為納西族東巴經的〈破地獄經〉原圖，翻譯者將圖中文字手繪呈現如圖 3-2。該文書內容意指超度什羅亡靈儀式，是送什羅祭司亡靈到格補命在的地方經；翻譯者將內容直譯為「超度什羅亡靈列達二十二個象地」。該檔案詳細後設資料著錄可參考節錄於〈後設資料內涵分析報告：西南少數民族一文書後設資料分析書〉[31] 一文的「附錄一、納西族東巴經之〈破地獄經〉文書翻譯資料」。

圖 3-1：納西族東巴經的〈破地獄經〉

30 全字庫，檢索：2012 年 2 月，http://www.cns11643.gov.tw/AIDB/welcome.do。

31 中央研究院歷史語言所，〈後設資料內涵分析報告：西南少數民族一文書後設資料分析書〉，數位典藏技術彙編 2007 年版，檢索：2011 年 1 月，http://www2.ndap.org.tw/eBook08/showContent.php?PK=220。

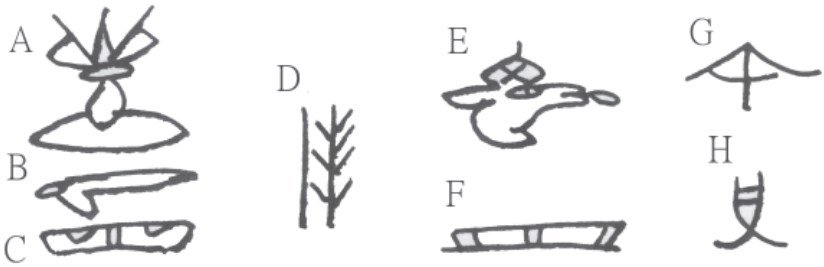


圖 3-2：納西族東巴經的〈破地獄經〉手繪文字

3. 掃描技術

掃描是運用電子設備讀取資料或圖像並將之轉為數位影像，一般平面的文字資料如圖書、期刊、報紙、檔案等物件，多透過掃描方式進行數位化。掃描的原則在於依據典藏及應用等不同目的，選擇合適的設備與儲存格式。掃描之前應先有前置作業，進行資料盤點、建立掃描清單、物件狀況檢查、資料編碼。於執行掃描時應制定掃描規範標準，標準內容包括物件取用原則、工作人員配合事項、影像製作規格、資料掃描原則、選定掃描設備等，掃描完成後進行校對工作，並將原資料歸檔及進行數位檔案後續整理。透過一致性的掃描作業流程（如下圖 **3-3** 所示），以確保掃描輸出之品質。

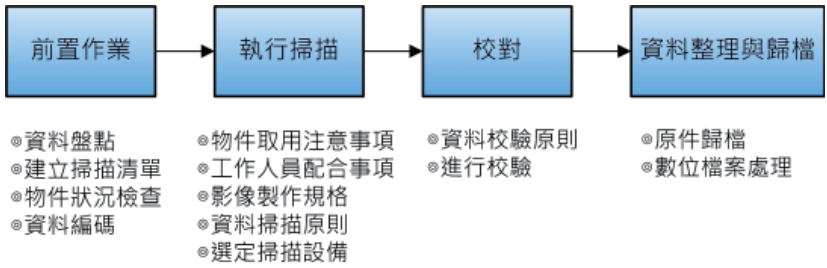


圖 3-3：掃描作業流程
資料來源：拓展台灣數位典藏計畫重整繪製

整體來說，在執行掃描時，普遍性的掃描作業流程可歸納如下 ：[32]

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，若有缺漏或無法開啟的檔，加以修改或補齊。
9. 歸檔。
10. 清潔掃描器。

由於各種資料物件狀況不同、各數位化單位對產出品質之要求不同，故進行掃描時需針對物件特性制定工作流程與規範。以善本古籍為例，傅斯年圖書館針對善本古籍館藏數位化工作制定〈傅斯年圖書館全彩影像掃描、拍攝及校驗相關作業標準〉[33]，該標準規範書況檢查、清點檔案並登錄註記、判定掃描或拍攝方式、掃描人員配合事項、原件掃描原則、原件校驗原則等內容。掃描前之書況檢查包括：逐冊與逐頁檢閱、判定是否修裱（焦脆 / 剝落）、處理書況（摺角、摺痕、破洞、木屑、遮字、書簽、頁次裝訂錯誤等項目）、判定掃描或拍攝方式；其中掃描方式則依書況分為以下幾種方式，並分別註記：

1. 依原書。
2. 攤開。

---

32 王雅萍、謝筱琳，《漢籍全文數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2011 年 6 月，頁 31-32。

33 傅斯年圖書館，〈傅斯年圖書館全彩影像掃描、拍攝及校驗相關作業標準〉，檢索：2011 年 11 月，http://lib.ihp.sinica.edu.tw/pages/03-rare/DAP/contentp/03-4-2.pdf。

3. 兩頁併掃。
4. 夾紙另掃。
5. 紙條隨內文掃存。
6. 不掃存。
7. 前後夾紙。
8. 襯紙。
9. 分段做釘（前後有護葉，但中間護葉不掃存）。
10. 包角（依原書掃描）。
11. 單頁掃存（原件雙幅大於 A3）。
12. 封面掃存（書籤，古字，圖畫等）。
13. 解析度 600 dpi 掃存。

在掃描器材的選用上，常用來掃描紙質文獻的設備包括饋紙式掃描器、平台式掃描器與平床式掃描器，通常依據數位化物件尺寸大小、紙張狀況與用途來選擇適合的掃描設備。若是原件為微縮資料則採用微縮膠捲掃描器進行掃描；當原件不宜掃描且適合數位拍攝者，則採用數位拍攝操作流程。各類型掃描器之功能與使用方式略有不同，說明以下：

(1) 饋紙式掃描器

常見的饋紙式掃描器為桌上型的自動進紙機制，一般可掃描尺寸最大到 A3，掃描人員只需將整疊欲掃描的資料放在饋紙槽中，掃描機將自動依序逐張進行掃描。此掃描方式適用於紙質狀況良好之資料，且紙張分離、大小一致。優點為掃描速度快；缺點為檔案必須拆卷分頁才可進行掃描、影像品質較不易控制，且不適合掃描脆弱紙張，如古籍線裝書。

除了桌上型饋紙式掃描器外，亦有大型饋紙式掃描器，供 A3 以上大小之檔案以一比一之比例且不接圖方式掃描儲存。如國史館

臺灣文獻館執行「典藏日據與光復初期史料數位化計畫」時，即以大型饋紙式掃描器來進行大尺寸檔案文書數位化工作。[34]

(2) 平台式掃描器

平台式掃描器主要用於掃描 A3 大小以內的紙張、書籍，掃描人員將原始素材平放在玻璃上由從下面經過的光源擷取影像，每掃一頁即需重新放置資料、操作掃描動作。此掃描方式適用於紙質狀況良好之資料、版面大小一致，掃描書籍時需特別注意不要損傷書籍裝訂部位。優點為機器價格相對較為便宜、體積小，人工每掃描一頁即可確認掃描品質；缺點為較為費時，不適合掃描書況不佳之書籍。

以「臺南大學日治時代日文珍本數位典藏計畫」[35] 為例，依據藏品的保存狀況而定，分別採用低速 A3 規格、低速 A4 規格平台式掃描器及高速度 USB 2.0 規格高階掃描器進行文獻掃描數位化。

(3) 平床式掃描器

平床式掃描器的掃描幅面較大，除了一般紙張尺寸外，亦主要用於掃描大尺寸的資料，如 A1 大小。掃描時資料面朝上，藉高處投射光源從機器上方擷取影像進行掃描動作，不直接接觸原稿；掃描書籍時，換頁亦不需將書本拿起，僅在書籍原位翻頁即可。優點為適用於掃描幅面大、脆弱的紙張書籍，相對於平台掃描節省時間人力；缺點為價格較昂貴。

古籍原書一般為了避免善本書受損，故掃描時大多不拆線，採用特殊平台，使書不必拆線而能平置於掃描器上掃描，如國家圖書

34 溫淳雅、劉華珍，〈國史館臺灣文獻館檔案大尺寸圖檔數位化工作流程簡介〉，拓展台灣數位典藏計畫，檢索：2012 年 2 月，http://content.teldap.tw/index/?p=1118。

35 張鳳吟、辜雅婷、陳美智，〈臺南大學日治時代日文珍本數位典藏計畫〉，拓展台灣數位典藏計畫，檢索：2012 年 2 月，http://content.teldap.tw/index/?p=1034。

館的〈古籍文獻典藏數位化計畫〉執行方式。[36] 再以〈淡新檔案〉數位化掃描來說，考量資料本身較為脆弱且多件資料互相黏連，且相鄰文件資料尺寸大小不一，大多數大於 A3 尺寸，故臺灣大學在進行〈淡新檔案〉掃描作業時採用平床式掃描機進行掃描，避免損壞到原稿，且能大面積完整掃描而不需後續拼貼檔案。然而若文件過長則需分區掃描，並重疊掃描區域以利後續拼貼；若檔案中含浮貼時，須於浮貼蓋著與掀開時分別掃描。

(4) 微縮膠捲掃描器

微縮資料是利用微縮攝影機，把比較大的檔案、文獻、圖書等資料，以數倍甚至數百倍的縮小比例，攝製於微縮軟片，經過化學或物理加工製成微縮品，所用化學藥劑，極微微細。[37]

微縮資料數位化主要是以副本進行掃描，將微縮軟片的片盤安裝在校正過的掃描器上，掃描器的各種參數，依據掃描試驗結果輸入及調整，並以一種連續的模式開始掃描。掃描時依影像排列區分掃描方式，一般影像組織排列方式包括單行橫式、單行直式、雙行單向式、雙行雙向式與複合影像式等多種組合排列方法。掃描時須特別注意的是每幅兩頁以上之軟片，應該被分成單獨的影像檔案。[38] 採用微縮膠捲掃描主要是因為許多早期的典藏文件為微縮捲片，另一方面則是可避免對原始文件再次掃描造成拆裝原件的傷害。

以國立中央圖書館臺灣分館執行之「館藏日文臺灣資料數位典藏計畫」為例，其採用微捲影幅資料為主轉製成數位檔案，而非以原件資料掃描或翻拍，其主要的目的是基於原件的保存與資料的利

36 張瑞芸、洪嘉培，〈國家圖書館古籍文獻典藏數位化計畫〉，拓展台灣數位典藏計畫，檢索：2012 年 2 月，http://content.teldap.tw/index/?p=1117#5。

37 林彥宏、程婉如、張思瑩，《微縮資料數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2011 年 6 月，頁 12。

38 同註 37，頁 36-38。

用。然而，由於微捲無法顯示彩色效果，所以針對彩色（2 色以上）印刷之圖書資料或者微捲拍攝不清的資料，在不拆圖書資料的原則下，圖書館則對其進行原件資料掃描或翻拍。[39]

4. 拍攝

一般文字資料在進行數位化工作時，除了掃描之外大多搭配拍攝來輔助進行。拍攝是運用攝影設備將物體形像紀錄成數位影像，一般立體物件多以拍攝方式進行數位化，文字資料如記載於甲骨、金石等載體上，或是文件本身過於脆弱、檔案尺寸太大而不易以掃描方式數位化者，即透過拍攝方式進行數位化輸入。

前述傅斯年圖書館進行館藏數位化作業，除了掃描之外亦會採以拍攝的數位化方式進行。當文字資料被判斷為適合數位拍攝及不宜掃描者，則進入數位拍攝操作流程。以傅斯年圖書館館藏拍攝標準為例，區分館藏資料為文字次原件和圖像式原件，並規範以不同的解析度儲存，其文字式原件解析度應為 300 dpi，圖像式原件解析度則應達 600 dpi。而不同之文字資料又依類型區分拍攝重點，如單張原件、卷軸類原件及文書類原件之拍攝方式大同小異，各別有需注意的拍攝要項，如表 3-1 所示。而各類型檔案之拍攝方式則依裝訂型式及尺寸大小進行判別，其標準詳如「附錄二、傅斯年圖書館原件拍攝原則」。

39 蘇倫伸，〈日治時期日文臺灣文獻數位典藏計畫概述〉，《臺灣圖書館管理季刊》，第 4 卷第 4 期，2008 年 10 月，頁 75-81。

表 3-1：傅斯年圖書館館藏各類資料拍攝重點標準 [40]

單張原件	1. 全幅拍攝（全幅小於 140*140 公分） 2. 多幅拍攝（全幅大於 140*140 公分） 3. 原件外觀拍攝（正面、背面） 4. 浮貼另拍（揭開） 5. 原件另拍局部一比一 600 dpi 影像
卷軸類原件	1. 全幅拍攝（全幅小於 140*140 公分） 2. 多幅拍攝（全幅大於 140*140 公分） 3. 單幅拍攝 4. 原件外觀拍攝（全卷、半卷、側面） 5. 浮貼另拍（揭開） 6. 原件另拍局部一比一 600 dpi 影像
文書類原件	1. 依原書拍攝（原件雙幅合拍高廣小於 80*90 公分） 2. 依原書拍攝（包角） 3. 單頁拍攝（原件雙幅高廣大於 80*90 公分） 4. 單頁拍攝（中縫過緊） 5. 攤開 6. 原件外觀拍攝（正面、背面、側面） 7. 分段做釘（前後有護葉，但中間護葉不拍攝） 8. 前後夾紙 9. 襯紙 10. 浮貼另拍（揭開） 11. 原件另拍局部一比一 600 dpi 影像

在拍攝人員配合原則方面，傅斯年圖書館之規定如下：

(1) 必須每日進行拍攝工作前，執行數位機背色彩校正程式。

(2) 每兩週必須執行螢幕色彩調校。

(3) 請注意是否有重要紙屑因未粘妥而掉落。

(4) 破損之原件（或人為因素）需修補者，仍請館員處理。

(5) 情況特殊之原件（如：無頁碼原件…），拍攝後交館員處理。

(6) 全件原件拍攝時，立即進行校驗，確認拍攝方式無誤後，再執行批次拍攝。

40 節錄於傅斯年圖書館，〈傅斯年圖書館全彩影像掃描、拍攝及校驗相關作業標準〉一文中內容，傅圖數位典藏計畫網站，檢索：2011 年 11 月，http://lib.ihp.sinica.edu.tw/pages/03-rare/DAP/contentp/03-4-2.pdf。

(7) 配合原件拍攝各項流程操作：

A. 當日拍攝之原件影像暫存於 Mac 電腦，並傳送至 PC 電腦之磁碟陣列。

B. PC 電腦影像校驗無誤後，燒錄 DVD 影像。

C. 當日拍攝之影像需當日校驗，故攝影師需視拍攝情況安排當日之拍攝及校驗時程。

而像是非紙質文獻之文字資料，如甲骨、拓片，也多以拍攝方式進行數位化。例如中央研究院歷史語言研究所、國立故宮博物院、國家圖書館、國立歷史博物館及國立臺灣大學在執行拓片與古文獻之數位化時，皆以數位機背拍攝搭配掃描方式進行。[41] 以中央研究院歷史語言研究所為例，其數位化方式依原件大小進行選擇，若原件小於 A3 則採用掃描方式，若大於 A3 則以數位機背拍攝。而其數位化之影像功能有二，包括：單張影像放大縮小等編輯功能，提供拓片、摹本和甲骨實物的影像；以及影像比對功能，能在同一個畫面做拓片與摹本、拓片與甲骨、正面與背面比對等。其拍攝之標準採用高階的影像處理方式，各種檔案規格如下：

表 3-2：甲骨拓片影像拍攝處理標準 [42]

	解析度	影像類型	檔案類型
典藏級	600 dpi	全彩	TIFF
商務級	300 dpi	全彩	JPG
公共資訊級	72 dpi	全彩	JPG

此外，中央研究院歷史語言所曾進行簡牘文書數位化工作，包括原簡、《居延漢簡—圖版之部》黑白反體照、《居延漢簡補編》黑白照片三個部分，然而因簡牘文書文字褪色，肉眼無法辨識，故原簡影

41 陳秀華、溫敏宇，《金石拓片數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2011 年 6 月。

42 柯維盈，〈歷史語言研究所藏甲骨文拓片資料庫〉，金石拓片數位典藏研討會，檢索：2011 年 11 月，http://rub.ihp.sinica.edu.tw/~oracle/05/01.pdf。

像攝影以紅外線攝影儀設備與數位相機，產出紅外線數位影像與彩色數位影像；《居延漢簡—圖版之部》黑白反體照片與《居延漢簡補編》黑白照片，則以高階平台掃描器與底片掃描器進行數位化工作，設備與規格如表 3-3 所示：

表 3-3：漢簡數位影像規格與器材參考表 [43]

項目	數位化方式	設備	規格
原簡	紅外線影像數位攝影	紅外線影像處理器電腦影像強化及儲存系統	512x480pix BMP 72dpi
	彩色影像數位攝影	數位機背	10,013x8,000pix 3,000dpi TIFF 全彩
《居延漢簡．圖版之部》黑白反體照	照片掃描	高階平台掃描器	600dpi TIFF 全彩
《居延漢簡補編》黑白照片	照片、底片掃描	底片掃描器 高階平台掃描器	1. 以 35mm 底片掃描者 4,000dpi/TIFF/ 灰階 2. 以 120mm 底片掃描者 4,000dpi/TIFF/ 灰階 3. 照片掃描者 1,200dpi/ 灰階

（二）光學文字辨識（OCR）

文字資料的數位化方式，在輸入作業的程序上可選擇人工輸入、手繪文字內容，或者是掃描、攝影等方式數位化。由於文字資料的數位典藏常需處理大量書籍剪報，建置龐雜的文字內容，若以上述的方式輸入資料恐需耗費較多的人力與時間，因此光學文字辨識系統的產生，即有效提升資料輸入的數位化過程效率。

光學文字辨識系統（Optical Character Recognition，簡稱 OCR），主要是對於既有已存的文件（包括印刷文件或手寫於紙上的圖片資訊），透過掃描器或

43 遲恒昌、陳秀慧、洪嘉培、張瑞芸、吳淑鈴、黃如足，〈漢代簡牘數位化工作流程〉，拓展台灣數位典藏計畫，檢索：2012 年 2 月，http://content.teldap.tw/index/?p=1104。

數位相機等光學輸入設備做文字識別，將其分析文字型態的模式演算轉換成電腦可識別的電子訊號。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），可快速將文字資料轉換成電子資料，以供資料庫做檢索的運用。

歐洲是最早使用光學技術來解決文字辨識問題的地區，因為起步較早發展研究者眾，加上歐美國家的拼音文字相對於中文字的辨識度來得簡易，已有不錯的成果。由於中文字的字數多，其字形架構與字形變化也有其複雜度，國內的中文 OCR 對於解決傳統書面資料轉換成電子資料的開發與研究，近幾年才逐步邁入實用階段。目前光學辨識系統技術運用的領域已十分廣泛，舉凡大型圖書館的文獻資料庫或企業內部文件等，一些需要透過數位化方式保存與管理的資訊，皆已利用光學辨識系統技術來增進資料比對與管理。從事文字資料方面的數位典藏單位，也有不少是透過光學辨識系統來提升數位化工作的效率，更能節省大量人力與時間的支出。

一般中文 OCR 辨識的流程包括下列幾項：

1. 影像輸入
   (1) 輸入文件：欲進行 OCR 辨識的資料必須先利用掃描器，將資料文章掃描成圖像格式檔，掃描的解析度越高，越有利於文字的辨識工作。為避免掃描品質不佳而使得黑白文件影像檔中的字元產生破碎或模糊不清，目前的 OCR 辨識系統已能允許彩色或灰階文件的影像輸入，並能利用影像處理技術求得較佳的黑白影像檔，以利提高辨識的準確性。
2. 影像處理
   (1) 清除雜訊：由於輸入文件的表面可能不乾淨，或是掃描器本身掃描時造成失真現象，甚至部分書籍的原文有非文字的符號或注釋標記，皆有可能使輸入的影像存在一些污點或獨立點，造成辨識的困擾。因此在進行文字辨識前，影像檔案應先清除這些雜點、獨立點，產生一個新的清晰圖檔才能進入辨識系統。

(2) 字體修整：掃描器本身造成的失真現象，或解析度太低，而導致掃描後的字體產生不完整的成像，例如：字元不連續、有鋸齒狀或字體有缺角破洞、歪斜等，亦可能造成文字辨識的困難，因此還需進行字體的修整。

(3) 色彩管理：物件本身的文字與底色反差明顯較宜進行 OCR，必要時仍得將全彩的圖檔轉成黑白或灰階，抽離多餘的色彩干擾，將有利於辨識的正確度。

3. 文件分析

(1) 區塊屬性分析：OCR 辨識系統通常只辨認單一字元，因此文件影像都需經過版面文件分析，其主要的分析區分為圖形、表格、文字三種區塊，先進的技術會將文件中所有的圖形、表格和文字分離出來。

