教育資料與圖書館學

Journal of Educational Media & Library Sciences

http://joemls.tku.edu.tw

Vol. 46 , no. 3 (Spring 2009) : 323-349

數位典藏資訊系統之長期保存規劃與實施

Planning and Implementing Long-term Preservation for Digital Archive System

林信成 Sinn-Cheng Lin
Associate Professor
E-mail: sclin@mail.tku.edu.tw

鄭國祥 Kuo-Hsiang Cheng
Ph. D. Student
E-mail: khcheng@mis.tsint.edu.tw

English Abstract & Summary see link at the end of this article

教育資料與圖書館學　46 : 3 (Spring 2009) : 323-349

數位典藏資訊系統之長期保存規劃與實施

林信成*

副教授
淡江大學資訊與圖書館學系
E-mail: sclin@mail.tku.edu.tw

鄭國祥

淡江大學資訊工程學系博士研究生
北台灣科學技術學院資訊管理系兼任講師
E-mail: khcheng@mis.tsint.edu.tw

摘要

本文以國科會數位典藏公開徵選計畫之一的「台灣棒球文化資產數位典藏計畫」所建置的資訊系統為實際案例，針對該系統所面臨之數位資訊長期保存問題，規劃適當的保存方法並採系統實證方式進行實作，以使該系統日益累積的珍貴資訊能長期為使用者檢索與利用。過程中共採用了轉存、標準化、詮釋資料、轉置、重複一套系統建置等作法進行系統升級，再透過網站記錄檔分析檢視系統升級之成效。本研究並建議國科會應儘速成立數位典藏資訊長期保存計畫辦公室，擬定長期保存策略、培訓技術人才、加強產學合作，以因應日益迫切的數位資訊保存議題。

關鍵詞：數位典藏，長期保存，轉置，詮釋資料

前　言

自20世紀末起，數位科技為人類文明帶來衝擊與轉機，各國政府皆致力於以數位化方式保存重要文化資產。如美國國會圖書館自1989年開始推行「美國記憶」(American Memory)計畫，1995年起延續的「國家數位圖書館計畫」(National Digital Library Program)，旨在集合美國歷史與文化的第一手資源，以支援相關的研究(註1)；英國則於1995年由大英圖書館研發部、英國「聯合

* 本文主要作者兼通訊作者。

2008/12/26投稿；2009/03/09修訂；2009/04/04接受

資訊系統委員會」(JISC)及「博物館、檔案館與圖書館委員會」(The Council for Museums, Archives and Libraries)共同贊助設立「英國圖書資訊網路辦公室」(The UK Office for Library and Information Networking，簡稱UKOLN)，是英國數位化計畫的重要單位(註2)；歐盟委託執行之「資訊社會科技計畫」(Information Society Technology，簡稱IST)源自於歐洲「基礎建設計畫」(Framework Programme)，十分注重資訊科技與人文的整合(註3)；加拿大博物館界為了創造豐富的數位化公共資源，於1995年成立「加拿大文化資產資訊網」(Canadian Heritage Information Network，簡稱CHIN)(註4)；澳洲則自1996年起由圖書館界結合商界和研究機構，開始進行「澳洲數位圖書館先導計畫」(註5)；亞洲的日本則自1997年起由國會圖書館訂定「電子圖書館構想工作指導方針」，並成立推動委員會，至2000年推行的「電子圖書館服務實施基本計畫」(註6)。

在這一波數位化潮流中，我國亦不落人後。行政院國家科學委員會自1997年即進行數位博物館先導計畫，並於2002年起開始推動「數位典藏國家型科技計畫」(National Digital Archives Program)，旨在將官方或民間典藏之重要文物、史料、藝術作品……等數位化，建立國家級數位典藏庫，進而促進人文與社會、產業與經濟的發展(註7)。文建會亦在行政院之「網路文化建設發展計畫」之下，推動了「國家文化資料庫」(National Repository of Cultural Heritage)之建置，目的是要系統性的、計畫性的進行文化藝術資源之蒐集、整理和保存，並藉由資訊科技將其數位化典藏，留下文化資源的長久記錄(註8)。

然而，數位典藏系統中的數位資訊(Digital Information)或數位物件(Digital Objects)之生命週期(life cycle)，包括創造(creation)、獲取(acquisition)、編目(cataloging)、識別(identification)、儲存(storage)、保存(preservation)和取用(access)等，皆異於傳統出版品(註9)。不僅如此，由於數位資訊具備修改容易、載體脆弱、資料易流失、機器依存度高、技術生命週期短……等特性，面臨了不同於傳統實體物件的長期保存問題。陳昭珍認為數位資訊保存不易的原因主要受資訊科技的影響，包括：數位資訊無法獨立存在、硬體及軟體易於損壞或過時作廢等因素(註10)。對於近年來經由大量產出的數位物件，Kuny曾提出警語，認為若無法予以妥善管理並長期保存，一旦流失便難以重建，人類的數位文明將可能進入所謂的黑暗期(註11)。因此目前各國產、官、學各界都極重視數位資訊長期保存的議題。

本文以國科會數位典藏公開徵選計畫之一的「台灣棒球文化資產數位典藏計畫」所建置的資訊系統為實際案例，針對該系統所面臨之數位資訊長期保存課題，規劃適當的保存方法並採系統實證方式進行實作，以使該系統日益累積的珍貴資訊能長期為使用者檢索與利用。在本文第一節概述之後，第二節歸納常見的數位資訊保存方法，第三節簡介該系統之概況並描述其所面臨的問題，第四節詳述該系統長期保存所採行之方法及實作，第五節從記錄檔分析、比較

新舊系統的使用狀況與效能，以檢視實驗成果，第六節為結論。

二、數位資訊保存方法

如前所述，數位資訊長期保存是資訊系統所面臨的重要課題之一。依據歐陽崇榮歸納Kranch、Muir、Lawrence、Rothenberg、Wiggins、Cresp、Lorie、Waugh、Pace等眾多學者在數位資訊保存議題方面的研究，至少有九種常見的數位資訊保存方法，如表1所示（註12）。其中第一類的基礎層，包括轉存（refreshing）、標準化（standardization）與詮釋資料（metadata）等，是最基礎的工作，若要實現第二層或第三層策略，基礎層最好先完成；第二類的核心層，包括轉置（migration）、模擬（emulation）與封裝（encapsulation），是數位資訊保存技術中最為重要的，其中的轉置策略是目前政府機關或及企業界最常使用的策略之一；第三類的輔助層，包含系統保存（system preservation）、重複一套系統建置（preservation through redundancy）與印成紙本或其他可瀏覽媒體，其目的就是使用核心層技術保存數位資訊時，有特別的困難或其他因素的考量下，所採取的保存策略。

表1 數位資訊長期保存策略

類型	第一類	第二類	第三類
層次	基礎層	核心層	輔助層
保存方法	轉存	轉置	系統保存
	標準化	模擬	重複一套系統建置
	詮釋資料	封裝	印成紙本或其他可瀏覽媒體

資料來源：歐陽崇榮，數位資訊保存策略（台北市：文華，2006），100。

以下概略描述這九種方法：

㈠轉存

轉存亦可稱為「更新」，指的是儲存媒體的更新，亦即將數位資訊從舊媒體轉存到新媒體上，例如：將磁碟片資料轉存到USB隨身碟，或將磁帶資料轉存到光碟片。此方法可以解決因資訊技術發展快速導致儲存媒體老化、過時，或讀取設備不再存在或不堪使用等問題，避免數位資訊的遺失或無法取用。

㈡標準化

標準化著重於資料內容的標準格式上，包含字碼、檔案格式、Metadata標準、資料交換規格、資料庫結構……等，標準化通常搭配其他方法使用。然而，因各種標準也可能隨時間演進而改變，因此標準化最困難的是如何選擇出合適的、穩定的標準規格，以支援多面向的應用，並避免數位資訊之內容因標準轉移而遺失或減損其原意。

㈢詮釋資料

詮釋資料是指對數位資訊加註說明與描述，以作為電腦系統存取、使用之依據，通常也須搭配其他方法使用。詮釋資料在數位資訊保存方面可分成描述性詮釋資料與保存性詮釋資料。描述性詮釋資料主要針對數位資訊本身內容的描述；保存性詮釋資料主要任務則是記錄數位資訊有關保存方面的資訊，例如：原始軟硬體環境、資料內涵、呈現形式、經歷的保存活動、使用權限……等，以支援保存方法的執行，增加數位資訊永久取用的可能性。

㈣轉置

轉置是目前最常用的長期保存方法之一。其作法是將資訊系統從舊的軟硬體環境轉移到新的軟硬體環境下；或將資料從舊的格式移轉到新的格式上，以利後續發展。轉置牽涉到數位資訊從前一代的軟硬體設備或資料規格轉移到新一代的設備、環境或規格上，必須將數位資訊的內容、架構與關聯性都保存下來，以確保其完整性，讓使用者能在不斷變遷的新科技中繼續使用。

