資料品質提升機制及常見錯誤樣態說明

王向榮

關於我

王向榮 Ronny Wang

現為李慕約公司共同創辦人，g0v 零時政府新聞小幫手、求職小幫手、台灣公司資料、開放政治獻金等專案發起人

行政院、國發會、陸委會開放資料諮詢委員

作品集 https://ronny.tw/data/

資料 和 資訊

教育部字典怎麼說...

資訊：電腦上指對使用者有用之資料和訊息的總稱。以別於未經處理過的資料
資料：計算機中一切數值、記號和事實的概稱。通常指未加以處理者。

差別在哪？

以肉絲炒麵為例...

資訊像是肉絲炒麵

資料像是 肉絲+麵條

以政府資訊來說...

人事行政局辦公日曆表　　資料

date	name	isHoliday	holidayCategory	description
2017/4/1		是	星期六、星期日	
2017/4/2		是	星期六、星期日	
2017/4/3		是	放假之紀念日及	全國各機關學校放假一日。
2017/4/4	兒童節、民族掃墓節（清明節）	是	放假之紀念日及節日	全國各機關學校放假一日(兒童節與民族掃墓節同一日時，於前一日即四月
2017/4/8		是	星期六、星期日	
2017/4/9		是	星期六、星期日	
2017/4/15		是	星期六、星期日	
2017/4/16		是	星期六、星期日	
2017/4/22		是	星期六、星期日	
2017/4/23		是	星期六、星期日	
2017/4/29		是	星期六、星期日	
2017/4/30		是	星期六、星期日	

資訊 vs 資料

- 資訊
  - 有明確目的，經處理過讓目標使用者可得到特定訊息
    - Excel 產生的報表
    - 依特定條件的查詢功能的網站
- 資料
  - 未經處理的原始資料
    - 做報表前收集的數據
    - 供查詢網站使用的資料庫

資料長什麼樣子

在資訊領域，資料可能長什麼樣子

整數：123
小數：3.14159
文字：「白日依山盡，黃河入海流，欲窮千里目，更上一層樓」
數列：[1,1,2,3,5,8,13,21,34]
日期：2018/3/12
經緯度：(23.9756500; 120.97388194)
名片資料：[姓名: 王小明, 電話: 0912345678, 公司: 哈哈公司, 職稱: 專案經理]
電子郵件：[收件者: alice@foo.com 寄件者: bob@bar.com 主旨: 生日快樂 內文:

非結構化資料 vs 結構化資料

非結構化資料

- 我叫阿明，男生，我身高 134 公分
- 我的名字叫小美，我是可愛的女孩，我身高 142 公分
- 我是小華，我是個一百四十四公分的男生

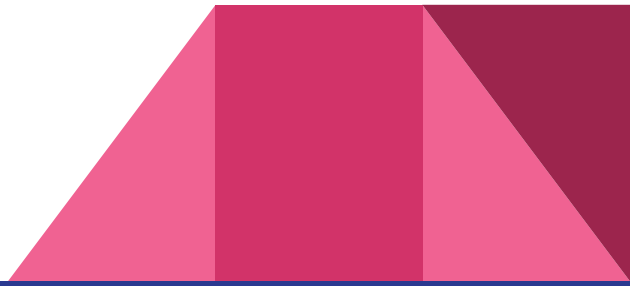
結構化資料

姓名	性別	身高
阿明	男	134
小美	女	142
小華	男	144

資料 vs 資料集

- 資料集 => 資料 的集合
- 集合通常是將相同種類或是相同屬性的東西合在一起
- 相同種類的東西通常可以有共通的屬性

在產生資料集時，或許可以思考看看，這資料集的集合放的是什麼，有哪些屬性是共通的？



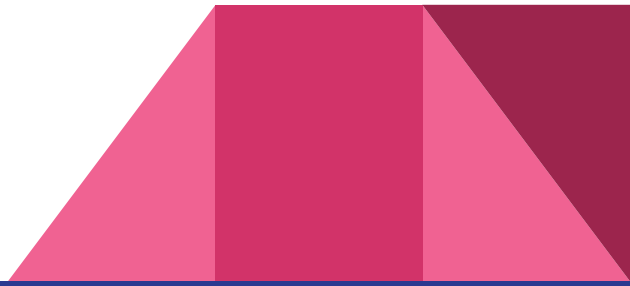
資料品質提升機制

為什麼需要資料品質提升機制

這是個資料越來越有價值的時代...

