都柏林核心集對減低檢索失誤率的實務研討

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

為了驗證元資料的實際效用，作者選用都柏林核心集做為著錄的元資料，並以選修作者所開設的研究所課程「元資料概論」的七名研究生為實驗者，設計了一個先導式的簡略實驗，實驗結果證實，都柏林核心集的確可以做為判斷文件是否為所需要的依據，因為其檢索失誤率僅有2.9%，相反的，國內外著名的七個檢索引擎則平均有高達七倍的檢索失誤率（20.7%）。同時也發現都柏林核心集確有達到創制者們預期的目標—易學好用和快速著錄，非常適合各種背景人士使用，達成「作者著錄」的目的。此初步的實驗顯示，都柏林核心集對減低資訊的檢索失誤率和提昇資訊檢索的準確度有很大的助益。

==========

Experiments on Using the Dublin Core to reduce the Retrieval Error Ratio

Abstract

In order to test the power of metadata on information retrieval, the author designed and conducted an experiment on a group of seven graduate students using the Dublin Core as the cataloging metadata. The experimental results show that, on average, the retrieval error ratio is only 2.9% for the MES system (URL: http://140.136.85.194), which utilizes the Dublin Core to describe the documents on Web, in contrast to 20.7% for the seven famous search engines including HOTBOT, GAIS, LYCOS, EXCITE, INFOSEEK, YAHOO, and OCTOPUS. The very low error ratio indicates that the users can use the information of the Dublin Core to decide whether to retrieve the documents or not.

==========

關鍵字：都柏林核心集，檢索失誤率，元資料，檢索，電子圖書館，Dublin Core，Retrieval Error Ratio，Metadata，Information Retrieval，Digital Library。

一、前言

1990 年代在人類資訊處理上最顯著的里程碑，首推 World-Wide Web（全球資訊網，簡稱 WWW）的盛行，WWW是起源於CERN中的一個增進高能物理學者間互動的實驗計畫 [註 1]，但WWW 藉著網際網路的無遠弗屆，親善的使用介面和易寫作的超文件標示語言（HyperText Markup Language，簡稱 HTML）格式，在短時間內形成一股風潮席捲全球，也無形中改變人們搜尋資料的習慣和期望。

網際網路和WWW的迅速興起和結合，對資訊傳播的方式產生了重大的衝擊。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。WWW則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。因此網際網路和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。

WWW盛行後，為因應檢索網頁內容的需要而有檢索引擎的產生，檢索引擎運作的方式，基本上是屬於全文檢索，主要是透過自動抓取程式在網際網路上抓取網頁，然後以自動拆字（或詞）作索引的方式來建立其資料庫，做為檢索的基礎，這種操作方式的特點是高運作效率和一網打盡，因此有高回收率與低精確率的特性。目前在使用 WWW 上的檢索引擎來查詢資料時，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。這個低精確率的缺點，隨著WWW網頁數量的急遽膨脹，成為無法忍受的致命傷。於是大家體會到對資料加以適當描述的重要性，這跟圖書館製作目錄的動機是一致的，這個古老的經驗又得到再一次的肯定。

元資料（Metadata）最常見的英文定義是 "data about data"，可直譯為描述資料的資料，主要是描述資料屬性的資訊，用來支持如指示儲存位置、資源尋找、文件紀錄、評價、過濾等的功能。以圖書館的角度來看，就其本義和功能而言，元資料可說是電子式目錄，因為編製目錄的目的，即在描述收藏資料的內容或特色，進而達成協助資料檢索的目的。因此元資料是用來揭示各類型電子文件或檔

案的內容和其他特性，其典型的作業環境是電腦網路作業環境。[註 2] 換言之，元資料是因應現代資料處理上的二大挑戰而興起的：一是電子檔案成為資料的主流，另外一個是網路上大量文件的管理和檢索需求。