(2) 文字分割或合併：版面分析將每一行或段落的文字切出後，仍須將每一文字元切割清楚。辨識文件中的每一字元間距夠明顯，即可提高字元切割或合併的效率與速度。

4. 文件辨識

(1) 辨識核心、萃取特徵：字元切割後，萃取特徵點是系統中最重要也是最困難的技術。辨識引擎以各種表示法或描述法進行辨認，將字元影像與資料庫中每個中文字的影像比對，計算相對位置的顏色是否相同，找出差異最小者為辨識結果。

5. 校正比對

(1) 進行文字比對、校正：當文字被辨識與編碼後，還需執行比對動作，比便找出與辨識字體相符或相近的中文字。中文 OCR 系統需有一中文資料庫，並針對辨識內容特定需求與用途，可內建辭典以供候選字做更正，降低辨識系統誤認的情況。

(2) 前後文相關辭校正：中文字形繁複，相似字形亦繁多，OCR 無法百分之百析出正確的內文字形，難免會有錯字的產生。若統計錯字出現的前後文相關聯字，蒐集整理大量的常錯字字串表，讓 OCR 系統具

備自動學習關聯字的功能，便能以正確字串快速取代常錯的字串，減輕校對的不便。

6. 輸出

(1) 將文字資料辨識結果輸出。部分 OCR 辨識軟體還可指定辨識結果以中文繁體或簡體字輸出，並直接將辨識結果儲存為 WORD、EXCEL、PDF、純文字等格式之檔案。

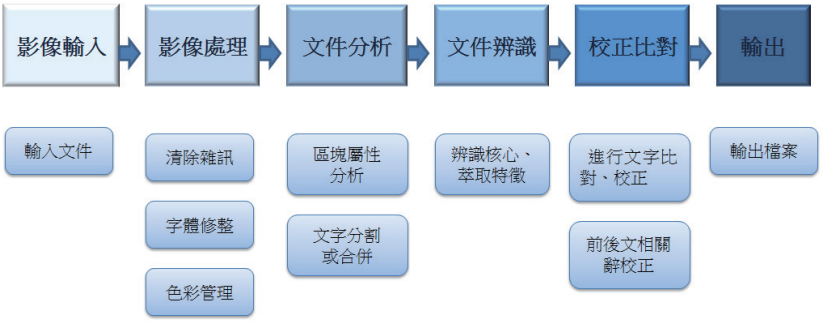


圖 3-4：OCR 辨識流程圖
資料來源：拓展台灣數位典藏計畫重整繪製

透過 OCR 系統技術，除了可解決文件辨識輸入的問題，部分軟體還能保有原書面文件的文字內容、字體大小、顏色甚至圖片、表格及相對位置的電子文件，不但可減輕資料輸入的工作還可提高數位化的效率。目前的主要用途也以建立中文文字資料庫為大宗，包括輸入處理各類報紙、書刊、雜誌等；或是將舊有的書籍、文章以 OCR 辨識輸入至資料庫，可重新管理編排、儲存或檢索等。進階的還能以機器翻譯文章、結合語音輸出等技術。此 OCR 技術未來亦可結合擴展至其他平板電腦裝置或手機等硬體上，可增加資料流通的便利性。[44]

關於各種 OCR 系統目前市面上大致有丹青中英日文文件辨識系統、蒙恬認識王專業系統、全景軟體等，有關這些系統的介紹，本計畫於西元 2009 年出版

44 蒙恬科技，〈光學辨識技術原理概述〉，蒙恬科技網站，檢索：2012 年 1 月，http://www.penpower.com.tw/technology-OCR.asp。

之《期刊報紙數位化工作流程指南》[45] 有詳細的討論與比較，本文不再贅述。

（三）校對

文字資料內容的正確率攸關數位化成果的品質，錯誤率盡可能越低越好，其中資料的校對工作就是重要的一環。校對的方式大致也以人工校對和軟體程式來進行校對，彼此相互配合更可接近原文的電子全文並提升產出的品質。然而，校對工作並非僅針對輸入文字的錯誤與否，部分文字問題仍是因為漢文字的變異與電腦系統缺字引發的各類缺字、異體字、避諱字等因素，這些都是校對過程中需進一步解決處理的。

1. 人工校對

人工校對通常是採逐字逐頁的傳統方式，雖然技術門檻較低但所耗費的人力與時間成本相對較高。負責校對的程序除了輸入者或廠商的第一校以外，計畫人員還需進行二次的校對。

因為人難免會有疲乏的狀況產生，所以仍可能出現誤判的情況。為求資料的準確率，「中華電子佛典協會」曾在進行電子藏經的數位化工作時，採「雙人同工」（是指「同一份工作，由兩個人各做一次」）的方式，將兩人的工作結果再進行比對、形成差異、解決差異，亦是一種更為嚴謹的實務工作。[46]

在執行步驟與步驟之間記得「新檔與舊檔的比對工作」，將這一步驟執行完畢的新檔與上一步驟執行後的前一舊檔相互比對，亦可統計其差異，觀察有無特殊差異的問題，即可再次校正人工錯誤的地方。

2. 軟體程式校對

目前的校對工作多採人工校對，但是文字資料龐大或人力有限時，軟體校對即可節省許多作業時間與人工校對的疏漏。

---

45 李佩瑛、程婉如，《期刊報紙數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2009 年 04 月，頁 33-44。

46 吳寶原，〈創意靈感一關於電子藏經的輸入、校對及編輯〉，《佛教圖書館館訊》，第 24 期，2000 年 12 月，檢索：2012 年 02 月，http://www.gaya.org.tw/journal/m24/24-main2.htm。

(1) 檔案比對：「檔案比對」顧名思義是指利用程式對同一份文件有兩個或兩個以上的不同檔案版本（例如人工輸入和 OCR 光學文字辨識兩種輸入方式不同的檔案）相互比對出差異。由兩個檔案匯入程式，並依其字形、文字編碼等對應出差異的地方來進行訂正。雖然檔案比對也並非百分百的方式，但相較於同一份文件多輸入一至兩次，節省人工校對逐字尋錯的時間，確實是比傳統作業方式更能達到文字的正確性。

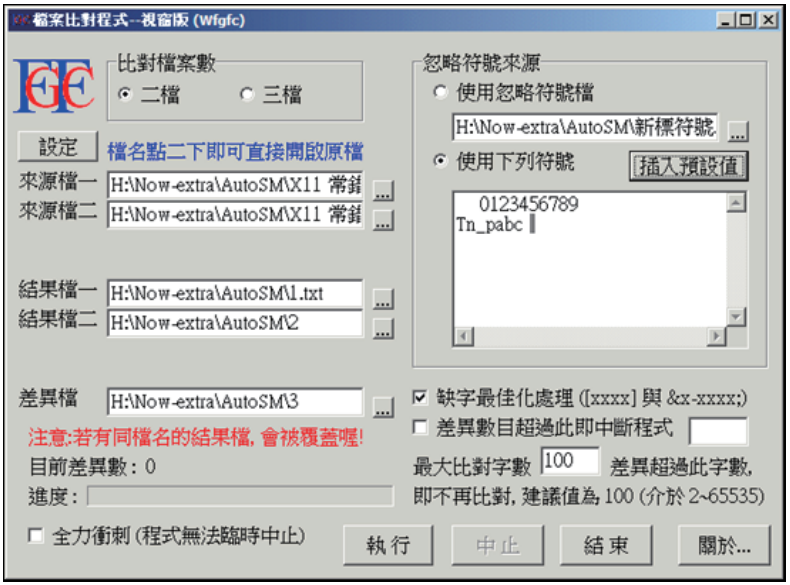


圖 3-5：中華電子佛典協會檔案比對程式畫面 [47]

47 王雅萍、謝筱琳，《漢籍全文數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2011 年 6 月，頁 41。

(2) 看圖比對：將原書掃描成圖檔，透過原書圖檔的結構分析，輸入檔每行予以定位，把這些特殊符號的位置算出來，讓文字檔在相對應的地方形成差異，再以看圖校對的方式來辦定差異。[48]

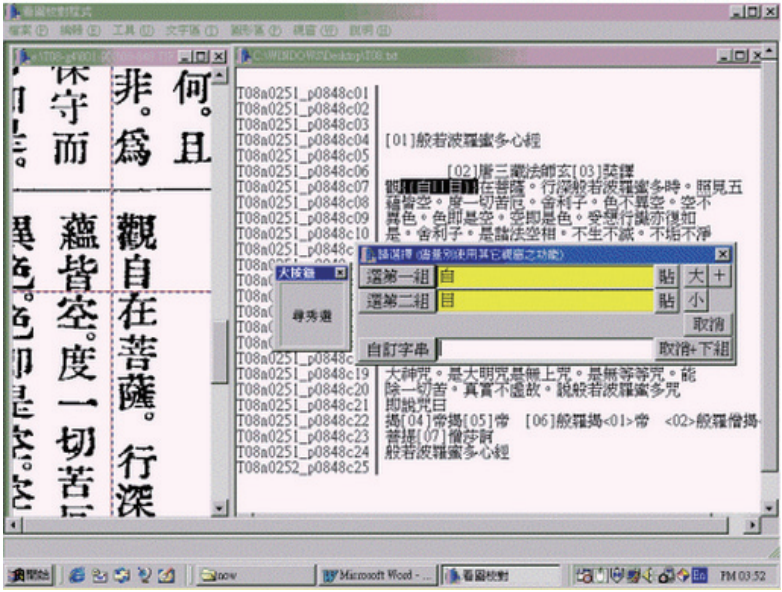


圖 3-6：看圖校對程式[49]

(3) OCR 校對工具：目前光學文字辨識技術 OCR 已是大宗全文輸入的選擇之一。如果強化 OCR 之後的校對機制，其效率將可優於人工的校對作業。以中央研究院資訊所設計的 OCR 校對工具為例，引用張復老師研發的 OCR 核心程式，以 Microsoft Visual C# User Control 作為 OCR 校對工具的發展語言。[50] 使用 User Control 設計介面，能讓影像與文字

48 同註 47，頁 39-40。

49 同註 46，吳寶原著。

50 陳信文，〈以 Microsoft Visual C# UserControl 實作 OCR 校對工具〉，中央研究院計算中心通訊電子報，第 11 期，檢索：2012 年 2 月，http://newsletter.ascc.sinica.edu.tw/news/read_news.php?nid=1878。

分行並列，對於區塊設定、標誌、橫直向設定、上下標等字型設定更為便捷；OCR 文本辨識系統的校對也更為快速，降低校對後的錯誤率。

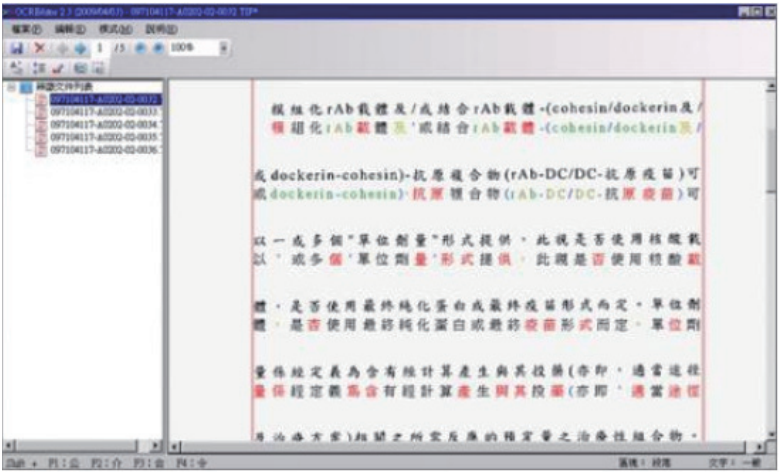


圖 3-7： OCR 校對工具 [51]

(4) 其他：除了上述的校對程式外，中華電子佛典協會也曾針對不同的案例研發適合的校對檢查程式。例如「行首資訊」（比對經號、頁、欄、行有無問題）、「對稱符號」（如 []、() 等符號有沒有不對稱情形）、「標記」（如段落標記是否遺漏）等程式，都是可快速得到可能的錯誤報告。在龐雜的文字大海裡，是取代人工檢查作業提升效率的選擇。

（四）數位化規格

文字資料之型態，大多為印刷品或手稿，其數位化大多是透過掃描及拍攝的方式，將原件內容儲存下來以利後續應用。另有一些文字資料為電子檔，除了一份原始檔外，亦建議轉成 PDF 、RTF 或 ODF 等檔案格式典藏利用。數位化單位有的會自行制定數位化規格標準，或參考其他單位的建議規格，以確保數

51 同註 50。

位化產出成果在各種不同用途下的品質。

因應現今數位化技術之成熟，「數位典藏與數位學習國家型科技計畫」修訂了相關的數位典藏數位化規格，以供未來執行計畫的依據與技術發展之參考規範，確保數位化成果之品質。此規格分為典藏保存與網路瀏覽兩個等級，前者是以數位化方式保存藏品物件的形象以供研究或加值之用，因此在數位化時必須達到典藏保存等級的最低數位化規格要求，日後在運用時可以此最高品質的數位化檔案進行降階轉檔處理，以確保整體數位典藏品質；後者則是供使用於網路媒體的非商業性公開釋出版本，典藏單位需達到其最低釋出規格之要求，可針對其需求製作不同尺寸品質之影像內容供預覽，以滿足國內對數位化內容的需要。

就文物數位化規格中，與文字資料相關的規格如下：

1. 文字資料

(1) 原始資料為電子檔：

若原始資料是以電腦打字的電子檔，除儲存一份原始檔外，建議轉成 PDF、RTF 或 ODF 等的檔案格式供長期保存之用。

(2) 原始資料為印刷品或手稿：

原始資料為手稿或印刷品資料者，其數位化規格如下：

表 3-4：文字資料數位規格

數位檔用途	典藏保存	網路瀏覽
說明	將資料數位化典藏保存其原有風貌	提供非商業性使用者網路觀看、預覽或列印之用
檔案格式	TIFF 6.0	JPEG、PNG
壓縮方式	非破壞性壓縮或不壓縮	不限制
色彩模式	RGB（24bit / pixel 以上）	RGB（24bit / pixel 以上）
解析度及尺寸	解析度：至少 300dpi 以上（依原始資料品質及重要性選擇適當的解析度，一般印刷品建議採 300dpi） 尺寸：短邊至少 1280 pixel 以上	影像尺寸：短邊至少 1024 pixel 以上

2. 影像資料

文字資料有時亦利用拍照方式進行數位化，其輸出之影像品質規格與上述文字資料規格同樣依目的區分等級標準。如原始資料為照片、正負片、幻燈片、圖片、地圖等，或是以數位化設備直接拍攝成影像者，其數位化規格如下：

表 3-5：影像資料數位規格

數位檔用途	典藏保存	網路瀏覽
說明	將資料數位化典藏保存其原有風貌	提供非商業性使用者網路觀看、預覽或列印之用
檔案格式	TIFF 6.0	JPEG、PNG
壓縮方式	非破壞性壓縮或不壓縮	不限制
色彩模式	RGB（24bit / pixel 以上）	RGB（24bit / pixel 以上）
解析度及尺寸	解析度： 至少 600dpi 以上（依原始資料品質及重要性選擇適當的解析度，照片正負片與幻燈片建議至少須 2400dpi） 影像尺寸： 短邊至少 2048 pixel 以上	影像尺寸： 短邊至少 1024 pixel 以上

以上資料來源：「數位典藏與數位學習國家型科技計畫—數位典藏文物數位化規格版本 2.0（2012 年 3 月）」[52]

52 本規格主要參考資料來自英國及美國數位典藏計畫的數位化規格，參考網址如下：

The National Archives. *Guidance*. Retrieved June 1, 2012, from http://www.nationalarchives.gov.uk/information-management/projects-and-work/guidance.htm

The National Archives and Records Administration. *Reformatting approaches*. Retrieved June 1, 2012, from http://www.archives.gov/preservation/products/definitions/reformatting.html

三、文字處理

文字資料數位化存取於電腦之後，為了能在電腦環境中正確地輸入、顯示與交換文字，需透過一套電腦可讀的編碼標準來進行文字處理。在中文字中，由於字型結構較複雜且字型眾多，常遇到缺字與異體字問題，故需透過相關缺字解決方案來處理文字。此外，為進一步展示作品的屬性與內涵，可透過標記（markup）的方式將文件資訊與文件內容的重點標示清楚；另可透過斷詞技術進行中文詞彙的擷取並建立索引詞，以利後續資訊檢索之利用。本小節將依文字資料數位化處理之相關議題進行介紹，包括中文編碼標準、缺字與異體字的處理，以及斷詞和標記等文字處理方法。

（一）中文編碼標準

中文編碼所指的是依特定的規則，將漢字以一組符號或文字編入位元組中，使電腦能夠處理及顯示漢字。中文字為表意文字，相對於字母拼音文字，採用兩個以上位元組來描繪出編碼空間的大字集。早期臺灣與香港地區多用 Big5 碼進行漢字編碼，後因 Big5 字碼量不足，故使用國際發展的 Unicode 碼，而其他各個國家亦有各自常用的編碼系統。然而由於每個編碼所包含的中文字數不同，編碼方式也不相同，且大部分都沒有標準規格，為了能解決不同國家間電腦字元資訊交換的困難，故發展出全球共用的交換碼標準，如 ISO10646、CNS11643。

文字資料藏品數位化轉入電腦後，需透過編碼進行全文電子檔之處理。在數位典藏工作的執行上，最後的步驟便是將成果輸出於網路上供社會大眾參考利用，在網頁資訊的呈現上即需透過字碼來處理文字的呈現。以下分別介紹常用之中文編碼標準：

1. Big5　大五碼

Big5 碼是由資策會於 1984 年策劃制定，擁有 13,053 個中文字、408 個符號及 33 個控制字元的字集，是我國早期中文電腦的業界標準，也是中文社群最常用的電腦漢字字集標準。而後隨著電腦擴充需要，業界各

中文系統廠商推出了不同版本的 Big5 碼，為統一標準，經濟部標準檢驗局在 2003 年委託財團法人中文數位化技術推廣基金會修訂 Big5 編碼字元表，重整為 Big5-2003 版本。

2. GB2312　信息交換用漢字編碼字符集基本集

GB2312 是中國大陸地區廣用的漢字編碼標準，全稱為《信息交換用漢字編碼字符集基本集》，由中國國家標準總局 1981 年發布實施，因應簡體文獻的處理需求。然而由於 GB2312 應用於處理歷史文獻上仍有字碼不足之問題，故中國國家標準局 1995 年重新修訂編碼，制定了編碼擴充的 GBK 標準，能夠用來同時表示正體字和簡體字。

3. Unicode　國際通用碼

由於各個國家有各自常用的編碼標準，每個編碼所包含的中文字數不同，編碼方式也不相同，為了解決不同國家間電腦字元資訊交換的困難，國際標準組織與國際電工聯盟合組的第一聯合技術委員會下的 SC2/WG2 工作小組（ISO/IEC JTC1/SC2/WG2）提出的 ISO 10646 標準草案；國際 Unicode 組織也設計適用全球的廣用碼，即 Universal Code，簡稱 Unicode，中文稱為統一碼或萬國碼，是目前使用最廣泛的跨國際字碼標準。ISO 10646 及 Unicode 主要是提供全球語言文字與符號之表示、傳送、交換、處理、儲存、輸入和顯示的共同編碼標準，在 Unicode 標準下以 UTF-8 與 UTF-16 兩種方式來定義電腦存取 Unicode 編碼的轉換格式，以便容納更多字元，解決不同電腦字元資訊交換困難的問題。Unicode 最新版本已發展至 6.1 版，有關 Unicode 之介紹與發展可參考 Unicode 官方網站：http://www.unicode.org/。

4. CCCII　中文資訊交換碼

中文資訊交換碼（Chinese Character Code for Information Interchange，簡稱 CCCII），緣起於 1979 年美國急需使用電腦處理東亞語文資料，故在加州史坦佛大學召開了一個籌劃東亞圖書館自動化的會議，希望訂定中文交換碼標準作為自動化之根據。我國召集文字學家、圖書館學家及電

腦學者進行研究並提出中文資訊交換碼，並陸續擴充編碼字集，且為了方便電腦上的文字處理，編製了「中國文字資料庫」（Chinese Character Database，簡稱 CCDB），列出每個字屬性如部首、筆畫、讀音以及各種對應和輸入碼。此標準現主要用於國內外圖書館系統居多。

中文編碼標準眾多，但由於 Unicode 容納了世界各種語言的字元和符號，字集較完整且收納的中文字遠多於 Big5，並支援支援台語白話字之呈現，故在國內數位典藏工作中，除因特殊需求，則多採用 Unicode 方式編碼。然而數位典藏所處理之文字資料除了一般繁體字之外，尚包括古文字與異體字等不同字體，儘管 Unicode 已納入超過十萬個字元，但仍無法完全避免缺字問題，故尚需透過缺字解決方案輔助進行文字處理。以下分別探討數位典藏中常遇到之缺字與異體字問題。

