建置「臺灣歷史人物傳記資料庫」（TBDB）的嘗試與初步成果

張素玢*

一、建置 TBDB 的動機

人物為歷史學研究重要的一環，從人物的各種面向與牽引出的出身經歷、社會階層、人物流動、婚姻網絡、政治網絡，都是研究歷史的關鍵議題。過去對人物相關網絡、家族譜系的研究主要依循傳統史學方法，以人工方式從資料文獻加以分析排比，或可滿足對單一人物、單一家族的研究，然而涉及到龐大的資料與複雜的網絡，往往力有所未逮。哈佛大學首先集合學界力量，積極經營中國歷代人物傳記資料庫（China Biographical Database，以下簡稱 CBDB），期待透過建立基本檢索系統和文本分析的功能，使人物傳記資料庫成為人文學者之研究與分析工具，以跨越個別人物傳記之局限，對人物的互動產生更動態的觀察。

目前臺灣的學界對人物研究主要遵循傳統史學的文獻研究法或輔以口述歷史，對單一人物或家族的研究都有不少成果。但是傳統的研究方式，每一主題皆需經年累月，一一爬梳文獻，比對資料，處理極為費時。儘管各研究者對某地區的家族、士紳或農村菁英有所深究，卻難以梳理眾多人物的網絡關係。若研究範圍擴大至全臺灣，其關聯性複雜度將至少與人物數量之平方成正比。若要使臺灣史之人物研究可支持人物、時間、空間的擴展性，且能使臺灣史學者於同一平臺上合作、分享研究成果，臺灣的歷史人物傳記資料庫與相關分析工具的建置已刻不容緩。

臺灣史的研究經過幾十年的積累，研究主題由小區域、細緻，慢慢推向更寬廣與更長時距的議題。方志的纂修、人物家族、社會網絡的研究，都有可觀的成果。過去 20 年政府積極推動資料庫建置工作的完成，也使資料的流通與檢索更方便。儘管群體人物傳研究的主觀、客觀條件日臻成熟，臺灣卻一直沒有

* 國立臺灣師範大學臺灣史研究所教授

具有勘考、文本分析功能的歷史人物資料庫。發展此種歷史人物資料庫，事實上牽涉到對資料庫涵蓋範圍之歷史脈絡具備深入的瞭解。例如定義人物於資料表中屬性時，必須考量資料庫涵蓋範圍之時空背景，而不能一體適用。筆者於是動心起念，想建置一個具有勘考與查詢功能的臺灣歷史人物傳記資料庫，在得到科技部的補助之下，開始邁開第一步。本計畫「臺灣歷史人物文本探勘與社會網絡分析工具：以《新修彰化縣志．人物志》為對象」由於基本成員為《新修彰化縣志．人物志》之撰寫團隊，對人物之屬性種類與可能之屬性值瞭解深入，因此以《新修彰化縣志．人物志》為基礎，從彰化縣人物開始建置 TBDB。

TBDB 是在中國歷代人物傳記資料庫（CBDB）的基礎上，依照臺灣近現代人物與時代的特色，修訂並擴充 CBDB 的類別架構，以《新修彰化縣志．人物志》作為臺灣歷史人物傳記資料庫（Taiwan Biographical Database，以下簡稱 TBDB）人物屬性建置與文本探勘的起始。TBDB 應用文本探勘與社會網絡分析工具於兩個主題，並評估工具之效度和信度，使資料庫能確實符合歷史人文學者的研究需求。完成 TBDB 初步成果後，開放學界利用查詢，並進行除錯。日後繼續擴充文本資料，加入其他縣市之人物志、日治與戰後的各類人物辭典，逐漸延展到全臺灣的人物傳記。

