Metadata與OAI-PMH在新聞數位典藏之整合應用

林信成

副教授
淡江大學資訊與圖書館學研究所
E-Mail: sclin@mail.tku.edu.tw

康珮熏

研究生
淡江大學資訊與圖書館學研究所
E-Mail: 693070038@s93.tku.edu.tw

摘要

本文探討所建置的台灣棒球新聞資料庫與數位典藏聯合目錄的後設資料整合問題。首先，我們設計一個後設資料轉換系統，可轉出DC、NITF、RSS和DAC等四種國內外普遍盛行之後設資料格式；本系統不僅能支援後設資料自動轉換，也能衍生出RSS訂閱和聯合目錄大批匯出等應用。其次，基於OAI-PMH架構，我們以此新聞資料庫作為Repository，建置符合OAI-PMH協定之Data Provider。它可供包含數位典藏聯合目錄在內的任何Service Provider查詢，然後進行命令集剖析，再以特定的後設資料格式回應資料集給Service Provider。就實驗結果而言，本系統可有效促進數位化新聞的傳播與分享。

關鍵詞：數位典藏，後設資料，標示語言，聯合目錄，OAI-PMH協定

前　言

我國由國科會推動的「數位典藏國家型科技計畫」(National Digital Archives Program，簡稱NDAP)(註1)，為彰顯新聞數位典藏之重要性，特別成立了新聞主題小組(註2)，旨在「將新聞媒體裡的新聞事件5W1H加以分析，並編制新聞標題索引目錄，建立新聞資料的加值利用…」(註3)。歷年來與新聞主題相關的計畫有：「北平世界日報內容數位化開發計畫」(註4)、「國家圖書館期刊報紙典藏數位化計畫」(註5)、「電視新聞多媒體資料館」(註6)、「蘭嶼原住民媒體資料庫建置與數位典藏計畫」(註7)，以及由我們所提的「台灣棒球運動珍貴新聞檔案數位資料館之建置」(註8)等。

傳統上，報紙數位典藏系統大多採取將所有新聞內容納入資料庫的作法，其優點是涵蓋面較廣，可以收錄各式各樣的主題；缺點則是資料庫龐大，要進行深度加值不易。因此，在我們先前的研究中，提出以「單一主題」進行新聞數位典藏與加值利用是較為可行的方式，並實際以聯合報的台灣棒球運動歷史性新聞為例，先進行系統分析與建置，再探討內容語意描述及Metadata著錄格式，使其成為更具利用價值的數位典藏庫（註9）。本文則為後續之研究成果，仍以參與國科會數位典藏計畫之臺灣棒球新聞資料庫為基礎，進行兩個不同層面的整合研究：

（一）設計一個Metadata轉換系統，並採用泛用型之DC、專用型之NITF與RSS、以及數位典藏聯合目錄所制定的DAC等四種國內外普遍盛行之Metadata格式為標準，開發支援此四種規格的Metadata自動轉換系統，並衍生出RSS訂閱和聯合目錄大批匯出等應用，以提升新聞的利用價值，有助於新聞資源的管理、典藏與加值利用；

（二）以原始新聞資料庫作為Repository，建置符合OAI-PMH協定之Data Provider，以供扮演Service Provider之數位典藏聯合目錄取用，便於使用者透過數位典藏聯合目錄獲取數位化新聞資源，便於數位化新聞的傳播與分享。

二、相關研究

（一）新聞Metadata標準

在先前的研究中，我們將可應用於數位新聞的Metadata格式區分為三大類：

1. 國際通用的新聞Metadata格式

國際上已發展出許多專用於新聞事件的Metadata格式，且各有不同用途。例如：由國際新聞通訊協會IPTC（International Press Telecommunication Council）（註10）所制訂的NITF（News Industry Text Format）（註11）著重在新聞內文的描述；NewsML（News Markup Language）（註12）則著重封裝多種不同的媒體，用於描述電子出版、傳送、典藏的新聞檔；SportsML（Sports Markup Language）（註13）則用於運動項目紀錄；ProgramGuideML（Program Guide Markup Language）（註14）是廣播與電視新聞節目專用。再者，如IDEAlliance發佈的PRISM（Publishing Requirements for Industry Standard Metadata）（註15）主要為滿足雜誌、新聞、目錄、書籍和期刊等平面媒體出版者的商業需求而設計。又如XMLNews.Org所研擬的XMLNews（註16）主要在描述新聞報導的實質內容，是借用NITF而來的。至於衍生自Netscape推播技術（Push）的RSS則是一種用於互通新聞和其他Web內容的資料交換規格，目前已普遍應用於入口引擎、新聞網站、Blog和WiKi等系統。

2. 海峽兩岸中文新聞Metadata格式

為應用於中文新聞資源，海峽兩岸也發展出數種新聞Metadata格式。如台灣方面有文建會制訂的「新聞紀錄Metadata格式」（News Records Metadata Format，

本文簡稱為NRMF）（註17）；新聞業界則有聯合報系正進行中的「聯合新聞標示語言」（UDN Markup Language，簡稱UdnML）（註18）。大陸方面則有新華社所發展的「新華標示語言」（Xinhua Markup Language，簡稱XinhuaML）（註19），以及由中國報業協會所制訂的「中國報業電子新聞文稿格式」（Chinese News Text Format，簡稱CNTF）（註20）。

3. 泛用型Metadata格式

除了上述數種新聞專用Metadata外，也有泛用型Metadata，例如由OCLC（Online Computer Library Center）和NCSA（National Center for Supercomputing Application）所推動的都柏林核心集（Dublin Core，簡稱DC）（註21），可用於描述各種電子資源，其格式簡單易懂，具延伸性與互通性，可因應不同的需求，亦適用於數位化新聞的描述。國科會制訂的「數位典藏聯合目錄Metadata格式」（Digital Archive Catalog，簡稱DAC）（註22）則為數位典藏計畫共通標準。而由全球資訊網協會（World Wide Web Consortium，簡稱W3C）所制訂的「資源描述框架」（Resource Description Framework，簡稱RDF）（註23），更可作為各種Metadata整合機制。

（二）OAI-PMH分散檢索協定

OAI-PMH（Open Archives Initiative Protocol for Metadata Harvesting）（註24）是由OAI協會（Open Archives Initiative）自1999年開始發展，至2002年2.0版才較為完備的一個協定，為國際上數位化資源的交換標準之一。最初目的是作為學術性電子期刊預印本之互通性檢索，後則發展成Metadata分散整合機制，在資訊傳播過程中提供互通的標準架構，以將分散的資源加以匯整，因此可作為數位圖書館、數位博物館間之通訊協定，達成分散式數位典藏品整合檢索目的。OAI-PMH協定具實作容易、開放性，採用XML與HTTP等開放標準，相容性高…等優點，在歐美已有許多單位進行建置與使用（註25）。

