國史館數位典藏資料庫查詢系統規劃與建置

張瓊月*

國史館自民國91年參與國科會數位典藏國家型科技計畫，審選「國民政府檔案」、「蔣中正總統文物」、「資源委員會檔案」、「臺灣省政府地政處檔案」與「蔣經國總統文物」等5個讀者調閱較多之全宗作為數位化的內容。本館數位典藏工作可分為系統開發、檔案文物編目建檔、影像掃描、數位典藏專屬網站規劃與建置及加值應用等方面，本文將僅就系統開發規劃建置之經過與成果作介紹。

民國91年由中央研究院計算中心、數位典藏國家型科技計畫後設資料工作組（以下簡稱「後設資料工作組」）技術支援共同開發「國民政府檔案」與「蔣中正總統文物—檔案及照片」數位典藏系統。92、93年分別完成「資源委員會檔案」及「臺灣省政府地政處檔案」數位典藏系統。

民國94年10月以前，本館的數位典藏系統均係為一個全宗開發一個著錄系統與資料庫及查詢系統，因此有4個全宗著錄系統及其相對應的4個查詢系統。在此4個查詢系統中，僅有進階查詢的功能較健全，且只有「國民政府檔案查詢系統」的「不分欄位全文檢索」功能是可以正常執行的，符合ＥＡＤ階層架構的瀏覽查詢功能則付之闕如。

* 國史館審編處科員

有鑑於資料分散的查詢系統不利於讀者查找檢索所需史料目錄，且無法提供簡便的查詢功能，因此本館於92年即開始著手規劃整合式資料庫與跨全宗查詢系統，93年2月份數位典藏成果展時，在中研院計算中心的協助下完成了符合檔案層級架構的階層化瀏覽查詢介面，但簡易與進階查詢介面仍無法提供跨欄位、跨全宗的查詢功能。直到94年底，始完成著錄系統與資料庫整合工作，95年2月份數位典藏成果展時，「國史館數位典藏資料庫查詢系統」正式上線提供跨全宗、跨欄位並可階層化瀏覽的查詢服務。

壹、國史館數位典藏系統與資料庫開發建置流程

以下就數位典藏系統與資料庫開發建置流程做說明：

一、數位典藏系統與資料庫開發建置流程

(一)檔案文物內涵分析及後設資料（metadata）需求表撰述

本館針對檔案文物類型、分類系統、內涵屬性、外觀形式、各全宗檔案機關之組織沿革及執掌業務等進行分析，配合內容描述、編目建檔、使用權限設限及影像掃描等實務作業，尊重保存檔案文物之原有層次與架構關係，提出後設資料（metadata）欄位需求表單。[1]

後設資料工作組規劃設計的「後設資料（metadata）欄位需求表單」，[2]所需填寫的表單計有「metadata需求確認表單」、「藏品單元（unit）層級關係圖」、「metadata藏品元素需求表單」、「metadata元素代碼表單」、「metadata著錄範例表單」、「metadata系統功能需

1 黃淑瑛：〈資源委員會檔案整編與數位典藏系統建置述要〉，《國史館館刊》，復刊第37期（民國93年12月），頁169。

2 後設資料作業表單與填寫範例，可參考後設資料工作組網頁實務規劃單元，網址http://www.sinica.edu.tw/~metadata/design/design-frame.html。

求表單」等，供該組再次評估和分析資料內涵，同時建議採用國際檔案描述標準及製作後設資料需求規格書之用。

(二)研擬、確認後設資料（metadata）需求規格書

藉由參訪同質性之典藏機構，深入了解其採用後設資料標準之類型、研發方法及系統管理機制，請後設資料工作組提供國際間最常使用之檔案描述標準規範，經由雙方評估分析後，取得共識採用美國檔案館界及國會圖書館所通用之檔案描述編碼格式EAD（Encoded Archival Description）為藍本，研發適合本館藏品內涵特性描述格式之EAD修正版，由後設資料工作組進行需求規格書（或內涵分析報告書）的撰寫，歷經數次溝通、確認與後設資料元素的修正調整，方有正式版本產生，同時也完成本館後設資料元素與DC（Dublin Core）、MARC 21之後設資料標準比對，作為國際後設資料接軌及國際間XML資料交換之關鍵機制。

