

10-6 資料數位化與命名原則

研發單位：教育部、國家圖書館

提供單位：國家圖書館

使用單位：國家圖書館

一、前言

就主題而言，各種數位圖書館、數位博物館或文化資料庫所包括的資料類型可能有：音樂、戲劇、舞蹈、美術、文學、電影、建築、傳統文化等各種主題；然而就媒體類型而言，則可分為：文字、影像、聲音、視訊等四種型態。由於原始資料之來源不同，因此也需要不同的數位化方式。

各種資料數位化時，必須考量到使用者的設備、使用的便利性、資訊檢索的求、網路上資料的傳輸速度、資料的永久保存等問題，以分別建立幾個不同使用目的的檔案。各種資料數位化的處理原則、檔案格式及驗收規範，詳細說明如下，以供各單位數位化時之參考。

二、處理原則

1. 數位化方式依照原件之媒體類別及下列檔案格式處理。
2. 原件為電子檔之文獻，除將原檔儲存外，建議轉成 RTF、PDF 或 HTML 三種檔案格式。
3. 原件內容為影像者，建議掃瞄一永久檔，永久檔以原型資料大小進行數位化，以不壓縮方式貯存（1：1 模式），並於需要時，再以此永久檔轉成其他目的之檔案。
4. 原件若無法直接掃描，可成拍成光學透射正片，建議尺寸為：4in*5in 或 8in* 10in。
5. 解析度之選擇以原件之品質及數位化之目的為依據。如一般印刷品及照片的品質為 250dpi 至 300dpi，若為重製使用，則需要依原件尺寸取 300dpi 以上來計算作為數位化掃描的各種參數依據（即若原件為 8in*10in 拍成 4in*5in，則解析度須設為 600dpi 或放大兩倍後之解析度為 300dpi）。
6. 原件為聲音及視訊資料者，需輸出兩種型式之數位檔案：一則供永久存檔，一則提供即時播放使用。
7. 對於各輸入設備必須做色彩校正，校正之規範採 IT8 規範色彩導表 7.1、7.2、7.3 版之規範處理。

三、檔案格式

文字資料

1. 原始資料為電子檔

若原始資料是以電腦打字的電子檔，除存一份原始檔外，建議轉成 HTML、PDF 或 RTF 三種檔案格式。

2. 原始資料為印刷品或手稿

原始資料為手稿或印刷資料，如果要做全文檢索，則需重新打字，其餘處理方式同上；若不做全文檢索，則只須建立詮釋資料，並將原件掃描，並以不壓縮格式，儲存一份永久檔，再視須求轉成其他目的之格式，如網路下載格式及預覽格式，此三種格式之規格建議如下：

| 檔案目的 | 說明 | 建議檔案規格 |
|----------|---|---|
| 資料永久保存格式 | 將資料數位化典藏，保持原有風貌。提供使用者作為重製、壓縮處理或其他圖像處理交換之用 | 檔案格式：TIFF 色調深度：黑白；灰階-每像素 8-bits；彩色-每像素 24-bits 壓縮：不壓縮 解析度: 300~600 或更高 dpi (依原始資料品質及重要性選擇適當解析度，一般印刷品可採 300dpi) |
| 網路下載格式 | 提供使用者網路上觀看及列印用 | 檔案格式：JBIG or JBIG2 色調深度：黑白；灰階-每像素 8-bits；彩色-每像素 24-bits 壓縮：JPEG(灰階壓縮比約 10:1，彩色壓縮比約 10:1) 解析度: 150dpi~300 dpi 或影像大小：從 500x400 至 1000x700 pixels |
| 預覽影像 | 提供使用者預覽及選擇欄位用 | 檔案格式：GIF 色調深度：每像素 8-bits 壓縮：原生影像至 GIF 解析度或影像大小: 72dpi 或影像大小從 150x100 到 200 x 200 pixels |