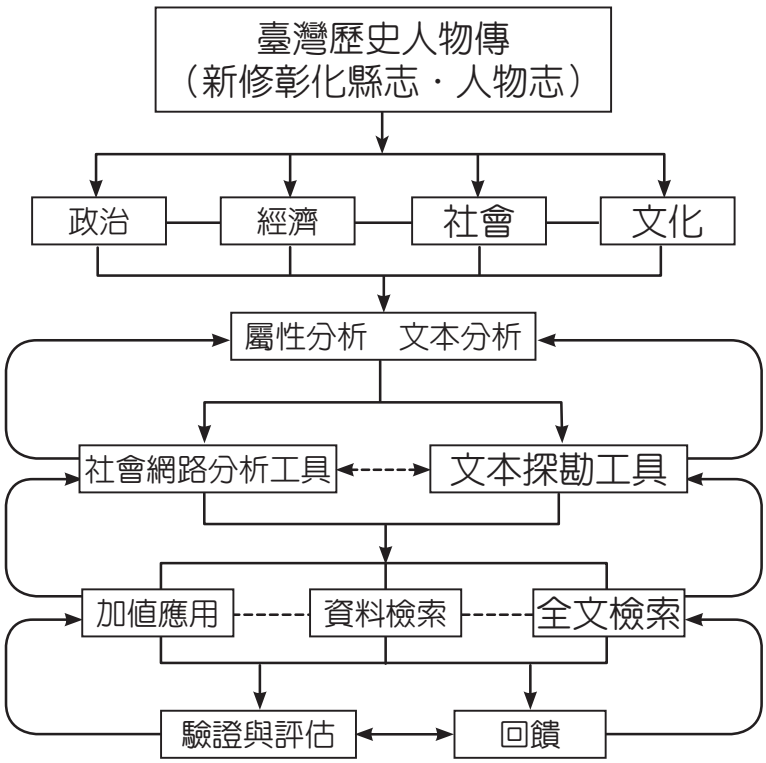
二、人文與資訊學者的合作

臺灣歷史人物傳記資料庫（TBDB）的建置由歷史人文與資訊科學兩領域的學者共同合作。歷史學者為《新修彰化縣志．人物志》的班底，包括筆者、中研院臺灣史研究所顧雅文助研究員、中興大學歷史學系李毓嵐副教授、朝陽科技大學李昭容助理教授，資訊科學領域的學者有臺師大圖書資訊學研究所柯皓仁教授、中央大學資訊工程學系蔡宗翰教授，臺師大歷史學系李宗翰副教授因有參與建置 CBDB 的經驗，其分工在資訊組。

歷史人文學者運用領域知識，負責分析資料層中的臺灣歷史人物志文本，藉以產出人物類別、屬性、關聯等知識本體（ontology），以及權威控制（authority control）、索引典（thesaurus）、主題詞表（subject terms）和命名實體（named entity）等輔助資料庫（auxiliary database）的建立。並提出功能與使用介面需求，以及探勘分析結果的品質控管，以使 TBDB 能確實符合歷史人文學者的研究需要。

資訊科學學者主要負責 TBDB 系統的開發，包含設計人文學者分析產出知識本體、輔助資料庫的關聯式資料庫系統，根據臺灣歷史人物志文本的全文資

料進行文本處理、文本探勘工作，以發掘傳主的類別、屬性、關聯等資訊並自動填入 TBDB；設計社會網路分析與檢索功能，以利使用者探索傳主間的社會關係；最後並以使用者中心（user-centered）概念設計 TBDB 全文檢索、資料庫檢索介面以及更多的加值應用服務。以上的過程都透過密集會議，每個月定期討論，以獲得具體結論，維持資料庫建置與計畫執行的進度。



圖一：TBDB建置基本架構與流程圖

三、TBDB 檢索系統的建置

TBDB 的知識本體來自《新修彰化縣志・人物志》，各篇傳文雖有共同的行文方式及段落重點，但不像 CBDB 引用已有格式化體例的「傳記資料索引」等資料，能夠較為快速將人物資料輸入並結構化；TBDB 則需要藉由文本探勘技術，從《新修彰化縣志・人物志》的傳文中挖掘資料，再加以分類及結構化。因此本資料庫系統建立探勘規則的關鍵字進行測試，確立人物屬性類別的骨架，再協助進行權威控制、除錯檢驗等工作，使探勘結果能正確置入資料庫，檢索系統更加完善。

資訊組提出一個半自動方式，擷取《新修彰化縣志．人物志》中的特定專有名詞。以詩社為例，經由詩社名稱之輔助資料庫，求得輔助資料庫內平均詩社長度、詩社最大相同字數、相似字串對等資料。並引入停用字詞表，過濾部分不具意義字元，以修正結果。

