數位典藏國家型科技計畫 OAI-based 聯合目錄建置規畫

陳昭珍
國立臺灣師範大學圖書資訊學研究所副教授
cc4073@cc.ntnu.edu.tw

陳雪華
國立臺灣大學圖書資訊學系教授
sherry@ccms.ntu.edu.tw

陳亞寧
中央研究院計算機中心組長
arthur@sinica.edu.tw

摘　要

建置聯合目錄時，除需考慮應如何有效地擷取詮釋資料(metadata)外，還需考慮的數位物件命名原則及詮釋資料與數位物件的正確性連結問題。本文主要在說明說明選擇 OAI 做為國家數位典藏聯合目錄建置機制的原因，以及國家數位典藏聯合目錄系統之功能需求與系統架構。

關鍵字：國家數位典藏聯合目錄、開放典藏計畫詮釋資料擷取協定、命名原則、數位物件連結管理系統、互通性、詮釋資料

Keywords： National Digital Archives Program、Open Archives Initiative (OAI)、Naming、Handle System、Interoperability、Metadata

Abstract

To design the union catalog of digital libraries, in addition to the efficient mechanism of harvesting metadata, we must consider the digital objects naming principle and the persistent connection method between metadata and the digital objects. In this article, we define the functions and the system architecture of the OAI-based union catalog and explain the reason why the OAI (Open Archives Initiative) protocol was chosen as the mechanism to create the Union Catalog of National Digital Archives Program.

一、前言

數位典藏國家型計畫是臺灣政府提倡知識經濟之後，開始重視數位內容的重大施政。目前數位典藏國家型計畫參與的單位有中央研究院、國立臺灣大學、國家圖書館、國立故宮博物院、國立科學博物館、國立歷史博物館、國史館、臺灣省文獻會、文化建設委員會等單位及數十個學位團體。這些單位所建置的數位資訊系統應如何分享、如何讓使用者從一個介面檢索到所有典藏機

構的資料、以及如何讓民眾看到數位典藏的全貌，是一個非常重要的問題。

要分享各單位所建置的數位資源，首先，聯合目錄的建置是當務之急；其次如何透過metadata將全文、影像、聲音、視訊等數位物件展現出來，是隨之而來的問題。聯合目錄的建置模式可分為兩種：一種是集中式的聯合目錄，另一種為分散式查尋的虛擬聯合目錄，集中式聯合目錄的優點是使用者檢索起來效能佳，但其缺點是建置成本昂貴。[1]虛擬聯合目錄的優點是建置成本低廉，但是檢索效能差。為保留上述集中式聯合目錄及虛擬聯合目錄的優點，且避免其缺點，在數位時代，一個新的分散擷取metadata的協定-- OAI就應運而生了。OAI詮釋資料擷取協定（Open Archives Initiative Protocol for Metadata Harvesting），是於 2001 年 1 月，由開放性資料庫發展協會（Open Archives Initiative，OAI）所發展的協定，它提供一個簡單的自動、批次、分散擷取不同機構資料庫之詮釋資料、及建立集中式聯合目錄的解決方案。