（二）缺字

台灣早期的全文數位化計畫除了採用上述 BIG5（大五碼）外，也逐漸改用字集更為龐大的 Unicode（標準萬國碼）。但不管是以哪一種編碼作業都仍有些字的字形無法呈現，造成「缺字問題」的現象出現，尤其以古籍文字資料特別顯著。普遍的解決方式都是沿用「缺字即造字」的原則，但是新增的造字可能無法共享，也難以管理缺字，甚至造成檢索文件時面臨異體字檢索等難題。對於數位典藏系統之缺字需求，通常需能夠銜接古今文字，例如楷書、小篆、繁體、簡體等同一個漢字，卻在不同時間、空間的相異字型；典藏系統可以檢索、著錄查詢的缺字；甚至不需安裝軟體即可於網頁查看的系統。基於上述的缺字問題，以下簡要介紹幾個漢字缺字的處理方案：

1. 全字庫 [53]

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

53 同註 30，全字庫。

全字庫」（簡稱全字庫）網站，其主要目的除了當初為建設我國的中文電腦應用環境外，也為解決個人電腦中文字數不足問題、解決自造字交換問題，並維護各機關單位之間造字後的「同字同碼」原則，有效管理整合所有造字，及網頁上的造字顯示等問題。

此系統的主要功能，包括：中文碼查詢、字型下載、中文碼轉換、共用（相同）造字集安裝、機關企業團體自造字集整合及管理、網頁上自造字顯示、內部網路複製全字庫、安裝 BIG-5E 字集、轉碼匣門、新增字申請等服務。

2. 漢字構形資料庫

中央研究院資訊科學研究所之「文獻處理實驗室」建立了「漢字構形資料庫」（2008 年起更名為「中央研究院漢字部件檢字系統」），其架構是遵循漢字構形原理，將漢字拆解為最小單位的基礎部件，再由部件拼湊而成，利用部件來表達數以萬計的漢字，大大減少交換碼不足之問題。亦整理出各代文字的部件資料庫，分析所有漢字在文字學上的合理組成，提供小篆、楷書、異體字之相對應。

構字式亦即字形結構表達的方式，例如「謝」字的構字式為「言射」、「霜」為「雨  相」、「圓」為「口員」。構字式主要是對於漢字字形結構包含有漢字、部件、字根、連結符號、構字規則和得以拆解完成之漢字構字式，利用有限的部件及字根的組合方式來表達任一漢字，定義了數個「構字符號」。[54]

數位典藏計畫單位建置典藏系統時，因內容主題不同對於缺字的問題，需求欄位也有所差異，例如動物學類的昆蟲分類、昆蟲中文學名；書畫典藏的印記、釋文、題跋內容；金石拓片的青銅器資料、人名、器名；善本古籍中的書名、題記、全文、釋文等一些罕見用字或古字。「文獻

54 數位典藏與數位學習國家型科技計畫，〈中文缺字技術〉，數位典藏與數位學習計畫百科，檢索：2012 年 2 月，http://goo.gl/tkOzT。

處理實驗室」進而與各個典藏單位加以合作整合，以中央研究院「傅斯年圖書館善本古籍數位典藏系統」為例，[55] 由於該圖書館收藏了大量的古籍，其遭遇的古字缺字問題也不少，其系統亦建有「缺字查詢」的功能，提供使用者輸入構字式關鍵字，即可查詢有關於該關鍵字的缺字。目前主要提供字形查詢（單一或多個部件查詢，如圖 3-8）、構字式複製（如圖 3-9）、製作缺字圖形（如圖 3-10），[56] 及自定大小、顏色、字體，等相關功能。只要在善本典藏系統中輸入該缺字的構字式，當含有構字式的資料被著錄時，缺字系統即會自動判別並自動將構字式的字進行對應轉換。若不清楚某個缺字的構字式，也只需要使用缺字構字式查詢工具亦可解決此一問題。

圖 3-8：輸入單一或多個部件做字形查詢

圖 3-9：構字式複製

55 缺字系統，檢索：2012 年 2 月，http://char.iis.sinica.edu.tw/index.htm。
56 同註 55，缺字系統。

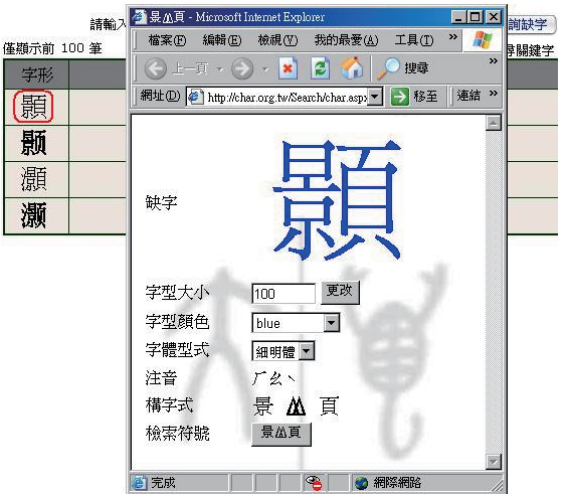


圖 3-10：製作缺字圖形（自訂大小、顏色、字體）

缺字的處理解決了數位典藏系統資料的著錄、顯示與查詢等缺字問題，也無須擔心交換碼空間不足造成缺字數量的限制等。使用者無須另外加裝軟體的便利性，可透過網路直接瀏覽查詢等，也是這些大量文字資料能流通與整合的重要意義所在。

（三）異體字

所謂的異體字，是指讀音、意義與正體字相同，但寫法、字形不同的兩個字，例如「體」的異體字有「体」、「骵」、「骵」、「躰」……等字。在使用時，異體字是可以彼此替代的，如「體育館」與「体育館」的意思是一樣的。

漢字的總數龐大，歷代皆有一些新增字，有相當數量都是異體字，主要原因不外乎是個人書寫漢字因隨意性而產生的變異，或前代不同形制的漢字積澱到後代而產生的差異。漢字存在著大量的異體字，例如簡化字只是其中的一種，這些異體字也是缺字問題得處理的狀況之一。[57]

57 莊德明，〈漢字數位化的困境及因應：談如何建立漢字構形資料庫〉，文獻處理實驗室，檢索：2012 年 2 月，http://cdp.sinica.edu.tw/service/documents/T960507.pdf。

因 Unicode 字碼表中這些異體字的字碼並不相同，處理這些異體字時也常造成中文資訊處理一定程度的混淆。異體字的形成時間、使用的地域和原因也不盡相同，例如有些古字在古籍中「然」、「燃」可能是相通的兩個字，但現今使用的意義已截然不同。異體字的使用與上下文內容有關，因此異體字處理亦顯得相對困難。

目前異體字關係的相關資料主要出處包括：教育部異體字典、康熙字典、漢語大字典、簡化字總表、Unihan 資料庫、兩萬漢字中日韓越英俄讀音釋義字典等。而上文提及的「漢字構形資料庫」，針對異體字所產生的缺字現象亦是研究的重點，其收錄不同歷史時期的異體字表，記錄不同時期的漢字結構，並使用構字式來解決古今漢字的編碼問題等。例如「員」字本為方圓的「圓」的本字。下表為「員」的歷代古文字形體。（如圖 3-11 ～圖 3-14）[58]

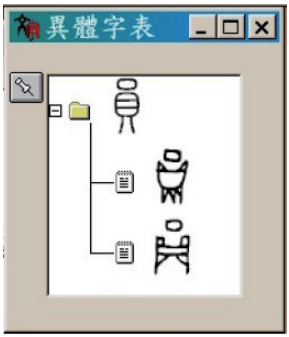


圖 3-11：甲骨文異體字表（員）

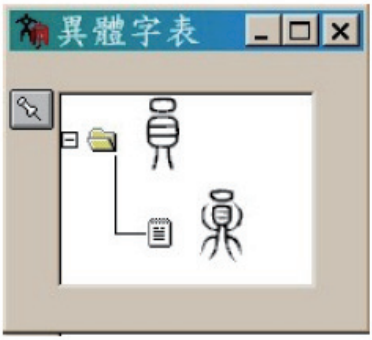


圖 3-12：小篆異體字表（員）

58 同註 57。

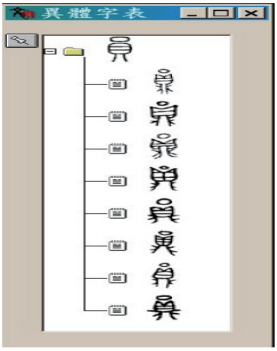


圖 3-13：金文異體字表（員）

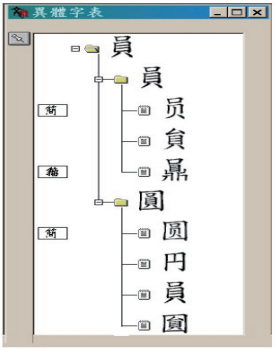


圖 3-14：楷書異體字表（員）

此外，「國際漢字電腦及異體字知識庫資料庫」（http://chardb.iis.sinica.edu.tw/）亦設計了多種異體字查詢介面，分別為單字、部首、部件、相似字、編碼查詢等。

圖 3-15：國際漢字電腦及異體字知識庫

就文字資料數位化有豐富經驗的中央研究院「漢籍工作室」為例，其建置的漢籍電子全文工作，對於異體字的處理辦法大致為：

1. 若遇意義、用法相當的其正體字的純異體字，則輸入的時候以其正體字表示。
2. 若判讀上下文仍無法確認異體字之義與用法是否合於正字，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」。為了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。若無法判讀異寫字的用意，則同樣保留原書字形。

（四）標記（Markup）

標記指的是為了突顯文字中某部分訊息而做的記號，在數位化過程中，為了在電腦上有效地表達文本結構，故透過文字和數字定義出標記的規則，使電腦能依照設定的需求來處理資訊。標記可分為程序性標記（Procedural Markup）與描述性標記（Descriptive Markup），程序性標記是針對文件的呈現外觀進行標示，而描述性標記是針對文件的內容和語義結構進行標示。

國際標準組織（ISO）於 1986 年制定「標準通用標記語言」（Standard Generalized Markup Language，簡稱 SGML），定義了一組標記的規則，可用來描述文件結構以便電子文件能在不同系統間交換資訊，並適用於任何複雜的資訊處理。然而由於 SGML 功能複雜、開發成本高、不易在網路環境中使用，故隨之出現「超文本標記語言」（Hyper Text Markup Language，簡稱 HTML）的應用。HTML 是一種製作網頁的標準資料格式，使用固定的標籤來定義資訊內容結構，著重版面編排與外觀格式，簡單易用故廣泛被接受。然而也由於其標籤集固定造成結構上的限制，且在資料交換上難以對每一項所要交換的資料作清楚的描述，於是促使「可延伸標記語言」（eXtensible Markup Language，簡稱 XML）的

誕生，以補其不足之處。XML 是全球資訊網聯盟（World Wide Web Consortium，簡稱 W3C）在 1996 年底所提出的 SGML 簡化格式，能描述各種複雜的文件結構及內容語意，且 XML 沒有定義任何固定的標籤，而是提供一個架構，讓使用者自行定義標籤，故較 HTML 更具彈性與延伸性。然而也由於 XML 的這些特性，故在數位典藏中廣泛被使用於文件資料結構之描述。

除了上述常用之標記語言，在文字資料方面，常用「文件編碼組織後設資料標誌標準」（Text Encoding Initiative，簡稱 TEI）[59] 來進行數位文本的標記。TEI 概念出現於 1987 年，並不停與時俱進，至 2007 年已發表至第五版。TEI 能對文本本身做任何面向的編碼，如章節結構、正文、頁碼、醒目字句、附註、參照與連結…等資訊的標記，讓電腦可以讀得懂文件內容。為方便使用者進行標記，TEI 協會研發出一套 Roma 工具（http://www.tei-c.org/Roma/），使用者能透過 Roma 依需求來製作 TEI 文件模型，使用方式可參考《TEI 使用指南》一書。[60]

國外已有許多使用 TEI 之實例，如維吉尼亞大學圖書館（University of Virginia Library）已使用 TEI 為館藏編碼；紐西蘭威靈頓維多利亞大學（Victoria University of Wellington）成立的紐西蘭電子文本中心（New Zealand Electronic Text Centre）也使用 TEI 為數位化的歷史文本進行全文編碼；[61] 美國國會圖書館（Library of Congress）所執行的著名的 American Memory 計畫，亦是基於 TEI 的原則建立 American Memory DTD[62]（AMMEM.DTD），透過 TEI 標誌 American Memory 中豐富的美國歷史與文化的數位化資料，以便大眾使用這些全文內容。[63]

---

59 TEI 網站，檢索：2012 年 2 月，http://www.tei-c.org/index.xml。

60 魯 · 伯納、麥克 · 蘇寶麥昆、馬德偉 著，謝筱琳、黃韋寧 譯，《TEI 使用指南—運用 TEI 處理中文文獻》，台北市：數位典藏拓展台灣數位典藏計畫，2009 年 4 月。

61 Victoria University of Wellington. *About The New Zealand Electronic Text Centre.* Retrieved February 20, 2012, from http://www.nzetc.org/tm/scholarly/tei-NZETC-About.html

62 DTD（Document Type Definition）為文件描述類型，概念緣於 SGML，功能在於定義文件所包含的元素（Element）以及每個元素的內容與屬性。

63 Library of Congress. (2009). *American Memory DTD for Historical Documents.* RetrievedFebruary 20, 2012, from http://lcweb2.loc.gov/ammem/amdtd.html

在國內，TEI 在佛典數位典藏中已經有相當成熟的應用。以中華電子佛典協會執行之「佛典數位典藏內容開發之研究與建構」計畫為例，該計畫使用 XML 做為佛典電子檔的標記語言，並採用 TEI 做為基礎標籤集，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[64]。除了以上這些使用實例，TEI 組織整理了一份使用 TEI 之計畫清單，有興趣者可至 TEI 網頁（http://www.tei-c.org/Activities/Projects/）瀏覽，另外有關 TEI 後設資料之處理則將於本書第肆章做進一步的說明。

（五）斷詞

在中文中，詞是具有最小意義的語言單位，當文字數位化成全文電子檔案進入資訊系統中，詞便成為全文檢索的關鍵。由於中文的語法結構與英文不同，詞與詞之間沒有間隔，電腦系統在進行資訊擷取時需要透過斷詞技術來辨別全文中的詞彙，才能進一步進行資訊處理。

中央研究院資訊所與語言所共同指導的詞庫小組在建構「現代漢語平衡語料庫」時即開始研擬分詞的規範，並在上百萬的語料分析中整理出分詞標準的細節規定。其分詞規範於 1999 年經由經濟部標準局通過為〈中文分詞處理原則〉（編號 CNS14366），並成為中央研究院詞庫小組進行詞類分析、定義及確定之工作依據。[65, 66]〈中文分詞處理原則〉語意和語法兩方面來規範分詞的依據以及分詞單位的定義，可做為詞庫收詞之標準與分詞設計上的參考，其基本原則與輔助原則如下：

64 吳寶原、謝筱琳，〈TEI- 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〉，拓展台灣數位典藏計畫，檢索：2012 年 2 月，http://content.teldap.tw/index/?p=1096。

65 賴佳旻、盧秋蓉、邱智銘，〈現代漢語平衡語料庫數位化工作流程簡介〉，拓展台灣數位典藏計畫，檢索：2012 年 2 月，http://content.teldap.tw/index/?p=1116。

66 經濟部標準局，〈中文分詞處理原則〉，國家標準（CNS）檢索系統，檢索：2012 年 2 月，http://goo.gl/LqlTe。

表 3-6：中文分詞處理原則

基本原則	
1. 語意無法由組合成分直接相加而得到之字串合併原則應該合為一分詞單位	合併原則
2. 詞類無法由組合成分直接得到，應該合為一合併原則分詞單位	合併原則
輔助原則	
1. 有明顯分隔標記應該切分之	合併原則
2. 附著語素盡量和前後詞合為一個分詞單位	合併原則
3. 使用頻率高或共現率高的字串盡量視為一個分詞單位	切分原則
4. 雙音節結構之偏正式動詞盡量視為一個分詞單位	合併原則
5. 雙音節加單音節之偏正式名詞盡量視為一個分詞單位	合併原則
6. 內部結構複雜之詞盡量切分之	切分原則

舉例來說，基本原則中兩條合併原則指的是若組合後語意改變者皆應視為一個分詞單位，如「飛黃騰達（成語）」、「十二萬分（定量結構）」、「辛辛苦苦（重疊結構：程度加強）」、「中山南北路」…等；或是詞類無法由組合成分直接得到者應該合為一分詞單位，如「好喝」、「很棒」。其他合併和切分原則之說明實例請參考〈中文分詞處理原則〉。在台語斷詞方面，雖台語語言特性與中文略有不同，但同樣適用中文分詞處理的原則。例如「仔」為台語中極為常用的詞綴，遇到「桌仔」、「椅仔」之類的詞彙可遵循「附著語素盡量和前後詞合為一個分詞單位」之原則來處理。有關台語斷詞之討論與範例說明，可參考「台語斷詞原則討論」（http://iug.csie.dahan.edu.tw/TG/CompLing/hunsu/hunsu.htm）。

系統自動分詞大多是利用詞庫中收錄的詞和文本做比對，找出可能包含的詞，但由於中文語言的特性使得不同字元組合會產生不同詞義，故造成分詞岐義問題，也影響分詞切分結果。除了分詞岐義的問題之外，因中文詞集是一個開放的集合，詞庫無法涵蓋所有的詞彙；再者各種領域存在一些特殊用詞與專有名詞，也不時產生新的詞彙，詞庫無法全面收錄這些詞，而這些未知詞也是造成斷詞困難的原因之一。

進行未知詞與關鍵詞的抽取辨識能增加詞彙的搜集並加強斷詞成效，中央研究院資訊科學研究所詞庫小組開發的「中文斷詞系統」（http://ckipsvr.iis.sinica.edu.tw/）即具備自動抽取新詞建立領域用詞與線上即時分詞的功能。該系統是國內在中文斷詞方面較成熟的系統，其詞彙庫包含約十萬個詞彙及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料；在基本詞彙庫外，使用者可依需要附加領域專屬詞庫，也可透過詞類標記功能，附加文本中切分詞的詞類，解決詞類歧義並猜測新詞之詞類。[67]

中文斷詞系統處理未知詞的步驟依序為：初步斷詞、未知詞偵測、中國人名擷取、歐美譯名擷取、複合詞擷取、由下而上合併演算（bottom-up merging algorithm），並在最後進行重新斷詞。初步斷詞是透過長詞優先演算法取出未知詞的詞素[68]，再透過未知詞偵測去判定初步斷詞後的單字哪些是詞素以及哪些是獨用詞彙，再從詞素中合併出未知詞。此外，針對特定類型如中國人名、歐美譯名、複合詞等進行詞構分析，最後運用由下而上合併演算做剩餘未知詞的擷取動作，並將擷取出的詞搭配原始的辭典再做一次重新斷詞以得到最後的結果。[69] 學術研究以及非營利目的之使用皆可向中研院申請中文斷詞系統進行應用。

67 中文斷詞系統，檢索：2012 年 2 月，http://ckipsvr.iis.sinica.edu.tw/。

68 詞素，即構成詞的最小成分，且在意義上不能再分析的單位。

69 馬偉雲，〈未知詞擷取作法〉，中文斷詞系統，檢索：2012 年 2 月，http://ckipsvr.iis.sinica.edu.tw/。

肆、後設資料規劃與資料庫系統建置

Planning for Metadata and Database Construction

後設資料，譯自「metadata」，即 data about data，為描述資料的資料。後設資料以結構化的方式展現資料的內涵，將不同資料透過一致的標準結合整理，協助電子資源之定義、描述與指示資料之位址，促進資料的控制與交換。對管理者來說，後設資料是管理數位資料控制的機制，使數位典藏品的共享與互通；對使用者而言，後設資料則幫助其進行資源之找尋、辨識、選擇與獲取。後設資料之規劃是數位典藏工作中重要的環節，並與資料庫設計息息相關。本章節將先簡述建置後設資料之目的，再就後設資料欄位之規劃、著錄，及後設資料與資料庫檢索系統發展說明之。

一、後設資料規劃

本書第參章提到「數位化物件挑選」是數位化工作流程前置作業中重要的程序，其工作流程包括對資料的瞭解、清查與製作清單和相關表格、訂定數位化規格、作業標準等項目，此一步驟即是建置後設資料的一部分，透過資料盤點建立典藏品的完整資料清冊，有助於之後數位化工作之進行與資料管理。

建置後設資料一方面可以進行藏品徹底地整理、描述與登錄，另一方面可以協助發展資料庫系統。後設資料著錄典藏品各方面的資訊，將有利於典藏單位未來對實體物件進行查核，或是進行原件與複製品之間的比對，在文字資料數位典藏方面，也有助於進行文獻版本學之研究。後設資料可說是數位檔案與實體物件之間連接的橋樑，完整的後設資料著錄，有助於建置高回收率（recall）與高精確率（precision）的資料庫檢索系統，再與全文影像檔案進行連結後，更能直接地向社會大眾顯示這些珍貴資料，提高文物的可及性，發揮數位典藏的價值。[70]

依後設資料之重要性可歸納建置後設資料之目的包括：增進原始資料數位化管理的容易度、提供藏品實體資料資訊、說明數位化後的相關資料（如授權範圍）、使資料易於處理並提高資料檢索的效益。