後設資料（metadata）需求規格書包括：前言、計畫簡介、系統說明、著錄欄位架構、資料結構表（說明是否為多值欄位、有無使用代碼選單、欄位型態與長度、是否為必填欄位、預設值、欄位值系統自動產生等）、後設資料比對表（與EAD及Dublin Coed等國際標準比對）、功能需求說明（系統功能說明、建檔功能說明、查詢功能說明）、代碼表（下拉式選單的內容）等。

「國民政府檔案」、「蔣中正總統文物」、「資源委員會檔案」、「臺灣省政府地政處檔案」、「蔣經國總統文物」等5個全宗後設資料（metadata）需求規格書均可詳見國史館數位典藏網站中「相關文件彙編」單元。[3]

(三)系統雛型開發、測試及維護

後設資料工作組將需求規格書交由中央研究院計算中心進行研發

3 國史館數位典藏網站—「相關文件彙編」，網址：http://dftt.drnh.gov.tw/digi_doc_big5.htm。

建置「數位典藏系統」雛形，再由本館進行功能測試與評估，進而提出相關之修正需求，裨益系統更臻完善，藉以進行檔案文物編目建檔及推展瀏覽檢索開放應用業務。[4]

(四)整合式資料庫的開發與建置

本館於94年開始積極辦理整合式資料庫的開發與建置工作，在後設資料工作組的指導下，由同仁自行完成5個全宗資料庫欄位架構與系統資訊的比對工作。比對工作需將5個全宗資料庫之中英文欄位名稱、資料型態、欄位大小等做最大聯集的整合，並針對欄位註明必填、多值、屬性、連結等功能；勾選提供使用者進階查詢之欄位（簡易查詢為不分欄位全文檢索，因此不另外標誌）；勾選查詢結果之簡要顯示款目、詳細顯示款目等。將完成比對的表單文件交由後設資料工作組確認無誤後，委請中央研究院計算中心開發「國史館典藏國家檔案與總統文物數位化中程計畫—著錄管理系統」及「國史館數位典藏資料庫查詢系統」。

貳、國史館數位典藏系統架構與功能

國史館數位典藏系統包含「國史館典藏國家檔案與總統文物數位化中程計畫—著錄管理系統」、「國史館數位典藏資料庫查詢系統」及影像圖檔瀏覽系統。此三個子系統均是利用道邇資料庫製作工具（DORE）搭配MySQL資料庫管理系統開發而成的，為使查詢系統功能更為健全，中研院計算中心選用Lucene為全文檢索引擎的架構開發可全文檢索的查詢系統。

4 黃淑瑛：〈資源委員會檔案整編與數位典藏系統建置述要〉，《國史館館刊》，復刊第37期，頁169。

一、國史館數位典藏系統架構

(一)道遜資料庫製作工具（DORE）[5]

DORE是資料庫應用程式的快速開發工具。可讓程式人員以更為直覺、概念化的方式，自資料庫綱要出發，藉由WEB網頁的驅動，追求雙重的效益：(1)程式師能夠按照需求，不偷斤減兩地快速製作資料庫應用系統；(2)使用者享有足夠的功能及操作便利。DORE支援以下四種常用網頁的運作：

1.管理網頁：以填表方式執行各筆資料的新增、修改、刪除及查詢。

2.查詢網頁：設定檢索條件，實施填表查詢（Query by example）。

3.條列網頁：查詢結果逐條分頁表列。

4.報表網頁：展現各筆查詢結果的詳細資料。

DORE在民國91年中漸趨成熟，學術研究用資料庫系統，多半以DORE快速重製。DORE所使用的作業系統為Linux，語言界面為繁體中文，Web Application Server是Apache，程式語言包含PHP、SQL、Java Script，資料庫管理系統為MySQL。

