論述 》統計 · 調查

光學閱讀辨識系統在普查業務的應用概況

行政院主計總處為處理量大繁雜之普查資料，爰自民國 88 年起開發「光學閱讀辨識系統（OCR）」，並持續應用於各次普查進行普查表資料之輸入、校登及檢誤等作業，有效提升資料處理時效及品質，並節省大量人力及經費，充分發揮普查 e 化效益。

陳建名、江文瑩（行政院主計總處國勢普查處科長、視察）

壹、前言

為推動統計調查資料處理 e 化作業，行政院主計總處（以下簡稱本總處）爰參酌世界各國作法，於民國 88 年與工業技術研究院合作辦理「普查與抽樣調查應用光學閱讀辨識系統（Optical Character Recognition, OCR）研究計畫」，並首度應用於 89 年戶口及住宅普查，大幅縮短普查表紙本資料電子化時間，提升資料品質，並減少作業人力及經費；後持續發展通用性表單及自動註號功能，加強辨識效率及作業流程優化，並應用於各次普查作業，有效提升資料處理時效及品質，彰顯普查 e 化作業效能。

貳、各國光學閱讀辨識系統運用情形

隨資訊科技發展，運用網路填報、電腦輔助調查及公務檔案連結等取得資料的方式，在統計調查方法上漸顯重要，但多數國家仍採傳統問卷方式辦理普查，無論採行面訪、留置填表、郵寄問卷等方式，如何提升回表資料處理作業效率及品質，實係各國努力研究改進的工作。早期光學符號辨識（OMR）居技術主流，迄 1990 年代隨硬體設備與辨識方法進步，OCR 技術快速發展，其不僅可辨識符號及字元，更可將表件以掃描後影像儲存，節省表件存放空間且便於管理及重覆調閱，有效改善後續資料檢誤及更正作業效率，爰廣為先進國家（地區）所採用，茲舉

數例如下：

一、日本

於 2000 年人口普查首次採用 OCR 技術處理 5,600 萬張問卷，平均辨識速度每小時 5,600 張，辨識正確率達 98.4%，錯誤及無法辨識率分別低於 0.3%及 1.3%，資料處理作業效率大幅提升；2010 年普查採用郵寄問卷、實地訪查及網路填報等方式辦理普查，實地訪查之普查表仍採 OCR 掃描方式整理表件。

二、美國

2000 年人口普查採用 OCR 技術進行勾選及文數字填答資料之辨識作業，OCR 辨識正確率達 99.9%，資料輸入正確率 97%；2010 年人口普查則於實地訪查及未回表追蹤期間，提供訪問員使用具全球衛星定位功能之數位化個人助理（PDA），取代部分 OCR 普查表。

三、香港

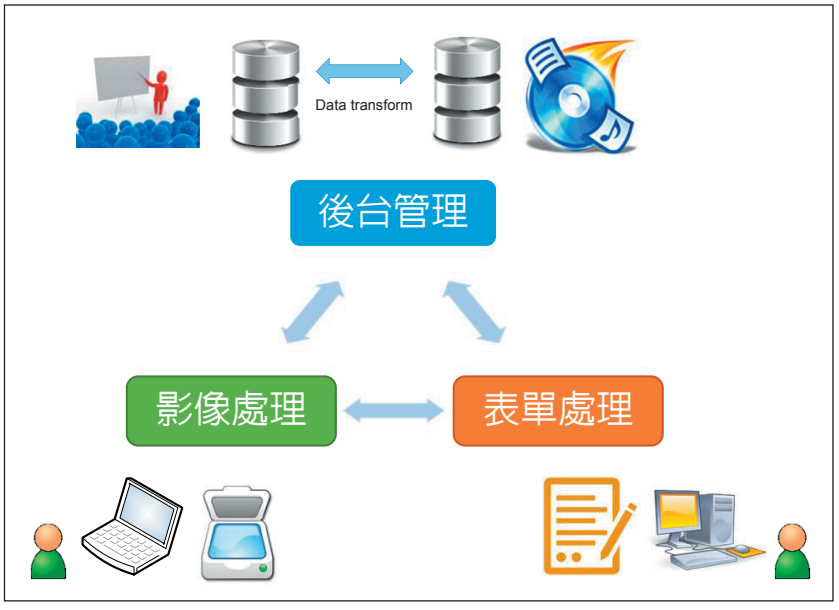
於 2000 年人口普查採用 OCR 技術進行資料審核、校登及邏輯性檢查等作業；2006 年普查增加網路填報（Electronic Questionnaires, e-Q）方式，約有 2%的受訪戶運用此方式完成問卷，其餘表件仍採 OCR 方式辦理；2011 年普查採網路填報及實地訪查方式辦理，後者回表仍採 OCR 掃描表件。

參、我國光學閱讀辨識系統

我國 OCR 系統歷經多次普查作業更迭及改進，目前軟硬體有 Windows 伺服器、Kodak i1860 Fujitsu fi-6770、Kodak9500/7500 高速掃描器、Kodak 1500 平台掃描器及 HS-SQL server。整體架構依實際作業需要可區分多個處理單元，每單元均具備獨立之工作站與伺服器，以分散式處理普查表影像；並採取平均工作量原則，各處理單元均能獨立運作，不影響系統整體執行效能。

本系統具有後台管理、影像處理及表單處理三大主要功能（附圖），每一功能下分別包括 3 個子系統，系統功能說明如下：

附圖　系統處理單位組成圖



資料來源：行政院主計總處國勢普查處。

一、後台管理功能

包括作業管理、資料轉換及光碟製作等 3 個子系統，管理者可藉由後台管理功能，適當分派工作量及監控作業進程，進行系統效能分析，管理及維護 OCR 系統資料轉換及後續光碟備份等功能。

