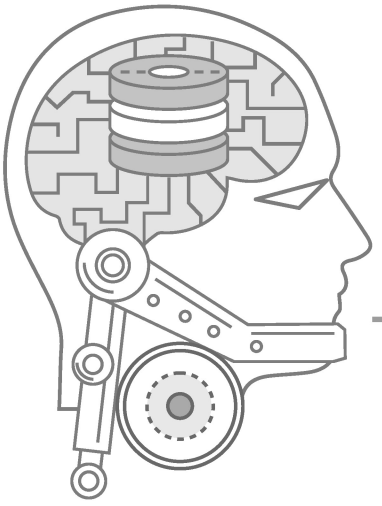
CHAPTER 1



認識 HTML、CSS 和網路爬蟲

1-1 網路爬蟲的基礎

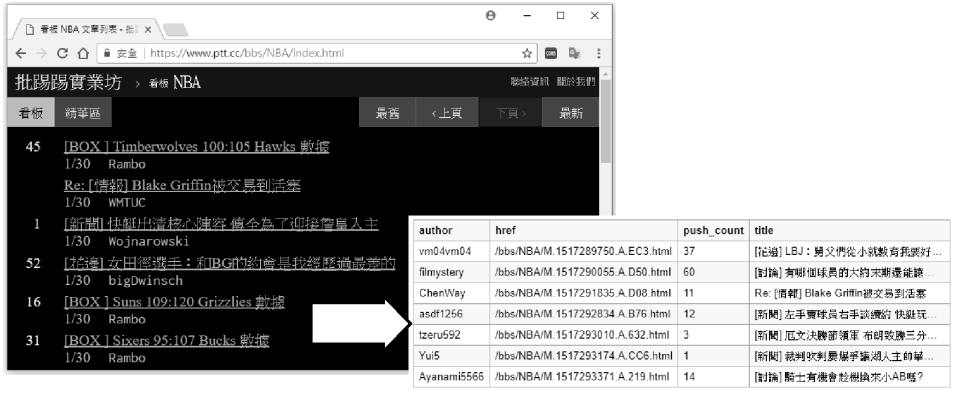
「網路爬蟲」（Web Scraping）是一個從 Web 資源擷取資料的過程，我們可以使用瀏覽器檢視所需資料的網頁後，就可以直接從 Web 資源取得所需資訊，而不是使用網站提供的現成 API 存取介面。

網路爬蟲或稱為網路資料擷取（Web Data Extraction）是一種資料擷取技術，可以讓我們從 Web 網站的 HTML 網頁取出所需的資料，其過程包含與 Web 資源進行通訊，剖析文件取出所需資料和將資料整理成資訊，即轉換成所需的資料格式。

認識網路爬蟲

網路爬蟲（Web Scraping）是一種針對目標 Web 網站自動擷取資訊的技術，雖然我們可以手動自行使用複製和貼上方式來收集和擷取資訊，但是網路爬蟲可以自動幫助我們收集和擷取資訊。

一般來說，Web 網站內容很多都是從資料庫取出結構化資料來產生網頁內容，但是因為網站內容編排的版面配置，在網頁會新增標題、註腳、選單、巡覽列和側邊欄等其他資訊的區段，造成網頁內容反而變成一種結構不佳的資料。網路爬蟲可以讓我們從 Web 網站取出非表格或結構不佳的資料後，將之轉換成可用且結構化的資料，如下圖：



上述圖例可以看出直接從 PTT NBA 板的網頁中爬取資料，然後轉換成結構化資料（即使用表格呈現的資料），這就是網路爬蟲的目的：轉換 Web 網站的特定內容成為結構化資料，例如：轉換輸出成關聯式資料庫、Excel 試算表或 CSV 檔案等。

Memo

以電腦資訊科技來說，網路爬蟲是一種反向工程，HTML 網頁的資料來源大多是伺服器資料庫的記錄資料，一種結構化資料，HTML 網頁內容將這些資料轉換成非結構化資料來呈現，網路爬蟲是將 HTML 網頁內容的非結構化資料，再度轉換成結構化的表格資料。

