以 OAI 架構設計數位典藏共同目錄系統：中央研究院發展經驗

范紀文 [1]　　王祥安 [12]　　張益嘉 [1]

[1] 中央研究院資訊科學研究所

[2] 台灣科技大學資訊工程系

{fann, sawang, eiga}@iis.sinica.edu.tw

摘要

數位典藏國家計畫由 8 個典藏機構共同執行，總計有 42 個子計畫，典藏內容分屬 12 個主題，正在使用和規劃設計中的數位典藏系統超過 90 個。各機構基於本身業務或研究需求，這些典藏系統往往採用不同的資料結構，典藏品的目錄也往往各不相同。本文探討的問題在於如何建立一套共同目錄，包括分類標準和一致的典藏品後設資料，以建立單一窗口提供分類瀏覽和檢索功能，呈現整體數位典藏國家計畫的數位化成果。讓使用者透過此項目錄服務，接觸各單位豐富的數位典藏和相關知識。

考量檢索效率、資料即時同步更新等效能指標，我們採用 OAI-PMH（Open Archives Initiative Protocol for Metadata Harvesting）作為資料交換標準規範。本研究分享以 OAI 架構為基礎，建立共同目錄的經驗。例如，在建立典藏分類標準時，我們採用整合性質相近的典藏領域的策略，逐一協助各個典藏系統，從其內部資料結構，摘取適當欄位內容，自動轉換出語意明確的 Dublin Core 後設資料。我們也提出一套儲存架構，以維護 OAI data provider 中從典藏系統轉換而來的後設資料，並提昇 service provider 的查詢效率。本文也描述我們利用共同目錄系統雛型，進行營運測試的結果和各項因應方案。

關鍵字：

資料分類架構、共同目錄、OAI-PMH

1.前言

典藏數位化是當前世界各國資訊化的發展趨勢，我國數位典藏國家型科技計畫[5]包含的層面相當的廣泛，包含自然科學與人文科學兩大類領域，總共超過 90 個以上的數位典藏系統。針對各典藏單位的需求與藏品特性所設計的典藏系統，可符合各典藏單位的需求，但各典藏系統彼此的規格差異甚大，無法從單一的典藏系統檢索到所有的藏品資料，如此，使用者無法知道其它典藏系統中是否有相關的藏品資料；而數位典藏計畫辦公室亦無法確切地掌握藏品數位化的進度。基於研究、使用與計畫管控的需求，在兼顧資料互通與合理安全使用的前提下，透過共同目錄系統將散佈在各典藏系統中的藏品資料予以彙集整合，使各機構的藏品資料都能在單一的檢索界面中被查詢，以解決資料分散在各典藏系統不易檢索的問題。

2.電子資源整合技術發展的現況

從早期圖書館界的集中式書目系統如 MELVYL[2]演變至以 Z39.50 為主的分散式檢索系統，單純的電子資源整合方式已不符需求。因此促成諸如 SFX[12]以結合 OpenURL 整合學術性資源；或是 SOAP 以 XML 為基礎，透過 HTTP 協定系統相互溝通，以避免使用分散式物件存取協定（DCOM、CORBA）造成的系統開發複雜與整合上的問題。

網路的快速發展，各種資料散佈於網路的各個角落，限制了資料的存取與流通性，以單一的技術如 Z39.50[3]著實無法處理各種資料整合所面臨的問題[1]，為解決此問題，Paul Ginsparg. Rick Luce. Herbert Van deSompel 等人於 1999 年 10 月在 Universal Preprint Service 會議中提出 OAI 的概念，並於 2001 年 1 月正式公佈 OAI-PMH 1.0 版，進一步的 OAI-PMH 2.0 版於 2002 年 6 月被釋出[4]；透過 HTTP 簡易的通訊機制與 Metadata 的技術，匯集不同數位典藏單位間 XML 格式的資料以提供資料檢索的服務。如 DSpace[7]便是 MIT Libraries 為保存數位化資料所發展的一套整合儲存管理、權限管理、內容管理、工作流程管理、檢索瀏覽等的系統，採用 OAI-PMH 做為聯合服務系統整合 7 所大學圖書館藏資料的資料整合方案，以長久保存、維護與共享數位化資料。CDS[8]則是 CERN 為整合各種學術研究資料所發展的數位化文件典藏系統。ePrints[9]則是採取封閉式的 OAI 電子資源整合方式，針對其內部的 Repository 使用 OAI-PMH 進行整合，在 Repository 與 Portal Service 間則以 Z39.50 或是 SOAP 進行整合以達到其最佳化的存取。目前以 OAI 為架構的電子資源整合計畫皆正在發展 Data/Service Provider 解決方案，如 DP9[11]係為解決 Data Provider 與 Web Crawler 整合問題所發展；OAICAT 係 OCLC 所發展，提供檔案、資料庫與 JDBC 等不同整合方式的 Data Provider；在 Service Provider 方面，由於具有商業營運價值，因此有較多相關的系統在發展，諸如 OAIster、Cyclades、Arc、Perseus 、ePrints 等皆是針對不同需求所發展出來的服務系統，根據 Michael L. Nelson[10]的研究顯示，未來服務提供者間的競爭將更為激烈並且將成倍增性的成長。

