數據新聞的實踐

如今是數據時代。

天文台說剛過去的大除夕是二十九年來最寒冷的大除夕，因為他們比較了二十九年來的氣溫紀錄。

政府公布新政策時，往往說因有數據支持。大家聽了都會問：是甚麼數據？政府不如公開那些數據，好讓大家分析，才決定是否支持這政策。

而很多新聞，也是類似這些數據的分析結果。

「新聞的未來，是分析數據。」發明互聯網的 Tim Berners-Lee 爵士三年前這樣描劃新聞未來的發展方向。他發表此番言論之前數天，英國政府公開了政府部門、每項超過 25000 英鎊的開支紀錄，他有感而發說：記者需要對大量數據有敏銳的觸覺，及懂得利用不同工具去分析這些數據，讓公眾清楚了解這個國家、政府發生甚麼事。

而這數年，外國很多大學及訓練記者組織都舉辦課程，培訓記者分析數據，而數據新聞學（Data Journalism）亦成為新聞學的新趨勢。

港大數據新聞研究項目

我自去年初，參與香港大學新聞及傳媒研究中心有關數據新聞的研究項目（http://datalab.jmsc.hku.hk），我們開發了給公眾使用有關土地規劃的資料庫，也做了幾個新聞報導。過程中，我深深感受到，數據新聞不單屬於新聞行業或新聞專業培訓，而是全民社會的，因為涉及政府的透明度、向民眾問責，是大家的知情權，此乃為重要的公民權利。

港大的數據新聞研究項目，首個階段是有關香港的土地資源如何使用，土地資源茲事體大，影響經濟、政治、民生，也是發掘新聞的主要泉源。這方面的資料，分佈幾個政府部門及法定委員會，包括城市規劃委員會、地政總署及屋宇署，大部份都是公開的，大家都可以上網查看。

我在一零年三月就在本刊提及城規會紀錄的參考價值：

「有很多公共紀錄，都和你擁有的、計劃購置的、及居住的房地產，息息相關，關係到的是樓價、租金及居住環境。原來所有房地產都有其指定的用途，好像住宅、酒店、餐廳、寫字樓等等，可建樓層高度也有規定，假如對面大廈的業主要改建起多幾層樓，或你樓下鄰居要（合法地）把某單位由住宅改為骨灰龕，或有人要把新界買少見少的魚塘填平，改作貨櫃車場，其實公眾都可以預先知道，並有權就此提出意見。

「以上情況，業主都必要向城市規劃委員會申請。城規會內部除了有討論外，還需諮詢公眾，收集意見。諮詢期間，業主向城規會提出申請的主要文件，包括圖則，公眾都可查閱，而公眾所提出的意見概要，也是公開的。查詢這些資料，公眾可以前往城規會的辦事處，而瀏覽城規會的網頁（www.tpb.gov.hk），就更方便，過去廿年的申請紀錄，也可找到。」

可是，城規會的紀錄都以pdf存檔，檔案以「閱讀文件」為主，例如，某一申請個案的幾頁文件，大家都可以讀到及下載到。但若果我們想了解某一個區在特定期間內總共有多少個申請個案、申請類別是甚麼、某類別的成功率等，或各區申請情況的比較，只靠那些一頁頁的文件，我們就要人手逐一數算和紀錄，大家可以想像，過程費時又會有錯漏。

pdf 檔 Vs csv 檔

於是我們向城規會要求索取申請個案資料的另一個版本：CSV，即是申請個案的資料，以試算表 excel 存檔，好像地區、申請類別、城規會決議、決議日期等等，變成一個一個欄目，而每項申請個案的資料，就按欄目填在表格上。

如果大家對以上兩個檔案形式不太清楚，不妨看看自己的網上銀行信用咭賬單，一些銀行提供兩種賬單檔案給客戶查閱，一個是 pdf，另一是 csv，pdf 檔就像一份份印刷文件，上面的資料你不可以移動；而csv檔就是一份試算表表格，你可以利用試算表功能計算所有簽賬的總和、平均值、中位數等，你也可以把全年所有餐廳的簽賬儲存在一個新試算表上，再利用運算功能，你就可以計算到全年你花在餐廳的消費是多少。

而對於我們調查過去二零零九年一月至零三年三月全港共五千多個的規劃申請個案，有了這個試算表檔，分析就容易很多了。我們抽出當中涉及農地改變用途的七百多個申請個案，之後，我們仔細研究那幾百個個案，再運用試算表的篩選、排序及樞紐分析表（pivot table）功能，就可得出以下幾個重要分析結果：

1. 最多申請個案的首五個地區（括號內為數目）：錦田南（117）、龍躍頭及軍地南（84）、坪輋及打鼓嶺（79）、八鄉（68）、錦田北（64）。

2. 申請最多的土地用途（括號內為數目）：新界豁免管制屋宇（279）、露天存放（178）、停車場（37）、康樂場所（37）、訓練教育場所（33）；露天存放中，存放車輛最多（66）。

3. 獲批准的規劃許可所申請的土地用途，地盤佔地面積的五大：露天存放（161975 平方米）、康樂場所（86196 平方米）、屋宇（52894 平方米）、停車場（35692 平方米）、訓練教育（19940 平方米）。

