圖書資訊學研究20：1 (December 2025)：47-85
DOI:10.30177/JLISR.202512_20(1).0002

檔案文本探勘研究：以特種檔案爲例

The Study on Archival Text Mining: A Case Study of Te-Zhong Record Series

邵僅

Chin Shao

國立政治大學圖書資訊與檔案學研究所研究生
Graduate Student
Graduate Institute of Library, Information and Archival Studies
National Chengchi University

林巧敏*

Chiao-Min Lin

國立政治大學圖書資訊與檔案學研究所教授
Professor
Graduate Institute of Library, Information and Archival Studies
National Chengchi University

【摘要 Abstract】

本研究以中國國民黨「特種檔案」為對象，進行數位人文研究，採用詞頻統計、實體辨識、社會網絡分析與地理資訊系統等功能，探討檔案中人物、組織、事件之核心詞彙分布及共現關係，解析檔案隱藏的內容重點與人物關係網絡，亦對應地理資訊系統呈現各時期重要地點的分布。本研究透過量化數據與視覺化技術，展現特種檔案中無法以人工直觀的重要資訊，並結合質性

*通訊作者：林巧敏 cmlin@nccu.edu.tw
投稿日期：2025年2月4日；接受日期：2025年7月9日

說明，為檔案內容的分析與詮釋提供新的方法與應用模式，期許能為檔案學與數位史學提供不同的研究取向。

This study focuses on the Kuomintang's "Te-Zhong Archives," employing digital humanities methods such as word frequency analysis, named entity recognition, social network analysis, and geographic information systems. The research explores the distribution of core vocabulary related to individuals, organizations, and events, uncovering key themes and relational networks within the archives. Geographic information systems are utilized to visualize the spatial distribution of significant locations across different periods. Through quantitative data analysis and visualization techniques, this study reveals critical insights embedded in the archives that are difficult to discern through traditional manual examination. By integrating qualitative explanations, it provides new approaches and applications for the analysis and interpretation of archival content, aiming to offer alternative research approaches for archival studies and digital historiography.

【關鍵詞 Keywords】

檔案文本分析；數位人文；詞彙共現分析；特種檔案；文本探勘
Archives Text Analysis; Digital Humanities; Co-Occurrence Analysis; Te-Zhong Records; Text Mining

壹、前言

在當前數位化與資訊爆炸的時代，數量龐大的紙本檔案經由數位掃描轉換，成為能以電腦處理的各種數位資訊，且其數量在持續增加中（蔡永橙、黃國倫、邱志義，2007），促使以數位工具進行人文研究的「數位人文學」誕生。歷史、文學、政治、檔案等人文領域研究者，可跳脫傳統考證方式限制，透過計算機技術輔助，進行以往難以達成的研究，提升龐大資料量的檢索效率及跨領域應用；亦能藉助如主題模型、社會網絡分析以及地理資訊系統（geospatial information systems, GIS）等工具，從檔案中發掘隱含主題、還原歷史人物的社會互動，或再現事件的時空脈絡。此類研究，不僅有助快速抓取宏觀圖景，也能深入微觀細節，對文獻資料進行多層次解讀，更能以可視化等嶄新、直觀理解方式，展現檔案中的歷史脈絡與人文記憶。

對於檔案文本的內容分析，存在兩種研究取向，一是藉由諮詢內容專家對於特定研究問題進行議題探究與分析，二是對於文本內容進行歷史研究關注的人、事、時、地等發展脈絡，進行鳥瞰式整體分析。若是屬於後者的研究，係著眼於研究方法論的革新，希望透過數位人文研究的力量，協助研究者重新審視檔案資料的價值，揭示其隱藏的意涵，並賦予靜態資料嶄新的生命力，有望揭示學術研究的新議題，也能為當代社會在歷史反思與文化保存上注入新的可能性。

有鑑於近年來社會大眾及輿論關注黨務相關檔案之應用，其中「特種檔案」深受矚目。此檔案文件系列為民國17年至民國38年（1928～1949）中國國民黨中央秘書處及地方、海外黨務之歷史檔案。中國國民黨發源自孫中山先生於1894年創立之興中會，迄今已逾125年，為近代中國歷史最悠久、執政時間最長的政黨。記錄中國國民黨黨務發展過程之檔案資料也因其歷史價值，成為研究中華民國史與中國近代史重要之史料（林巧敏，2022）。

為能揭示此文件系列檔案內容，本研究試圖借助數位工具，以「文字探勘」方式進行檔案內容分析，探討國立政治大學所藏之國民黨「特種檔案」內容重要人物、組織、事件等核心詞彙分布、共現關係及地理資訊，盼能提供各界瞭解此文件系列內容梗概，並提供不同的研究取徑，以量化分析呈現檔案內容之重要概念、人物社會網絡與地理資訊分布。由於當前檔案學科與歷史學科中，針對檔案文本內容探勘與國民黨特種檔案之研究仍屬少見，因此，本研究亦盼望探索出對於此類研究相

對友善的資料分析流程與工具應用、分析結果呈現方式，對於未來檔案文本探勘研究與數位史學研究有所助益。本研究目的在於：

一、探討「特種檔案」文本內容中重要人物、組織等核心詞彙之分布與意義。

二、分析「特種檔案」文本內容中重要人物、組織、事件等詞彙間的共現關係。

三、結合地理資訊系統探討「特種檔案」文本內容之地理分布與當代地緣政治之關聯。

貳、文獻探討

一、文字探勘應用於文本分析之研究

進入21世紀後，電腦科技成為日常，文字探勘技術更於諸多領域受到重視，應用廣泛。在商業方面，可透過文字情感探勘評估顧客滿意度，如對新上映電影的好評程度（Pang & Lee, 2008）；公共方面，可透過文字探勘監控新聞、部落格、推特等網頁文本，確保國家安全（Zanasi, 2007）；生物醫學方面，可透過探勘大量病歷資料中之相似案例，促進臨床研究並提供更精準的醫療（Van Le, Montgomery, Kirkby & Scanlan, 2018）。

Asllani、Halstead與Taylor（2020）自美國基督教組織名錄*Shepherd's Guide*中選出明確為基督教信仰之60所企業，包括30所醫療服務業與30所金融服務業，加上2017年《富比士》（*Forbes*）雜誌選出之96所最佳創業組織（top entrepreneurial organizations），共156所機構作為標的。先以網路爬蟲獲取企業官方網頁文本，再以文字探勘中的詞彙及詞彙層級文本探勘法（word-level text mining），檢出其網站中的宗教語言，據此設定檢索關鍵字，再以LIWC（Linguistic Inquiry and Word Count）工具，對具有心理、情感意義的單字進行分析統計，最後於LIWC中創建基督教宗教語言類別字典（Christian religious language category），包括信仰（faith）、上帝（god）、救世主（savior）、聖經（bible）、教堂（church）等高頻詞彙，釐清宗教友好機構是否更常使用宗教用語，以探討此類機構網站中宗教用語之使用趨勢、使用程度與情感強度。

金觀濤、劉青峰與邱偉雲（2019）自收錄近代中國重要思想文獻一億餘言的「中國近現代思想史專業數據庫（1830 ~ 1890）」中，對於「主權」一詞進行搜尋統計，得出該詞彙使用高峰期（1860 ~ 1898）

與使用比例變化圖，並依高峰期撰寫者身分屬性（外國人、中國官員、中國士人等）和語料性質（官方文書、士人論著、報刊等），分析該詞彙之語境變化，接續文字探勘，得出數據庫中其他與「主權」相關之要素（國家、國民、民族主義）相關之語料，再以數據驅動法（data-driving）找出與主權要素相關的共現詞彙，並引入統計學中的累積和計算法（cumulative sum）最終將不同時間點上的共現詞彙軌跡製成圖表，釐清近現代中國從傳統社會的「帝國」，轉變為先進社會的「國家」之明確時間點。

上述兩項研究雖領域迥異，卻同樣依賴文字探勘中的詞頻統計與語境關聯分析。這種方法論的共通性，可深刻體現文字探勘技術，如何透過量化語言特徵，檢驗不同文本背後的社會文化脈絡。而隨著自然語言處理技術的成熟，後續研究將情感分析從單純的詞典比對，提升至模型訓練層次，顯示文字探勘應用從靜態統計邁向動態學習。

湯雅媜（2020）以過往兩岸關係為藍本，分別制定中共對臺之恫嚇威脅指標，包括武統威脅、外交威脅、反臺獨威脅、仇臺威脅、中美爭霸威脅，以及善意友好指標，並尋找與二指標概念接近之文章，整理為代表性文本。再以網路檢索與爬蟲技術（web crawler）爬取2018年1月至2020年9月間，共七個中共官媒與官網新聞，從中擷取有關臺灣之訊息，進行斷詞、特徵萃取（feature selection）、文字袋（bag of words model）、字典及語料庫（corpus），力求精簡文字，並導入TF–IDF（term frequency–inverse document frequency）模型進行權重統計，最後利用奇異值分解（singular value decomposition）進行降維（dimension reduction），使數據純化，依重要性排列，捨棄多餘或不重要的雜訊，保留下來的訊息則建立潛在語義索引（latent semantic analysis），透過Similarities套件計算兩向量間的相似程度，尋找與前述二類代表性文本相似性高的文章，以文本產生時間點之兩岸情勢、舉措進行可視化與對比分析，藉以觀測中共於各時間段之對臺態度和威脅趨勢變化。

張奕萱與林巧敏（2022）針對網路論壇「風聞社區」自2019年1月至2020年12月間之檔案主題貼文進行研究，先以網路檢索與爬蟲擷取文本，再以NLPIR進行斷詞、過濾無意義詞彙與詞頻統計等預處理，利用Excel函數功能，統計貼文中詞彙之詞性與詞頻，得出檔案議題貼文之核心主題與內容。接著，運用Access軟體，將斷詞後的所有詞彙與中文情感詞彙本體庫進行比對，提取出具有情感意義的詞彙與各詞彙的對應情感分數，最終計算出單篇貼文的情感分數總和。最後，根據情感分類結果建立訓練集與測試集，利用WEKA（Waikato Environment for

Knowledge Analysis）進行情感分析，以釐清檔案主題貼文中呈現的正負向情感、討論議題特點與關聯性。

二、檔案史料之數位人文研究

歐美國家數位人文學發展較早，以檔案為對象之研究案例亦相對多元，以數位化進行史料保存、資訊視覺化最為多見，藉由建置計畫專屬網頁並詳細提供參考資訊，對於人文研究貢獻匪淺，亦有運用數位科技進行毀損文本之復原辨識。