(二)全文檢索（Full-text Search）

由於本館目錄資料量龐大、欄位結構多，並且需要透過網路進行多人同時且快速的查詢，因此無法利用資料庫管理系統內建的查詢語法進行資料庫內容的查詢與檢索作業，而需要建置「全文檢索引擎」。全文檢索需事先對資料建立索引，全文索引與書籍後面所附的索引頁的資料類似，協助我們很快的可以找到該書的一些重要字詞。因此全文檢索引擎可以根據設定，將資料庫的內容事先掃描一遍，並

5 中央研究院計算中心http://www.ascc.sinica.edu.tw/center/（檢索日期：2006/10/25）。

進行全文演算法處理，於資料庫外部建置索引檔，或稱為索引資料庫，以便使用者進行搜尋時可以在最短的時間內得到答案，而且搜尋時與後端的資料庫是分開的，不會增加後端資料庫的任何負載，讓資料庫可以更專心於處理應用程式的資料匯入、移轉、檢調等作業，也讓資料庫可保持更高、更穩定的效能。[6]中央研究院計算中心所採用並運用在國史館數位典藏資料庫查詢系統的全文檢索引擎為Lucene。

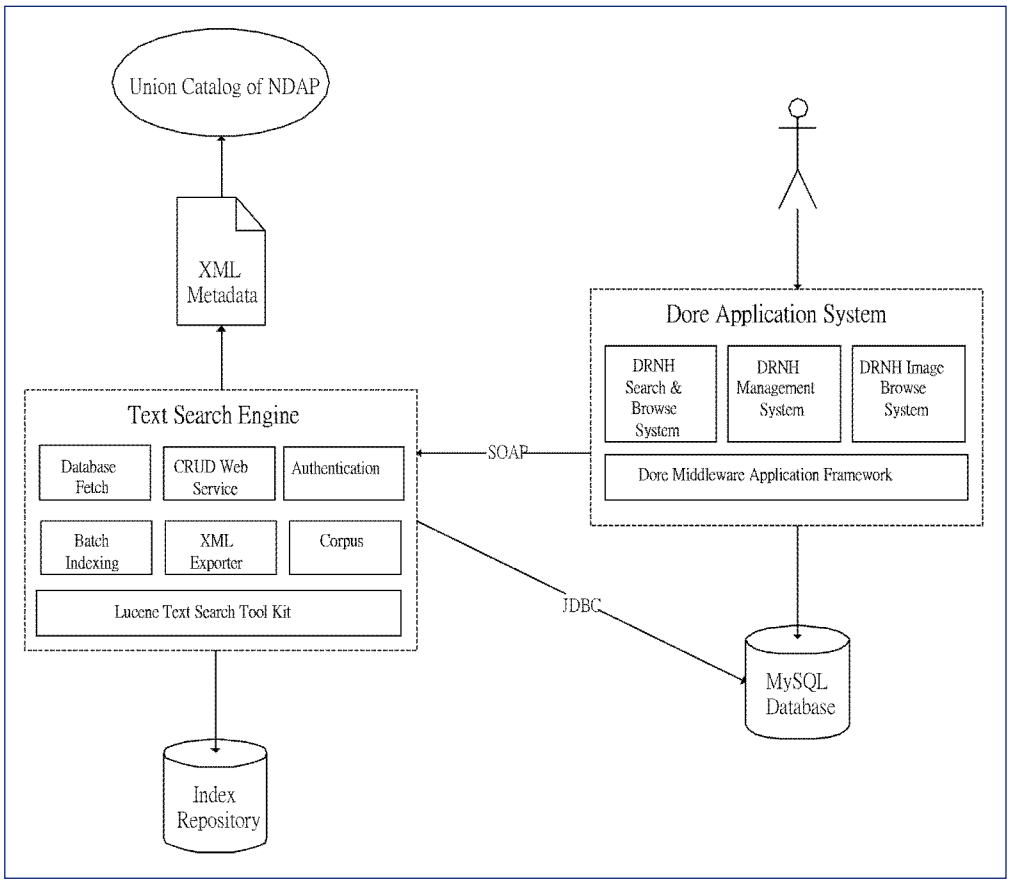
Lucene是Apache軟體基金會Jakarta項目組的一個子項目，是一個開放原始碼的全文檢索引擎套件，它不是完整的全文檢索引擎，而是一個全文檢索引擎的架構，提供了完整的查詢引擎和索引引擎及部分文本分析引擎。Lucene的目的是為軟體開發人員提供一個簡單易用的工具包，以方便在目標系統中實現全文檢索的功能，或者是以此為基礎建立起完整的全文檢索引擎。

