政府開放資料基於時間與空間的跨檔案應用之初探性研究

林其緯
國立暨南國際大學資訊管理學系
s110213514@mail1.ncnu.edu.tw

陳彥錚
國立暨南國際大學資訊管理學系
ycchen@ncnu.edu.tw

摘要

隨著資通訊技術發展進入大數據時代，開放資料已成為為全球資料應用主流，目前政府開放資料平台雖有提供大量資料集提供大眾使用，但目前平台上許多資料格式不一或無嚴謹架構，資料品質不一，難以直接加以利用，此問題在針對不同時間或不同行政區域的相同資料項目統計的跨檔案應用更為嚴重。這些應用需要對數據進行預處理，但預處理過程相當複雜。本論文對基於時空特徵的政府開放資料跨檔案應用進行初步研究。本研究旨在實現一個開發平台，可用於開發各種開放資料的跨檔案應用。本論文首先探討開發跨檔案應用的可能問題，然後提出可能的解決方案，以期降低應用開發的成本與門檻。此初探性研究也將實作一個名稱解析工具，用於萃取政府所有開放數料集名稱上的時間和空間屬性，以利跨檔案應用的開發。本論文並認為未來更需要一個具有各式輔助工具的開發平台，協助開發政府開放資料的跨檔案應用。

關鍵詞：開放資料、跨檔案應用 、時間性資料、空間性資料

聯絡人： 陳彥錚 教授
國立暨南國際大學 資訊管理學系
南投縣 545 埔里鎮大學路 470 號
TEL: (049) 2910960 轉 4654
FAX: (049) 2915205
E-mail: ycchen@ncnu.edu.tw

壹、導論

為了因應大數據時代的來臨，越來越多企業、民間團體、社群網站等都在各領域累積龐大資料並進行加值應用，其中政府開放資料也引起學術界以及各公、私營部門關注，由政府提供的開放資料，不僅可以促進跨機關間的資料流通，也可以增進政府施政透明度，更能提升民眾生活品質，同時也滿足產業需求，而且有了穩定公開的資料源，大眾才有機會使用、分析這些數據，進行利用資料產生新的應用，創造新的社會價值以及促進經濟成長。

臺灣提供開放資料已經行之有年，且政府為了確保資料集具有一定品質，也訂定各項準則來規範各機構、地方政府，以期所釋出的資料集能夠符合相關規定，例如各個資料集皆須提供通用的資料集屬性，包含欄位名稱、檔案格式、編碼格式、資料下載網址、主題分類、服務分類等屬性。儘管如此，本論文研究發現目前平台上資料集仍有諸多問題，由於資料來自中央與地方政府各機關，資料品質不一，且多為沒有經過嚴謹結構化的資料，導致相關類型的資料集可能會因為來源或時間不同，在資料內容的定義及詮釋上會有差異(鍾致浩, 2020)，當相關應用需取用這些資料時，常無法看到全面、整合性的資料，通常都需要進一步的解構與分析，此問題不僅增加整合應用的複雜度，對普羅大眾在使用上也多了一道門檻。此外，許多開放資料的加值應用需要基於不同時間或空間場域的資料統計與分析，政府開放資料中有許多資料集是基以時間或空間資訊呈現(陳宏洋, 2018)，例如某一年度每個月份各高速公路休息站營業額統計。如果這些時間或空間資料欄位沒有一致性，政府開放資料的每一資料集都自成一個資料孤島，很難有跨年度跨縣市的大的時間與空間軸度的有效應用，例如過去十年台灣各縣市老年人口或薪資所得變化與比較。這些都需要跨不同時間與行政區域的多重資料集資料分析與統計，一旦這些來自不同單位或同一單位不同年度資料集在時間空間資料表示的格式或尺度不一致，這類具整體性長時間統計分析的開放資料應用便會窒礙難行，失去了政府開放資料的目的與價值。假使能夠制定共通的資料格式標準，用以鏈結不同時空間的開放資料集，並將其中具高度相關性的資料清理、整合，賦予資料新的意義，對於資料加值應用會有更好的價值。

儘管政府有許多開放資料集提供大眾使用，但目前大部分使用方式仍然停留在檔案下載或提供 API 串接，沒有更進一步的加值應用是相當可惜的，例如，在疫情期間，像是口罩、快篩、疫苗等防疫物資的缺乏，並且在政府嚴格的控管下，民眾為了取得物資，就必需花時間四處尋找且排隊等待，雖然政府有即時資訊公告至資料開放平台，但一般人不會為了瞭解附近據點的物資現況，時時關注平台上資料集的變化，並且在經由人工處理來獲取自己所需的資訊。倘若，這時有一個便民的即時物資查詢服務，民眾就不用如此大費周章且提心吊膽的四處奔波。而本實驗室之陳彥錚教授為處理該窘境，利用相關開放資料集開發即時物資查詢服務「COVID-19 家用快篩試劑數量即時查詢平台」(陳彥錚, 2022)，如(圖 1 **錯**

誤! 找不到參照來源。)。該平台使用者透過縣市別和鄉鎮別的選項查看附近服務據點或是使用 Web GPS 定位方式來取得當前位置的鄰近據點並顯示其物資數量，不只提高查詢方便性，更能快速了解附近據點狀況。平台服務範圍最初只有埔里鎮的居民，但經過更新後納入更多台灣地區相關資訊，讓全台民眾皆能使用該服務。圖 1：COVID-19 家用快篩試劑數量即時查詢平台(平台作者:陳彥錚教授)