1. OAI-PMH組成元件

OAI-PMH主要有五個組成元件：

(1) 資料提供者（Data Provider）：提供符合OAI-PMH協定之 Metadata格式資料；

(2) 服務提供者（Service Provider）：經由OAI-PMH協定向Data Provider取得資料，並以所獲得的Metadata建立加值服務；

(3) 資料儲存器（Repository）：透過HTTP協定接受OAI-PMH命令集，以回應資料存取需求的資料庫伺服器；

(4) 資料錄（Record）：依據OAI-PMH協定，從資料儲存器內將資料以XML編碼傳回前端的Metadata紀錄。

(5). 資料集（Set）：資料儲存器中歸屬於某個資料類別的資料錄所成的集合；

2. OAI-PMH命令集

OAI-PMH定義了六個命令集（Verbs），其運作模式為Service Provider送出Verbs

至Data Provider，以獲取Repository中的Record。這六個命令集為：

(1) GetRecord：檢索Repository中單筆的Metadata資料錄；

(2) Identify：取得Repository的識別資訊；

(3) ListIdentifiers：取得Repository中Metadata資料錄識別明細；

(4) ListMetadataFormats：檢索Repository中所支援的Metadata格式；

(5) ListRecords：列出Repository中指定範圍的所有Metadata資料錄；

(6) ListSets：檢索Repository中的資料集結構。

由於「數位典藏國家型科技計畫」所包含的眾多參與機構與計畫因內容主題的多樣性，不但分散且相互獨立，為求匯集各計畫之成果，供民眾查詢取用，乃產生以OAI-PMH建置「國家數位典藏聯合目錄」作為服務提供者（Service Provider）之構想，以整合各自分散的眾多數位典藏資訊系統。

三、Metadata轉換系統設計

(一) 新聞Metadata比較分析

在前述眾多的新聞Metadata規格中，經分析評估後，我們選用DC、DAC、RSS和NITF四種Metadata格式作為實作轉換機制，主要原因為：

1. DC簡單易用，不但是目前數位資源描述最普遍的Metadata格式，且為其餘若干格式之基礎，使用上具有豐富彈性，搭配RDF可融合各種Metadata，更可配合OAI-PMH協定進行分散性資源檢索，以利資源分享；

2. DAC為聯合目錄系統共通標準，乃國家數位典藏計畫指定規格，用以整合數位典藏資源，供公眾查詢利用；

3. RSS為線上新聞訂閱格式，可讓使用者獲得即時新聞資訊，且資料內容以摘要方式呈現，便於使用者過濾所需新聞，以利進一步閱讀；

4. NITF具豐富的內文語意標示，用於描述新聞內容，其元素規定嚴謹、完整，可詳述5W1H要素。

為建置Metadata自動轉換之系統，我們以DC為基礎，與RSS、NITF與DAC三種Metadata之元素，以及台灣棒球新聞資料庫內部Metadata欄位進行比對分析，歸納出常用元素與欄位對應關係，作為Metadata轉換之用，表1為比較分析之後的對應結果。

(二) 新聞Metadata轉換系統設計

依表1的元素對應分析，接著進行棒球新聞管理系統規劃與實作。本系統著重於資訊組織與管理層面，主要是在原有檢索子系統中加入Metadata轉換模組和顯示模組，再另開發RSS訂閱子系統，而後端管理子系統則加入聯合目錄大批匯出模組、棒球大事記管理模組及棒球名人錄管理模組，如圖1所示。圖中虛線部分為目

表1　四種Metadata與資料庫對位對應分析

（資料來源：本研究比較分析。▶：表示下一層元素；⇨表示該元素屬性值）

外部Metadata					內部Metadata	說明
DC	DAC	RSS 1.0	RSS 2.0	NITF	資料庫欄位	
identifier	DACatalog ▶ AdminDesc ▶ DigiArchiveID	rdf:RDF ▶ item ▶ dc:identifier	rss ▶ channel ▶ item ▶ guid	nitf ▶ head ▶ docdata ▶ doc-id⇨id-string	autoid (主索引)	識別碼（自動編號）
	DACatalog ▶ MetaDesc ▶ Identifier					
date	DACatalog ▶ MetaDesc ▶Date	rdf:RDF ▶ item ▶ dc:date	rss ▶ channel ▶ item ▶ pubDate	nitf ▶head ▶ docdata ▶ date.issue⇨norm	date	新聞日期
				nitf ▶head ▶ pubdata⇨ date. publication		
	DACatalog ▶ AdminDesc ▶ Catalog ▶ Record ／時間架構			nitf ▶body ▶ body.head ▶ hedline ▶date-line		
title	DACatalog ▶ MetaDesc ▶ Title	rdf:RDF ▶ item ▶title	rss ▶ channel ▶ item ▶ title	nitf ▶head ▶title	hotline	主標題
				nitf ▶body ▶ body.head ▶ hedline ▶hl1		
				nitf ▶body ▶ body.head ▶ hedline ▶hl2	subhotline	副標題
creator	DACatalog ▶ MetaDesc ▶ Creator	rdf:RDF ▶ item ▶ dc:creator	rss ▶ channel ▶ item ▶ author	nitf ▶body ▶ body.head ▶ hedline ▶ byline	author	記者
publisher	DACatalog ▶ MetaDesc ▶ Publisher	rdf:RDF ▶ item ▶ dc:publisher	rss ▶ channel ▶ item ▶ title	nitf ▶body ▶ body.head ▶ hedline ▶ rights	paperid	報別
contributor	DACatalog ▶ MetaDesc ▶ Contributor	rdf:RDF ▶ item ▶ dc:contributor	—	nitf ▶body ▶ body.head ▶ distributor	—	—
description	DACatalog ▶ MetaDesc ▶ Description	rdf:RDF ▶ item ▶ description	rss ▶ channel ▶ item ▶ description	nitf ▶ body ▶ body.content	document	新聞內容
subject	DACatalog ▶ MetaDesc ▶ Subject	rdf:RDF ▶ item ▶ dc:subject	rss ▶ channel ▶ item ▶ category	—	pagename	版名 (若無版名則置入固定值：體育新聞)

外部Metadata					內部Metadata	說明
DC	DAC	RSS 1.0	RSS 2.0	NITF	資料庫欄位	
source	DACatalog ▶ AdminDesc ▶ Hyperlink	rdf:RDF ▶ item ▶ link	rss ▶ channel ▶ item ▶link	—	—	來源 (該篇新聞之URI)
language	DACatalog ▶ MetaDesc ▶ Language	rdf:RDF ▶ channel ▶ dc:language	rss ▶ channel ▶ language	nitf ▶body ▶ body.end ▶ tagline ▶lang	—	語文 (固定值：zh-tw or 中文)
rights	DACatalog ▶ MetaDesc ▶ Rights	rdf:RDF ▶ channel ▶ dc:rights	rss ▶ channel ▶ copyright	nitf ▶head ▶ docdata▶ doc.copyrights⇨ holder	—	版權 (固定值：聯合報系)
format	DACatalog ▶ MetaDesc ▶ Format	—	—	—	—	資料格式 (固定值：text/xml or XML檔)
type	DACatalog ▶ MetaDesc ▶ Type	—	—	—	—	資料類型 (固定值：text or 文字)
coverage	DACatalog ▶ MetaDesc ▶ Coverage	—	—	nitf ▶head ▶ docdata ▶ doc-scope⇨ scope	—	涵蓋時空 (固定值：Taiwan, ROC or 台灣)
	DACatalog ▶ AdminDesc ▶ Catalog ▶Record ／空間架構：			nitf ▶body ▶ body.end ▶ tagline ▶ location		
relation	DACatalog ▶ MetaDesc ▶ Relation	rdf:RDF ▶ item ▶ dc:relation	rss ▶ channel ▶ item ▶ source	nitf ▶ head ▶ docdata ▶ doc.rights⇨ owner	—	關聯／來源 (固定值：http://ndap.dils.tku.edu.tw or 台灣棒球運動珍貴新聞檔案數位資料館)

