從都柏林核心集看未來資料描述格式的發展趨勢

吳政叡 (Cheng-Juei Wu)
輔仁大學圖書資訊系專任副教授
Associate Prof.
Department of Library & Information Science
Fu-Jen University
E-mail: lins1022@fujens.fju.edu.tw

中文摘要

網際網路和WWW的結合，大幅降低了資訊傳播的障礙，於是全球單一資訊網的架構已在逐漸形成中，但這引發了資訊量過多的問題。為了有效率的過濾和處理大量資料，一套適合眾多非圖書館專業人員的資料描述格式（或元資料），乃成為當務之急，都柏林核心集即是設計來解決這一問題，其所立下的原則，如無必須項原則、可重覆原則、可修飾原則等，正是未來資料描述格式的發展方向。本文簡介了都柏林核心集的13 個資料項，並且針對上述各原則加以分析。

關鍵字：都柏林核心集，元資料，目錄，電子圖書館。

一、前言

近幾年來，網際網路（Internet）和全球資訊網（World-Wide Web，簡稱WWW）的迅速興起，對資訊傳播的方式產生了重大的衝擊。由於網際網路是連結全世界的巨大網路，透過此網路，資料得以日夜不息的在全世界流動。另一方面，WWW以其易寫作和方便連結文件的優點，在短時間內蔚為風潮，從全球性跨國公司到個人，莫不爭相建立自己的首頁，來善用這二十四小時不停的訊息傳播工具。同時網際網路和WWW的相互結合，大幅降低了資訊傳播的障礙，其所引發的效應之一，即是造成資訊量的激增。以前雖然是知識爆炸，但由於資訊傳播管道的障礙甚多，還不至於讓人覺得壓迫甚重，因為你無法在短時間內接觸到很多的資料。但是現在的網際網路和WWW，卻在一瞬間將全部隱藏的資料引爆出來呈現在你面前，這下可真讓人感受到資訊爆炸的威力。

資訊傳播障礙的移除，引發了二個看似迥異卻又相關的問題，一是如何來有效率的過濾資料，一是如何來有效率的描述資料。就前者而言，

目前在使用 WWW 上的搜尋引擎（如Locys [註 1] 等）來收集資料時，大家經常會面臨到的問題之一，是所得到的資料回覆量太多，經常可有上萬條款目，實無法一一來加以過濾，更糟的是，排在前面的款目，又往往不是你所真正需要的，頗使人進退維谷，祇有瞎猜亂挑。很明顯的，我們需要更多的資訊，來從回覆的款目當中，挑選我們真正需要的資料，而這些資訊必須由資料提供者來提供，因此如何制定一套資料描述格式，來有效率的描述收藏的資料，成為一個重要的課題，這正是元資料（Metadata）日漸受到重視的原因。如經由微軟一些與網際網路相關的軟體所製作的 HTML 文件中，已在文件開頭處，加入許多元資料項目，來紀錄此文件是由那個軟體製作的。

在資訊的傳播方式上，網際網路和WWW盛行前，是主要以下面的方式進行:

資料提供者-->圖書館-->資料使用者

因此圖書館可以說是主要的媒介者，來溝通資料提供者（如出版社）和資料使用者（如個人），所以圖書館扮演了資料儲存和傳播者的主要角色。為了有效達成其做為媒介和橋樑的角色，使圖書館能夠有效率的來管理擁有的資料，以便使用者可以很快找到所需的資料，圖書館須要有一套很好的方法，來描述所收藏的資料，於是有目錄的興起。其後隨著資料處理科技的進步，從卡片目錄演進到目前的機讀編目格式（MARC），其編製手法和處理方式，或有人工與電腦操作的差別，但它們的基本功能和扮演的角色卻是相同的。

如今網際網路和WWW提供了一條直接的管道，使資料提供者和資料使用者可以直接接觸，毋須透過圖書館來作為媒介者。這固然降低了資訊傳播的障礙（少了一個中介機構），但另一方面，資料提供者如今必須自己擔負起圖書館所提供的一些功能，其中之一是對所擁有的資料加以描述（著錄）。但圖書館所發展出來的資料描述格式，雖然完整和嚴謹，但卻較適合圖書館專業人員使用，對大多數的非專業人員而言，是過於繁瑣和不易學習的。都柏林核心集（Dublin Core）[註 2] 即是在這一背景下興起的產物，試圖提供一套簡易的資料描述格式，來滿足大多數非圖書館專業人員的需求。本文擬藉由對此一資料描述格式的深入探討，來分析未來資料描述格式的發展趨勢，也提供圖書館界在制定未來資料描述格式時的一個參考。