- 開放資料、大數據、資料視覺化、機器學習、人工智慧...
- 工具越來越多：Excel、Google 試算表、R、Python ...

提供好用的資料是有幫助的



資料品質三面向

- 可直接取得
- 資料易於處理
- 資料易於理解

可直接取得

可直接取得之前，要先可取得...

- 避免打錯網址
- 避免資料不預期被移除

如果資料無法被取得就什麼都別談了...

可直接取得的重要性...

https://docs.google.com/spreadsheets/d/10ah9qtu-IMAVuXyvapcYW19z2TC-U8rQ01TfYbnciEg/edit#gid=0

如果資料是可直接取得的，Google Spreadsheet 可以作到直接載入資料並即時互動

＊若想把這個 Speadsheet 抓去自己玩玩，可以登入自己的 Google 帳號，點選「檔案」-> 「建立複本」

webservice (or API) 怎麼處理...

API 是什麼？

應用程式介面（英語：Application Programming Interface，簡稱：API），又稱為應用編程介面，就是軟體系統不同組成部分銜接的約定

簡單說就是電腦跟電腦之間溝通的管道

（電腦很笨的，要溝通一定要用很精準的方式溝通）

範例一:

查詢起造人為「王小明」的資料

查詢方法:

欄位名稱1=起造人代表人	查詢條件1=王小明

http://build.kinmen.gov.tw/opendata/OpenDataSearchUrl.do?d=OPENDATA&c=BUILDLIC&Start=1&起造人代表人=王小明

API 沒有標準... 但是 API 說明書有...

- 2010 年一家軟體公司 SmartBear Software 公開了一套 Swagger 的開放原始檔工具，幫助開發者可以拿來描述並測試　API
- 2015 年 Linux 基金會與 Google, Microsoft, IBM 等公司 組成了 OpenAPI Initiative
- 2016 年 Swagger 規範被 OpenAPI Initiative 使用，並改名為 OpenAPI Spec

所以 API 現在說起來是有說明書的規範的....

唐鳳政委也正在推廣 OpenAPI Spec

一個 API 會有什麼資訊?

範例一:

查詢起造人為「王小明」的資料

查詢方法:

欄位名稱1=起造人代表人	查詢條件1=王小明

http://build.kinmen.gov.tw/opendata/OpenDataSearchUrl.do?d=OPENDATA&c=BUILDLIC&Start=1&起造人代表人=王小明

如上圖，會有

1. API 的開頭網址：http://build.kinmen.gov.tw/opendata/OpenDataSearchUrl.do?d=OPENDATA&c=BUILDLIC
2. 有相關的參數：
   a. Start=1 ：從第 n 筆開始
   b. 起造人代表人=王小明 ：想要查起造人是誰的資料
3. API 會回傳什麼內容，長什麼樣子？

使用 OpenAPI spec 的好處...