前已完成者，虛線外則納入未來的研究中。

以上各子系統之間需適度整合才能良好運作，圖2為整合後之檢索子系統與RSS訂閱子系統之活動圖，使用者可經由Web介面選取欲進行之動作。主要可分為：全文檢索、圖片檢索、訂閱歷史上的今天，以及結束作業。

1. 檢索子系統

檢索子系統除先前已完成的部分外，於本研究中加入了Metadata轉換模組和顯示模組。由圖2可知當使用者輸入查詢條件後，程式即對後端資料庫進行查詢動作，並顯示符合條件的資料清單，使用者除可閱讀新聞全文外，也可進一步選擇欲轉

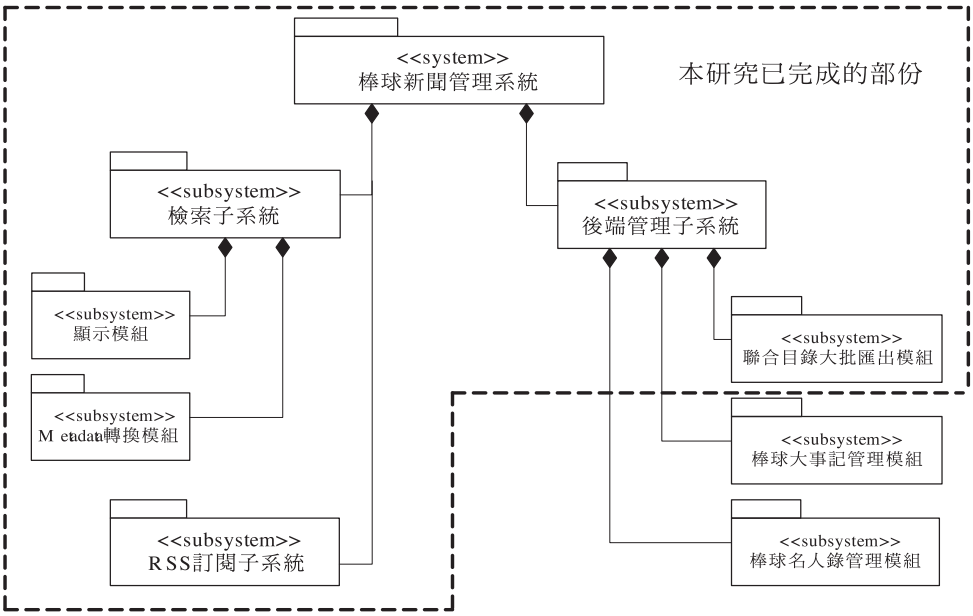

圖1 棒球新聞管理系統

換的Metadata格式，程式即將該篇新聞內容轉換成所選定之Metadata格式並顯示於Web介面上。目前已開發完成的轉換模組有：DC轉換子模組、DAC轉換子模組、RSS 1.0/2.0轉換子模組，以及NITF轉換子模組。

2. RSS訂閱子系統

本子系統乃RSS格式之衍生應用，主要是將數位典藏庫中的歷史性新聞以「棒球史上的今天」方式提供線上訂閱。RSS規格除了可作為單篇新聞發佈的Metadata格式，也可包含多篇新聞資源於單一RSS文件，將「台灣棒球運動珍貴新聞檔案數位資料館」作為發佈頻道（channel），每日更新，每個item是一篇新聞，item的新聞內容描述（description）只截取前150個字，使用者透過系統提供之RSS Feed進行訂閱，利用RSS Reader即時獲取摘要型新聞資訊，使用者可從其摘要資訊選取欲進一步閱讀之新聞，再以RSS提供之URI加以連結至新聞本文，即可得知台灣棒球史上當日發生的棒球新聞事件。

3.後端管理子系統

本子系統是為管理者與維護者所建置，用以管理新聞資源之修改、更新，以提供前端使用者最佳之新聞資源。目前「聯合目錄大批匯出模組」已完成，並順利將資料匯出至聯合目錄。此模組遵循數位典藏聯合目錄系統所定義之DAC格式，採用半自動化作法，將資料匯至聯合目錄：首先由本模組程式依年代為基準，結合Metadata轉換模組，自動將資料庫中的新聞資料大批轉出符合DAC格式之XML檔，存放於不同資料夾，再交由數位典藏聯合目錄小組人員整批匯入數位典藏聯合目錄系統中，以供公眾檢索。