國史館數位典藏資料庫查詢系統建構在Lucene全文檢索引擎套件的基礎上，Authentication模組提供驗證與授權的功能，為索引的建置與查詢提供權限控管的機制；Web Service模組透過Web Service提供CRUD的介面供外部系統操作索引檔；為了滿足使用者在繁體中文環境下資料檢索的需求，本系統採用正向最大匹配（Forward Maximum Match, FMM）演算法來進行中文斷詞，試圖解決使用者在資料檢索上可能遇到的中文同義詞與異體字的問題。Corpus模組提供中文斷詞功能所需之字典檔，目前字典檔的產生與維護有兩個來源：透過程式訓練建置或是整合中央研究院平衡語料庫Batch Indexing模組提供批次建置索引的功能。

6 黃國倫：〈資料庫開發與相關技術應用〉，《數位典藏專業培訓課程（八）講義》（民國95年7月），頁7。

下圖為本館數位典藏系統架構圖，由負責開發國史館數位典藏系統的中央研究院計算中心資訊技術人員廖祐宏先生繪製提供。

圖1：國史館數位典藏系統架構圖

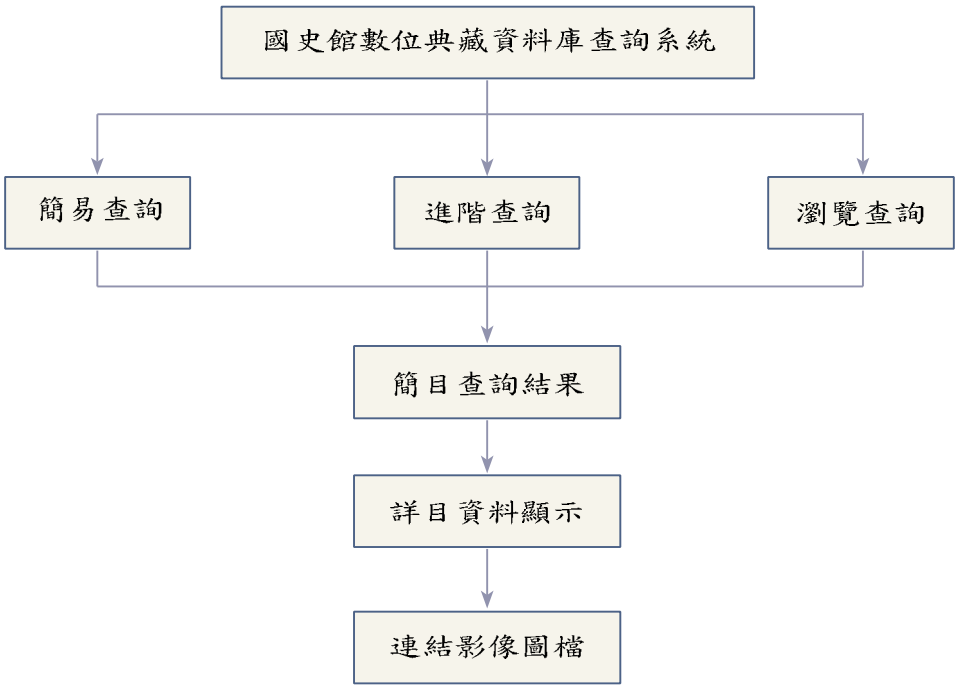


二、數位典藏資料庫查詢系統功能

國史館數位典藏資料庫查詢系統提供「簡易查詢」、「進階查詢」及「瀏覽查詢」三種功能介面，使用者可依需求選擇查詢介面。簡易查詢具有「不分欄位全文檢索」功能，使用者可鍵入欲查詢的關鍵詞，即可進行跨資料庫、跨欄位的檢索並獲得大量的相關資料。為了滿足進階使用者的更精密搜尋需求，系統亦提供進階查詢的功能，

透過多條件欄位的條件設立，讓檢索的結果更為精確，可以方便使用者篩選多餘的資料。瀏覽查詢以階層性、結構化的方式呈現數位典藏資料庫中的目錄資料，以利使用者了解本館史料的層級架構，與史料間的關連性。在查詢結果的顯示上，預設先提供簡目式的摘要顯示，當使用者已經發現真正需要的一筆或數筆資訊時，則可以經由典藏號連結進入詳細目錄的資訊，並可由詳細目錄連結影像圖檔。整體之系統介面設計是以簡單、直覺為訴求，減少使用者的適應與學習障礙。