為了驗證元資料的實際效用，作者選用都柏林核心集做為著錄的元資料，並以選修作者所開設的研究所課程「元資料概論」的七名研究生為實驗者，設計了一個先導式的簡略實驗，來比較都柏林核心集和國內外一些著名的檢索引擎的效能。但是傳統上用來衡量資訊檢索效能的回收率和精確率，在現代商業資料庫或檢索引擎動輒上百萬筆（甚或更大）資料的規模下，在應用上有實際的困難或不足之處。事實上除了由於計算上的困難而窒礙難行外，更因為全文檢索相關技術的發展和一網打盡的特性，回收率和精確率已逐漸喪失其使用價值和意義。

事實上如果我們仔細觀查使用者的檢索過程和行為，可以發現無論是使用古老的卡片目錄、圖書館自動化系統、WWW的檢索引擎，使用者在查到目錄資料或者是檢索引擎的回覆款目後，所必須做的共通抉擇，是判斷此資料是否為所需，接下來的行動是直截了當的二分法：取得原文或者忽略跳過。根據經驗，取得原文的過程往往甚為耗時費力，這使得行動之前的判斷益顯重要，反而是評估檢索系統效能的重要依據，能協助使用者做出正確判斷的系統，其最終和整體的效能才是最佳的。

基於以上的認知，作者製作了一個新的衡量標準--檢索失誤率（retrieval error ratio，簡稱RER），用來評估檢索系統的效能，檢索失誤率是以使用者最後看到原文後的判定為基礎，來比較和評估檢索系統在提供（目錄）資訊與判斷資料重要性（即所謂ranking能力）的整體表現。因此本文以檢索失誤率（RER）為衡量標準，來比較都柏林核心集和國內外一些著名的檢索引擎的效能。

二、都柏林核心集的發展現況和欄位簡介

都柏林核心集（Dublin Core）創始於1995 年 3 月由國際圖書館電腦中心（Online Computer Library Center，簡稱OCLC）和 National Center for Supercomputing Applications（NCSA）所聯合贊助的研討會，是五十二位來自圖書館、電腦、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。研討會的中心問題是--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？[註 3] 主要的目標是發展一個簡單有彈性，且非圖書館專業人員也可輕易了解和使用的資料描述格式，來描述網路上的電子文件。

都柏林核心集最近一次的研討會為第五次研討會，於1997年10月6-8日在芬蘭的赫爾辛基舉行，以下根據澳洲國家圖書館的一位與會者--Bemal Rajapatirana

的報告，介紹第五次研討會的情況與成果。

根據Bemal Rajapatirana的報告，與會者達成了如下的幾項共識：[註4]

(一) 加快標準化的腳步。

(二) 區分簡單和複雜兩種都柏林核心集格式。

(三) 語法上採用HTML和RDF格式為主—HTML的格式目前是使用4.0版本，寫法請參見作者的另一篇文章 [註 5]，RDF格式的寫法，作者已撰寫一文加以介紹。

(四) 成立工作小組。

(五) 規範次項目（或類別）修飾詞的制定原則。

在都柏林核心集的欄位方面，因為其目標是定位於一個簡單有彈性，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，最初規範有 13 個資料項 [註 6]，在 1996 年 OCLC所舉辦的一場研討會上，根據與會影像處理專家的建議，都柏林核心集新增了二個資料項--簡述（Description）和版權規範（Rights Management），並修改了部分資料項名稱 [註 7]，在此以扼要的方式，將 1997 年 10 月公布的資料著錄項目列表如下：[註 8]

(一) 主題和關鍵詞（Subject）：作品所屬的學術領域，控制語彙用 scheme 註明出處如 LCSH，亦可包含分類號如杜威十進分類號（Dewey Decimal Number）。