3.藏品共同目錄架構之設計

3.1 藏品資料分類目錄

對於數位典藏眾多的藏品資料而言，階層式的藏品資料分類方式非常適合用於資料的重新組織與呈現，透過層次性的資料分類結構，提供簡便的數位化藏品瀏覽環境以解決大量資料呈現與組織的問題。

以書畫藏品為例，經整理故宮博物院書畫資料的分類架構，如圖 1（左）所示，國立歷史博物館的書畫資料分類架構，如圖 1（右）所示。經比對雙方彙整後的資料分類架構，可發現以藏品類型為分類的架構中，雙方存在著較高的相似性，且不會發生衝突，因此可將雙方的分類予以合併；但在作品內容的部份，由於故宮書畫處的分類較史博館細緻，透過分類與比對的方式找出雙方共同的資料項目（如山水、人物、草蟲），接著比對其餘資料項目中，是否與已找出的共同資料項目相近似的資料項目（如人物與佛道人物），若有則再次予以合併，之後再逐步地將雙方意義相近似的分類項目找出並予以合併（如樹木、花草、其它植物則合併為植物），剩下不能合併的部份可選擇直接採用該分類項目或將其統稱為「其它」一項，如此便可形成一個雙方對外皆可使用的共同目錄架構，如圖 2 所示。

從以上的例子可知，一個適切的資料分類體系建構不易，彼此的典藏目的與觀點的不同，建立領域認同的藏品資料分類架構是項極困難的工作，必須透過不斷地與各計畫溝通，將同領域的資料分類架構進行比對彙整，產生出領域的藏品資料分類架構，再整合各領域的藏品資料分類架構，方能產生較為適切的數位典藏共同藏品資料分類架構。