圖2：國史館數位典藏資料庫



(一)簡易查詢

簡易查詢介面所使用的即是以Lucene為架構的全文檢索引擎，具有「不分欄位全文檢索」功能，可檢索資料庫中全部的欄

位，使用者可鍵入欲查詢的關鍵詞，執行查詢即可獲得大量的相關資料，適合初學者使用。

以檢索「蔣中正」為例，鍵入關鍵詞後，按下「查詢」鍵。

查詢結果首先顯示條列式簡目，使用者可由「典藏號」連結詳目資料。

詳目資料如右圖。使用者可利用「下一筆」按鈕，繼續瀏覽其他詳目資料或按「回目錄」回到簡目。

(二)進階查詢

進階查詢即「欄位檢索」，亦是透過Lucene全文檢索引擎進行檢索，可針對內容描述、人名、地名、時間、關鍵詞、照片尺寸、卷名、副副系列名、副系列名、系列名、副全宗名、典藏號、來源、語文等不同的欄位查詢，不同欄位間或同一欄位中可做交集（AND）、聯集（OR）、排除（NOT）等組合查詢，適合研究者篩選目錄資料，縮小查詢範圍使用。

進階查詢下拉式選單各選項說明：

1.全部：具不分欄位全文檢索功能。

2.內容描述：史料內容摘要描述。

3.人名：史料內容中重要的人名，未於內容描述欄位中著錄時，在人名資訊欄位著錄相關人名。

4.地名：史料內容中重要地名，各重要地名所屬之「省」分非常明確時，以（）註記在該地名之後。

5.時間：史料內容的相關時間，查詢時間時請用西元年查詢，可輸入「年」（yyyy）、「年-月」（yyyy-mm）、「年-月-日」（yyyy-mm-dd）進行查詢，例如：1926或1926-12或1926-12-31。

6.關鍵詞：依史料內容所擇取之關鍵詞。

7.照片尺寸：照片尺寸以英吋為單位，格式為「？X？」，即「長X寬英吋」，X為大寫。查詢時，請輸入「3X5、4X6、5X7……」等。

8.卷名：史料的題名。

9.副副系列名：史料分類層級。[7]

10.副系列名：史料分類層級。

7 關於本館史料分類層級，可參考國史館數位典藏網站http://dftt.drnh.gov.tw/digi_doc_big5.htm「相關文件彙編」之「檔案控制層級表」。