二、OAI 協定簡介

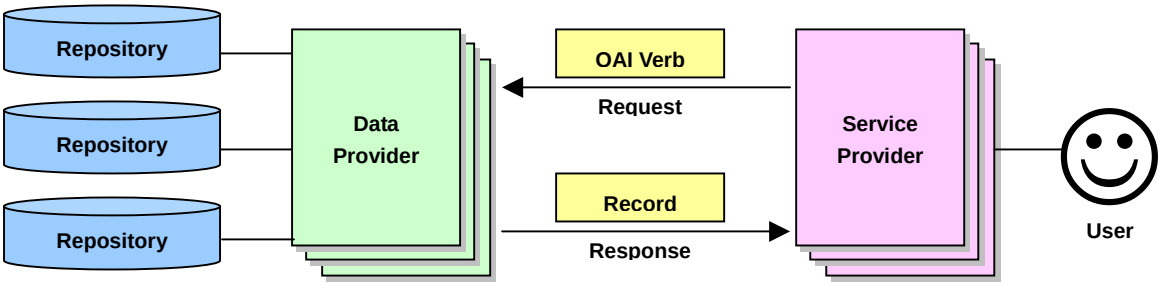
開放典藏計畫(Open Archives Initiative ，簡稱OAI)，最初是由Paul Ginsparg, Rick Luce, Herbert Van de Sompel等人，在 1999 年 10 月於Santa Fe的Universal Preprint Service會議中所促成的。有鑒於各資料庫系統與系統之間，彼此互不隸屬，相關資料、或不同領域之資料，分散儲存而難以統整，使得資料的流通有所限制而未臻完善。該會議中與會代表認為有必要對於學術性之電子期刊之預刊本（electronic pre-print）及相關數位典藏，發展出一套可以互通（interoperability）的標準架構，因此成立OAI計畫[2]。

OAI 計畫最初目的是為學術性電子期刊預印本之互通性檢索而成立，但數位圖書館所遇到的互通性檢索問題與此相似，所以 2000 年上半年，OAI 計畫便將其適用範圍擴展至數位圖書館領域，由 Digital Library Federation 與 Andrew W. Mellon Foundation 在 Harvard University 所舉行的 Cambridge Meeting 中，討論如何將數位圖書館之館藏資訊散播到網路搜尋引擎上。會中代表們（包括圖書館及博物館界）一致認為：使用互通的方式傳佈詮釋資料（metadata），將是重要的關鍵。

2001 年 1 月，OAI發表了名為開放典藏計畫詮釋資料擷取協定(Open Archives Initiative Protocol for Metadata Harvesting)的網路通訊協定，提供網路資源互通搜尋一個可行的解決方案。OAI詮釋資料擷取協定運用網際網路（Internet）及詮釋資料（metadata）兩種技術，在增強功能及簡化實行難度上，有很好的效果。[3]。

OAI是一個簡單、容易設計程式的

協定，主要乃在透過指定的命令集，提供前端向後端儲存器提取所需資訊的協定，OAI元件主要分為OAI Service Provider與Data Provider。以OAI協定為基礎的聯合目錄架構，主要是由OAI的service provider（服務系統)定期向data provider（各資料庫系統)抓取metadata資料，建立集中式之聯合目錄。由於OAI是架構在HTTP之上的應用協定，因此其命令集即是透過HTTP所使用前端與後端傳輸之變數名稱與其內容，觸發後端對應之伺服器程式，依據變數內容處理後傳回之結果，並須遵照OAI協定XML Schema所規範的XML格式。Service Provider主要在集中維護從各系統擷取來之metadata，並將獲得的metadata在其上建立加值服務；Data Provider主要在維護資料倉儲（repositories），並且支援OAI協議來取得其倉儲內的資料內容。其關係以圖一表示：[4]



圖一：OAI技術架構組成元件

三、Metadata與數位物件連結管理問題

傳統的聯合目錄並不須處理metadata與數位資源的連結問題，但在數位時代，找到有何相關的書目資料還不夠，更重要的是須將數位檔案連結出來，這才是真正的數位典藏系統的聯合目錄。在Web環境中，這樣的連結主要告URL機制，雖然URL是一種容易實作及富彈性的機制，但也是是一種不穩定，不可靠的連結機制，因為URL是由hostname、path、filename加上取用這個檔案的通訊協定（http、ftp、gopher等）所組成，一旦主機移位、貯存路徑或檔案名稱變更時，URL就無法正確定位；此外，由於資料的實體貯存空間有限，在資料量不斷成長的情況下，當貯存空間飽和時，一定會產生將資料轉移到其他伺服器的需求。我們在瀏覽網頁時，常會出現HTTP v1.0/1.1 Error404訊息，這表示所點選欲連結的的資源已經被移除，以致無法更進一步利用。因此，網路資源的連結機制一直受到重視，尤其當數位圖書館、數位博物館、數位典藏等成為重要的研究與發展方向之後，在相關的計畫中，一定會處理此一問題。而數位資源的連結，一般而言，涉及兩個問題，一個是數位資源的命名，另一個是由資源名稱連結到儲存位置的系統。資源的命名在網路環境中為何會成為個重要的問題，除了