由維吉尼亞大學、紐約州立大學賓漢姆頓分校（State University of New York at Binghamton）與國家人文基金會（National Endowment for the Humanities）共同進行的研究，是對於1692 ~ 1693年間發生之賽勒姆審巫案（The Salem Witch Trials）相關史料進行內容分析，並為受害者平反。此案為美國史上最嚴重的政治迫害事件，並間接導致麻薩諸塞（Massachusetts）清教徒政府的垮臺（Ray, 2022）。該計畫自波士頓地區博物館與檔案館徵集約155位受害者檔案，包括庭審紀錄、起訴書與判決書等，進行數位化、人名權威控制，建立開放檢索之數位資料庫。此外，亦為18世紀賽勒姆地區之地形、道路、房屋分布等環境建置GIS，輔以17世紀獵巫古籍之數位文本，以幫助使用者理解該事件之時空、文化與地理背景。

歷史檔案的數位化不僅重現單一事件，更能系統性建構長期社會變遷紀錄。「英國和澳洲追蹤倫敦囚犯」（Tracing London Convicts in Britain & Australia）的研究計畫是由利物浦大學（University of Liverpool）、牛津大學（University of Oxford）、謝菲爾德大學（University of Sheffield）等校組成研究團隊，接受藝術與人文研究委員會（Arts and Humanities Research Council）贊助，旨在為1780 ~ 1925年間的英國籍罪犯建立資料庫，以利歷史學、犯罪學、社會學等相關人文學科之研究。該計畫先自英國國家檔案館（The National Archives United Kingdom）、新南威爾斯州檔案館（New South Wales State Archives）以及家譜網站Ancestry、Findmypast等機構取得與英國籍罪犯相關檔案，包括姓名、圖像、生卒年、庭審紀錄甚至經警政司法單位登記之身體特徵，再以光學字元辨識（optical character recognition）與人工辨識繕打兩種方式，針對印刷檔案與手寫檔案進行文本數位化，將文字精確度控制於98% ~ 99%，並由謝菲爾德大學數位人文研究所開發的檢索系統，提供包括以犯行分類之罪犯傳記、以地區分類之犯罪防治與司法史、相關議題研究指南等查詢。該網頁亦提供常見問題之可視化圖庫，例如能夠以視覺化圖示回答「1763

年至1868年間判處死刑之罪犯執行結果」（Punishment outcomes for convicts sentenced to death, 1763-1868）、「最常見的30個刺青部位」（Top 30 body parts）等問題（Digital Panopticon, n.d.）。

除史料與歷史事實的保存，數位人文方法亦可應用於歷史檔案修復，由倫敦大都會檔案館（London Metropolitan Archives, LMA）與倫敦大學學院數位人文中心（University College London Centre for Digital Humanities, UCLCDH）等機構合作進行的修復計畫，即「偉大的羊皮紙書」（The Great Parchment Book）。該書成於1639年，此為登記愛爾蘭阿爾斯特省（Ulster）一帶土地財產的古老文獻，包括農地莊園之贈與、租賃、使用細節、協議合約等，可幫助英國皇室釐清當地不動產歸屬。1786年，該書因火災嚴重受損，因其對於愛爾蘭歷史的重要性，故其殘存部分仍受到妥善保存，但至2012年著手修復時僅餘165頁，且順序混雜，羊皮紙也因受熱嚴重皺縮。修復過程中經LMA反覆試驗，先以最妥善的加濕、清理和張力烘乾（tension drying）等流程撫平皺縮部分，讓紙張盡可能展開並進行初步修復，再由UCLCDH等學術機構支援數位修復方法與經驗合作進行，讓帶有皺褶、彎曲之立體頁面模型可呈局部攤平，以2D方式閱讀，最後輔以半自動方法進行筆跡線條檢測，生成完整2D頁面。該計畫成功復原現存頁面中絕大多數內容，成為愛爾蘭史學研究重要資產，並於2016年列入聯合國教科文組織「世界記憶計畫」（Pal et al., 2017）。

18世紀間，阿爾卑斯山以北的歐洲富人興起赴義旅行風潮，以英國、愛爾蘭人最多，藉此學習文化、拓展交流，並與當地菁英名流來往，對歐洲的藝文與政治發展產生深遠影響，並留下大量日記、書信、藝術品與官方紀錄。史丹佛大學的「壯遊計畫」（The Grand Tour Project）參考John Ingamells於1997年出版的《英國與愛爾蘭旅人義大利詞典（1701-1800）》（*Dictionary of British and Irish Travelers to Italy 1701-1800*），為5,213名旅者的行程與生平建立資料庫、繪製時間軸，並以GIS工具進行視覺化，開發出Grand Tour Explorer工具，供研究者以更靈活方式，進行多維度檢索與視覺化分析，例如可查詢特定時間與地點的旅客，或觀察整體旅遊趨勢。由於該時期史料分散且龐大，計畫團隊亦在開發之餘，持續徵求將未收錄之相關史料納入資料庫，以利研究（Stanford University, n.d.）。

臺灣地區結合檔案之數位人文研究案例，較常以公文書或歷史文獻之內容分析為主，近年亦開始以數位方法修復古代文書並架設資料庫，結合資訊科技、史學、文字學等多元學科，為檔案研究開啟新的篇章。

郭乃華（2014）以《中國古代法典及其事例之自動化整合——以乾隆朝〈大清會典〉為例》，進行清代檔案之自動化整合研究，以乾隆朝（1736 ~ 1795）官方修編之典章制度文獻《欽定大清會典》及《欽定大清會典則例》文淵閣四庫全書本為對象架設檢索系統。該研究首先針對二書之數位化檔案進行分類擷取、條目切割命名與時間主題資訊擷取，按原書劃分之部門、職掌分析內容，得出3,695筆《會典》條目與21,250筆《則例》條目，並依書中相似內容之編排規範，以「分類」、「主題」與「條目」整理為三大層面，比較《會典》與《則例》之整體異同。由比對結果可知，《會典》與《則例》雖都為記載清代體制典章，但彼此間並無完整對應，因此再以條目為單位，採最長共同字串（longest common subsequence）將《會典》條目與《則例》條目進行對照，得出13,908組對應關係，於系統網頁中予以連結參照工具呈現。研究建置的系統除了提供全文檢索，亦可於檢索結果中進行部門篩選、條目統計、相關條目檢視等功能，研究者可在短時間內綜覽所需之會典資料，可降低研究過程中反覆翻閱比對兩書的工作量。

而運用數位檔案系統進行之研究，有王麗蕉（2018）以總督府職員錄系統觀察臺灣女力公職之發展，該研究結合檔案數位化風潮，以中央研究院臺灣史研究所檔案館建置之「臺灣總督府職員錄系統」為對象進行資料分析，該系統收錄日本殖民統治時期由臺灣總督府編印發行之《臺灣總督府職員錄》，檔案產生期間為1898 ~ 1944年共95萬筆。該研究先經比對與補齊本籍資料缺漏者，排除補齊率較低與因特殊情況記載不全之年分，篩選出本籍為臺灣者占總名錄之百分比，再以日治時期臺灣女性之姓氏下加註「氏」字，檢索臺籍職員中女性歷年人數概況、出身地（各州廳）與總人數，以可視化圖表繪製臺灣女性與臺籍人士公職就業比例變化、公職女力來源的州廳分布與領域職業類別分析，再依從事比例最高之教育工作、通信工作等條件進一步分析，得出日治時期日、臺籍初等學校男女教師比例與女性通信人員人數之變化圖，可觀察日治時期臺灣女性於公務機關的職場發展趨勢，代表檔案學界可將數位檔案資料庫應用於歷史研究的可能性。

以數位人文進行古代檔案復原與釋讀者，當屬中央研究院歷史語言研究所（以下簡稱史語所）「數位人文學與簡牘研究：文字釋讀與簡冊復原」計畫。2019 ~ 2020年，史語所、數位文化中心與奈良文化財研究所合作進行「簡牘字典」開發，為史語所藏13,000餘枚漢簡開發研究工具「影像字元定位及標註模型」，並依此為基礎，自2021年起以數位人

文工具輔助文字釋讀與簡冊復原。「簡牘字典」計畫中，無法辨識之字形多是以前後文脈絡假設、字典查詢、回查出處、拆解查詢等方式進行釋讀，該研究經數位文化中心支援，以文字模型得出約300個字形釋讀假設，並透過簡牘字典驗證比對，得出與學者過往結論相符之成果，經驗證數位化釋讀功能具有學術上的可信度後，依此進行簡牘復原。簡牘復原過程先以中研院數位人文研究平臺進行相似段落比對，篩選出多字重複的簡牘，再依簡牘字典著錄內容對比遺址、形式、筆跡之特徵相符程度，得出簡牘間的關聯性，並尋找其他具關聯性的簡牘。該研究仍在持續進行中，且將簡牘字典收錄的資料同步公開於臺日合作建置之「歷史文字資料庫統合檢索系統」，免費提供簡繁體中文與日文漢字查詢功能與高解析度文字圖像，持續造福漢字文化之古文字研究族群（開放博物館，n.d.）。

數位人文不僅處理難以閱讀的文本，更能解析未被充分解讀的史料價值。與民國史相關之研究中，葉韋君（2019）進行〈後五四時期的知識婦女：《婦女共鳴》的社會網絡（1929-1944）〉研究，採用《婦女共鳴》刊物，研究1922 ~ 1944年知識婦女的社會網絡，該刊物由國民黨婦女創辦，獲政府特許及政黨補助，歷經訓政到抗戰時期，是當時發行時間最長的婦女刊物之一，也是研究該時期婦女的重要文本，但在婦女刊物史上，卻被低估為「國民黨婦女的喉舌」，甚少獲得深入探討。該研究利用中央研究院近代史研究所設置的「近代婦女期刊資料庫」及「婦女期刊作者研究平臺」蒐集《婦女共鳴》作者群的傳記資料，並進行該期刊內外部社會網絡分析。外部分析包括參與至少三個主要婦女組織，如國民黨婦女會、南京市婦女會、南京婦女救國會之人物與關聯，以觀察作者群關係活動與傾向；內部分析則以歷任編輯在任時期進行分期，並整理出投稿五篇以上的主要作者群，看出各期內部人物變化。最後，根據擷取出的關係數據，繪製出社會網絡關係圖，得出人物間的關係鏈結，進一步探討五四運動後，《婦女共鳴》的作者們如何在黨政結構下，運用資源生產論述，發展後五四知識女性的社會網絡。

運用民國檔案資料進行分析者，另有鄧志松、郭迺鋒與林崑峯（2011）〈國父行腳之空間分佈與社會網絡分析之應用〉，以記錄國父孫中山一生行跡的「國父行腳資料庫」，搭配Super GIS地理資訊系統與社會網絡分析軟體NodeXL，計算資料庫中資訊的對內程度中心性（in-degree centrality）及中介程度中心性（betweenness centrality），除剖析國父革命時期於各國奔走的路線與停駐點、繪製國父在各國之間的移動

絡線，亦提出清末民初建國組織與革命人士之關聯網絡，並以視覺化方式呈現。從婦女期刊到革命領袖行蹤，社會網絡分析不僅驗證將社會關係數據化的可行性，更展現數位工具如何將人物互動之抽象概念，轉化為具時空座標的視覺論證。