---

70 洪淑芬，《文獻典藏數位化的實務與技術》，台北市：數位典藏國家型科技計畫 訓練推廣分項計畫，2004 年 2 月，頁 55-56。

（一）後設資料需求評估

後設資料的規劃需視藏品特性與典藏單位之需求而定，在進行後設資料規劃時，通常需先依數位化藏品類型進行需求確認，再建立後設資料欄位。

數位典藏國家型科技計畫與中央研究院機構計畫共同成立了「數位典藏計畫後設資料工作組」（Metadata Architecture and Application Team，簡稱 MAAT），協助數位典藏計畫在後設資料方面的推動與規劃，並設計「後設資料生命週期作業模式」（如圖 4-1），提供建置後設資料時的參考。

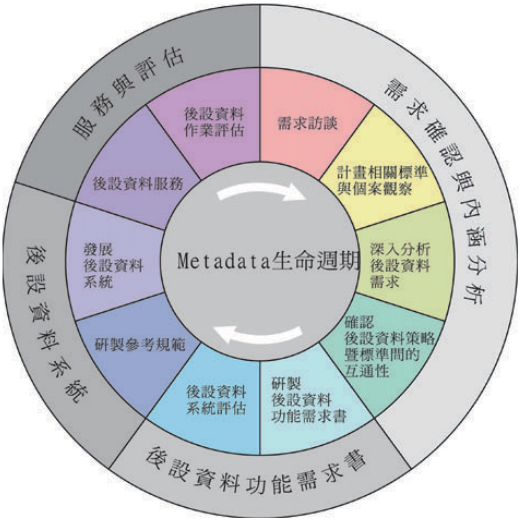


圖 4-1：後設資料生命週期作業模式 [71]

根據「後設資料生命週期作業模式」，後設資料的建置可分成四大階段，包括「需求確認與內涵分析」、「研製後設資料功能需求書」、「發展後設資料系統」，以及「後設資料服務與評估」。

71 後設資料生命週期作業模式（Metadata Lifecycle Model），簡稱為 MLM。檢索：2012 年 4 月，http://metadata.teldap.tw/design/lifecycle_new2.htm。

本節參考後設資料生命週期，依數位典藏工作流程過程，先就「後設資料需求評估」與「後設資料欄位建立與著錄」進行說明，下一節再介紹「後設資料系統之發展」，即資料庫系統之建置。

在數位典藏進行後設資料的規劃時，首先需要進行需求確認與內涵分析，瞭解藏品資料的屬性，擬定描述藏品特性時應包括的元素，同時分析相關類型藏品之應用個案，做為採用其應用標準實作的評估，以便進行較完整的藏品識別與描述。

由於藏品特性不同，許多數位典藏單位會依藏品設定專屬的後設資料著錄規範。為輔助進行後設資料之規劃，後設資料工作小組彙整數位典藏的實務經驗，設計 12 種「後設資料作業表單」[72]，供執行數位典藏之單位填寫，以協助該單位規劃適用之後設資料。

「後設資料工作表單」之功能，在於瞭解不同後設資料類目的範圍、關聯、關係性及屬性，不只可向數位化單位提出功能需求之彙整，亦可提供系統開發人員快速建置系統的相關資訊，為許多執行數位典藏之單位所利用。中央研究院臺灣史研究所在進行楊雲萍文書之數位典藏工作時，即運用「後設資料工作表單」製作〈楊雲萍文書後設資料功能需求書〉；在數位化日治時期古籍資料時，也同樣依此後設資料工作表單建立〈日治時期臺灣研究古籍後設資料功能需求書〉，以利數位化工作之執行。此外，舉凡中央研究院歷史語言研究所、近代史研究所與國史館，也皆參考此後設資料工作表單，輔助進行後設資料之結構、欄位之規劃。

以下介紹需求表單包括之項目與內容，並以文字資料檔案所填寫之需求表單實例輔助說明：

1. Metadata 需求確認表單：包括計畫名稱與目標、數位化藏品資料類型與數量、藏品數位化規格需求…等，是對數位化工作內容的基本說明。其表單格式如下：

72 數位典藏與數位學習國家型科技計畫，〈後設資料作業表單與填寫範例〉，後設資料工作組，檢索：2012 年 4 月，http://metadata.teldap.tw/design/worksheet/worksheet.htm。

表 4-1：Metadata 需求確認表單 [73]

主題計畫名稱（中 / 英文）：	計畫主持人： （姓名、電話、電子郵件）
主題計畫所屬單位：	
計畫說明：	
計畫目標：	
參與 Metadata 作業的主題計畫同仁：（姓名、電話、電子郵件）	
Q1	是否希望將 Metadata 需求納入計畫內？ ☐是　　☐否（原因：_____）
Q2	是否希望中研院「Metadata 工作組」協助規劃與分析典藏品 Metadata ？ ☐是（請至 Q3）☐否（原因：_____）
Q3	系統目標
Q4	系統範圍
Q5	預計時程與期限？
Q6	預期的需求重點與成果？（GIS system ？ Exhibition Management……）
Q7	藏品資料類型與數量（若不同資料類型需不同的 Metadata 著錄格式，請註明，並標明計畫進行的優先順序）
Q8	目前典藏品目錄資訊的狀態？ 8-1 使用的系統名稱、建檔數量、資料庫系統相關文件（規格書，Schema…等） 8-2 是否使用任何 Metadata 標準？ 8-3 是否有任何既有之著錄表單？（如：文物基本資料表、登記表…等 8-4 是否有公開網站，可供查詢使用？ 8-5 是否有其他相關參考資訊（如：其他單位網站、著錄規範…等）
Q9	藏品數位化規格需求？（解析度、檔案類型、開放程度…等） 其他說明與建議：

填表者：_____　　填表日期：_____

73 數位典藏國家型科技計畫，〈Metadata 需求確認表單〉，數位典藏技術彙編 2007 年版，檢索：2012 年 4 月，http://www2.ndap.org.tw/eBook08/showContent.php?PK=6。

以〈楊雲萍文書後設資料功能需求書〉[74] 為例，其「Metadata 需求確認表單」詳細說明該計畫執行內容、數位典藏系統目標、範圍與預期成效，以供後設資料規劃者參考。其填寫內容如下：

表 4-2：Metadata 需求確認表單

主題計畫名稱（中／英文）： 臺灣省諮議會及中研院臺史所史料典藏數位化計畫	計畫主持人
主題計畫所屬單位：中央研究院臺灣史研究所	
計畫說明：楊雲萍先生本名楊友濂，因以「士林雲萍生」一名於《臺灣民報》發表文章，自此遂以「雲萍」之名著於世，因而本藏以「雲萍文書」為名。雲萍先生為日據時期臺灣文學舉足輕重之名家，開臺灣白話文學之先河，其後更跨足於史學研究，為文史雙棲之哲人。本藏數量約 1,000 多件，主要為雲萍先生往來信件，藏品年代約始於昭和 4 年，直到民國 70 年左右。所藏之往來信件尤以日治時期，與文友西川滿、金關丈夫、立石鐵臣和林獻堂等私人信函，甚為珍貴，得以窺見日治背景之下文學家結社之實況；此外，與《臺灣時報》、奉公會、總督府等信函，更提供了當時文人之處境與時代背景等資訊。因此，本藏對於臺灣文學史或日治時期臺灣史，是十分重要之一手資料。本計畫將所藏之書信數位化，建置資料庫，並期與雲萍先生年表結合，希望能將文書典藏與時代歷史相結合。	
計畫目標：建置數位典藏資料庫，不只能達到數位典藏保存文物資料的目標，並且能提供日後加值應用，提供研究、教學及文化推廣之目標。	
Q1、是否希望將 Metadata 需求納入計畫內？　是。	
Q2、是否希望中研院「Metadata 工作組」協助規劃與分析典藏品 Metadata ？　是。	
Q3、系統目標： 1. 著錄系統需具有新增、查詢、修改、刪除、複製等維護功能。 2. 需與人名權威、機關團體權威連結。	
Q4、系統範圍： 就計畫執行人員而言，此系統需具有資料的著錄建置、維護等基本功能，針對使用者，滿足其檢索、調閱影像全文的需要。整體而論，「雲萍文書數位典藏系統」與各權威檔所提供的資訊，互相連結、補充。	
Q5、預計時程與期限？ 期望在 93 年 6 月完成 Metadata 分析，並完成後設資料需求書在 12 月有系統雛形可進行著錄測試。	
Q6、預期的需求重點與成果？（GIS system ？ Exhibition Management…） 1. 著錄系統與檢索系統 2. 影像管理系統 3. 權威檔建置	
Q7、藏品資料類型與數量：文書資料（約 1500 件左右），其資料類型有書信、明信片、電報、照片、圖書報刊……等數量：琺瑯器共 1000 件、銅器 700 件、瓷器 850 件…	

74 數位典藏國家型科技計畫，〈楊雲萍文書後設資料系統功能需求表單〉，數位典藏技術彙編 2007 年版，檢索：2012 年 4 月，http://www2.ndap.org.tw/eBook08/showContent.php?PK=9。

Q8、目前典藏品目錄資訊的狀態？ 8-1. 使用的系統：系統名稱、建檔數量、資料庫系統相關文件 （規格書，Schema…等） 無 8-2. 是否使用任何 Metadata 標準？ 參考 Dublin Core 8-3. 是否有任何既有之著錄表單？（如：文物基本資料表、登記表…等） 建置中（Excel 表單） 8-4. 是否有公開網站，可供查詢使用？無 8-5. 是否有其他相關參考資訊（如：其他單位網站、著錄規範…等） 國內：臺史所古文書 Metadata 需求書、臺大伊能手稿 Metadata 欄位、真理大學馬偕與牛津學堂 國外：Their Own Words http://deila.dickinson.edu/theirownwords/ American Journey http://www.americanjourneys.org/index.asp Trails of Hope http://overlandtrails.lib.byu.edu/
Q9、藏品數位化規格需求？（解析度、檔案類型、開放程度…等） 圖檔解析度分為： 文字：使用全彩（bits/pixel 或以上）光學解析度 300dpi 掃描存檔，以 TIFF 及 JPEG 格式存檔。 照片：使用全彩（bits/pixel 或以上）光學解析度 600dpi 掃描存檔，以 TIFF 及 JPEG 格式存檔。
其他說明與建議：無

2. 藏品單元（unit）層級關係圖：以圖示說明藏品單元（unit）間的層級關係，以檔案為例，其層級關係為全宗、副全宗、系列、副系列、卷……。以〈外交經濟重要檔案數位典藏計畫 經濟重要檔案後設資料需求規格書〉為例，其藏品單元層級關係圖依檔案層級從最大藏品單位至最小藏品單位繪如下圖所示：[75]

75 數位典藏國家型科技計畫，〈檔案類：外交經濟重要檔案數位典藏計畫—經濟重要檔案後設資料需求規格書〉，數位典藏技術彙編 2007 年版，檢索：2012 年 4 月，http://www2.ndap.org.tw/eBook08/showContent.php?PK=231。

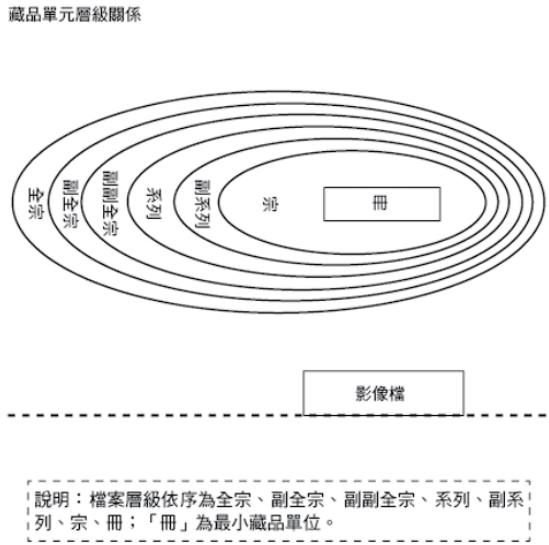


圖 4-2：外交經濟重要檔案藏品單元層級關係圖與說明

再以〈國家歷史資料庫 戰後臺灣的初期發展（1945-1954）後設資料功能需求書〉為例，其藏品單元層級關係圖也是依檔案層級來繪製，並呈現檔案內容與區分類別，全宗以整體計畫區分、副全宗以時代區分、系列以史事專題區分、副系列以不同來源資料區分、卷以案卷或書籍名稱區分、並以單件為最後一層級。此藏品單元層級關係圖使「國家歷史資料庫」架構一目瞭然，如下圖所示：[76]

76 數位典藏國家型科技計畫，〈檔案類：國家歷史資料庫一戰後臺灣的初期發展（1945-1954）後設資料功能需求書〉，數位典藏技術彙編 2007 年版，檢索：2012 年 4 月，http://www2.ndap.org.tw/eBook08/showContent.php?PK=235。

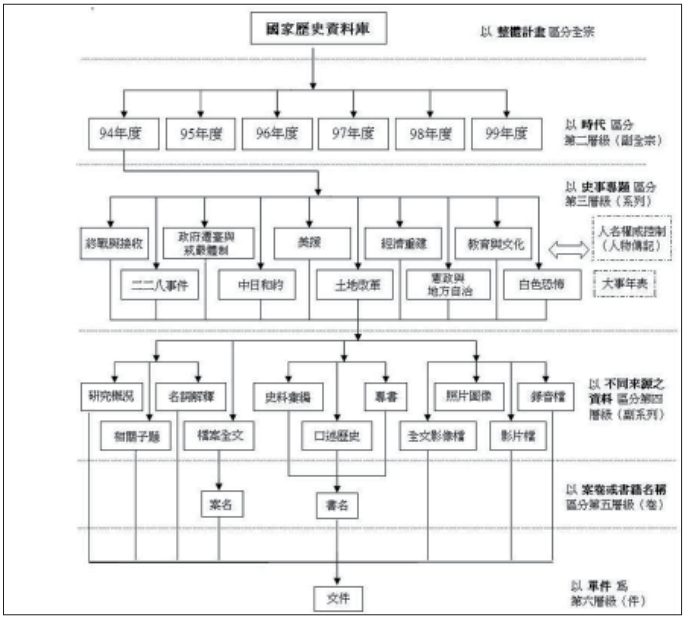


圖 4-3：國家歷史資料庫藏品單元層級關係圖與說明

3. 藏品單元（unit）群組關係圖：以圖示說明藏品單元（unit）間著錄之群組關係，如：兩個以上之藏品單元（unit）所組成之有意義的藏品組件、套件或附件等關係，並標明是否須著錄為一筆後設資料紀錄。例如一本書之全文分章節掃描成許多「件」，則可將這些「件」群組。
4. Metadata 藏品元素需求表單：分別說明藏品所需之中文與英文元素（即欄位），並以文字描述各欄位所代表之意義與著錄規範，如日期格式規範為 yyyy/mm/dd。
5. Metadata 元素代碼表單：為藏品各元素制訂代碼，可在建檔時以代碼作為詞彙控制之用，例如文字資料藏品之保存狀況，可設代碼包括：現況良好、破損、送修、鏽蝕…等，或是使用權限可設代碼為開放、不開放、經同意後開放。
6. Metadata 著錄範例表單：依據上述所列之層次、欄位，分別填上實例紀錄。

以〈臺灣省行政長官公署檔案後設資料需求規格書〉[77]為例，該計畫依全宗、系列／副系列／宗、卷／件層次建立後設資料需求欄位表，內容包括項目中英文名稱，並列出系統屬性功能需求，如：資料型態[78]、欄位所需之大小空間、欄位屬性、提供者…等項目。其全宗需求欄位總表與著錄範例如下：[79]

表 4-3：臺灣省行政長官公署檔案全宗需求欄位總表

項目名稱		英文名稱		資料型態	大小	必填	多值	屬性	提供者
類型		Type		Varchar	4	*		固定值：檔案	系統
機關代碼		Institution Number		Varchar	6	*		預設值：th	填表者
全宗號		Record Group Number		Varchar	3	*		固定值：003	填表者
全宗名		Record Group Name		Varchar	18	*		預設值：臺灣省行政長官公署	填表者
典藏資訊	典藏地	Collection	Location	Varchar	16	*		預設值：國史館臺灣文獻館	填表者
	典藏位置		Stack Area	Varchar	18	*		預設值：文獻大樓戰後檔案室	填表者
入藏資訊	來源	Acquisition	Acquisition	Varchar	12	*		預設值：臺灣省政府	填表者
	取得方式		Resource	Varchar	6	*		預設值：移轉	填表者
	入藏時間		Acquisition Date	Varchar	10	*		預設值：19990611	填表者
編目紀錄	登錄者	Cataloging Records	Cataloger Name	Varchar	8	*		系統自動產生	系統
	修改者		Modifier Name	Varchar	8			系統自動產生	系統
	建檔日期		Cataloging Date	Varchar	10	*		系統自動產生	系統
	修改日期		Modifier Date	Varchar	10			系統自動產生	系統

77 數位典藏國家型科技計畫，〈檔案類：臺灣省行政長官公署檔案後設資料需求規格書〉，數位典藏技術彙編 2007 年版，檢索：2012 年 4 月，http://www2.ndap.org.tw/eBook08/showContent.php?PK=75。

78 資料型態中，Int 表示存放有效數字型態的整數資料，Varchar、Text 是存放純文字型態的資料。

79 臺灣省行政長官公署檔案之系列／副系列／卷／件需求欄位及著錄範例，請見〈臺灣省行政長官公署檔案後設資料需求規格書〉。檢索同註 77，數位典藏國家型科技計畫。

表 4-4：臺灣省行政長官公署檔案全宗層次著錄範例

項目名稱		著錄範例
類型		檔案
機關代碼		th
全宗號		003
全宗名		臺灣省行政長官公署
典藏資訊	典藏地	國史館臺灣文獻館
	典藏位置	文獻大樓戰後檔案室
入藏資訊	來源	臺灣省政府
	取得方式	移轉
	入藏時間	20000331
編目紀錄	登錄者	系統自動產生
	修改者	系統自動產生
	建檔日期	20021101
	修改日期	20021104

7. 元素關係結構圖：以圖表表示元素間之結構性（如：欄位間的層級關係）、對外資料庫的連結（如：與地理資訊系統之連結）…等。
8. Metadata 系統屬性功能需求表單：填寫於系統中顯示之欄位類別、欄位順序、資料型態、欄位最大字元數、必填欄位、多值欄位、功能連結與預設值等。較為常用的資料型態如下：[80]

80 數位典藏國家型科技計畫，〈後設資料需求表單填表說明與範例〉，數位典藏技術彙編 2007 年版，檢索：2012 年 4 月，http://www2.ndap.org.tw/eBook08/showContent.php?PK=7。

表 4-5：Metadata 系統屬性功能常用的資料型態

類別	資料型態	說明
數值欄位	INT	存放整數型態的數值資料，範圍：-2147683648 至 2147483648（SIGN）或 0 至 4294967295（UNSIGN）。例如：25、747……
	FLOAT	存放浮點數型能的數值資料，例如：8.64、23.9361 ……
字串欄位	VARCHAR	存放不定長度且不超過 255 個字的字元串資料。例如：陳菁豐、中央研究院……
	TEXT	存放不定長度且不超過 65535 個字的字元串資料。例如：「 國民政府檔案內容涵蓋國民政府時期所經歷的內憂外患及其推行重大建設的珍貴史料，共計……略……。 在中華民國國民政府時期歷經北伐戰役、寧漢分裂、中原大戰、五次剿共、西安事變和國共內戰，以及……略……。」
日期時間欄位	DATE	以”YYYY-MM-DD”格式存放日期資料，如：2003-12-25。
	TIME	以”HH:MM:SS”格式存放時間資料，如：13:25:24
	DATETIME	以”YYYY-MM-DD HH:MM:SS”格式存放日期時間資料，如：2003-12-25 13:25:24
	YEAR	以”YYYY”格式存放年份資料，如：1998。

9. Metadata 系統查詢功能需求表單：設定查詢顯示之欄位名稱與檢索方式，可包括簡要查詢、進階查詢、簡要顯示款目、檢索結果欄位、詳細顯示款目、付費使用之項目…等。
10. Metadata 系統紀錄建檔流程：以圖表或文字表示計畫實務上的紀錄建檔流程，即著錄後設資料之工作要點。
11. Metadata 系統主鍵元素架構格式：主鍵元素（primary key）為具有唯一性之資料值，主要是作為紀錄辨識與紀錄連接之用。此部分可用圖表或文字來表示主鍵元素之架構格式。
12. 系統使用群組與其使用功能表：此表主要目的為調查系統使用者之大略群組區分與其擁有權限，以方便日後系統分析與開發參考使用。可依單位需求將群組分為：系統管理人員、研究人員、研究助理…等。

（二）後設資料欄位建立與著錄

一般進行後設資料之規劃時，會參考既有之標準，若符合需求，則直接使用之。在文字資料數位典藏中，常用的後設資料標準包括整合性的「都柏林詮釋資料核心集（Dublin Core，簡稱 DC）」、檔案館常用的「檔案編碼描述格式（Encoded Archival Description，簡稱 EAD）」、圖書館的機讀格式（Machine Readable Cataloguing Record，簡稱 MARC），以及「文件編碼組織後設資料標誌標準」（Text Encoding Initiative，簡稱 TEI）…等標準。以下將簡述這幾種文字資料常用之後設資料標準。