例子：Subject = Digital Geospatial Metadata。

例子：Subject = 都柏林核心集。

(二) 題名（Title）：作品名稱。

例子：Title = Geospatial Support Staff Metadata Tutorial。

例子：Title = 都柏林核心集與元資料實驗系統。

(三) 著者（Creator）：作品的創作者或組織。

例子：Creator = Abeyta, Carolyn。

例子：Creator = 吳政叡。

(四) 簡述（Description）：文件的摘要或影像資源的內容敘述。

(五) 出版者（Publisher）：負責發行作品的組織。

(六) 其他參與者（Contributors）：除了著者外，對作品創作有貢獻的其他相關人士或組織。

〔註: 如書中插圖的製作者。〕

(七) 出版日期（Date）：作品公開發表的日期，建議使用如下格式–YYYY-MM-DD 和參考下列網址：http://www.w3.org/TR/NOTE-datetime。在此網頁中共規範有六種格式，都是根據國際標準日期暨時間格式 – ISO（國際標準組織）8601制定而成，是ISO 8601的子集合（subset），現在列舉和解說如下以供參考：[註 9]

(1) Year（年）-- YYYY。

例子：1997（西元1997年）。

(2) Year and Month（年、月）-- YYYY-MM。

例子：1997-09（西元1997年9月）。

(3) Complete date（完整日期）-- YYYY-MM-DD。

例子：1997-09-07（西元1997年9月7日）。

(4) Complete date plus hours and minutes（完整日期加時、分）-- YYYY-MM-DDThh:mmTZD

〔註：T用來隔開日期和時間，TZD表示本地時間和國際格林威治時間的差距（時間差）。〕

例子：1997-09-07T19:05+08:00（西元1997年9月7日台灣下午7點5分，而台灣所屬的中原標準時區與國際格林威治時間差8小時）。

(5) Complete date plus hours, minutes, and seconds（完整日期加時、分、秒）-- YYYY-MM-DDThh:mm:ssTZD

例子：1997-09-07T19:05:25+08:00（西元1997年9月7日台灣下午7點5分25秒）。

(6) Complete date plus hours, minutes, and seconds（完整日期加時、分、秒）-- YYYY-MM-DDThh:mm:ss.sTZD

例子：1997-09-07T19:05:25.25+08:00（西元1997年9月7日台灣下午7點5分25又1/4秒）。

(八) 資源類型（Type）：作品的類型或所屬的抽象範疇，例如網頁、小說、詩、技術報告、字典等，建議參考下列網址：http://sunsite.berkeley.edu/Metadata/types.html。在上述網頁中將作品的類型粗分成六種，現在列舉和解說如下：[註 10]

(1) Text（文字）-- 作品的內容主要是文字（可夾帶影像、地圖、表格等），例如書籍、文集、技術報告、小冊子等。

例子：<META NAME="DC.type" CONTENT="Text">。

(2) Image（影像）-- 相片、圖形、動畫、影片等。

(3) Sound（聲音）-- 各式各樣的聲音，例如演講、音樂等。

(4) Software（軟體）-- 可執行的程式（二進制檔）和程式的原始檔，但不包括各種互動式應用程式。

(5) Data（資料）-- 各種文字或數據資料的集合體，例如地理資料、書目記錄、統計數據、遙測資料等。

(6) Interactive（互動式應用）-- 設計給一個或多個使用者的互動式應用，例如遊戲軟體、線上聊天服務、虛擬實境等。

以上的六種類型又以第一種類型（Text）最為繁複，可再細分如下：

(1) Abstract（摘要）-- 其他文件的簡要敘述。

例子：<META NAME="DC.type" CONTENT="Text.Abstract">。

(2) Advertisement（廣告）-- 如徵人啟事。

(3) Article（論文）。

(4) Correspondence（書信）-- 可再細分為討論、電子郵件、信件、明信片四類。

例子：<META NAME="DC.type" CONTENT = "Text.Correpondence.Email">。

(5) Dictionary（字典）。

(6) Form（表格）。

(7) Homepage（WWW首頁）。

(8) Index（索引）。

(9) Manuscript（手稿）。

(10) Minutes（會議紀錄）。

(11) Monograph（專論） -- 如書籍。

(12) Pamphlet（小冊子）。

(13) Poem（詩）。

(14) Proceedings（會議論文集）。

(15) Promotion（促銷文件）。

(16) Serial（連續性出版品） -- 可再細分為期刊、雜誌、報紙、時事通訊四類。

(17) TechReport（技術報告）。

(18) Thesis（學位論文） -- 可再細分為碩士、博士二類。

例子：Type = Text.Dictionary。

例子：Type = 文字.技術報告。

(九) 資料格式（Format）：告知檢索者在使用此作品時，所須的電腦軟體和硬體設備，例如 text/html（MIME格式）、ASCII、Postscript（一種印表機通用格式）、可執行程式、JPEG（一種通用圖像格式）。亦可擴展至非電子文件，例如book（書本）、叢書、期刊。

