論述 》管理 · 資訊

運用網路爬蟲工具提升稽核效益

為引導機關多元運用資訊技術執行內部稽核，本文以旅宿業合法性比對為例，藉由網路爬蟲工具，針對各訂房網站刊登廣告之國內旅宿業者進行大範圍資料擷取，輔以「通用稽核模組」與臺灣旅宿網公開之合法旅宿資料進行整理及比對，主動勘探疑似未完成登記而刊登營業訊息之旅宿業者，以降低稽查成本，提升稽核效能。

林思岑（行政院主計總處綜合規劃處科員）

壹、前言

在大數據時代，透過網路此一巨大資料庫蒐集資料，進而分析作爲決策參據已是當今趨勢，其中網路爬蟲（Web crawler）係運用網路自動化抓取（Web scraping）技術，抓取所需網頁之數筆資料，包括文字內容、圖片等，抓取者可藉由取得之資料做後續的應用。爲協助機關內部稽核人員運用網路爬蟲技術辦理稽核作業，本文以該技術於網站抓取旅宿業者資料，結合行政院主計總處開發之「通用稽核模組」[1]執行資料整理及比對，勾稽疑似未依發展觀光條例規定領取營業執照或登記證（以下簡稱未合法登記）而刊登營業訊息之旅宿業者作爲稽查作業之參考，以提升整體稽核效率。另進行資料抓取及應用時，應注意著作權法等相關規定及資料即時性，俾利後續資料分析及加值應用之妥適。

貳、緣由

臺灣爲促進觀光事業，著力於發展在地多元文化特色，包括歷史故事、名勝古蹟、傳統技藝、種族風俗、自然景觀、農場互動、溫泉文化及生態體驗等，旅宿業者亦藉著旅遊群聚效應，於觀光勝地周邊拓展住宿商機吸引各地觀光客前往。依發展觀光條例第 21、24 及 25 條規定，經營觀光旅館業者、旅館業者、民宿經營者

應分別向中央或地方主管機關申請，領取證照始得營業，又同條例第 55 條之 1 規定，未依規定領取營業執照或登記證而經營觀光旅館業務、旅館業務或民宿者，以廣告物、電腦網路或其他媒體等，散布、播送或刊登營業之訊息者，處以罰鍰。

爲確保旅客住宿安全與消費者權益，以及保障合法旅宿業者之經營權，近年來各地方政府致力於稽查取締未合法登記之旅館、民宿及日租套房等旅宿業者，而現行稽查方式多爲被動接獲民眾檢舉或由稽核人員定期至訂房網站逐筆核對是否屬交通部觀光署臺灣旅宿網之合法旅宿業者，此種稽查方式不僅耗力費時，亦無法適時遏止未合法登記之旅宿營業訊息。網路爬蟲可模擬使用者瀏覽網站重複性模式，將人工手動逐筆複製、貼上資料之作業轉由電腦自動擷取網頁資訊，進而運用取得之資料進行各項處理及分析，例如：透過不同線上平臺，尋找最優惠之商品價格、利用股票網站取得公司財報資料以追蹤股價趨勢、自動下載新聞內容，掌握最新時事等。本文結合前開便捷之自動抓取技術及通用稽核模組等工具，針對刊登於各訂房網站之國內旅宿是否合法登記營業，進行大範圍且有效率之比對勾稽作業。

參、稽核步驟

爲有效清查國內未合法登記之旅宿，經下載免費 Web Scraper 應用程式，藉由圖像化方式擷取訂房網站相關資料，並將擷取之資料連同交通部觀光署臺灣旅宿網公開之合法旅宿資料，透過「通用稽核模組」進行整理、比對，以獲取適切之稽核證據，茲將作業步驟說明如下：

