政府機關網站典藏策略與方法

許芳銘
國立東華大學資訊管理學系教授
fmhsu@mail.ndhu.edu.tw

何祖鳳
銘傳大學資訊工程學系副教授兼圖書館館長
tfho@mail.mcu.edu.tw

范秋足
東南科技大學企業管理學系助理教授
jofan@mail.tnu.edu.tw

摘要

網站已成爲政府機關提供民眾資訊與服務的重要管道，乃屬機關之重要記憶。然而此珍貴之機關記憶卻因不定期的網頁改版而逐漸演化消失，因此已成爲各國積極推動之檔案管理重點工作。網站典藏牽涉法令、發展策略、管理、技術等問題，然而機關可在考量成本、風險、優先順序、可得資源及資訊科技設備等因素之後，將機關網站內容納入檔案管理範疇。機關可以在不花費額外軟硬體建置費用之下，進行網站典藏，擴大保存機關之珍貴記憶。

關鍵詞：政府機關、網站、策略、數位典藏

一、 前言

國家記憶分散於社會、人民及各級政府機關記憶之中，因此電子型式國家寶藏來自於各政府機關之數位典藏（digital archives）。網站已成爲政府機關提供民眾資訊或服務的重要管道，極具有歷史價值。網站乃是機關活動的歷史記錄或文化軌跡，若能妥善保存網站資料，則可瞭解機關的活動事蹟或某事件的發展歷程，供日後見證或研究參考。另外，歷史網頁資訊記載著某個事件的處理程序及成果，機關人員在接觸類似公務時，可參考前人作法，有效提升決策速度，因此典藏網站內容除了可以保存更多具證據、歷史、文化等資料之外，亦可提升機關處理公務效率。再者，網站中的圖表與資料都是精心收集資料與設計的成果。若能善加運用現有資源，則可達節省時間及人力成本之效。

網站典藏（web archive）乃是一個組織在接受長期保存與應用的責任之下，所保有在全球資訊網發佈的網頁材料（web-published materials）（Murray & Hsieh, 2008）。在保存上，網站典藏面臨法令、內容變化、技術障礙等問題。在法令上，著作權爭議是網站典藏所需克服的障礙。在內容變化上，網站的涵蓋範圍日趨廣泛，且因著網站的動態特性，數量與內容急遽成長、更動或消失，許多內容由動態資料庫產生，使網站的長期保存困難。在技術障礙上，因應設計網站的資訊技術快速更新，使得目前仍無法保存以 Flash 等工具所撰寫的動態網頁（Day, 2003）。

二、 政府機關網站典藏策略

政府機關若欲進行網站典藏，則其所保存的內容應以與公務有關且具保存價值之網頁爲範圍，因此與機關業務範疇相關、機關用以處理公務以及提供民眾相關資訊之有價值網頁均應加以保存。在考量成本、風險、優先權、可得資源及基礎設備等因素之下，機關可採用以下四種方式進行網站典藏：1、自製（in-house）：由機構負責資源規劃、管理與實行，使得機關具有較多的掌控彈性。2、外包（contracted-out）：機構負責規劃，由廠商協助完成。機關應挑選及管理適當之網頁，並考量網站主機及存放空間。3、協同合作（collaboration）：多個機關之間共同合作，共享技術與資源以達到相同目標。4、聯盟（consortium）：由機關之聯合單位負責，營造一個互助合作的環境，提供機關建立持續網站典藏。

協同合作是蒐集與典藏網站的關鍵成功因素（Day, 2003）。單一機構很難能蒐集全域的網站資料，僅能針對特定主題及領域進行網站典藏。因此，不同機關推行網站典藏時，彼此合作與溝通是很重要的。進行網站典藏時，應規劃典藏策略，並建立永續發展的合作基礎。典藏策略與現存之資訊技術息息相關，需參照資訊技術才能制定合適的網站典藏策略、範圍及方式。在協同合作努力下，才能使網路資源備份保存計畫順利執行。藉由協同合作，網站典藏可從檔案主管機關與各級機關兩個方向進行。檔案管理局藉由修改法規將網站納入保存規範範圍，建立網站典藏相關作業指引與配套措施，推動各機關之網站典藏；各級機關則依據個別業務與資源特性，考量成本以及對機關執行公務的長遠效益，實際落實網站之保存。

典藏網站內容時，機關必須決定網頁典藏之優先順序及其執行方式。因爲保存「每個」網頁幾乎是不可能，因此建立優先順序是典藏網站資料的重要基礎。在管理上，機關應制定明確之網站典藏政策與保存目標。因爲任由檔案管理人員決定，可能面臨檔案管理人員無法精確挑