㈤模擬

模擬方法是讓新電腦系統模仿舊電腦系統的運作，主要是藉由模擬器（emulator）作為新舊電腦之間的中介橋樑，讓新一代的電腦可以虛擬運作如舊電腦呈現的畫面和內容。理想上，模擬不但可以確保資料不會遺失，且數位資訊的外觀（look）、感覺（feel）和特有的行為（behavior）都將被一一模擬出來，讓數位資訊回復其最原始的面貌，使數位資訊在新電腦上完整重現。

㈥封裝

封裝方法是將被保存的數位資訊及其相關資料如文件說明、組織活動……等包裹在一個封包裡，融合了Metadata的內涵。換言之，是將數位資訊及描述其內容的Metadata一起包裹起來。封裝用Metadata所包含的除數位資訊基本資料外，尚有支援該數位資訊之原始軟體環境的功能，及組織使用該數位資訊的目的……等相關資料，便於爾後透過解譯、轉換、模擬等方式讀取與取用。

㈦系統保存

系統保存又稱為技術保存（technology preservation），是非常簡單的保存方法，作法是將處理數位資訊的相關電腦系統依現況完整的保存下來，包含電腦硬體及周邊設備、作業系統及應用軟體、資料庫管理系統及數位資訊本身。

㈧重複一套系統建置

此一作法是將整個資訊系統，包含電腦硬體、軟體、資料庫、數位資訊……等，於另一地重複建置一套或設立鏡射場地（mirror site），以做為異地備援之用，是目前許多企業組織的作法。此作法著眼在數位資訊的保護上，以防止數位資訊的毀損。

㈨印成紙本或其他可瀏覽媒體

此方法是將數位資訊列印成紙本、輸出成微縮片或其他類比媒體保存。然而，數位典藏之原始目的是將類比資訊轉化成數位資訊，以便電腦儲存、處理與應用，若又轉回類比資訊，又將遭遇類比媒體保存的問題，且品質也比不上原件；況且有些數位化物件也無法轉回類比物件，例如：實體器物的數位照片或3D模型，頂多印成紙本照片。所以此保存方法有頗多限制。

三、案例探討：台灣棒球文化資產數位典藏系統

㈠系統概況

「台灣棒球文化資產數位典藏計畫」是由淡江大學資圖系數位典藏研究小組執行的國科會數位典藏計畫。該計畫自2004年起開始，歷經不同階段完成了「台灣棒球運動珍貴新聞檔案數位資料館」（註13）、「台灣棒球維基館」（註14）與「台灣棒球數位文物館」（註15）三個不同性質的數位典藏子系統，整體系統架構概述如下：

1. 子系統一：台灣棒球運動珍貴新聞檔案數位資料館。這是一個以台灣棒球歷史性新聞及老照片為主的數位典藏系統。採用Windows＋IIS＋ASP＋JSP的系統開發方式，在系統建置過程中，也一併探討了數位新聞內容語意描述及分析數位化新聞適用的各式Metadata，並開發Metadata轉換及OAI-PMH協定模組，主要在於解決與國家數位典藏聯合目錄的資料交換與整合問題。

2. 子系統二：台灣棒球維基館。這是一個以開放原始碼（Open Source）的Wiki平台所建置的社群協作知識匯集系統，整合知識組織技術並開放網路社群共同編輯，使內容更豐富多元，以彌補上述新聞資料庫的不足，成為更具利用價值的棒球知識庫。系統開發採用了Windows＋Apache＋MySQL＋PHP的方式。

3. 子系統三：台灣棒球數位文物館。這是一個棒球文物管理展示系統，旨在尋求棒球文物典藏單位、棒球文史工作者或收藏家及棒球界人士的合作，期望能達成將分散各地之棒球文物數位化建檔，供社會大眾查詢應用的目的。系統開發採用了Windows＋Apache＋MySQL＋PHP的方式。

圖1為該計畫於2007年資訊展時，整合了上述三個子系統參展之傳單。

圖1 台灣棒球數位典藏計畫2007年資訊展參展傳單

㈡問題描述

在歷經數年運作之後，該計畫雖於2007年資訊展時成功展出，並獲得參觀者的好評，但該系統在數位資訊的保存實務與技術方面，卻開始遭遇一些問題：

1. 初期的電腦設備是配合當時系統規劃而購置，如今隨著系統的擴大與軟體技術更新，原先硬體便不勘負荷。伴隨瀏覽人數的增加，網頁的呈現速度緩慢，無法滿足使用者的期待；加上資料持續增加，資料備份時間過長，導致狀況百出，例如「台灣棒球維基館」的資料備份，每次都須停機超過4小時。

2. 作業系統過時，系統商無法提供修補檔，安全性一再出現漏洞。以「台灣棒球運動珍貴新聞檔案數位資料館」為例，初始建站時的系統是採用Windows 2000 Server，曾有多次被駭客植入惡意軟體（如木馬程式或釣魚程式）的記錄，讓系統維護作業人員疲於奔命，防不勝防。

3. 資料格式更新快速，而維護人員礙於經費、設備、技術無法有效配合，致使系統成長緩慢。例如「台灣棒球維基館」在建站時的Mediawiki舊版資料庫結構與新近版本差異頗大；又如早期編碼採用Big-5，有不少中文字無法在網頁呈現，這已不符合現行的Unicode潮流。

4. 資料庫軟體須儘速更新，如「台灣棒球維基館」的資料庫軟體MySQL，初期採用的版本迄今已過於老舊，其編碼方式是用Latin-1的編碼，與UTF-8的編碼方式有所不同。

5. 應用程式過時且程式語言紛雜，如「台灣棒球運動珍貴新聞檔案數位資

料館」曾採用ASP、JSP等程式語言開發，這與後來建置的兩個系統所採用的PHP程式不同，除了有跨平台的問題外，也增加維護的人力；又如「台灣棒球維基館」的MediaWiki程式沒有升級，許多功能無法擴充。

除了以上較重大的問題外，更有許許多多經常發生的小問題，時時刻刻威脅著系統內數位典藏品的保存與取用，在硬體損壞或軟體過時作廢之前，宜儘速擬定長期保存策略與方法，並著手進行長期保存作業。

四、系統實證研究

為使系統中日益累積的珍貴資訊能長期為使用者檢索與利用，在眾多數位資訊保存方法中，採用了轉存、標準化、詮釋資料、轉置、重複一套系統建置等作法，並進行系統實證，以下將闡述實作結果。

㈠基礎層

1. 轉存

本系統資料保存的基礎作法有二：⑴將所有資料轉存至一部專用的網路儲存設備（Network-Attached Storage，簡稱NAS）內；⑵將資料庫內的數位典藏資料採用XML格式轉匯到數位典藏聯合目錄，如圖2所示。以上兩種作法皆可避免本機儲存媒體將來因老舊而無法讀取之問題，有利資料的長期保存。但若線上運作的主機出了狀況，並沒有備援系統可取代，因此決定額外購置一部伺服器，採用「重複一套系統建置」的方式作為備援方案。

圖2　資料轉存至數位典藏聯合目錄可作異地備份也有利長期保存

2. 標準化與詮釋資料

字碼由早期的Big-5及Latin-1改換成UTF-8的格式；檔案格式則採用目前通行的標準格式如XHTML、XML、JPEG、GIF等；詮釋資料亦採用標準規

格，除了聯合目錄制訂的DAC格式外，還採行了國際通用標準如DC、NITF、RSS等，有利於資訊的交換共享和長期保存。網站程式語言和資料庫方面則採用目前最通行的PHP和MySQL。圖3所示為本系統採用標準化Metadata之一例。

```
–<DACatalog>
  –<AdminDesc>
     <Project Creator="淡江大學資訊與圖書館學研究所" GenDate="">台灣棒球數位文物館</Project>
    –<Catalog>
      –<Record>
          典藏機構與計畫:公開徵選計畫:淡江大學:資訊與圖書館學研究所:台灣棒球文物館與數位典藏系統之建置與整合研究(I)
       </Record>
       <Record>內容主題:器物:棒球文物</Record>
       <Record>地理架構:台灣</Record>
       <Record>時間架構:2005-05-19-2005-05-22</Record>
     </Catalog>
     <DigiArchiveID>MUSEUM1</DigiArchiveID>
     <Hyperlink>http://163.13.175.47/dc.php?autoid=1</Hyperlink>
    –<ICON>
       http://163.13.175.47/collection/2001-2010/B2005-001N.jpg
     </ICON>
   </AdminDesc>
  –<MetaDesc>
    –<Title>
       2005年第23屆亞洲棒球錦標賽亞軍獎牌: 23rd ASIAN CHAMPIONSHIP in MIYAZAKI Second Place
     </Title>
     <subject>國際賽事</subject>
     <subject>獎盃</subject>
    –<Description>
       2005年5月19日-22日，第23屆亞洲棒球錦標賽，在日本宮崎縣舉辦，中華隊獲得了第二名，參加的國家有日本、台灣(中華隊)、中國、南韓、菲律賓、泰國等國(按名次排列)。
     </Description>
     <Publisher/>
```