什麼不是網路爬蟲

在實務上，並非所有從網路取得資料的操作都可以稱為網路爬蟲，如果沒有資料擷取操作過程，可以直接從 Web 網站取得機器可讀取的資料，這些操作並不能稱為網路爬蟲，例如：

- 直接從網站下載資料檔：有些網站已經提供現成結構化資料的檔案可供下載，例如：Excel 檔案、CSV 檔案或 JSON 檔案等。
- 應用程式介面 Web API：很多公司都會提供 Web 基礎的 API 介面，例如：REST API，我們可以透過 REST API 來下載結構化資料，例如：JSON 資料。

Memo

請注意！上述應用程式介面 Web API 如果是公開 API，基本上，並不能算是網路爬蟲，如果不是公開的 API，而是自行透過分析瀏覽器的 HTTP 請求來找出的 Web API，從廣意來說，也可稱為網路爬蟲。

1-2 了解瀏覽器瀏覽網頁的步驟

網路爬蟲就是一個從 Web 網頁擷取資料的過程，可以讓我們從瀏覽的網頁中擷取出所需的資訊，在實務上，我們一定需要先從瀏覽器的網頁內容中看到目標資料後，才能使用網路爬蟲將目標資料擷取回來。

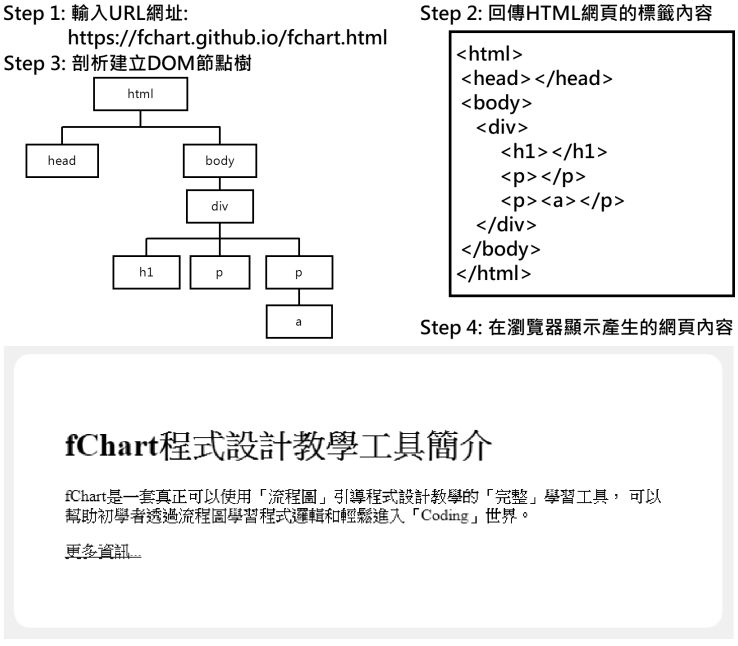
相信每一位讀者都曾在瀏覽器輸入 URL 網址來瀏覽網頁，這些看起來十分簡單的操作，就是建立網路爬蟲第一步，打開瀏覽器的窗口找到你要的資料，然後擷取出有興趣的資料，其基本步驟如下：

1. 在瀏覽器輸入 URL 網址就是向 Web 伺服器送出 HTTP 請求，這是 GET 請求（即取得資源的請求），使用的是第 1-3-1 節的 HTTP 通訊協定。

2 Web 伺服器依據瀏覽器送出的 HTTP 請求來回應內容至瀏覽器，通常就是 HTML 網頁。

3 瀏覽器接收到伺服器回應的 HTML 網頁後，將網頁內容剖析建立成樹狀結構，每一個 HTML 標籤是一個節點，這就是 DOM（Document Object Document）。

4 瀏覽器依據 DOM 產生內容，成為我們在瀏覽器看到的網頁內容。



URL 網址

基本上，在瀏覽器輸入的 URL 網址是由幾個部分所組成，例如：fChart 教學工具簡介頁面的測試網頁，其 URL 網址：

```
https://fchart.github.io/fchart.html
```

上述 URL 網址的目的是指出你需要取得哪一個 Web 伺服器的哪一個資源，資源有很多種，最常見的就是 HTML 網頁和圖檔，詳細 URL 網址的說明請參閱第 1-3-2 節，如下：