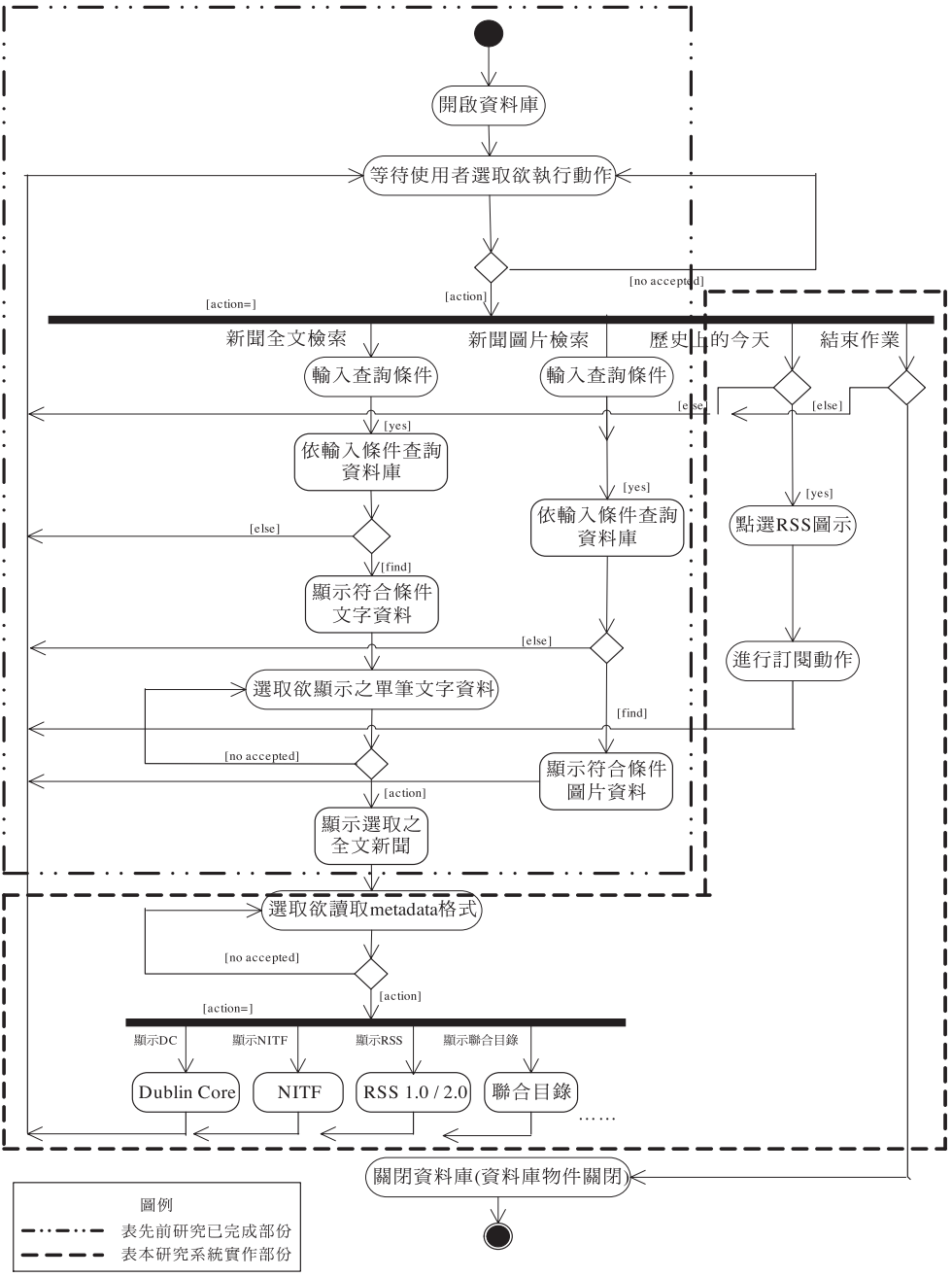

圖2 檢索子系統與RSS訂閱子系統之活動圖

此外，「棒球大事記管理模組」和「棒球名人錄管理模組」列入後續系統改進計畫中，預計採用維基協作系統（Wiki collaboration system）進行實作「台灣棒球維基館」；藉助Wiki易於分類、管理之特性，我們除了可將台灣棒球新聞以及所收集到的其他資料，整理成預定的「台灣棒球大事記」、「台灣棒球名人錄」之外，亦將延伸出「台灣棒球發展史」、「台灣棒球影像館」…等附加資訊；此外，由於Wiki具備開放協作之精神，所以也預計開放給所有對台灣棒運發展有興趣的人士來參與編寫工作。台灣棒球維基館目前已略具雛形，網址為：http:twbsball.dils.tku.edu.tw，但因採用Wiki開放協作模式，故其內容僅作為純學術研究之用，藉以探討資訊架構（Information architecture）設計相關議題，如索引典、控制詞彙、分類、索引等實務應用，並不納入本研究計畫正式授權產出之數位化內容。

四、 OAI-PMH檢索系統設計

雖然本系統之大批匯出模組已能順利將Metadata匯出至聯合目錄，然而，聯合目錄之終極目標是所有參與數位典藏的計畫皆能採用OAI-PMH架構，以便完成即時、全自動、高效率的資料彙集。有鑑於此，我們於是接著進行符合OAI-PMH協定之系統開發工作。

(一) 資料庫類目規劃

在數位典藏聯合目錄中，規劃以數種不同之「分類架構」進行計畫管理（註26）。本研究收錄之台灣棒球體育新聞，原以專卷方式進行新聞事件分類，其原始目的乃為協助報社從業人員查詢資料（註27），故分類細微，且部分類目以事件為名稱，各類目收錄篇數多寡不一，不但較不適宜一般使用者檢索查詢，對聯合目錄之分類檢索而言也顯得過於繁瑣。因此本研究針對資料庫之新聞事件設計符合棒球新聞之類目，以便於透過OAI-PMH「資料集」滿足聯合目錄「分類架構」之應用。並在原有系統中增設分類管理功能以便於管理者使用。

為求類目清楚簡單，我們捨棄起初為求詳盡完整採用主類目之下又再複分的樹狀結構，而改為單層類目，以免新聞類目過於繁多，歸屬不易，不便管理者分類和使用者查找。且資料經歸納整理後，發現棒球新聞事件多以收錄賽事為主，故將各層級賽事獨立成為類目名稱，其餘類目則以較泛稱之描述為名，總共分為十四類。本計畫之分類架構及其對應至數位典藏聯合目錄之整體架構如圖3所示。此外，因本研究以新聞資訊為主，故單篇新聞事件採取多重分類方式，以滿足使用者從不同角度搜尋新聞資訊。

（資料來源：本研究設計繪製）

圖3　本研究對應於聯合目錄之分類架構

(二) OAI-PMH之Metadata對應分析

OAI-PMH可支援多種Metadata標準，其中以DC為必備格式，稱為OAI_DC。基於前述對於新聞Metadata轉換與對應之結果加以延伸，將本資料庫內部Metadata欄位與OAI_DC各元素進行對應分析，歸納元素與欄位之注錄內容，以便回應符合OAI-PMH所規範之Metadata記錄集，對應分析結果彙整如表2所示。

(三) 實作與驗證

依據前述分析結果與OAI-PMH 2.0版規範，本研究開始著手建置臺灣棒球新聞檔案數位資料館之Data Provider伺服器，以支援數位典藏聯合目錄依Service Provider規範所傳送之命令集及參數。系統主要功能是剖析（parse）六個命令集及其參數，再依前述新聞類目作為紀錄集，以OAI-PMH定義之XML Schema格式封裝新聞Metadata作回應；此外，因資料量極大，也一併考慮流量控制（Flow Control）問題；而為了能與聯合目錄的Service Provider順利介接，也進行了系統驗證。

表2　OAI_DC與內部Metadata欄位對應表

OAI_DC	內部Metadata欄位	說　　明
dc:identifier	autoid（主索引）	唯一識別碼（自動編號） 格式為「oai:ndap.dils.tku.edu.tw:新聞事件之autoid」
dc:date	date	新聞日期
dc:title	hotline	主標題
	subhotline	副標題
dc:creator	author	記者
dc:publisher	paperid	報別
dc:contributor	—	—
dc:description	document	新聞內容
dc:subject	pagename	版名 (若無版名則置入固定值：體育新聞)
dc:source	—	來源 (該篇新聞之URI)
dc:language	—	語文 (固定值：中文)
dc:rights	—	版權 (固定值：聯合報系)
dc:format	—	資料格式 (固定值：XML檔)
dc:type	—	資料類型 (固定值：文字)
dc:coverage	—	涵蓋時空 (固定值：Taiwan, ROC)
dc:relation	—	關聯／來源 (固定值：淡江大學資訊與圖書館學研究所—台灣棒球運動珍貴新聞檔案數位資料館 http://ndap.dils.tku.edu.tw/)