影像資料

若原始資料為照片、圖片、地圖等，則需以掃描器掃成影像檔，並以不壓縮格式儲存一份永久檔，再視需求以永久檔轉存成其他目的之格式，如網路下載格式及預覽格式，此三種格式之規格建議如下：

| 檔案目的 | 說明 | 建議檔案規格 |
|-----------|---|---|
| 資料永久保存格式 | 將資料數位化典藏，保持原有風貌。提供使用者作為重製、壓縮處理或其他圖像處理交換之用 | 檔案格式： TIFF 色調深度：灰階-每像素 8-bits；彩色-每像素 32-bits 壓縮：不壓縮 色彩濃度值 4.0D 以上(color)，3.2D(B&W) 解析度: 300~600 或更高 dpi (依原始資料品質及重要性選擇適當解析度，一般印刷品可採 300dpi，美術品供複製畫使用建議採 600dpi，供印刷出版使用採 350dpi) |
| 資料服務/參考格式 | 提供使用者網路上觀看及列印用 | 檔案格式： JFIF(JPEG 交換格式) 色調深度：灰階-每像素 8-bits；彩色-每像素 24-bits 壓縮： JPEG(灰階壓縮比約 10:1，彩色壓縮比約 20:1) 解析度: 150dpi~300 dpi 或影像大小：從 500x400 至 1000x700 pixels |
| 縮圖影像 | 提供使用者預覽及選擇欄位用 | 檔案格式： GIF 色調深度：每像素 8-bits 壓縮： 原生影像至 GIF 解析度或影像大小: 72dpi 或影像大小從 150x100 到 200 x 200 pixels |

聲音資料

若原始資料為錄音帶、CD 或 LP 等有聲資料媒體，則需將該媒體轉成數位檔案，並儲存一份不壓縮之永久檔，再視需要轉成其他目的之檔案，如下表所建議：

| 檔案目的 | 說明 | 建議檔案規格 |
|--------------------|---|---|
| 資料永久保存格式 | 將資料數位化典藏，保持原有風貌。提供使用者作為重製、壓縮處理或其他圖像處理交換之用 | 檔案格式：WAVE (Microsoft 格式) 取樣率：44.1 kHz 16~ 24 bit/sample 聲道：立體音 |
| 串流(streaming) 寬頻格式 | 提供以寬頻方式上網之使用者線上聽取 | 檔案格式：MP3 取樣率：44.1KHZ 聲道：立體音 頻寬：128kbps |
| | | 檔案格式：RA 格式 編碼種類：雙頻 ISDN 立體音 取樣率：16KHZ 頻寬：100Kbps |
| 串流(streaming) 窄頻格式 | 提供以窄頻方式上網之使用者線上聽取 | 檔案格式：MP3 取樣率：22.05KHZ 聲道：立體音 頻寬：56kbps |
| | | 檔案格式：RA 格式 編碼種類：ISDN 立體音 取樣率：8KHZ 頻寬：50Kbps |

視訊資料

若原始資料為錄影帶，則需以視訊擷取軟體將之轉為數位視訊檔，並存成下列五種檔案格式。

| 檔案目的 | 說明 | 建議檔案規格 |
|--------------------|---|--|
| 資料永久保存格式 (1) | 將資料數位化典藏，保持原有風貌，此為 VCD 的品質 | 檔案格式：mpg 壓縮方式：MPEG-1 圖片大小：352x240 像素 視訊解碼畫面變化率：29 個畫面/秒以上 音效解碼為立體雙聲道 CD 音質 資料傳輸率(Data rate)：約 1.2 megabits/秒 (約 150 kilobytes/秒) Jitter 標準畫面時間：9 mSec 或更優 平均同步位移：1 mSec 或更優 標準裝置同步位移：15 mSec 或更優 音效資料串流(nChannels)：2 或更優 音效資料串流(nSamplesPerSec)：44100 或更優 |
| 資料永久保存格式 (2) | 將資料數位化典藏，保持原有風貌，此為 DVD 的品質，檔案大小為格式(1)的 3~100 倍，適合需高解析保存物件 | 檔案格式：mpg2 壓縮方式：MPEG-2 圖片大小：720x480 像素 視訊解碼畫面變化率：29 個畫面/秒以上 音效解碼為立體雙聲道 AC3 資料傳輸率(Data rate)：約 6~8 Megabits/秒 (約 150 kilobytes/秒) Jitter 標準畫面時間：9 mSec 或更優 平均同步位移：1 mSec 或更優 標準裝置同步位移：15 mSec 或更優 音效資料串流(nChannels)：2 或更優 音效資料串流(nSamplesPerSec)：44100 或更優 |
| 串流(streaming) 寬頻格式 | 提供以寬頻方式上網之使用者線上觀看 | 檔案格式：mpg4 壓縮方式：MPEG-4 音效解碼為 MP3 Bit-rate(位元傳輸率)：250Kbps 檔案格式：RM Bit-rate(位元傳輸率)：250Kbps Audio Format：22 kHz, stereo |