此做法可用全自動的方式進行，並適用於各種擷取情況。透過專家（人文組）建置的輔助資料庫自動運算並產出規則以進行擷取。但由於電腦演算法尚無法妥善處理語意關聯，僅就語法、拼字組合進行擷取，而易導致偏差，故實務上，仍需適時的人工介入，並進行結果的修正。另一方面，系統自動判斷的「應社、旗津吟社、鹿港詩社、斐亭鐘會」等詩社卻未能被人工檢出，這也顯示以人工擷取專有名詞，仍可能存在一定的失誤與不一致，故而以系統自動化辨識專有名詞輔助人工複檢乃是較可行的作業方式。經過一年多的努力，目前完成以下檢索系統與分析工具：

1. TBDB 資料庫檢索系統
   （1）四項檢索點：人物類別（文化、經濟、政治、社會）、姓名、出生地、傳文。
   （2）後分類功能：提供人物類別、出生地和詩社三層面（facet）限縮檢索結果。
2. 社會網絡分析功能
   （1）「另有傳」人物網絡化。
   （2）詩社成員的地理分布：展現詩社、地點、詩社成員間的關係。
3. 人物空間資訊呈現
   （1）詩社成員的出生地分布情況。
   （2）傳主活動地分布視覺化（熱度圖）。

四、社會網絡關係圖的製作與分析

計畫執行的第一年，為提供資訊組開發文字探勘及社會網絡分析工具的作業目標，人文組負責收集整理重要社團、組織及人名錄清單，作為測試對象。包括中部地區重要詩社參與人士、彰化銀行大股東、臺中中學校設立捐款人、二林蔗農事件相關人士、臺灣議會設置請願運動參與人士、臺灣文化協會彰化支部成員、戰後員林客運股東名錄等名單。

資訊組將這些人物的節點串聯製圖後，由人文組檢驗、解讀及分析，從中發現：

(一)文化性社團成員與政治結社運動之關係

以臺灣議會設置請願運動為中心觀察，可發現不同詩社成員參與政治結社運動的人數差異，「應社」成員參與政治活動的程度相當高，由此提醒研究者要注意文化性社團在政治活動中扮演的角色。

(二)彰化銀行大股東與臺中中學校設立捐款人的重疊性高

兩者皆為臺灣中部地區資本雄厚人士，此圖的連結關係符合歷史認識。

(三)經濟菁英與政治社會運動之關係

從圖中可瞭解臺灣在日治時期的政治與經濟活動之間，還有一層較隱晦的聯繫關係，需要特別留意文化協會彰化支部中政治活躍分子在其中所扮演的仲介功能。

五、建置經驗與心得分享

建置一個新的資料庫，在發展各種功能或工具以前，其實人文與資訊學者的磨合，人腦與電腦的對話更為重要。在研究在建置資料庫的過程中，有下列經驗和心得——

1. 計畫團隊由歷史學者和資訊科學家所組成，這兩個領域使用不同的認知語言，最初須克服相互瞭解的困難，以開放的心胸並定期聚會討論，促進雙方的理解。
2. 絢麗的工具不見得是適用的工具。在建置 TBDB 初期，資訊科學成員熱心地實作了許多工具，例如社會網絡分析、時間軸、熱度圖，這些工具也的確讓歷史學者讚嘆。然而當歷史學者進一步檢視，發現這些工具不僅使分析更為複雜，甚至難以解釋。一個好工具最重要的是能符合歷史學者的需求和研究流程、才能發揮其研究效益。
3. 隨著數位與資通訊科技時代的來臨，許多人文社會學者急於跨入數位人文領域並喜於運用許多資料庫和工具，而可能淪於「為數位人文而數位人文」。研究者認為僅有在人文社會學者擁有問題意識的前提下，資料庫和工具才能夠發揮它們最大的效用。
4. 數位與人文無縫接軌是個理想，事實上存有相當大的困難度。團隊中歷史背景的學者，要在一年內充分瞭解設計工具的需求，而提供類別、屬性、關聯等知識本體，還要瞭解到怎樣才是電腦能夠理解的方式，對人文學者而言，是要費力跨越的學科鴻溝。

5. 計算工具常常有很多限制，人文學者必須瞭解這些障礙，找出兩者可溝通的方式。同樣的，開發工具的資訊學者，找出規則或勘考關鍵字，呈現出來的視覺效果似乎頗具效能，但事實上並非合理的設計。因此團隊中的兩組人馬必須透過經常性的討論，讓彼此的思考頻率能共振；歷史學者分享其研究經驗與成果，提供資訊學者瞭解的研究議題，萃取必要的元素。