範例：交通部 PTX
http://ptx.transportdata.tw/MOTC

Parameters

Parameter	Value	Description	Parameter Type	Data Type
IATA	(required)	機場代碼	path	string
$select		挑選	query	string
$filter		過濾	query	string
$orderby		排序	query	string
$top	30	取前幾筆	query	string
$skip		跳過前幾筆	query	string
$format	JSON	指定來源格式	query	string

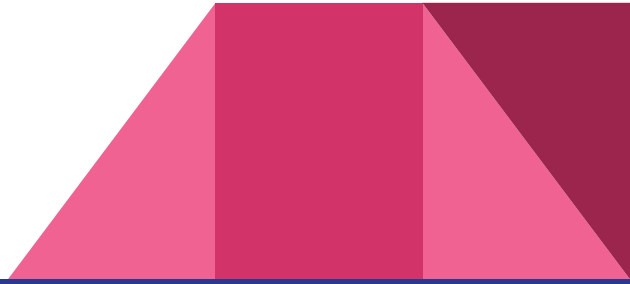
Try it out!

所以 WebService 要怎麼通過品質檢測機制呢？

OpenAPI Spec 所產生的會是一個 YAML 或是 JSON 格式的檔案

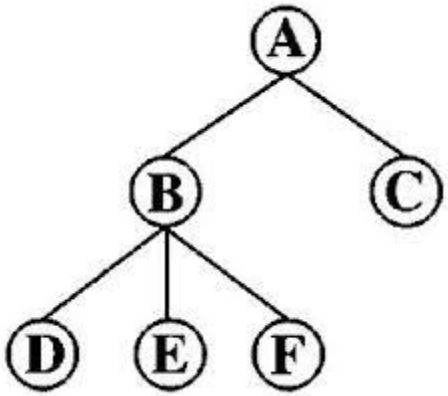
在平台上上架這個產生的檔案就行了

詳細資料可參考：https://swagger.io/

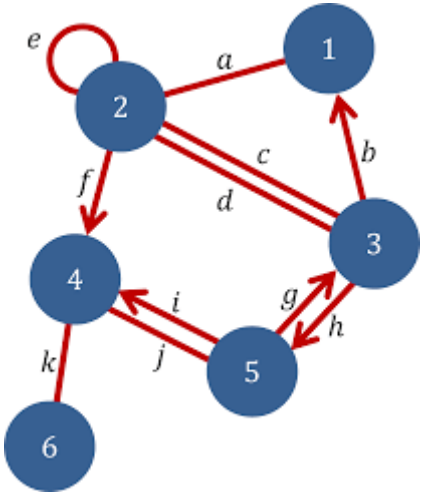


資料易於處理

常見抽象資料結構



樹狀



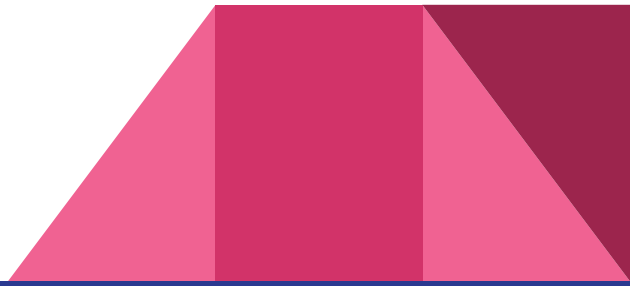
圖狀

表格

常見資料格式能存的資料結構

- EXCEL, ODS, CSV：表格
- JSON, XML：表格、樹狀

政府機關資料分為既有系統和人工收集整理兩種，後者多數資料是以 EXCEL 方式儲存，因此表格狀資料是今年「資料易於處理」的重點



政府資料主要來源

- 委外廠商開發的資訊系統
- 承辦人員人工整理的資料

政府資料的主要來源

- 委外廠商開發的資訊系統
  - 對外網站或是內部系統
  - 資料存放在資料庫系統中
  - 資料已經結構化有明確欄位定義
- 承辦人員人工整理的資料
  - 年度報告或是因特別需求產製報告或文件
  - 大部份資料存放在 Excel 檔案中
  - 資料不一定有結構化