1. 都柏林詮釋資料核心集（Dublin Core, DC）

都柏林詮釋資料核心集（以下簡稱 DC）是一套跨領域的資訊資源描述標準，可以提昇資源在跨領域、跨主題的可見度，同時由於其易用、易懂之特性，因此具有廣泛的使用者。DC 可以對資源做一般性的描述，也可進一步深入的描述，以提供語意較豐富的描述服務。DC 的內容分為 15 個欄位，DC 的每一個欄位都是非必備且可重複，欄位無先後順序之分，且可視使用者需求，重複多次相同欄位。由於其著錄方式簡單、容易擴展、可適用於眾多領域，並可直接處理網路資源，為數位資源資訊組織帶來很大的便利性。

在數位典藏中，DC 也是最為廣泛使用之後設資料標準，「數位典藏與數位學習聯合目錄」為了統合各學術領域資源，採用 DC 欄位定義，將後設資料歸納成以下主要欄位結構，這 15 個欄位即是做為聯合目錄跨領域整合資料的核心欄位，不論哪一類型的數位典藏藏品皆可適用，使用者亦可不受拘束的運用這些欄位做組合查詢，以達到更準確的資料檢索。[81]

81 數位典藏與數位學習國家型科技計畫，〈通往數位典藏豐碩藏品的窗口—目錄導覽〉，數位典藏與數位學習成果入口網，檢索：2012 年 4 月，http://digitalarchives.tw/about.jsp。

(1) 標題（Title）：給予資源的名稱
(2) 著作者（Creator）：編輯資源內容的主要負責人
(3) 主題 / 關鍵字（Subject & Keywords）：資源內容的標題
(4) 描述（Description）：資源內容的解釋
(5) 出版者（Publisher）：使資源能廣泛的使用者
(6) 貢獻者（Contributor）：對於資源內容形成貢獻者
(7) 日期（Date)：資源週期的事件日期
(8) 資料類型（Resource Type）：資源內容的性質或類型
(9) 格式（Format)：關於資源的實際或是數位的形式
(10) 資料識別（Resource Identifier）：可以明確的指示出該資源
(11) 來源（Source）：敘述目前資源的參考來源
(12) 語言（Language）：資源所使用的語言
(13) 關連（Relation）：說明相關的資源
(14) 範圍（Coverage）：資源內容的廣度或範圍
(15) 管理權（Rights Management）：描述資源權利相關的資訊

2. 檔案編碼描述格式（Encoded Archival Description, EAD）

檔案描述格式（Encoded Archival Description，以下簡稱 EAD）是一種結構化的檔案檢索工具，其發展之目的是為了支援檔案和手稿的收集保存，提供一個永久編碼標準，以利檔案資源在網路上容易取得，EAD 是以機讀方式展現檔案描述，標籤文件類型定義（Tag DTD）是根據 SGML DTD 而發展，架構為階層式結構，且 EAD 元素與屬性的結構方式是依使用者的需求而訂，各元素與屬性的使用與層級並沒有一定的限制，能詳實呈現檔案和圖書館的目錄系統。

EAD 是目前檔案界最常使用的詮釋資料標準，包含以下 3 大結構[82]：

82 王麗蕉，〈檔案描述標準 MARC AMC 與 EAD 之對映〉，《圖書與資訊學刊》，第 51 期，2004 年 11 月，頁 112-113。

(1) EAD 標目（Eadheader）：用來辨識查檢工具，可產生電子形式或印刷形式的題名頁，其中又包含 <eadid> 唯一性的辨識號碼或代碼、<filedes> 查檢工具的書目性資訊、<profiledesc> 記錄查檢工具的使用語文以及文件編碼創造者的資訊、與 <revisiondesc> 修訂的描述等。

(2) 前面事項（Frontmatter）：提供題名頁與其他正式出版有關的資料描述，作為檢索之用，此項目為選擇性，非必備項目。

(3) 檔案描述（Archdes）：此大項為必備項，是檔案單元主要結構項目的描述，包含下列幾項：

- 辨識描述 <did>：包含描述檔案單元時所需的資訊，例如單元名稱、日期、範圍長度等。
- 附屬資料描述 <add>：作為協助檔案資料的使用，而非檔案本身的描述，例如書目、檔案計畫、索引等資訊的描述。
- 行政資訊 <admininfo>：提供有關機構採訪、管理的資訊，包含查檢限制、採訪資訊、其他可使用形式、典藏歷史、使用限制等。
- 編排 <arrangement>：描述檔案編排的依據與情形。
- 傳記與歷史 <bioghist>：提供傳記資料與檔案歷史等資訊。
- 檢索控制 <controlaccess>：作為控制檢索資訊之描述，例如團體名稱、家族名稱、功能、類型、地理名稱、個人名稱、主題、題名等檢索資訊之描述。
- 附註 <note>。
- 其他描述資料 <odd>。
- 組織 <organization>。
- 範圍與內容 <scopecontent>。
- 附屬成分描述 <dsc>：提供對檔案單元進一步的描述，深入分析到文件系列、案卷、甚至是項目的描述，以 <c01> 表示第一層分析，<c02> 為第二層描述，以此分層類推，可細分至第 12 層的附屬成分描述，描述的項目包含辨識資料、行政資訊、檢索控制、範圍與內

容等前述檔案單元資訊。

在文字資料數位典藏中，文書檔案大多皆採用 EAD 為後設資料著錄標準，主要是考量 EAD 為針對檔案資料結構描述而發展的標準具階層性的架構，且獲多項國外大型圖書館及檔案計畫採用。前述提到的楊雲萍文書數位典藏、外交經濟重要檔案數位典藏、國家歷史資料庫—戰後臺灣的初期發展（1945-1954）數位典藏、臺灣省行政長官公署檔案數位典藏…等計畫，當後設資料工作組為其規劃後設資料時，皆是以 EAD 為基準，並將 EAD 之元素與計畫後設資料需求製作成分析比對表，能看出各層及使用狀況，提供主題計畫參考。有關文書檔案數位典藏計畫 EAD 標準實際應用實例，請參考〈數位典藏技術彙編 2007 年版〉檔案類後設資料需求規格書：http://www2.ndap.org.tw/eBook08/showContent.php。

3. 機讀編目格式標準（Machine Readable Cataloguing Record, MARC）

機讀編目格式標準又稱為機讀格式（以下簡稱 MARC），為電腦系統可以閱讀及處理編目紀錄的格式。其最初的發展目的是作為圖書館間書目資料的交換標準，現在已成為圖書館界用來為各種資料進行編目的標準，MARC 依國情發展出不同格式，目前國內常用的為「中國機讀編目格式」（CMARC）與 MARC21。

MARC 的組成要素包括「紀錄結構」、「內容標識符號」與「資料內容」，以三位數欄號為標示，並以分欄識別、分欄代號協助電腦辨識與處理編目資料。MARC 欄位從 0_ _ 至 8_ _ 各有不同的代表意義：0 _ _ 識別段、1_ _ 代碼資料段、2 _ _ 著錄段、3 _ _ 附註段、4 _ _ 連接款目段、5 _ _ 相關題名段、6_ _ 主題分析段、7 _ _ 著者段、8_ _ 國際使用段。舉例來說，題名與作者資訊會著錄於「200 題名與著者敘述項」、出版資訊著錄於「210 為出版項」、檔案頁數與大小著錄於「215 為稽核項」…等，若有電子全文連結則著錄於「856 電子資源位址及取得方法」。有關 MARC 各欄位之著錄標準，可參考國家圖書館整理的相關規範：http://

catweb.ncl.edu.tw/portal_e3_cnt.php?button_num=e3&folder_id=25。

由於 MARC 主要用於書目紀錄之整理，故成為文字資料數位典藏常用的格式之一，特別是常用於處理善本古籍等具有圖書型式之檔案。國立台中圖書館在將其館藏日文舊籍數位化時即直接採用「中國機讀編目格式」為後設資料標準，中央研究院歷史語言研究所傅斯年圖書館在將善本古籍數位化時，也將其後設資料需求對照為 DC、MARC、MARC21 等格式。[83]

4. 文件編碼組織後設資料標誌標準（Text Encoding Initiative, TEI）

文件編碼組織後設資料標誌標準（以下簡稱 TEI）是一種數位文本的標記語言，它是一個開放、集體發展的標誌標準，適用於所有時期、任何語言的文本，也是文字資料常用的後設資料標準之一。TEI 允許任何種類的文本資訊，例如：出處、結構、異讀、省略、註腳、索引、表格等，也包含編輯者本身加入的解釋、註解及修訂資訊，甚至包括記錄某種資訊確定度的可能性。其擁有超過 500 個元素、屬性、集及模組，應屬人文學科中包含最廣的標誌標準。[84] 現在流通的版本大致有 TEI P4、TEI P5 以及 TEI Lite 等版本，皆可於官方網站（Text Encoding Initiative，http://www.tei-c.org/）瀏覽之。

TEI 的文件結構可分成兩大部分：包括標頭（以元素 <TEI 標頭 >（<teiHeader>）標誌）和文件（以元素 < 文件 >（<text>）標誌）。TEI 標頭部分提供相當於印刷文件題名頁所提供的資訊，如機讀書目的敘述文字、文件編碼敘述的方式、文件的非書目性敘述（文件背景資訊）、以及修訂歷史。文件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。例如，每一份 TEI 文件有相同的基本架構：

83 數位典藏國家型科技計畫，〈善本類：善本圖籍後設資料需求規格書〉，數位典藏技術彙編 2007 年版，檢索：2012 年 4 月，http://www2.ndap.org.tw/eBook08/showContent.php?PK=78。

84 同註 60，《TEI 使用指南—運用 TEI 處理中文文獻》，頁 16-17。

```
<TEI>
    <teiHeader>…</teiHeader>
    <text>……</text>
</TEI>
```

兩個主要的段落與 HTML 中的 <head> 和 <body> 相似。<teiHeader> 包含有關建立數位文件的後設資料，例如它的來源、編輯者、版權、修訂歷程等， 而 <text> 則包含文件內容本身。

中華電子佛典協會（CBETA）的佛典數位化工作，多年來即是採用 TEI 的標記語言為數位化工具之一。CBETA 電子佛典主要目的，就是要利用資訊科技的易於保存、複製、傳播、及再製等電子媒體的便利性，期能保存紙本中的資訊。如《大正藏》的編排格式，經名、譯者、作者、品名、揭頌、附文、校勘等資料，必須在電子檔裡用「標記」的方式記錄《大正藏》中的各種資訊。其中有關標頭相關的標記就有如下數種，都是電子檔書目資訊描述（尚有其他多種類型的資訊描述）：[85]

表 4-6：標頭相關標記

標籤名稱	英文說明	中文說明
< fileDesc >	file description	檔案描述
< titleStmt >	title statement	標題陳述
< respStmt >	statement of responsibility	關於負責人的陳述
< resp >		關於負責性質的描述片語，例如某人「編譯」
< editionStmt >	edition statement	版本陳述
< edition >	Edition	版本
< publicationStmt >	publication statement	出版發行陳述
< distributor >		出版者
< availability >		使用限制、版權聲明
< sourceDesc >	source description	本電子檔所依據的來源說明

85 周邦信，〈標記語言的應用〉，《佛教圖書館館訊》，第 24 期，2000 年 12 月。

範例：

表 4-7：標記範例

```
< fileDesc >
  < titleStmt >
    < title >Taisho Tripitaka, Electronic version, No. 001 長阿含經 < /title >
    < respStmt >
      < resp >Electronic Version by< /resp >
      < name >CBETA< /name >
    < /respStmt >
  < /titleStmt >
  < editionStmt >
    < edition >Version 1.0 (Big5)< date >1999/12/10< /date >< /edition >
  < /editionStmt >
  < publicationStmt >
    < distributor >
      < name > 中華電子佛典協會 (CBETA)< /name >
      < address >
        < addrLine >cbeta@ccbs.ntu.edu.tw< /addrLine >
      < /address >
    < /distributor >
    < availability >
      < p >Available for non-commercial use
      when distributed with this header intact.< /p >
    < /availability >
    < date >< /date >
  < /publicationStmt >
  < sourceDesc >
    < bibl >Taisho Tripitaka Vol. 1, No. 001 &desc;< /bibl >
  < /sourceDesc >
< /fileDesc >
```

二、資料庫系統建置與檢索

建立後設資料的另一目的，即是為了建置一個有效的資料查詢系統。透過上述有關後設資料的規劃評估及欄位建立、著錄等介紹，尚須利用這些相關資料延伸建置資料庫，使其提高檢索的效益。據數位化工作流程簡圖的概念，承接上述的工作後，其中一環即是「資料保存」階段。為確保每一數位資料能繼續被取用的「長久性」，其資料保存工作已被視為數位化工作中重要的課題。所有即將進入資料庫存放資料的地方，也就必須有一定的格式標準、儲存檔案等系統之因應對策。因此資料保存與資料庫管理系統亦即相輔相成的環節步驟。

（一）資料庫建置

資料庫的系統建置完善，與資料庫檢索功能的效益是相互影響的關係。檢索刊載的資訊豐富，對於使用者必然是助益良多，除了取決於後設資料的欄位和建入欄位內的資料外，資料庫的檢索功能與資料呈現等方面一開始即需詳盡的規劃。

1. 資料庫的基本概念[86]

在本章節開始之前，我們先來了解資料庫設計的基礎：資料與資訊。資料（data）是指未經過處理無法顯示它們意義的原始事實，例如使用 Google 文件進行網路問卷設計與調查，在後端表格中所顯示的可能是由一些數值所組成（圖 4-4），你無法一下子就能理解這些數值所代表的意思，但為了讓這些原始資料所內含的意義呈顯出來，就必須經過資料處理，而就成為資訊（information），例如將問卷表單資料轉換為圖形，這些統計圖表便能解釋問卷調查的結果，也就是可以呈現意義的資訊（圖 4-5）。

86 Peter Rob、Carlos Coronel 著，張世敏 譯，《資料庫系統：設計、實作與管理》，台北市：新加坡商聖智學習，2009 年 1 月，頁 4-11。

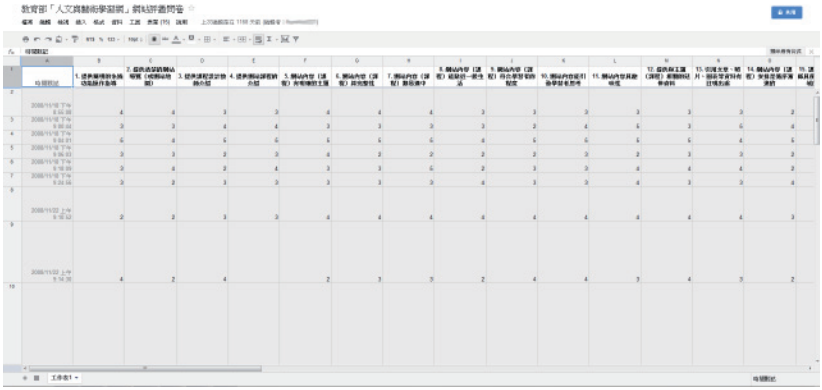


圖 4-4：Google 文件裡的表單

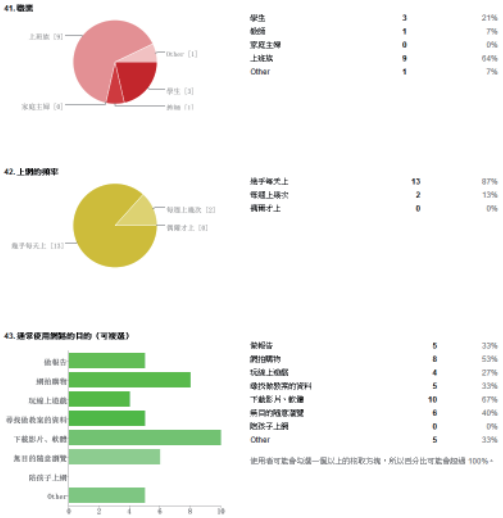


圖 4-5：Google 表單的統計分析圖

要有正確的資料，就需以容易存取與處理的檔案格式儲存，有了正確的資料，才能提供正確有用的資訊。也因此如何建構一個適當的資料管理模式，便是在資料或資訊處理上很重要的課題。

大量的文字資料轉換為電子格式後，為能便於資料的流通使用與管理，便會透過資料庫管理系統，連結使用者與資料之間的互動關係。一般資料庫儲存的內容包括使用者有興趣的原始事實的「最終使用者資料」，以及用來詮釋內容與進行資料整合與管理的「後設資料」。而資料庫管理系統可以看作是橋樑，它是一組程式，來管理資料庫的結構，以及控制使用者利用應用程式所下達的指令，以進行資料庫中內容的存取（圖 4-6）。

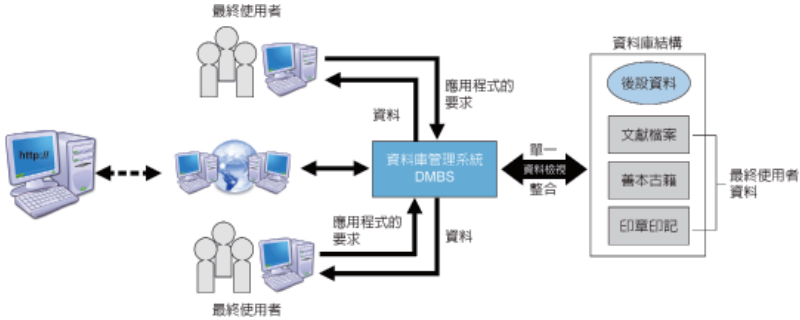


圖 4-6：使用者、資料庫管理系統與資料庫之間的互動關係
參考資料：《資料庫系統：設計、實作與管理》，Peter Rob、Carlos Coronel 著
拓展台灣數位典藏計畫重繪

前文也提到，有正確的資料，才能提供正確的資訊，而資料庫與其結構的設計的優劣也是同等的重要，設計優良的資料庫可以加強資料的管理，並且產生與提供正確的資訊給使用者，也有助於提升相關研究、教育等工作的效率。

2. 資料庫的選擇[87]

如何選擇適合的資料庫系統呢？除了根據各建置單位在資料管理的需求外，也要考慮數位化後資料量規模、機構預算、作業系統平台、資料庫功能等實際數位典藏的專案需求。表 4-8 提供幾個評估因素，在選擇

87 黃國倫，〈資料庫初體驗 (2)〉，拓展台灣數位典藏計畫，檢索：2012 年 5 月，http://content.teldap.tw/index/?p=494。

資料庫產品時可以先進行自我需求分析，以了解資料庫應具備的特性，表 4-9 就是根據評估結果所建議可使用的資料庫類型：

表 4-8：資料庫評量特性表

評估因素	說明
資料複雜度	是否支援有多對多的關係？ 是否提供欄位格式限制？日期、數字、長文字？
資料量	最大的資料儲存筆數？
資料查詢需求	是否支援 SQL 查詢？ 是否提供 AND、OR、部份符合、大於、小於條件查詢？
使用者數量	同一時間最多使用人數？
跨平台要求	是否能在 Windows、Linux 或其他平台運作？
商業用途	是否用於公司營利之目的？

資料來源：〈資料庫初體驗 (2)〉，黃國倫。

表 4-9：資料庫選擇建議表

評估結果	選擇建議
資料簡單、資料量少、無查詢需求	可採用 Microsoft Office 文書工具，如：Word、Excel 等
資料簡單、資料量少、簡單的查詢需求	可採用類似 Microsoft Access 工具
複雜度高、資料量多、複雜的查詢需求、同時多人連線使用	交由系統開發人員評估
需安裝在 Linux 上	Microsoft 產品皆無法使用
非商業用途	可採用 MySQL、PostgreSQL 或其他 Open Source 資料庫產品

資料來源：〈資料庫初體驗 (2)〉，黃國倫。

3. 資料庫建置流程[88]

數位典藏所使用的資料庫系統的建置流程，會經由規劃、分析、設計與開發等四個步驟（圖 4-7）：

88 蔡永橙、黃國倫、邱志義等著，《數位典藏技術導論》，台北市：台大出版中心，2007 年 11 月，頁 68-85。

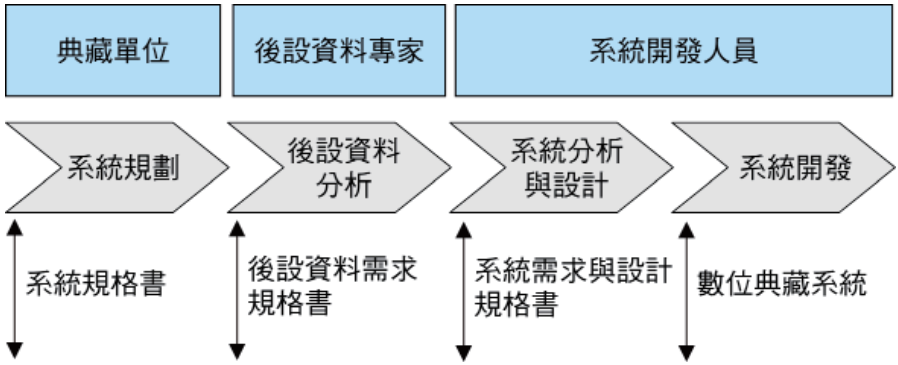


圖 4-7：數位典藏系統建置流程
資料來源：《數位典藏技術導論》，蔡永橙、黃國倫、邱志義等著[89]