受上述開發經驗啟發，本研究進一步探討當一個政府開放資料的應用是需要使用多個政府開放資料集時，開發相關應用時可能會遇到哪些問題。跨資料集的資料應用涉及不同單位層級與不同時間軸度，預期會有許多待克服的問題，本論文先進行初探性研究，從分析目前政府所有開放資料集的分析，找出可能可以做為跨檔案應用的標的，初步先聚焦時間與空間資料格式與表達一致性問題，指出目前政府開放資料集，並尋求解決之道，以促使政府開放資料跨檔案應用的發展。

貳、文獻探討

一、政府開放資料

開放資料（open data）指的是任何人都可以自由獲取、重複使用且再次發布的資料，且應滿足「開放授權」、「方便近用」、「開放格式」等三項條件(中華民國數位發展部, 2023)，這種資料不僅不受著作權、專利權及其他管理機制所約束，更不應對任何人或團體在取用上有所限制。而在眾多領域當中，其中尤以政府掌握大多數與人民利益、國家發展的相關資料備受各界關注。配合時代發展進步所需，政府除了最基本的資訊公開外，也更一進步的推動資料開放，而「政府資料開放」（Open Government data）就是由政府公部門、機關將其所蒐集持有的大量資料進行數位化後，以資料集為單位並採開放格式(CSV、XML、JSON)放置於網路平台上，方便民眾取得所需之資料，且可以在不受到任何限制的情況下，進行資料編輯、資料分析及資料的公開傳輸，如此一來，不僅促使跨機關的資料流通，也可使民眾更積極參與政府推動的決策與施政的方向，並且在政府有限的能力下，結合民間無限的創意與力量，開發出多樣性的應用程式，加深這些開放資料的加值應用(陳泰銘, 2018)。

政府機關開放資料的做法有助於提高政府透明度以及社會參與度，讓公民更加了解政府的運作，同時也促進了創新和經濟發展，但政府開放資料的實現與品質提升，更是需要政府的積極參與和支持，包括制定開放資料政策、建立資料集的標準、提供資料的品質保證、保障資料的隱私和安全等，有好品質的開放資料，不但方便資料的流通，更可以讓學校、企業以及民間團體更好的做為學術研究利用、數據分析的基底，發揮資料本身之價值，讓社會大眾更容易參與大眾事務，打破以往民間與公部門間的隔閡。

二、 政府開放資料詮釋資料

詮釋資料的在定義上，是指任何一種可以用來增加網路電子資源辨識、描述與定位的資料，隨著開放政府資料政策的實行，除了強調資訊公開透明外，也重視詮釋資料的建立和分享。詮釋資料是指對原始資料進行整理、解釋和標註後所產生的資料，它能夠讓使用者更容易理解原始資料的意義和價值，並且方便資料的應用和再利用。而詮釋資料的建立可以分為兩個階段：第一階段是資料的整理和標註，這是基於資料本身的特性所進行的工作，包括資料的格式、屬性、語義等。第二階段是資料的解釋和說明，這是針對資料使用者的需求進行的，包括資料的背景、來源、範圍、限制等。這些詮釋資料可以被包裝成為資料目錄、資料說明文件、使用手冊等形式進行分享，讓使用者更容易地找到和理解資料，進而提高資料的價值和應用效益。

在開放政府資料的背景下，詮釋資料的建立和分享已成為一個重要的課題。政府應該針對不同的使用者需求，進行適當的詮釋資料建立和分享，並且提供多樣的詮釋資料形式和分享方式。例如，對於學術研究者來說，資料的語義和來源是非常重要的，政府可以提供詳盡的資料說明文件和相關文獻資料；對於普通使用者來說，資料的應用和操作更為關注，政府可以提供使用手冊和教學影片等形式的詮釋資料。此外，政府也應該加強詮釋資料的品質管理和監督，以確保詮釋資料的準確性和可靠性。政府可以建立詮釋資料的標準和指南，並且定期進行詮釋資料的審核和更新，以保證詮釋資料與原始資料的一致性和時效性。

從使用者觀點而言，目前釋出的開放資料最主要問題之一，是如何找到且取用合適的資料集，在歐俐伶等人的研究中表示目前對於開放資料的相關詮釋資料相當缺乏(歐俐伶 & 楊東謀, 2016)，導致資料集的內容及定義相當明確，雖然已有許多探討與應用詮釋資料的文獻，但利用詮釋資料來管理政府開放資料的相關文獻卻相當稀少，因此，若能引進或發展合適的詮釋資料架構應用於開放資料領域，輔助資料集的取用與管理，就能讓使用者更方便的取用開放資料，不僅間接促進開放資料的推廣與發展，同時也能對資源的取用、加值更有所助益。

三、 網路爬蟲

網路爬蟲是透過模擬使用者瀏覽目標網頁，自動抓取所需資訊。在一般情況

下，使用者在搜尋引擎輸入關鍵字後需要透過點擊搜尋結果的連結，進入網站後，才能獲取所需的資訊。不過，網路爬蟲可透過程式自動化地擷取網頁上特定類型及命名格式的資料，以快速且有效地提供使用者所需的資訊，透過此技術，使用者可以節省時間與精力，並且更快速地找到符合其需求的資料。而在實際應用中，網路爬蟲技術可以分為「動態爬蟲」和「靜態爬蟲」，主要的區別在於爬蟲程式是否針對特定網站進行指定，若有指定網站，靜態爬蟲可直接從該網站獲取資訊；若未指定網站，動態爬蟲則需模擬使用者的操作，例如輸入資料、點擊按鈕等，網頁完成載入後，才會開始爬取資料。透過適當選用不同的爬蟲方式，可以更有效地取得所需的資訊並節省資源。

