

從CBDB到TBDB

從中國歷代人物傳記資料庫（CBDB）說起

文・圖——祝平次 清華大學中國文學系副教授

編者按：在本館通訊的55期，讀者們從〈擴增研究：臺灣人物傳記資料庫工作坊小記〉中，已讀到祝平次老師所介紹的資料庫概念，以及「彈指之間」的文學史料探勘，其前置作業以及運作的方式，而這些都是「數位人文研究」概念，為了讓各位讀者更加了解數位人文的內涵，特別邀請祝平次老師專文為大家介紹這個花費十餘年、結合跨國人才完成的「中國歷代人物傳記資料庫」，而臺灣如何運用其功能，從而開始規劃「臺灣人物傳記資料庫」，這將是臺灣文學與文化研究裡，一個重要的工程。

在本專題中，我們從外在的科技展示，到內在的人文探究，嘗試一步步與讀者一起進入臺灣文學研究的新領域，那會是更多元、更精采的世界。

本文要介紹的「中國歷代人物傳記資料庫」，英文名稱為China Biographical Database，簡稱為CBDB。（為了行文簡便，以下使用英文簡稱。）CBDB由哈佛大學費正清中國研究中心、中央研究院歷史語言研究所及北京大學中國古代史研究中心三個單位合作開發。¹目前收有約35萬筆的人物資料，主要集中在唐到清的人物。經過幾年的推廣，CBDB目前已為漢學界所熟知，是一非常著名的資料庫。下文將介紹它的組成和功能，並探討將它的設計理念轉化為臺灣人物傳記資料庫（Taiwan Biographical Database，簡稱為TBDB）的可能性。

資料庫的建置概念，就是把需要的資料擷取出來，賦予它一定的結構，以便之後可以很快地查詢、擷取資料。例如，《維基百科》〈臺灣歷史年表〉有以下條目，「1593年：日本豐臣秀

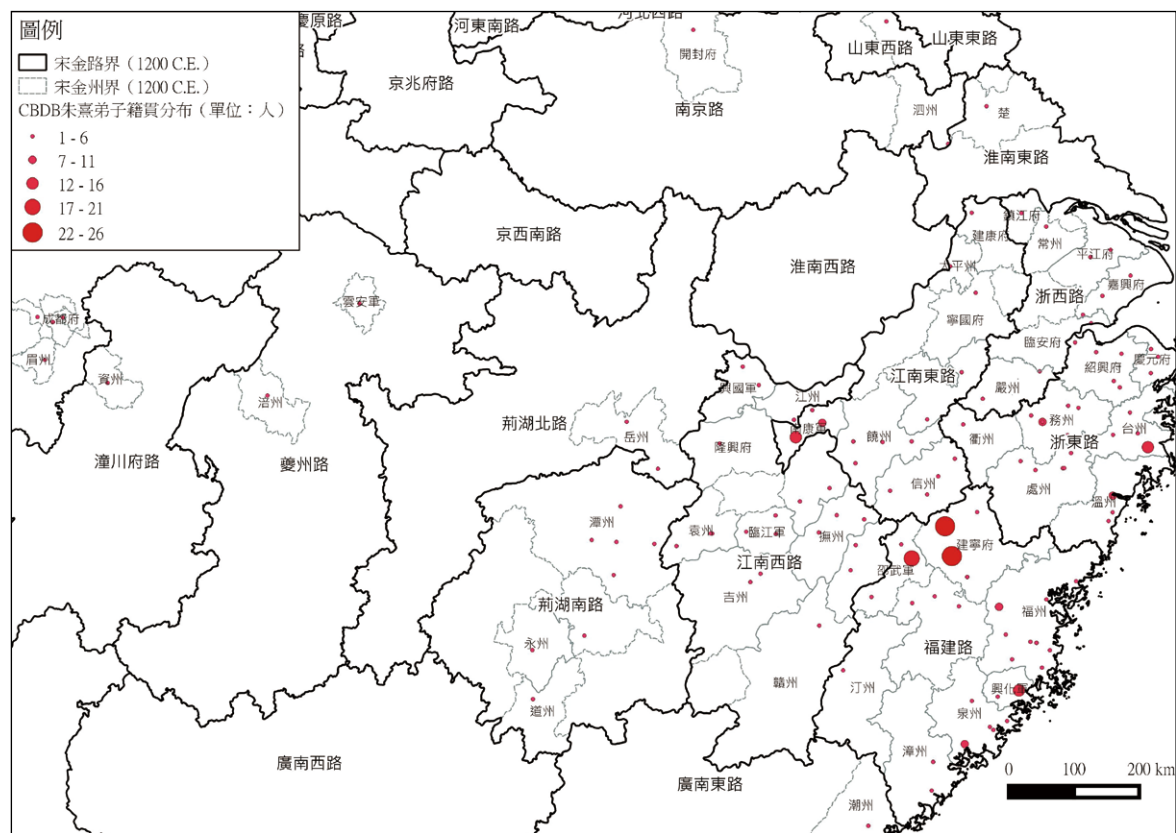
吉派家臣原田孫七郎到臺灣催促納貢未成。」這樣的條目可以變成：哪一年、什麼國家的什麼人派什麼人到什麼地方做什麼。經由這樣的轉化，就可以得到「時間」、「國家」、「人」、「地方」、「事務」等重要資料單位做為建置資料庫的結構。資料庫的概念很早就存在，像中國以前的類書、地方志可以說都具備了資料庫的形態。但因為是紙本的，所以就無法達到前面所說很快查詢擷取資料的作用。電腦的發明使快速利用調度資料成為可能，但就過去的事實來看，資料庫卻很少為人文學者所利用，一直到近十數年來數位人文的興起，才改變了這樣的狀態，而CBDB就是其中一個很好的例子。

