2002/11/7

台灣視覺記憶計畫主題報告－唯一識別碼

數位典藏國家型科技計劃 後設資料工作組

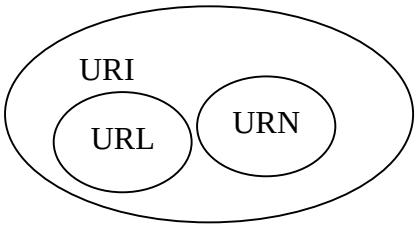
2002/10/25

數位檔案命名規則整理

1. 唯一識別號(URI)的定義與作用

早期的網路資源都採「直接定址」的方式，亦即採用我們最熟知的 URL 來作為網路資源存取的途徑，URL 全名 Uniform Resource Locator，是我們所謂的網址，是已經定義好的格式，例如：ftp:、http:等通訊協定。然而在網路資源以光速成長的情況下，以及資料可能不斷轉移的過程當中，我們很容易就失去了與原來網路資源的連結。因此為重要的網路資源訂定出唯一的識別碼成為勢在必行的趨勢。唯一識別號全名為 Uniform Resource Locators，用來識別 WWW 世界中文件、檔案等各類型的網路資源，URI 是 URN 與 URL 的集合。URN(Uniform Resource Name)是經過特定機構所承諾、永久有效的資源名稱，他用 URN 的機制來標明某一項資源，例如某一本書籍的 URN 為 urn:ISBN:0-201-433357。

其 URL、URN、URI 三者之間的關係如下圖所示[1]：



因此我們在進行數位典藏時，賦予數位資源唯一的檔名，也是保證珍貴的數位資產不會因為儲存位置改變而資料遺失的缺憾，為達成上述之目標，編碼規則的制定是勢在必行的，MAAT 小組提出彙整【技術彙編】以及一些相關技術文件，將目前一般使用的編碼訂定原則整理如下：

2. 數位典藏編碼原則[2]

(1) 命名目的

a. 資料數位化過程與 Metadata 的建立可分開執行。
b. 依檔名可回溯找到數位化物件。
c. 未來加入國際既有之命名系統時，能直接由此檔名加上國家識別碼，而成為國際間唯一的號碼。

(2) 命名原則

---

[1]張錦堂 2001 “ XML 的名稱領域(Namespaces) ” available at http://sinica.edu.tw/~ctchang/mydoc/namespaces.html

[2]計劃辦公室 2002 ”「數位典藏國家型科技計畫」技術彙編 ” available at http://www.ndap.org.tw/TechReport/intro.shtml

數位資源由各單位分別數位化之後，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或是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哪個單位建立的；此外，每一原始物件為不同之目的，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哪一物件的哪一種檔案格式。簡而言之，數位資源的命名原則包括：

a. 可以由檔名辨識此資料是由哪一個單位所提供
b. 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
c. 依命名方式在整個系統中，每一數位資源皆有唯一之檔名
d. 檔案名稱與 Metadata 結合
e. 符合各種網路資源之命名規則：
   - 使用 ASCII code 命名
   - **檔案名稱英文字大小寫不作區分**[3]
   - 不使用%、/、?、#、*、-字元

3. 本計劃檔案編碼目前情況以及建議處理方式

(1) 目前情況

	台史所檔案命名情況	北藝大檔案命名情況	MAAT 建議處理方式
藏品原件	採十五碼制，前三碼為英文字母，分別為DTW(原件)、DTS(正片)、DTN(負片)、DTP(相片)，第二段四碼數字為入藏批號，第三段四碼為件數，第四段為破折號加上三碼數字代表頁次。如：DTW13200012-019。	原件：共八碼，由前二碼為民國年代，加破折號之後是當年同一批入藏的流水號。如：86-003840。	原始著錄號能夠顯示出當時主題計畫處理藏品的方式，建議保留，著錄在影像實體層次的「原始典藏號」欄位中。
數位檔案	未開始進行數位化。	若已經過數位化的藏品還會有個數位典藏號，共五碼，第一碼為分類代號 A-L，分類表如下所示： **A 刑律　E風物**	台史所希望能夠遵循北藝大的數位化編碼方式，因此請兩造討論北藝大原提供之分類法是否適用?

