BIBLID 1026-7220(2001)90 : 3 p.1-5

數位化檔案命名原則

陳昭珍　國家圖書館輔導組兼閱覽組主任
國立臺灣師範大學社教系圖書資訊學組副教授

陳立原　國家圖書館管理師

張文熙　國家檔案局資訊組科長

一、前言

WWW 是網際網路主要的服務，而 URL（Uniform Resource Locators）則是 WWW 運行的基礎。由於網路資源與日遽增，使得資源的定址應用出現困擾，因爲 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，在資料量不斷地成長的情況下，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。我們經常在瀏覽網頁的時候會出現 HTTP v1.0/1.1 Error404 訊息，這表示我們所點選欲連結的資源已經被移除，以致於無法更進一步利用。

如果網路資源能夠提供永久性的（persistent）命名方式，不會受到外界貯存體的變更時而改變，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，索書號是根據分類號及作者號賦予，而不是根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，因爲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），或稱檔案名稱（file name）時，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，因此數位資源必須有一個唯一的名稱，可以被使用者從各種存取途徑參照。

除爲解決數位資源的位置問題而需有唯一的識別碼外，爲了數位資源的電子商務應用，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爲例，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，當要銷售的數位物件比一本圖書、期刊或錄音錄影資料的單位更小時，也需要有適合這種單件的編碼方式才行。

二、直接定址與間接定址

目前網路資源主要是採用直接定址方式，直接定址方式最大好處是簡單、快速，但當資源有異動或修改時，則需要通知並修改相關聯之資料。找尋網路資源就像找尋友人一樣，需依友人所留之電話與之連絡，若友人更換電話號碼，就無法依原碼找到人。同樣的，當某人更換電話號碼時，也須將新碼通知所有親朋好友，通知不到者，可能就會失去聯絡了。

爲了避免上述問題，電話公司推出個人隨身碼，若某人換了工作或換電話號碼，不需通知其他

【專論】

人，只要在電話公司資料庫更改對應之電話即可。網路資源之間接定址原理亦同，透過一台記載檔名與網址對照的主機，若網址更改，只要通知該主機即可，這種主機一般稱爲 handle system。

三、智慧型編碼及非智慧型編碼

除了直間定址與間接定址的觀念外，有關數位物件的命名問題還有智慧型的編碼及非智慧型的編碼之別，而到底要採那一種編碼，也是出版界、圖書館界、網路界等爭論不休的議題。所謂的智慧型編碼是指識別碼有意義，如 ISBN 即爲一種智慧型的編碼，在其十個號碼中共分四部分，第一部分代表出版國，第二部分代表出版社，第三部份是該書的流水號，最後是檢查號。所謂非智慧型的編碼是指識別碼無任何意義，只是一個指向中央資料庫的隨機號碼而已。

到底應給予數位資源那一種編碼方式，各家說法皆不同，反對智慧型編碼者最主要的理由爲：數位資源越來越多，要給予智慧型的編碼有實質上的困難，只要該碼具有「供應性」(affordable) 即可。

四、數位資源命名及定址方式

目前歐美國家已發展出來的物件編碼方式主要有：一致性資源命名（Uniform Resource Names，以下簡稱 URN）和數位物件識別碼（Digital Object Identifier，以下簡稱 DOI）及 SICI code and BICI code 等三種方式，茲說明如下：

1.URNs

由網際網路協會（IETF）1993 年 3 月所提出的一致性資源命名（Uniform Resource Names，簡稱 URN）計畫，用於解決網路資源在連接上的問題，不再只是網路資源位址的指定，而是眞正給予網路資源一個永久性的名稱，以符合目前網路資源發展的需求。URN 的主要觀念是將網路資源名稱與網路資源實體位址獨立開來，透過命名定址系統轉置名稱與位址。