選具保存價值之網頁。若篩選機制寬鬆，保存目標可能過於繁雜；若篩選機制過於嚴謹，則會有遺珠之憾。在決定某網頁應否保存時，可利用 MoSCoW 分類方式將網頁分爲以下四類，以確認其保存優先順序：1、Must（M）：必須保存之網頁；2、Should（Should）：應該盡可能保存之網頁；3、C（Could）：可考慮保存之網頁；4、W（Won't）：不需保存之網頁。

另外，機關應考量財務、人員、電腦軟體工具、電腦硬體儲存空間等作業所需的資源，亦應針對網站典藏建立明確之資訊技術政策，以規範確立現階段採取網站典藏與未來存取之資訊技術工具，以確保所典藏之網站內容在未來之可及性。再者，機關進行網站典藏之前，須挑選網站的保存途徑與範圍，並達成同意保存網站的協議，再進行抓耙（crawl）與保存的工作。網站典藏所需之軟硬體設備主要是儲存空間與系統維護之成本，各機關可依照本身的需求自行調整。網站典藏所涉及的技術範圍主要在於軟體系統部分，檔案管理人員須有挑選具保存價值網站之能力，並根據價值差異定義合宜的保存年限、抓耙頻率、抓耙深度等。就資訊人員而言，最大的挑戰在於因應軟體平台與語系等特性需求，架設網站典藏管理系統環境。

爲達到長久典藏網站資源，應將網站的相關描述如詮釋資料等，一併納入網站典藏範圍。網站典藏管理系統必須有一套完整規劃的詮釋資料格式用以描述與管理網站資訊，並詳細記錄網站的內涵與產生方式。產出數位典藏資源的機關，應仔細審視並修正其網站之圖形利用、有效連結、詮釋資料等內容之完整，使得機關網站典藏符合 ISO 15489 所提及之電子檔案完整性（integrity）與可及性（accessibility）。

三、 網站典藏方法

網站典藏可分爲二種方式：集中式及分散式。集中式典藏是由單一機關徵集特定範圍內所有機關中與保存主題相關之網站內容；分散式典藏則由一主要機關收集所有所屬機關內與保存主題相關之網站內容。因此，上級機關需先決定網站典藏的方式，進而規範所屬相關機關的作業方式。在作業流程上，集中式或分散式網站典藏僅在網站授權上有差異。