(1) 系統規劃與後設資料分析

數位典藏單位根據典藏或管理需求，與後設資料專家、系統開發人員三方溝通，規劃合適的系統規格、功能、後設資料結構與屬性，同時考量整理系統的擴充性、互通性與技術性等，進行初步的可行性評估，產出系統規格書與後設資料需求規格書。

(2) 系統分析與設計

系統開發人員透過與數位典藏單位溝通，分析目標使用者、系統需求的系統功能需求後，進行系統的流程與功能設計，產出系統需求與設計規格書。

(3) 系統開發

系統開發人員根據系統分析所產出的文件中，使用與整合相關的資訊技術與工具，建構出完整的數位典藏系統。一般數位典藏系統基本需要滿足前端檢索呈現，以及後端資料管理等功能，因此從典藏導向與使用導向的角度，所設計的數位典藏系統，除了資料庫外，還可包含權限管理、多媒體檔案管理、檢索展示與缺字等子系統（如圖 4-8）。

89 同註 88，頁 70。

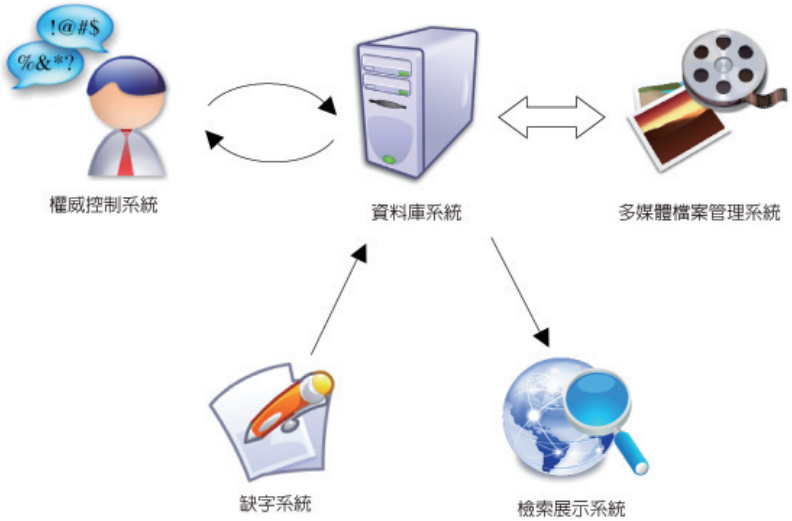


圖 4-8：數位典藏系統架構
資料來源：《數位典藏技術導論》，蔡永橙、黃國倫、邱志義等著 [90]

多數進行文字資料數位化的典藏單位，因為典藏品特性，以及為能提供使用者更多元且豐富的呈現方式，除了文字資料庫外，有的也會搭配原始文字資料的載體影像，如書冊影像、手稿影像等進行呈現，或者與文字資料有關聯之其他資訊進行關聯檢索。以中央研究院「史語所數位知識總體經營計畫－分支計畫：傅斯年圖書館善本古籍國際學術知識網絡建置計畫」為例，其數位典藏系統（http://lib.ihp.sinica.edu.tw/pages/03-rare/system/index.htm）之規劃採取五道程序（圖 4-9）[91]：

(1) 確立數位化主題及清單：

依據各館之館藏特色加以集結或使用者使用頻繁之珍藏加以數位化，並須事先檢視館藏是否可承受數位化之操作，不宜為數位化而將

90 同註 88，頁 80。

91 高芷彤，《古籍線裝書數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2009 年 4 月，頁 19-20。

原始典藏損壞。

(2) 實體典藏之整理及管理：

經由數位化之清單可同步檢視原始典藏並進行修護，珍藏圖籍必須同時考量對外之公用檢索及對內之行政管理機制。

(3) 實體典藏之數位化：

須兼顧典藏及利用二種需求，典藏版的目的是作為典藏之用，以因應原始典藏日益損壞後仍保有其數位化影像（即使虛擬典藏仍無法取代原始典藏），利用版則是考量在網路時代應提供更快速之個人化服務，故另製標誌檔以因應網路傳輸檢索之用。

(4) 虛擬典藏之檢索應用：

實體典藏之數位影像呈現，另依據使用者之各項需求規劃建置其主題屬性之資料庫，並串連各系統之關連性。

(5) 檢索應用之反饋、調整數位化順序：

經由使用者實際檢索利用後同步進行各種系統之使用評估並修正系統設計，並經由使用者之使用反饋調整數位化順序，以取得「藏」與「用」之平衡。

該計畫網站除了一般使用者可查詢檢索的文字資料相關資料庫系統，尚包括「善本圖籍全文影像系統」、「人名權威資料系統」、「空間地理資訊系統」、「館藏印記資料庫系統」、「善本古籍附圖影像系統」、「善本古籍電子全文系統」、「善本圖籍全文出版系統」等，亦有典藏單位內部數位化與藏品管理系統，即「館藏線上公用目錄」、「珍藏書庫庫房管理系統」、「圖籍掃描校驗管理系統」等。此數位典藏系統提供了不同使用者需求的多元資料庫檢索平台，同時內部也能強化數位典藏管理的機制。

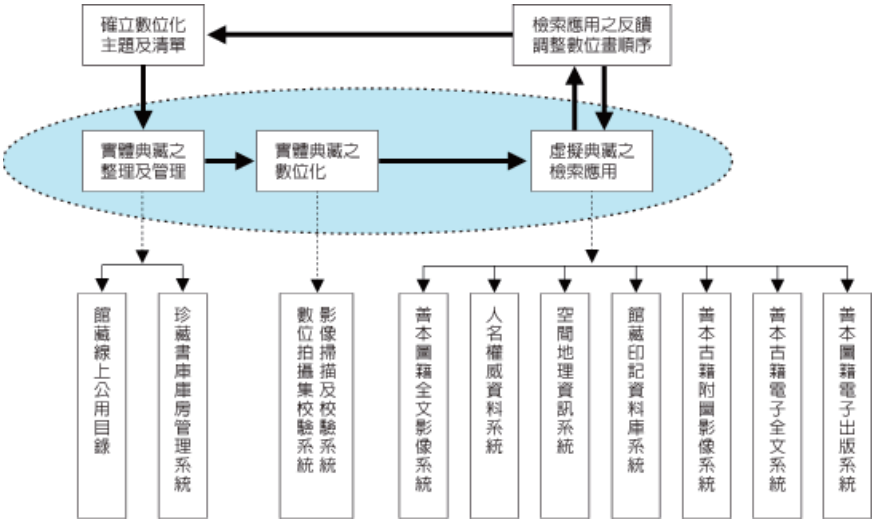


圖 4-9：傅斯年圖書館數位典藏系統規劃圖

資料來源：《古籍線裝書數位化工作流程指南》，高芷彤[92]

（二）資料儲存管理

自 18 世紀首次工業革命以來，人們不斷的發展出能改變社會型態、人類生活方式的科技技術，直到了電腦的發明，到了 20 與 21 世紀的數位革命，藉由數位化與網際網路，資訊能以極低的成本快速地複製與流傳，不僅有效促成各類媒體科技的發展，也改變人們接受、討論與傳播資訊的方式。不論是文字、聲音、圖片、影像等，透過 0 與 1 的編碼，快速地在各種媒體中流通與應用。數位檔案雖然有著方便檢索、應用的優點，但這些檔案也受到資訊科技發展的影響，面臨到更多儲存管理的問題，包括：[93]

1. 數位媒體容易損壞變質：

即便是在恆溫恆濕等理想環境條件下，這些數位媒體較傳統紙本資源的壽命仍來得短，且是屬於易碎易變質的媒材。據推測，光碟的壽命從

92 同註 91。

93 陳雪華、洪維屏，〈數位資訊資源長久保存之探討〉，檢索：2012 年 5 月，http://tech2.npm.gov.tw/faimp/speakers/may4-e1_ch.pdf。

10 年到 100 年不等，但隨著讀寫技術的改變，可讀取光碟中資料的壽命，可能更短。[94]

2. 技術的過時：

數位科技的迅速發展下，硬體、軟體，以及相關的技術，以約 3 到 5 年的週期，被新的產品所取代，新的技術與軟硬體裝置也可能產出新的格式與資源類型，而在新產品之前所產出的數位檔案是否能相容，誰也無法預測。

3. 數位檔案無法獨立存在：

數位檔案是需要有相對應的軟硬體才能讀取與呈現在使用者面前。然而現在科技進度迅速，無法確保新的軟硬體設備是否能讀取現在所數位化的內容，以及相關的產出。

在許多國內外文獻裡，論及到數位資訊保存的策略，可分為系統保存、重複一套系統建置或其他可瀏覽媒體、轉存、標準化、詮釋資料、轉置、模擬與封裝等方式。根據這些策略的保存技術與特性，還可以分成三類（表 4-10）：一是隨著時間必須經常性、週期性進行基礎的保存工作的「基礎層」；二是數位資訊保存技術中最核心，也最重要的「核心層」；第三則是「輔助層」，就是當核心層技術的保存工作有困難或是其他特別的因素下，所需使用的輔助保存技術。[95]

94 陳昭珍，〈電子資源的長久保存〉，《佛教圖書館館訊》，第 25/26 期，2001 年 6 月，檢索：2012 年 2 月，http://www.gaya.org.tw/journal/m25-26/25-main3.htm。

95 歐陽崇榮，《數位資訊保存策略》，台北市：文華，2008 年 3 月，頁 9-10。

表 4-10：數位資訊保存策略

層次	保存策略	簡介
基礎層	轉存	將數位檔案從舊媒體複製到新媒體，例如將硬碟資料轉存到光碟片。
	標準化	選擇出合適的資料標準規格，搭配轉置或轉存策略，所有資料也遵循標準規格重置，以避免遺失或遺漏的問題。
	後設資料	對數位資訊的內容、格式、結構、使用方法 ..... 等的說明與描述，作為電腦系統存取使用的依據。在保存上一般搭配轉置、模擬或封裝等策略使用。
核心層	轉置	將數位資訊的內容、架構與關聯性保存下來，再轉移到新系統中，使用者可在新系統中使用。
	模擬	利用電腦軟體來模擬被保存的數位檔案的軟硬體及其資料原始的運作與呈現方式。
	封裝	將數位資訊與描述該數位資訊的內容的後設資料一起包裹起來，再透過解譯、模擬或轉換的方式，了解數位檔案的內容。
輔助層	系統保存	將數位檔案與其相關的軟硬體或系統一併保存起來。
	重複一套系統建置	將數位資訊重複複製數份，並將這些複製的檔案異地存放，或設立鏡射場地。
	印成紙本或其他可瀏覽媒體	將數位檔案列印成紙本或輸出成微縮膠捲。

資料來源：拓展台灣數位典藏計畫彙整

各種數位檔案的保存方法，有著不同的優缺點，可視各典藏單位的人力、物力等資源狀況與需求，搭配使用，除了定期將資料進行檢測與重新備份外，同時也盡可能在不同地方儲存備份資料，以避免因人為或自然災害導致所有數位檔案的損失。

（三）資料檢索運用

電腦發明之前，人們透過手工的方式，對大量資訊進行分類與建立索引，隨著科技的發展，相關軟硬體技術的進步，使用者幾乎可以在彈指之間，搜尋到想要的資料，像是常見的網路搜尋引擎便是資料檢索的典型代表。許多領域也透過資料檢索系統，來幫助使用者找到正確且符合需求的資訊（如圖 4-10）。

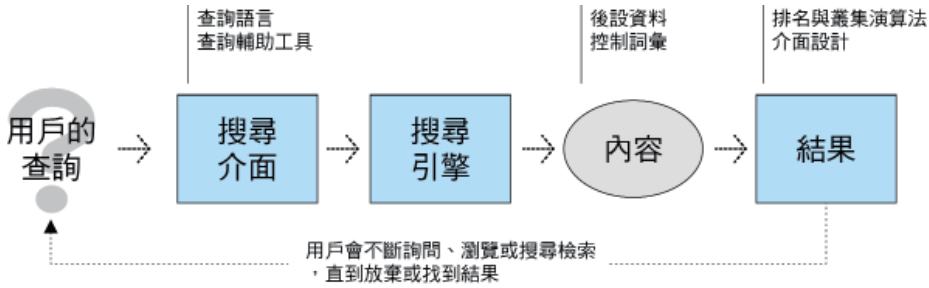


圖 4-10：搜尋系統基本解析
資料來源：《資訊架構學：網站應用．第三版》，Peter Morville、Louis Rosenfeld 著，陳建勳譯 [96]

洪淑芬在《文獻典藏數位化的實務與技術》一書中，也以台大圖書館文獻典藏數位化工作的經驗，匯整出表 4-11 的理想珍貴文獻文物資料庫基本功能。

96 Peter Morville、Louis Rosenfeld 著，陳建勳 譯，《資訊架構學：網站應用第三版》，台北市：歐萊禮，2007 年 7 月，頁 153。

表 4-11：資料庫檢索功能與數位工作關係表

資料庫功能	說明	數位化的相關工作
靈活的檢索功能	可以進行互動式檢索，包括運用運算邏輯，做跨欄位的交集或聯集之檢索	Metadata 欄位規劃 Metadata 建置 檢索程式撰寫
資料庫提供人名與地名權威檔	可幫助使用者查得所有相關資料	同上
提供主題詞或關鍵詞清單	因為具歷史性或學術性之文獻文物，其資料相關用語、地名、人名等往往較為深澀不為人知，因此，資料庫如能提供主題詞或關鍵詞清單，可幫助使用者選擇適用的詞彙進行查詢，增進資料庫利用效益	同上
提供選單檢索功能	使用者於主題詞或關鍵詞清單點選詞彙後，系統自動進行檢索，列出含有該詞彙的資料清單	Metadata 建置 檢索程式撰寫
資料庫於使用者檢索結果中，提供「閱覽簡目」、「閱覽詳目」、「閱覽全文」、「閱覽影像」等選項	手稿古文書之類的資料因文字生澀或部分篇潦草，為便利使用者閱讀，可能同時進行全文數位化，而於資料庫中，與元件影像同時提供，以便使用者對照閱讀	影像數位化 全文數位化 檢索程式撰寫
資料庫提供某一筆（某一頁）資料，連至具有相關性的上屬、下屬、平行等他筆資料，或前後翻頁功能	古文書或書刊資料，往往一筆資料包含多頁，前後筆資料亦可能相關，透過這樣的功能，較重新鍵入關鍵詞檢索，更容易找到相關的下一筆資料	檔名規範 檢索程式撰寫

資料來源：《文獻典藏數位化的實務與技術》，洪淑芬。[97]

以下便介紹幾種文字資料的檢索應用方式，欲建置相關典藏資料庫者，可依據自身典藏的文字資料的特性，選擇一種，或是搭配多種檢索方式來呈現數位內容，來增加典藏數位化的應用效果。

97 洪淑芬，《文獻典藏數位化的實務與技術》，台北市：數位典藏國家型科技計畫 訓練推廣分項計畫，2004 年 2 月，頁 48。

1. 全文資料庫[98]

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。全文資料庫可應具備下列幾種功能：

(1) 全文檢索功能

一般圖書館或是資料庫的文字檢索功能多數僅在書目、作者名以及關鍵字等的搜尋，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。

(2) 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。若能建立層級檢索功能，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聯。

(3) 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，以建立檔案資料的聚集及一致，控制並提高檢索精準率。以中央研究院歷史語言研究所的「人名權威資料查詢」為例，若您要查詢清朝乾隆皇帝的相關資料，在姓名的欄位不論輸入的是乾隆、弘曆、十全老人等，都會指向同一筆的資料。

(4) 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為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

98 王雅萍、謝筱琳，《漢籍全文數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2011 年 6 月，頁 63-66。

排格式，就像瀏覽原書一般。

臺灣大學數位典藏研究發展中心所開發之「台灣歷史數位圖書館」（Taiwan History Digital Library，簡稱 THDL，http://thdl.ntu.edu.tw/）即是一個以「明清時期的臺灣歷史」為主題的數位資料庫，收錄了來自國立臺灣大學圖書館與國立臺中圖書館所全文數位化的第一手臺灣史料，並提供全文檢索。由於此全文檢索系統只要史料全文中有出現該關鍵字的內容，即使不是使用者需要的也都會全數被找出來，而增添使用者困擾。所以該建置團隊為了提供「能幫助歷史研究者有效利用大量史料進行研究」的工具，發展出「檢索後分析」的概念—系統在檢索後，會主動分析這些被檢索資料的特徵與關聯性，例如分布的年代、相關的人物地點、相似文書資料等，並提供分類讓使用者便於再縮小範圍進行查詢（圖 4-11）。[99] 稍微可惜的是該數位圖書館並未完全解決人名或地名的權威控制問題。[100]

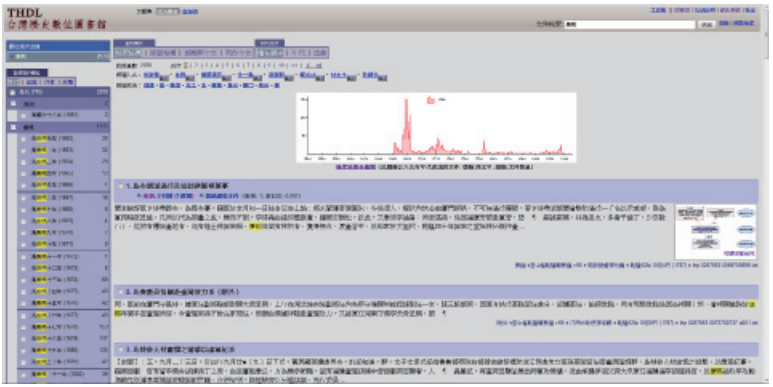


圖 4-11、台灣歷史數位圖書館檢索畫面

99 陳詩沛、杜協昌、項潔，〈史料整理分析工具之幕後—介紹「臺灣歷史數位圖書館」的資料前置處理程序〉，《從保存到創造：開啓數位人文研究》，台北市：台大出版中心，2011 年 11 月，頁 51-66。

100 陳志豪，〈臺灣歷史數位圖書館與歷史研究的實際應用—以「淡新檔案」為例〉，《從保存到創造：開啓數位人文研究》，台北市：台大出版中心，2011 年 11 月，頁 67-94。

2. 時空與地理資訊檢索[101]

許多文獻資料的內容中，可能提及到許多舊時的時間與地理背景，對現在的使用者來說，更添閱讀理解上的困難，而搭配地理資訊系統來進行內容檢索，便是一種可以讓使用者更容易閱讀這些古文獻的方式。

地理資訊系統因為需要較高的硬體運算能力，所以過去較多僅應用於處理空間資料的專業領域。隨著科技的進步，如今一般個人電腦也能輕易地處理地理資訊系統所需的運算，加上程式處理思維的創新，讓地理資訊系統相關應用與呈現也更加多元。

法鼓佛教學院自西元 2007 年起，便開始以地理資訊系統技術來輔助資料呈現，分別有「漢傳佛教高僧傳地理資訊系統」專案，以及「臺灣佛教地理資訊系統」專案。其中「漢傳佛教高僧傳地理資訊系統」專案所建置的漢傳佛教高僧傳地理資訊系統（http://dev.ddbc.edu.tw/biographies/gis/interface/），便是將《梁高僧傳》、《唐高僧傳》、《宋高僧傳》、《明高僧傳》的文字資料，利用 XML/TEI 文獻處理技術以及地理資訊系統（GIS）整合時間、人名、地名等資訊，轉換為時空地理資訊系統的圖像化數位資源，網頁上利用 Google Earth 所提供的圖資資料，同時將文字內容與時間呈現給大眾（圖 4-12）。使用者可以在該系統進行圖文參照或是時間的檢索，其中空間資料的部分，是以高僧的出生地、弘法地、遊化地、駐錫地地理位置的標定為主，可進一步經由地理座標上的統計，提供佛教流布、宗派教區等資訊，了解各個佛教高僧走過各地的歷程。[102]

101 廖泫銘，〈漢傳佛教高僧傳之時空資訊系統〉，檢索：2012 年 5 月，http://gis.rchss.sinica.edu.tw/google/?p=1600。

102 洪振洲、馬德偉、張伯雍、李志賢、黃仁順，〈佛教數位典藏與 GIS 技術應用經驗分享〉，《從保存到創造：開啓數位人文研究》，台北市：台大出版中心，2011 年 11 月，頁 147-167。

圖 4-12：漢傳佛教高僧傳地理資訊系統介面

在數位典藏工作中，除了將藏品以掃描、拍攝等方式數位化為數位物件儲存，達到永久典藏之目的外，更重要的，是要讓這些數位檔案可以供各界透過網路查詢、閱覽與使用。透過後設資料之描述，能完整呈現藏品之原始特性與數位化資訊，並且也有助於進行資料庫系統設計。而現今資料庫的設計也愈趨多元，讓使用者不僅可以檢索查詢基本功能外，也結合了不少跨領域系統的運用，使典藏資料庫的內容應用更加廣泛，資訊更為豐富。