例子：Format = text/html。

例子：Format = 叢書。

(十) 資源識別代號（Identifier）：字串或號碼可用來唯一標示此作品，例如URN、URL、ISSN、ISBN等。

例子：Identifier（scheme = URL） = http://www.blm.gov/gis/meta/barney/tut_met1.html。

(十一) 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 = http://www.blm.gov/。

(十二) 來源（Source）：作品從何處衍生而來（同內容範疇），例如莎士比亞的某個電子書出自那個紙本。

(十三) 語言（Language）：作品所使用的語言，建議遵循 RFC 1766 的規定，請參考下列網址：http://ds.internic.net/rfc/rfc1766.txt，RFC 1766 是使用 ISO 639的二個字母的語言代碼。[註 11]

例子：Language = en。（English）[註 12]

(十四) 涵蓋時空（Coverage）：作品所涵蓋的時期和地理區域。

(十五) 版權規範（Rights）：作品版權聲明和使用規範。可能值如下：[註 13]

(1) 空白（Null）：無特別聲明，使用者須自行參考其他來源。

(2) 無限制（No Restriction on Reuse）：可複製再傳播。

(3) 參考處（URI or Other Pointer）：使用的相關說明，在所指定的出處。

例子：Rights = 無限制。

以上的15個資料項中，某些是針對電腦作業環境而設計的，如資料格式（Format），其他如資料類型（Type）、關連（Relation）、來源（Source）等，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容易使用，幾乎所有的資料項都有自我解釋的功能，大部份人在短時間內就知道如何來使用，根據作者最近一次所做的研究實驗，在經過短暫的練習和熟悉系統後，平均1-3分鐘可完成一篇網頁的著錄工作。

在都柏林核心集的修飾詞發展方面，由於1997年3月在澳洲坎培拉（Canberra）舉辦的都柏林核心集的第四次研討會，已正式確立所謂的「坎培拉修飾詞」（Canberra Qualifier），主要包括三種修飾詞 -- 語言修飾詞（lang）、架構修飾詞（scheme）、次項目修飾詞（subelement）。同時根據『Syntactic Considerations for the Dublin Core』一文，HTML 4..0中也加入兩項屬性 – LANG 和 SCHEME [註 14]，由此可見修飾詞的使用，已是柏林核心集發展中的一個必然趨勢。語言修飾詞（lang）如上所述是遵循RFC 1766的規定，例如<META name="DC.creator" lang="zh" content="吳政叡">。架構修飾詞（scheme）和次項目修飾詞（subelement）則不如語言修飾詞（lang）的發展來得明確，有關修飾詞的相關內容與著錄標準，請參考作者的著作『都柏林核心集與元資料系統』一書。[註 15]

三、研究動機與實驗設計

為了改善檢索引擎高回收率但低精確率的缺失，元資料須扮演過去圖書館中類似目錄的功能，過去目錄在的圖書館運作中所扮演的重要角色之一，即是透過對書籍（或資料）的適當描述，使讀者可以快速的從目錄中找到所欲找尋的資料，而不必到書架上一本本的來翻閱，因此可以提高檢索的效率。另一方面，仔細觀察今日使用者在操作檢索引擎的過程，不難發現使用者最受困擾的地方有二：一是回覆款目過多，無法一一來加以過濾；一是回覆款目的資訊過少，不易判斷是否為所需的資料，因此須一一取回原文來加以檢視，但是取回原文的過程，又因網路擁塞，常常是費力又耗時。