- fchart.github.io：Web 伺服器的名稱。
- fchart.html：資源名稱，即 HTML 網頁的檔案名稱。

HTML 網頁

Web 伺服器在接收到瀏覽器的 HTTP 請求後，就會依據請求回應 HTML 網頁內容，即回應資源，例如：當在瀏覽器輸入 URL 網址 https://fchart.github.io/fchart.html，可以在瀏覽器看到回應的網頁內容，如下圖：

上述網頁內容是瀏覽器已經剖析 HTML 網頁+CSS 樣式產生的網頁內容，可以看到在中間區塊顯示的網頁內容：標題、段落和超連接。

我們可以檢視網頁的原始程式碼，請在瀏覽器的網頁內容上，點選滑鼠【右】鍵，執行快顯功能表的【檢視網頁原始碼】命令，可以看到 Web 伺服器回傳的 HTML 網頁內容，這是 HTML5 標籤，在<style>標籤中是 CSS 樣式（HTML 標籤的外觀描述），如下圖：

```
<!doctype html>
<html>
<head>
   <title>fChart程式設計教學工具簡介</title>
   <meta charset="utf-8" />
   <meta http-equiv="Content-type" content="text/html; charset=utf-8"/>
   <style type="text/css">
   body {
       background-color: #f0f0f2;
   }
   div {
       width: 600px;
       margin: 5em auto;
       padding: 50px;
       background-color: #fff;
       border-radius: 1em;
   }
   </style>
</head>
<body>
<div>
   <h1>fChart程式設計教學工具簡介</h1>
   <p>fChart是一套真正可以使用「流程圖」引導程式設計教學的「完整」學習工具，
   可以幫助初學者透過流程圖學習程式邏輯和輕鬆進入「Coding」世界。</p>
   <p><a href="https://fchart.github.io">更多資訊...</a></p>
</div>
</body>
</html>

```

上述回應內容是 HTML 標籤和 CSS 樣式組成的 HTML 標籤碼。

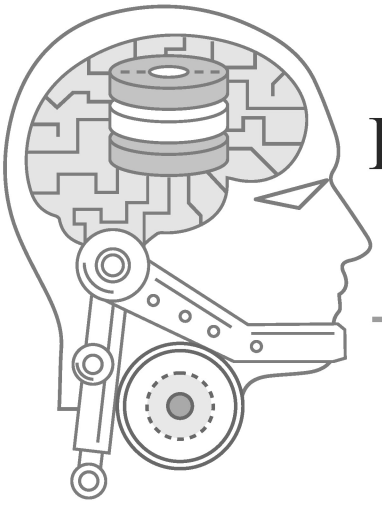
樹狀結構的節點

瀏覽器首先會將 Web 伺服器回傳的 HTML 標籤，建立成樹狀結構的節點，即 DOM 節點樹，這是一種階層結構的標籤。因為 HTML 標籤是成對的：「<標籤名稱>...</標籤名稱>」（在結尾標籤名稱前有「/」），而且在標籤中可以擁有其他 HTML 標籤，如下：

```
<div>
    <p></p>
    …
</div>
```

在上述 標籤中擁有<p>標籤， 標籤是父標籤；<p>標籤是子標籤，這是巢狀標籤的階層結構。在 DOM 節點樹就是 HTML 標籤的階層結構，每一個 HTML 標籤是一個節點（Node）。

CHAPTER 14



Power Automate Desktop 辦公室自動化

14-1 自動化檔案與資料夾處理

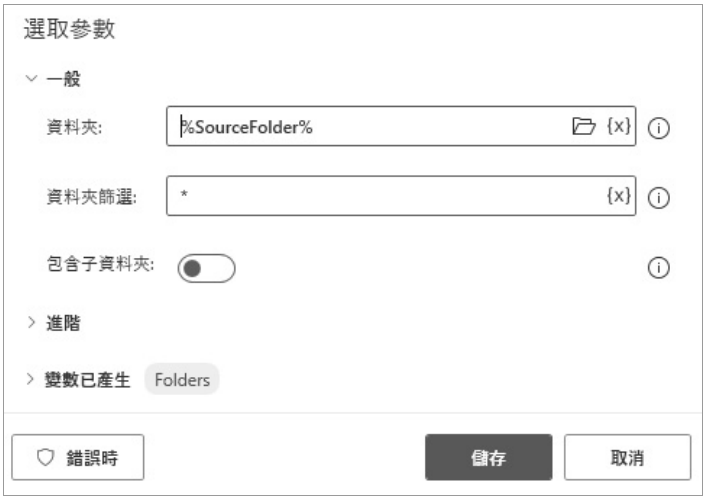
Power Automate Desktop 檔案與資料夾處理的相關動作是位在【檔案】和【資料夾】分類，可以執行資料夾與檔案的複製、移動、重新命名和刪除等相關的自動化操作。