一、瞭解資料

瞭解訂房網站架構、抓取目標（旅宿名稱、地址）、臺灣旅宿網之合法旅宿資料，以及相關網站資料使用限制。

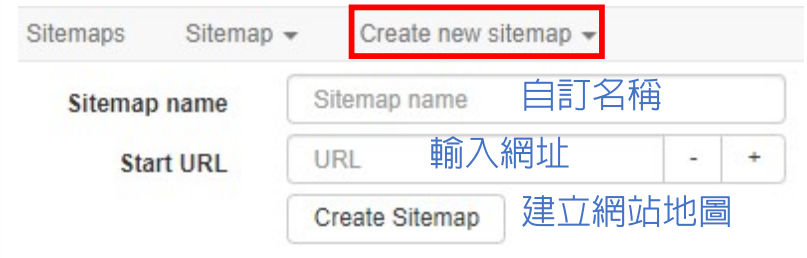
二、抓取資料（Scrape）

（一）首先在欲抓取的訂房網站資料頁面開啓 Web Scraper 應用程式。

（二）建立網站地圖（Create sitemap）：命名並輸入欲抓取網頁資料之網址後建立網站地圖（圖 1），建立完成後會出現新增新選擇器（Add new selector）選項。

（三）設定及選擇欲抓取之目標資料

圖 1　建立網站地圖



資料來源：擷取 Web Scraper 應用程式畫面，作者自行繪製。

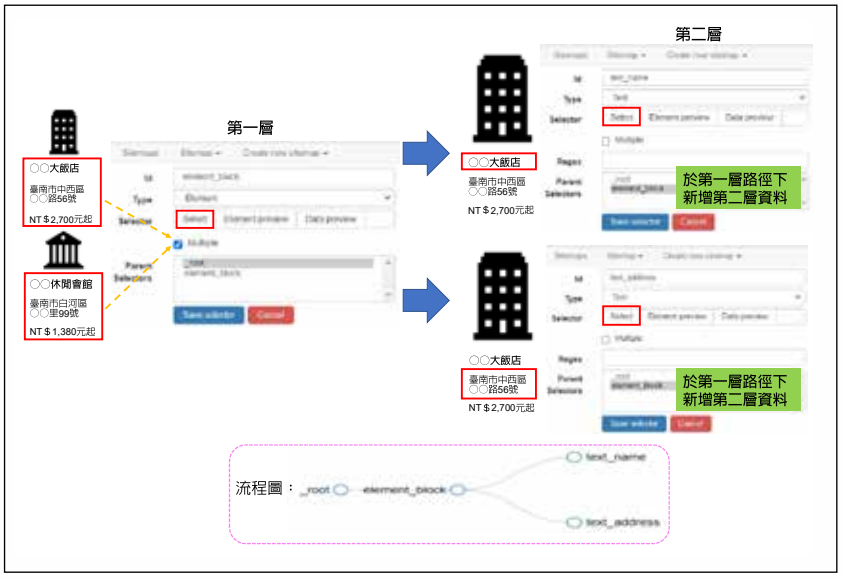
論述 》管理 · 資訊

如所有旅宿及目標資料（旅宿名稱、地址）皆呈現於同一頁面，首先點選新增新選擇器（Add new selector），爲抓取各旅宿資料，第一層類型（Type）選擇元素（Element），再選取（Select）各旅宿資料區塊，選取後點選完成選取（Done selecting），並勾選多筆數（Multiple）後點選儲存選擇器（Save selector），次爲抓取旅宿名稱及地址，接續點選前開已建立之第一層選擇器進入第二層，第二層須分別建立抓取旅宿名稱及地址 2 個選擇器，類型選擇文字（Text），再選取（Select）旅宿名稱（或地址）（圖 2）。設定抓取流程後，可先至選擇器圖解（Selector graph）檢視抓取邏輯是否正確，再點選抓取（Scrape）執行指令（圖 3）。又如旅宿資料呈現於不同頁面，亦可藉由 Web Scraper 應用程式設定及選擇擷取目標資料。