以下針對網站典藏階段與作業流程、網站典藏工具、網站抓耙方式、網頁典藏格式等四方面加以說明。

(一) 網站典藏階段與作業流程

機關進行網站典藏時，可分爲網站挑選、網站授權取得、網站典藏三個階段。其作業流程，如圖 1。

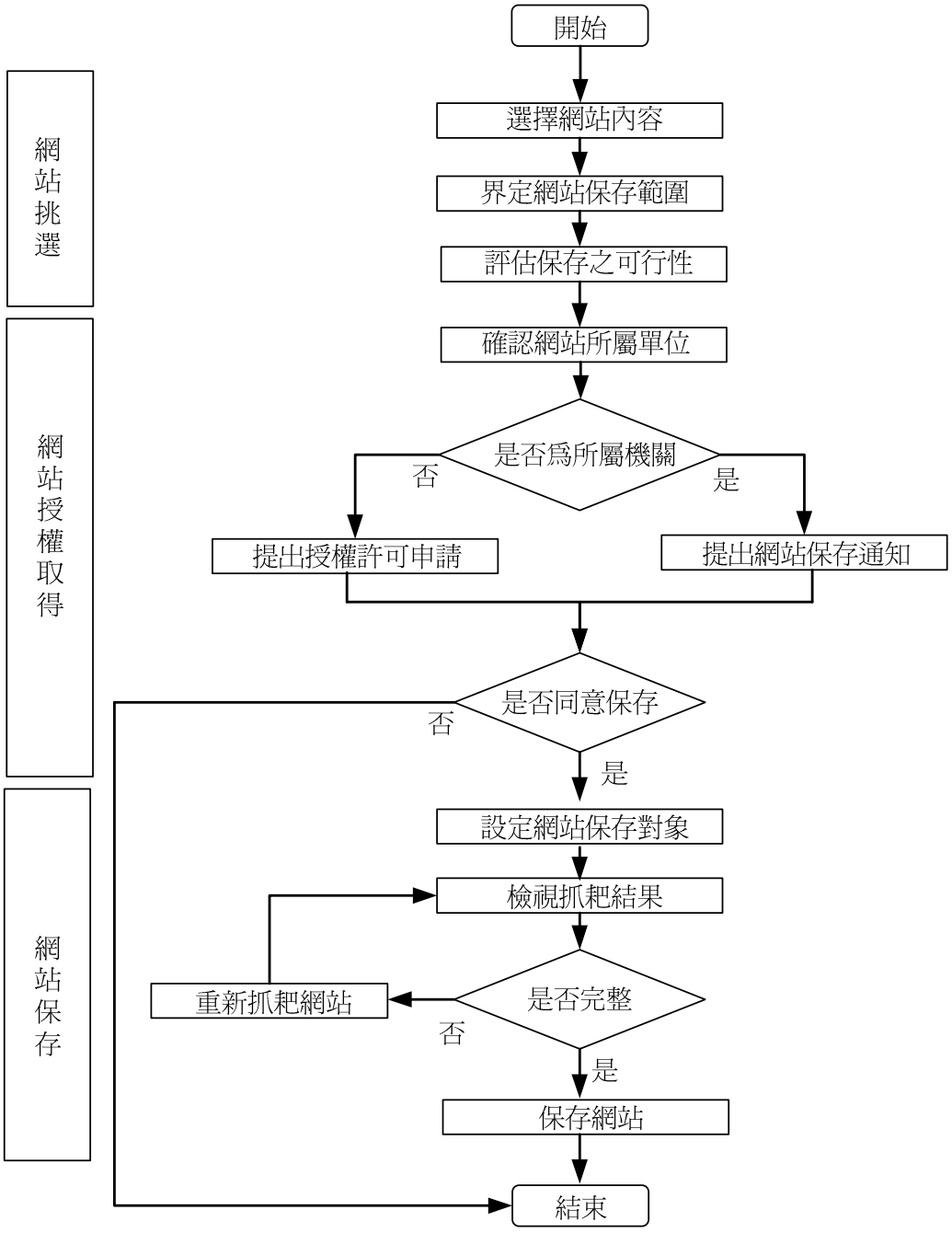

圖 1　網站典藏作業流程

各階段之作業內容，說明如下：

1、 網站挑選階段

(1) 檔案管理人員選擇網站內容時，可依照下列原則進行挑選：具有證據性、具有歷史價值、與機關業務範圍相關的某特殊事件有關、具有再利用價值、經專家推薦等。

(2) 挑選出待保存之網站後，檔案管理人員需界定網站典藏範圍，至少包含以下項目：網站抓耙深度、網站抓耙頻率等。

(3) 界定網站典藏範圍後，應再針對以下項目進行評估：

- 欲抓耙此網站多少內容：抓耙前先瀏覽網站以確認是否僅需此網站的網頁內容或包含其外部連結頁面，以確立抓耙之範圍。
- 此網站連接多少伺服器及主機：多數提供大量多媒體及 PDF 檔案的網站均包含外部連結網頁，應確認抓耙工具有能力抓取所有需要的 URL 種子（seed URLs），亦即抓耙工具開始之處。
- 此網站是否被保護：可檢視此網站，以確認是否限制抓耙。
- 抓耙工具的限制：此網站是否包含抓耙工具所無法抓取的部分，例如資料庫及 FTP 裡的資料等。

2、 網站授權取得階段

集中式或分散式網站典藏僅在「網站授權取得」階段有差異，因集中式網站典藏會徵集到外部機關之檔案，因此需向外部機關提出授權許可之申請；分散式網站典藏主要是徵集所屬機關之檔案，因此僅需向所屬機關提出網站典藏通知即可。集中式網站典藏可能包含提出授權許可申請及網站典藏通知兩種作業，分散式只需提出保存通知即可。

3、 網站典藏保存階段

(1) 檔案管理人員或資訊人員應設定網站典藏對象，包含網站名稱、URL 種子、範圍、時間、排程選項、描述資料及網站權益資料等資訊。

(2) 檔案管理人員或資訊人員可透過網站典藏管理系統抓耙網站。

(3) 當網站典藏管理系統抓耙資料完成後，檔案管理人員或資訊人員需檢視網站內容，以確保網站內容之完整性與可及性。此外，檔案管理人員或資訊人員

也可藉由與該網站前次的抓耙內容相互比對，以確認採相同或相異的設定方式進行。

(4) 若經檢視後，發現網站正確或如預期時，則由檔案管理人員或資訊人員透過網站典藏管理系統將此網站妥善保存。若經檢視後，發現網站內容並不完整或不如預期，則由檔案管理人員或資訊人員透過網站典藏管理系統重新抓耙網站。