圖3 系統採用標準化Metadata有利於資訊交換共享和長期保存

㈡核心層：轉置

系統轉置是數位典藏系統長期保存工作的核心，也是本研究關注的重點任務。本研究小組於2007年8月開始實施系統轉置計畫，2008年3月上線測試，2008年5月新系統逐漸趨於穩定運作，順利完成系統轉置。本系統之轉置工程共包含四大工作項目：1.電腦硬體轉置，2.作業系統轉置，3.資料格式轉置，4.應用程式轉置。以下詳細說明轉置過程所遭遇的問題及解決方法。

1.電腦硬體轉置

為提供使用者更優質的環境以取用數位典藏內容，本系統先從電腦硬體轉置工程著手，規劃如圖4所示之系統架構，預計將先前草創時期以直立式伺服器所建置的硬體平台悉數轉換為機架式伺服器，不但易於集中管理，也提升了系統穩定度與可靠度。

另一方面，為了加速網站之連線速度，紓解網路傳輸速率過慢之問題，乃透過校內程序向淡江大學資訊中心申請一條Giga-bit網路專線，並將原有較低頻寬之交換器、路由器一併升級，使網路連線頻寬由原本的100Mbps速率升級為1Gbps。

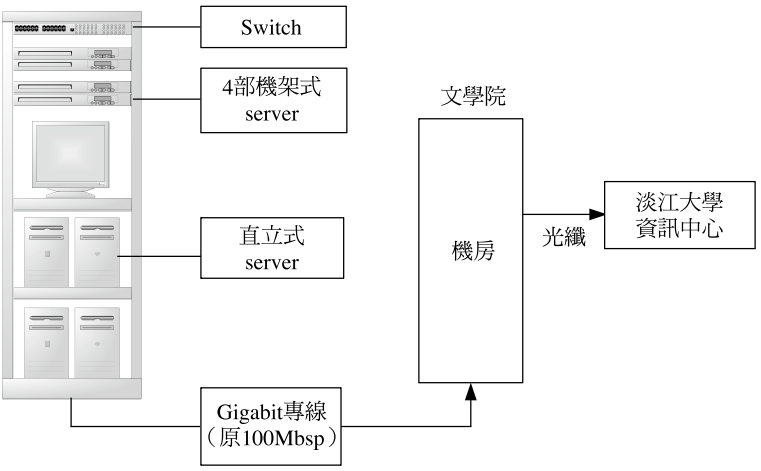

圖4 硬體轉置架構圖

相較於直立式伺服器，機櫃式伺服器最大的優點在於節省空間，另外整齊的堆疊也提升管理方便性，此次硬體轉置的規格如下：

- 中央處理器：2顆AMD Opteron 2214（2.2GHz）
- 記憶體：4GB DDRII 667 ECC Registered
- 硬碟：2顆熱抽取250GB SATA II 7200rpm
- 網路：主機板內建二組Broadcom BCM 5721網路晶片，支援10/100/1000 Mbps
- 顯示晶片：主機板內建ATI ES1000 PCI-based
- 風扇：6個可變速散熱風扇（4大2小）
- 電源供應器：單顆500W
- 尺寸：670 mm（L）×445 mm（W）×43.6 mm（H）

以上規格之新伺服器共採購四部供上述三個子系統使用，其中因「台灣棒球維基館」的流量最大，所以規劃雙主機架構，採用「重複一套系統建置」的方式，以利於數位資訊的保存與利用；而「台灣棒球運動珍貴新聞檔案數位資料館」和「台灣棒球數位文物館」因流量較小，所以規劃單主機架構。

2. 作業系統轉置

⑴選擇作業系統

若將硬體比喻為電腦的軀殼，作業系統則是電腦的靈魂！當今架設網站常見的作業系統以Windows和Linux為最大宗。然而，建站時究竟採用Windows或Linux是見仁見智的問題。本次系統轉置決定將作業系統由Windows轉換為Linux，主要考量的原因如下：

A. 穩定度：Linux穩定度較佳，Yankee Group曾調查比較各種作業系統當機小時數，顯示出unix-like的作業系統遠比Windows伺服器穩定（註16）。

B. 安全性：Linux遭受病毒感染或駭客入侵的機率遠遠低於Windows，安全性威脅較輕。

C. 經濟性：Linux是自由軟體，是免費的，且大部分應用軟體都可自由獲取，對於經費有限的學術研究較適合。

然而，由於Linux的發展呈現百家爭鳴，因此尚須對幾個常見的Linux系統進行比較，才能選擇較佳的方案。

Linux最早是由芬蘭人林納斯．托瓦茲（Linus Torvalds）根據Andrew Tanenbaum教授於1986年完成的教學用小型Unix核心—Minix，搭配bash、gcc等工具，於1991年創造的一套可在Intel的386機器上運作的簡單版Unix-like作業系統。林納斯．托瓦茲後來將Linux以通用公共授權（General Public License，簡稱GPL）方式釋出，讓所有人都可免費使用，之後許多程式設計師陸續為Linux做了重大修改，使Linux隨著時間的演進愈趨完善，如今已成為舉世重要的作業系統之一。

Linux只是個核心程式，需要其他程式搭配。在這之前，任教於麻省理工學院的理查．史托曼（Richard Stallman）教授，於1984年創建自由軟體基金會（Free Software Foundation，簡稱FSF）與GNU（GNU's Not Unix）專案，並不斷編寫創建GNU程式，因此被譽為自由軟體之父，先前提及的bash、gcc及現今世上許許多多自由軟體都出自GNU專案。GNU與Linux完美結合，成就了一套完整的作業系統，現今所說Linux作業系統，應稱為GNU/Linux會更適當。

GNU/Linux目前非常盛行，許多組織和企業基於GNU/Linux研發了自有品牌的Linux發行版，至少已超過百種（註17），無法一一詳述。以下分述較常見者及其特色，作為選擇作業系統時參考。

A. Red Hat系列

國內外Linux用戶最耳熟能詳的Linux發行版當屬紅帽公司的Red Hat系列。紅帽公司最早由Bob Young和Marc Ewing於1995年創建。現在Red Hat Linux發行版分兩系列：收費的Red Hat企業版（Red Hat Enterprise Linux, RHEL）及社群研發的免費Fedora Core（事實上，Fedora除由志願者組織參與外，也有許多紅帽公司正式員工參與）。其最大的長處在擁有數量龐大的使用者，優秀的社群技術支援，並有許多創新，譬如它的yum/rpm就是大家所熟知的套裝軟體管理工具，透過yum/rpm可以有效避免軟體相依性的問題，讓套件安裝更加簡易。

B. Debian系列

Debian由德國人Ian Murdock於1993年創建。它最大的特色是在最遵循GNU規範，100%免費，擁有優秀的網路和社群資源，它也發展出自有的套裝軟體管理工具apt-get/dpkg，媲美紅帽公司的yum/rpm。另外，目前極負盛名的Knoppix及Ubuntu等便是基於Debian的另外一個Linux發行版。

C. SUSE系列

SUSE也是德國的一個發行版。它由四位德國人於1992年末創建，在2004年1月被網威（Novell）收購。現在的SUSE Linux發行版也分成兩個系列：收費的SUSE企業版及由社群研發的免費的openSUSE（由網威資助）。它最大的特色是開發出專業，易用的YaST套裝軟體管理工具。

D. Mandriva

Mandriva原名Mandrake，最早由一位法國青年Gaël Duval於1998年創建，在2005年被拉丁美洲最大的Linux廠商Conectiva收購，之後Mandrake就更名為Mandriva。它最大的特色是擁有友善的操作介面並部分兼容於Red Hat Linux。

雖是百家爭鳴，但由於有Linux Standard Base（LSB）（註18）以及目錄架構的File system Hierarchy Standard（FHS）（註19）標準來規範開發者，另外所有的Linux發行版都是基於相同的Linux核心包裝而成，整體來說差異並不大，較大的差別在於各家所開發出的管理工具及套件管理工具的不同！在台灣由於Fedora較普及，技術交流也較頻繁，加上它是免費的Linux發行版，所以此次作業系統轉置決定採用Fedora。

⑵作業系統轉置之實施

由於Linux有非常多不同的發行版，而欲購入的新伺服器能否順利安裝這些不同的發行版，並順利運作尚未可知。以現況而言，幾家大伺服器廠商如IBM、HP、DELL、ASUS等，僅保證可安裝商業版的Red Hat Enterprise Linux或SuSE Linux Enterprise Linux，其它Linux發行版則不在保證範圍之內，預定採用的Fedora免費版並不在硬體廠商保證範圍，造成了選購伺服器的困擾。解決方式是先考量預算，再鎖定幾種預算內之伺服器，隨後與代理商接洽先借用伺服器進行測試，確定伺服器硬體完全可搭配Fedora之後再開始採購流程。