14-1-1 取得目錄下的檔案和資料夾清單

在【取得目錄下的檔案和資料夾清單】流程（流程檔：ch14-1-1.txt）共有 4 個步驟的動作，可以取得指定目錄下的檔案和資料夾清單，如下圖：

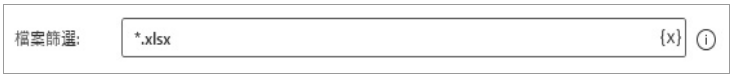
1 【變數＞設定變數】動作可以指定變數 SourceFolder 的路徑是「D:\WebScraper\Ch14\教育訓練成績」（請自行修改路徑）。

2 【資料夾＞取得資料夾中的子資料夾】動作可以取得指定路徑下的資料夾清單儲存至 Folders 變數，【資料夾】欄位是目標路徑，其值可以是變數，或點選欄位後第 1 個資料夾圖示來選擇路徑（第 2 個圖示是選變數），【資料夾篩選】欄是過濾條件，「*」符號表示所有資料夾，開啟【包含子資料夾】可包含子資料夾下的資料夾，如下圖：

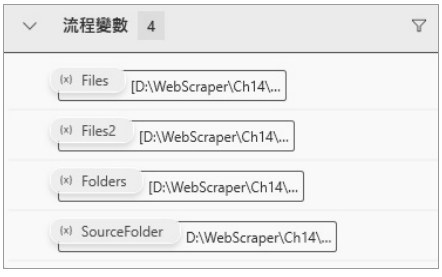


3 【資料夾>取得資料夾中的檔案】動作可以取得指定路徑下的檔案清單儲存至 Files 變數，在【資料夾】欄位是目標路徑，【檔案篩選】欄是過濾條件，「*」符號表示所有檔案，開啟【包含子資料夾】可包含子資料夾下的檔案，如下圖：

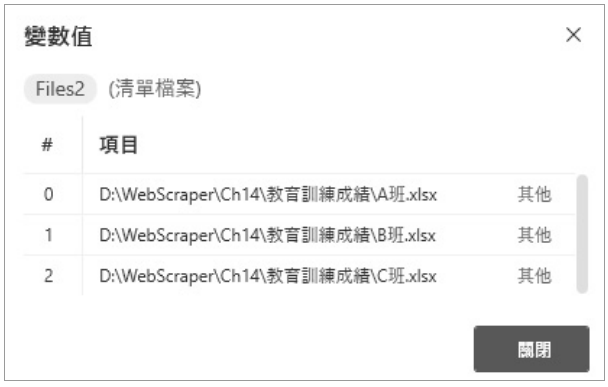
4 【資料夾>取得資料夾中的檔案】動作和步驟 3 相同，檔案清單是儲存至 Files2 變數，檔案的篩選條件是【*.xlsx】，可以過濾取出副檔名是.xlsx 的 Excel 檔案，如下圖：



上述流程的執行結果，可以在「變數」窗格檢視流程變數的值，其值就是取得的檔案和資料夾清單，如下圖：



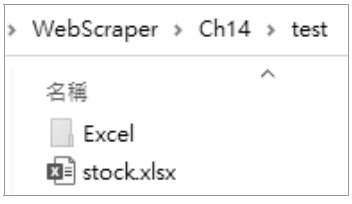
上述 Files、Files2 和 Folders 變數的資料型態是清單，請雙擊變數名稱，例如：Files2，可以看到此清單的項目，每一個項目就是一個 Excel 檔案，如下圖：



上述變數 Files2 因為有篩選，所以只有.xlsx 檔案，如果是 Files 變數，就可以看到.xlsx 和.csv 檔案清單；Folders 變數是資料夾清單。