二、都柏林核心集簡介

都柏林核心集這個元資料格式，是 1995 年 3 月由 Online Computer Library Center （OCLC）和 National Center for Supercomputing Applications（NCSA）所聯合贊助的研討會，在邀請五十二位來自圖書館、電腦和網路方面的學者和專家，共同研討下的產物。目的是希望建立一套描述

網路上電子文件特色的方法，來協助資訊檢索。因此在研討會的報告中，將元資料定義為資源描述（resource description），而研討會的中心問題是 [註 3]

如何用一個簡單的元資料記錄來描述種類繁多的電子物件？

根據研討會的報告，都柏林核心集處理的對象將祇限於類文件物件（document-like objects，簡稱 DLO）[註 4]，那何謂 DLO 呢？簡言之，是可用類似描述傳統印刷文字媒體方式，加以描述的電子檔案。同時因為研討會的目標是發展一個簡單有彈性，且非專業人員也可輕易了解和使用的資料描述格式，所以都柏林核心集祇規範那些在大多數情況下，必須提及的資料特性，總共有 13 個資料項，在此我們以扼要的方式列表如下：[註 5]

資料項一. 主題（Subject）：作品所屬的學術領域。
例子：Subject = Digital Geospatial Metadata

資料項二. 題名（Title）：作品名稱。
例子：Title = Geospatial Support Staff Metadata Tutorial

資料項三. 著者（Author）：作品的創作者或組織。
例子：Author = Abeyta, Carolyn

資料項四. 出版者（Publisher）：負責發行作品的組織。

資料項五. 其他參與者（OtherAgent）：對作品創作有貢獻的相關人或組織。

資料項六. 出版日期（Date）：作品公開的日期。
例子：Date=1995/05

資料項七. 資料類型（ObjectType）：作品的類型或所屬抽象範疇，可用來幫助資料檢索。
例子：ObjectType = tutorial

資料項八. 資料格式（Form）：告知檢索者在使用此作品時，所須的電腦軟體和硬體設備。
例子：Form=html

資料項九. 識別代號（Identifier）：字串或號碼可用來唯一標示此作品。
例子：Identifier（scheme = URL）=http://www.blm.gov/gis/meta/barney/tut_met1.html

資料項十. 關連（Relation）：與其他作品（不同內容範疇）的關

連，或所屬的系列和檔案庫。
例子：Relation（type = ContainedIn）（identifier=URL）= http://www.blm.gov/

資料項十一. 來源（Source）：作品從何處衍生而來（同內容範疇）。

資料項十二. 語言（Language）：作品所使用的語言。
例子：Language = English

資料項十三. 涵蓋時空（Coverage）：作品所涵蓋的時期和地理區域。

檢視上述架構，可清楚看到某些資料項，是針對電腦作業環境而設計的，如資料項八（資料格式，Form），其他如資料項七（資料類型，ObjectType）、資料項十（關連，Relation）和資料項十一（來源，Source），也和網路或電子作業環境有密切的關係。同時此資料描述格式可說是非常簡單和易使用，幾乎所有資料項都有自我解釋的功能，大部份人在短時間內就知如何使用，比起機讀編目格式的深奧難懂，此元資料在鼓勵非專業人士的自行著錄所收藏資料，以及縮短製作時間和成本上，可說是成功的。

由於都柏林核心集祇是一套最小的核心資料項，須要有一個機制來與現存的其他較完整描述格式，如 USMARC [註 6] 和美國聯邦地理資料委員會 （Federal Geographic Data Committee，簡稱 FGDC）的地理電子元資料（Digital Geospatial Metadata）[註 7]等標準，來做資料的對照和轉換，使都柏林核心集的資料，能在最少成本下，轉換成更完備的描述格式。在這方面，有關 USMARC 和都柏林核心集的連結已在發展中，有興趣的讀者請參考 [註 8]，在那一篇文章中，對如何將都柏林核心集的 13 個資料項，對映到 USMARC 的相關欄位，有詳盡的分析和解說，如都柏林核心集的 Subject 資料項應對映到 USMARC 的欄位 653 或欄位 650 等。