身處資訊量如此龐大的時代，如何有效的收集資料是相當重要的工作之一，如果透過人工搜尋、複製的方式來收集資料，除了效率不佳外，還會花費相當多的人力成本及時間，所以透過網路爬蟲技術來協助進行資料蒐集，只要在開始蒐集前先制定好規則，就可以依照規則來蒐集和擷取資料並整理出我們所需的內容及格式。本論文將使用網路爬蟲技術對政府開放資料進行分析。

四、 其他相關研究

本研究之實驗室最近對政府開放資料的相關研究(黃雅琳, 2022)是探討政府開放資料的檔案格式以及編碼問題，彙整目前政府開放資料格式之相關樣態，包括資料編碼與檔案格式兩個主要問題，分別研究其不同樣態，並針對以上問題提出階層式架構之表格資料定位與萃取方法，以解決現有工具無法滿足使用者欲利用表格式資料進行應用開發的情況。

而如果單以時間及空間資料集應用，鄒惠貞等人(鄒惠貞、葉信伶、江威誼、江博煌, 2015)利用台南市含有地理資訊的台南市本土登革熱病例數資料集，空間資料整合則利用 TWD97 座標系統的鄉鎮區用來分析台南市流動人口集中相關點位資訊，包含鐵路交通、古蹟、市場、夜市與醫院點為資料，分析登革熱爆發後之後三個月的人口分布資訊，並透過地理資訊系統之經驗貝氏克利金法(Empirical Bayesian Kriging, EBK)及反距加權法(Inverse Distance Weighted, IDW)，分別進行大台南地區之登革熱案例分布與風險人口的擴散模式之推估。

另一相關研究(劉仲鑫 & 林昀葆, 2017)則是利用政府開放資料中的運動城市調查各縣市規律運動現況，以運動人口、運動次數、規律運動頻率來探討國人的運動狀態，分析運動時的生理狀況且搭配腦波儀及行動裝置，擷取腦波專注和放鬆的資料，並且利用機器學習，採用回歸演算法進行預測分析，已決定係數做為評估模型的依據，給予人們適當的運動方向及建議。

參、政府開放資料集在跨檔案應用之相關問題

直到本論文的分析時間 2023 年 4 月 10 日為止，我們可從政府開放平台所獲取的資料數量共有 55,985 筆，提供共 7 種主題與 18 種服務分類，我們將各資料集所提供的檔案格式進行統計(如圖 2 所示)，發現有超過 5 成的資料採用 CSV 格式，本研究將以 CSV 格式的資料為探討對象，實際地從使用者操作資料的角度，探討其中可能需要額外防範與解決的錯誤或例外情況。

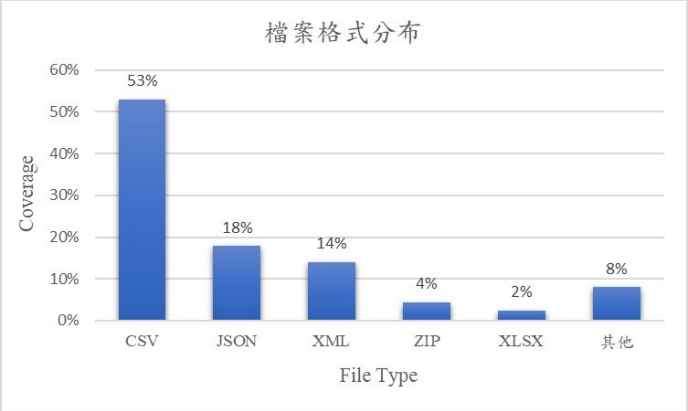

圖 2：檔案格式作分布統計圖

而依據我們目前的觀察，整理在跨檔案應用初期開發過程首先會遇到的問題，包含平台資料集整理方式、資料提供的時間與空間及資料內的欄位定義等，雖然政府對於開放資料的結構及提供方式都有規範，但實務上，可能因沒有考量跨檔案應用的需求，不同的資料提供單位對於相同或相似主題的資料並沒有一致的欄位定義及呈現方式，導致資料內容及品質參差不齊。以下是本研究以跨檔案應用開發者觀點，指出台灣政府開放資料待改善的問題，包括平台資料集管理、時間與空間資料類型一致性、以及欄位定義與內容語義問題。茲於以下章節分別說明。

一、平台資料集管理

關於平台資料集管理問題，政府開放資料平台上有來自各機關及各地政府的開放資料，這些資料來源、發佈時間都相當廣泛，但沒有統一的資料管理方式。例如，以黨團補助季報表(圖 3)及臺南市兒童及少年寄養家庭戶數(圖 4)的搜尋結果為例，兩者都是以季為單位進行資料統整，但前者是將該年度每季資料放置於平台上，以此為例，一年共有四季，在平台上顯示為四筆季報；而後者雖然也是季報，但將每季資料包含在年度裡面，只在平台上只顯示為一筆年度報表。雖平台上有提供關鍵字搜尋，但要在無明顯標示的情況，實務上很難快速、精確地找到所需資料。本論文研究認為對於具有時間性的資料集管理上需要統一規範，如果是以季或以月為最小時間單位進行資料整理時，釋放資料時，例如能統一以年為時間單位整併在一起，使用者在取用資料時，可以更容易以一致方式掌握所有