作業系統由Windows NT Server轉置為Linux後，除節省購買作業系統費用外，最大差別是之前的Windows NT Server是32bit的作業作統（目前中文版Windows NT Server都是32bit的作業系統），其記憶體的使用值理論值是4GB，但實際使用值僅有3GB左右，而所採用的Fedora則選擇64bit版本，記憶體的使用值就沒這樣的限制。轉置前後新舊硬體的重大差別如表2所示。

表2 新舊伺服器硬體比較表

	舊伺服器	新伺服器
中央處理器	單顆32bit Intel單核心CPU	雙顆64bit AMD雙核心CPU
記憶體	1GB SDRAM	4GB DDR II ECC Reg.
硬碟	單顆120GB IDE HDD	雙顆可熱插拔250GB SATA II HDD
網卡速率	100Mb	1000Mb

另一項作業系統轉置的重要考量是人才的培訓，畢竟Windows NT Server

與Linux的管理有非常大的不同，轉置後能有多一些人員來分工負責，遇上突發狀況才能在最短時間內解決。因此在轉置的過程中，也同步舉辦了幾次Linux研習班，培訓Linux系統管理人才，當然除Linux作業系統外，其它幾個重要的服務如MySQL、Apache、PHP都涵蓋在Linux的培訓範圍內。

3. 資料格式轉置

⑴資料庫結構與資料編碼格式

資料格式轉置部分主要包含資料庫結構與資料編碼兩部分。由於三個子系統建置時間不一，應用程式也不同，雖同樣採用MySQL資料庫，但其資料庫結構皆不相同，必須一一檢視、對映，並撰寫程式加以剖析、轉換，以便將整個資料庫連同資料一併從舊版的系統轉移至新版的系統中。

至於資料編碼部分，原有三個子系統是採ISO/IEC 8859-1(Latin-1)和Big-5方式。ISO/IEC 8859是國際標準組織ISO(International Standard Orgnization)和國際電工協會(International Electrotechnical Commission)聯合制定的一系列8位元字集標準(註20)。其中之一的ISO/IEC 8859-1是以ASCII為基礎，在空置的0xA0-0xFF的範圍內，加入192個字母及符號，可供大多數使用拉丁字母的西歐語系使用(註21)。而Big-5碼係由中華民國財團法人資訊工業策進會於1984年策劃制定，最初宗旨是為配合國人自製的五大套裝軟體使用，所以稱為Big-5(註22)。後來由於市面上絕大多數繁體中文套裝軟體都是在Big-5內碼系統上發展，促使Big-5逐漸成為業界標準，目前已普及於台灣、香港與澳門等使用繁體中文的地區。

然而，傳統編碼方式在支援多語言的處理方面有其侷限。為容納全世界各種語言的字元和符號，ISO一些會員國於1984年發起制定新的國際字元集編碼標準，稱為通用字元集(Universal Character Set，簡稱UCS)，編號為ISO/IEC 10646。但草案初稿公佈後，其編碼結構卻遭美國部分電腦業者反對。直到1988年初，美國Xerox公司倡議以新的編碼結構，另外編訂世界性字元編碼標準：Unicode，並由一群來自Xerox公司和Apple公司的工程師組成工作小組負責原始設計工作。1991年元月，十多家電腦硬軟體、網路和資訊服務業者，包括：IBM、DEC、Sun、Xerox、Apple、MicroSoft、Novell名公司，共同出資成立非營利組織Unicode協會(The Unicode Consortium)，專責設計與推動Unicode成為國際標準等工作。1991年10月，在歷經數月努力後，ISO和Unicode協會達成協議，將Unicode併入ISO/IEC 10646(註23)。基於效率與容量的考量，Unicode常見的三種編碼方式為：UTF-32/UCS-4、UTF-16/UCS-2和UTF-8(註24)。目前Unicode主要仍由Unicode協會推動，其宗旨在將世界上既有數百種字元編碼系統，以Unicode方案加以取代、整合，使軟體或網站能夠貫穿多個平臺、語言和國家，不須重建也能將資料傳輸到不同系統而無損壞(註25)。

為使Unicode與已存在和廣泛使用的舊有編碼互相兼容，尤其大多數電腦系統都支援的基本拉丁字母部分，所以Unicode的首256字元仍保留給ISO 8859-1所定義的字元，使既有的西歐語系文字不須特別轉換；另方面因相同的原因，Unicode把大量相同的字元重複編到不同的字元碼中去，使得舊有紛雜的編碼方式得以和Unicode編碼間互相直接轉換，而不會遺失任何資訊（註26）。

由此可知，目前國際趨勢是以Unicode作為標準的資料編碼格式，以便數位資訊能在不同的系統、平台、語系間互通。因此，決定將原有的三個子系統中的資料編碼格式由ISO/IEC 8859-1（Latin-1）和Big-5方式轉置成Unicode的UTF-8編碼。以下說明在資料格式轉置過程所遭遇之問題及解決之道。

⑵子系統一之資料轉置

首先，「台灣棒球運動珍貴新聞檔案數位資料館」最初的資料庫及網頁是使用Big-5編碼，為了因應未來仍能順利取用此網站的數位典藏資訊，將其轉換為Unicode是必要且合理的。所幸在Linux作業系統，只要使用iconv指令，即可順利將網頁部分從Big-5編碼轉成UTF-8編碼，但MySQL資料庫部分則稍為繁瑣一些，經過反覆幾次實驗，終於成功完成轉換。茲歸納步驟如下。

A. 資料匯出：先使用MySQL提供的指令mysqldump將資料庫的資料匯出成一個SQL檔。然而，因本系統舊有的MySQL版本並未提供UTF-8的編碼機制，所以先匯成Latin-1的編碼格式再進行後續處理。

B. 切割檔案：在上一步驟所匯出的SQL檔須要編輯過才能正確匯入新版的MySQL資料庫，但由於本系統之數位典藏資料檔共有5萬篇歷史新聞及1千2百幅老照片，檔案太大（約有394MB），以致匯入的過程出現多處錯誤訊息，因此必須將這個匯出的SQL檔切成若干個小檔再一一處理。但要切割的檔案必須以行為單位才行，若在一個單一的SQL語法切開會造成匯入的錯誤，因此利用Perl語言寫了一支切割檔案的程式，將原始匯出的SQL檔每一萬行即加以切割，另儲存成一個較小的SQL檔，如此就從一個大的SQL檔變成了55個SQL小檔。

C. 修改檔案：這步驟主要是將資料表的編碼從Latin-1改成UTF-8的編碼，切割成的小檔雖多，但用搜尋及取代的方式很快就可完成。

D. 匯入資料：將之前已修改過的55個SQL檔一一匯入新的MySQL資料庫，但發現有幾個檔案在匯入時產生錯誤，探究原因發現是有些特殊符號（如「\」及「'」等）的組合構成不合法的SQL語法，因此必須手動地加以修改產生錯誤的SQL檔。完成之後，舊有的Big-5的資料庫就成功轉換到UTF-8的新資料庫了。

⑶子系統二之資料轉置

完成上述子系統一的資料轉置後，接著進行子系統二「台灣棒球維基館」

的資料轉置。該系統之資料庫是使用Latin-1編碼，因為Latin-1的字集大於UTF-8，所以在轉換成UTF-8的過程略有不同，也經過多次實驗才成功地轉換。轉換的步驟歸納如下。

A. 資料匯出：用MySQL提供的指令mysqldump將資料庫的資料匯出成一個SQL檔。由於台灣棒球維基館擁有1萬多頁面，且因採用Wiki系統，每個頁面的編輯歷程皆儲存下來，造成這個SQL檔相當龐大，約有1GB。

B. 切割檔案：同樣地這個匯出的SQL檔必須修改之後才能正確地匯入新的MySQL資料庫，使用之前寫好的Perl程式以每10萬行作切割，如此這個SQL檔變成了10個較小的SQL檔可方便編輯修改。此處採用Perl的原因是該語言屬較高階之直譯式語法，用來切割檔案非常適合，程式碼約10行左右就可完成。

C. 過濾非UTF-8的字碼，原本的Latin-1字集大於UTF-8，因此有非UTF-8字碼的網頁上會造成整個網頁的不正確，如整個網頁會出現一大片問號，因此必須設法過濾非UTF-8的字碼，在找不到合用的工具後，用C語言根據UTF-8的編碼方式寫了一支過濾程式，將非UTF-8的字碼用問號替代，這些UTF-8的字碼約有數千個，都是一些符號。此處採用C語言的原因是判別字碼須要許多位元運算，採用較低階且具強大位元運算能力的C語言來處理相對容易。