14-1-2 批次重新命名和移動檔案

Power Automate Desktop 可以建立桌面流程，將整個目錄下的 Excel 檔案重新命名來加上日期後，移動這些 Excel 檔案至另一個全新目錄。首先請開啟 Windows 檔案總管，自行複製「Ch14\examples」目錄成為「Ch14\test」目錄，如下圖：



上述 Excel 檔案是位在「Excel」子目錄，共有【營業額 1~4.xlsx】四個 Excel 檔案，流程可以將這些 Excel 檔名後加上日期後，全部移動至新建的「Ch14\test\Output」目錄。

14-4 實作案例：自動化下載網路 CSV 檔和匯入 Excel 檔

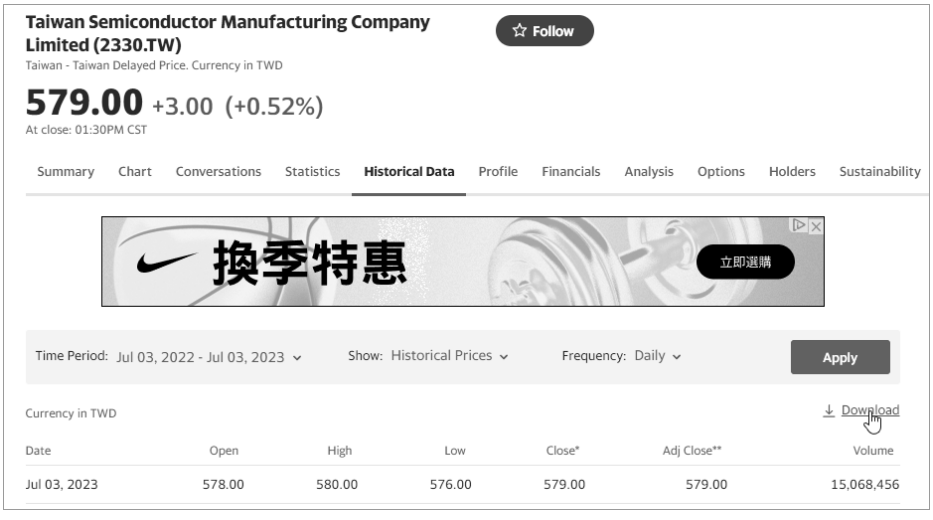
因為目前有很多 Web 網站或政府單位的 Open Data 開放資料網站都提供直接下載資料的按鈕或超連結，除了自行手動下載資料外，只需找出下載的 URL 網址，我們就可以建立 Power Automate Desktop 桌面流程來自動下載 CSV 檔案，並且匯入儲存成 Excel 檔案。

下載美國 Yahoo 的股票歷史資料

在美國 Yahoo 財經網站可以下載股票的歷史資料，例如：台積電網址：

- https://finance.yahoo.com/quote/2330.TW

上述網址最後的 2330 是台積電的股票代碼，.TW 是台灣股市，如下圖：



請在上述網頁選【Historical Data】標籤後，在下方左邊選擇時間範圍，右邊按【Apply】鈕顯示股票的歷史資料後，即可點選下方【Download Data】超連結，下載預設以股票名稱為名的 CSV 檔案。

自動化下載網路 CSV 檔和匯入 Excel 檔：ch14-4.txt

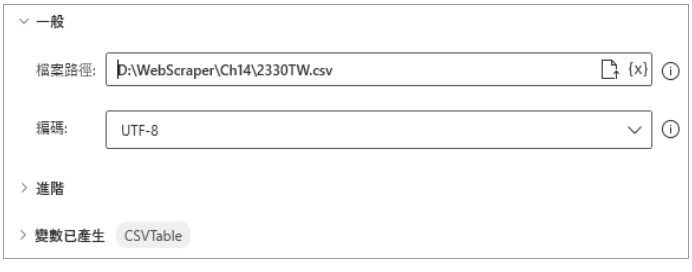
請在 Yahoo 股票資料的【Download Data】超連結上，執行【右】鍵快顯功能表的【複製連接網址】命令，可以取得下載 CSV 檔案的 URL 網址。