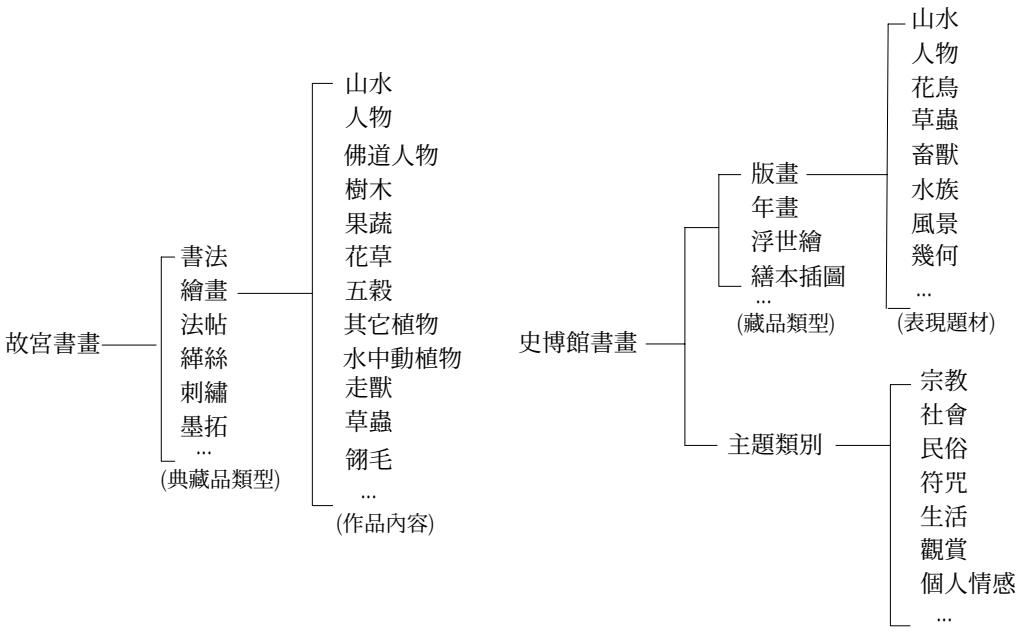

圖 1　故宮與史博館書畫藏品資料分類

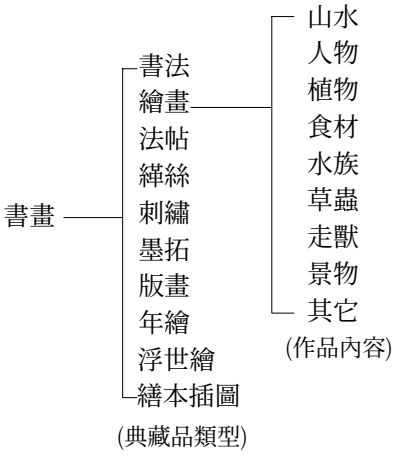

圖 2　書畫領域資料分類架構

3.2 藏品資料與分類目錄對應

藏品資料分類目錄是領域內不同藏品資料分類方式的彙整，係數位典藏共同目錄系統中藏品資料分類篩選的準則，因其異於典藏單位內部的資料分類架構，因此在訂定領域共同的分類目錄時，必需保留兩者間的對應關係，才能確保藏品資料與分類目錄能確實對應，以做為日後匯出藏品資料的參考依據。

以書畫分類目錄而言，係依據雙方藏品的分類資料為基礎（如圖 1 所示），經過一連串的歸類與合併過後產生出共同的藏品分類目錄（如圖 2 所示），在此過程所有參與分類目錄訂定的典藏單位，皆必須仔細核對每一個藏品分類項目是否確實能夠對應到自己的藏品資料

分類，及對應到哪些的資料分類項目，若有任何的項目無法對應到自己的藏品資料分類項目，則應重新檢示該項分類是否適當或是有無其它的對應方式；若確認該分類皆無任何的問題，則應將該對應關係予以記錄彙整產生如表格 1 的對應關係，以作為後續藏品資料匯出或匯入時的參考依據。

表格 1　藏品資料與分類目錄對應關係表

領域藏品資料分類目錄	故宮藏品資料分類目錄
1. 第 0 層內定為內容主題，表示係以藏品的內容為分類的依據。 2. 第 1 層內定為書畫，表示係書畫類型的藏品。 3. 第 2 層的分類係對應到故宮後設資料中的藏品類型。 4. 第 3 層的分類係對應到故宮後設資料中的作品內容下的主題代碼一。	
內容主題／書畫／繪畫／山水	山水
內容主題／書畫／繪畫／人物	人物、佛道人物
內容主題／書畫／繪畫／植物	樹木、花草、其它植物
內容主題／書畫／繪畫／食材	果蔬、五穀
內容主題／書畫／繪畫／水族	水中動物植物
內容主題／書畫／繪畫／草蟲	草蟲
內容主題／書畫／繪畫／走獸	走獸
內容主題／書畫／繪畫／景物	風景、翎毛

3.3 共同目錄資料匯入格式

要將所有的藏品資料整合到共同目錄系統實為極大的挑戰，因此必須設法降低資料轉換過程的複雜度與成本，其中以共同目錄資料系統（Data Provider）與典藏系統的關係是最為的密切，但由於 OAI 對於共同目錄資料系統（Data Provider）資料匯入的格式並無明確的規範，典藏系統可用任意的格式將藏品資料匯給共同目錄資料系統，為了處理不同格式的資料必需額外增加能夠處理該格式的介面，如此勢必造成資料轉換成本大幅的增加。因此我們透過設計一個共同的資料匯入格式，以達到降低資料轉換成本與簡化系統複雜度的目的。