伍、加值應用

Value-added application

將珍貴的文化資產進行數位化保存、管理與檢索應用後，更重要的是能透過各種加值應用的方法，使這些資產擁有最佳的附加價值。陳雪華（2002）將數位典藏成果之應用分為「靈感的啟發」與「素材的應用」兩種模式，靈感的啟發指的是將典藏內容數位化後做有系統地整理，並讓使用者對這些內容有更進一步的瞭解；素材的應用則是將產出之成果加值於其他商品之中。[103] 本章節依此內涵將加值的應用模式分為「數位學習」與「商業運用」兩種類型，透過這些方式讓數位典藏內容更深化於教育、研究、生活與產業發展之中。以下將輔以實例作分別的說明：

一、數位學習

數位學習（e-Learning）指的是學習者透過數位媒介進行學習的歷程。由於學習者可依個人的喜好選擇適當的數位媒介（如光碟、電視、廣播、錄影帶、遊戲機、網際網路等），以同步或非同步的方式進行學習，使學習不再受限於時間、空間，並具有個人化的特性。隨著科技的發展和網際網路的無遠弗屆，數位學習已是今日普及且便利的學習型態，而目前常見的數位學習形式大致有「主題網站」、「電子書」和「線上數位學習課程」等方式。下文將略述數位化文字資料於此三方面的加值應用。

（一）主題網站

主題網站是透過多媒體或多個網頁呈現特定主題的知識內容，學習者可藉由瀏覽網站獲取自己所需的知識或資訊。以故宮為例，過去館藏的文化歷史資產相當豐富且多元，經由資料授權與多媒體製作後，讓受制於時間、空間而無法親自前往館內觀賞的大眾，亦能在家透過網際網路進行學習。目前故宮博物院的主題網站依照知識類別的不同，分為圖書文獻類（如圖 5-1）、書畫類、器物類與科技類四種類別。在圖書文獻類「皇城聚珍—殿本圖書欣賞」網站中，

103 陳雪華、項潔、鄭惇方，〈數位典藏在數位內容產業之應用加值〉，《博物館典藏數位再造理論與實務研討會—人與自然論文集》，2002 年 11 月，頁 17-24。

學習者可透過動畫、經數位化處理的殿本照片及文字說明，了解殿本的特色和刊印流程，並賞析過去難能可見的《四庫全書》、《周禮》、《勸善金科》、《太宗皇帝大破明師於松山之戰書事文》等古籍。最後，亦可藉由網站的小遊戲，檢視自己對於殿本知識的吸收程度。若學習者對佛經典籍有興趣，也可於「千華臺上—佛經文獻圖說」網站裡，點選欲學習的標題文件和佛籍照片，一窺珍貴的手寫佛經內容及相關佛典譯本。

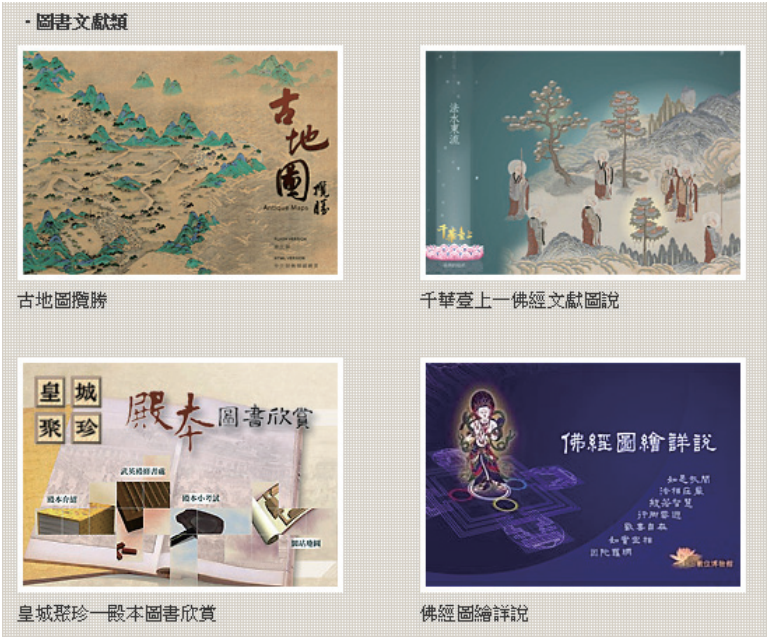


圖 5-1：主題網站加值應用範例—國立故宮博物院
資料來源：國立故宮博物院[104]

此外，中央研究院數位典藏資源網的「筆墨譚心—延閣日記」[105]（http://digiarch.sinica.edu.tw/tan/）亦是頗具特色的主題網站。我們不僅可在網站裡閱讀

104 數位故宮，檢索：2012 年 4 月，http://www.npm.gov.tw/digital/archives/b04.html#。
105 筆墨譚心—延閣日記，檢索：2012 年 5 月，http://digiarch.sinica.edu.tw/tan/。

譚延闓先生過去的生平與經歷，也可藉由其已數位化的日記文稿中，從不同角度了解民初的政壇名流與他們的交互關係，有助於增進我們對於重要歷史事件與特色人物的認識。

圖 5-2：主題網站加值應用範例一「筆墨譚心—延闓日記」
資料來源：中央研究院數位典藏資源網

（二）電子書

電子書是藉由將資料數位化，使之得以用結構化的形式進行編輯、管理，並透過不同的載體讓學習者可以閱讀書籍的形式進行瀏覽。原先電子書專指電子化或數位化的圖書，隨著資訊的快速流通和演變，亦可泛指由多媒體編輯軟體所製作，具有動態聲光效果並以書籍形式揭示的內容物。電子書的載具可依使用者的閱讀習慣，為手機、電腦、平板電腦、電子書閱讀機等設備。在格式方面，早期是以 PDF 格式為主，現今廣為使用的電子書格式為 EPUB 格式，是 2007 年由國際數位出版論壇（International Digital Publishing Forum，簡稱 IDPF）所提出的自由開放標準，此外亦有 AZW、LRF、MOBI 等格式（詳見「附錄三、電子書格式簡介」）。

目前市面上有許多製作電子書的軟體（如 Flash Page Flip、AZARDI、eScape、epubBuilder、EpubSTAR、Adobe Indesign 等），也有網站提供線上製作

服務（如 Issuu 線上服務網站[106]），使用者可自行將現有的檔案資源匯入軟體中，即可完成互動式電子書的製作。透過閱讀擬真的紙本書和生動詮釋的書籍內容，達到寓教於樂的效果。

接下來，將針對國內幾個推動電子書的機構和其案例內容作介紹：

1. 國家圖書館

國家圖書館是國內典藏豐富善本古籍資源的單位，在古籍文獻數位化方面行之有年，近年來更積極挑選古籍製作電子書籍，讓更多讀者貼近這些古代經典作品，並藉此推動數位閱讀。[107] 此外，在國家圖書館製作「漢學通覽經典系列」與「臺灣記憶系列」電子書中，我們可以看到明朝陽瑪諾傳教士描述天文觀察與宗教觀的《天問略》，也可以看到我國第一部描述西方靜力物理學與機械工程學的《奇器圖說》，抑或是從學琴手勢到琴譜、樂理詳細記載的《太古遺音》等古籍。這些電子書有的以 PDF 檔案格式呈現原書內容，有的以 FLASH 動畫製作，展現出一本書在數位閱讀中的各種樣貌。

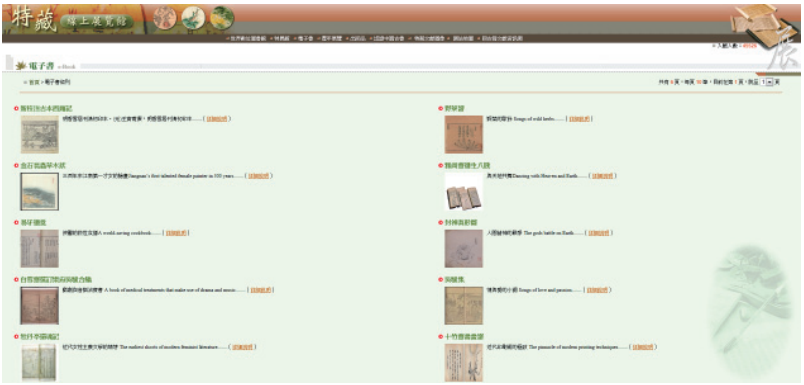


圖 5-3：電子書加值應用範例 1 —國家圖書館特藏線上展覽館
資料來源：國家圖書館特藏線上展覽館

106 Issuu 網站，檢索：2012 年 4 月，http://issuu.com/。
107 國家圖書館特藏線上展覽館，檢索：2012 年 4 月，http://rarebook.ncl.edu.tw/rbookod/exhibition/hypage.cgi?HYPAGE=ebook/ebook_list.htm。

2. 中華電子佛典協會

在佛典方面，中華電子佛典協會以收集所有的漢文佛典、建立電子佛典集為宗旨，利用電子媒體之特性，促進佛典保存與流通。該協會將大正藏、卍續藏、嘉興藏、金藏、中華藏…等眾多佛典經文製作為 EPUB 格式的電子書，只要進入「CBETA 漢文大藏經網站」電子佛典資料庫[108]作搜尋，就可以找到相關的電子佛典，同時也可透過行動載具下載閱讀。若學習者想在智慧型手機或平板電腦等裝置中閱讀 CBETA 電子書，可參閱中華電子佛典協會的使用說明（http://www.cbeta.org/epub_opds.php）。

圖 5-4：電子書加值應用範例 2 — CBETA 漢文大藏經網站
資料來源：CBETA 漢文大藏經網站

108 CBETA 漢文大藏經—電子佛典集成資料庫，檢索：2012 年 4 月，http://tripitaka.cbeta.org/index.php。

3. 數位典藏與數位學習國家型科技計畫

數位典藏與數位學習國家型科技計畫於 2011 年建立「網上書上網數位典藏與學習電子書庫」電子書平台[109]，與國內典藏單位如中央研究院傅斯年圖書館、國家圖書館、台大圖書館、農業委員會林業試驗所…等機構合作，將重要的古籍、書冊製作為 EPUB 格式的電子書，內容包括文學、史地、藝術、哲學、科學等類別。由於善本古籍具有珍貴且不易保存的性質，在數位化之前，一般民眾較不易接觸到這些經典的真實樣貌，但當善本古籍以數位化方式製作成電子書後，便跨越紙本典藏的限制，提供大眾閱覽古籍經典的機會。

圖 5-5：電子書加值應用範例 3 一網上書上網一數位典藏與學習電子書庫的電子書平台
資料來源：網上書上網一數位典藏與學習電子書庫

109 網上書上網一數位典藏與學習電子書庫，檢索：2012 年 4 月，http://ebook.teldap.tw/index.jsp。

（三）線上數位學習課程

線上數位學習課程為透過網路資訊平台展示知識內容或資訊，使學習者能在網際網路上進行學習。由於在進行數位學習課程製作時，大多會採用系統化的教學設計模式進行課程的規劃，故學習的目的性或互動性可能較主題網站或電子書明顯。以下提供兩個與文字資料有關的線上數位學習課程案例做為參考：

1. 淡新檔案學習知識網

《淡新檔案》為清乾隆四十一年至光緒二十一年間淡水廳、臺北府以及新竹縣的行政和司法檔案，對於研究我國法制史、地方行政史、社會經濟史和農業經濟極具學術價值。臺大圖書館於數位典藏計畫中將《淡新檔案》影像數位化並進行 Metadata 建檔，將數位化成果建置網站供民眾使用。但由於《淡新檔案》具有相當的深度與複雜度而不易利用，為推廣《淡新檔案》之學術教育應用，臺大圖書館於民國 96 年至 99 年間執行「臺灣文獻數位典藏教學研究應用計畫：《淡新檔案》學習知識網」計畫，製作「臺灣大學淡新檔案學習知識網」[110]。站內提供「文書類別」、「版面格式」、「主題故事館」、「行政流程」和「測驗卷」等單元學習課程，讓學習者更深入地了解珍貴的文獻檔案。

110 臺灣大學淡新檔案學習知識網，檢索：2012 年 4 月，http://www.digital.ntu.edu.tw/tanhsin/description/1.html。

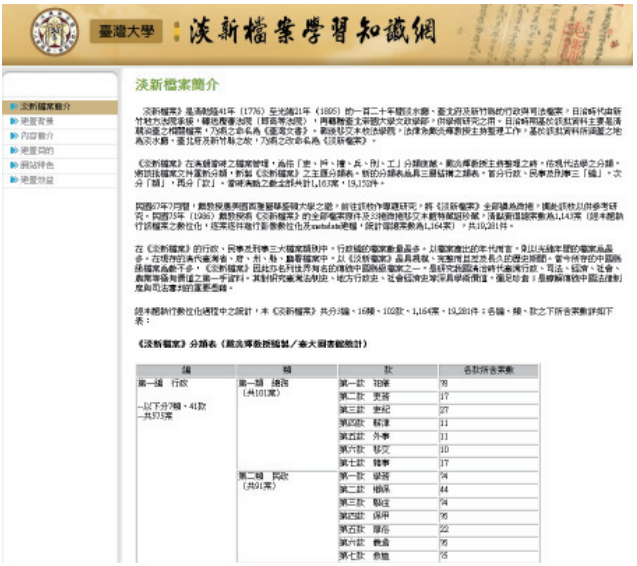


圖 5-6：線上數位學習課程加值應用範例 1 —淡新檔案學習知識網
資料來源：淡新檔案學習知識網

2. 國立歷史博物館「千古一草聖于右任的書法世界」

書法是文字書寫的藝術，其包含線條、美感與思想的平衡呈現。藉由書法的學習，可讓學習者了解過去文字的演變。但對於許多初學者而言，學習書法前的準備工作相當繁複，書寫後亦須花時間清理與善後，使其對學書法望之卻步。在「千古一草聖于右任的書法世界」[111] 線上數位學習課程中，透過將書法資料數位化，學習者可同時觀賞到不同書法名家的作品，並從中理解作家的時代背景和創作理念。此外，每個子單元前面皆以生動活潑的動畫引領學習者進入單元內容，在單元尾端亦安排多媒體測驗，讓學習者可階段性地檢核自己的學習狀況。

111 千古一草聖于右任的書法世界，檢索：2012 年 4 月，http://webtitle.nmh.gov.tw/yuyouren/index_welcome.html。

圖 5-7：線上數位學習課程加值應用範例 2 一千古一草聖于右任的書法世界
資料來源：國立歷史博物館

二、商業運用

商業運用意即將數位化文字資料轉化成產品，使之可應用在經濟相關的範疇。透過商業的應用，能賦予數位資料嶄新的價值，進而開創商機，帶動數位創意產業的發展。目前數位資料的商業應用型態多採「網路販售」和「產業合作」兩種模式，下面將依序提供相關的說明：

（一）網路販售

因應數位化時代的來臨，人們購物的習慣逐漸發生轉變，不再只侷限於過去面對面的實體交易，取而代之的是以網路、電視、電話等數位媒介為主的虛擬交易。由於網路購物十分便利，深受許多從事買賣行為的人士喜愛，各式各樣的網路商城亦紛紛建立。在國立故宮博物院的網路商城[112]中，我們可以看到古代名家的複製墨跡或字畫所製作而成的手卷、框畫、飾品、文具或禮品等商

112 故宮精品網路商城，檢索：2012 年 4 月，http://www.npmshops.com/main/modules/MySpace/index.php?sn=npmshops&cn=ZC687。

品；亦可看到記載歷史文化或特展活動的期刊、光碟、書籍等影音圖書。此外，故宮所珍藏的清康熙朝泥金藏文寫本《龍藏經》[113]，即是由故宮授權典藏文字資料予廠商印製出版、販售的案例。此《龍藏經》收集釋迦牟尼佛所說的顯密經典約 1,100 部，包含圖像 2 冊、經文 108 冊和檢索 1 冊，共 111 冊套書。藉由各家網路商城的建置，可讓有興趣的購買者先於網站瀏覽購買資訊，或透過網路與商家聯繫的方式，以決定是否進一步購買，使購買過程比以往更加方便。

（二）產業合作

除了將數位資料作直接的販售外，亦可用既有的資源與廠商合作開發商品，以最大化商業效益。「數位典藏橋接計畫」即為「數位典藏與數位學習國家型科技計畫」總計畫辦公室與「財團法人資訊工業策進會」產業支援處一同合作規劃執行的計畫。「數位典藏與數位學習國家型科技計畫」可提供執行多年所累積的豐富數位典藏品，數位創意產業則可協助篩選藏品中能作為商用的素材，予以設計、開發和製作商品，進而國際授權展售、推廣和行銷。藉由與產業合作，不僅能彰顯我國數位典藏多元且豐富的價值，亦可大幅推動數位創意產業的蓬勃發展。在該計畫中，「數位典藏加值商用平台」的建立是重要的執行項目之一。我們可於「數位典藏加值商用平台」[114] 裡搜尋數位典藏的素材，或是瀏覽相關的加值設計作品等服務。透過系統化的加值模式建立，提供我們一套可依循的典藏素材商用化篩選機制、國際授權及商業推廣之模式。

113 此《龍藏經》指的是由國立故宮博物院授權民間廠商所印製出版的圖書，檢索：2012 年 5 月，http://www.npmshops.com/main/modules/MySpace/index.php?sn=npmshops&cn=ZC323881。

114 數位典藏加值商用平台，檢索：2012 年 4 月，http://www.teldapbridge.org.tw/teldap/bridge/index.php。

圖 5-8：商業運用加值應用範例—數位典藏加值商用平台
資料來源：數位典藏加值商用平台

文字資料之數位加值應用層面已趨漸廣，從以往可運用的檢索資料庫至今日各種創新的推廣服務，都隨著資訊科技的發展有了更多形式的展現。由於科技和網際網路的發達，人們今日常以平板電腦、手機等行動載具，進行電子書、主題網站、社群網站或是 APP 應用程式的瀏覽。從事數位內容推廣工作者可依循大眾平時接收資訊的途徑，規劃並建構推廣服務（如資源網站或應用程式的內容建置和下載等服務），或是將數位化的文字資料素材，以創新的概念運用於與社會大眾相關的交流服務，使數位典藏能更加融入於日常生活之中。

然而，文字資料本身蘊藏著數千年來的歷史文化精華，除了透過諸多模式的傳承發揚外，加值應用的規劃亦須回歸本質上的考量，例如：瞭解計畫的加值需求、釐清加值本身的動機與目的，甚至相關的著作權益、後續管理機制等問題，皆是不容輕忽的工作。透過審慎的事前評估，才能使數位化資料的推廣效益最大化。

陸、結語

Conclusion

透過文字的紀錄，人類的文明發展與知識的累積得以有機會地被留存下來。這些珍貴的文獻刊載各式的重要資料，若以傳統媒體保存，固然有其歷史意義，假如能以科技技術轉換為電子數位檔加以保存並讓遠端多方的人們使用，將使這些資料的傳播更具傳承意義。因此，近十幾年來國內的數位典藏計畫，對於如何規劃執行數位化工作、建立與運用文字資料皆已成為不容輕忽的工作。

數位化的方式或技術因科技時代有所汰換，隨著設備的提升也相對減低人力成本的支出，並更能有效地將圖書館館藏或其他大量文獻資料數位化。由於文字的結構不同，記載的載體也因時代有異，進行文字資料的數位化工作，得依其不同性質採取適合的執行方式。此《數位化工作流程指南：文字資料》，即試著從文字資料的範圍，再依不同的記錄載體與文字資料的類型，針對其文字的特殊性如何進行數位化工作，介紹各類文字資料的處理方式。除了最初的物件挑選，在文字資料數位化程序上，著重「輸入」與「文字處理」兩大數位化步驟。輸入的方式從早期的人工輸入、拍攝甚至掃描技術也不斷地演進，增進不少文字辨識的效率提升數位化工作的品質。其中文字處理最常遇到的缺字、異體字、斷詞等問題，也隨著部分研究機構的不斷實驗研究，開發了一些便捷有利的處理系統，更是後續資料庫建置的一大助力。這些處理系統解決了不少數位典藏資料著錄的難題，讓更多的文字資料能透過網路直接查詢、瀏覽，使大量的文字資料更有效地流通與整合資源，達到資料庫共建共享的更深一層意義。

文字資料的數位化，不僅是達到資料分享的意義外，進一步的運用與加值亦是傳承與發揚這些珍貴資產價值的一途。尤其是隨著網路科技的發達、行動裝置的盛行等趨勢，許多的文字資料開發出的應用學習或數位商品也日益增多，例如電子書、電子書包等等。就數位典藏的角度而言，在文字資料數位化的階段，不僅是輸入文字內容、建置詳盡的後設資料，在後續的數位學習階段或應用加值方面，都可能需要因應不同對象或場域製作相應的功能。此本指南在前面章節也都略舉出目前國內外已有的應用網站或加值商品，尤其是漢文字的獨

特性，相較於其他國家的文字資料更能呈現其特色風貌，也是不少技術人員或研發內容的團隊一再挑戰的對象。希冀這些特有的文字資料能以各種運用層面，予以大眾利用並加以傳承下去。