在 RFC 1737 文件中，列出了 URN 的功能需求，主要有下列八項：

(1) 全球性：URN 的命名是以全球網路環境爲應用範圍，而非以區域爲主，因此在任何地點均需有相同意義。

(2) 唯一性：相同的 URN 不會指定給二件不同的資源。

(3) 永久性：URN 的存在是永久的，URN 的存在甚至比所指向之資源更久。

(4) 包容性：可以爲目前所有可能在網路上出現的資源命名。

(5) 相容性：URN 的命名方式必須支援現有的命名系統並滿足他們的需求。

(6) 延展性：任何 URN 的命名方式必須具有延展性，以提供未來發展。

(7) 獨立性：命名方式與解譯系統之間相互獨立，命名方式不會被特定的解譯系統所限制，同樣地解譯系統也能解譯命名方式所指定 URN 的能力。

(8) 解譯性：能將 URN 的名稱轉換爲網路資源位址 URL。

1997 年 5 月 IETF 協會在 RFC 2141 文件中詳細描述 URN 命名語法，URN 命名的開頭字元爲 urn：，分析其結構主要可分爲三部分：

(1) 命名方式：由參與 URN 計畫的各個單位與相關機構自行決定命名方式，包括 hdl、lifn、path、inet 等方式。

(2) 解譯機構：解譯機構爲每一種命名方式的管理 URN to URL 主機位址，並提供相關服務。

(3) 文件名稱：個別文件的名稱。

所有的 URNs 都遵循下列語法及編碼規則：

<URN>：：="urn："<NID>"："<NSS>

<NID> 爲 Namespace Identifier，表示命名方式

<NSS> 爲 Namespace Specific String，爲網路資源的位址（含解譯機構位址及文件路徑及名稱）

1999 年 6 月 IETF 訂定 URN 命名空間機制（詳如 RFC 2611），由 IANA 組織接受各單位申請註冊，命名空間分爲三種層級：Formal（須經由 IETF 組織討論訂定）、Informal、Experimental（不須向 IANA 組織登記），下爲至 2001 年 6 月 7 日已登錄的命名空間：

Registered Formal URN Namespaces	Value	Reference
IETF	1	[RFC2648]
PIN	2	[RFC3043]
ISSN	3	[RFC3044]
OID	4	[RFC3061]
NEWSML	5	[RFC3085]

OASIS	6	[RFC3121]
XMLORG	7	[RFC3120]

Registered Informal URN Namespaces	Value	Reference
urn-1	1	[urn-1]
urn-2	2	[urn-2]
urn-3	3	[urn-3]

註冊 Inforaml URN 命名空間可看 http://www.isi.edu/in-notes/iana/assignments/urn/

整個 URN 系統的實際運作目前並無一完整系統（含解譯、註冊、管理服務）建立，但系統各功能已提出理論或方案，如與現有Internet作業環境結合，則可藉由 DNS（詳見 RFC 2168）和 THTTP（HTTP 功能增強版，詳見 RFC 2169）方式使用，未來期待提供一整合解譯及註冊環境。

URN 的應用原則主要如下：

(1) 並非每一項文件或是資源都要使用 URN，要確認這個文件穩定性高，訊息內容相當有意義才需取得 URN。

(2) 一份文件只能有一個 URN，如果一份文件有多個檔案時，應該視作多份文件，分別給予URN。

(3) 相同內容的文件之複本應該使用相同的 URN，但會有多個 URL 存在。

(4) 不同檔案格式的資源版本應該給予不同的 URN。如同一個文件有 MS-word 及 html 的版本應該給予不同的URN號碼。

(5) 當文件被修改時若只有拼字錯誤之修正，不涉內容修改時，URN 維持不變。如果文件本身已經具有其他的識別系統（ID system），即用原識別系統做爲 URN 中的 NID。如有 ISBN 則其 URN 應爲 URN:ISBN:<ISBN-number>。