三、特性分析與未來發展趨勢

首先，就柏林核心集的創造背景而言，『需要為創造之母』，如果現今做為圖書館自動化核心的機讀編目格式，能滿足現在興起的全球單一網路作業環境特性，能有效解決目前在WWW檢索上所面臨的問題，則人們必不會捨近求遠的再去創造其他資料描述格式。那人們為什麼不用機讀編目格式去描述所擁有的電子文件呢？主要原因之一是機讀編目格式太複雜難懂，其欄位結構複雜，項目太多，著錄規則有些又厚達數巨冊（如AACR2），很明顯的，其紀錄的製作成本太高，無法為全球眾多的WWW站台所接受。

對一般小型或私人的站台而言，其工作人員大多數是非圖書館專業人員，雖然規模不大，但若站台不對其所擁有的資料加以適當的描述，則我們無法有效率的來過濾、管理、和檢索資料。因此都柏林核心集一再強調其基本目的，是提供眾多非圖書館專業人士，一套簡單好用的資料描述格式，並且盡量降低紀錄的製作成本，來應付快速增加的資料量，

這可由其祇規範了十三個項目得知。作者個人非常同意這一作法，因為再好和完備的格式，若用的人越來越少，則終有一天會變成博物館內的古董。因此圖書館未來所使用的資料描述格式，不應越來越精細和複雜，若無法同時兼顧專業和非圖書館專業人員，則寧可盡量遷就非專業人員的需求，畢竟從著錄的角度來說，他們將是多數者，專業人員覺得不足的部分，再用其他方式來加以彌補。

再者就項目的基本設計原則而言，基於與會者認為沒有任何單一的元資料格式，足以適用於任何作業環境的認知，他們主張先建立一套描述資料的最小核心資料項。因此元資料的設計原理，是使此元資料的資料項，同時擁有意義明確、彈性和最小規模三種特色。在設計上所秉持的原則是：內在本質原則、易擴展原則、語法獨立原則、無必須項原則、可重覆原則、和可修飾原則。以下是它們的簡要敘述：[註 9]

（一）內在本質原則（Intrinsicality）：祇描述跟作品內容和實體相關的特質，例如主題（subject）屬於作品的內在本質。但是收費和存取規定則屬於作品的外在特質，原則上不屬於核心資料項，將透過其他機制來加以處理。

（二）易擴展原則（Extensibility）：應允許地區性資料以特定規範的方式出現，也應保持元資料日後易擴充的特性，以及保有向後相容的能力。

（三）語法獨立原則（Syntax-Independence）：在此元資料成熟前，將盡量避免制定特定語法。

（四）無必須項原則（Optionality）：所有資料項都是可有可無，以保持彈性和鼓勵非專業人士參與製作。

（五）可重覆原則（Repeatability）：所有資料項均可重覆。

（六）可修飾原則（Modifiability）：資料項可用附加限定語（qualifier）來進一步修飾其意義。

現在我們就針對以上各原則分析如下：

（一）內在本質原則：因為著錄資訊全來自資料本身，並不須要再額外去找其他的參考來源，很顯然的可以大幅減輕著錄者的負擔，對非專業人士來說，也是較可被接受的一種方式。

（二）易擴展原則：此原則是為了適應全球網路的作業環境，因眾多的站台各有自己獨特的資料種類和需求，因此必須有適當的彈性。

（三）語法獨立原則：這祇是因應都柏林核心集目前的發展階段而設的。

（四）無必須項原則：這可能使得某些人覺得非常驚異和不適應，傳統的圖書館著錄格式，如 MARC，和很多的元資料格式，如

FGDC的、GILS [註 10]、DIF [註 11]等，都有必須著錄項，如題名項和作者項等，主要不外乎是要維持一定的著錄品質。但為了鼓勵著錄，和強調有資料總比沒資料好的原則，都柏林核心集決定不硬性規定任何必須著錄項，作者頗認同此一原則。為了能適應各式非圖書館專業人員的背景和能力，必須著錄項若不能全部免除，也應盡量減少，以減輕著錄者的負擔。

（五）可重覆原則：此原則進一步簡化許多著錄規則，如在此一原則下，將不區分作者的排名。傳統上為了決定第一作者或是題名，著錄規則中往往有很多的篇幅來規範。事實上，從檢索的角度來看，讀者何嘗在意一本書內的排名次序，眾多的題名，也可藉由電腦的輔助，輕易來加以檢索或處理，實無在著錄格式上，加以嚴格區分的必要。這些從卡片目錄時代為了排片需要所遺留下的產物，有必要加以檢討和去除。