在【自動化下載網路 CSV 檔和匯入 Excel 檔】流程共有 8 個步驟的動作，可以下載網路的 CSV 檔和匯入儲存成 Excel 檔案（關於 Excel 操作的進一步說明，請參閱第 15 章），如下圖：

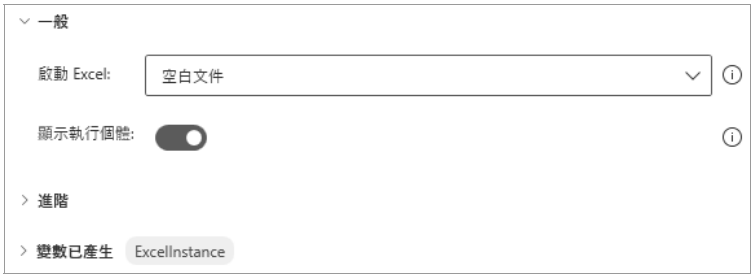
1 【HTTP>從 Web 下載】動作可以使用 HTTP 通訊協定以 URL 網址來下載檔案，如同瀏覽器瀏覽網頁一般，其下載資料是儲存在 DownloadedFile 變數，在【URL】欄位是取得的下載網址，【方法】是【GET】請求，在【儲存回應】檔選【儲存至磁碟(適用於檔案)】來下載檔案，【檔案名稱】欄選指定完整路徑，即可在【目的地檔案路徑】欄指定下載檔案的完整路徑，如下圖：

2 【檔案>等侯檔案】動作是等侯檔案直到檔案已經建立或刪除，在【等候檔案完成】欄是完成條件，檔案建立是選【建立日期】；檔案刪除是選【已刪除】，【檔案路徑】欄就是等待的檔案，以此例是在等待下載檔案的建立，即完成檔案下載，如下圖：

3 【檔案>從 CSV 檔案讀取】動作是讀取 CSV 檔案內容成為 CSVTable 變數的資料表資料，然後才能存入 Excel 工作表，在【檔案路徑】欄是 CSV 檔案路徑；【編碼】欄是 UTF-8 編碼，如下圖：

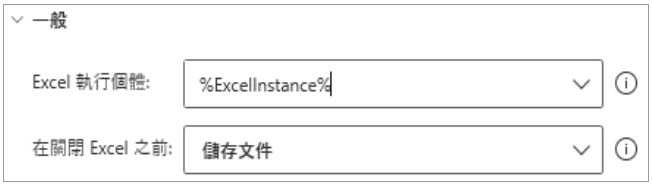


4 【Excel>啟動 Excel】動作可以啟動 Excel 開啟存在或建立空白的活頁簿，在【啟動 Excel】欄選【空白文件】是建立空白活頁簿，ExcelInstance 變數值就是 Excel 軟體的執行個體，如下圖：



5 【Excel>寫入 Excel 工作表】可以將讀取的 CSV 資料寫入 Excel 工作表，在【要寫入的值】欄就是之前讀取 CSV 資料的 CSVTable 變數，因為是空白活頁簿，【寫入模式】欄請選【於目前使用中儲存格】，如下：

6 【Excel>關閉 Excel】動作是關閉 Excel，在關閉前可以指定是否儲存 Excel 檔案，請在【在關閉 Excel 之前】欄選【儲存文件】，可以在儲存檔案後再關閉 Excel，如下圖：



7 【檔案>移動檔案】動作是移動流程建立的 Excel 檔案，因為步驟 4 是開啟空白活頁簿，預設是儲存在登入使用者的「文件」目錄，檔名是【活頁簿 1.xlsx】，在【要移動的檔案】欄的路徑中，hueya 就是使用者名稱，

請自行修改成你的使用者名稱，【目的地資料夾】欄是移動的目的地路徑，如果檔案已經存在就覆寫檔案，如下圖：

8 【檔案>重新命名檔案】動作是將移至「D:\WebScraper\Ch14」資料夾的【活頁簿 1.xlsx】檔案改名成為【2330TW.xlsx】，如下圖：

上述流程的執行結果，可以在「D:\WebScraper\Ch14」資料夾看到使用下載 CSV 檔案 2330TW.csv 匯入建立的 Excel 檔案：2330TW.xlsx。