如果使用簡單易用的元資料（如都柏林核心集），即可提供適當的資訊給使用者做為判斷是否須取回原文的依據，則上述兩個問題將可獲得解決。一來使用者不必浪費心力去取回不需要的文件，減少網路的交通和擁塞；二來這些元資料所提供的資訊，將來也可以做為機器自動判斷和過濾的基礎，來減少回覆款目的數量，達成提高檢索準確度的目的。

為了驗證都柏林核心集在減低檢索失誤率（RER）上的可能效用（有關檢索失誤率的詳細說明，請參考作者的另外一篇著作『資訊的檢索失誤率探討』 [註16]），作者設計了以下的實驗，來比較與探討都柏林核心集和檢索引擎所提供的資訊，對使用者判斷文件的影響和差異。

參與實驗的實驗者背景資料如下：參與實驗者為選修作者在輔仁大學圖書資訊研究所開設的「元資料概論」課程的研究生，總人數為7人，包含4男和3女。其中4人大學主修是非圖書資訊相關科系，但對電腦的使用較為熟悉；另外3人則為圖書資訊相關科系畢業或圖書館的工作者，但對電腦的操作熟練程度不如非圖資科系的4人。所有7名研究生均為初次接觸都柏林核心集，實驗前所受的相關訓練有二：一是大約一小時關於都柏林核心集15個欄位的簡略解說，一是大約二小時關於所使用系統--MES（作者自行建立的「元資料實驗系統」）的示範操作和解說。

實驗過程如下：研究生二人為一組，一人扮演讀者，一人扮演參考館員，由扮演讀者的研究生出一個題目，然後扮演參考館員的研究生自行選擇一個檢索引擎來搜尋，參考館員從檢索引擎所回覆的款目中挑選20筆（從回覆款目的第一筆挑選起，但是扣除純介紹性的網站首頁等不適合的款目），將這20筆的回覆款目印出，這些由檢索引擎所提供的資訊，組成實驗中的對照組。然後參考館員將這20筆款目的原始文件一一下載取回，然後根據原始文件，利用作者建立的元資料實驗系統（MES），使用都柏林核心集來加以著錄，並將這20筆的都柏林核心集資料做為實驗組。

首先，扮演參考館員的研究生將對照組的資料（即檢索引擎所提供的資訊），拿給扮演讀者的研究生勾選，選出他或她認為需要的文件（或款目）有那些。接著再給讀者實驗組的資料（即都柏林核心集所提供的資訊），請其勾選需要的文件。最後給讀者看原始文件，請其勾選需要的文件。

四、實驗結果整理與分析

前面所提的三組資料中，原始文件無庸置疑是最準確的，讀者看到原文後，自然可以知道此文件是否為其所需，因此可以做為實驗組和對照組資料比較的依據。同時由於柏林核心集所提供的資訊，較檢索引擎所提供的豐富，所以理論上實驗組與原始文件的差異，會較對照組與原始文件的差異來得小，此研究希望透過這個粗略的初步實驗，可以窺知實驗組和對照組的差異有多大，以及都柏林核心集所提供的資訊，是否充分到可替代原始文件，做為判斷需要的依據。