網站典藏管理系統之內容，如圖 2 所示。管理者（administrator）係指負責執行網站典藏工作的人員，而使用者則是指瀏覽網站者。管理者透過管理工具（curator tools）設定典藏系統來自動抓耙與呈現網站，而使用者則是透過存取工具（access and finding aids）瀏覽有興趣的網站內容。就系統運作流程而言，首先，管理者操作管理工具依據其設定，透過蒐集工具（acquisition）蒐集網站內容。然後，將蒐集回來的網頁快照（snapshot）成 WARC（Web ARChive）網頁典藏格式，加以保存。最後，使用者來檢調應用時，網站典藏管理系統將此網站典藏資料藉由存取工具（access and finding aids）呈現在使用者面前。

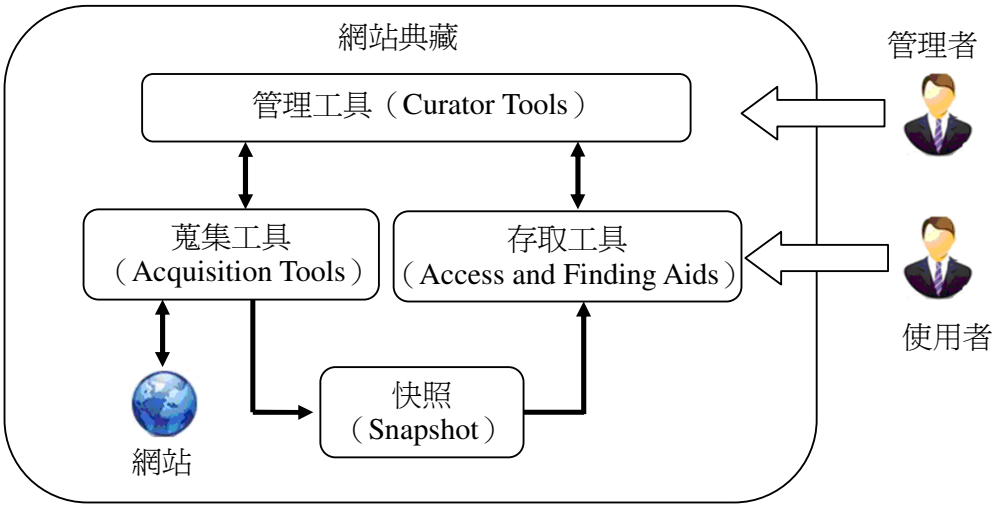

圖 2 網站典藏管理系統

進行網站典藏作業時，應先將網站內容匯出，藉由事先設計之網頁詮釋資料格式，將所匯

出之網頁加以著錄相關詮釋資料內容，以成爲對機關而言之一般化電子檔案，儲存於網站典藏管理系統，亦即特殊類型之檔案管理系統中，使網站內容納入一般電子檔案管理的範疇中。其概念如圖 3 所示。

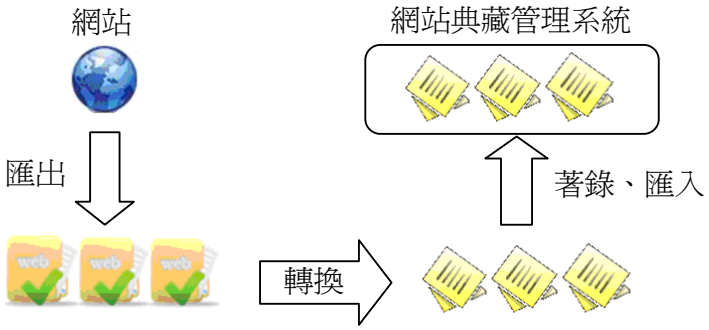

圖 3　網站典藏作業

(二) 網站典藏工具

常見之網站典藏工具，列舉如下：在蒐集（acquisition）方面，包含 Heritrix、HTTack 等工具；在存取（access）方面，包含 Wayback、WERA、NutchWAX 等工具；在管理（curator） 方面，包含 Web CuratorTool、PANDORA Digital Archiving System、NetarchiveSuite 等工具。