D. 修改檔案：將表單的編碼指定成UTF-8的編碼，之後再匯到新的資料庫時還是有錯誤，原因是所採用的MySQL在舊版的Latin-1儲存方式與新版的UTF-8儲存方式，其表單的字碼定義及限制有所不同，因此在匯到新的資料庫之前必須修正兩個資料表。

E. 匯人資料：將之前已修改過的10個SQL檔一一匯入新的MySQL資料庫，即大功告成。雖然在之前過濾非UTF-8的字碼步驟中將非UTF-8的字碼替換成問號，這造成不少網頁的原來特殊符號變成了問號，但因「台灣棒球維基館」是共同協作的網站，在使用者的共同協助之下，有發現異樣的網頁也都能在很短的時間內就修正完畢。

F. 資料庫升級：這裡所謂的升級是將資料庫由Mediawiki 1.4.0版的資料庫結構提升至Mediawiki 1.11.1版的資料庫結構；兩者有非常大的差異，難度頗高。實現的方式是根據Mediawiki官方網站的建議分成兩個步驟，先將1.4.0版的資料庫結構升級至1.5.8版的資料庫結構，再由1.5.8版的資料庫結構升級至1.11.1版的資料庫結構。Mediawiki官網雖有提供升級工具，但升級之後的資料庫仍要修正，原因是先前曾因研究需要，在舊有的資料庫表格新增自定的欄位，如使用者的註冊時間、使用者的編輯次數等。這些在升級之後的資料庫表格中恰好都有相對應的欄位，只是欄位名稱不同，因此須將這些欄位逐一更

名；另外也發現升級之後的使用者預設語言都變成了英文，因此又寫了一支程式將所有使用者預設語言改成中文繁體。

⑷子系統三之資料轉置

在上述子系統一和子系統二的資料轉置完成後，接著進行子系統三「台灣棒球數位文物館」的資料轉置工作。由於該館的系統建置是較近期的事，資料庫系統原本即已採用MySQL，版本也較新（5.x版），且建置時就直接採用了UTF-8的編碼，所以問題較少。在資料庫轉置時僅須在原來的平台（Windows NT Sever）將資料匯出，再複製到新的平台（Fedora）後將資料匯入就可完成，難度並不高。

4. 應用程式轉置

⑴子系統一之程式轉置

原有「台灣棒球運動珍貴新聞檔案數位資料館」於四年前開始建置至今，因不同階段的研究、開發需求，而混用了包含JSP、ASP及PHP等程式語言，如今要從Windows平台轉換到Linux平台最大困難出現在ASP，因ASP程式只能在Windows之下運作，雖在Linux平台也可安裝虛擬機器程式來跑ASP，但這並非徹底解決問題的方式，所以為了一勞永逸，最後作法是將每個ASP程式用PHP語法一一改寫。

⑵子系統二之程式轉置

原有「台灣棒球維基館」應用程式是基於Mediawiki 1.4.0程式，再依據研究需求進行若干改寫。當升級到1.11.1版之後，這些曾修改過的程式則必須根據新版本的運作架構再重新改寫，這是在整個轉置過程中最困難，花費時間最長的，無法一一列舉，以下僅條列舊應用程式在轉換過程所碰到的若干問題及解決方式。

A. 保留已修改過的功能：在舊版軟體中有許多基於研究需求而修改之處，在新版軟體中是沒有的，因此必須將原始Mediawiki 1.4.0程式碼與「台灣棒球維基館」修改過的程式碼一一比對，找出哪些程式碼是修改過的，之後再到Mediawiki 1.11.1版的程式碼找到相對應的地方修改。

B. 修正網頁顯示不正常的地方：因為Mediawiki新舊版的程式差異頗大，許多網頁在程式轉置後，反而無法正常顯示，例如：出現簡體字，出現$1、$2等符號（這些是沒有被正確的字串替換）……等，因此，必須一一從原始程式碼中追蹤這些錯誤的起因加以修正。

C. 新舊版的外掛程式不相容問題：例如在舊版使用的圖片認證程式是使用「CAPTCHA」程式，但這個程式已無法在新版的程式運作，因此決定改換成較新的圖片認證程式「ConfirmEdit」。這造成使用者介面與舊版不太一樣，使用者需要一些時間適應。

D. 修改功能上的錯誤：例如中文的搜索在新版並不正確，追蹤程式碼之後發現是斷詞程式出問題。例如：在搜索「中華職棒」的字串時，新版的Mediawiki會將中文字斷開如「中 華 職 棒」來搜索，這造成了搜索中文的不正確，找到斷開中文程式碼的地方後加以修改即可解決；另外在圖片上傳的地方也出現了問題，還好這些程式都是明碼，最終都可找到原因加以解決。

E. 增加新的外掛程式：Mediawiki通常是較新版本有較多外掛程式可用，未升級前，許多又新又好用的外掛程式無法使用；升級後，許多新的外掛程式就都能使用了，目前系統新增了「FeedImport」及「EmbedVideo」兩支外掛程式。前者是RSS剖析程式，目前在本系統中用來整合外界部落格文章，而開闢出一個「棒球部落」專區；後者是可內嵌YouTube或Google Video等影片，目前在本系統中用來整合珍貴的歷史球賽鏡頭。

⑶子系統三之程式轉置

「台灣棒球數位文物館」由於最晚建置，網頁程式也是採用PHP，因此轉置到新平台時僅須複製程式到新的平台，再設定正確的環境變數即可。

㈢輔助層：重複一套系統建置

目前的三個館都重複作了一套系統建置，這是當作備援系統。現行的系統每日自動將資料作備份並轉送到備援系統，這除了可取代網路硬碟的轉存外，好處是在原系統萬一有意外發生時，備援系統可立即取代現有系統繼續運作。

五、成果分析

本研究小組於2007年下半年開始查覺原有的舊系統陸續出現系統不穩、頻頻當機、頻寬不足、負荷過重等現象，遂於2007年8月開始實施系統轉置計畫，2008年3月上線測試，2008年5月新系統逐漸趨於穩定運作，順利完成系統轉置。本節將透過網站記錄檔分析（Transaction Log File Analysis）來檢視系統轉置之成效。

網站記錄檔是當使用者利用瀏覽器輸入網址或對網頁中的連結進行點選時，存取網頁資源的行為紀錄，通常包含了對網站伺服器所提出的請求（Request）（註27）。它會逐筆記錄在伺服器上，因而網站記錄檔可提供許多珍貴的資訊；藉由紀錄檔分析可在不干擾使用者的情形下進行測量，所測得的資料也具客觀性。

本文用以分析的數據取自台灣棒球數位典藏計畫網站之記錄檔，自2007年1月至2008年11月止，計有瀏覽量和檢索量。由圖5及圖6可明顯看出2007年3月至8月期間，不論瀏覽量或檢索量縱有高低起伏，但整體趨勢線是往上升的。使用量的增加同時代表系統負荷的增加，2007年8月時上升趨勢減緩，表