資料集與時間相關的資料統計，不用再花費大量時間搜索所需的資料，也可以促進開放資料平台整體的一致性。

圖 3：黨團補助季報表

圖 4：臺南市兒童及少年寄養家庭戶數

二、時間與空間資料類型一致性

截至目前本論文統計政府開放平台資料，我們可以發現雖然有些資料集的主題屬性相似，但因為由不同政府機關或單位所提供，導致各資料集名稱的表達方式皆不盡相同。例如(圖 5-圖 8)，在時間命名方式上，分別有採用以年度命名(臺南市政府民政局, 2021)、以季命名(臺南市政府社會局, 2021)、以月區命名計(臺南市政府社會局, 2023a)、以月命名(臺南市政府社會局, 2023b)或者有些資料更是將歷年資料都整理於同一份檔案內，無法直接得知檔案時間，此外，在時間表達方式上有些則是使用民國年，而有些是使用西元年；有些加上時間單位文字，但有些只有純數字；有些使用特殊符號作區隔，有些則不區隔；在表達時間範圍時，更會遇到符號上與文字上的差異，最經常用到的符號為「-」與「~」，而文字則為「至」或「到」。在空間表達方式上，縣市名稱有「台」與「臺」用法混亂、縣市分級標示不清及非所有資料集皆有紀錄或顯示區域別。

圖 5：以年度命名

圖 6：以季命名

圖 7：以月區間命名

圖 8：以月命名

經本研究進一步詳加觀察，整理平台上資料集名稱大部分的時間型態，針對上述問題，在時間的判斷上，我們建議採用正規化表達式(Regular Expression)的方式進行檢查，將所觀察到的不同時間表示分別以不同的正規化規則分別列出；而在空間資料表示的不一致上，由於目前政府開放資料平台本身即有一個台灣縣市鄉鎮列表的資料集，本論文建議直接引入台灣縣市鄉鎮的開放資料作為判別規則來源。本論文基於時間與空間的跨檔案應用開發需求，在針對所有的開放資料集進行分析時，便是以此建議方式萃取每一資料集名稱上所標示的時間與空間資訊，並個別紀錄，資料集名稱其他剩餘文字也須保存，作為跨檔案關聯的重要依據。

三、欄位定義與內容語義問題

單以一個資料集審視資料欄位定義與內容語義，基本上不會有一致性的問題，然而對跨檔案應用而言，當處理的對象為多個資料集檔案，這些資料集彼此在欄位定義與內容語義沒有統一時，便非常不利於跨檔案應用的開發。本研究以「iTaiwan 熱點」主題為例(圖 9-圖 12)，分別取用高雄 iTaiwan 熱點資料(高雄市政府研究發展考核委員會, 2023)、嘉義 iTaiwan 熱點資料(嘉義市政府, 2023)、雲林 iTaiwan 熱點資料(雲林縣政府, 2021)、南投 iTaiwan 熱點資料(南投縣政府, 2023)四個縣市所提供的資料集為例，說明跨檔案應用面臨的問題。

(1) 欄位數量不一致：從(圖 9)、(圖 10)、(圖 11)、(圖 12)，我們可以清楚發現，同樣是各縣市 iTaiwan 熱點資料，但各個資料集的欄位數量不盡相同，欄位數量不一致可能來自提供資料多寡與詳細程度不一，例如雲林 iTaiwan 熱點資料只有熱點名稱與地址兩個欄位，沒有提供經緯度資料。

(2) 欄位規劃不一致：上述欄位數量不一致，除了提供資料詳細程度不一，也有可能是因為欄位規劃本身的不一致，例如南投 iTaiwan 熱點資料以個別欄位

顯示郵遞區號，其他縣市則包含於地址欄位，此外，南投 iTaiwan 熱點資料也多了地區欄位。另外，嘉義與南投有主管機關欄位，高雄與雲林則沒有，這些欄位規劃不一致都不利於跨檔案應用的開發。

(3) 欄位名稱問題：即使不同的資料集包含相同語義資料，但欄位名稱卻不一致。例如南投 iTaiwan 熱點資料(圖 12)在欄位名稱上特地加上欄位備註說明，與其他縣市明顯不同。另外，即使去除備註文字後，嘉義、南投、雲林都有熱點名稱欄位，然而高雄卻以"地點"為欄位名稱。

(4) 欄位語系問題：相同的資料欄位以不同語言進行命名也可能不利於跨檔案應用在資料欄位的對應。例如經緯度欄位，高雄 iTaiwan 熱點資料使用英文來做為欄位名稱，但嘉義 iTaiwan 熱點資料和南投 iTaiwan 熱點資料卻是用中文。

以上是目前本論文對開放資料跨檔案應用在欄位定義與內容語義上初步研究的發現，從以上四個問題我們可以清楚了解，雖然都是提供 iTaiwan 熱點的資訊，但在沒有強制規範欄位定義及內容語義指引下，很容易因為每一個資料提供單位對於開放資料的理解不同，導致欄位定義與內容語義的不一致。