上述原因外，從下列幾個角度來看，數位資源的命名確為欲長久典藏資訊的數位資訊系統中必需處理的問題：[5]

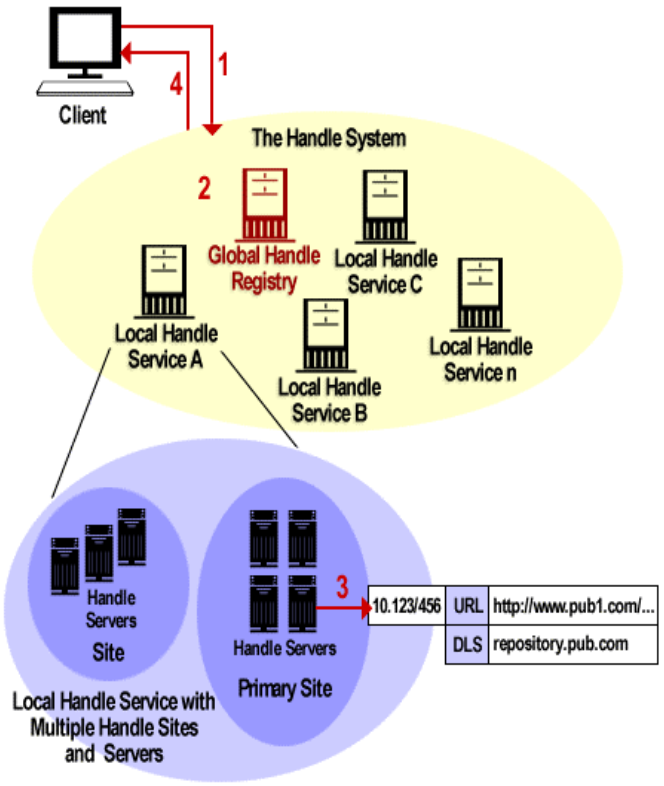
1. 從數位資源儲存管理的角度而言，命名問題也非常重要。如在一機構內，進行資料數位化的單位可能不只一個，各單位的檔案，各種類型都有，如果各單位對於數位資源名稱不能事先協調，而有各自的命名方式，則可能產生重複、系統不一致等問題，時間久了之後，將難以辨識，硬體資源也不易有效分配。
2. 此外，當一個區域或一個國家，數位資源日漸豐富後，民眾將不再滿足於分別連結到各個資料庫查尋的服務，而希望有一整合性的使用方法。此時，可能需將相同領域的資源予以集中，而若各機構事先有一致的命名方式，也會減少重複、難以辨識、不一致、書目資料與數位資源連結不易等困擾。
3. 再者，從電子商務的角度視之，也需要賦予數位資源唯一的識別碼。以目前圖書、期刊或錄音錄影資料為例，若非有 ISBN、ISSN 及 ISRC 等號碼，則圖書、期刊或錄音錄影資料的國際銷售將會困難重重。而今，當要銷售的資料單元是書中的一個章節、一個圖、一個表而不是整本書，則其編碼方式也需細到章、節、圖、表才行。