2.DOI

DOI 是於 1997 年建立的數位資料命名標準，由在1998年法蘭克福成立的 International DOI Foundation（簡稱 IDF）負責運作，舉凡政策的制定、技術支援、註冊及繳納規費、維護線上的使用指南等，均由該基金會負責執行。系統主要功用就是對網上的內容能作唯一的命名與辨識，藉以保護智慧財產。

目前有二百個公司位使用 DOI 系統，並有四百萬筆以上 DOI 資料註冊，註冊中心（Registrant Agency）有兩個，分別爲 IDF 和 CrossRef。IDF 於 2001 年 2 月提出 The DOI Handbook ver 1.0.0 供全球使用，內容收集 DOI 的技術、建置、管理方式，爲有意加入者提供一入門手冊。DOI 命名的語法主要是遵照 ANSI/NISO Z39.84 標準，其編碼規則如下：

<DOI>=<DIR>.<REG>/<DSS>

<DIR>=10

<REG>Registrant's Code

<DSS>DOI Suffix String

Character set is Unicode 2.0

Case sensitive

<DSS> 的起始字元不能爲*/

.<REG> 碼是由註冊中心發給各要註冊單位

Prefix Suffix

10.1000/123456

DOI

DOI 系統的實際運作目前是採用 Handle System 技術，瀏覽器所需要內嵌（embed）軟體及系統運作軟體可從 http://www.handle.net/ 網址下載。

3.SICI code與BICI

Serial Item and Contribution Identifier（SICI）- Z39.56 -1996 主要是用來與ISSN配合，以辨識某一種期刊或該期刊的某一期或某一篇。SICI code對於目前美國圖書館界在推動的館際合作計畫NAILDD program，以及文獻傳遞服務。除了SICI code 以外，美國的出版界也有BICI code（Book Item and Contribution Identifier）用來辨識套書中的某一冊及某一篇章。

五、URN與DOI的比較

(一) DOI的限制

URN 和 DOI 雖然對統一資源命名的希望是一致的，但這兩大系統生成背景相當不同，所以在本質上及架構上都有所不同。DOI 是由美國出版者協會（AAP, Association of American Publisher）及其所約定的技術合作夥伴 CNRI, Corporation for National Research Initiative 所制定[1]。在運作管理上是以國際 DOI 基金會（IDF，International DOI Foundation）來主持的。IDF 基金會董事會則是由美國出版者協會（AAP）主要的大出版商所組成，如 Microsoft, Elsevier, John Wiley & Sons 等等。由於缺乏圖書館學會及大學在基金會主體內運作，因此圖書館界僅能透過 NISO 組織對 DOI 的發展表達意見，缺乏像美國圖書館學會（American Library Association）或是專門圖書館學會（Special Library Association）等具官方色彩濃厚的單位

加入參與 DOI 制定，在這樣的環境下發展出來的系統，自然難定出符合圖書館界需求之規範。(http://www.press.umich.edu/jep/04-02/davidson.html)。實際上，在 1997 年 IDF 也明確指出了 DOI 系統乃爲迎合出版者的需求而制定，雖然出版者的需求和圖書館界的需求會有重疊，但顯然兩者不能劃上等號。使用 DOI 系統的原因，其中大部份是基於 DOI 可以讓使用者直接從某家出版商的產品連到其他出版商的數位化產品如書目資料庫（Bibliographic database）、引用文獻（Article citation）、摘要（Abstract）或全文（Full-text）。其識別碼系統可以使得傳統的數位產品具有進一步的行動能力（Actionable），意即從識別碼可以指引到資源本身，不再是靜態的表現，這是DOI編碼系統最有價值之處。

不過 DOI 既是出版商所制定的規則，所以它著重在智慧產權（Intellectual Property）的控制[2]，藉以確保出版者的權益。因此實際操作時，使用者必須取得啓始端及被連接端的存取權，或需付費給啓始端。從另一角度來說 DOI 系統是出版社電子商務及電子版權（Copyright）之管理機制，這也美國出版者協學 AAP 之所以對統一的資源名稱有興趣的主要原因。