上述案例展現了文字探勘與數位人文技術進行檔案分析的應用潛力，過往有不少應用文字探勘、視覺化與GIS等技術的實例，但以核心詞彙分布、共現網絡、地理資訊關聯等多維角度分析政黨檔案之研究仍屬罕見。本研究基於前人研究基礎，探討數位人文在近現代中國政治史與黨務發展史研究的相關實例，可提供檔案內容分析與地緣政治研究方法論上的參考，有助於推展數位史學與檔案學交會的新研究方向。

參、研究設計與實施

一、研究對象

國立政治大學圖書館所藏之國民黨特種檔案，為中國國民黨中央秘書處及地方、海外黨務之歷史檔案，該批檔案涵蓋時間始自北伐結束（1928年），終於中國國民黨退守臺灣（1949年）。內容包括東北黨務幹部會議卷、中央及各省黨務工作計畫、外交部特種檔案、有關各地奸偽卷、戰時中外關係、中央秘書處檔案等33類，11,922件，共計約120萬字，目錄則由國家歷史博物館著錄、國立臺灣大學圖書館整理製作、政治大學圖書館增補。目錄資料欄位包含：編號、題名、產生者、資料類型、出版日期、西元出版日期、內容描述、產生地、出版者、相關人名資訊、相關地名資訊與製作單位。

二、研究架構

本研究設計架構圖示（如圖1）及說明如下：

第一部分文本預處理中，將特種檔案之Excel目錄剔除無意義欄位，包括全為「原件」之「資料類型」、全為空白之「出版者」以及全為「國立臺灣大學圖書館」之「製作單位」，將剩餘9個欄位進行後續分析，並轉存為TXT以方便作業，再使用CKIP tagger進行斷詞、詞性標註與實體辨識，並將斷詞結果的名詞進行人工確認及合併統計。

第二部分內容分析，根據研究目的使用不同數位工具進行內容分析，各工具之功能及其研究作用如表1。

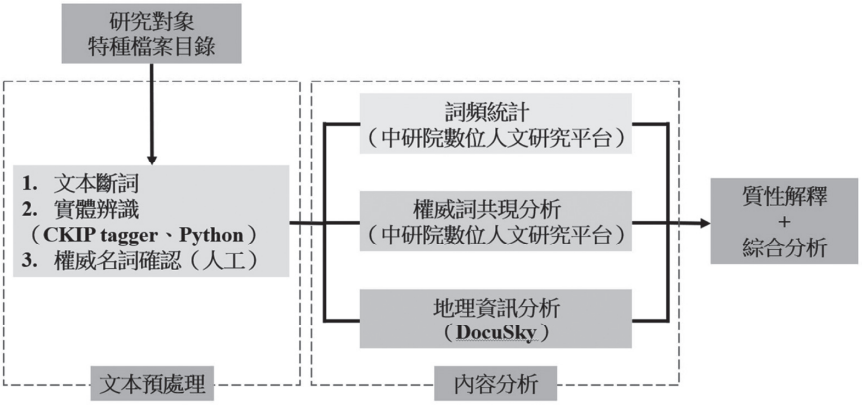

圖1　研究架構圖

表1

分析內容使用的工具說明

分析內容	主要工具	輔助工具
斷詞、詞性標註、實體辨識、詞頻統計	CKIP	Python
權威詞共現分析	中央研究院	
數位人文研究平臺		
地理資訊分析	DocuSky	Map Location、Google Earth

（一）詞頻統計

將人工整理之不同類型權威詞匯入中研院數位人文研究平臺，使用權威詞統計功能進行統計，釐清檔案中出現之人物、組織詞彙頻率與比率。

（二）權威詞共現分析

將人工整理之不同類型權威詞匯入中研院數位人文研究平臺，使用權威詞共現分析與統計功能進行各類型詞表之共現分析，並繪製社會網絡圖，可觀察出特種檔案中常見人名、組織、事件等核心詞彙之間的關係。

（三）地理資訊分析

先以CKIP進行特種檔案分期內容之地名辨識，應用線上工具Map Location取得地名經緯度後，再使用DocuSky平臺建置之DocuGIS功能，將經緯度進行世界地圖可視化，比較不同時期特種檔案中地理資訊變化。

三、研究工具

本研究自文獻探索與網路評比意見中，挑選出數項具備本研究所需功能、使用便利之免費數位工具，或參考開發者測試，或經試用後以表2比較各功能之性能、特點，進行最終選擇。文本預處理與詞頻統計部分參考中研院CKIP小組測試結果，自CKIP Tagger、CKIPWS與Jieba-zh_TW（Jieba繁體中文版）中，選用精確度最高（達到97%）的CKIP Tagger。該工具奠基於語言模型、統計方法與深度學習技術，是第一個具備未知詞偵測與句法詞類預測能力的中文分詞系統，擁有斷詞、詞頻統計、61種POS詞性、11類NER一般領域專有名詞及7類NER數量詞、自定義詞典等功能，以準確率與功能性聞名（中央研究院，2023）。本研究於預處理步驟中採用「斷詞」與「實體辨識」，標記出檔案中人物、地名、組織與事件，提供後續分析基礎。

表2

斷詞工具比較（ASBC 4.0測試集）

工具	CKIP Tagger	CKIPWS (classic)	Jieba-zh_TW
(WS) prec	97.49%	95.85%	90.51%
(WS) rec	97.17%	95.96%	89.10%
(WS)f1	97.33%	95.91%	89.80%
(POS) arc	94.59%	90.62%	—

資料來源：中央研究院（2023）。結合斷詞、詞性標記、實體辨識的中文處理套件（CkipTagger）。

詞頻統計與權威詞共現部分依研究者試用過程，自微詞雲、CORPRO與中研院數位人文平臺中，選擇繁體中文精確度較高、可自訂權威詞並直接繪製共現詞網絡圖的中研院數位人文平臺。該平臺以「一站式服務」形式，讓研究者同時取得歷史文獻、數位工具與合作夥伴，並持續更新功能、匯入文本。截至2021年，已匯集逾90億字中文文本及數百類人名、地名開放權威檔。本研究於詞頻統計部分採用「權威詞彙統計」功能，進行檔案中人物、組織權威詞之詞頻統計；再於權威詞共現部分採用「權威詞共現統計與視覺化」功能，進行人物、組織、事件之共現狀態視覺化。

地理資訊分析部分採用臺灣大學數位人文中心建置之DocuSky數位人文學術研究平臺，該平臺為全球第一個專為人文歷史學者設計、可建

立個人化資料庫的研究工具，提供文本標記、格式轉換、詞頻分析、文本對讀、風格分析、分類樹等基本與進階功能，使用者可以自行建置與管理資料庫，靈活處理不同研究主題與需求。本研究將運用其中的「DocuGIS」功能，搭配人工整理的地名與經緯度資料，將不同時期的地名分布描繪於地圖上，以呈現空間上的變化。

四、研究程序與資料分析

歸納出適用於特種檔案之探勘方法與數位工具後，首先將xlsx格式之特種檔案目錄文本整理過濾，成為適用於數位分析的Excel與TXT文本，並利用CKIP Tagger進行內容斷詞與實體辨識，獲取檔案中人物、組織與事件之詞彙，經人工檢視確認後，設定為後續分析之人物、組織、事件三大類型權威詞，分別儲存為CSV。最後，將完成斷詞之內容轉存為不同分期與全文之TXT與CSV，分別去除非中文字符（如括弧、破折號、問號等出現於檔案附件或說明月日不明情況時出現），以降低後續分析可能產生之誤差。

因CKIP會將所有可能詞彙全數斷出，且檔案中人名內容十分繁雜，所以為確保分析準確，在建立權威詞表過程中，不建議直接採用實體辨識結果，務經人工整理確認。有疑慮者也需要回溯原文，再決定是否修正並決定如何修正。如實體辨識結果中同時存在的「陳素」與「陳素電／函」、「十八集團」與「十八集團軍／兵」，經確認原文使用次數、方式無疑義後，保留「陳素」與「十八集團」，以字詞相同、便於分析工具辨識者為合併處理的原則，減少詞頻統計之誤差率。

接著進入文本探勘分析流程，採用中央研究院數位文化中心推出之數位人文研究平臺，以預處理過程中建立之人物、組織與事件權威詞進行共現分析，並繪製社會網絡圖，釐清檔案中人與事之關聯。處理過程是將資料以正確分段形式匯入中研院數位人文研究平臺，需先將檔案全文轉換為以空行分段之TXT檔，每筆檔案需間隔一個空行，以利平臺正確判讀檔案筆數，並以TXT形式匯入文本。接著，以人工整理CKIP實體辨識後，標記為PERSON（人名）與EVENT（事件）之內容，去除重複詞並建立為每詞一行之權威詞TXT檔，分為「人名」、「組織」、「事件」三大類。最後，將整理完畢的權威詞匯入中研院數位人文研究平臺，採權威詞共現統計與視覺化功能，以「人名對人名」、「人名對組織」、「人名對事件」、「組織對組織」、「組織對事件」、「事件對事件」六種組合，得出人物、組織與事件在特種檔案中的共現情形，但本研究闡述其中比較有意義之分析結果，解讀研究結果呈現的意義。

因檔案內容中不乏提及地理位置，為進行地理資訊分析，先將文本依1928 ~ 1949年間現代史大事記，以重大政治事件或戰爭為主分為四期，各期內容再經CKIP實體辨識，挑選標記為LOC（地理區）與GPE（行政區）之名詞，再以ChatGPT輔以Google Earth進行地名經緯度確認。最後，將各期辨識結果分別整理為「編號」（id）、「地名」（place）、「經度」（x）、「緯度」（y）四欄位共四組CSV表格，分別匯入數位人文研究平臺DocuSky中之地理資訊工具DocuGIS，取得特種檔案中各時期地理位置分布。

最後，結合相關史學文獻之分析，以質性方式對結果進行解釋。最後，回顧研究階段曾遭遇之瓶頸與限制，提出對於未來研究的建議。

肆、研究結果分析

一、核心詞彙分析

（一）人物高頻詞彙數量分布

本研究將特種檔案後設資料案名與內容描述欄位，以CKIP Tagger進行斷詞與實體辨識後，共得93,581條具實體意義詞彙，分為18類。運用中央研究院數位人文研究平臺「權威詞詞頻統計」功能，將實體辨識為人名（PEOPLE）與組織（ORG）之詞頻加以整併並統計後，獲得高頻人物100條（圖2）、組織30條（圖3），惟限於篇幅，本研究對於此兩類