(三) 網站抓耙方式

機關可依其需要決定在伺服器端備份網站資料，在瀏覽器端抓耙網頁畫面，或是使用抓耙軟體典藏網站。從伺服器端抓耙網站資料，將有機會完整地抓耙到網站的所有內容。然而，越來越多的網頁會配合使用者需求而動態產生，例如瀏覽器的樣式、螢幕大小，桌上型電腦或智慧型電腦等。要保存何種版本之網站內容，亦是一個需考慮的議題。若從瀏覽器端著手抓耙網站，則採用快照（snapshot）軟體進行抓耙，但僅能抓耙到網站中不會變動的網頁內容。若透過抓耙軟體來典藏網站，則或許能解決前兩者之部分問題，但依然可能會遺失外部資源，例如就無法囊括由資料庫所產生之內容。

(四) 網頁典藏格式

從伺服器端進行典藏網站是以鏡射（mirror）方式完全備份原始網站檔案，此種方式保存下來的檔案格式會與原有網站的檔案格式完全相同。從瀏覽器端的保存方式會以圖檔的方式將網站頁面儲存下來，檔案格式是以常見的 JPG、GIF 及 PNG 等影像檔案格式爲主。若使用抓耙軟體進行網站之典藏，則建議採用 ISO 28500 所提出之 WARC 網頁典藏國際標準格式，以符合國際間的發展趨勢。

四、 結論與建議

網站資訊記載著機關發展的軌跡，並且見證歷史的脈動，然因政策之更迭或作業程序之改變，導致政府機關網站改版，致使今日所看見之網站內容在未來可能消失。由於資訊科技蓬勃發展且大量應用於辦公室之文書處理作業之中，因此許多先進國家已將電子郵件與網站內容納入電子檔案之定義範疇，以維護檔案管理業務之完整。然而，若要順利推動網站資訊之典藏，則須結合檔案主管機關與各級機關協同合作，修訂相關法規，將網站內容納入機關檔案管理之範疇，使機關首長瞭解網站典藏之目的與效益，並使網站典藏作業與檔案管理作業流程相仿，亦併入日常作業。藉由教育訓練宣導網站典藏之目的與方法，推廣網站典藏概念與做法，消彌檔案管理人員與資訊人員心中的疑慮，避免推行網站典藏時之阻礙。

在法令上，著作權爭議是網站典藏所面臨的最大法律障礙。因此要謹慎挑選網站來源，結合有效的權利管理政策（rights management policy），確保僅通過授權的特定人士才能取得具爭議或機敏的資料，並且在發生抵觸法律及關係人權利時，立即移除相關網頁內容，以有效地在網站典藏與著作權保護中取得平衡（Charlesworth, 2003）。

因爲保存「每個」網站幾乎是不可能，因此建立優先順序是成功典藏網站資料的重要基礎。在決定何者該保存或不該保存時，可採用 MoSCoW 分類方式以標明其重要性。在考量機關之成本、風險、優先權、可得資源及基礎設備等因素後，機關可採行自製、外包、協作、聯盟等方式進行網站典藏。機關將網站內容納入電子檔案保存範疇時，其所將要負擔的建置成本其實是相對低廉的。機關可依需要，選擇在伺服器端備份網站資料、瀏覽器端抓耙網站畫面，或是使用抓耙軟體典藏網站。在硬體方面幾乎無需增購任何設備，而也可以採用免費軟體，以節省經費，例如可採用 Heritrix 做爲網站典藏的蒐集工具，採用 Wayback 做爲存取工具，採用 Web Curator

Tool 做爲管理工具。同時，應將網站的詮釋資料納入保存範圍，以便網站被抓取之後進行自動編目。機關也可利用此詮釋資料進行檢核，掃描並檢查確認所設計的網站是否符合標準理想。過程中，可藉由網站版本的比對，進行版本控制，以節省大量的儲存空間。

因爲可以在不花費額外軟硬體設備費用之下進行網站的典藏，因此只要得到機關首長的支持，就能制定網站典藏相關政策與制度，提升檔案管理人員與資訊人員的網站典藏概念與能力，規劃推動網站典藏的流程，以使機關網站獲得良好保存，擴大典藏機關之珍貴記憶。

參考文獻

1. Charlesworth, A., “A Study of Legal Issues Related to the Preservation of Internet Resources in the UK, EU, USA and Australia,” 2003, http://library.wellcome.ac.uk/projects/archiving_legal.pdf
2. Day, M., “Collecting and Preserving the World Wide Web: A Feasibility Study Undertaken for the JISC and Wellcome Trust,” 2003, http://library.wellcome.ac.uk/projects/archiving_feasibility.pdf
3. Murray, K. R. and I. K. Hsieh, “Archiving Web-published Materials: A Needs Assessment of Librarians, Researchers, and Content Providers,” Government Information Quarterly, Vol. 25, 2008, pp. 66-89.