（資料來源：本研究分析）

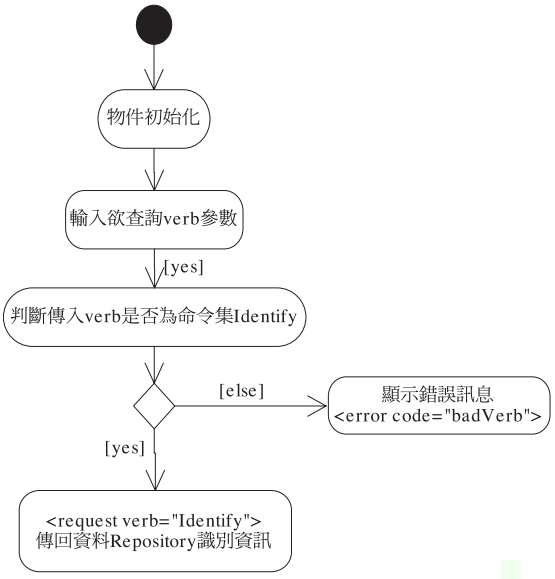

圖4　Identify活動圖

（資料來源：本研究繪製）

1. 命令集剖析與回應

針對六個命令集及其參數的剖析，本文以個別活動圖說明其運作原理，實作的回應結果則可直接參閱本研究網站之實際輸出（註28）。

(1) Identify

Identify乃為傳回該Repository的識別資訊，並無伴隨其他參數，因此剖析過程單純，只有當命令錯誤時，回應badVerb之錯誤訊息，其活動圖可參考圖4。以本研究為例，repositoryName注錄「臺灣棒球運動珍貴新聞檔案數位資料館」，而baseURL則為「http://ndap.dils.tku.edu.tw/oai/oai.asp」，earliestDatestamp則為收錄新聞資料的最早日期「1968-01-01」，採用之協定版本protocolVersion則為2.0版。詳細回應結果可連至本系統取得：http://ndap.dils.tku.edu.tw/oai/oai.asp?verb=Identify。

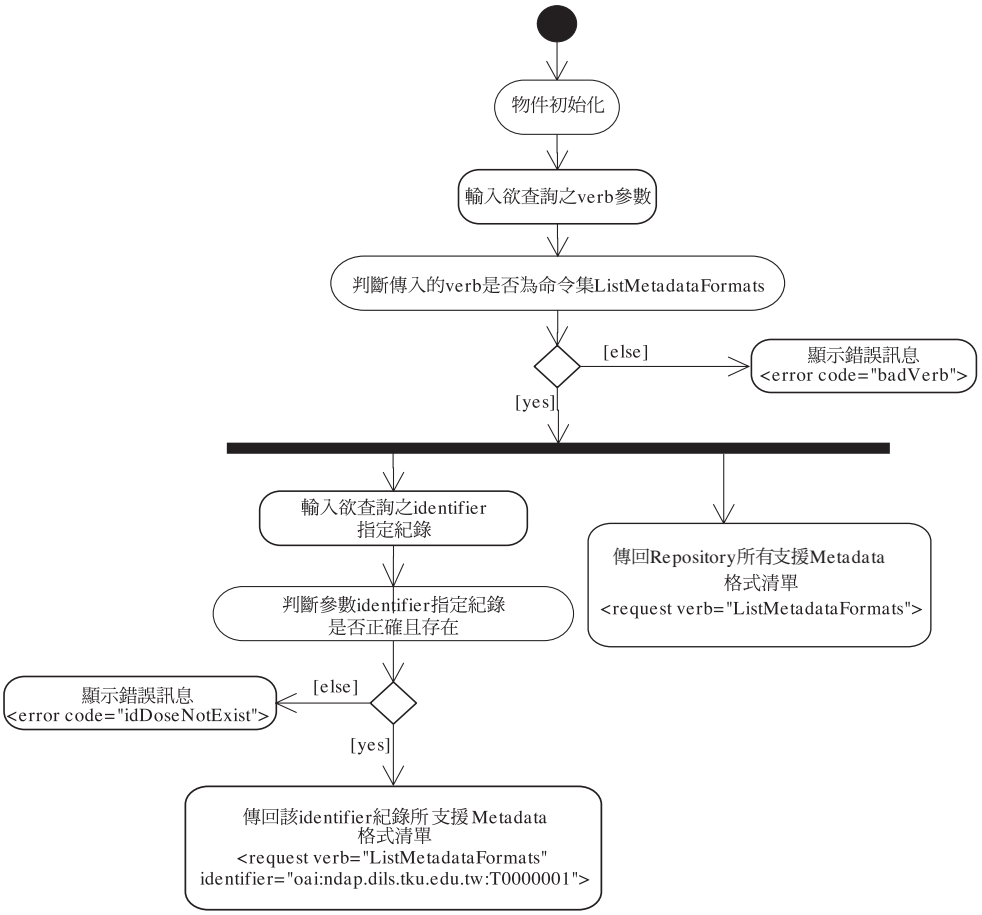

圖5 ListMetadataFormats活動圖

（資料來源：本研究繪製）

(2) ListMetadataFormats

ListMetadataFormats是用以查詢該Repository所支援之Metadata格式，若無指定特定之紀錄，則顯示所有支援之Metadata格式；若有指定單筆紀錄，則顯示該筆紀錄所支援之Metadata格式，圖5為其活動圖，實作結果可連至本系統取得：http://ndap.dils.tku.edu.tw/oai/oai.asp?verb=ListMetadataFormats。本研究目前只支援oai_dc一種Metadata格式，將來可再予以擴充。

(3) ListSets

ListSets為該Repository中所有「資料集」清單，亦即資料儲存器中所有分類類目，setSpec為該類目之標誌，如有複分須遵循OAI-PMH協定以「:」區別下層類目，

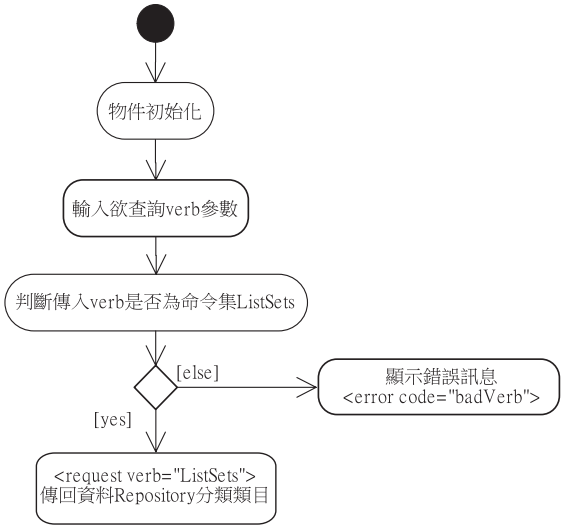

圖6 ListSets活動圖

（資料來源：本研究繪製）

setName為類目名稱，其活動圖參見圖6。以本研究為例，其採用單層類目如圖2所繪，因未超過15類，故不提供控制流量參數resumptionToken，回應結果可連至本系統取得：http://ndap.dils.tku.edu.tw/oai/oai.asp?verb=ListSets。

(4) ListIdentifier

ListIdentifier為取得Repository中紀錄標記，其必備參數為metadataPrefix或流量分頁參數resumptionToken，其中有可進階限制回傳資料範圍之參數，若查詢之參數錯誤將回傳錯誤訊息，其剖析流程可參考圖7。以http://ndap.dils.tku.edu.tw/oai/oai.asp?verb=ListIdentifiers&metadataPrefix=oai_dc為例，此URL乃為要求回傳metadataPrefix為oai_dc之紀錄Metadata格式，未進階限定範圍，而Repository中之紀錄過多，故顯示流量控制參數resumptionToken，讓使用者可以再次查詢接續之新聞紀錄。