	地點	地址	latitude	longitude
2	高雄市新興區公所	800高雄市新興區中東里中正三路34號4樓	22.631253	120.309992
3	高雄市政府交通局本部	800高雄市新興區中正三路25號16樓	22.630656	120.30986
4	高雄市稅捐稽徵處新興分處	800高雄市新興區中正三路25號1樓	22.630549	120.30952
5	高雄市政府交通局停管中心3樓	800高雄市新興區中正三路25號3樓	22.62252	120.291235
6	高雄市新興區戶政事務所	800高雄市新興區中正三路34號2樓	22.631322	120.309949
7	高雄市立圖書館新興分館	800高雄市新興區中正三路34號3樓	22.631163	120.309879
8	高雄市政府地政局新興地政事務所	800高雄市新興區中正三路34號6樓	22.631072	120.310021
9	高市府資訊中心-逍遙園	800高雄市新興區六合一路55巷15號	22.631915	120.309953
10	高雄市立圖書館新興民眾閱覽室	800高雄市新興區民生一路271號	22.627272	120.306843
11	高市府資訊中心-城市光廊(公車站西)	801高雄市前金區中山二路與五福三路公車站	22.623365	120.30093
12	高雄市政府警察局	801高雄市前金區中正四路260號	22.627856	120.290712
13	高雄市勞工博物館1F	801高雄市前金區中正四路261號1F	22.62713	120.2899
14	高雄市勞工博物館3F	801高雄市前金區中正四路261號3F	22.62713	120.2899
15	高雄市勞工博物館4F	801高雄市前金區中正四路261號4F	22.62713	120.2899
16	高雄市勞工博物館6F	801高雄市前金區中正四路261號6F	22.62713	120.2899
17	高雄市立中醫醫院	801高雄市前金區中華三路68號1-4樓部份樓層-新盛街入口	22.627658	120.29701
18	高市府資訊中心-城市光廊(公車站東)	801高雄市前金區五福三路21號前公車城市光廊站	22.622563	120.299647

圖 9：高雄 iTaiwan 熱點資料

	主管機關	熱點名稱	地址	緯度	經度
2	嘉義市政府警察局	嘉義市政府警察局第一分局竹園派出所	600嘉義市西區博愛路二段455號	23.47294617	120.4323273
3	嘉義市政府警察局	嘉義市政府警察局第二分局北門派出所	600嘉義市東區民權路268號	23.48283005	120.4520645
4	嘉義市政府警察局	嘉義市政府警察局第二分局南門派出所	600嘉義市東區民族路230號	23.47664452	120.4553757
5	嘉義市政府警察局	嘉義市政府警察局第一分局長榮派出所	600嘉義市西區中山路567號	23.47917747	120.4431839
6	嘉義市政府警察局	嘉義市政府警察局第二分局新南派出所	600嘉義市東區吳鳳南路75號	23.47204208	120.4542389
7	嘉義市政府警察局	嘉義市政府警察局第一分局北興派出所	600嘉義市西區博愛路一段506號	23.48508263	120.4422684
8	嘉義市政府警察局	嘉義市政府警察局第二分局公園派出所	600嘉義市東區中山路2號	23.482711	120.463219
9	嘉義市政府警察局	嘉義市政府警察局第二分局後湖派出所	600嘉義市東區忠孝路537號	23.49850464	120.4504471
10	嘉義市政府警察局	嘉義市政府警察局第二分局長竹派出所	600嘉義市東區民權東路34號	23.48258972	120.480751
11	嘉義市政府警察局	嘉義市政府警察局第一分局八掌派出所	600嘉義市西區世賢路三段451號	23.46189308	120.4378052
12	嘉義市政府	ITWI000001嘉義市政府1樓	嘉義市東區中山路199號(1樓)	23.481224	120.453725
13	嘉義市政府	ITWI000002嘉義市政府文化局圖書館	嘉義市東區忠孝路275號(1F參考室及嘉義市政府文化局2F普通閱覽室)	23.487295	120.452583
14	嘉義市政府	ITWI000003嘉義市世賢圖書館	嘉義市西區世賢路一段685號(2樓電梯口上方天花板)	23.48986	120.428536
15	嘉義市政府	ITWI000004嘉義市黃資圖書館	嘉義市西區延平街328號(1樓服務台旁配線箱上方天花板)	23.476653	120.446969
16	嘉義市政府	ITWI000005嘉義市市立博物館	嘉義市忠孝路275-1號(1、2、3F大廳)	23.487233	120.451758
17	嘉義市政府	ITWI000006嘉義市史蹟資料館	嘉義市公園街42號	23.481106	120.467468
18	嘉義市政府	ITWI000007嘉義市射日塔	嘉義市公園街46號	23.481367	120.469028

