臺灣學數位圖書館發展、應用與開展

文／吳宇凡（淡江大學資訊與圖書館學系兼任助理教授）　圖片提供／國立臺灣圖書館

國立臺灣圖書館（以下簡稱國臺圖）自2007年開始，在數位典藏計畫以及館務的支援與推動下，構建「臺灣學數位圖書館」，累積十六項特色資料庫，打造以臺灣學為核心的數位資源館藏，讓讀者得以跨越時空隔閡應用豐富的館藏臺灣資料。

國臺圖於2007年3月奉核成立「臺灣學研究中心」，著手進行「館藏日文臺灣資料數位典藏計畫」，戮力各項藏品的數位化與知識結構的建立。除館藏的數位資源外，更蒐集市面上各項出版與灰色文獻，構建「臺灣學數位圖書館」（Taiwan Studies Digital Library），充分運用館藏資源，支援各學術單位及相關人員在研究上的需求，厚實臺灣學研究與教學之基礎。

國臺圖數位典藏成果的積累，不僅為社會大眾提供不同於以往應用文獻的方式，改變讀者使用文獻的媒介，館藏數位化帶來的新技術、數位內容、詮釋資料（metadata）、物件脈絡（context），在現今數位人文、人工智慧的衝擊下，同樣為我們帶來不同於以往的研究範式（paradigm）或方法（method）。

▲ 2007 年 3 月，國臺圖奉核成立臺灣學研究中心。

值此背景之際，臺灣學數位圖書館如何因應，是本文期待探討的議題。

數位典藏豐碩成果

臺灣學數位圖書館的20項特色資料庫，根據資料庫的內容特徵與來源進行區分，可分作「日治時期館藏數位化」、「臺灣學資源整合與徵集」、「目錄索引與加值」，以及「圖書館館刊與其他文獻徵集」等四種類別，分述如下：

日治時期館藏數位化之資料庫，主要係以國臺圖館藏數位化為基礎建置，包括「日治時期圖書影像系統」、「日治時期期刊影像系統」、「館藏古文書影像系統」、「地圖資料庫」，並從前述資料中再聚焦完成「寫真帖資料庫」、「館藏南方資料影像系統」、「館藏舊籍日本文獻影像系統」。這些館藏尤其以日治時期國臺圖前身臺灣總督府圖書館、南方資料館所典藏者最為難得，對於過去使用者無法接觸之珍貴文獻，在這些資料庫中得以一親芳澤。

▲日治時期期刊影像系統。

▲臺灣文獻期刊論文索引。

其次為臺灣學資源整合與徵集之資料庫，厚實臺灣學研究之底蘊，以臺灣學為核心所建置之整合型資料庫，諸如「臺灣學電子資源整合查詢系統」、「臺灣學數位電子書資料庫」、「臺灣學數位典藏查詢系統」，將國臺圖及各典藏機構、來源之數位資源整合，讓社會大眾得以了解現階段國臺圖數位典藏成果、各典藏機構、來源臺灣學相關數位資源，並擇選數位典藏成果轉製之電子書，俾利進一步的利用與觀看。

目錄索引與加值型之資料庫，一直以來都是圖書館工作的核心，而臺灣學數位圖書館所建置此類型之資料庫，同樣係以臺灣學為出發點，打造國臺圖所典藏或徵集世界各地出版、未出版之臺灣文獻目錄，包括「日文舊籍臺灣文獻聯合目錄」、「臺灣文獻期刊論文索引」、「臺灣文獻書目解題」（已導入國臺圖數位典藏查詢系統），讓讀者得以迅速掌握臺灣學研究資源，並由館內外專家學者，共同針對文獻書目進行說明，從而更深入文獻內容。

▲臺灣學數位電子書資料庫。

最後一項圖書館館刊與其他文獻徵集之資料庫，除「臺灣學通訊資料庫」係以圖書館所發行刊物《臺灣學通訊》為基礎建立之資料庫外，並含括國臺圖逕行徵集之其他文獻類型所構建之資料庫，如「臺灣政經人文剪報資料庫」、「臺灣名家手稿資料庫」，將臺灣研究基金會、黃光男、陳鴻瑜等所提供資料，抑或各專家學者手稿，整合成為可供檢索應用之系統，從而完善臺灣學文獻之保存與提供。

臺灣學數位圖書館所累積數位文獻完整度高且類型多元，諸如文字、圖像／照片、插畫、地圖、手稿、廣告，並有結構式的詮釋資料，俾利進一步應用與分析。在數位化內容的進程上，考量

▲臺灣政經人文剪報資料庫。

▲館藏古文書影像系統。

有限的成本及供使用者查找之目的，多以非文本全文方式進行內容建置，換言之，前揭資料庫所含數位資源，以數位物件與詮釋資料內容為主，也因此在日後應用與發展上，亦僅能就現有資料進行拓展，許多涉及本文內容分析者，另待進一步資源投入。