二、影像處理功能

包括影像掃描、影像瀏覽及影像查詢等 3 個子系統，普查表經人工整理後即進行影像掃描，若有定位失敗、歪斜、影像缺損或影像出現線條、框線等品質不佳情形，即予刪除重新掃描，系統提供使用者依普查編號、判定編號、統一編號等索引值查詢影像。

三、表單處理功能

包括表單閱讀、文字校登及問題表單處理等3個子系統，係截取普查表影像中已設定之欄位進行辨識，未能成功辨識之符號、字元、文字等，即移至檢誤校登作業程序，若未能完成文字模式之校登資料，或於字元與檢誤模式校登過程中出現錯誤之資料，則採異常件處理流程處理重新作業。

肆、光學閱讀辨識系統近期運用情形

我國於 89 年戶口及住宅普查使用 OCR 系統後，歷次普查均運用該系統執行資料處理作業，惟爲因應資訊技術提升及相關作業需求，發揮系統效益，歷經多次改進。因重要普查資料涵括中文，爲提高辨識能力，爰於 92 年委外開發「多專家文字辨識組合技術」，建置行職業中文詞庫，針對本總處人力資源調查之行職業代碼進行轉換推導，目前行職業 2 碼推導正確率已達 9 成。另爲使各普查均能共用 OCR 軟硬體設備，爰於 98 年委外辦理通用性表單系統升級作業，對於不同普查表單，僅需請廠商協同調整系統表單模組，以達到資源共享，設備充分利用之目的；迄 104 年增設普查表遮罩功能，及系統更新作業，以解決未來系統升級後客戶端作業相容性問題。

因持續精進系統功能，作業效能大幅提升，以近期 2 次普查爲例，100 年工商及服務業普查， 20 個工作天即完成 105 萬張普查表掃描校登作業，字元校登率精進至 0.30%；104 年農林漁牧業普查於 17.5 天內即完成約 93 萬張普查表單之掃描校登作業，字元校登率更達 0.07%（附表）。

爲發揮系統效益，爰將 104 年農林漁牧業普查大部分檢誤作業納入系統執行，俾利後續編製統計報表及分析撰擬作業，主要運用情形說明如下：

附表　歷年普查 OCR 字元校登一覽表

	99 年 人口 普查	99 年 農林漁牧業 普查	100 年 工商 普查	104 年 農林漁牧業 普查
掃描量	210 萬張	95 萬張	105 萬張	93 萬張
字元校登率 （校登量/辨識量） *100	0.61%	0.22%	0.30%	0.07%

資料來源：行政院主計總處國勢普查處。

一、精進普查資料檢誤功能，擴大系統效益

往年普查資料僅利用 OCR 系統執行資料掃描及校登工作，檢核作業均另行撰擬檢誤程式額外處理；經改進後，顯著提升普查資料掃描、校登至檢誤等作業之一貫性，使系統使用效益最大化。

二、即時影像查詢，快速調閱資料

早期資料檢核作業，須調閱紙本普查表供檢誤使用，OCR 系統將普查表影像以電子檔形式儲存，藉由鍵值即可查閱該表相關資料，資料調閱便利、快速且不占空間；至另行開發之「共通性普查資料檢誤系統」，用於處理較複雜之檢誤條件，並介接 OCR 影像檔，可線上即時調閱處理，增進作業效率。

三、提供普查表遮蔽功能，提升資訊安全

為加強資訊安全，於 104 年開發 OCR 影像遮罩功能，提供自行定義表單遮罩位置之介面設計工具，每張表可設定至少 3 種顯示類別，每種表單可遮蔽至少 5 個位置，作業人員在進行資料校登時，依不同權限及實際需求可遮蔽普查表中特定敏感資料，以提升普查資料隱蔽性。

伍、未來精進方向

OCR 系統雖歷經多次改進，惟綜合各項作業問題，未來改進及努力之空間仍存，歸納如下：

一、提升辨識效能及自動註碼推導

為利機器正確辨識普查資料，表件須保持整潔及書寫工整，造成調查人員的負荷，未來應善用資訊技術，促進填表規範及辨識效能的平衡性，減輕填表負擔。另現行系統建置之「多專家文字辨識組合技術」，行職業 2 碼代碼推導正確率達 9 成以上，至 4 碼代碼推導的部分，將持續進行效率及確度修正，以改進系統自動註碼功能。

二、精進硬體設備

現行 OCR 系統硬體設備歷經 99 年人口及住宅普查、99 及 104 年農林漁牧業普查、100 年工商及服務業普查與多次專案調查作業，已超過使用年限，作業可靠性明顯降低，為維持系統穩定性，應於定期檢視 OCR 系統，並進行汰舊換新評估。

三、增加利用率

目前 OCR 系統設備用於普查及專案調查處理資料，利用率仍有提升空間，其文件掃描及影像管理功能，可擴展應用至各類文件的儲存與管理。

陸、結語

OCR 系統確實能有效提升普抽查業務工作效率，節省大量人力及經費。近年雖然許多國家運用多元資訊技術辦理普查 e 化作業，如結合網路填報、電腦輔助面訪（CAPI）等技術，直接輸入資料，簡化後續資料處理作業，增進資料品質及減省經費，但 OCR 系統掃描、校登及檢誤功能仍有其不可取代的優勢，故如何針對不同資料，採取並整合有效之作業模式，使普抽查 e 化效益最大化，厥為當前發展及努力之方向。❖