來自資訊系統的處理方式

- 已經是結構化存在資料庫中，大部份不需要資料清理
- 需要透過廠商直接從資料庫匯出符合開放資料格式
- 可能會有 API 可以使用

來自承辦人員人工整理資料的處理方式

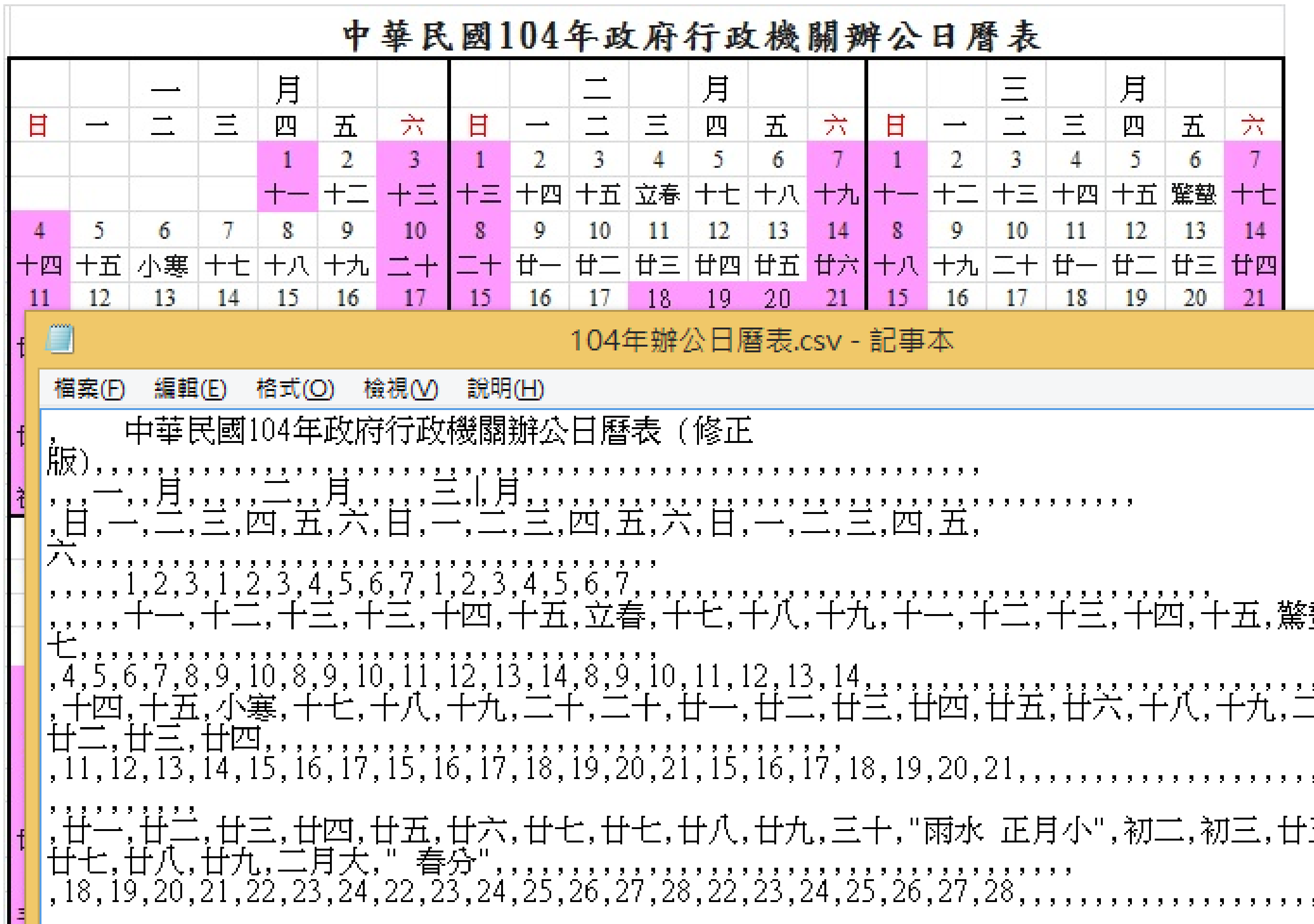
大部份是 Excel

Excel 格式如何呢？

EXCEL、ODS、CSV 比較...