CBDB為一關聯性資料庫（relational database）。它的主要的建置單位是一張張的表格。再藉由表格中相同的欄位，把不同的表格連

1. Harvard University, Academia Sinica, and Peking University, China Biographical Database (November 1, 2017), <https://projects.iq.harvard.edu/cbdb>.

結起來。例如有一張人物的資料表，有一張地名的資料表，再藉由第三張有人物識別碼（識別碼就是平常講的id）、地名識別碼的表，就可以把人物和地名進行關聯，表示出在一個地方曾經有哪一些人出現過；或一個人曾經去過哪些地方。如果再有一張職官表，就可以再用一張有著人物識別碼、地名識別碼、職官識別碼的表，來表示什麼人在什麼地方當什麼官。也就是說，有越多關於主要資料的表格，就能將它們任意關聯起來，描摹越多事物。

CBDB主要的資料，除了人物的姓名別號、生卒年、籍貫外，也記錄了人物的入仕方式、仕宦經歷、親屬關係與社會關係。另外，也記錄了各朝的地名與官職系統。在這些主要資料表格的基礎上，CBDB預設了三種主要功能：集體傳記學研究、社會網絡分析以及地理空間分析。集體傳記學不以一個個個別的人為研究對象，而是以一個特定的群體為研究對象。例如臺南20到40歲的男性市民或某家公司的成員有什麼共同的生活背景與經歷。在CBDB裏，可以很快地檢視宋



GIS (Geographic Information System 地理資訊系統) 圖例

代的宰執官員是否有明顯的共同經歷、明代進士的平均年齡是幾歲等等集體傳記學的問題。社會網絡分析則著重在任何的人都是活在特定的關係網絡當中，所以透過各種網絡關係，則可以幫助我們了解一個人在這個網絡中的角色、地位以及重要性。在現實社會裏，網絡分析可以被利用來分析學童霸凌的問題、企業人員組織的問題。在CBDB裏，則可以查閱宋代不同的士人利用書信這個媒介所建立起來的網絡，也可以查詢、比對士人的政治網絡與親屬網絡之間的關係。CBDB中的人物傳記資訊，還可以配合地理資訊系統（Geographical Information System, 簡稱GIS）做成各式各樣的地理資訊分析，例如進士多寡與人口密度之間的關係。也可以利用，不同時段的資料，對於地理分布的變遷進行觀察。這三種功能對於文史研究都非常重要，以前的學者雖然也意識到這些問題，但沒有適當的工具，在資料數量上並沒有辦法進行這麼大規模的處理。