圖 10：嘉義 iTaiwan 熱點資料

	熱點名稱	地址
2	雲林縣虎尾鎮公所虎尾廳	624雲林縣虎尾鎮行政路6號
3	自來水公司斗南營運所	630雲林縣斗南鎮中山路172號
4	臺鐵局斗南站候車室	630雲林縣斗南鎮中山路2號
5	斗南新光郵局	630雲林縣斗南鎮光興路78號
6	臺鐵局斗南站第2月台行車室	630雲林縣斗南鎮南昌里中山路2號
7	斗南郵局	630雲林縣斗南鎮南昌里中山路77號
8	高速公路局斗南工務段會議室	630雲林縣斗南鎮大業路119號
9	中油公司雲林縣斗南鎮斗南站	630雲林縣斗南鎮延平路二段165號
10	衛生福利部雲林教養院會客區	630雲林縣斗南鎮忠孝路157號
11	雲林縣斗南鎮公所	630雲林縣斗南鎮文昌路200號
12	臺鐵局石龜站月台候車區	630雲林縣斗南鎮石龜里(無地址)
13	斗南石龜郵局	630雲林縣斗南鎮石龜里中和路54號
14	大埤郵局	631雲林縣大埤鄉中山路6號
15	豐田(兼元長)工業區服務中心-豐田	631雲林縣大埤鄉嘉興村豐田路67號
16	台糖公司雲林區處辦公大樓	632雲林縣虎尾鎮中山路2號
17	虎尾郵局	632雲林縣虎尾鎮中山里林森路一段495號
18	華南商業銀行虎尾分行	632雲林縣虎尾鎮中正路50號

圖 11：雲林 iTaiwan 熱點資料

	主管機關（所屬機關名稱）	地區（所屬縣市名稱）	熱點名稱（依熱點所在地命名）	郵地區號	地址	緯度	經度
2	南投縣政府	南投縣	南投縣綜合大樓		540南投縣南投市三和一路8號	23.904083	120.689396
3	南投縣政府	南投縣	南投縣綜合大樓電腦教室		540南投縣南投市三和一路8號	23.904083	120.689396
4	南投縣政府	南投縣	南投縣風景區管理所		540南投縣南投市三和一路8號	23.903741	120.689602
5	交通部	南投縣	南投三和郵局		540南投縣南投市三和二路30號	23.904367	120.688465
6	南投縣政府	南投縣	南投縣南投地政事務所		540南投縣南投市三和二路60號	23.904142	120.687046
7	經濟部	南投縣	中油公司南投縣南投市南投三和站		540南投縣南投市三和里中興路549號	23.906699	120.690651
8	交通部	南投縣	南投中山街郵局		540南投縣南投市三民里中山街259號	23.910518	120.683924
9	國家發展委員會	南投縣	青少年活動中心		540南投縣南投市中學路1號	23.949902	120.696208
10	財政部	南投縣	第一商業銀行(南投分行)		540南投縣南投市中山一街2號	23.9103767	120.6845506
11	財政部	南投縣	臺灣土地銀行南投分行		540南投縣南投市中山街202號	23.910014	120.685425
12	交通部	南投縣	南投中興郵局		540南投縣南投市中興新村中學路4號	23.949518	120.694397
13	臺灣省政府	南投縣	臺灣省政府資料館第1棟		540南投縣南投市中興新村中正路2號	23.956521	120.686445
14	行政院主計總處	南投縣	行政院主計總處(中部辦公園區)		540南投縣南投市中興新村光明路25號	23.938997	120.698673
15	國家發展委員會	南投縣	中興會堂(室內)		540南投縣南投市中興新村光榮北路1號	23.949308	120.696249
16	國家發展委員會	南投縣	中興會堂(戶外)		540南投縣南投市中興新村光榮北路1號	23.949308	120.696249
17	教育部	南投縣	國立公共資訊圖書館-中興分館		540南投縣南投市中興新村光華路123號	23.949744	120.694163
18	國家發展委員會	南投縣	國家發展委員會中興辦公室		540南投縣南投市中興新村府西路71號1樓	23.959833	120.686531

圖 12：南投 iTaiwan 熱點資料

針對上述問題，本研究認為如果能先將資料集內的標題欄位先行萃取並進行文字解析，篩選具代表性意義的文字，透過特徵標記方式，例如判斷是否涵蓋人、事、時、地、物資訊，或是否涵蓋中文、英文或數值資訊等，標註出每份資料表現突出的特徵後，利用這些特徵記錄，找出各資料集中可能的共通特徵進行融合，預期可以降低跨檔案應用開發的門檻，然而這些都有待相關輔助開發工具的支援方可實現。

肆、基於時間空間的政府開放資料集名稱解析

從前一章節探討政府開放資料跨檔案應用的可能問題，本研究嘗試以自動化的方式促進跨檔案應用發展的可行性，其首要的工作是先解析目前台灣所有的政府開放資料集，找出可以進行跨檔案應用的資料集，本研究認為資料集之間的時間及空間關聯最有可能是跨檔案應用的主要類型，例如跨年度或跨縣市的開放資料應用。因此，本研究針對資料集名稱中的時間與空間資訊，嘗試以正規化表達方式進行撰寫自動解析程式，解析程式在軟體方面使用 Python(3.10)語言並結合 pandas(1.4.2)套件進行開發，硬體部分則使用 ASUS 之桌上型電腦，表 1 為本研究的資料集名稱自動解析實驗設備規格。

表 1 資料集名稱解析實驗設備規格

	規格
處理器	11th Gen Intel(R) Core(TM) i5-11400 @ 2.60GHz 2.59 GHz
記憶體	16 GB
固態硬碟	1 TB
作業系統	Windows 11 64 位元作業系統

本研究使用的測試資料為政府資料開放平台上，截至 2023 年 4 月 10 日為止的 55,985 筆，我們先利用網路爬蟲將平台上的資料集進行蒐集，而後開始進行資料結構分析。首先，我們先針對時間資訊進行解析，透過第參章的時間與空間類型之問題所觀察到之情況，制定對應的 18 種正規化規則(如圖 13 所示)，如果資料集名稱符合這 18 種規則之其中一種，此解析程式會依據相符的規則萃取出