在實驗組和對照組與原始文件比較時，可能發生的失誤有兩種：第一型失誤是在閱讀實驗組或對照組的資料時，讀者認為需要，但在閱讀原始文件後，判

定非為其所需要的資料；第二型失誤剛好相反，是在閱讀實驗組或對照組的資料時，讀者認為不是他或她需要的資料，但在閱讀原始文件後，發現是需要的資料。

實驗結果經過歸納整理後，列表如下：（每一題目有20篇文件）

表1. 實驗結果。

組別		一	二	三	四	五	六	七
題目		Unicode	圖書館利用教育	Distribution Searching	Information Filter	Asia Financial Crisis	Z39.50	Metadata
檢索引擎		HOTBOT	GAIS	八爪魚	LYCOS	EXCITE	INFOSEEK	YAHOO
原始文件中需要文件的數目		14	12	0	2	17	14	17
實驗組中需要文件的數目		14	14	0	2	16	14	18
對照組中需要文件的數目		13	17	8	2	9	15	19
實驗組	第一型失誤	0	2	0	0	0	0	1 (5%)
	第二型失誤	0	0	0	0	1 (5%)	0	0
	總失誤	0	2 (10%)	0	0	1 (5%)	0	1 (5%)
對照組	第一型失誤	2 (10%)	5 (25%)	8(40%)	0	0	1 (5%)	2 (10%)
	第二型失誤	3 (15%)	0	0	0	8 (40%)	0	0
	總失誤	5 (25%)	5 (25%)	8 (40%)	0	8 (40%)	1 (5%)	2 (10%)

上表中的數據簡要解釋如下：

(一) 第一類數據中的需要文件數目，是讀者在看過三組資料後，認為符合其所需的數目，每一題目有20篇文件（或網頁）。為避免影響讀者判斷的準確性，三組資料的閱讀順序是依其所含資訊的多寡而定（由寡到多），因此閱讀順序為：對照組（檢索引擎）-> 實驗組（都柏林核心集）-> 原始文件。

(二) 第二類數據是失誤筆數和檢索失誤率（RER）的統計，實驗組和對照組分開統計，檢索失誤率是計算失誤筆數佔總筆數（20）的百分比，公式如下：

$$檢索失誤率 = \frac{失誤筆數}{總筆數} \times 100\%$$

例如：失誤筆數為 3 時，檢索失誤率（RER）為15%。

根據這個簡略的實驗，得到以下的初步觀察結果：

(一) 檢索失誤率（RER）隨著題目和所使用檢索引擎的不同，而有很大的差異，這暗示有些檢索引擎的設計方法和所收集資料，祇適合某類資料的查詢。

(二) 以實驗組的都柏林核心集而言，在140筆中祇有3筆第一型失誤，即在看都柏林核心集的記錄時認為相關，但後來閱讀原文時，卻發現不是所需的文件，此外則祇有1筆第二型失誤，整體而言，總共有4筆失誤，檢索失誤率（RER）是2.9%，可以說是非常的低，因此都柏林核心集足以做為過濾文件是否為所需的依據。

(三) 檢索引擎整體有16筆第一型失誤和11筆第二型失誤，總共有29筆失誤，檢索失誤率（RER）是20.7%。由於本實驗的設計是祇採取回覆款目的前20筆，並且已先行扣除一些不相干的款目，所以實驗中的數據，可以說是檢索引擎可能有的最好表現，在實際情境操作時，檢索引擎的表現將會較差。

(四) 某些檢索引擎如Infoseek已採用一些較好的設計，來進一步縮小搜尋的範圍，因此在查尋特定學術性的題目時，已能得到近似都柏林核心集的效果，例如第六組的實驗，題目是Z39.50，使用進階功能，限定Z39.50出現在題名（title），而非文件的任何地方。

(五) 本實驗的規模較小，實驗者的背景屬於高學歷的研究生，題目大多是學術性關聯的，因此祇能視為一個先導性的研究，實驗中所得到的結論，並不能視為是最後的定論，作者將陸續進行一系列的實驗，來做更進一步的探討。

五、結語

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。而如何有效率來過濾和處理大量資料，乃成為亟待解決的課題，作者以為這個問題的解決方案，必須仰賴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，使檢索系統（或檢索者）有足夠的資訊來加強對資料的過濾和處理。綜觀目前大多數的檢索引擎，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「電子式目錄」，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。

總結來說，元資料是因為全球資訊網的作業環境，和電子檔案逐漸成為資料主流等趨勢而興起的資料描述格式。元資料除了負起傳統目錄指引資料和協助檢索的功能外，在格式的設計上，也須能顧及電子檔案所獨有的一些特性，如檔案格式的種類繁多、資料轉換需求頻繁、版本辨識困難等問題。