地圖定位及搜索

我們還把申請個案定位在地圖上，加上搜索功能，大家就可在這個數據庫看到申請個案的分布，也可以按自己的選擇，搜索每區或每種土地用途的數目和分佈。

除了城規會紀錄，屋宇署也提供很多有用資料，每月該署會在網上（www.bd.gov.hk）發布多項資料，其中有三項跟新樓地盤有關：獲批圖則、獲批施工同意書及獲發入伙紙的數量、地盤地址、樓宇類別及地盤申請人及建築師等。

大家經常走過一些地盤，但你未必知道地盤會建甚麼或者發展商是誰，屋宇署公布的資料，本來是可以為大家解答以上問題，但問題是資料也是 pdf 檔，且是按署方批發文件的日期排列，不是按分區或分樓宇種類，對民眾查閱並不方便。

港大的研究項目，於是把資料轉換為可運算數據，及在地圖定位，目的是方便民眾按自己選擇查閱。

政府公開資料政策

正如上述所言，能否從資料中發掘新聞，或者讓民眾了解土地資源如何使用，很大程度視乎那些資料是否可用電腦運算及分析的，因為我們面對的，不是幾十個數據那麼簡單，而往往是數以千或萬計的，若我們要動手計算，或者把資料自行輸入在電腦運算表格，大費周章之餘，又難免有錯漏。

即使有不少網上免費工具，可以把 pdf 檔案轉換為文字檔或者試算表，也並不是全部穩妥。

因此，最好當然是資料本身是可用電腦運算形式。目前政府很多部門都在網上上載很多資料供市民參考，但大多數是以 pdf 或類似的檔案，大家在網上閱讀、列印是可以的，但要運算、分析，就很困難。

我曾問某些政府部門：在他們網站的一些

資料（以 pdf 檔形式），可否給我相同資料的數據檔？負責有關方面的官員知我是為研究而來，都答應給我。而以我所了解，部門其實早已有這些數據檔，所以他們給我一份，他們不用花很多時間和資源。

這一點也不出奇，正如我們大家一樣，政府用試算表儲存數據，其實是普遍不過的事，問題是，為何政府不把這形式的數據放在網站，好讓公眾（包括記者在內）可以使用？

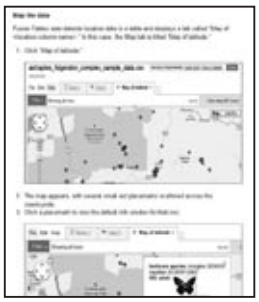
申訴專員公署上月宣布，會主動展開調查公開資料守則及政府管理檔案的手法，我期望無論是政府部門，或是申訴專員公署，研究那些資料需要公開給民眾之外，同時要兼顧公開資料的形式，是否便利民眾使用的。

媒體人手及資源配合

媒體要多做些數據新聞項目，除了有賴政府是否發放可運算的形式的數據外，媒體還要在人手及資源上，作出一些配合，以下是我的三個建議：

一. 電腦運算專才

雖然很多人普遍對電腦運算軟件已有一定認識，但當中一些巧妙的運算、篩選、分類、排序等，並不是三時兩刻就可掌握及好好運用，我去年初參與的「美國調查報導記者及編輯」（Investigative Reporters and Editors，簡稱 IRE）舉行的全國電腦輔助調查新聞大會，四天的會期中有多場 excel 訓練班，每班都坐滿人，廿多歲初出道的、至五、六十歲的大記者、大編輯都在課堂上專心學用excel。



所以要從事新聞專業新領域的記者，需在電腦運算軟件方面，下點功夫，依我及其他數據新聞記者的經驗，學好 excel 是最基本的，而 Google fusion tables 包含更多運算功能兼備地圖定位，能夠學好，當然最好。

媒體主管也可以聘用這方面的專才，和記者合作，各施所長。同時，媒體的網頁也需要重新設計，提供更多可跟讀者互動的搜索、地圖等介面，所以有關電腦和網頁的專才，未來媒體不可缺乏。

在進一步，新聞工作也需要編程員（programmer）懂得挖掘和處理大量的數據，我的研究團隊就有一位編程員，把政府的 .pdf 文檔改編成為可供電腦軟件自動化分析的檔案。

我認識有大學生看準了行業大趨勢，而選擇了讀新聞及電腦雙學位。

二. 地理資訊專才

雖然現在流行的網上及手機地圖搜索，幾乎沒盲點，但很多沒開發的地方，沒有讓人識別的地標，最強勁的地圖搜索器也無法準確把要找的地方定位。以香港為例，新界農村及離島就有不少地圖盲點。

我進行港大的數據項目時，一名同事是地理資訊系統（Geographical Information Service）的專業人士，她對獲取國際通用的經緯度或坐標，有嫻熟的技巧及充沛的資源，對我們把分散在很多不知名地方的農地規劃申請個案定位，提供了很重要的解決方案。

很多新聞和地區結合，往往會是另一條大新聞的頭緒，例如，為甚麼大埔汀角的農地改建村屋的申請個案最多受城規會否決。因此，地理資訊方面的專才，也是未來新聞發展的重要部份。

三. 記者的思考模式

有人爆料，當然是做大新聞的方法，但記者其實也可以從資料中找出一些趨勢、地區或時間上的分佈，例如，某些案件多數會在某地區或者某便利店外發生，這已是新聞，再配合記者採訪，例如訪問警察，新聞就更深入。

記者應視資料和數據是消息源，是提供新聞的重要線索。

陳貝琼
香港大學新聞及傳媒研究中心名譽講師、
「促進政府公開資料」(Open Government)
研究計劃研究員