示系統已出現過載警訊，於是在8月起展開系統轉置規劃。

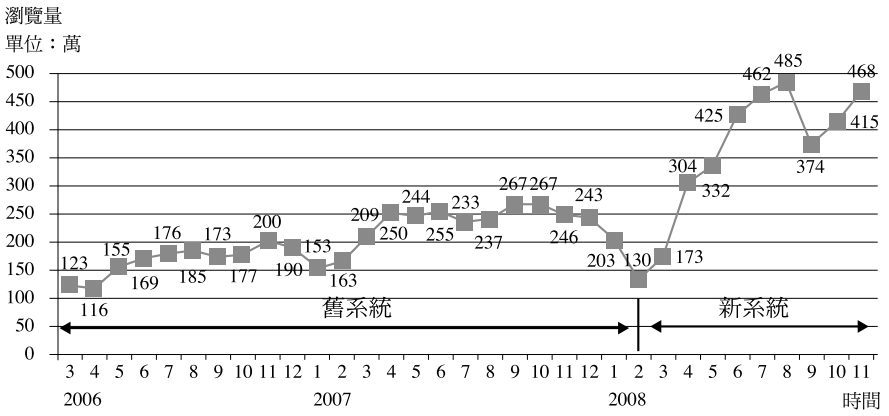

圖5 網站瀏覽量

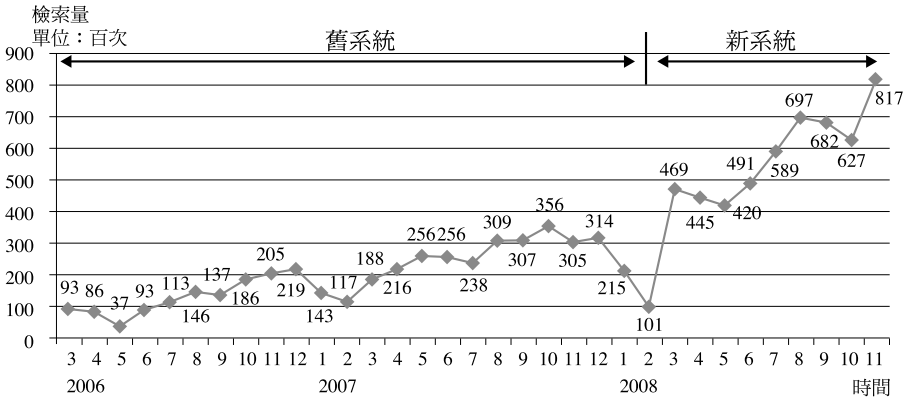

圖6 網站檢索量

2007年9月至12月系統瀏覽量趨勢線呈現持平現象，代表系統已達飽和極限，無法處理更多的使用需求，除非新系統儘速上線運作，否則系統的使用率已無法有效提升。此期間研究小組雖已向學校提出購置新伺服器和擴展頻寬等申請，但由於行政作業需要時間，新伺服器於2008年1月底購進，2月開始進行軟硬體安裝、測試，但上線日期仍有待學校資訊中心支援佈設寬頻專線。從圖中可明顯觀察到此期間（2008年1月～2月）不論是瀏覽量或檢索量都出現了嚴重衰退現象。此乃由於系統負荷過重、頻頻當機，加上頻寬不足、網路塞車使然。幸好2月底、3月初，終於陸續完成Gigabit專線佈設、主機測試和機櫃安裝，新伺服器終於可以正式上線運作了。因此，從圖中可看到2008年3月份起，系統效能明顯改善，使用狀況順暢，瀏覽量和檢索量不但大幅回升，還遠遠超越前波高點，代表寬頻專線和新伺服器發揮了效用，可處理更大量的使用

者需求，穩定的新系統也能有效保存珍貴的數位資訊並順利提供使用者存取利用。

六、結　論

從數位資訊長期保存觀點來看，轉置是最核心的工作，歐陽崇榮教授曾依據國外相關機構發表之研究報告（註28）以及國內專家學者的看法，也同樣歸納出轉置策略在現階段是數位資訊保存較為恰當的方法（註29）。本研究雖然順利完成台灣棒球數位典藏系統之轉置工作，使系統中日益累積的珍貴資訊能長期為使用者檢索與利用。然而，從本研究小組的經驗來看，學術研究單位除非經費允許，可將轉置工程外包予資訊廠商進行，否則若由研究人員負責執行轉置工作，勢必耗費大量時間與人力，若研究團隊本身又缺乏技術人力，則可能無法順利完成。因此，本文提出幾點建議，以供國科會持續推動數位典藏計畫之參考：

㈠成立數位典藏資訊長期保存計畫辦公室：國科會數位典藏計畫雖為我國唯一兼顧人文與科技的國家型計畫，但許多研究團隊都是由人文社會學背景的師生所組成，缺乏資訊技術方面的人力，在技術方面亟需計畫辦公室之指導與協助，因此儘速成立數位典藏資訊長期保存計畫辦公室，可提供各計畫在進行數位資訊長期保存時的協助。

㈡儘速培訓相關技術人員：如上所述，許多研究團隊都未有專屬的技術人力，因此「數位典藏資訊長期保存計畫辦公室」成立之後的首要任務，便是儘速了解各項長期保存的方法與技術，並培訓相關技術人員，以因應各計畫案之所需。

㈢研擬系統保存標準作業程序：若能儘速歸納出一套有關系統軟體、硬體及資料之轉置與保存的標準作業程序，提供各研究計畫參考，則可收事半功倍之效。以本研究之經驗而言，基礎層之轉存是必要之例行工作，而軟硬體、資料格式與Metadata則在系統建置初期盡量採用符合業界標準之規格。至於核心層的轉置工作通常是系統運作若干年限後再著手進行，屬於較大的工程，需要較多的資源投入。此外，進行轉置的過程可一併將舊系統保存下來，以備萬一。

㈣加強產學合作：學界一向擁有堅實的學理基礎與研發能量，而業界則有強大的技術資源與創製能力，若能有效整合學術界與產業界力量，必能促進知識之累積與擴散，發揮教育、訓練、研究、服務之功能，有助於國家教育與經濟發展。對於數位典藏資訊長期保存之推動亦能有所助益。

註　釋

註1　The Library of Congress, “American Memory,” http://memory.loc.gov/ammem/index.html (accessed May 20, 2008).

註2　UKOLN, “About UKOLN,” http://www.ukoln.ac.uk/about/ (accessed May 20, 2008).

註3　Information Society Technology, “About IST,” http://cordis.europa.eu/ist/about/about.htm (accessed June 2, 2008).

註4　CHIN, “Canadian Heritage Information Network: Digital Content Development and Heritage Resources,” http://www.chin.gc.ca/English/index.html (accessed May 20, 2008).

註5　Renato Iannella, “Australian Digital Library Initiative,” *D-Lib Magazine* 2, no. 12 (December 1996), http://www.dlib.org/dlib/december96/12iannella.html (accessed May 20, 2008).

註6　国立国会図書館総務部企画課，「電子図書館サービス実施基本計画」，国立国会図書館，http://www.ndl.go.jp/jp/aboutus/elib_standardproject.html（檢索於2008年5月20日）。

註7　行政院國家科學委員會，「經典意像 珍藏台灣—數位典藏國家計畫」，http://www.ndap.org.tw/（檢索於2008年5月20日）。

註8　行政院文化建設委員會，「關於國家文化資料庫」，http://nrch.cca.gov.tw/cca-home/index.jsp（檢索於2008年5月20日）。

註9　Gail M. Hodge, “Best Practices for Digital Archiving: An Information Life Cycle Approach,” *D-Lib Magazine* 6, no.1 (January 2000), http://dlib.ejournal.ascc.net/dlib/january00/01hodge.html (accessed May 20, 2008).

註10　陳昭珍，「國家檔案數位典藏面臨的挑戰與發展方向」，檔案季刊1卷，1期（2002）：61-68。

註11　Terry Kuny, “The Digital Dark Ages? Challenges in the Preservation of Electronic Information,” *International Preservation News* 17 (May 1998), http://www.ifla.org/VI/4/news/17-98.htm#2 (accessed May 20, 2008).

註12　歐陽崇榮，數位資訊保存策略（台北市：文華，2006），8-9。

註13　淡江大學資訊與圖書館學系，「台灣棒球運動珍貴新聞檔案數位資料館之建置」，http://ndap.dils.tku.edu.tw/（檢索於2007年1月31日）。

註14　淡江大學資訊與圖書館學系，「台灣棒球維基館」，http://twbsball.dils.tku.edu.tw（檢索於2007年1月31日）。

註15　淡江大學資訊與圖書館學系，「台灣棒球數位文物館」，http://museum.dils.tku.edu.tw（檢索於2007年1月31日）。

註16　Laura DiDio, “Yankee Group 2007-2008 Server OS Reliability Survey,” Institute for Advanced Professional Studies, http://www.iaps.com/exc/yankee-group-2007-2008-server-reliability.pdf (accessed August 16, 2008).

註17　DistroWatch.com, “Put the Fun Back into Computing. Use Linux, BSD.,” http://distrowatch.com (accessed August 10, 2008).

註18　The Linux Foundation, “Linux Standard Base,” http://www.linuxfoundation.org/en/LSB (accessed August 10, 2008).

註19 Daniel Quinlan, “Pathname: Filesystem Hierarchy Standard,” http://www.pathname.com/fhs (accessed August 10, 2008).

註20 Wikipedia Contributors, “ISO/IEC 8859,” Wikipedia, http://en.wikipedia.org/w/index.php?title=ISO/IEC_8859&oldid=222744795 (accessed July 27, 2008).

註21 ISO/IEC 8859-1, “8-bit Single-byte Coded Graphic Character Sets, Part 1: Latin alphabet No.1,” (12 February, 1998), http://anubis.dkuug.dk/JTC1/SC2/WG3/docs/n411.pdf (accessed August 10, 2008).

註22 行政院主計處電子處理資料中心，「BIG-5碼介紹」，http://www.cns11643.gov.tw/web/word/big5/index.html（檢索於2008年7月27日）。

註23 曾士熊，「Unicode與ISO10646（上）」，http://www.ascc.sinica.edu.tw/nl/89/1610/02.txt（檢索於2008年5月20日）。

註24 曾士熊，「Unicode與ISO10646（下）」，http://www.ascc.sinica.edu.tw/nl/89/1611/02.txt（檢索於2008年5月20日）。

註25 The Unicode Consortium, “What is Unicode?” http://www.unicode.org/standard/WhatIsUnicode.html (accessed August 10, 2008).

註26 Wikipedia contributors, “Unicode,” Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Unicode (accessed July 27, 2008).