透過以上過程，才能讓資訊工具端成員確認人文學者所需，而不會設計出有技術但不實用的工具，有一些研究議題也常在共同討論中被發掘出來。

六、TBDB 未來之路

臺灣歷史人物傳記資料庫才剛邁出艱辛的第一步，未來的工作可分短期和長期目標：

(一)短期目標

1. 本研究已完成雛形系統的建置，並對外開放使用，將根據使用者回饋的意見改善系統功能。
2. 為自動化填入 TBDB 人物的欄位，未來將繼續辨識所有命名實體，並開發自動偵測親屬與社會關係的功能。
3. 人物的親屬和社會關係錯綜複雜，需要設計良好的社會網絡分析和視覺化工具。除了地理資訊外，時間資訊對歷史研究也是十分重要，未來將發展有助於探索時間資訊的工具。
4. 研究團隊未來將從 TBDB 構思新的研究議題，並展現 TBDB 能如何協助歷史研究。

(二)長程規劃

1. 擴大資料庫的量

日後繼續擴充文本資料，加入其他縣市之人物志、日治與戰後的各類人物辭典，逐漸延展到全臺灣的人物傳記、日記等，目前已完成的各種相關資料庫整合更是各單位必須共同努力的目標。在擴充過程也因新資料、新人物類型不斷加入，對資料庫欄位（即資料庫所收錄的資訊），以及分析工具從事滾動式修正，才能成為對研究有實質幫助的資料庫和勘考分析工具。

2. 提升資料庫的質

臺灣歷史人物傳記資料庫才建置一年多，仍處於草創階段，每一項檢索功

能與探勘工具都有待更全面的檢視和修訂。未來資料量不斷擴增時，資料歧異性的問題也會隨之產生，使原有檢索系統和工具不符期待，系統必需不斷「進化」。

3. 新工具的繼續開發

TBDB 建置的第二年重點放在「社會網絡分析」，因為近 900 位的彰化人物，有著綿密複雜的關係，已超出人力所能處理，卻是數位工具大可發揮之處。其次家族之內或不同職業群的「關聯性分析」，也是工作的重點之一。再者，時間資訊對強調時間軸線的歷史研究相當重要，若能在空間資訊以外再加上時間資訊，對研究可產生何種效益值得期待。

4. 人臉辨識

近現代的人物，除了文字資料，影像圖片也不少，老照片更是對人物的生平作為有詮釋作用。本團隊發現，老相片中的團體照往往是一群存在某種社群關係的人物，在具有說明作用的建築物前，在某個特定時間留下合照。這些人臉若能一一被辨識出來，其社群網絡就勘考完成，再加入建築物的辨識更能增加效益。人臉辨識技術已廣泛運用在科技產品，若是能克服技術層面的問題應用於老照片，則是建構社群網絡的另一種突破。

5. 研究議題的發想

數位工具建置的目的在為研究服務，而非展現絢麗的操作效果，所以研究議題的發想，才是工具開發的動力。臺灣學界多年來的人文研究成果，若能透過數位科技的方法和工具，可增廣研究主題和研究效益；同樣的，研究議題的提供，也使數位工具能持續開發進化。

6. 與國內、外數位人文學群接軌

TBDB 未來的發展也將努力與最新的數位人文發展接軌，亦即將 TBDB 建設成一個開放式的資料庫，可與其他相關資料庫或數位工具對接。例如臺大數位人文研究中心在項潔教授的主持下，由杜協昌博士所開發出來的 DocuSky 資料庫，提供可讓使用者針對自身需求建置客製化的資料庫與分析工具群。若能將 TBDB 與 DocuSky 結合，當能對學術研究以及數位人文的發展做出更多的貢獻。

在國際上，也可透過 TBDB 建立起一個與國際相關研究的平臺，除了和 CBDB 對話，未來系統將建立 API，提供 BiogRef.org、TextRef.org 等平臺串接資料，並與其他已建置類似勘考、分析工具的系統互相比較學習（如 Ctext），吸取更多經驗，使臺灣歷史人物傳記資料庫 TBDB 更完善成熟。