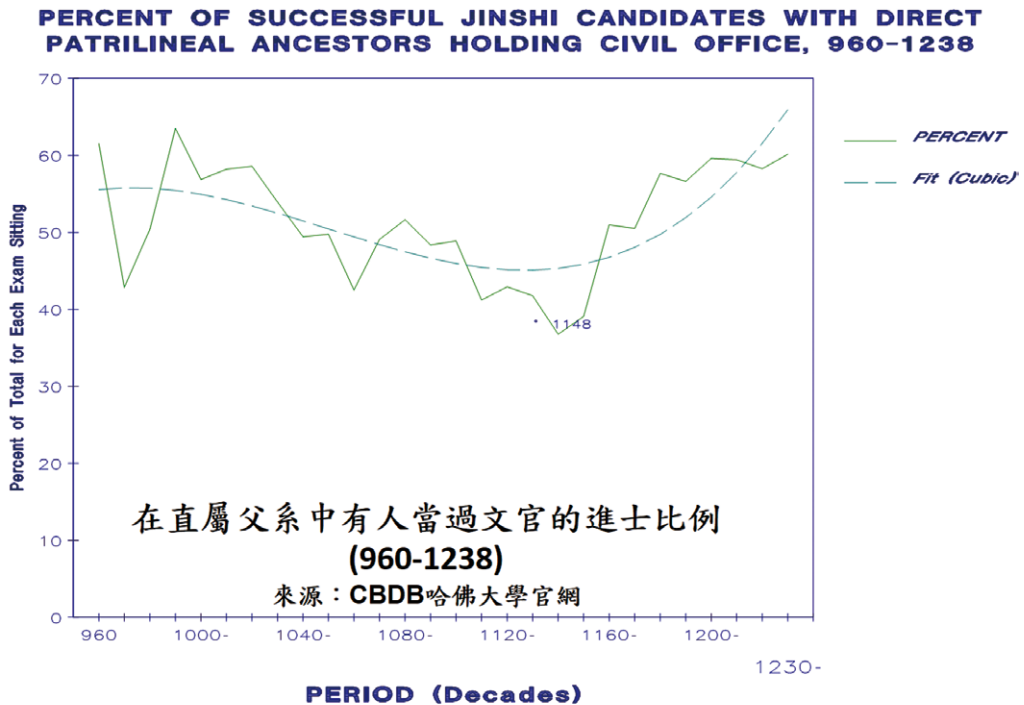
在現代化的管理而言，集體傳記學的方法並不陌生，舉凡稅收、公共衛生、人口學等等都與之密切相關。而在臺灣常講的政商關係，尤其透過聯姻的方式，也正是社會網絡分析的長項。至於地理資訊系統的運用，則可以幫助我們觀察家族遷移的現象。又如在2014年發表的「15萬名人出生地到墓地的視覺化」（<https://www.youtube.com/watch?v=4gIhRkCcD4U>），來探視世界史的遷移現象也是一個很好的例子。這也顯示理解古代人類社會和理解人類現代社會其實是可以利用同一種方式去進行，也使得文史研究的現代意義

更為突顯。而且因為它能處理的資料量，是之前學者很難想像的，自然也就有更多型態的研究，因而受到學界的重視，自然也容易引發別的領域的學者想要開發類似資料庫的想望。

的確，如果各地學者能夠合作，為自己研究的文史領域都建立一個類似的資料庫，對於學界的重大影響，一定非常巨大。因為資料庫可以不斷修訂、增補，而其分析、統計成果卻可以馬上呈現，對於知識的建置、再確認、擴展都不是紙本時代所可比擬的。哈佛大學的開發團隊原本也希望能夠釋出資料庫建置的架構，最後因為種種原因沒有完成。但資料庫的運用在現今的社會是一個普遍的概念，就像照相、繪製地圖一樣，在過去都是一種專業，現在卻是人人能做的事一樣。所以，即使不利用CBDB的資料庫結構，如何為研究臺灣的學者打造一個類似的資料庫，還是一個值得令人期待的嘗試。

今年5月12日，在國立臺灣文學館就聚集了一群這樣的學者：施懿琳長期蒐輯編寫《全臺詩》，從《全臺詩》詩人提要中，提取了八百多位的人名；張素玟則正在進行把《彰化縣志人物志》轉成臺灣歷史人物傳記資料庫的計畫；捷克學者路丹尼（Tá aDluhošová）則正在從事日治末期與戰後初期，臺灣文人的轉型研究，也著手建置資料庫。文學館也邀請了幾位對於數位工具比較沒有經驗，但是對於傳統研究方法卻非常嫻熟的學者，以及中央研究院臺灣史研究所的學者與會。雖然對於資料庫的架構，還沒有共同的看法；但對於應該要有資料庫來促進臺灣研究的進