- CSV 只是純粹儲存資料的格式
- EXCEL、ODS 可以拿來存資料，也可以拿來建立報告
  - 建立報告常用到的功能：
    - 資料分頁
    - 合併儲存格
    - 顏色框線
    - 背景顏色
    - 函式
    - 利用排版加表頭和表尾放備註

拿 CSV 直接轉換報告的悲劇



理想的報告與資料關係...

人事行政局辦公日曆表　　資料

date	name	isHoliday	holidayCategory	description
2017/4/1		是	星期六、星期日	
2017/4/2		是	星期六、星期日	
2017/4/3		是	放假之紀念日及	全國各機關學校放假一日。
2017/4/4	兒童節、民族掃墓節（清明節）	是	放假之紀念日及節日	全國各機關學校放假一日(兒童節與民族掃墓節同一日時，於前一日即四月
2017/4/8		是	星期六、星期日	
2017/4/9		是	星期六、星期日	
2017/4/15		是	星期六、星期日	
2017/4/16		是	星期六、星期日	
2017/4/22		是	星期六、星期日	
2017/4/23		是	星期六、星期日	
2017/4/29		是	星期六、星期日	
2017/4/30		是	星期六、星期日	

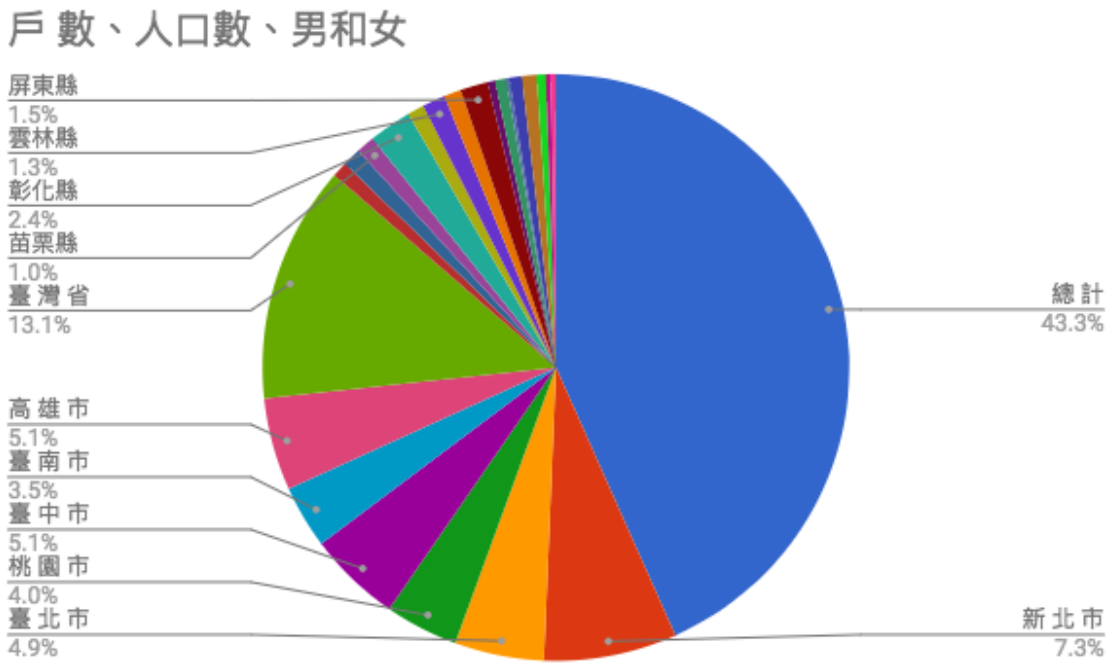
好的表格型資料

1. 第一行是表頭，只放欄位名稱
    a. 欄位名稱不會重覆
    b. 欄位盡量不需要成長變動
2. 第二行以後開始是資料
    a. 資料的欄位數不會比表頭多，每一格都要對應的了他的欄位
3. 除了表頭跟資料，沒有其他東西
    a. 有備註、資料時間等資訊可以另外放在檔案說明或平台說明
4. 只有原始資料，沒有總計、平均之類被運算出來的資料