基本上，此一資料匯入格式係採用 Dublin Core 做為描述藏品資料的核心項目，並加入管理性的控制資料（如表格 2 所示），如計畫描述資料（Project）元素描述此筆藏品資料屬於哪個計畫的產出，此文件係由誰或哪個系統所建立與何時被建立，其中文件建立時間係做為資料是否需要更新的判斷依據；而資料分類目錄（Catalog）元素則是記錄此筆藏品資料被分類到哪幾個的分類，以方便共同目錄資料系統建立起藏品資料與分類目錄的對應關係；數位藏品編號（DigiArchiveID）則描述此筆藏品資料的識別編號，以做為判斷此筆資料是否已經被建檔；及描述相關的藏品資料鏈結，以增加共同目錄系統資料的完整性。雖然在 Dublin Core 的 15 個元素中亦有可描述管理性資訊的項目，但由於將管理性描述資訊放置於 Dublin Core 中會與藏品內容產生混淆的情況，例如將

藏品展示網頁的鏈結與藏品圖像的鏈結放到 Dublin Core 的關連（Relation）元素內，會產生無法區別何者係藏品展示網頁鏈結的問題，尤其是將多項的藏品資料項目對應到單一的 Dublin Core 元素時尤為嚴重。為了避免產生類似的問題，因此在設計此資料匯入介面時，便將管理性描述資訊與藏品資料區分在不同的位置，以更明確定位出何者係管理性的描述資訊。

表格 2　管理性共同資料匯入格式 DTD

```
<!ELEMENT AdminDesc (Project, Catalog?, DigiArchiveID, Hyperlink?, ICON?)>

<!--計畫資料描述-->
<!ELEMENT Project (#PCDATA)>
<!ATTLIST Project
        GenDate CDATA #REQUIRED
        Creator CDATA #REQUIRED
>

<!--相對應的資料分類目錄-->
<!ELEMENT Catalog (Record+)>

<!--分類目錄,目錄與目錄之間請以":"符號分隔-->
<!ELEMENT Record (#PCDATA)>

<!--數位藏品編號-->
<!ELEMENT DigiArchiveID (#PCDATA)>

<!--數位藏品資料的超鏈結-->
<!ELEMENT Hyperlink (#PCDATA)>

<!--數位藏品小圖示的超鏈結-->
<!ELEMENT ICON (#PCDATA)>
```

4.系統設計與實作

本研究採用 OAI 架構設計共同目錄系統，以整合各計畫的藏品資料與檢索介面，並藉由 OCLC 所發展 OAICAT 之基礎，因此本研究著重於 Data Provider 與典藏系統間之整合及 Service Provider 分類瀏覽界面之設計。以下僅就這兩部份做簡要之說明。

4.1 藏品資料匯入模組

本模組（如圖 3 所示）係 Data Provider 的一部份，負責將藏品資料匯入 Data Provider，主要係由 Files、DACatalog、DBStore 與 DACatalogReader 等組成，DACatalogReader 係本模組的主體，負責整合三者之功能；Files 負責將特定目錄下的所有 XML 檔案列出，以供 DACatalog 解析 XML 檔案的內容；DACatalog 則將解析後的典藏品識別號、分類目錄資訊及相關的藏品描述資料，透過 DBStore 介面儲存到資料庫，最後 DACatalogReader 再透過 Files 介面將 XML 檔案移動到特定目錄，以供後續 Service Provider 請求該筆藏品目錄資料時取用，依此程序匯入 XML 檔案直到所有的藏品資料皆匯入為止。

典藏系統依據共同目錄資料匯入格式之規範匯出藏品資料，因使用相同的格式匯出藏品資料，簡化了藏品資料匯入模組之設計，透過

此模組可快速地將藏品資料予以解析、儲存到 Data Provider。為確保每一筆藏品資料可對應到多個的分類目錄，本研究修改 Data Provider 資料儲存之方式，將藏品基本資料、分類目錄與其對應關係儲存在資料庫，藏品資料則存放在 XML 檔案，同時運用兩者之優勢以強化 Data Provider 藏品資料管理的能力。