| 檔案目的 | 說明 | 建議檔案規格 |
|--------------------|-------------------|--|
| | | 檔案格式：ASF Bit-rate(位元傳輸率)：250Kbps Audio Format：22 kHz, stereo |
| 串流(streaming) 窄頻格式 | 提供以窄頻方式上網之使用者線上觀看 | 檔案格式：mpg4 壓縮方式：MPEG-4 音效解碼為 MP3 Bit-rate(位元傳輸率)：56Kbps |
| | | 檔案格式：RM Bit-rate(位元傳輸率)：56Kbps Audio Format：11 kHz,mono |
| | | 檔案格式：ASF Bit-rate(位元傳輸率)：56Kbps Audio Format：11 kHz, mono |

四、物件數位化驗收規範

1. 雇主(以下簡稱甲方)於接獲廠商(以下簡稱乙方)交付之工作項目與產品後，除核對交付清單所列數量及項目是否相符外，就數位化檔案之品質，依中國國家標準（CNS）2779 Z4006（數值檢驗抽樣程序及抽樣表）之規定，採用 III 級一般檢驗水準，正常檢驗單次抽樣計畫型式，允收上限 AQL=1.0。
2. 文件影像品質驗收基準
 - (1) 符合各資料數位檔案規格
 - (2) 在原稿清楚的前提下，影像文字清楚。本項所稱之「清楚」係以雷射印表機在 300dpi 解析度下所印出之原件影像為判斷依據。
 - (3) 影像去除因掃描引起的黑邊、污點。
 - (4) 影像歪斜不可超過 1 度。
 - (5) 檔案名稱編碼符合甲方指定之檔案命名方式。
3. 聲音、影片驗收基準
 - (1) 符合各資料數位檔案規格。
 - (2) 原物件與數位化永久保存級檔案同時播放比較，無明顯差別。
 - (3) 檔案名稱編碼符合本契約中檔案命名方式。
4. 甲方驗收本工作時，若認為本工作之內容、品質或數量有未符合規定時，得要求乙方於一定之工作天內，就不合格之項目重新製作，並完成相關交付作業，再行驗收。
5. 驗收所需之器材及設備由乙方準備。

五、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，為使數位資源也有唯一的檔名，在資料數位化前即需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. 未來加入國際暨有之命名系統時，如 URN、DOI 等，能直接由此檔名加上國家識別碼，而成為國際間唯一的號碼。

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，大部份的數位資源都會以分散及集中的方式各存兩套以上，所以，必需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，每一原始物件為不同之目的，也會轉換成不同的檔案格式，因此由檔名必需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：

1. 可以由檔名中辨識此資料是由那一個單位所提供
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
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。
4. 檔案名稱與 **Metadata** 結合
5. 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

六、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

1. 命名方式如以 URN 方式則為 urn，DOI 則為 doi。
2. 註冊機關代碼如為 URN informal 方式，則由申請機關向註冊中心 (IANA) 申請分發為 urn-d (d 為數字)，若為 DOI，則向註冊中心 (FDI 或 CrossRef) 申請分發一代碼。

3. 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為唯一代號。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。
4. +為區分碼，如 URN 為“：”，DOI 為”/”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為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