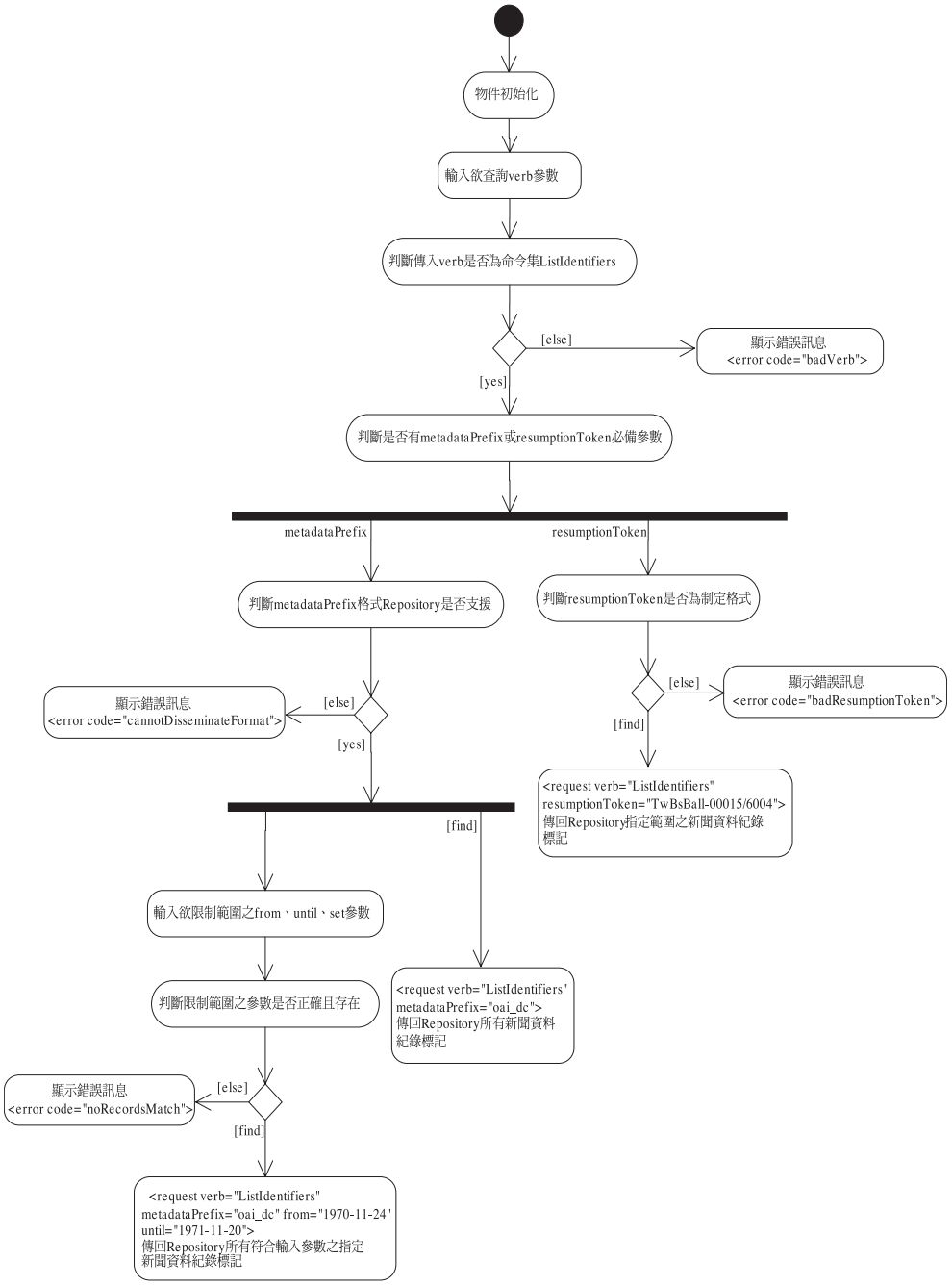

圖7 ListIdentifiers活動圖

（資料來源：本研究繪製）

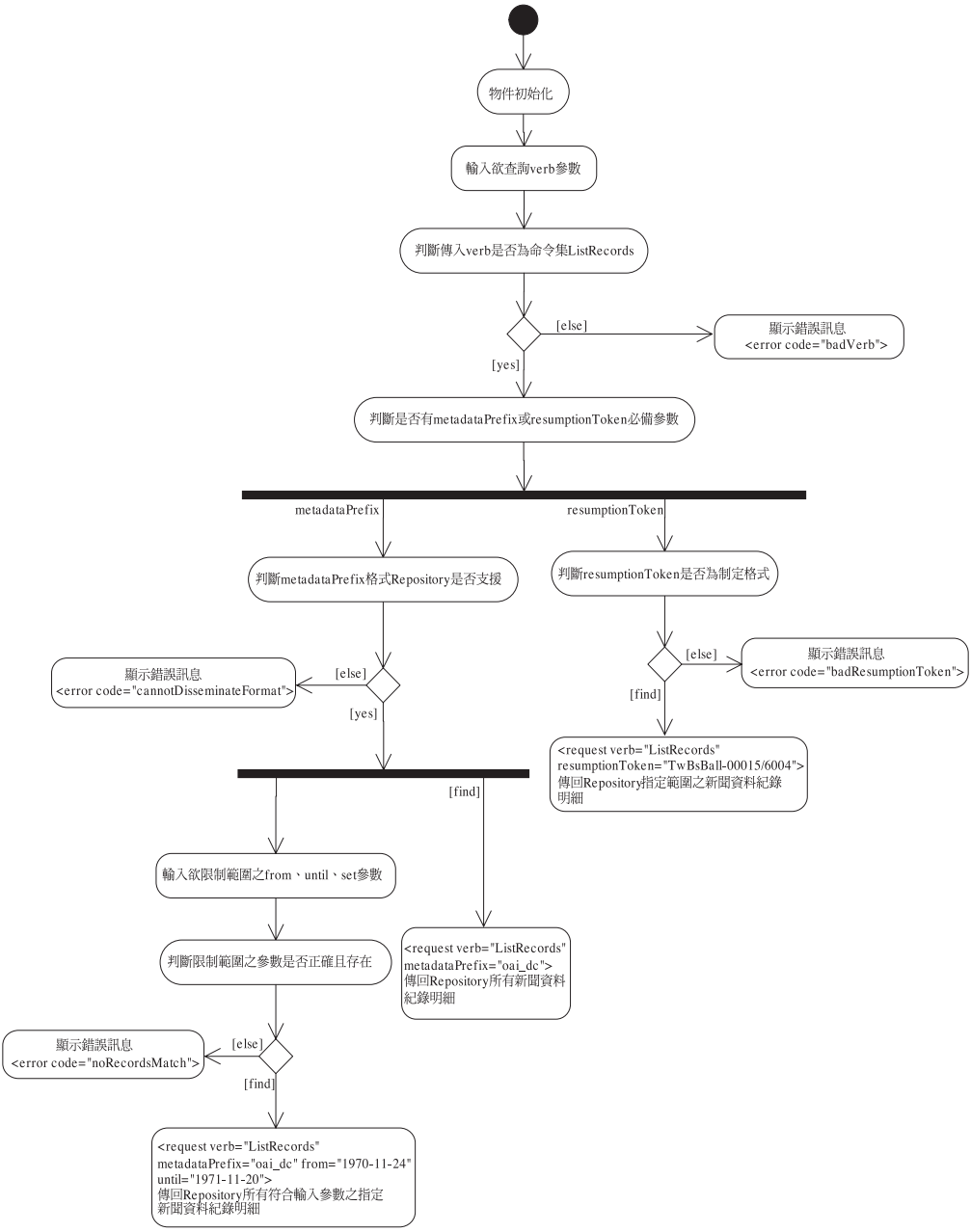

圖8　ListRecords活動圖

（資料來源：本研究繪製）

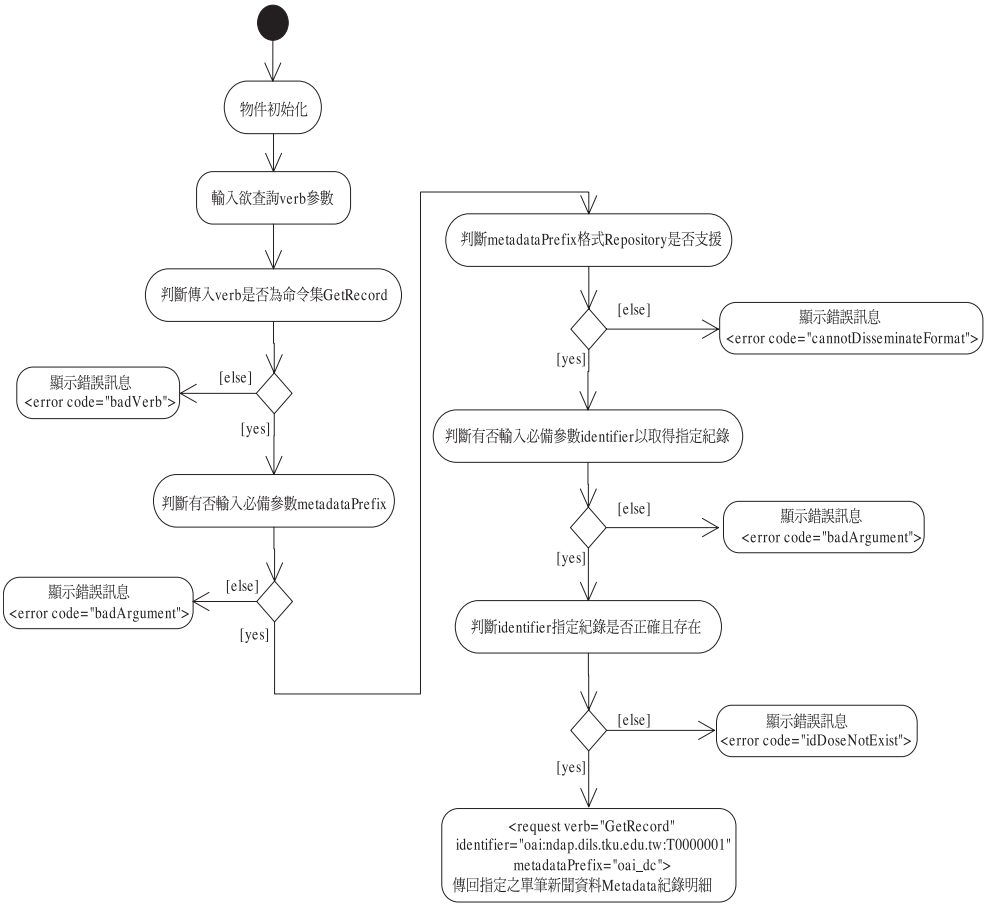

圖9 GetRecords活動圖

（資料來源：本研究繪製）

(5) ListRecords

ListRecords乃用以回傳Repository中指定範圍的所有紀錄明細，其必備參數與可選參數同ListIdentifier之規定，其差異在於回傳之資料為完整明細，而不同於ListIdentifier只有資料之標記部份，其剖析活動圖如圖9所示。本研究依表2所分析之內部Metadata與oai_dc對應項目，將新聞事件以OAI-PMH規定之格式回傳，例如：http://ndap.dils.tku.edu.tw/oai/oai.asp?verb=ListRecords&metadataPrefix=oai_dc，是要求回傳紀錄清單，而無進階限定查詢範圍，因Repository中之紀錄過多，故顯示流量控制參數resumptionToken，讓使用者可以再次查詢接續之新聞紀錄清單。

(6) GetRecord

GetRecord用以檢索Repository中單筆符合查詢範圍之完整Metadata格式紀錄，Identifier與metadataPrefix為必備參數，以Identifier作為查找Repository之憑據，而

依metadataPrefix之指定Metadata格式顯示，其剖析流程可參閱圖9之活動圖。本研究之Identifier格式採用「oai:ndap.dils.tku.edu.tw:新聞事件之autoid」，查找符合該Identifier的新聞事件後，依表2所整理與OAI_DC對應之項目，以XML格式回傳該指定Identifier的紀錄明細。例如http://ndap.dils.tku.edu.tw/oai/oai.asp?verb=GetRecord&Identifier= oai:ndap.dils.tku.edu.tw:T0000001&metadataPrefix=oai_dc為查找Identifier為oai:ndap.dils.tku.edu.tw:T0000001，指定以OAI_DC的Metadata格式回傳紀錄之結果。