表示時間的部分，並且保留不包含時間的其餘部分。為提供一致的時間表達方式，萃取出來的時間部分會另外轉換成相同的時間格式，例如，如果是 109 年第一季會轉換成以起始時間 2020 年 1 月及結束時間 2020 年 3 月的表達方式加以記錄以利往後資料查詢或合併時，能更精確的判斷。其次，我們將已經經過時間處理後的字串，再進行空間資訊的文字解析，我們引入台灣縣市鄉鎮的開放資料(圖 14)，利用 pandas 讀取檔案，取用檔案內的「縣市名稱」及「鄉鎮市區」資料且整理成陣列，而後開始進行正規化表達式處理，如果有符合空間正規化表達式的字串資料時，解析程式會把空間的部分從字串中切割出來，先記錄在陣列裡面，同時也會保留不包含空間的部分，待全部資料完成後，將全部結果回傳記錄到資料庫。本研究經由解析處理，建立開放資料集名稱在時間與空間表示更為精確的後設資料(Metadata)。

本研究以網路爬蟲分析 55,985 筆資料之名稱，經統整後，其中包含時間資訊的資料計有 22,542 筆，包含空間資訊的則有 31,139 筆，而同時包含時間及空間資訊的資料則有 15,758 筆，此統整結果顯示政府開放資料有許多是可以基於時間或空間來進行跨檔案的資料聚合計算或資料關係查詢相關應用，此外，政府開放資料原資料名稱在時間空間表達不明確問題，經由名稱解析後所獲致的一致性，也有利於未來跨檔案應用開發。

```
condition = ['(\d+)年(\d+)月至(\d+)年(\d+)月', #0
             '(\d+)年(\d+)月至(\d+)月',#1
             '(\d+)年(\d+)~(\d+)月',#2
             '(\d+)年至(\d+)年',#3
             '(\d+)年(\d+)月(\d+)日',#4
             '(\d+)年(\d+)-(\d+)月',#5
             '(\d+)年(\d+)月',#6
             '(\d+)年至第(\d+)季',#7
             '(\d+)年第(\d+)季',#8
             '(\d+)年-(\d+)年',#9
             '(\d+)-(\d+) 年',#10
             '(\d+)-(\d+)年',#11
             '(\d+)年~(\d+)年',#12
             '民國(\d+)年',#13
             '(\d+)年度(\d+)月',#14
             '(\d+)年度',#15
             '(\d+)年',#16
             '(\d+)-(\d+)',#17
             '(\d+)'#18
            ]
```

圖 13：時間正規化規則

```
file_name = './file/city.xls'
df = pd.read_excel(file_name, sheet_name=0, usecols=['縣市名稱', '鄉鎮市區']) # c
df_city = df.drop_duplicates(subset=['縣市名稱', '鄉鎮市區'], keep="first")
city_dict = {k: list(v) for k, v in df_city.groupby(by='縣市名稱')['鄉鎮市區']}
```

圖 14：空間正規化規則

伍、結論與未來展望

本研究基於過去開發「COVID-19 家用快篩試劑數量即時查詢平台」(陳彥錚, 2022)之經驗，深知在短暫的時間限制下開發開放資料應用，需要一個系統化並能提供各式輔助工具的開發平台。對於基於時間或空間關聯的跨檔案應用，開發的困難度預期會更高，本論文對於此問題的初探性研究，首先聚焦於資料時間空間屬性的自動解析，未來計畫更進一步利用這解析後的時間空間屬性，建立一套完整的開發平台，以期降低跨檔案開放資料應用開發成本，促進跨檔案開放資料之加值應用，讓使用者能更便利地運用政府開放平台之資料。

本研究針對開放資料跨檔案應用可能問題進行探討，也針對資料集名稱的時間空間資訊進行自動解析，未來計畫深入探討並進行實作也已有具體的想法:

- **資料集名稱分析機制**：鑒於目前初探政府開放資料領域，先利用正規化表達式來處理資料集名稱時空間上的問題，因為此法能暫時處理目前開放平台上大部分的表達方式，然而我們深知此法必須依靠人工整理、觀察，且難免會有遺漏未處理的狀況或新的表達方式出現，如果需要一直關注資料集的變化，無法讓分析流程自動化，對開發者來說也是件相當耗費精力的事情，所以未來應結合自然語言與機器學習，讓時間與地理資訊之萃取能更加準確，使整個分析機制更為完善、精確。

- **資料編碼及格式**：如果欲開發整合系統，必先重新架構與整合大量零散的開放資料集，但據我們目前的觀察，在處理不同檔案間的編碼及格式上，勢必會遇到一些困難，例如政府開放資料平台規範提供的編碼格式應為 UTF-8，但實際上仍有部分資料集採用其他編碼格式，例如，BIG5、UTF-8-SIG 等，此外，檔案格式更是有 10 幾餘種，例如，CSV、JSON、XML、XLSX 等，如果跨檔案應用使用的資料集具不同編碼及檔案格式，經由本研究未來計畫開發的平台工具，使用者所取用的是開發平台處理後的資料，不用分神擔憂資料編碼及格式問題，可以將時間與精力花費在思考如何進行跨檔案應用。