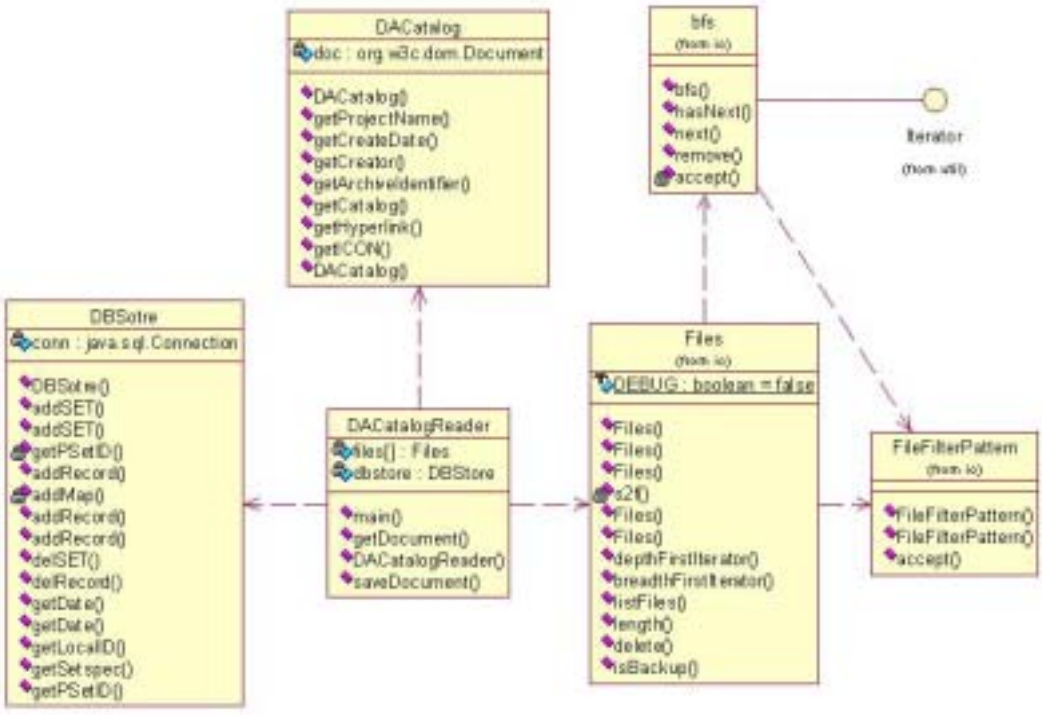

圖 3　藏品資料匯入模組類別架構圖

4.2 共同目錄檢索瀏覽模組

依據需求分析共同目錄系統至少必需具備分類目錄瀏覽、Dublin Core 檢索與全文檢索等功能，以提供跨越典藏單位的資料檢索，與瀏覽各館藏單位的藏品內容。據此設計本系統，其類別架構如圖 4 所示。Catalog 類別負責維護階層性分類目錄的完整性，透過此類別可存取本系統中任意層次的目錄資料，將 Catalog 類別所取得的分類目錄代碼，傳入 ArchiveRecord 類別中的 getRecord 方法便能取得特定目錄下的藏品資料，據此便可設計分類目錄瀏覽功能；至於檢索功能，因 Search 類別具有基本檢索的特性，可快速實作不同的檢索方法，因此可透過呼叫相對應的檢索介面，完成 Dublin Core 與全文檢索功能的設計。整個系統採用元件化角度規劃，資料庫存取、運作邏輯與視覺界面採分層設計，因此單一功能或視覺界面的改變並不會影響到其它子系統的運作，如此的設計除了可降低維護成本外，並可大幅提升系統的再用性與擴充性。參與本研究的包含考古、器物、書畫、檔案、植物等主題內容，目前各主題計畫已分別匯入 50 筆藏品資料供測試（網址：http://catalog.ndap.org.tw），未來陸續將匯入更多的藏品資料，讓使用者可輕鬆地瀏覽、檢索數位典藏中的藏品物件。