圖2　前100高頻人名詞彙文字雲

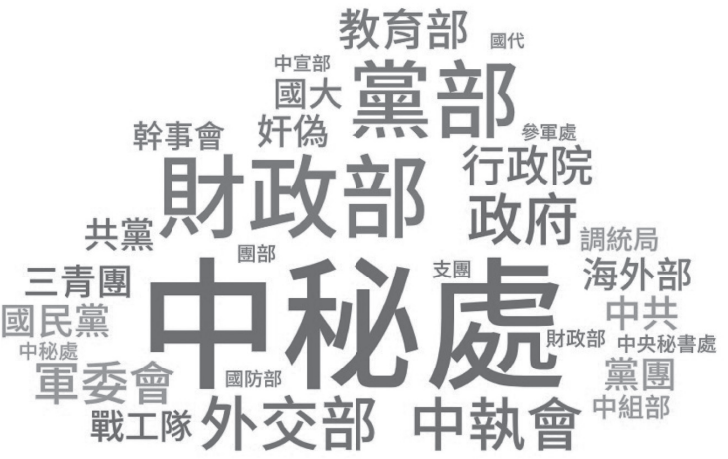

圖3　前30高頻組織詞彙文字雲

中各前10條詞彙進行說明，分析核心人物、組織之分布與職務，並解釋檔案內容呈現之意義。本分析之詞頻多寡，僅代表人名與組織名在文本的出現頻率，並非等同活躍程度，也不試圖以頻率高低詮釋該人物、組織之重要性，或對史事進行論斷。此項分析在於以量化觀察文本隱含的資訊，作為史學者進一步探究之參考。

因特種檔案時間範圍為1928 ~ 1949年，本研究觀察前10高頻人物（表3）中，出現頻率最高的人名是吳鐵城，其自1940年起輾轉歷任國民黨中央海外部部長、國民外交協會理事長、中央秘書長，幾乎所有涉及外交關係、黨內會議之檔案都經其手，直到1949年國民政府遷臺轉任總統府資政為止（劉壽林、萬仁元、王玉文、孔慶泰，1995）。而作為駐東南亞重要人物、曾任越南總領事的邢森洲，直接接受吳鐵城指揮，在1943 ~ 1946年間電訊往來甚密，多以越南、泰國國內消息為主，可呈現此期與南向國家往來密切（吳淑鳳、許順昇，2011）。國民黨最核心人物蔣中正詞頻位列第三，最早出現在1929年第三次全國代表大會期間訂立之建設、外交、財政、國防等南京政府下轄委員會名單，其後則較平均分布在1939 ~ 1947年間，1945年抗戰勝利時達最高峰，詞頻僅次於吳鐵城、邢森洲。1943年起接任港澳總支部主任委員的陳素，於抗戰勝利前回報大量當地戰況，後留任廣東（李盈慧，2013）。張壽賢相關檔案始於1941年，其擔任國民大會籌備委員會幹事至1946年轉任秘書，並於1948年擔任戡亂建國動員委員會秘書處處長，所涉內容以會內事務、海外黨務為主。

表 3

人物前10高頻詞表

排序	詞彙	詞頻
1	吳鐵城	9,356
2	邢森洲	1,657
3	蔣中正	829
4	陳素	716
5	張壽賢	669
6	徐恩曾	580
7	朱家驊	450
8	何應欽	370
9	陳立夫	360
10	王天雄	280

素有中統「特工老闆」之稱、曾任中統局長的徐恩曾，在1941年兼任交通部次長後較常出現在檔案中，1945年因涉嫌中印緬邊境交通線走私被蔣中正免職（李福鐘，2021）。朱家驊相關檔案始自1937年起出任浙江省政府主席，隔年改任中國國民黨中央執行委員會秘書長兼中央調查統計局局長，1944年後再任教育部長回歸學術體系（徐友春，2007）。何應欽相關檔案於1941 ~ 1945年抗戰期間最為密集，遍歷國民政府軍事委員會委員、參謀總長、同盟國中國戰區陸軍總司令等職，所在檔案內容多為各戰區之情況與黨務。陳立夫曾於1935年任軍事委員會調查統計局長，檔案則始於1938年任教育部長期間，1941 ~ 1948年間遍歷國大籌備委員、中央政校代校長、國大代表、立法院副院長等職，留下許多學界、學潮、抗敵有關的檔案內容，1949年後因黨內派系糾葛，日漸退出政治舞臺。至於在1942 ~ 1943年前後高頻出現的王天雄，因在該時期派赴加爾各答，成立中國留印海員戰時工作隊，並任總隊長，負責留印船員的消防、運輸、守衛等訓練（陳宗嶽，2012），歸國後仍持續擔任國防工作。

（二）組織高頻詞彙數量分布

本研究觀察前10高頻組織統計（表4），不乏國民黨的內部組織，例如中秘處、黨部（含中央及各地黨部）、中執會等屬於國民黨之組織；也包括政府組織，例如外交部、軍委會、財政部、教育部、行政院等，可知當時主政的國民黨與政府組織互動頻繁。

表4

組織前10高頻詞表

排序	詞彙	詞頻
1	中秘處	3,316
2	財政部	1,217
3	黨部	1,068
4	中執會	929
5	外交部	941
6	政府	862
7	軍委會	602
8	教育部	504
9	行政院	502
10	中共	355

觀察高頻組織的屬性，發現特種檔案內容除與國民黨核心黨務密切相關，也留下大量財政、外交、軍政事務紀錄。在特種檔案原有分類中，除外交部特種檔案，也包含海外黨務、韓國、越南、馬來亞、泰國等類別，足見其與外交相關程度高，亦為此系列檔案重要的詞頻。在財政部、教育部與軍委會的高頻詞中，除了有自行產生之檔案，也存在大量相關事務之副本資料，如財政部亦包括組織間款項流動、內戰期間經濟犯罪等一切可能與財政相關事務之紀錄；教育部詞頻中除教育部產生的檔案，亦包括學生運動、校園政治宣傳等紀錄；觀察軍委會的高詞頻，則包含一切國內外的敵軍動向，此皆躋身於特種檔案前10大高頻詞彙之列。

二、核心詞彙共現分析

（一）人物與人物之網絡關係

本研究以中研院數位人文研究平臺進行檔案核心詞彙共現分析，先整理完成特種檔案文本與人物、組織、事件三類權威詞表匯入平臺後，繪製社會網絡圖（如圖4至圖7）。為便於網絡圖判讀並篩選重要資料，本研究嘗試不同關聯強度和節點數組合，以呈現有意義之核心詞彙共現關係，最終採用最大關聯強度為100以下者，關聯強度取1，100以上者則取10；總節點數大於300者約保留前10%，介於60 ~ 300間者約保留前20%，60以下者則全數保留，亦即總節點數約不超過60，此為最有利於閱讀判讀的經驗

數值。惟因工具功能限制，字面上難以共同方式呈現之同義詞彙（如「中共」與「中國共產黨」），無法進行合併計算。因此選擇保留此類詞彙原樣，避免人為判斷的誤差。此外，社會網絡分析係以數位方法、系統性方式進行現象發掘，獲取不同類型詞彙於檔案文本中之共現、互動情形。本研究結合檔案內容引用或是該時代背景脈絡的解釋，意在提供讀者初步瞭解文本內容資訊，但最終之評價與解釋仍屬於歷史研究之範疇。

圖4根據人物與人物社會網絡關係分析結果，呈現出吳鐵城、邢森洲、蔣中正、陳素、徐恩曾、張壽賢、何應欽、王天雄等人是特種檔案社會網絡中較重要的人物，且彼此間均存在密切關係。其中於東南亞擔任總領事的邢森洲與曾任海外部長、密切掌握國民黨外交事務的吳鐵城，因存在直接上下級關係，共現次數最高，其與蔣中正和張壽賢有明顯的關聯；吳鐵城則與多位人物有高頻共現，除邢森洲、蔣中正，也包括於港澳活動的陳素、陸軍將領何應欽和中統局徐恩曾等，顯見外交、軍事與情報此三領域的密切往來；蔣中正呈現的共現對象除了有主理外交、軍事的吳鐵城、何應欽、張發奎等人，也包括接任調統局的朱家驊、在華活動的韓籍革命家金九等；陳素主要與吳鐵城和徐恩曾等人存在共現；曾任內政部常務次長等職的張壽賢主要與吳鐵城、蔣中正、何

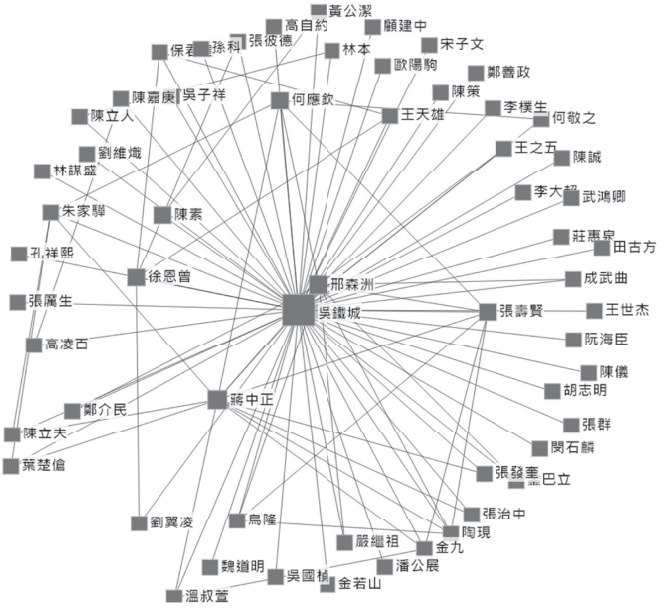

圖4　人物間社會網絡關係圖

應欽等人存在共現；至於留印海員戰工隊任總隊長的王天雄，與主理外交事務、情報的吳鐵城、徐恩曾等人也存在共現。人物共現分析結果可提供人物之間的往來關係，後續可提供歷史學者探究曾密切往來者與史事之間的連結。

（二）人物與事件之網絡關係

圖5根據人物與事件之社會網絡分析，可看出多位重要人物與重要事件的關聯性。作為主要事件的抗戰，與時任國民政府最高領導人的蔣中正、主掌黨國外交活動的吳鐵城，以及參與國共合作抗日、多次與蔣中正會面的中國共產黨領導人之一的周恩來等人，均有密切關聯；朝鮮民族革命則連結了多位相關人物，如韓國獨立運動家金若山、韓國臨時政府副主席金奎植、何應欽與張壽賢等，顯示國民黨對於朝鮮獨立運動的參與；至於以南洋閩南華僑為主體的閩僑大會，因其對外交甚至海外抗戰工作有重要影響力，故擔任南洋閩僑總會主席的新加坡企業家陳嘉庚，也與吳鐵城、蔣中正等國民政府高層有共現關係。