11.系列名：史料分類層級。

12.副全宗名：史料分類層級。

13.典藏號：辨識史料的唯一號碼。典藏號為全宗號、副全宗號、系列號、副系列號、副副系列號、卷號、件號等合併產生的字串，具唯一性。

14.來源：史料取得來源或移轉單位。國民政府檔案與蔣中正總統文物來源均為「總統府」；資源委員會檔案來源為「經濟部」。

15.語文：史料記載文體的語文類別，如「中文、英文」等。

以查詢「何應欽」在「1945」年的相關資料為例。下拉式選單中選擇「人名」輸入「何應欽」；選單中「時間」輸入「1945」；選單中「內容描述」輸入「何應欽」，按下「查詢」。

查詢結果首先顯示條列式簡目，使用者可由「典藏號」連結詳目資料。

詳目資料如右圖。使用者可利用「下一筆」按鈕繼續瀏覽其他詳目資料或按「回目錄」回到簡目。

(三)瀏覽查詢

瀏覽查詢介面是以檔案階層瀏覽為主要方向，在檢索結果的呈現上，於檔案各層級提供相關資料，包括全宗檔案的整理描述（全宗號、全宗名、傳記歷史註、範圍與內容、典藏單位、典藏位置），於系列層級則有該系列的描述資訊，其下再有卷等檔案的相關資訊。

瀏覽查詢以階層性、結構化的方式呈現國史館數位典藏資料庫中國民政府檔案、蔣中正總統文物、資源委員會檔案、臺灣省政府地政處等全宗的目錄資料。瀏覽查詢讓全宗的呈現如同一本書的章節，可以從樹狀結構圖中看出各全宗下系列、副系列、副副系列、卷、件的史料編排層級架構，讀者得以鳥瞰整個全宗獲取整體概念，並能清楚明白史料間的關連性。

中央研究院計算中心的程式設計人員為因應國史館龐大的目錄資料量，所設計的瀏覽查詢介面最大特點為其可大量開展的樹狀結構表，例如蔣中正總統文物檔案中的特交檔案系列下的一般資料中六百多卷的卷名亦可一次展開。

由國史館數位典藏資料庫查詢系統首頁點選「瀏覽查詢」，或由該系統各頁連結「瀏覽查詢」後，由左方之「數位典藏資料庫目錄」點選欲查詢全宗之資料夾，如國民政府檔案，可藉由層層點選展開樹狀圖，並藉由右方條列式簡目中之典藏號連結詳目。

由條例式簡目之「典藏號」進入詳目，則會顯示該筆目錄各項欄位的詳細資料。

參、結　語

由於加入數位典藏國家型科技計畫，本館在經費、人力、技術上都獲得長足有效的支援；在檔案管理作業上，也由傳統管理作業方式蛻變為科技管理的新方向，對於史料檔案查詢系統的往前邁進亦提供極大的助力。本文藉由「國史館數位典藏資料庫查詢系統」規劃與建置的經驗，為本館參與第一期數位典藏國家型科技計畫在系統開發方面的經驗與成果做一紀錄，並期望可提供國內各歷史檔案主管與典藏

單位在規劃與建置查詢系統時之參考。

數位典藏系統與資料庫的建置猶如為本館豐富的典藏史料架構出一空間配置完善、動線順暢且設備功能新穎的大樓，而本館奉命執行數位典藏計畫的審編處同仁，每天兢兢業業編目建檔的檔案描述資料及戰戰兢兢掃描檢驗的影像圖檔則為其注入人文氣息。本館91至95年數位產出成績，在編目建檔方面共有369,561筆目錄資料（597.2MB），影像圖檔掃描方面合計4,564,338頁（1,257,099.9MB），數量龐大且品質精良的目錄資料及影像圖檔，恰與數位典藏系統互相輝映、相輔相成。

運用資訊科技與網際網路使本館典藏的珍貴國家重要文化資產在保存上更具安全性與長遠性，在應用上更趨普及化與效率化，使用者可以不受時空限制地輕易上網查詢資料，便利使用者查找資料或進行研究等，在這重視讀者服務的時代裡，本館對於史料檔案查詢系統的建置極為重視，期能在現有的基礎上將系統功能、編目建檔數量與品質及影像圖檔掃描開放等均再向上提升。