為什麼要拿掉平均、總計呢？

以縣市人口資料為例，做圖表

	A	B	C
	區域別	戶數	人口數
	總計	8,656,728	23,571,408
	新北市	1,544,255	3,986,501
	臺北市	1,051,436	2,681,375
	桃園市	792,337	2,193,098
	臺中市	960,800	2,790,381
	臺南市	684,225	1,886,074
	高雄市	1,092,766	2,776,366
	臺灣省	2,488,001	7,107,086
	宜蘭縣	168,592	456,259
	新竹縣	191,590	552,488
	苗栗縣	187,946	552,807
	彰化縣	387,326	1,281,304
	南投縣	178,027	500,449
	雲林縣	240,269	689,463
	嘉義縣	182,842	510,498



乾淨的表格

座號	姓名	身高	體重
1	王小明	160	61
2	林小美	151	40
3	吳小華	168	59

資料易於理解

很乾淨，但這樣子能知道資料是什麼嗎？

COUNTY_ID	COUNTY	FLD01	FLD02	FLD03	FLD04	FLD05
65000	新北市	13829532	12992035	0	2303638	714624
63000	臺北市	48265667	45641689	0	12017654	236100
68000	桃園市	11999876	10908946	0	1438568	130235
66000	臺中市	11766450	10968525	0	1451550	254240
67000	臺南市	4355934	4268873	0	515447	24970
64000	高雄市	11247415	11063048	0	1274758	327299
10002	宜蘭縣	1012644	993920	0	125651	5730
10004	新竹縣	2614622	2590730	0	599126	149009

主要欄位說明 - 105年1月全國賦稅收入實徵淨額

COUNTY（縣市名稱）、FLD01（總計）、FLD02（稅課收入）、FLD03（關稅）、FLD04（所得稅小計）、FLD05（營利事業所得稅）、FLD06（綜合所得稅）、FLD07（遺產及贈與稅）、FLD08（遺產稅）...

COUNTY_ID	COUNTY	FLD01	FLD02	FLD03	FLD04	FLD05
65000	新北市	13829532	12992035	0	2303638	714624
63000	臺北市	48265667	45641689	0	12017654	236100
68000	桃園市	11999876	10908946	0	1438568	130235
66000	臺中市	11766450	10968525	0	1451550	254240
67000	臺南市	4355934	4268873	0	515447	24970
64000	高雄市	11247415	11063048	0	1274758	327299
10002	宜蘭縣	1012644	993920	0	125651	5730
10004	新竹縣	2614622	2590730	0	599126	149009

壓縮檔處理方式...

為什麼需要壓縮檔？

- 原始資料過大筆數過大，想節省下載速度
- 雖然資料都長一樣，但是有很多不同年份（或不同縣市）的資料，不同年份（或不同縣市）的資料想各別一個檔案
- 有些不同性質的資料，但常常同時被一起使用，希望能方便使用者同時一起取得，所以用壓縮檔包在一起

壓縮檔在易用性可能遇到的問題...

- 可能包含跟資料內容無關的檔案（像 README.txt）
- 只從檔案名稱中可能看不出這個是什麼資料
- 可能同時包含好幾種不同的資料集在同一個壓縮檔中

解決壓縮檔的問題

- 加上 manifest.csv ，說明這個壓縮檔中哪些檔案是資料，以及簡單說明是什麼樣的資料
- 加上 schema.csv ，說明這個壓縮檔中是否有不同的主要欄位