就學術界的觀點而言，DOI 的系統設計並不完美[3]，因爲 DOI 系統不是提供穩定、可靠、標準化而且免費使用的系統，同時並不是任何人都可使用，需要經過註冊及付費的手續。爲了同時兼顧商業往來及智慧產權的保護，這些出版商也不會把所有的產品都在網際網路釋出。另外，在 DOI 系統問世之前，其實也有其他的控制方式存在，這使得 DOI 系統對出版商的約束力變小。再者，IDF 對參與成員的審核標準也相當嚴格，因此也使得小型出版商不願再付費加入DOI系統。

在編碼規則上 DOI 系統可以長達 128 個字元[4]，對數位世界而言，這種命名長度太長而難以應用，以 Publisher Item Identifier（PII）爲例，只有 17 個字元的長度，就足以識別出版品。根據 Norman Paskin 估計約用 10 的 11 次方個物件就足以完成每一項出版品的唯一識別碼，目前 DOI 的命名法對系統記憶體而言是相當浪費的。

(二) URN的限制

URN 是 IETF 所提出來的計畫，在性質上 IETF 是非營利性質的單位，因此工作小組成員來自不同領域之群體（Community），這和 DOI 的成員組成方式明顯不同，所以 URN 的設計理論上比較能夠包含不同社族的需求。也因爲考慮之適用範圍較大，因此發展的速度就比較緩慢，也不像 DOI 已經有了簡易而統一的運作機制可提供使用者使用。就概念而言，URN 也是爲了使網際網路的資源有單一的識別碼而設計的，而 URN 的解譯（Resolution）服務，就是使 URN 可以轉成適當的 URL 及文件所相關的詮釋資料（Metadata）以便取得資料，但是全球性（Global）的 URN 解譯系統的基礎建設，目前並未建置完成，同時在瀏覽器端與URN解譯行爲之交互動作協定尚未標準化，所以 URN 可能還要再過一段時間才能發展爲成熟的運作環境。

六、American Memory的命名方式

在美國國會圖書館的 American Memory 系統中的每一數位資源，都有一個包含兩個部分的邏輯名稱，此外還有一套嚴謹的規則，用來將檔案儲存在階層式目錄中，以便由邏輯名稱衍生出實際儲存資料的位置。

American Memory 的每一個全集（collection）都有一個不超過八碼的唯一名稱，全集中的每一資料都有一個少於八碼的唯一名稱。例如大部分的影像資料都存成三種檔案格式，此影像檔相關檔案名稱如下：

此影像資料的邏輯名稱爲 "detroit/4a32371"，

而其 thumbnail 名稱則爲 " 4a32371t.gif"

爲經常性下載（routine access）而壓縮的檔案名稱爲 "4a32371r.jpg"

未壓縮（uncompressed）檔案名稱爲 "4a32371u.tif"

而一本小冊子或書的相關檔案名稱如下：

此書或小冊子檔名爲 " nawsa/n7111"

此書有一套 SGML 檔，檔案名稱爲 "n7111.sgm 及 n7111.ent"

此書另有一影像檔，每一頁影像連續命名爲 " n7111001.tif, n7111002.tif ..." 而其插圖及表格又另外命名。

上述的邏輯名稱會記錄存在 MARC 856 的 $d 及 $f，若由一 SGML 文件要連到影像檔或插圖，則使用用此邏輯名稱做爲連結點。而 LC 會將和一 SGML 相關的檔都放在同一目錄下。

目前這些檔案都以階層模式放在 Unix 目錄下，

例如一張照片 detroit/4a32371 的 thumbnail 檔乃存在 /4a/4a30000/4a32000/4a32300/ 目錄下的 4a32371t.gif 檔。