（六）可修飾原則：這原則使都柏林核心集非常有彈性，可同時滿足圖書館專業和非專業人員的需求。對於非專業人員來說，他們基本上不須要去查專業書籍來進行著錄的工作，這將大大減輕項目的著錄成本和時間。另一方面，對欲維持一定品質的專業人員而言，透過在（ ）內加修飾語，可明確指出所使用的資訊來自何處，如：Subject（=LCSH）=UNIX（Computer System） [註 12]。作者非常贊同這個可同時兼顧專業和非專業人員的設計理念，由於未來圖書館勢必與全球網路的資訊傳播系統緊密結合，成為全球網路資訊系統的一份子，自不可能採用獨特的資料描述格式，所以一套能同時兼顧專業和非專業人員的資料描述格式，將是時勢所趨。

四、結論

因為網際網路和WWW的緊密結合，資訊傳播的障礙已大大的降低，兩者的結合提供一條非常方便和快速的傳播管道，使資料得以日夜不息的在全球流動，透過一些著名的搜尋引擎，人們可以很方便的搜尋資料。但是新的科技也引發新的問題，其中一個急迫須要解決的問題，是從搜尋引擎上得到的資料回覆量往往過多，而無法有效的加以過濾和處理。

作者以為這個問題的解決方案，將主要依賴二個方法，一是資料提供者運用元資料來提供與文件相關的充分資訊給檢索者，一是檢索系統採用模糊邏輯原理，來加強對資料的過濾和處理能力。以前者而言，綜觀目前大多數的搜尋引擎，在資料的回覆畫面上，都祇有顯示標題、密合百分比、簡短的數行文字、URL（路徑+檔名）、有些系統有附上檔名大小和製作時間。如此簡略的設計，無怪乎檢索者無法判斷某筆資料到底是否為其所需，而惟有將整個檔案下載，直接閱讀後才能得知。這種操作是很沒有效率的，因為網路的傳輸部分，往往是系統最慢的一個環節，解決之道應是透過元資料來對資料加以適當的描述，提供給檢索者更多的資訊來做判斷，而達到減少不必要傳輸的目的，事實上，這正是目錄的基本功用。

元資料對電子文件（或檔案）所扮演的角色，正可對比於目錄之於傳統的印刷媒體資料，因此元資料可說是『電子目錄』，正如目錄過去所

扮演的角色一樣，元資料將可大幅減少不必要的檔案傳輸次數，提高資料檢索的效率。

都柏林核心集雖然祇是一個很簡略的料描述格式（或元資料），但它是圖書館界試圖解決電子文件處理難題上的一個新嘗試，其所立下的原則和典範，如無必須項原則、可重覆原則、可修飾原則等，是令人印象深刻和激賞的。以作者個人的觀點，都柏林核心集對未來資料描述格式所揭示的方向和途徑是正確的，但它目前尚在一個初期的發展階段，仍有待圖書館界加以繼續補充和推廣。

註釋：

註 1：Infoseek Corp., "Infoseek Home Page," <http://www.infoseek.com/> (18 Feb. 1996).

註 2：Stuart Weibel, Jean Godby, Eric Miller, and Ron Daniel, "OCLC/NCSA Metadata Workshop Report," 1995, <http://www.oclc.org:5047/oclc/research/publications/weibel/metadata/dublin_core_report.html>.

註 3：同註 2，頁 2。

註 4：同註 2，頁 3。

註 5：同註 2，頁 7-11。

註 6：Library of Congress, USMARC Format for Bibliographic Data. (Washington, DC: Library of Congress, 1994).

註 7：GDC, "Content standards for digital geospatial metadata -- FGDC," 1994, <http://fgdc.er.usgs.gov/fgdc.html>.

註 8：Rebecca Guenther, "Mapping the Dublin Core Metadata Elements to USMARC", 1995, <gopher://marvel.loc.gov/00/.listarch/usmarc/dp86.doc>.

註 9：同註 2，頁 5-6。

註 10：”Guidelines for the Preparation of GILS Entries,” March 1995, <http://gopher.nara.gov:70/0/managers/glis/guidance/gilsdoc.txt>.

註 11："Directory Interchange Format (DIF) Writer's Guide, Version 5.0a," Oct. 1996, <http://gcmd.gsfc.nasa.gov/difguide/difman.html>.

註 12：同註 2，頁 7。