- **地理資訊系統**：如前所述，政府開放資料平台上的資料集其實富含許多時間空間資訊，如果經過資料清洗、整併，有共通的標準格式完成資料鏈結後，將資訊中的 GPS 座標取出，搭配 Google Maps API，進行資料視覺化及地圖功的結合，再設計時間流動功能，便可按照時間的變化，查看政府開放資料、時間與地點三者的變化與趨勢。

參考文獻

1. 中華民國數位發展部. (2023). 政府資料開放平台-關於平臺. 中華民國數位發展部. Retrieved Apr. 10 from https://data.gov.tw/about

2. 南投縣政府. (2023). 政府資料開放平台-南投縣 iTaiwan 熱點. 南投縣政府. Retrieved Apr. 10 from https://data.gov.tw/dataset/38366

3. 高雄市政府研究發展考核委員會.(2023). 政府資料開放平台-高雄市 iTaiwan 地點與經緯度. 高雄市政府研究發展考核委員會. Retrieved Apr. 10 from https://data.gov.tw/dataset/43878

4. 陳宏洋. (2018). 結合地理資訊系統與機器學習於政府開放資料的應用與挑戰 國立中興大學]. 台中市. https://hdl.handle.net/11296/8u227t

5. 陳彥錚. (2022). COVID-19 家用快篩試劑數量即時查詢平台. 暨大資管系.

Retrieved Apr. 4 from https://ycchen.im.ncnu.edu.tw/RTestQuery.html

6. 陳泰銘. (2018). *開放資料即時性之自動檢測機制* 元智大學]. 桃園縣. https://hdl.handle.net/11296/2d6wfb

7. 雲林縣政府. (2021). 政府資料開放平台-雲林縣 iTaiwan 熱點. Retrieved Apr. 10 from https://data.gov.tw/dataset/27639

8. 黃雅琳. (2022). 政府開放資料之表格式應用開發的統一性方法之研究.

9. 鄒惠貞、葉信伶、江威誼、江博煌. (2015). 登革熱疫情的空間趨勢分析. *醫療資訊雜誌, 24:4*, 39-48.

10. 嘉義市政府. (2023). 政府資料開放平台-嘉義市 iTaiwan WiFi 無線網路熱點. 嘉義市政府. Retrieved Apr. 10 from https://data.gov.tw/dataset/160219

11. 臺南市政府民政局. (2021). 政府資料開放平台-嬰兒出生數按生母年齡分. 臺南市政府民政局. Retrieved Apr. 10 from https://data.gov.tw/dataset/140077

12. 臺南市政府社會局. (2021). 政府資料開放平台-110 年臺南市兒童及少年寄養家庭戶數. 臺南市政府社會局. Retrieved Apr. 10 from https://data.gov.tw/dataset/143480

13. 臺南市政府社會局. (2023a). 政府資料開放平台-109 年度臺南市兒少保護個案處遇及結案情形. 臺南市政府社會局. Retrieved Apr. 10 from https://data.gov.tw/dataset/140009

14. 臺南市政府社會局. (2023b). 政府資料開放平台-109 年高雄市政府所屬機關 iTaiwan 無線網路統計資料. 高雄市政府研究發展考核委員會. Retrieved Apr. 10 from https://data.gov.tw/dataset/128446

15. 劉仲鑫, & 林昀葆. (2017). 使用雲端機器學習預測運動時專注力與放鬆度影響之研究 [A Study of the Prediction of Attention and Meditation Using Cloud Machine Learning on Exercise]. *Electronic Commerce Studies*, *15*(3), 427-451.

16. 歐俐伶, & 楊東謀. (2016). 台灣政府開放資料之詮釋資料建置探討 [The Construction of Metadata for Open Government Data in Taiwan]. *教育資料與圖書館學*, *53*(1), 63-102. https://doi.org/10.6120/JoEMLS.2016.531/0043.RS.AM

17. 鍾致浩. (2020). 開放資料之融合、搜尋引擎與智慧化 LOD API 建置 國立暨南國際大學]. 南投縣. https://hdl.handle.net/11296/hb8uk6

A Preliminary Study on Temporal and Spatial Cross-file Applications of Taiwan's Government Open Data

Chi-Wei Lin
Department of Information Management, National Chi Nan University, Taiwan
s110213514@mail1.ncnu.edu.tw

Yen-Cheng Chen
Department of Information Management, National Chi Nan University, Taiwan
ycchen@ncnu.edu.tw

Abstract

With the development of information and communication technology, we live in the era of big data, with a great amount of open data. In Taiwan, the government provides an open data platform with many datasets for public use. However, the data formats of the open datasets on the platform are not all consistent and the data structures are not standardized. As a result, it is difficult to make use of these data directly. Especially, this issue will be more crucial for multi-file applications, e.g. statistics of the same data items in different time periods or in different government departments. These applications require the pre-processing of the data, but the pre-processing is quite complicated. This paper conducts a preliminary study on cross-file applications of government open data based on temporal or spatial characteristics. The study considers the issues in the development of a platform for implementing cross-file applications of various open data. In this preliminary study, we will first identify the issues in developing cross-file applications, and then propose possible solutions to reduce cost and overheads in the implementation of the applications. We will also present a name resolution tool to extract the temporal and spatial information of all the open data datasets provided by Taiwan's government. This paper concludes that a development platform with comprehensive tools is needed to help the development of cross-file applications of government open data.

Keywords: Open data, Cross-file application, Temporal data, Spatial data.