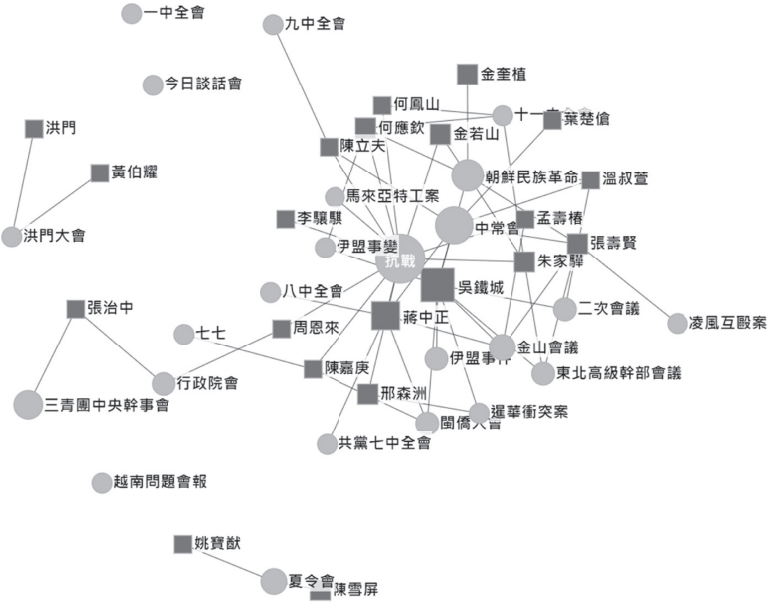

圖5　人物與事件間社會網絡關係圖

（三）組織與組織之網絡關係

圖6依據組織之間的社會網絡分析，可看出多個組織間的連結和互動頻率，如中秘處與軍委會之間的互動最為緊密，反映了兩單位間在國防、情報領域的密切合作；中秘處和外交部之間的連結次數較多，顯示其在外交政策和國際事務中亦具重要角色；財政部和行政院之間的互動次數表明兩單位作為財政管理龍頭，在政府運作中的緊密聯繫；中宣部作為負責宣傳和意識形態工作的機構，與教育部的頻繁互動可見在宣傳教育政策、思想推廣、組織宣傳活動等方面的合作；中央海外部與越盟、越南國民黨之間的互動，亦反映了中國與越南的外交和區域政策之間的關係。

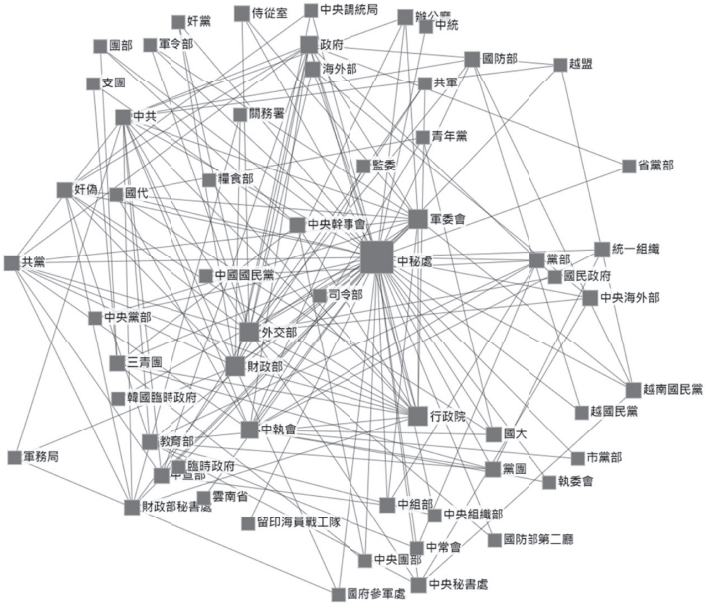

圖6　組織與組織間社會網絡關係圖

（四）組織與事件的網絡分析

圖7依據組織與事件之社會網絡分析，可得知不同政治組織和重要事件之間的關聯。如朝鮮民族革命與朝鮮民族革命黨、中執會、中秘處等組織均有關聯，可佐證該運動亦受到來自中方的推動與支持；網絡圖示呈現中日抗戰與軍委會、中秘處、共黨等組織皆有密切關係，表明這

些組織在抵禦外侮與當下政局中的重要角色；新四軍事件與引發事件的皖南新四軍、對新四軍發出移防命令的軍委會均有相關；金山會議（舊金山聯合國國際組織會議，因應檔案中部分紀錄為金山會議，為避免缺漏，以該名稱作為權威詞）則與作為國際會議代表核心的外交部，以及協調黨內事務、差旅經費的中秘處、財政部等組織存在共現關係。組織與事件分析所呈現的往來關係，對於歷史研究探查事件中相關人物的介入和協調，可提供進一步研究的起點。

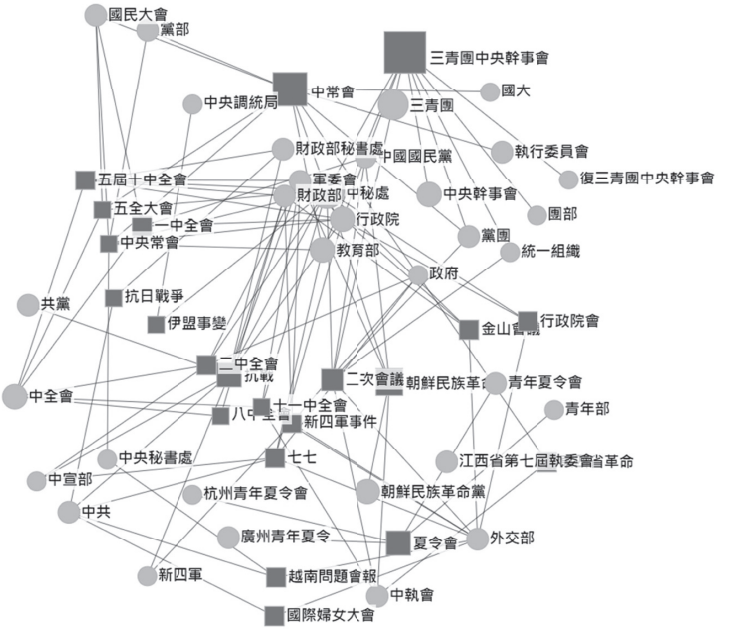

圖7　組織與事件間社會網絡關係圖

三、文本內容地理資訊分布

特種檔案經CKIP tagger進行實體辨識後，標示為LOC（地理區）的詞彙共1,113條，標示為GPE（行政區）的詞彙共23,250條，亦即特種檔案全文中共包含24,363條地名。

因數量繁多，且各年度地理資訊量差異龐大，本研究根據檔案內容特性，將文本依1928 ~ 1949年間重大事件或戰爭為時間切分考量，進行分期處理，未註明年度之53條詞目則不計入。經初步人工確認、分期並刪除各分期內重複地名後，獲得不同分期與地名數如下：

（一）1928 ~ 1939年（包含北伐結束、訓政開始、九一八事變、西安事變、第二次國共合作開始、七七事變、遷都重慶開始抗戰等）共294條。
（二）1940 ~ 1944年（包含汪精衛國民政府成立、新四軍事件、開羅會議等）共623條。
（三）1945 ~ 1947年（包含對日抗戰勝利、臺灣光復、第二次國共內戰開始、二二八事件等）共983條。
（四）1948 ~ 1949年（包含三大會戰、渡江戰役、國民政府遷臺等）共323條。

整理完畢之各期地名分別匯入DocuGIS後，即可得出地理分布圖（如圖8至圖11）。

（一）1928 ~ 1939年地理資訊分布

本時期可謂為國民政府經歷北伐結束，至全面抗戰的關鍵轉型期。北伐結束後，國民政府推動訓政體制與國家建設，並面臨內部政治挑戰與區域局勢升溫。1931年起，則陸續經歷了九一八事變、西安事變與其帶來的第二次國共合作等多番變故，至1937年七七事變後，抗日戰爭全面爆發，國民政府方遷都重慶，開啟漫長的戰時體制。此時期檔案分布雖顯零散，但對全國統一後的黨政運作與戰時動員均有所記錄，包括

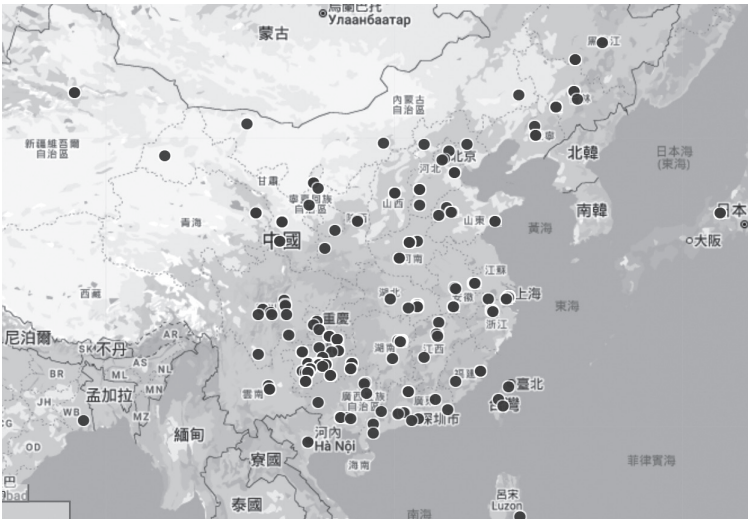

圖8　1928 ~ 1939年地理資訊

中央與地方間的治理、軍事行動、宣傳與文化策略等皆在檔案中留下足跡。

1928 ~ 1929年之檔案內容，集中於河北、南京地區，多為政府規章、條例及組織調整等，包括《中華民國無線電管理條例》、《國民政府內政法規》與《政治工作報告》等法律條例，以及北伐統一後行政綱要文本、平奉平綏鐵路建設、海外支部名冊、加緊訓政工作延緩召開國民大會與山東五三慘案事宜等。1931 ~ 1935年檔案涵蓋黨務委員會工作報告、會議紀錄與宣傳資料，集中於南京、安徽、山東一帶。南京檔案包括《剿匪報告決議案》、《中執委政治委員會組織條例》等行政成果，以及新生活運動、丘元武被刺案等地方事件。地方政府檔案則記錄各省黨務及行政工作概況。1936 ~ 1937年檔案涵蓋憲法草案審議、地方自治、教育調查等，主要集中於河北、南京、廣東、西北地區。南京檔案重點在憲法草案會議及新法規推行，並記錄七七事變後的抗日戰爭動員文書與國共合作宣言。地方檔案則多為工作報告，如察哈爾與廣東的地方事務回報。1938 ~ 1939年隨著重慶遷都，檔案紀錄更加完整，反映國民黨全國組織與管理情況，涵蓋偽邊區政府、財政、黨務、教育等領域。重慶檔案以「國民精神總動員」為主，號召全國抵禦外侮，並有大量中央與地方部門函件。其他地區，如貴州省，詳細記載黨務人事與組織運作。臺灣檔案首次出現，聚焦「爭取臺胞解放」，展現出國民政府從統一、建設到抗戰的歷史脈絡與政治行動。

（二）1940～1944年地理資訊分布

1940 ~ 1944年為抗日戰爭關鍵時期，也是國共關係、乃至中國政治格局劇變的重要階段。1940年，汪精衛主導、日本扶持的南京國民政府成立，讓國民政府領土再度出現小範圍分裂，而隨著抗戰進入關鍵階段，新四軍事件、大動作查禁左翼書籍等國共衝突，也逐步推動雙方走向決裂。與此同時，國際局勢也深刻影響中國戰局，1943年的開羅會議確立盟軍對日作戰的共同立場，並重塑中國的國際定位，對戰後中國的國內外政策產生了深遠影響。