2. 流量控制

當資料庫資料過多時，流量控制可用於分批取得資料，OAI-PMH協定中控制流量之參數為resumptionToken，但並未硬性規定此參數之注錄格式，而由各系統設計者彈性運用。本研究分析OAI官方網站中已註冊之Data Provider對於此參數之注錄格式，歸納出三種主要方式：

(1) 號碼牌：由Data Provider給予檢索端一個號碼牌作為註記，檢索端可以resumptionToken參數回傳該號碼牌進行再次檢索時，Data Provider則回應檢索端接續於上次檢索的紀錄。號碼牌注錄方式有給予隨機號碼、流水號或檢索時間等方式。

(2) 資料錄編號：Data Provider回應所有符合檢索之資料錄總數，與目前查詢回應之第一筆資料錄為全部資料錄之第幾筆，亦即以「當前資料錄/全部資料錄」之格式為之。

(3) 分頁編號：Data Provider將回應資料錄加以分頁顯示，告知檢索端目前為第幾頁，注錄方式通常為各Data Provider自訂之頁碼格式。

將此三種方法從便利、易懂等角度加以評估後，本研究決定採取「資料錄編號」之方式，利用「當前資料錄/全部資料錄」之資料錄編號方式管控流量，對於檢索端而言較為容易理解、使用。

3. 系統驗證

Data Provider建置完成需通過系統驗證之過程，才能確保符合OAI-PMH協定的規範，也才能與Service Provider順利介接，本研究利用OAI-PMH官方網站所提供的線上驗證功能（註29）進行檢測。起初一直無法通過OAI驗證系統，雖經研究者不斷修改程式與自行檢驗，確認並無問題卻仍無法順利過關。經由反覆的研究，最後推測可能因本研究之資料庫內容為中文繁體，採用Big5編碼，而OAI線上驗證系統並未支援此項編碼格式。於是我們另行建置英文版測試資料庫，更改XML編碼為Unicode（UTF-8），其餘回應之結構、語法與格式則維持不變，再次進行驗證後，便可完全通過其檢測。由此可確定本研究建置之Data Provider其回應結果符合OAI-PMH協定之規範。