（四）匯出資料（Export data）

可選擇以 XLSX 或 CSV 檔案格式匯出資料（下頁圖 4）。

圖 2　抓取流程示例



資料來源：擷取 Web Scraper 應用程式畫面，作者自行繪製。

圖 3　檢視邏輯並執行抓取

資料來源：擷取 Web Scraper 應用程式畫面，作者自行繪製。

三、整理資料

運用通用稽核模組之「資料整理」功能，將所擷取訂房網站資料及臺灣旅宿網取得之旅宿名稱及地址（市縣、鄉鎮市區、里、鄰、郵政區號、中英文地址等）進行資料齊一化處理（下頁附表），以利後續資料比對。

四、資料比對

運用通用稽核模組之「跨表比對」功能，將整理後之訂房網站及臺灣旅宿網資料進行二階段比對，步驟如下：

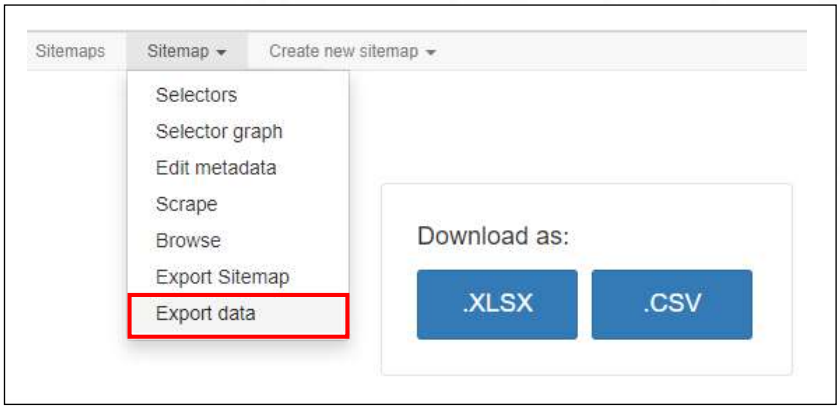
（一）比對旅宿名稱

將訂房網站資料列主檔，臺灣旅宿網資料列次檔，以旅宿名稱爲關聯欄位，市縣、鄉鎮市區、地址分別填入資料欄位，點選「有對應成功」，匯出確認兩者資料是否爲同一旅宿（圖 5），再點選「未對應成功」，匯出並保留此檔案（下頁圖 6）。

（二）比對旅宿地址

將前一步驟之「未對應成功」資料列主檔，臺灣旅宿網資料維持次檔，以旅宿地址爲關聯欄位，旅宿名稱、市縣、鄉鎮市區分別填入資料欄位，點選「有對應成功」，匯出確認兩者資料是否爲同一旅宿，再點選「未對應成功」，匯出疑似未合法登記之旅宿資料。

圖 4　匯出資料



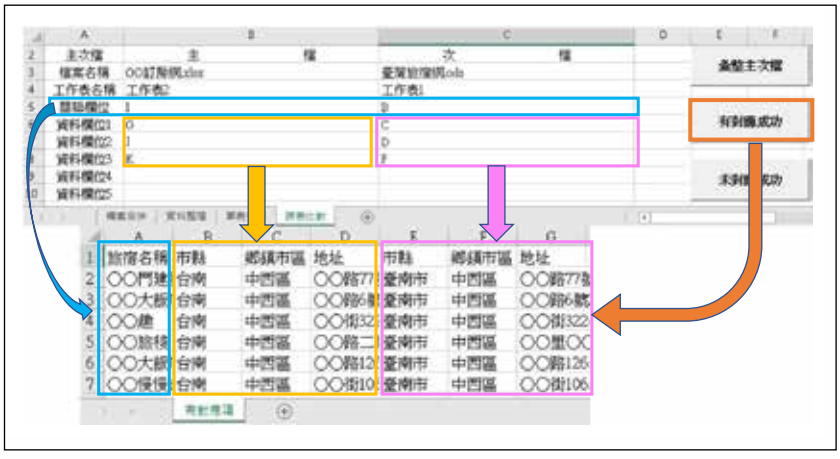
資料來源：擷取 Web Scraper 應用程式畫面，作者自行繪製。

附表　齊一化處理後資料

旅宿名稱	市縣	鄉鎮市區	地址
○○大飯店	臺南市	中西區	○○路 1 號
○○休閒會館	臺南市	白河區	○○街 2 號
○○商旅	臺南市	永康區	○○一路 3 號
○○民宿	臺南市	安平區	○○路 4 巷 5 弄 6 號