此期間是特種檔案內容最豐富的期間之一，不同地區檔案皆可見大幅度的成長，尤其中原與廣東、東部沿海一帶，臨時首都重慶相關內容以各地學運、黨務、政壇人事為主，亦有對中共及敵偽（日本）情形匯報、《開羅會議公報》全文等。由日本帝國扶植的汪精衛國民政府在此時期成立，可見稱其為「南京偽府」之檔案，對當地政府活動進行回

報。雲南、廣西在檔案中的內容除黨務外，也涉及中共動向與越南政局，包括曾破獲廣西省奸偽省委組織、越南國民黨提交年度行動綱要草案等。此外，從1940年修正《檢查書店發售違禁出版品辦法草案》後，全國各地如廣東、江西、四川、甘肅、湖北等地，均出現查禁具左翼色彩的連鎖書局「生活書店」或「變相類似書店」之回報，為國共徹底決裂埋下伏筆。

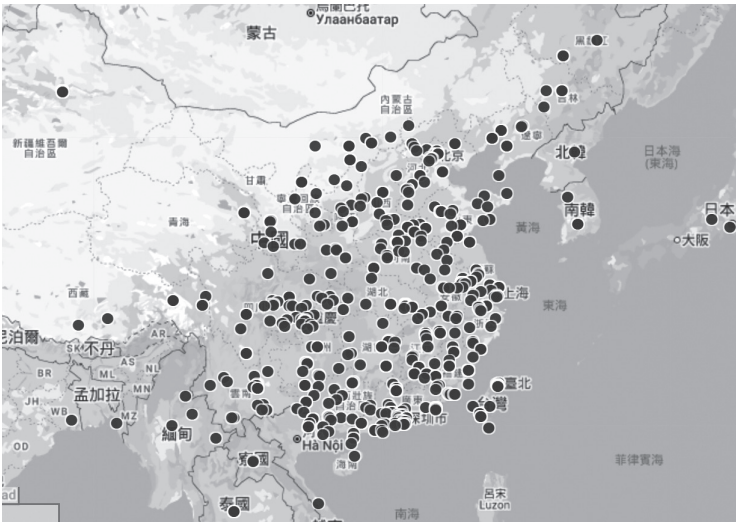

圖9　1940 ~ 1944年地理資訊

（三）1945 ~ 1947年地理資訊分布

本時期重要事件包括二戰結束、臺灣光復與國共內戰的全面爆發等。1945年，隨著對日抗戰勝利，中華民國恢復了對臺統治權，然國共雙方為接掌正式政權與受降儀式代表權等事僵持不下，終究引發全面內戰。1947年，臺灣爆發二二八事件，成為臺灣社會的深刻創傷，影響至今未平。

此時期的中國大陸與臺灣，分別處於戰後重建、政治對立與軍事對抗的艱困中，但特種檔案內容未見減少，仍涵蓋大量各地產生檔案。因本時期是接管臺灣之初，可見臺灣地名出現率有所提升，1945年檔案即出現《臺灣接管計畫綱要草案》並多次提及「協助（因二戰滯陸）臺胞返臺」等指令，1947年發生的臺灣二二八事件亦有留下〈中央當局對臺

灣二二八事件之反應與措施〉紀錄。汪精衛政權隨二戰結束後，南京、重慶、四川在檔案中高頻出現，並存在與各地黨務、選舉、工作進度匯報的相關地點。廣東、甘肅、雲貴等地相關內容多是有關黨支團事務與教育部門黨政工作，如寧夏、甘肅各地均曾電送多所院校教職員黨籍調查表等。此時期為第二次國共內戰之初，也出現許多相關檔案，如稱中共受蘇聯策動、煽動境內外人民、於多地印製仿幣破壞金融體系等情報與指摘，南京亦留下雙方試圖和談破局的紀錄。

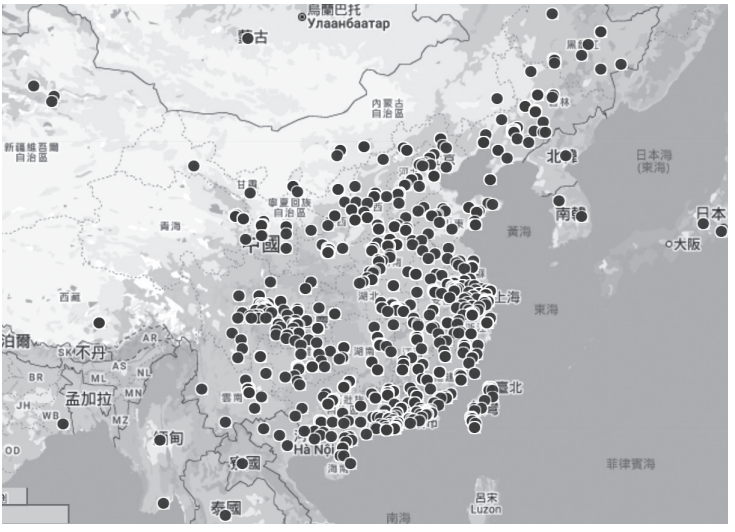

圖10　1945～1947年地理資訊

（四）1948～1949年地理資訊分布

作為國共內戰決勝階段，本時期政局對戰後迄今的東亞地緣政治秩序產生極大影響。1948年起，共軍陸續發動遼瀋、淮海、平津三大會戰，重創國軍主力；1949年渡江戰役後，國民政府節節敗退，最終遷往臺灣，加之二二八事件後臺灣的社會動盪，統治根基嚴重動搖。中共則透過土地改革與軍事勝利逐步掌控中國大陸，最終在1949年建立中華人民共和國，奠定兩岸分治局面。

1948年起，因國民政府已在內戰中節節敗退，時局動盪下可明顯看出整體檔案量減少與重心遷移，內陸地名被提及次數尤為降低，除了有河南地區以鄉鎮單位回報參議會、參政員選舉事務之外，其餘檔案出現

地點多往東南城市及沿海移動。檔案出現較集中之省市如江西、廣東與首都南京，仍有執委會會議、黨務幹部講習、青年夏令營等活動進行，其餘多為零散簡要之政務。遷臺行動逐步開始後，臺灣與福建沿海共11條地名顯示，其內容多是臺灣工會和黨團工作匯報。

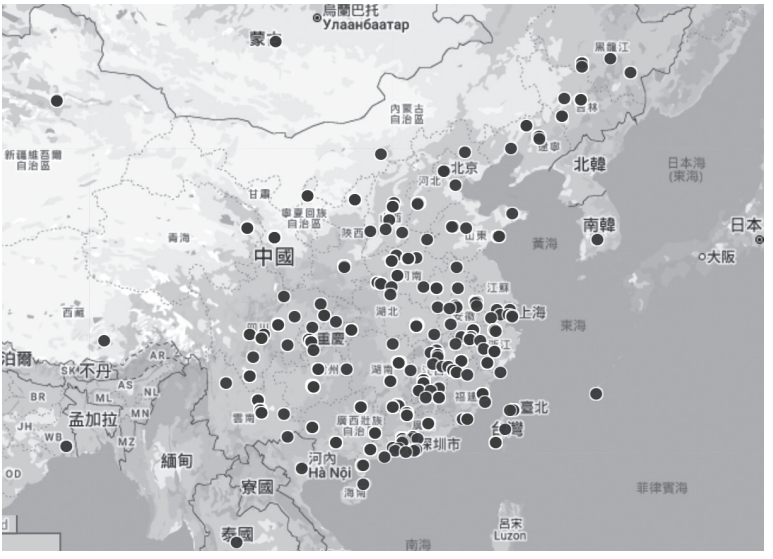

圖11　1948 ~ 1949年地理資訊

綜上所述，特種檔案中地理資訊的分布結合歷史事件與檔案原文，可反映出中華民國從1928 ~ 1949年間統治範圍的變化與權力中心的遷移。早期檔案內容集中於華北、華東與南京等地，可見國民政府在北伐成功後，逐步進行較大範圍建設與行政規劃，並開始應對日本侵華挑戰。而隨著中日戰爭發展與政權穩定，重慶成為臨時首都，情報、外交、各地黨務運作等檔案內容在全國皆有明顯增長，各地區國共對立狀態亦日漸升級。隨著汪精衛政權建立、日本戰敗、臺灣光復與國共內戰爆發等事件，檔案重心逐步往南部與東部沿海移動，各地黨部之工作回報亦明顯減少，臺灣相關檔案數則逐步增加，可看出內陸地區的影響力漸漸減弱，權力核心向東南沿海轉移的過程。

四、綜合討論

本研究運用三種不同數位工具，對國民黨特種檔案進行文本內容分

析，包括以CKIP Tagger進行文本預處理，自萬餘筆檔案目錄辨識、擷取出其中人物、組織、事件、地名等詞彙；再透過中研院數位人文平臺的權威詞共現分析功能，挖掘檔案中人物、組織與事件之間的共現關係，構建社會網絡，釐清複雜的歷史脈絡；並結合地理資訊分析，描繪不同時期檔案中地理位置的分布與變遷。

共現分析結果展現了如蔣中正、吳鐵城、何應欽、邢森洲、徐恩曾等重要人物在外交、軍事與情報領域的核心角色；而中日抗戰、朝鮮民族革命等歷史事件與人物、組織間的聯繫；乃至軍委會、中秘處、外交部等機構在國防、外交與財政事務中的協同合作。讓國民政府時期，國內外複雜政局的關鍵層次、關聯，可一目了然。從地理資訊分析結果中，則可看出1928 ~ 1949年間國民政府在南京、重慶、西南、東南沿海至臺灣的活動範圍，並充分佐證了越南、韓國、印度等國在檔案中的重要性，除反映國民黨政府黨務、內政、外交活動範圍的變化，也存在對於鄰近國家革命運動的參與。本研究結合數位工具進行檔案文本內容多重分析，可使難以全面解讀的萬餘筆檔案得以從多角度剖析，進一步提升對此段歷史過程的整體概觀，並展現數位工具在檔案研究中的潛力與價值。

伍、結論與建議

一、結論

（一）以詞頻統計與視覺化釐清不同人物、組織之活躍程度，可深化對於檔案內容的解讀

本研究以詞頻統計與視覺化為基礎，呈現了特種檔案中重要人物和組織於行政、外交、軍事等領域的活躍程度，並輔以歷史事實之質性解釋，展現高頻人物與組織在此批檔案中扮演的角色或可能產生之影響。不僅為歷史研究者提供更具體的時間軸索引，亦幫助研究者可更為直觀理解該時期國民黨政府之內政運作、外交策略、戰時活動之變化和發展。

（二）以社會網絡分析釐清文本中人物、組織與事件的關聯脈絡，有助於觀察歷史背景當中的角色互動

本研究透過核心詞彙共現分析，系統性呈現特種檔案中文本與人物、事件、組織之間的互動結構，並依據詞彙關聯強度及節點數進行資料篩選，確保圖像可讀性與解釋性。如人物與人物共現網絡中，顯示外