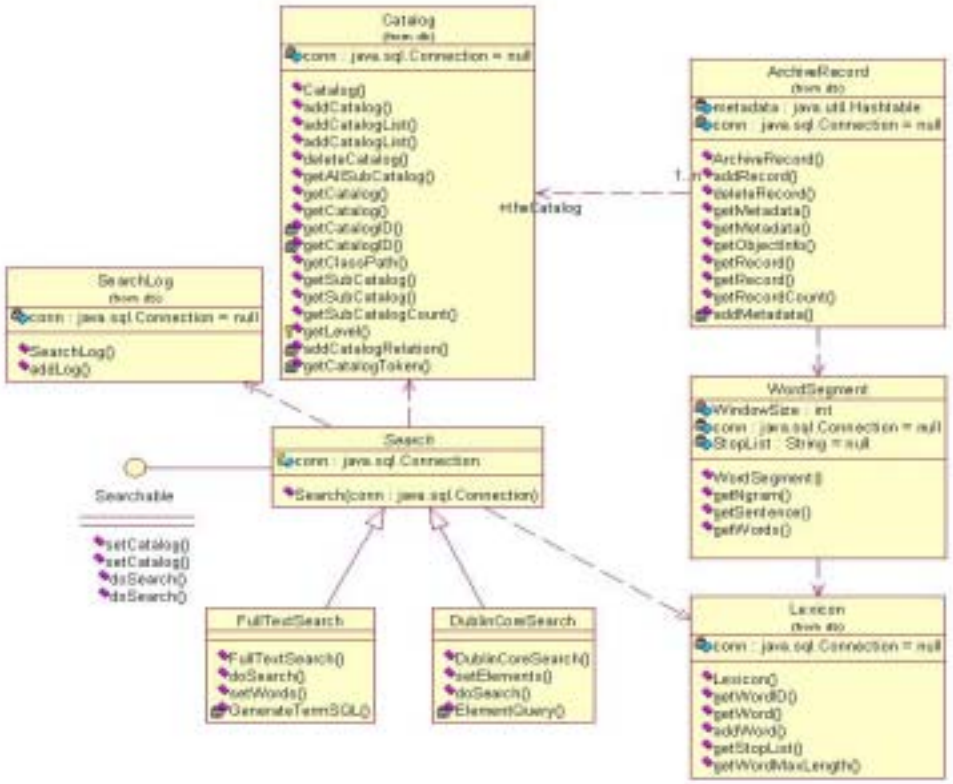

圖 4　共同目錄檢索瀏覽模組類別架構圖

5.經驗分享

隨著數位典藏相關計畫陸續完成典藏資訊系統的開發，藏品資料得到有效地管理，但因系統彼此獨立，無法進行跨系統間的藏品資料檢索，阻礙了資訊的流通。有鑑於此，本研究透過 OAI 架構的導入將不同類型的藏品資料與檢索服務整合，研究進行中除了面臨技術上的問題外，藏品資料的整合則是最大的挑戰，就本研究的經驗而言，充分地與典藏單位溝通是整合典藏資料的不二法門，以下就從幾方面分享本研究發展共同目錄系統的經驗。

(1) 藏品分類目錄的整合

因典藏觀點的不同，藏品分類目錄亦不盡相同。過去數位典藏計畫因考量人力成本的問題，第一期計畫採取訂定所有計畫皆相同的分類目錄，結果遭致各典藏計畫極大的爭議，第二期計畫雖先徵詢各典藏計畫分類目錄的意見，但因缺乏相同藏品間分類目錄的整合，造成典藏計畫不易加入共同目錄的問題。因此建議在發展共同目錄系統時，系統建置人員應扮演技術支援的角色，以協助各計畫整合藏品資料與服務為出發點，並從分類目錄的使用者觀點切入，使用前述所提的分類目錄設計方法協助整合，在不拘泥於單位內部表達形式的前提下產生共同目錄，方可避免上述問題的發生且較為符合一般使用者的需求。

(2) 藏品資料的匯出

典藏系統與 Data Provider 間必需訂定明確的資料互通格式以協助雙方的整合，此資料格

式除了需訂定嚴格的 XML 資料格式規範外，轉出資料項目的著錄格式規範亦相當的重要，否則極容易因彼此表達方式的不同，造成無法檢索或是形成不易閱讀或理解的資料。建議所有匯出資料項目的表達方式應以容易了解為首要，不應拘泥於內部的表達方式，並選擇最能表達藏品特色的資料項目轉出，如此在進行典藏系統與 Data Provider 整合時方能較為順利。

(3) Data Provider 的設計

目前 Data Provider 的設計無任何標準規範，且彼此差異甚大，由於各個典藏計畫皆需建置專屬的 Data Provider，將其設計成為一個標準的套裝軟體，可大幅降低系統的整合成本與加速整合進度。就本研究研發 Data Provider 的經驗而言，本研究建議強化 Data Provider 管理藏品資料來源的能力，使其能管理多個來源的藏品資料，與提供鏈結回到原藏品資料的功能，使 Service Provider 能辨別各個藏品資料的來源;及強化 Harvester 向 Data Provider 請求藏品記錄以進行資料更新時的處理效能，減少 Harvester 請求一些沒有變動過不需要更新 Service Provider 內藏品記錄的資料傳輸量。