資料來源：作者自行整理。

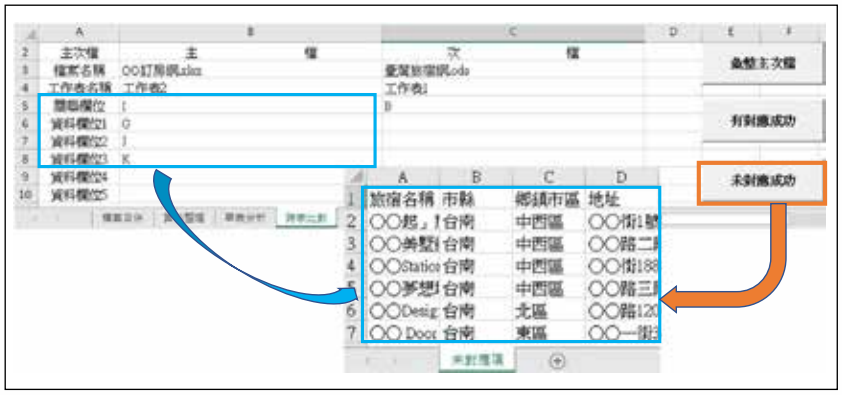
圖 5　比對旅宿名稱流程示例（1）



資料來源：擷取「通用稽核模組」執行畫面，作者自行繪製。

論述 》管理 · 資訊

圖 6　比對旅宿名稱流程示例（2）



資料來源：擷取「通用稽核模組」執行畫面，作者自行繪製。

肆、稽核發現與結論

經運用 Web Scraper 及「通用稽核模組」輔助稽核結果發現，國內有多家旅宿業者未列於臺灣旅宿網之合法登記名單，疑似爲未合法登記經營之旅宿業者，依據發展觀光條例第 55 條之 1 規定，未依規定領取營業執照或登記證而經營觀光旅館業務、旅館業務或民宿者，以廣告物、電腦網路或其他媒體等，散布、播送或刊登營業之訊息者，可處罰鍰。

建議主管機關就上述疑似未合法登記之旅宿業者加強查核取締，並輔導未合法登記之旅宿業轉型爲合法旅宿或伺機退場，進而確保消費者住宿安全與權益及保障合法旅宿業者之經營權，強化國內旅宿業管理效能。

伍、結語

當前政府施政致力於數位轉型，各機關行政業務亦朝向系統化管理及電子化處理，稽核人員面臨業務資訊數位化，如能以電腦技術輔助稽核工作將可提升稽核效率，惟多數稽核人員非資訊專業背景，應用電腦稽核技術須耗費大量學習成本，且過程尚存諸多困難與挑戰。

基於上述考量，本文先擇定適當且便捷之爬蟲工具，引導稽核人員透過圖像化而非撰寫程式之方式擷取網站目標資料，再透過行政院主計總處提供之稽核模組進行資料整理及跨表比對取得稽核結果，大幅降低資料錯誤率及提升資訊有用性。稽核人員倘能靈活運用科技工具及培養資訊素養，便能將稽核作業化被動爲主動，適時於網絡海量資料中有效率地發掘異常資訊，俾利機關決策之參據，同時達到產業永續發展之目標，爲整體社會帶來正面影響，創造多贏局面。

註釋

1.「通用稽核模組」操作方法請詳行政院主計總處官網 / 主要業務 / 政府內控與內稽 / 內部控制監督作業範例。❖