交、軍事、情報等領域的權力結構與跨部門合作的複雜性。在人物、事件、組織等相互關係中，則可見人物或組織在重大事件（如抗戰、朝鮮民族革命、閩僑大會、金山會議等）中的重要角色與影響力，或政府各部門間的合作脈絡及跨機構的行動聯繫，梳理文本中隱藏的人物、事件與組織間的關聯脈絡，補足傳統研究中較難直接掌握的權力互動與跨層次協作關係。

（三）以地理資訊分析判讀各時期黨務、外交等活動範圍，以及地理位置分布趨勢與重心變化

本研究透過檔案中地名的整理與空間可視化，展示了國民政府在特種檔案中不同時期的地理布局與政務重心的轉移。如1928 ~ 1939年間，檔案多集中於華北、華東與南京，反映出國民政府北伐統一後推動訓政、建設國家的決心。而抗日戰爭爆發、遷都重慶後，國家治理及戰時動員體系，漸往南移，亦呈現出多中心的分布特徵，特別是華中、華南與西南地區，均記錄大量黨務、軍事及敵偽情報。至國民政府敗退前夕，檔案中的內陸地名逐步減少，重心轉往於沿海與南方地區，顯示統治權力已由華北、華東逐步向東南沿海及臺灣轉移的歷史趨勢。不僅補足傳統文本難以掌握或全面呈現的政治活動範圍，也藉檔案內容突顯了歷史事件與地理場域之間的深層關聯。

（四）利用多種數位工具進行檔案內容不同角度的分析，有助於檢視和解讀檔案內容

本研究以不同工具與方法對特種檔案進行不同類型分析，以權威詞共現分析探討檔案中不同人物、組織、事件間的共現關聯；以語料庫分析挖掘檔案中重要人物、組織的詞語趨勢脈絡；更以地理資訊分析繪製出不同時期檔案中的地理位置分布狀態。透過結合多元分析，能以各種角度剖析檔案，讓人工難以一一閱讀整理的萬餘筆特種檔案更容易瞭解內容梗概。

綜上所述，本研究運用數位人文工具，分析了1928 ~ 1949年間，特種檔案中涉及的關鍵人物、組織和事件等核心詞彙，與不同時期檔案中涵蓋的地理資訊。可呈現不同核心詞彙間的關聯網絡，並輔以質性說明提供對於黨務運作及當時中國內外政局的理解。本研究從宏觀角度探索檔案資料的全貌，採由上而下的鳥瞰角度，檢視檔案內容的整體全貌，並非針對特定史事進行史料研究分析，將詞目量化統計，並非以詞頻高低連結重要性的解釋或以數位分析結果論斷史事，本研究目的在於提供

量化觀察文本隱含的資訊，作為史學者進一步探究的參考。然其呈現結果，足證檔案數位人文研究能提供內容專家（如現代史研究者）探詢研究主題的指引，深入挖掘檔案中的人物、事件、組織的互動關係與地理資訊。但對於分析結果的詮釋和應用，仍需內容專家根據文本與史實，進行更深度的研究解釋。

二、建議

（一）對於數位工具操作建議

DocuSky中的DocuGIS功能雖可匯入全球地名與經緯度，但如有特定觀察區域，建議每次可針對單一地區盡可能放大檢視，本研究只針對整體中國或特定省分進行分析，可利於觀察結果。另外，DocuGIS功能提供之底圖可選擇地圖樣式，包括不同色系、布局、地名語言等，如使用群集或熱區功能，可採深色底圖為佳。

（二）對於檔案處理流程之建議

以數位人文工具進行歷史檔案研究，能達成人工研究難以企及的廣度，並可大幅降低研究所需的人力與時間，協助研究者快速梳理大量檔案內容。就本研究所使用之特種檔案目錄而言，雖然多已由史學專業人員完成內容摘要，但如〈上總裁電〉、〈附軍委會函1件〉、〈附董顯光致朱家驊函1件〉等附件，仍多缺乏內容摘述。此情形可能影響文本完整性，進而對分析結果產生一定影響。未來若能於檔案處理流程中，針對附件內容進行摘述或全文數位化，將有助於文本內容的蒐集及研究分析。

（三）對於後續研究之建議

本研究以國民黨特種檔案為基礎，初步建立「關鍵詞—人物—地理」之多維分析架構，聚焦於主題分布、共現關係與地緣資訊的整體呈現。未來研究可在此基礎上，進一步引入進階文本探勘技術，如主題模型、語義網絡分析等語言處理方法，可更細緻解析歷史語境中潛藏的權力結構、敘事策略與意識形態運作，深化對歷史文本理解。同時，亦可與內容研究專家合作，針對特定研究議題，進一步以中心性指標或模組劃分進行社會網絡分析，辨識潛在的中介人物、跨機構連結或地方派系網絡，甚至擴展不同時期、類型的黨政檔案進行比較，建立跨時段、跨文本的政治網絡與歷史敘事演變圖譜，以鞏固文本探勘在歷史研究中的方法論地位，拓展數位人文在檔案學與歷史學領域的應用潛力。

誌謝

本研究之部分成果係承蒙國家科學及技術委員會專題研究計畫（NSTC 113-2410-H-004-112）經費支持，本文內容與觀點僅代表作者立場，與補助單位無涉。

參考文獻

中央研究院（2023）。結合斷詞、詞性標記、實體辨識的中文處理套件（CkipTagger）。檢自https://iptt.sinica.edu.tw/shares/928【Academia Sinica. (2023). *Jie he duan ci, ci xing biao ji, shi ti bian shi de zhong wen chu li tao jian (CkipTagger)*. Retrieved from https://iptt.sinica.edu.tw/shares/928 (in Chinese)】

王麗蕉（2018）。數位檔案系統在人文研究的應用：以總督府職員錄系統觀察臺灣女力公職發展。圖資與檔案學刊，*92*，44-66。doi:10.6575/JILA.201806_(92).0003【Wang, L.-C. (2018). The application of digital archives system in humanities: Taiwanese female power in the Taiwan Sotokufu personnel directory. *Journal of InfoLib and Archives*, *92*, 44-66. doi:10.6575/JILA.201806_(92).0003 (in Chinese)】

李盈慧（2013）。吳鐵城與戰時國民黨在港澳的黨務活動。在陳鴻瑜（編），吳鐵城與近代中國（頁65-88）。臺北市：秀威資訊。【Lee, Y.-H. (2013). Wu Tie Cheng yu zhan shi Kuo min dang zai Gang Ao de dang wu huo dong. In H.-Y. Chen (Ed.), *Wu Tie Cheng yu jin dai Zhong guo* (pp. 65-88). Taipei: Showwe Information. (in Chinese)】

李福鐘（2021）。蔣經國在情治系統中的角色。蔣經國總統資料庫。檢自https://presidentialcck.drnh.gov.tw/article.php?access=PATA00022/24d3807f【Lee, F.-C. (2021). Chiang Ching Kuo zai qing zhi xi tong zhong de jue se. *Chiang Ching Kuo Digital Archive*. Retrieved from https://presidentialcck.drnh.gov.tw/index.php?act=Archive/article/PATA00022 (in Chinese)】

吳淑鳳、許順昇（2011）。「無名英雄‧英雄無名：探索情報作戰的歷史」座談會會議紀實。國史研究通訊，*1*，63-74。【Wu, S.-F., & Hsu, S.-S. (2011). “Wu ming ying xiong. Ying xiong wu ming: Tan suo qing bao zuo zhan di li shi” Zuo tan hui hui yi ji shi. *Academia Historica Newsletter*, *1*, 63-74. (in Chinese)】

林巧敏（2022）。國立政治大學典藏中國國民黨檔案之整理與檢索

服務。漢學研究通訊，*41*(2)，33-39。【Lin, C.-M. (2022). The management and retrieval services of the Kuomintang archives at National Chengchi University. *Newsletter for Research in Chinese Studies*, *41*(2), 33-39. (in Chinese)】

金觀濤、劉青峰、邱偉雲（2019）。中國現代主權觀念形成的數位人文研究。二十一世紀，*172*，49-67。【Jin, G.-T., Liu, Q.-F., & Chiu, W.-Y. (2019). Zhong guo xian dai zhu quan guan nian xing cheng de shu wei ren wen yan jiu. *Twenty-First Century*, *172*, 49-67. (in Chinese)】

徐友春（2007）。民國人物大辭典（增訂版）。河北省：河北人民出版社。【Hsu, Y.-C. (2007). *Min guo ren wu da ci dian* (Rev. ed.). Hebei Province: Hebei People's Publishing House. (in Chinese)】

郭乃華（2014）。中國古代法典及其事例之自動化整合——以乾隆朝《大清會典》為例（未出版之碩士論文）。國立臺灣大學資訊工程學系，臺北市。doi:10.6342/NTU.2014.02971【Kuo, N.-H. (2014). *Automatic integration of historical Chinese codices and their explanatory books: A case study of Records of Laws and Systems of The Qing Dynasty during the reign of Qianlong* (Unpublished master's thesis). National Taiwan University, Taipei. doi:10.6342/NTU.2014.02971 (in Chinese)】

陳宗嶽（2012年3月7日）。始終堅奉「忠黨愛國」的王天雄將軍〔部落格文章〕。檢自https://blog.udn.com/cty43115/6185517【Chen, Z.-Y. (2012, March 7). Shi zhong jian feng "zhong dang ai guo" de Wang Tian Xiong jiang jun [Blog post]. Retrieved from https://blog.udn.com/cty43115/6185517 (in Chinese)】

張奕萱、林巧敏（2022）。運用數位人文工具探討檔案時事議論主題及其情感分類之實作。圖資與檔案學刊，*101*，164-190。doi:10.6575/JILA.202212_(101).0005【Zhang, Y.-X., & Lin, C.-M. (2022). Using digital humanities tools to explore the implementation of topic and sentiment analysis in archival affairs commentary. *Journal of InfoLib and Archives*, *101*, 164-190. doi:10.6575/JILA.202212_(101).0005 (in Chinese)】

開放博物館（n.d.）。數位人文學與簡牘研究：文字釋讀與簡冊復原。檢自https://openmuseum.tw/muse/curation/a54bb27fd85d3991e917622a383fa0f7【Open Museum. (n.d.). *Shu we ren wen xue yu jian du yan jiu: Wen zi shi du yu jian ce fu yuan*. Retrieved from https://openmuseum.tw/muse/curation/a54bb27fd85d3991e917622a383fa0f7 (in Chinese)】

湯雅媜（2020）。兩岸關係的紅色警戒：中共官媒訊息文字探勘，2018年1月至2020年9月（未出版之碩士論文）。國立臺灣大學國家發展研究所，臺北市。【Tang, Y.-C. (2020). *Red alerts for cross-strait relations: Text mining study on Chinese state-run media outlets from January 2018 to September 2020* (Unpublished master's thesis). National Taiwan University, Taipei. (in Chinese)】