若網路資源能夠提供永久性的(persistent)命名方式，不因外界貯存體的變更而改變，則在大量的網路資源應用時會方便許多。圖書館書架上的書每一本都有一個索書號，索書號是跟據分類號及作者號賦予，而不是跟據書架的架位來取號，每一本書的索書號都是唯一的，此索書號是書目記錄與書架上的圖書之鍊結點，讀者查到書目記錄後，可藉此索書號，在書架上找到書。如果圖書移架了，圖書館員也不用去更改書目記錄上的索書號，因為此索書號所表達的是資料的相對位置(relative location)，而非絕對位置。若網路資源所記載的是這種具有唯一性的辨識碼(identifier)，或稱為檔案名稱(file name)，而不是絕對位置，也可讓資源管理者不必因儲存數位資源的電腦改變，而須更改詮釋資料(metadata)上的檔案名稱及路徑。此外，數位資訊系統通常不會是一封閉型的連結系統，也不會是一個單一的資料庫系統，檢索數位資源的途徑也不會只有一個，因此數位資源必需有一個唯一的名稱，讓建檔者及使用者可以更便利建立連結或存取資料。

如上述所言，在 metadata 中，若記錄的是全文或多媒體檔案的 URL，則當主機改名或全文及多媒體資料移動了位置，就無法由查尋到的 metadata 連到該數位檔案。因此，一般而言，metadata 中所記錄的通常為該數位檔的檔名(或稱為識別碼)，而非實際儲存的位置，再透過數位資源的命名與解析管理系統(或稱為識別碼解析器)來管理檔案名稱及其相對應的 URL 資料，這就是所謂的 resolution 或是 Handle system，這是數位圖書館聯合目錄系統中不可或缺的一環。

Handle System 解析識別碼的方式是以分散式的模型為基礎，而客戶端方

面則有幾種可能：可以是具有這方面功能的客戶端、或是一般具有特殊外掛程式（plug-in）的 Web 瀏覽器、抑或是一個透過特定代理伺服器（proxy）的一般客戶端。不論是以上哪一種情形，這些客戶端都會透過 Handle System 所定義的通訊協定來與 Handle System 溝通，而這些通訊協定都是已經經過正式定義，甚至有些已經實作出來了。解析的流程步驟，可以由下圖來簡要表示之：



如上圖所示，我們可以將識別碼解析的程序，分為以下四個步驟（即上圖所標示的 1~4）來完成：

1. 某個客戶端（如 Web 瀏覽器）遇到了解析識別碼的需求（如 10.123/456），通常這樣的情形會發生在某個特殊的超連結，抑或是某種特別的連結參考情形下。不論是在網際網路上，或是在個別的內部網路中，這個客戶端都會針對 Handle System 送出解析這個識別碼的需求。而如上節所述，這項工作可以由客戶端自己直接完成，也可以透過特定代理伺服器（proxy）來執行。

2. 在 Handle System 中，包含著由識別碼服務所形成的集合，每項服務則包含了至少一個主要的主機群以及一定數量的次要主機群，而每個主機群也都是由一定數量的識別碼伺服器所構成。就解析的程序來看，每一個主機群都複製了隸屬於該項識別碼服務的所有識別碼。而其中一個特別的服務，也就是所謂的全域性識別碼註冊資料庫 GHR，則是負擔著指引出各項區域性命名服務的重責大任。每個區域性的服務也會知道如何去存取 GHR。這樣的關係允許識別碼解析請求在任何一處發生，同時也可以順利的將請求引導到負責該項識別碼解析的特定服務。

3&4. 每一個識別碼都會關聯到一個或多個具有特定型態的資料值。在本例中，識別碼 10.123/456 就會關聯到兩種資料值，其中一個是 URL 型態的資料，另一個則是一種新的協定叫做 DLS，這樣的資料最後會回傳到客戶端供其參考。請注意，這所謂多個具有特定型態的資料值，也有可能會是同一種資料型態，如多個 URL 等等。也就是說，一個單一的識別碼資料可以對應到多個 URL 資料值，以供後續的運用。但是在此同時，Handle System 並不會對這些資料的後續使用做任何的假設，也就是說，這些資料的運用完全取決於客戶

端的需求。在這樣的情況下，便可以允許應用程式擁有最大的彈性來利用 Handle System 所提供的命名服務。在本例中，該客戶端可以使用資料值得通訊協定來定位該項網路資源，當然了，這些都完全取決於客戶端的使用目的。