註27 Tomas C.Almind, Peter Ingwersen, “Information Analysis on the World Wide Web: A Methodological Approach to internetometrics,” Centre for informetric Studies, Royal School of Library and information Science, Copenhagen, Demark. (CIS Report 2), 2006.

註28 National Archives of Australia, “Managing Electronic Records,” http://www.naa.gov.au/recordkeeping/er/manage_er/append_3.html (accessed July 27, 2008).

註29 歐陽崇榮，電子媒體類檔案管理制度及保存技術之研究（台北市：檔案管理局，2002）。

Journal of Educational Media & Library Sciences 46 : 3 (Spring 2009): 323-349

Planning and Implementing Long-term Preservation for Digital Archive System

Sinn-Cheng Lin*

Associate Professor
Department of Information & Library Science, Tamkang University
Taipei, Taiwan
E-mail: sclin@mail.tku.edu.tw

Kuo-Hsiang Cheng

Ph. D. Student
Department of Computer Science and Information Engineering, Tamkang University
Adjunct Lecturer
Department of Information Management
Technology and Science Institute of Northern Taiwan
E-mail: khcheng@mis.tsint.edu.tw

Abstract

This paper focuses on the long-term preservation issues for the Digital Archive System of Taiwan Baseball Culture Assets. Funded by the Digital Archive Project of National Science Council in Taiwan, the system was created five years ago and became outdated. It suffered problems such as overloading and unstableness. In order to preserve the valuable information that accumulated in the system so it can be accessed by users for a long period of time, the researchers adopted the technologies of refreshing, standardization, metadata, system migration, and system redundancy to upgrade the system. Then, we analyzed the website's log files to examine the effectiveness of the new system. Finally, the article suggests the National Science Council to establish a long-term digital information archive office as soon as possible, to plan the long-term preservation strategy, to train the technical staff, and to encourage team work between industry and academic units in order to provide the needed assistance for digital archive projects.

Keywords: *Digital archive; Long-term preservation; Migration; Metadata*

SUMMARY

Introduction

Unlike the conventional real materials and objects, digital information is not easy to preserve for a long period of time simply because that it is easy to be modified; the digital mediums are fragile; data are easy to lose and disappear; it is highly machine-dependent, and its technological life span is short. Dr. Kuny has warned that if we cannot preserve digital information properly, it would be impos-

* Principal author for all correspondence.

sible to restore it ever again. The digital civilization could then turn into a dark age. Therefore, all over the world the issue of preserving the digital information is a big concern among scholars, government officials, and businessmen. To study the long term preservation challenges of digital systems and make the valuable information available for future users, the researchers examined the *Digital Archive System of Taiwan Baseball Culture Assets,* one of the digital preservation systems funded by the National Science Council, with a number of preservation methods as a result of this research.

Empirical Results on the System

Three subsystems were built for the *Digital Archive System of Taiwan Baseball Culture Assets,* which are *Digital Archive of Taiwan Baseball Historical Newspaper Records, Taiwan Baseball Wiki,* and *Digital Museum of Taiwan Baseball Culture Assets*. In order to preserve the valuable information in these systems and make them available for users, the researchers adopted different preservation methods, i.e., refreshing, standardization, metadata, system migration, and system copying. The results are described in the following sections.

1.Refreshing

The two basic methods to preserve data are: (1) transferring all data to a Network-Attached Storage (NAS); (2) using XML to export the digital archive in the database to the Union Catalog of Digital Archives.

2.Standardization and metadata

The researchers replaced the previous character code formats, Big-5 and Latin-1, to UTF-8. The file format is adopted the standardized XHTML, XML, JPEG, GIF, etc. The metadata of the system was also established in accordance with the metadata specifications. In addition to the DAC format setup by the Union Catalog of Digital Archives, international standards such as DC, NITF, RSS were adopted so that information is easy to share, exchange, and preserve.

3.System migration

The system migration included four projects: (1) the migration of computer hardware; (2) the migration of operating system; (3) the migration of data format; and (4) the migration of system applications.

(1) The migration of computer hardware

All the hardware platforms set up by tower servers in the early stage were replaced by rack mount servers. On the other hand, in order to upgrade the Internet connection to solve the problem of slow transmission, the researchers obtained a Giga-bit leased line from the Information Center of Tamkang University. We also upgraded the low-speed switch and router; as a result, the Internet transmission is speeded up from 100 Mbps up to 1 Gbps.

(2) The migration of operating system

Considering the factors of stability, security, and economy, we decided that Windows operating systems should be replaced by Linux. After comparing a number of Linux systems, we migrated current operating system to Fedora.

In the process of the migration of operating system, the purchase of server and staff training are two major issues. The researchers borrowed a server from the vendor for the purpose of testing so that we could be sure that the new hardware can be compatible with Fedora. After making sure the testing went well we then purchased the server. To train and prepare the staff for system management, we organized a number of Linux workshops during the process of system migration.

(3) The migration of data format

The migration of data format included two parts, mainly database structure and data coding. Although all the aforementioned three subsystems are MySQL databases, the database structures were different in various ways, such as the time they were created and the applications they possessed; therefore, we had to examine and compare the systems carefully. We also had to write program to conduct data analysis and migration, so that everything in the database can be migrated from the old version to the new one.

As for the data coding, we migrated the format from ISO/IEC 8859-1(Latin-1) and Big-5 to UTF-8 coding of Unicode.

(4) The migration of system applications

First, we rewrote every ASP programs of the first subsystem, *Digital Archive of Taiwan Baseball Historical Newspaper Records,* in PHP language, so that it can operate smoothly on the Linux system. The original applications of the second subsystem, *Taiwan Baseball Wiki,* are created with Mediawiki 1.4.0, therefore, to migrate to version 1.11.1 we needed to revise the portions that were rewritten before based on the new structure. The overall tasks included modifying the database structure, revising the malfunctioning websites, fixing the incompatibleness between different versions, fixing functional errors, adding new plug-in applications, etc. The third subsystem, *Digital Museum of Taiwan Baseball Culture Assets,* was created most recently and the websites were designed with PHP programs as well; therefore, we only needed to copy the programs to the new platform and reconfigure the environment variables.

4.System redundancy

All of the three digital archive subsystems have another set of redundant system as backups. Data can be automatically copied and transferred to the back-up system daily, which would work as the active system in case of an emergency.

Performance Analysis

The implementation of the system started in August, 2007; went life in March, 2008, and began to work fairly stable in May, 2008; thus, the entire system migration was completed. The researchers examined the transaction log files to see the effectiveness of the system migration. The data analysis included the number of website page views and searches from January, 2007 to November, 2008. The results are as follows:

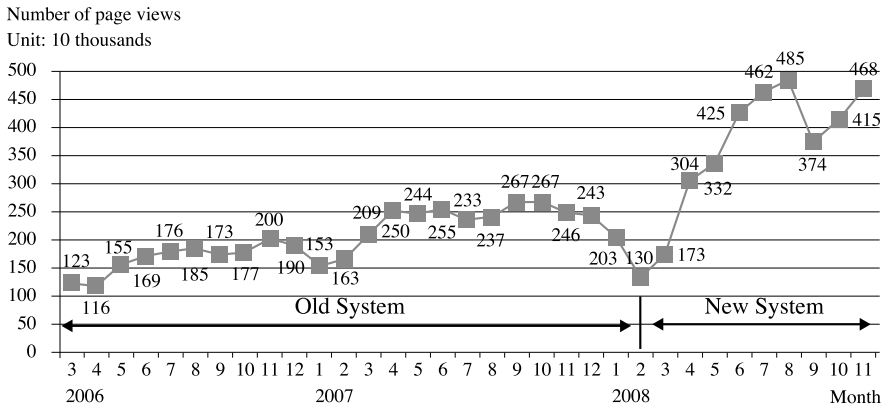

Figure 1 Number of Website Page Views

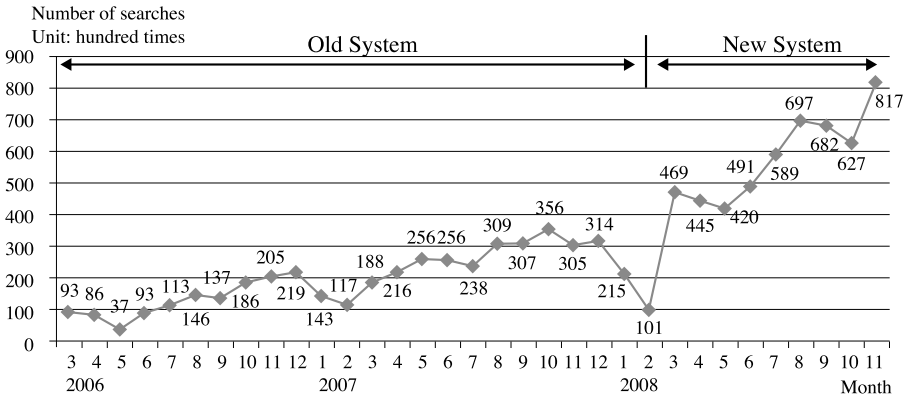

Figure 2 Number of Website Searches

As the figure shows, the lines of both numbers of website page views and searches went up and down from March to August, 2007, but in general they showed an upward orientation. In August, 2007, the usages slowed down, which was a warning sign for the system loading. From September to December, 2007, the usages kept almost flat, and could be interpreted that the system loading had reached its limit and the system could not process further requests from users. The migration task started in the end of January, 2008 by purchasing new server.