葉韋君（2019）。後五四時期的知識婦女：《婦女共鳴》的社會網絡（1929-1944）。近代中國婦女史研究，*33*，113-162。【Yeh, W.-C. (2019). Intellectual women in the Post-May Fourth Period: A social network analysis of women's resonance (1929-1944). *Research on Women in Modern Chinese History*, *33*, 113-162. (in Chinese)】

蔡永橙、黃國倫、邱志義（2007）。數位典藏技術導論。臺北市：國立臺灣大學出版中心。【Tsai, Y.-C., Huang, K.-L., & Chiu, C.-Y. (2007). *Introduction to digital archives technology*. Taipei: National Taiwan University Press. (in Chinese)】

鄧志松、郭迺鋒、林崑峯（2011）。國父行腳之空間分佈與社會網絡分析之應用。地理資訊系統季刊，*5*(3)，19-23。doi:10.6628/GIS.2011.5(3).4【Teng, C.-S., Kuo, N.-F., & Lin, K.-F. (2011). The spatial distribution of Dr. Sun Yat-Sen tracks and social network analysis. *Geographic Information System*, *5*(3), 19-23. doi:10.6628/GIS.2011.5(3).4 (in Chinese)】

劉壽林、萬仁元、王玉文、孔慶泰（編）（1995）。民國職官年表。北京市：中華書局。【Liu, S.-L., Wan, R.-Y., Wang, Y.-W., & Kong, Q.-T. (Eds.). (1995). *Min guo zhi guan nian biao*. Beijing: Zhonghua Book Company. (in Chinese)】

Asllani, A., Halstead, D., & Taylor, V. A. (2020). A text mining investigation of the presence and emotional impact of religious language in service organizations' websites. *Services Marketing Quarterly*, *41*(1), 68-87. doi:10.1080/15332969.2019.1707376

Digital Panopticon. (n.d.). *Tracing London convicts in Britain & Australia, 1780–1925*. Retrieved from https://www.digitalpanopticon.org/

Pal, K., Avery, N., Boston, P., Campagnolo, A., De Stefani, C., Matheson-Pollock, H., ... Weyrich, T. (2017). Digitally reconstructing the Great Parchment Book: 3D recovery of fire-damaged historical documents. *Digital Scholarship in the Humanities*, *32*(4), 887-917. doi:10.1093/llc/

fqw057

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135. doi:10.1561/1500000011

Ray, B. (2022). *The Salem witch trials digital archive: How and why*. Retrieved from https://salem.lib.virginia.edu/intro.html

Stanford University. (n.d.). Home. *The Grant Tour Project*. Retrieved from https://grandtour.stanford.edu/

Van Le, D., Montgomery, J., Kirkby, K. C., & Scanlan, J. (2018). Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. *Journal of Biomedical Informatics*, *86*, 49-58. doi:10.1016/j.jbi.2018.08.007

Zanasi, A. (Ed.). (2007). *Text mining and its applications to intelligence, CRM and knowledge management*. Southampton, UK: WIT Press.

The Study on Archival Text Mining: A Case Study of Te-Zhong Record Series

Chin Shao
Graduate Student
Graduate Institute of Library, Information and Archival Studies
National Chengchi University

Chiao-Min Lin
Professor
Graduate Institute of Library, Information and Archival Studies
National Chengchi University

Introduction

The exponential growth of digitized materials in the twenty-first century has fundamentally transformed the landscape of historical research. Traditional methods of manually reading and annotating archival records were once the gold standard of scholarly rigor, which is now being challenged by the overwhelming quantity, complexity, and diversity of historical sources available in digital form. Large-scale archival corpora, particularly those compiled by government or party institutions, can encompass millions of words across heterogeneous document types, far exceeding the cognitive and temporal limits of human readers. This situation is particularly evident in the context of modern Chinese history, where political upheavals, wars, and diplomatic exchanges have generated voluminous and multifaceted records.

In response to these challenges, this study adopts digital humanities methodologies to perform text mining on the Te-Zhong Archives (Special Record Series) produced by the Central Headquarters of the Kuomintang (KMT) between 1928 and 1949. The corpus comprises internal party communications, diplomatic dispatches, intelligence reports, and organizational directives, making it a rich yet methodologically demanding source for understanding the political, diplomatic, and military operations of

the Republican era. The study employs computational text analysis to extract named entities, analyze keyword frequency distributions, and identify patterns of co-occurrence among persons, organizations, and events. It integrates geographic information systems (GIS) to map the spatial dynamics of political activities over two turbulent decades.

The objectives of this study are twofold. First, it establishes a replicable analytical framework that can be applied to other large-scale historical archival collections. Second, it demonstrated how the combination of text mining and spatial visualization can enrich historical interpretation, providing macro-level structural insights and micro-level contextual readings. By bridging archival studies and digital humanities, this study addresses the increasing scholarly demand for methods that are computationally rigorous and historically meaningful.

Research Methods

The methodological design of this study comprises four main stages: data preprocessing, linguistic annotation, content mining, and spatial analysis.

1. Data Preprocessing

First, the Te-Zhong Archives catalog data were digitized and transformed into machine-readable formats (e.g., UTF-8 encoded plain text and CSV files) compatible with analytical tools. Care was taken to normalize character encodings, remove typographical inconsistencies, and reconcile historical orthographic variations (e.g., variant Chinese characters and early twentieth-century transliterations of foreign names). This step ensured that subsequent computational analyses would be accurate and reproducible.

2. Linguistic Annotation and Entity Extraction

Developed by Academia Sinica, the CKIP Tagger was used to perform Chinese word segmentation, part-of-speech (POS) tagging, and named entity recognition (NER). NER focused on extracting four categories of entities: persons, organizations, events, and place names. Given the historical nature of the corpus, additional custom dictionaries were created to capture archaic terms, political jargon, and proper nouns not found in standard lexicons.

This hybrid approach combines automated tagging with manual validation to maximize precision and recall in entity extraction.

3. Content Mining and Network Construction

After entity extraction, co-occurrence analysis was conducted on authority terms (standardized forms of names and organizations) and raw tokens to identify patterns of association. Term frequency–inverse document frequency (TF–IDF) measures were applied to identify significant yet non-trivial terms. The Academia Sinica Digital Humanities Platform was employed to construct social network diagrams, where nodes represent entities and edges represent statistically significant co-occurrence relationships. Network centrality metrics (e.g., degree centrality and betweenness centrality) were computed to identify key actors and mediating entities.

4. GIS Mapping

Place names identified during NER were geocoded into latitude–longitude coordinates using the DocuSky GIS tool. Historical place names were cross-referenced with contemporary maps and gazetteers to ensure spatial accuracy. Next, geocoded data were segmented into four distinct historical phases: Phase I (1928–1937), characterized by early Republican consolidation in Nanjing and northern cities. Phase II (1937–1945): Wartime relocation to Chongqing and expansion into South China. Phase III (1945–1947): Postwar reconstruction with increased activity in Taiwan, Vietnam, and Korea. Phase IV (1948–1949): Civil War retreat and relocation to Taiwan and the southeastern coast.

This temporal division enabled the visualization of shifting political geographies, highlighting correlations between archival production sites and broader historical events.

Research Results

1. Core Vocabulary Analysis

The frequency analysis revealed that Wu Tiecheng emerged as the most frequently mentioned individual in the corpus. This finding reflects his pivotal

role spanning diplomacy, the Secretariat, and central leadership. His presence in a wide range of document categories—from foreign correspondence to internal directives—underscores his strategic importance within the KMT political structure. Other high-frequency figures include: Xing Senzhou, associated with overseas diplomatic missions, Chiang Kai-shek, representing supreme military and political authority, Chen Su, associated with overseas party organizations, and Xu Enzeng, prominent in intelligence and security operations. These figures embody the core decision-making network of the party-state.

Organizationally, the most recurrent terms include the Secretariat, Party Headquarters, the Military Affairs Commission, and the Ministry of Foreign Affairs. The prevalence of these institutions underscores to the administrative origins of the documents and highlights the interconnected nature of internal governance, foreign policy, fiscal planning, military strategy, and propaganda during this period.

2. Co-Occurrence Network Analysis

The network graphs indicate that Wu Tiecheng shares dense co-occurrence links with Xing Senzhou, Chiang Kai-shek, and Xu Enzeng. This finding highlights his role in bridging diplomatic and intelligence networks. In event-based associations, Chiang Kai-shek appears strongly connected to the Second Sino-Japanese War, the New Fourth Army Incident, and the Jinshan Conference, signaling his direct involvement in military strategy and high-level political negotiations.

Notably, there is a strong link between Korean independence activists Kim Yak-san (金若山) and Kim Kyu-sik (金奎植) and events related to the Korean national revolution. This finding suggests that the KMT was not only engaged in domestic and Sino-Japanese affairs but also actively participated in shaping regional geopolitics. Organizational co-occurrence patterns further reveal institutional collaborations, e.g., the Military Affairs Commission working closely with the Secretariat, and the Ministry of Foreign Affairs coordinating with the Executive Yuan. These patterns illustrate the cross-departmental dynamics mobilized in response to crises such as war, decolonization, and territorial realignment.

3. Geographic Information Distribution

Divided into four chronological phases, GIS visualization provides a spatial complement to textual analysis. Phase I (1928–1937) comprises Nanjing and major northern cities, corresponding to the consolidation of the Nationalist Government and early state-building efforts. Phase II (1937–1945) focuses on Chongqing and South China, reflecting wartime relocation of the capital to the south, which resulted in intensified diplomatic activities. Phase III (1945–1947) records an upsurge in references to Taiwan, Vietnam, and Korea, highlighting the KMT's postwar geopolitical priorities, including the administration of newly acquired territories. Phase IV (1948–1949) exhibits a pronounced concentration in Taiwan and the southeastern coast, mapping directly onto the government's retreat and re-establishment following defeat in the Civil War. These spatial patterns mirror well-known historical developments while exposing lesser-known geographic nodes of political and diplomatic activity.

Conclusion

This study demonstrates the methodological potential of combining computational text mining, social network analysis, and GIS-based spatial visualization in examining large-scale historical archives. Applied to the Te-Zhong Archives, these techniques elicit the structural relationships among key actors, organizations, and events, while revealing the shifting geographies of political activity during a transformative period in modern Chinese history.

The findings confirm that digital humanities methods can supplement and, in some cases, surpass traditional manual reading in terms of efficiency and scope. By identifying central figures, reconstructing inter-organizational networks, and mapping the spatial dimensions of archival production, researchers can gain a macro-historical overview and targeted micro-historical insights. The process outlined in this study—comprising data preprocessing, linguistic annotation, content mining, and GIS mapping—can serve as a replicable and adaptable framework for other historical corpora.

Ultimately, the Te-Zhong Archives represent more than documenting the internal workings of the KMT; they reflect the intertwined histories of Chinese

diplomacy, military strategy, and state relocation. Interpreted through the lens of digital analysis, these records acquire renewed relevance. This provides new pathways for historical inquiry while contributing to the evolving methodological repertoire of archival and digital humanities scholarship.