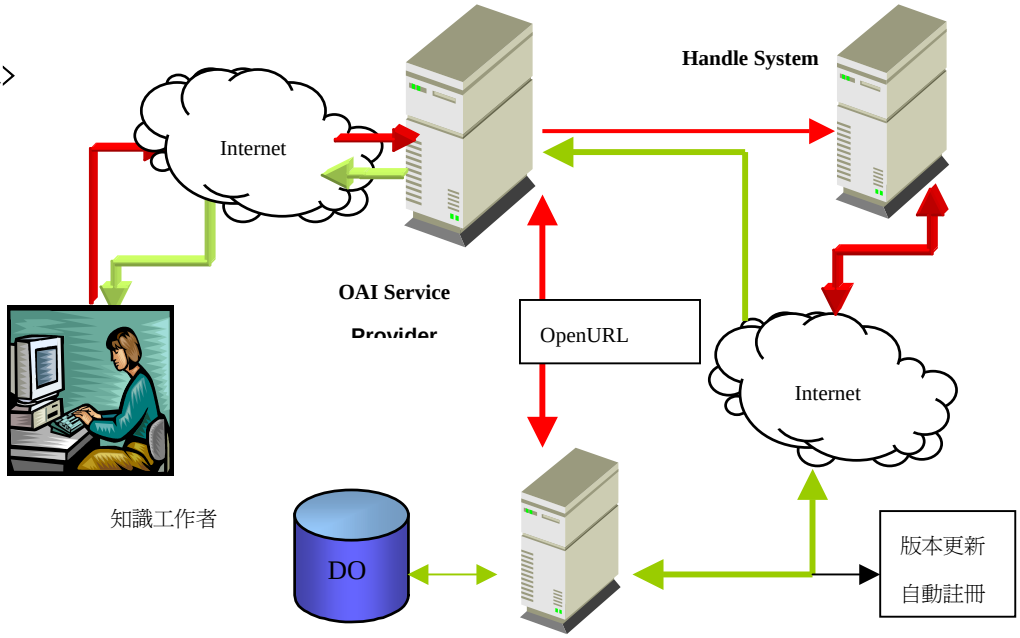
四、國家數位典藏 OAI-based 聯合目錄系統功能

為建立數位典藏國家型計畫之聯合目錄，數位典藏計畫已邀請國內參與數位典藏機構之代表成立 OAI test-bed 小組，以 OAI 及 handle system 技術建構國家數位典藏之聯合目錄，初期並將由台大、中研院、國立故宮博物院、自然科學博物館等機構參與測試計畫。OAI 是一個簡單、容易設計程式的協定，但是在實際的聯合目錄系統設計上，尚有一些是 OAI 未考慮到的問題，如與各單位之資料庫應如何連結、如何透過詮釋資料擷取數位物件、資料服務端之介面應如何設計等，都是 OAI 未訂定，也是無法訂定的部份，但在實際環境中，則是一定要處理的問題。以下即介紹我們為國家聯合目錄我定義的系統供能及系統架構，包括：整體性功能、資料服務端 (Service provider)、資料供應端 (Data provider)、Handle system、系統註冊、系統測試等。主要系統架構如圖三所示，聯合目錄的主要功能包括：

- **Metadata 擷取及維護功能**

1. 定期向各data provider抓取metadata資料，以更新聯合目錄中之metadata，抓取條件包括：set、date、identifier、metadata類型等。更新頻率可以參數設定。

<圖三>



2. 具備讓 data provider 登錄功能(register)，並以親和之介面維護基本資料，及上載 metadata、DTD、及 schema。
3. Service provider 為 Web services 之服務系統，提供使用者查尋、瀏覽功能。
4. Service provider 可儲存維護各典藏單位較詳細之 metadata，並具有 metadata cross mapping 之功能。