目錄 /4a/4a30000/4a32000/4a32300/

檔名 4a32371t.gif

因此當使用者透過查尋找到一筆 MARC 書目記錄時，系統就會抓到此書目記錄欄位 856 之 $d 及 $f，並結合一個位置表（locator table），產生最後的 URL。這套機制使得 LC 的數位資料檔得以與 MARC 書目資料個自獨立，然而它是由 custom coding，對於完全互通的數位圖書館之長久保存而言，並不合適，因此 LC 也研究 WWW 世界中已被提出來的 URN（Uniform Resource Names）機制，並且選用了 CNRI（Corporation for National Research Initiatives）的 Handle System 及儲存機制（Repository）。

七、數位檔案命名原則

近幾年來，國內相關數位計畫正如火如荼的展開，爲使數位資源也有唯一的檔名，在資料數位化前即須就檔案命名方式加以規範，此命名原則需能滿足下列目的：

（1）資料數位化過程與 Metadata 的建立可分開執行。

（2）依檔名可回溯找到數位化物件。

（3）未來加入國際既有之命名系統時，如 URN、DOI 等，能直接由此檔名加上國家識別碼，而成爲國際間唯一的號碼。

數位資源由各單位分別數位化後，可能會各自儲存在本機構之伺服器，或集中儲存到某一伺服器。換言之，大部分的數位資源都會以分散及集中的方式各存兩套以上，所以，必須能由檔案名稱辨識出這份資料是由那一單位所建立的；此外，每一原始物件爲不同之目的，也會轉換成不同的檔案格式，因此由檔名必須能知道該檔案是那一物件的那一種檔案格式。簡而言之，數位資源的命名原則主要包括：

（4）可以由檔名中辨識此資料是由那一個單位所提供。

（5）此命名方式可支援同一物件之多種檔案格式及其使用目的。

（6）依命名方式在整個系統中，每一數位資源皆有唯一之檔名。

（7）檔案名稱與 Metadata 結合。

（8）符合各種網路資源之命名規則：

1. 使用 ASCII code 命名
2. 檔案名稱英文字大小寫不作區分
3. 不使用 %、/、?、#、*、- 字元

八、與國際命名方式的結合

資源命名是一項複雜的議題，網路資源永久名稱的指定，將是網路資源管理重要的一環，而國內代表中華文化的數位資源未來也必定要往國際化發展。未來將各機關的命名與國際上各種命名方式加以結合其方式主要如下：

命名方式＋註冊機關代碼＋註冊資源代碼

* 命名方式如以 URN 方式則爲 urn，DOI 則爲 doi。
* 註冊機關代碼如爲 URN informal 方式，則由申請機關向註冊中心（IANA）申請分發爲 urn-d（d 爲數字），若爲 DOI，則向註冊中心（FDI 或 CrossRef）申請分發一代碼。
* 註冊資源代碼則由註冊單位內部自編，無一定格式但要內部爲唯一代號。如 URN 則需要提出內部編碼方式給 IANA 協會審查，而 DOI 只要資源識別碼註冊時不與現有重覆即可。
* ＋爲區分碼，如 URN 爲 "："，DOI 爲 "/" 等。

由上分析，不管加入那一個網路資源組織，其註冊資源代碼都是要由註冊機關自訂，因此目前我們設計的檔案命名方式，未來只要再加上註冊機關代碼即可爲國際間唯一的識別碼，如註冊單位是臺灣，則就再加上臺灣的代碼。故在未來不管國際間盛行那一種網路資源組織，都可以快速簡單的轉換成該組織命名方式，使其符合系統擴充性及未來性。

參考資料

RFC1737, 2288 ,2168 ,2169, 2276, 2141, 2276

1. DOI Handbook http://www.doi.org/handbook_200/toc.html
2. Clifford Lynch, "Identifier and Their Role in Networked Information Applications",http://cause-www.colorado.edu/ir/library/html/cem9743.html （1997 OCT）
3. Lloyd A. Davidson and Kimberly Douglas, "Promise and Problems for Scholarly Publishing", http://www.press.umich.edu/jep/04-02/davidson.html
4. ibid.