這些數位化工作程序皆是環環相扣，彼此步驟流程之間是互相關連的，在在影響其執行工作的效率以及品質是否合乎預期的標準。透過數位化工作流程指南，是規劃專案、執行計畫的參酌辦法之一，以期相關計畫工作能順利進展，達到數位化的完善成果。雖然此本指南的相關工作標準亦會隨著技術的更新有所異動，但在過程中發展的這些各種新的創意、新的應用，或許也就是數位典藏成果發酵的另一風景與見證。

參考文獻

References

專書

王泰升、陳詩沛、杜協昌等 / 合著，項潔編，《從保存到創造：開啟數位人文研究》，台北市：台大出版中心，2011 年 11 月。

王雅萍、謝筱琳，《漢籍全文數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2011 年 6 月。

李佩瑛、程婉如，《期刊報紙數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2009 年 4 月。

林尹，《文字學概說》，台北市：正中書局，2007 年 10 月。

林彥宏、程婉如、張思瑩，《微縮資料數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2011 年 6 月。

洪淑芬，《文獻典藏數位化的實務與技術》，台北：數位典藏國家型科技計畫訓練推廣分項計畫，2004 年 2 月。

海拉 • 哈爾門著，方奕 譯，《文字的歷史》，台中市：晨星出版，2005 年 4 月。

高芷彤，《古籍線裝書數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2009 年 4 月。

陳秀華、溫敏宇，《金石拓片數位化工作流程指南》，台北市：數位典藏拓展台灣數位典藏計畫，2011 年 6 月。

詹雅蘭，《OAIS 參考模式應用在國家檔案永久典藏機制之探討》，台北市：台灣師範大學圖書資訊學研究所，2004 年 6 月。

歐陽崇榮，《數位資訊保存策略》，台北市：文華，2008 年 3 月。

蔡永橙、黃國倫、邱志義等著，《數位典藏技術導論》，台北市：台大出版中心，2007 年 11 月。

魯 • 伯納、麥克 • 蘇寶麥昆、馬德偉 著，謝筱琳、黃韋寧 譯，《TEI 使用指南》，台北市：數位典藏拓展台灣數位典藏計畫，2009 年 4 月。

戴慶夏、趙小兵等主編，《中國少數民族語言文字信息處理研究與發展》，北京市：民族出版社，2010 年 6 月。

羅凡晟，《文字學數位內容加值應用之研究》，台北縣：花木蘭文化出版社，

2010 年 9 月。

Peter Morville、Louis Rosenfeld 著，陳建勳 譯，《資訊架構學：網站應用第三版》，台北市： 歐萊禮，2007 年 7 月。

Peter Rob、Carlos Coronel 著，張世敏 譯，《資料庫系統：設計、實作與管理》，台北市：新加坡商聖智學習，2009 年 1 月。

期刊論文

王麗蕉，〈檔案描述標準 MARC AMC 與 EAD 之對映〉，《圖書與資訊學刊》，第 51 期，2004 年 11 月，頁 112-113。

杜正民，〈佛學研究資源數位化作業標準與規範 Standards of the Digitalized Buddhist Research Resource Project〉，《漢學研究通訊》，第 96 期，2005 年 11 月，頁 7-16。

周邦信，〈標記語言的應用〉，《佛教圖書館館訊》，第 24 期，2000 年 12 月。

洪淑芬，〈紙質文獻類的雜誌書籍之數位化〉，《佛教圖書館館刊》，第 45 期，2007 年 6 月，頁 19-25。

陳雪華、項潔、鄭惇方，〈數位典藏在數位內容產業之應用加值〉，《博物館典藏數位再造理論與實務研討會—人與自然論文集》，2002 年 11 月，頁 17-24。

歐珠、普次仁、大羅桑朗杰、趙棟才、劉芳、邊巴旺堆，〈印刷體藏文文字識別技術研究〉，《計算機工程與應用》，第 45 卷第 24 期，2009 年，頁 166。

蘇倫伸，〈日治時期日文臺灣文獻數位典藏計畫概述〉，《臺灣圖書館管理季刊》，第 4 卷第 4 期，2008 年 10 月，頁 75-81。

網路資源

CBETA 漢文大藏經 電子佛典集成資料庫，檢索：2012 年 4 月，http://tripitaka.cbeta.org/index.php。

Issuu 網站，檢索：2012 年 4 月，http://issuu.com/。

TEI 網站，檢索：2012 年 2 月，http://www.tei-c.org/index.xml。

千古一草聖于右任的書法世界，檢索：2012 年 4 月，
http://webtitle.nmh.gov.tw/yuyouren/index_welcome.html。

中文斷詞系統，檢索：2012 年 2 月，http://ckipsvr.iis.sinica.edu.tw/。

中國西南少數民族資料庫，檢索：2011 年 11 月，
http://ndweb.iis.sinica.edu.tw/race_public/index.htm。

甲骨文數位典藏資料庫，檢索：2011 年 12 月，
http://ndweb.iis.sinica.edu.tw/rub_public/System/Bone/home2.htm。

全字庫，檢索：2012 年 2 月，http://www.cns11643.gov.tw/AIDB/welcome.do。

吳寶原，〈創意靈感—關於電子藏經的輸入、校對及編輯〉，《佛教圖書館館訊》， 第 24 期，2000 年 12 月，檢索：2012 年 02 月，
http://www.gaya.org.tw/journal/m24/24-main2.htm。

拓展台灣數位典藏計畫，檢索：2012 年 2 月，http://content.teldap.tw/index/。

金石拓片資料庫，檢索：2012 年 2 月，http://rarebook.ncl.edu.tw/gold/。

青銅器拓片數位典藏資料庫，檢索：2011 年 11 月，
http://rub.ihp.sinica.edu.tw/~bronze/index.htm。

故宮精品網路商城，檢索：2012 年 4 月，
http://www.npmshops.com/main/modules/MySpace/index.php?sn=npmshops&cn=ZC687。

柯維盈，〈歷史語言研究所藏甲骨文拓片資料庫〉，金石拓片數位典藏研討會，檢索：2011 年 11 月，http://rub.ihp.sinica.edu.tw/~oracle/05/01.pdf。

缺字系統，檢索：2012 年 2 月，http://char.iis.sinica.edu.tw/index.htm。

馬偉雲，〈未知詞擷取作法〉，中文斷詞系統，檢索：2012 年 2 月，
http://ckipsvr.iis.sinica.edu.tw/。

國立台灣師範大學台灣文化及語言文學研究所，〈台灣白話字發展簡介〉，台灣白話字文獻館，檢索：2011 年 11 月，

http://www.tcll.ntnu.edu.tw/pojbh/script/about-2.htm。

國家圖書館特藏線上展覽館，檢索：2012 年 4 月，http://rarebook.ncl.edu.tw/rbookod/exhibition/hypage.cgi?HYPAGE=ebook/ebook_list.htm。

莊德明，〈漢字數位化的困境及因應：談如何建立漢字構形資料庫〉，文獻處理實驗室，檢索：2012 年 2 月，http://cdp.sinica.edu.tw/service/documents/T960507.pdf。

陳信文，〈以 Microsoft Visual C# UserControl 實作 OCR 校對工具〉，中央研究院計算中心通訊電子報，第 11 期，檢索：2012 年 2 月，http://newsletter.ascc.sinica.edu.tw/news/read_news.php?nid=1878。

陳昭珍，〈文字資料的數位化〉，數位典藏學程—人文領域，檢索：2011 年 12 月，http://humanities.lis.ntu.edu.tw/md/20070314.pdf。

陳昭珍，〈電子資源的長久保存〉，《佛教圖書館館訊》，第 25/26 期，2001 年 6 月，檢索：2012 年 2 月，http://www.gaya.org.tw/journal/m25-26/25-main3.htm。

陳雪華、洪維屏，〈數位資訊資源長久保存之探討〉，檢索：2012 年 5 月，http://tech2.npm.gov.tw/faimp/speakers/may4-e1_ch.pdf。

傅斯年圖書館，〈傅斯年圖書館全彩影像掃描、拍攝及校驗相關作業標準〉，傅圖數位典藏計畫網站，檢索：2011 年 11 月，http://lib.ihp.sinica.edu.tw/pages/03-rare/DAP/contentp/03-4-2.pdf。

筆墨譚心—延闓日記，檢索：2012 年 5 月，http://digiarch.sinica.edu.tw/tan/。

黃國倫，〈資料庫初體驗 (2)〉，拓展台灣數位典藏計畫，檢索：2012 年 5 月，http://content.teldap.tw/index/?p=494。

經濟部標準局，〈中文分詞處理原則〉，國家標準（CNS）檢索系統，檢索：2012 年 2 月，http://www.cnsonline.com.tw/previewJPG.jsp?general_no=1436600&language=C&pagecount=315。

廖泫銘，〈漢傳佛教高僧傳之時空資訊系統〉，檢索：2012 年 5 月，

http://gis.rchss.sinica.edu.tw/google/?p=1600。

漢代簡牘數位典藏資料庫，檢索：2012 年 2 月，
http://rub.ihp.sinica.edu.tw/~woodslip/index.htm。

漢籍電子文獻瀚典全文檢索系統，檢索：2011 年 11 月，
http://hanji.sinica.edu.tw/。

網上書上網 數位典藏與學習電子書庫，檢索：2012 年 4 月，
http://ebook.teldap.tw/index.jsp。

臺灣大學淡新檔案學習知識網，檢索：2012 年 4 月，
http://www.digital.ntu.edu.tw/tanhsin/description/1.html。

蒙恬科技，〈光學辨識技術原理概述〉，蒙恬科技網站，檢索：2012 年 1 月，
http://www.penpower.com.tw/technology-OCR.asp。

數位典藏加值商用平台，檢索：2012 年 4 月，
http://www.teldapbridge.org.tw/teldap/bridge/index.php。

數位典藏技術彙編 2007 年版，檢索：2012 年 4 月，
http://www2.ndap.org.tw/eBook08/showContent.php。

數位典藏國家型科技計畫，〈北平世界日報內容數位化開發計畫之數位化工作流程圖文說明〉，《國家數位典藏通訊電子報》，檢索：2011 年 11 月，
http://www2.ndap.org.tw/newsletter06/news/read_news.php?nid=544。

數位典藏與數位學習國家型科技計畫，〈中文缺字技術〉，計畫百科，檢索：2012 年 2 月，http://goo.gl/tkOzT。

數位典藏與數位學習國家型科技計畫，〈後設資料生命週期作業模式〉，後設資料工作組，檢索：2012 年 4 月，
http://metadata.teldap.tw/design/lifecycle_new2.htm。

數位典藏與數位學習國家型科技計畫，〈通往數位典藏豐碩藏品的窗口—目錄導覽〉，數位典藏與數位學習成果入口網，檢索：2012 年 4 月，
http://digitalarchives.tw/about.jsp。

數位故宮，檢索：2012 年 4 月，

http://www.npm.gov.tw/digital/archives/b04.html#。

蔡耀明，〈網路上的梵文與梵文佛典資源〉，《佛學數位資源之應用與趨勢研討會論文集》，2005 年 09 月，檢索：2011 年 12 月，http://buddhism.lib.ntu.edu.tw/BDLM/seminar/book0.htm。

聯合新聞網，檢索：2011 年 12 月，http://udn.com/NEWS/main.html。

謝清俊，〈漢字的字形與編碼〉，文獻處理實驗室，檢索：2011 年 11 月，http://cdp.sinica.edu.tw/paper/1996/19961004_1.htm。

Library of Congress. (2008). *Structural metadata dictionary for LC digital objects.* Retrieved November 30, 2011, from http://memory.loc.gov/ammem/techdocs/repository/attdefs.html

Library of Congress. (2009). *American memory DTD for historical documents.* Retrieved February 20, 2012, from http://lcweb2.loc.gov/ammem/amdtd.html

Mengkuei Hsu，〈自製 ePub 電子書（上）：認識電子書格式之爭〉，T 客邦，檢索：2012 年 5 月，http://www.techbang.com.tw/posts/2467-home-epub-e-book-on。

The National Archives and Records Administration. *Reformatting approaches*. Retrieved June 1, 2012, from http://www.archives.gov/preservation/products/definitions/reformatting.html

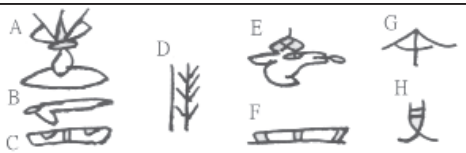
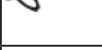
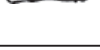
The National Archives. *Guidance*. Retrieved June 1, 2012, from http://www.nationalarchives.gov.uk/information-management/projects-and-work/guidance.htm

Victoria University of Wellington. *About the New Zealand Electronic Text Centre.* Retrieved February 20, 2012, from http://www.nzetc.org/tm/scholarly/tei-NZETC-About.html

附錄

Appendix

附錄一、納西族東巴經之〈破地獄經〉文書翻譯資料

文書翻譯資料：以標題或內文為著錄單位（以 MS-102 標題為例）[115]								
狀態		著錄						
使用限制		限制						
文書登錄號		MS-102						
翻譯者		和力民						
翻譯日期		2003-03-14						
原件類別		標題						
影像檔名		 NX0102CT001000.tif						
相關影像檔名								
內容	漢語意譯	超度什羅亡靈儀式。送什羅祭司亡靈到格補命在的地方經。						
	譯著者意譯	超度什羅亡靈列達二十二個象地。						
內文讀音與直譯	讀音	z" ‘	t§uš	pyÙ	na ‘ ‘tsaÙ	ts'lÙ	muš	me ‘
	直譯	祭儀	延長	祭儀	納札	建	是的	啊
內文分字讀音與解釋	影像檔名							
	讀音與字義	讀 t o • b A Ù 或 pyÙ，東巴也。	讀 §I•，肉也，這裏假借為人名，讀 §«rš。	讀 Nv•，死者亡靈木身，這裏假借為超度，讀 NvŠ。	讀 ts'oÙ，大象也。這裏製作形符，讀藏語象之音。	讀 luAÙ，牛軛也。這裏假借為人名，讀 ŽuA•。	讀 dyÙ，大地，地方也。	讀 me•，女性，母性，雌性。假借為語氣助詞「啊」，讀 meš。
備註		1/46						

115 MS-102 為傅斯年圖書館之少數民族文書編目號，此編號物件題名為納西族東巴經的〈破地獄經〉。

附錄二、傅斯年圖書館原件拍攝原則

裝訂型式		文書類原件		卷軸類原件	單張原件
拍攝流程	外觀	小於拍攝平台尺寸	**正面**：將文書置於拍攝平台尺規起始處，調整平台至書本與色彩導表等高，由正上方拍攝。 **背面**：同上。 **側面**：將平台調整為左右齊平，書本放置於平台上。取下掃描機背，將相機光圈開到最大 5.6，調整相機及原件位置，使書本在相機觀景窗上呈現 45 度角；小型色彩導表立於文書左側，使用對焦鏡在觀景窗上對焦於色彩導表，完成後將掃描機背放回，背蓋拉開，光圈調回 16。	**正面**：將卷軸置於拍攝平台尺規起始處，調整平台，使題名與色彩導表等高，由正上方拍攝，光圈調到 16，焦點對準題名處。 **背面**：同上。 **側面**：將平台調整為左右齊平，卷軸放置於平台上。取下掃描機背，將相機光圈開到最大 5.6，調整相機及原件位置，使卷軸在相機觀景窗上呈現 45 度角；小型色彩導表立於卷軸左側，使用對焦鏡在觀景窗上對焦於色彩導表，完成後將掃描機背放為，背蓋拉開，光圈調回 16。	**正面**：將單張原件置於拍攝平台尺規起始處，由正上方拍攝，焦點對準原件正中央。 **背面**：同上。
		大於拍攝平台尺寸	無。	未定（待後續測試討論）。	**正面**：以磁鐵固定於牆上，原件正下方放置色彩導表，焦點對準原件正中央。 **背面**：同上。
	本文	小於拍攝平台尺寸	書本左右兩頁分別置於拍攝平台上，水平位移平台，使書背置於左右平台之間；調整平台高低，使拍攝表面與色彩導表呈現水平。如中縫過緊，則翻開書本約 120 度拍攝單頁。翻頁後需重新調整平台，務必維持拍攝表面高度一致，以免成像面積改變。	**正面**：平台合併，原件置於平台中央靠下方尺規處拍攝。 **背面**：同上。	**正面**：單張原件之外觀即為本文，故不重複拍攝。 **背面**：不重複拍攝。
		大於拍攝平台尺寸	無。	未定（待後續測試討論）。	**正面**：不重複拍攝。 **背面**：不重複拍攝。

附錄三、電子書格式簡介

資訊科技促成數位出版的發展，閱讀載具的開發帶動數位閱讀的成長，電子書也成為數位學習中重要的資源。PDF 檔案能編排完整的圖文內容，也可嵌入聲音影片與 Flash 動態效果，既可以保護文件編排版面且內容無法被任意修改，故成為電子書早期的主要格式之一。然而由於 PDF 格式之缺點在於文字與版面為固定大小，無法依閱讀器設計自行調整閱讀版面，在不同尺寸的閱讀設備上會對閱讀有所影響。

2007 年國際數位出版論壇（IDPF）提出 EPUB 格式，其目的是希望做為電子書內容描述規範，以有效與閱讀系統溝通，使電子書便於傳播和使用。EPUB 格式的特性在於文字可依閱讀設備特性自動重新編排文字大小並調整成最適當的閱讀版面，而其缺點則在於無法像 PDF 進行精確的圖文排版，也缺乏許多功能的支援。但 EPUB 功能仍不斷改進中，不只讓電子書滿足各種閱讀需求，還能與聲音動畫結合，以期能更完美呈現各類型電子書不同的內容與樣式，現今已成為廣為使用的電子書格式。

隨著電子書載具的推陳出新，電子書的格式亦更加多樣（請詳見下表），透過不同格式所製作的電子書將呈現各種不同的風貌，未來多元且適性化的閱讀方式將指日可待。

表：電子書格式優缺點比較表 [116]

電子書格式	優點	缺點
.txt	1. 純文字檔案，格式單純。 2. 佔用資源較小，讀取速度快。	無法包含圖片和多媒體。
.pdb	Palm 上的閱讀檔案格式，本身格式適合行動閱讀。	格式推出較久，支援性受限，新書較少。
.pdf	1. 能編排完整的圖文內容，支援中文直排。 2. 可嵌入圖片、聲音、影片與 Flash 動態效果。 3. 可保護文件編排版面，內容無法被任意修改。	1. 文字與版面為固定大小，無法依閱讀器設計自行調整閱讀的螢幕版面。 2. 佔用資源大（檔案大），若包含大量圖片，則翻頁速度會大幅下降。
.azw	1.Amazon 專用格式，適合行動閱讀。 2.Amazon.con 上有大量英文電子書，市佔率高。 3. 搭配 Kindle 可支援語音功能。	1. 只有 Kindle 和 Kindle DX 可以用。 2. 目前只有英文電子書，沒有中文書。 3. 無法精確排版，不支援中文直排。
.epub	1. 可依閱讀設備特性自動重新編排文字大小，並調整成最適當的閱讀版面。 2. 開放標準格式、跨平台，流通性大。 3. 電子書廠商紛紛支援，未來發展可期。	1. 無法進行圖文的精確排版，不支援中文直排。 2. 中文 EPUB 電子書資源較少。
.lrf（BBeB）	檔案輕巧，翻頁快速。	僅 Sony Reader 支援。
.mobi、.prc	支援的裝置繁多，Kindle 亦支援。	無法精確排版，不支援中文直排。

116 Mengkuei Hsu，〈自製 ePub 電子書（上）：認識電子書格式之爭〉，T 客邦，檢索：2012 年 5 月，http://www.techbang.com.tw/posts/2467-home-epub-e-book-on。

國家圖書館出版品預行編目 (CIP) 資料

數位化工作流程指南 : 文字資料 / 王雅萍等作 .
-- 初版 . -- 臺北市 : 數位典藏拓展臺灣數位典藏計畫 , 民 101.07
面 ;　公分 . -- (數位典藏叢書 ; 5)
ISBN 978-986-03-3086-1(平裝)

1. 文獻數位化 2. 文物典藏 3. 文字處理 4. 工作說明書

028.026　　101013573

數位典藏叢書 05
數位化工作流程指南：文字資料

指導單位：行政院國家科學委員會
發 行 人：林富士
總 編 輯：邱澎生
作　者：王雅萍、張如瑩、陳秀華、蕭貴徽
執行編輯：王雅萍、張如瑩、高朗軒、陳秀華、蕭貴徽
審 稿 者：法鼓佛教學院　杜正民教授
發行單位：數位典藏與數位學習國家型科技計畫　拓展台灣數位典藏計畫
地址：115 台北市南港區研究院路二段 128 號
中央研究院歷史語言研究所
電話：886-2-2782-9555 轉 288
傳真：886-2-2786-8834
網址：http://content.teldap.tw
Email：content@gate.sinica.edu.tw
封面設計：丁錫卿
排版印刷：禾古精緻印刷有限公司

中華民國 101 年 7 月初版
ISBN 978-986-03-3086-1