(4) Service Provider 的設計

與一般 Search Engine 的定位極為相似，就本研究設計 Service Provider 的經驗而言，兩者的差別僅止於取得資料的來源與方式的不同，因此可利用既有的 Search Engine 加以調整，以提供更多元化的檢索服務，且透過 Search Engine 的 Cache 機制可達到將各典藏單位的藏品資料集中典藏的目的。

本研究未來將繼續強化本系統現有功能外，並針對安全性議題與資料交換議題深入的探討，研究 Data Provider 與 Harvester 間認證機制的設計，解決因資料模式的不同而必需個別設計藏品資料匯出模組的問題，希望透過上述做法加速數位典藏共同目錄系統環境的建置。

6.結論

國家數位典藏聯合目錄建置計畫於發展中遭遇之問題如下：

(1) 管理層面的困難度大於技術層面

由於聯合目錄涉及人文科學與自然科學兩大範疇並分屬於12個主題，其中典藏系統超過90多個，然而每個典藏系統發展的進度不一且數位化的成果不同，如何建立一個有效的管理模式與團隊，將是成功與否的關鍵。

(2) 時間與地理區兩項分類目錄建置不易

聯合目錄中包括數種藏品類型且分屬於不同時制與地理位置，如何制訂一套合適的時間與地理分類架構是有其困難度，另外藏品資料考證的詳盡程度，也會影響分類目錄的建置，例如在漢代簡牘中，每個竹簡都要學者花費相當長的時間考證其年代，才能建置完善的分類目錄。同樣地理區分類目錄也面臨相同的問題。

發展數位典藏共同目錄系統最困難的部份，在於整合不同領域的藏品資料，因每個領域有特有的知識表達方式，要取得共識極為不易，因此發展數位典藏共同目錄系統應首重藏品資料的整合，充分地與典藏單位溝通是進行本研究的關鍵，唯有深入各個典藏計畫了解藏品的特性與內涵，從領域知識的角度，協助各典藏單位與領域進行資料分類目錄的分析與建立，同時以資訊技術為輔，協助解決各單位在藏品資料匯出時所發生的問題，方能發展出各單位皆能認同的數位典藏共同目錄系統。

誌謝

1. 行政院國家科學委員會，數位典藏國家型科技計畫—技術研發分項計畫，計畫編號：NSC93-2422-H-001-0003

2. 行政院國家科學委員會，數位典藏國家型科技計畫—技術研發分項計畫—典藏系統建置與相關技術研發計畫, 計畫編號：NSC 93-2422-H-001-0004
3. 數位典藏技術發展組(DAAL)
4. 中央研究院資訊科學研究所－電腦系統與通訊實驗室

參考文獻

[1] A. Paepcke, C. Chang, H. Garcia-Molina, and T. Winograd. “Interoperability for digital libraries worldwide. Special Issue on Digital Libraries”, Communications of the ACM, 41(4), April 1998.

[2] Karen Coyle. “The Virtual Union Catalog: A Comparative Study” D-Lib Magazine Vol. 6, no. 3. March 2000.

[3] Maintenance Agency page for International Standard Z39.50, http://lcweb.loc.gov/z3950/agency/.

[4] Lagoze, C., Van de Sompel, H., Nelson, M., and Warner, S.“The Open Archives Initiative Protocol for Metadata Harvesting”, 2002. Available at: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[5] Introduction of National Digital Archive Program, http://www.ndap.org.tw/.

[6] Chi-Wen Fann, J.M. Ho, D.T. Lee, “Study Metadata Design and Standardize Problem From Digital Archive Data Exchange Viewpoint”,新世紀圖書館與數位博物館趨勢研討會, 2001.11。

[7] MacKenzie Smith, “MIT DSpace”, 2nd Workshop on the Open Archives, October 2002.

[8] Jean-Yves Le Meur, “Building OAI repository with the CERN Document Server Software”, 2nd Workshop on the Open Archives, October 2002.

[9] Stephen Pinfield, “Open Archiving in U.K. Universities”, 2nd Workshop on the Open Archives, October 2002.

[10] Michael L. Nelson, “Service Providers: Future Perspectives”, 2nd Workshop on the Open Archives, October 2002.

[11] Xiaoming Liu, “DP9 Service Provider for Web Crawlers”, D-Lib Magzine, December,2001.

[12] The Ex Libris group – SFX Overview, http://www.exlibrisgroup.com/sfx.htm