當數位內容成為基礎

數位人文可以說是一種方法或是一種研究取徑，利用統計學、資料探勘、資訊視覺化，以及各項數位內容相關技術，探掘資料與資料的關係中隱含的訊息，為過去各界苦思如何將數位典藏成果，得以從「藏」轉化為「用」的課題中，提供資料應用與學術性的方向。根據項潔所稱，數位人文是那些「唯有借助數位科技方能進行的人文研究」，換言之，數位人文研究的發展立基於數位典藏的基礎上，而臺灣在數位典藏相關計畫的推動下，自然具備發展數位人文的環境。

數位人文研究成果不勝枚舉，在此不多作贅述。根據相關研究所採用方法，多數係以文本標記、詞頻／共現詞分析、社會網絡分析、內容比對、地理資訊系統（Geographic Information System）等方式，進一步探掘資料背後的脈絡訊息；這些方法涉及細部技術，諸如自然語言、斷詞、視覺化、統計、演算法、地理座標採集與標示等，絢爛的成果與視覺化呈現，讓學界掀起一陣旋風。隨著理論與技術逐漸成熟，以及相關單位分析工具的建置與提供，從事數位人文研究的門檻逐漸下降，甚至成為一門顯學，相關研究如雨後春筍般出現在各類學術發表上。

當資料豐富且類型多元的臺灣學數位圖書館，成為從事臺灣學數位人文研究者的重要研究標的，人們覺得資料探勘、數位人文技術就是數位化館藏的核心應用方式時，另項劃時代的內容分析／應用方式出現在人們的日常生活中。2022年底，美國獨立研究機構OpenAI公開ChatGPT3.5的服務，「生成式人工智慧」（Generative AI）在人們驚嘆中，躍入數位內容的世界。

姑且不論生成式人工智慧對於各行各業帶來的巨大影響，人工智慧因其訓練之語料與基礎，將影響所生成內容之在地性及可信度，世界各國或各機構無不厚實自我人工智慧發展基礎與能力，除技術外，尤其關注各類型典藏機構在相關技術發展之因應，思考過去投入的館藏典藏在這波衝擊下的意義與功能。

打造臺灣語料庫／詞彙庫

無論係資料探勘、數位人文研究，抑或人工智慧的訓練，系統訓練使用的語料庫／詞彙庫，決定了斷詞、後續分析及自動生成的準確度與可信度。現階段臺灣並無適切之日治時期詞彙庫／語料庫，致使系統在斷詞與分析上要花更多時間進行摸索與試誤，自動生成更容易產生疑義之內容。

臺灣學數位圖書館的數位內容隱含大量日治時期臺灣語料、圖像，並有相對應之詮釋資料，若能解決現階段缺乏全文的問題，以此為基礎進行內容分析產生日治時期臺灣之語料庫／詞彙庫，從而供社會大眾作為相關研究與應用，將使事半功倍、獲益良多。

建立臺灣圖像基礎

數位典藏的特色在於每一項數位物件皆對應一項詮釋資料，透過這些詮釋資料，可以清楚知道數位物件之各項訊息，諸如產生者、產生年代、地點、來源等。這樣的訊息在過去僅為方便使用者資料查找，然在資料探勘、數位人文或人工智慧的應用上，則可作為系統訓練、自我學習的基礎。

臺灣學數位圖書館有豐富多元的數位內容，尤其在圖像上整合日治時期臺灣各項文獻中的影像資料，隨著圖像分析、自動生成的技術與應用逐漸成熟，諸如新加坡網站colourise.com利用新加坡國家檔案館大量影像進行訓練，得以更精準進行黑白照片上色與修復，相關資料的整合與分析得以使日治時期臺灣影像的概念更加明確，在這樣的基礎上，使用者得以藉此建立日治時期臺灣圖像分析的基礎，完成進一步應用。

深化主題與數位策展

提供社會大眾藉由主題性的資料整合，得以進一步建立特定議題的認識，成為當代圖書館知識提供的一種方式。

過去，國臺圖辦理各項以臺灣學為核心的實體展示，透過策展人的思維與脈絡，進行議題與展出內容的規畫與選擇；隨著數位內容的完善，資料探勘、數位人文及人工智慧技術的發達，探掘過去難以發現的資料間關係成為容易的事情，也因此相關技術得以應用於日治時期臺灣主題之深化，協助策展者發現過去沒有注意到的議題或視角，並藉由數位物件之匯聚、連結與整合，進行數位策展，讓圖書館的知識提供越加活潑、多元，甚而透過數位物件在多媒體與相互連結上的特性，令展覽得以更加深化且動人。

國臺圖在臺灣學相關文獻的深度與完整性，一直以來都是臺灣在地研究發展的核心。人工智慧並非憑空而生，藉由過去人們日常的積累，方能構建未來的智慧；無論是在資料探勘、數位人文發展，或是人工智慧的訓練與應用，在超大型類神經網路的演算之下，過去臺灣學數位圖書館累積之豐碩數位資源，成為日後應用與開展的根柢，「我的人工，你的智慧」，凸顯了臺灣過去數位化資源投入的價值與意義，更感念前人在數位化歷程的努力與付出。