群體傳記學圖例



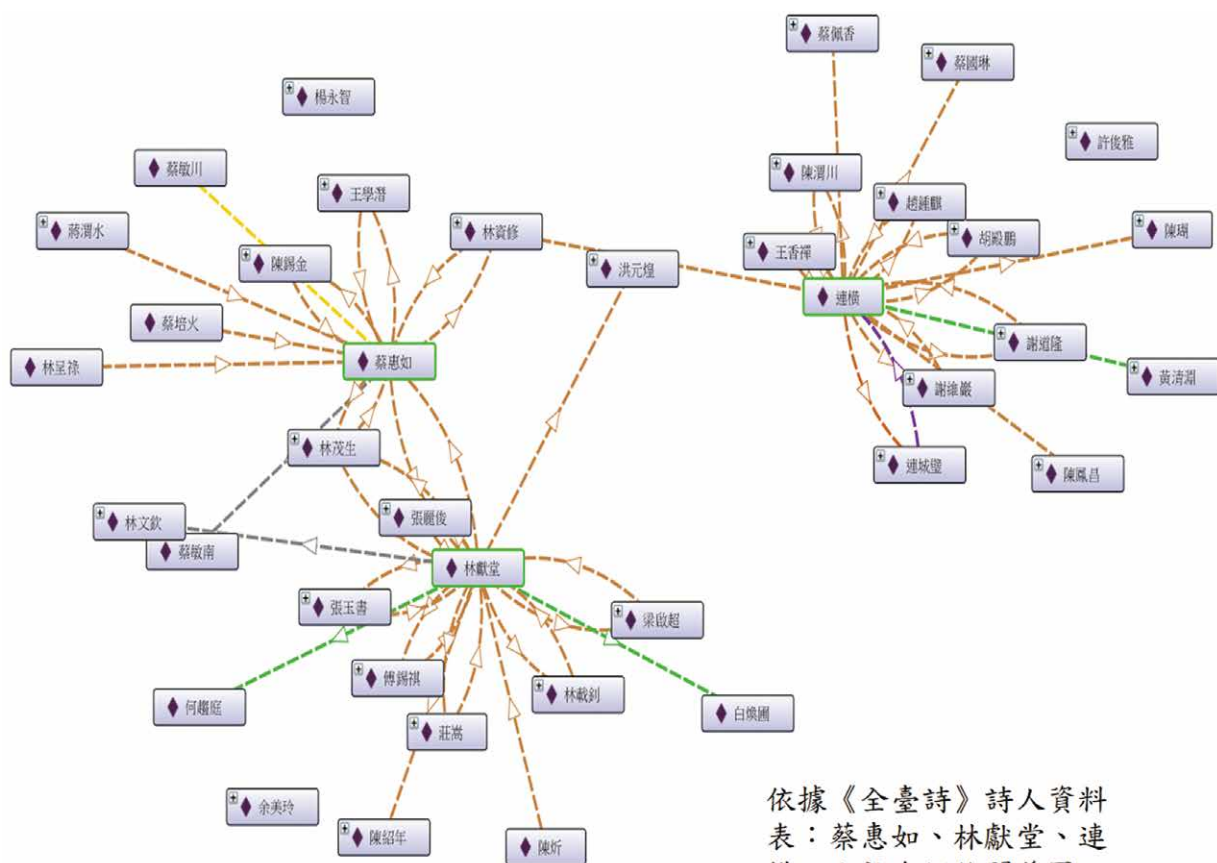
11:10:07 下午

2

程，則是有志一同。尤其資料庫中的資料，只要格式一致，將來要合併幾個不同的資料庫並不是問題。

就CBDB的成功來看，三個頂尖學術機構的決斷力、執行力與資源都不是其他機構可以比擬的。但即使如此，CBDB也會面臨繼續發展的一些難題，例如資金難以為繼、管理人員流動的問題等等。目前建置TBDB可能無法籌措到那麼多的人力與資源，但還是可以用其他的方式來達到同樣的效果。就資料建置來講，目前的重點應該是在整合現有的數位資源，尤其機構所擁有的各型

資料庫，如臺灣文學館、中央研究院臺灣史研究所以及近代史研究所等等。其實近十幾年各公家機構大量地數位化檔案資料，各式各樣的數位化資源應該不少。臺灣研究學者若能組織起來，與各機構協商，從現有的數位資料開始建置，就可以在比較短的時間內大量累積資料。至於管理方面，也同樣需要有個組織來進行，如果能有學會或跨國學術聯盟來進行，最為理想。若缺乏資金與人力，至少可以採取最低限度的管理，亦即組成線上學術社群，採取共建共享的方式，來慢慢擴充資料庫的規模。目前，各別研究者若不能一



依據《全臺詩》詩人資料
表：蔡惠如、林獻堂、連
橫三人社會網絡關係圖

社會網絡分析圖例

下子進行有規模的整合，也可以彼此交換資料，或設一個雲端硬碟，乃至於就直接放在GitHub (<https://github.com/>)，以開放的資料呼籲更多的投入者，畢竟學術原本就具有公開的性質，學術若不開放，只會扼制學術本身。若再考量現今個人電腦的效能越來越好，只要資料開放，對於怎麼使用資料還是有很多的可能性。嘗試怎麼用不同的方式，可以達成CBDB的效果，就學術群體的組織而言，也是在數位時代，應該進行的嘗試。

建置這個資料庫，也是省視我們對於臺灣知識的一個機會。如果學界可以分享一個這樣的資料庫，那臺灣研究的整體進程可以大大地前進一步。而且，這個資料庫的視覺化效果，也有助於學界外的民眾對於臺灣的過去與現況進行各式各樣的探勘，因此也能使得知識可以跨越一般職業的限定，成為大家共同文化的一部分，絕對是值得投入時間、人力、資源來實現的一件事。☒