manifest.csv 格式

- manifest.csv 的目的
  - 哪些檔案是重要的資料，哪些是可以不用理會的
  - 重要的資料的主要欄位是什麼
- manifest.csv 格式
  - name: 檔名
  - schema: 如果有自己不同的主要欄位，這邊要指定主要欄位存放的檔名
  - description: 這個檔案的更詳細的描述 (因為檔名可能是 A123985.csv 這種代碼性的檔名，如果沒描述對使用者來說會比較難知道他的用途）

https://data.gov.tw/faq/633

schema.csv 的格式

- schema.csv 的目的
  - 一個壓縮檔內有多個資料檔，有不同的主要欄位描述，單靠資料集本身銓釋資料的文字敘述不足夠
  - 欄位名稱宜精簡，但是精簡的欄位名稱很多事無法說清楚（Ex: 這欄位的單位、這欄位的格式、這欄位的集合...），需要另一個主要欄位的描述檔可以放更多資訊
- schema.csv 的欄位
  - name: 主要欄位名稱
  - title: 主要欄位描述
- 國際標準參考： JSON Table Schema

來講怎麼修吧...

一樣從這三點來講

- 可直接取得
- 資料易於處理
- 資料易於理解

可直接取得

可直接取得之前要先可取得...

可取得是最重要的，無法取得什麼都別談...

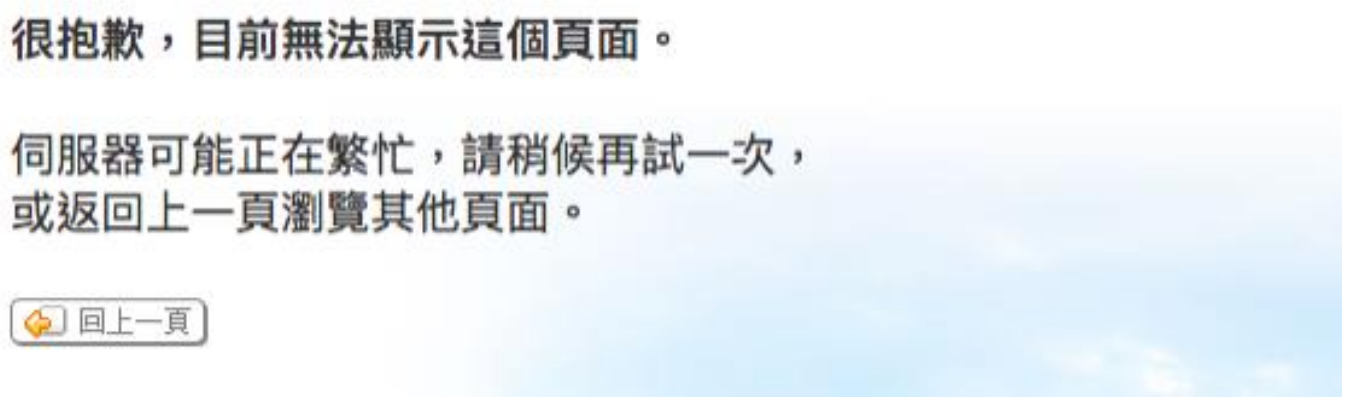
一些無法取得資料的案例

[002] HTTP 狀態非 200

[014] 資料資源URL格式錯誤

[015] 資料資源下載時錯誤

很抱歉，目前無法顯示這個頁面..



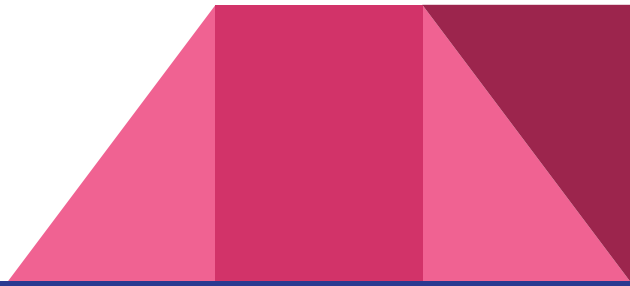
發生伺服器錯誤...