為了驗證元資料的實際效用，作者選用都柏林核心集做為著錄的元資料，並以選修作者所開設的研究所課程「元資料概論」的研究生為實驗者，設計了一個先導式的簡略實驗，實驗結果證實，都柏林核心集的確可以做為判斷文件是否為所需要的依據，因為檢索失誤率（RER）僅有2.9%，相反的，國內外著名的七個檢索引擎則平均有20.7% 的檢索失誤率。由於這祇是一個先導式的研究，須有一系列的實驗來使結論更為可靠。此次實驗的過程和詳細數據，請參考前面章節中的敘述。

本次實驗的其他發現，是都柏林核心集確有達到創制者們預期的目標—易學好用和快速著錄，非常適合各種背景人士使用，達成「作者著錄」的目的。參與實驗的研究生，祇接受過短暫的欄位解說和元資料實驗系統（MES）示範操作，即可開始進行著錄工作，研究生們反映，在經過短暫的練習和熟悉系統後，平均1-3分鐘可完成一篇網頁的著錄工作。因為幾乎無須打字輸入，祇須在電腦上開二個視窗，一個是網頁文件，一個是元資料實驗系統（MES）的著錄畫面，使用視窗中的剪和貼功能，即可完成所有的著錄工作。

另外一個意外的發現是有少數研究生反映，在下載的國外網頁中，曾有高達25%的網頁已經隱含有都柏林核心集的資料，這說明都柏林核心集在國外已日趨受到重視和被廣泛使用。

誌謝

作者感謝下列研究生參與這次的實驗（女士優先）：謝美玲、黃華明、陳怡佩、陳嵩榮、張政義、陳智泓、林柏吉。

註釋

註 1：T. Berners-Lee, L. Masinter, and M. McCahill, "Uniform Resource Locators (URL)," 1994, <ftp://ds.internic.net/rfc/rfc1738.txt>, p. 1.

註 2：吳政叡，「從元資料看未來資料著錄的發展趨勢」，資訊傳播與圖書館學 3 卷 2 期（民 86 年 12 月），頁44-45。

註 3：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>, p. 2.

註 4：B. Rajapatirana, "The 5th Dublin Core Metadata Workshop: a report and observations," 2 Dec. 1997, <http://www.nla.gov.au/nla/staffpaper/helsinki.html>.

註 5：吳政叡，「元資料實驗系統和都柏林核心集的發展趨勢」，國立中央圖書館臺灣分館館刊 4 卷 2 期（民 86 年 12 月），頁18。

註 6：同註 3，頁 7-11。

註 7：S. Weibel and E. Miller, "Image Description on the Internet: A Summary of the CNI/OCLC Image Metadata Workshop," D-Lib Magazine (Jan. 1997), <http://www.dlib.org/dlib/january97/oclc/01weibel.html>.

註 8："Dublin Core Metadata Element Set: Reference Description," 15 Jan. 1997, <http://purl.org/metadata/dublin_core_elements>.

註 9：M. Wolf and C. Wicksteed, "Date and Time Formats," 15 Sept. 1997, <http://www.w3.org/TR/NOTE-datetime>。

註10：R. Tennant, "Dublin Core Resource Types," 23 Sept. 1997, <http://sunsite.berkeley.edu/Metadata/minimalist.html>.

註11：H. T. Alvestrand, "Tags for the Identification of language," March 1995, <http://ds.internic.net/rfc/rfc1766.txt>, p. 2.

註12："Guide to Creating Core Descriptive Metadata," 13 April 1996, <http://www.ckm.ucsf.edu/people/jak/meta/mguide3.html>, p. 7.

註13：同註7，頁5。

註14：B. Marsh, "Syntactic Considerations for the Dublin Core," 2 Nov. 1997, <http://purl.oclc.org/metadata/dublin_core/syntax.html> , p. 7.

註15：吳政叡，都柏林核心集與元資料系統，（台北市：漢美，民國 87 年），出版中。

註16：吳政叡，「資訊的檢索失誤率探討」，中國圖書館學會會訊 109 期（民 87 年 6 月），出版中。