- **瀏覽功能**

5. 可設定將 metadata 中，典藏單位、作品類型、創作者、主題、時代、地點、典藏單位等欄位資料另外擷取出來，做為瀏覽點選功能之主題目錄。
6. 目錄編製功能應模組化，並可分辨已處理及未處理的資料。
7. 可將各單位之影像圖檔，以 thumbnail 存檔，作成圖示索引，讓使用者以作品類型、創作者、主題、時代、地點等點後，將點選結果以小圖示呈現，再連到各典藏機構，呼叫大圖。

- **查詢功能**

8. 系統能提供欄位檢索、模糊檢索(含同音、同義、簡繁俗體等)及不限欄位的全文檢索功能。
9. 共通的查尋模式包括，簡單查詢及進階查詢，進階查詢欄位包括：作品名稱、作品類型、創作者、主題、時代、地點、典藏單位等。
10. 系統具布林邏輯查詢、切截查詢(truncate)、完整查詢之功能，並以年代、資料類型、地點、檔案格式、館藏單位等條件縮小查詢範圍，限制條件盡量以點選方式處理，不需使用者鍵入資料。
11. 除整合檢索之共通介面外，使用者可選擇只要查尋某幾個單位之資料。查尋後之資料以共通之欄位呈現，並可設定是否要連結更豐富的原始 metadata。
12. 系統可讓使用者選擇呈現資料的排序方式，並可前後翻頁顯示。可選擇排序方式，含相關度、主題、創作者、時代、典藏單位等。此外可選擇以文字清單呈現或圖示清單呈現。

- **全文、多媒體連結管理功能**

13. 系統應提供正確的全文或多媒體連結、呈現或播放。連結機制包括：透過 handle system 轉為 URL 之連結。

- **權限及系統管理功能**

14. Service provider 具權限設定功能，可依據 service provider 系統管理者、目錄維護者、一般使用者，及 data provider 登錄編輯等，設定權限。

- **擴展功能**

15. 此 Service provider 也可擔任 data provider 角色，提供 Dublin Core 格式記錄，與國內外其他機構進行互通。
16. 具有 CCCII，big-5，UTF-8 轉換功能，可將各單位送來之資料轉為 UTF-8。
17. 首頁除了查詢與瀏覽外，尚可整合查詢 Internet 上的網

站，具入口網站功能。

五、結語

OAI-based 聯合目錄目前國外已有很多數位圖書館、博物館系統支援 OAI 服務，如法國國家圖書館、大英圖書館、UKOLN、美國 Digital Library Federation、UC Berkeley 等，所以 OAI-based 聯合目錄將是我我國數位典藏計畫與國外系統互通與合作之基礎，而多語言功能也是我們未來發展聯合目錄系統應重視的問題。

---

[1] 陳昭珍。電子圖書館的整合檢索理論與實務。台北：文華，民 89

[2] Carl Lagoze and Herbert Van de Sompel (2001). "The Open Archives Initiative: Building a low-barrier interoperability framework." Proceedings of the ACM/IEEE Joint Conference on Digtial Libraries, Roanoke VA, June 24-28, 2001, pp. 54-62. WWW= http://www.cs.cornell.edu/lagoze/papers/oai-jcdl.pdf

[3] Herbert Van de Sompel and Carl Lagoze (2000). "The Santa Fe Convention of the Open Archives Initiative." D-Lib Magazine 6(2). WWW= http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

[4] 國家圖書館，「研擬圖書資訊相關技術規範座談會會議資料」。台北市：國家圖書館，民 90 年。頁 3-1。

[5] 陳昭珍，陳立原，張文熙，「數位化檔案命名原則」，國家圖書館館訊，90卷3期 (民國90年8月)，頁1-5

感謝中研院陳淑君小姐提供意見及本計畫所有研究助理：蘇國盛、劉明杰、歐陽慧、張陳基、潘欣榮、顏學勇、陳昭汝、洪筱盈、張懷文、徐代昕等的參與。