五、結　論

本研究首先將四種目前較為盛行之新聞Metadata格式加以比較分析，綜合應用並加以轉換，以利數位化新聞的典藏、管理與傳播利用。未來本系統可再加入其他格式的Metadata，以滿足其他需求。再者，OAI-PMH是發佈與獲取Metadata的開放式標準，也是分散式資料庫整合查詢的利器之一，藉由此標準可使各數位典藏的計畫之資料加以匯集，且以DC為基礎，元素簡單易懂，融合性大，故可將所有不同典藏主題之計畫資料匯聚整合；本研究成功建置符合OAI-PMH協定規範之Data Provider，除可供數位典藏聯合目錄連結，提供數位典藏資源給使用者檢索、利用外，更可供全球符合OAI-PMH協定標準之Service Provider連結取用，提升數位化資訊的流通與傳播。

致　謝

本文為NSC 94-2422-H-032-002研究計畫部分成果，感謝國家科學委員會經費補助，聯合報授權使用所需新聞資料，研究生游忠諺、陳瑩潔、陳彥宇等協助系統建置、資料整理等工作，使本研究得以順利進行，特此致謝。

註　釋

註1　國科會，「數位典藏國家型科技計畫」，可得自http://www.ndap.org.tw/（上網日期：2005/1/25）。

註2　國科會，「數位典藏國家型科技計畫：內容發展分項計畫—新聞主題小組」，可得自http://content.ndap.org.tw/main/vision_brief.php?class_vision=16（上網日期：2005/5/25）。

註3　同註2。

註4　世新大學資訊傳播學系，「北平世界日報內容數位化開發計畫」，可得自 http://icd.shu.edu.tw/lipo/（上網日期：2005/7/4）。

註5　國家圖書館，「國家圖書館期刊報紙典藏數位化計畫」，可得自http://catalog.ndap.org.tw/dacs4/System/Organization/List.jsp?ContentID=6530&CID=7928 (上網日期：2005/7/4)。

註6　國科會數位典藏國家型科技計畫，「『電視新聞多媒體資料館』簡介」，電子通訊，第4期（2002年06月），可得自http://www.ndap.org.tw/1_newsletter/content.php?uid=239 (上網日期：2004/11/29)。

註7　國立交通大學傳播研究所，「蘭嶼原住民媒體資料庫建置與數位典藏計畫」，可得自 http://content.ndap.org.tw/main/vision_brief.php?class_vision=16 (上網日期：2005/7/29)。

註8　淡江大學資訊與圖書館研究所，「臺灣棒球運動珍貴新聞檔案數位資料館之建置」，可得自 http://ndap.dils.tku.edu.tw/ (上網日期：2005/7/19)。

註9　林信成，「主題式報紙新聞數位典藏之研究—以台灣棒球運動為例」，教育資料與圖書館學，45：3（民國94年3月），頁369-392。

註10　IPTC,“International Press Telecommunications Council,”available from http://www.iptc.org/pages/index.php（accessed 2005/3/22)

註11　IPTC,“News Industry Text Format,”available from http://www.nitf.org/ (accessed

2005/3/20).

註12 IPTC,“News Markup Language,”available from http://www.newsml.org/ (accessed 2005/3/20).

註13 IPTC,“Sports Markup Language,”available from http://www.sportsml.org/ (accessed 2005/3/20).

註14 IPTC,“Program Guide Markup Language,”available from http://www.programguideml.org/ (accessed 2005/3/20).

註15 IDEAlliance PRISM Working Group,“PRISM: Publishing Requirements for Industry Standard Metadata,”available from http://www.prismstandard.org (accessed 2005/3/22).

註16 XMLNews.Org,“NewsML,”available from http://xmlnews.org/NewsML/ (accessed 2004/12/20).

註17 行政院文化建設委員會國家文化資料庫知識管理系統，「News Records Metadata Format」，可得自 http://km.cca.gov.tw/download/rule.html (上網日期：2005/3/24)。

註18 聯合報系於2004年8月成立XML小組，對旗下各報社的新聞格式做規範與Metadata的制定。

註19 新華社技術局標準工組，「XinhuaMLv1.0功能說明書」，2003年1月18日。

註20 中國報業協會規範工作組，「中國報業電子新聞文稿格式」，2000年5月。

註21Dublin CoreMetadata Initiative,“Recommendation”s, available from http://purl.org/DC/documents/recommendations.htm (accessed 2004/12/17).

註22 國科會數位典藏國家型科技計畫，「數位典藏聯合目錄」，可得自 http://catalog.ndap.org.tw (2005/1/6)。

註23 W3C,“RDF Primer - W3C Recommendation 10 February 2004,”available from http://www.w3.org/TR/rdf-primer/ (accessed 2005/1/6)

註24 Open Archives Initiative,“The Open Archives Initiative Protocol for Metadata Harvesting,”available from http://www.openarchives.org/ OAI/openarchivesprotocol.html (accessed 2005/05/20)

註25 Open Archives Initiative,“OAI community,”available from http://www.openarchives.org/ community/index.html (accessed 2005/07/24).

註26「數位典藏國家型科技計畫_內容發展分項計畫：聯合目錄系統建置子計畫，『數位典藏聯合目錄分類架構』」，可得自http://catalog.ndap.org.tw/dacs4/System/Catalog/Catalog.jsp (2005/07/27)。

註27 孫正宜，「新聞專卷的數位化與加值應用—以台灣棒球報紙新聞數位典藏為例」，（淡江大學資訊與圖書館學系研究所碩士論文，民93）。

註28 國立交通大學傳播研究所，「蘭嶼原住民媒體資料庫建置與數位典藏計畫」。

註29 Open Archives Initiative,“Registering as a Data provider OAI-PMH version 2.0,”available from http://www.openarchives.org/data/registerasprovider.htm (2005/07/14).

Journal of Educational Media & Library Sciences 43 : 1 (September 2005) : 49-68

Integrated Application of Metadata and OAI-PMH to Digital News Archive

Sinn-Cheng Lin

Associate Professor
Department of Information and Library Science, Tamkang University
Taipei, Taiwan, R.O.C.
E-Mail: sclin@mail.tku.edu.tw

Pei-Shiun Kang

Graduate Student
Department of Information and Library Science, Tamkang University
Taipei, Taiwan, R.O.C.
E-Mail: 693070038@s93.tku.edu.tw

Abstract

This paper studies the metadata integration problem between the Taiwan Baseball News Database and the Union Catalog of National Digital Archives Program. First, we design a metadata transformation system which can transform the news articles to four popular metadata formats: DC, NITF, RSS and DAC. The system not only supports metadata transformation but also RSS subscription and batch exportation for Union Catalog. Secondly, based on the OAI-PMH architecture, the digital news archives plays a role as Repository, and then we extend the system to be a Data Provider. It can be queried by any Service Provider, parse the Verbs, and then respond the Record Sets to the Service Provider with certain metadata format. As a result, the digital news can be delivered more efficient.

Keywords: *Digital archive; Metadata; Makeup language; Union catalog; OAI-PMH*