In February, 2008, we began a series of installation and testing for hardware and software programs. During this time the system lost its efficiency dramatically. The new migrated system was up and running in the beginning of March. The figure shows that the new system greatly improved the system efficiency; the numbers of website page views and searches not only went up and were a lot more than when the old system was in use. This means the new system has successfully proved its value in processing a large number of user requests. Meanwhile, better system efficiency also means that the valuable digital assets stored in the database can be better saved and retrieved by users.

Conclusion

The researchers successfully accomplished the migration task for the *Digital Archive System of Taiwan Baseball*. As a result, the invaluable digital assets can be made available to the general public. The research team was responsible for the entire system migration project; therefore, spent a fairly large amount of time and manpower. For those research teams who are lack of the needed technology background and skills, system migration might be a difficult effort to pursue. Thus, the researchers conclude this paper with the following suggestions for the *National Science Council in Taiwan* as a referring resource for future digital archive preservation projects:

1. Establish a digital information archive office to provide the needed assistance for related long-term digital archive preservation projects.
2. Provide appropriate training sessions for the staff of the digital information archive office so that they are well prepared and understand various preservation methods and techniques for different types of preservation projects.
3. Develop a set of standard operating procedure with regards to the system software, hardware, data migration, and data preservation, so that the overall migration task can be accomplished effectively.
4. Team up with experts from the industry and academia in order to share and distribute the invaluable knowledge and to provide insights into education, training, research, and services.

ROMANIZED & TRANSLATED NOTES FOR ORIGINAL TEXT

註1 The Library of Congress, “American Memory,” http://memory.loc.gov/ammem/index.html (accessed May 20, 2008).

註2 UKOLN, “About UKOLN,” http://www.ukoln.ac.uk/about/ (accessed May 20, 2008).

註3 Information Society Technology, “About IST,” http://cordis.europa.eu/ist/about/

about.htm (accessed June 2, 2008).

註4 CHIN, “Canadian Heritage Information Network: Digital Content Development and Heritage Resources,” http://www.chin.gc.ca/English/index.html (accessed May 20, 2008).

註5 Renato Iannella, “Australian Digital Library Initiative,” *D-Lib Magazine* 2, no. 12 (December 1996), http://www.dlib.org/dlib/december96/12iannella.html (accessed May 20, 2008).

註6 国立国会図書館総務部企画課[National Diet Library]，「電子図書館サービス実施基本計画」[“The Initiative in Implementation of Digital Library”]，国立国会図書館[National Diet Library]，http://www.ndl.go.jp/jp/aboutus/elib_standardproject.html（檢索於2008年5月20日）[(accessed May 20, 2008)]。

註7 行政院國家科學委員會[National Science Council]，「經典意像 珍藏台灣—數位典藏國家計畫」[“Classical Images: Taiwan-National Digital Archives Program”]，http://www.ndap.org.tw/（檢索於2008年5月20日）[(accessed May 20, 2008)]。

註8 行政院文化建設委員會[The Council for Cultural Affairs]，「關於國家文化資料庫」[“About National Repository of Cultural Heritage”]，http://nrch.cca.gov.tw/ccahome/index.jsp（檢索於2008年5月20日）[(accessed May 20, 2008)]。

註9 Gail M. Hodge, “Best Practices for Digital Archiving: An Information Life Cycle Approach,” *D-Lib Magazine* 6, no. 1 (January 2000), http://dlib.ejournal.ascc.net/dlib/january00/01hodge.html (accessed May 20, 2008).

註10 陳昭珍[Chao-Chen Chen]，「國家檔案數位典藏面臨的挑戰與發展方向」[“The Development Trend of National Electronic Archives and Its Challenges”]，檔案季刊[*Archives Quarterly*]1卷，1期[1, no.1] (2002)：61-68。

註11 Terry Kuny, “The Digital Dark Ages? Challenges in the Preservation of Electronic Information,” *International Preservation News* 17 (May 1998), http://www.ifla.org/VI/4/news/17-98.htm#2 (accessed May 20, 2008).

註12 歐陽崇榮[James C. Ouyang]，數位資訊保存策略[*Digital Information Preservation Strategy*]（台北市：文華，2006）[(Taipei: Mandarin, 2006)]，8-9。

註13 淡江大學資訊與圖書館學系[Department of Information and Library Science, Tamkang University]，「台灣棒球運動珍貴新聞檔案數位資料館之建置」[“The Installation for Digital News Archives of Taiwan Baseball”]，http://ndap.dils.tku.edu.tw/（檢索於2007年1月31日）[(accessed January 31, 2007)]。

註14 淡江大學資訊與圖書館學系[Department of Information and Library Science, Tamkang University]，「台灣棒球維基館」[“WikiBaseball”]，http://twbsball.dils.tku.edu.tw（檢索於2007年1月31日）[(accessed January 31, 2007)]。

註15 淡江大學資訊與圖書館學系[Department of Information and Library Science, Tamkang University]，「台灣棒球數位文物館」[“Taiwan Baseball Digital Museum”]，http://museum.dils.tku.edu.tw（檢索於2007年1月31日）[(accessed January 31, 2007)]。

註16 Laura DiDio, “Yankee Group 2007-2008 Server OS Reliability Survey,” Institute for Advanced Professional Studies, http://www.iaps.com/exc/yankee-group-2007-2008-server-reliability.pdf (accessed August 16, 2008).

註17 DistroWatch.com, “Put the Fun Back into Computing. Use Linux, BSD.,” http://distrowatch.com (accessed August 10, 2008).

註18 The Linux Foundation, “Linux Standard Base,” http://www.linuxfoundation.org/en/LSB (accessed August 10, 2008).

註19 Daniel Quinlan, “Pathname: Filesystem Hierarchy Standard,” http://www.pathname.com/fhs (accessed August 10, 2008).

註20 Wikipedia Contributors, “ISO/IEC 8859,” Wikipedia, http://en.wikipedia.org/w/index.php?title=ISO/IEC_8859&oldid=222744795 (accessed July 27, 2008).

註21 ISO/IEC 8859-1, “8-bit Single-byte Coded Graphic Character Sets, Part 1: Latin alphabet No. 1,” (12 February, 1998), http://anubis.dkuug.dk/JTC1/SC2/WG3/docs/n411.pdf (accessed August 10, 2008).

註22 行政院主計處電子處理資料中心[Computing Center, Directorate General of Budget, Accounting and Statistics, Executive Yuan, R.O.C.]，「BIG-5碼介紹」[“Instroduction to BIG-5 code”]，http://www.cns11643.gov.tw/web/word/big5/index.html（檢索於2008年7月27日）[(accessed July 27, 2008)]。

註23 曾士熊[Shi-Xiong Zeng]，「Unicode與ISO10646（上）」[“Unicodeand ISO10646 (I)”]，http://www.ascc.sinica.edu.tw/nl/89/1610/02.txt（檢索於2008年5月20日）[(accessed May 20, 2008)]。

註24 曾士熊[Shi-Xiong Zeng]，「Unicode與ISO10646（下）」[“Unicodeand ISO10646 (II)”]，http://www.ascc.sinica.edu.tw/nl/89/1611/02.txt（檢索於2008年5月20日）[(accessed May 20, 2008)]。

註25 The Unicode Consortium, “What is Unicode?” http://www.unicode.org/standard/WhatIsUnicode.html (accessed August 10, 2008).

註26 Wikipedia contributors, “Unicode,” Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Unicode (accessed July 27, 2008).

註27 Tomas C.Almind, Peter Ingwersen, “Information Analysis on the World Wide Web: A Methodological Approach to internetometrics,” Centre for informetric Studies, Royal School of Library and information Science, Copenhagen, Demark. (CIS Report 2), 2006.

註28 National Archives of Australia, “Managing Electronic Records,” http://www.naa.gov.au/recordkeeping/er/manage_er/append_3.html (accessed July 27, 2008).

註29 歐陽崇榮[James C. Ouyang]，電子媒體類檔案管理制度及保存技術之研究[*Dianzi Meiti Lei Dangan Guanli Zhidu ji Baocun Jishu zhi Yanjiu*]（台北市：檔案管理局，2002）[(Taipei: National Archives Administration, 2002)]。