[3] 但在 DOI 在其規範當中，建議數位檔案在命名時必須遵守 CASE SENSITVE 的原則，亦即大小寫的區分是必須的，MAAT 小組建議主題計劃在進行數位檔案命名時，也須區分大小寫。

	台史所檔案命名情況	北藝大檔案命名情況	MAAT 建議處理方式
		I 藝文　B 軍政 F營繕　J 禮俗 C 運輸　G 厚生 K 高砂　D產業 H 教化L 其他 後四碼也為流水號，如：A0001。其數位檔命名分別為A0001.gif(thumbnail)、A0001.jpeg(提供瀏覽)、A0001.eps(不開放)	

4. 建議應用方法

綜合上述之數位檔案編碼規則，MAAT 小組提供對應之解決方法如下：

	編碼原則	建議應用方法
1	可以由檔名辨識此資料是由哪一個單位所提供	北藝大，台史所各取一個簡稱，如北藝大：ta(traditional art)，台史所 th(Taiwan history)，以茲辨別各自所建立的資料
2	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	主題計劃對於數位化檔案的規範如下: .gif 檔為縮圖、.jpeg 為提供瀏覽使用、.eps 則不開放使用。建議參考科博館(附錄一)，以及American Memory(附錄二)以影像檔功能上的區別來作命名。 如：taA0001u.eps(北藝大編號A0001 的未壓縮檔，不開放使用)
3	依命名方式在整個系統中，每一數位資源皆有唯一之檔名	在影像實體層次中的電子檔檔名欄位中分別著錄三種不同規格的數位檔檔名，以茲區別。
4	檔案名稱與 Metadata 結合	
5	符合各種網路資源之命名規則	

5. 與國際命名方式結合

未來將各機關的命名與國際上各種命名方式加以結合方式如下：

命名方式+註冊機關代碼+註冊資源代碼

- 命名方式如以 URN 方式則為 urn，DOI 則為 doi

- 註冊機關代碼如為 URN informal 方式，則由申請機關項註冊中心(IANA)申請分發為 urn-d(d 為數字)，若為 DOI，則向註冊中心(FDI 或 Cross Ref)申請分發一代碼。
- 註冊資源代碼則由註冊單位內部字編，無一定格式但要內部為唯一代號。URN 需要提出內部編法方式給 IANA 協會審查，而 DOI 只要資源識別碼註冊實不與現有重複即可。
- +為區分碼，URN 的區分碼為 ”:”`,DOI 則為 ”/”。

由上可知，不管加入哪一個網路資源組織，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為國際間唯一的識別碼，如註冊單位為台灣視覺記憶計劃，便加上台灣視覺記憶計劃的代碼。

附註：詳細 DOI 以及 URN 之介紹請見附錄三。

2002/11/7

附件二－Case of American Memory[4]

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為 "detroit/4a32371" ，

而其 thumbnail 名稱則為 " 4a32371t.gif" 為經常性下載（routine access）而壓縮的檔案名稱為 "4a32371r.jpg"

未壓縮（uncompressed）檔案名稱為 "4a32371u.tif"

而一本小冊子或書的相關檔案名稱如下：

此書或小冊子檔名為 " nawsa/n7111"

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

此書另有一影像檔，每一頁影像連續命名為 " 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為連結點。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）。

---

[4]以上節錄自 陳昭珍 2000 “數位化檔案命名原則” available at http://www.ncl.edu.tw/pub/c_news/89/01.html

附件三－URN 與 DOI 之簡介(節錄自陳昭珍 2000 “數位化檔案命名原則” available at http://www.ncl.edu.tw/pub/c_news/89/01.html)

以下列出兩種國際上較為著名之編法方式提供給主題計劃參考[5]：

(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, 而是真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為應用範圍, 而非以區域為主, 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：可以為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 的名稱轉換為網路資源位址 URL。

1997 年 5 月 IETF 協會在 RFC 2141 文件中詳細描述 URN 命名語法, URN 命名的開頭字元為 urn：, 分析其結構主要可分為三部分：

（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：解譯機構為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＞ 為 Namespace Identifier, 表示命名方式

＜NSS＞ 為 Namespace Specific String, 為網路資源的位址（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, 命名空間分為三種層級：Formal（須經由 IETF 組織討論訂定）、Informal、Experimental（不須向 IANA 組織登記）, 下為至 2001 年 6 月

[5]以上節錄自 陳昭珍 2000 “數位化檔案命名原則” available at http://www.ncl.edu.tw/pub/c_news/89/01.html

2002/11/7

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），即用原識別系統做為 URN 中的 NID。如有 ISBN 則其 URN 應為 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，分別為 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，為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　　Prefix Suffix

Character set is Unicode 2.0　　10.1000/123456

Case sensitive　　DOI

<DSS> 的起始字元不能為*/

<REG> 碼是由註冊中心發給各要註冊單位

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。