'/TSE' 應用程式中發生伺服器錯誤。

找不到資源。

描述: HTTP 404. 您要尋找的資源 (或其相依性的其中之一) 可能已經移除、名稱已經變更或是暫時無法使用。請檢閱下列 URL，並且確定它的拼寫無誤。

要求的 **URL:** /TSE/main/content/wHandMenuFile.ashx



無法連上這個網站


無法連上這個網站
www.bsmi.gov.tw 的回應時間過長。
請透過 Google 搜尋「bsmi gov tw wSite record file act」
ERR_CONNECTION_TIMED_OUT

判斷跟處理

判斷

- 可能有「找不到網頁」「404」「無法連上」「錯誤」「無法顯示網頁」「伺服器繁忙」

處理

- 檢查看看是不是網址寫錯
- 是不是檔案已經被刪除或是被下架
- 連絡承包廠商確認為什麼連不上
- 權限是不是忘了設定成公開

不能直接取得

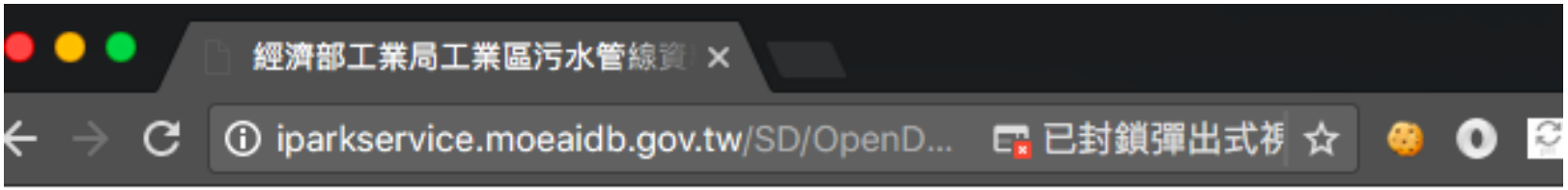
現在越來越多資料分析工具，如果不能直接取得，會讓使用資料更費力

通常錯誤訊息是 [001] 資料資源無法直接下載

「已封鎖彈出式視窗」

http://data.gov.tw/node/16388

會顯示「已封鎖彈出式視窗」



內文全部空白沒東西...

會到首頁，但是接下來不知道該怎麼辦....

http://data.gov.tw/node/11056

會到列表頁，但是還是需要人類判斷要下載哪個

http://data.gov.tw/node/17443

或者是導到了特別的下載頁面

判斷跟處理

判斷

- 需要多經過一個網頁才能下載到檔案
- 連過去出現全白網頁沒反應
- 雖然有自動下載，但是還有一個全白網頁留著

處理

- 連絡承包廠商請廠商調整

資料易於處理

好的表格型資料

1. 第一行是表頭，只放欄位名稱
    a. 欄位名稱不會重覆
    b. 欄位盡量不需要成長變動
2. 第二行以後開始是資料
    a. 資料的欄位數不會比表頭多，每一格都要對應的了他的欄位
3. 除了表頭跟資料，沒有其他東西
    a. 有備註、資料時間等資訊可以另外放在檔案說明或平台說明
4. 只有原始資料，沒有總計、平均之類被運算出來的資料

乾淨的表格

座號	姓名	身高	體重
1	王小明	160	61
2	林小美	151	40
3	吳小華	168	59

不乾淨的表格資料種類

[008] 此excel檔(xls, xlsx, ods)非固定欄位之excel格式

- 有分頁
- 有合併儲存格
- 有框線
- 有空行
- 有背景色

[013] 非RFC4180格式CSV

下面用幾個案例來說明怎麼處理

「報告」直接上架的案例..

範例網址：
https://docs.google.com/spreadsheets/d/1JfhSh5k5CKIVr2dF7lsudBBbIrOS8L3_mFpdKsySRoI/edit#gid=522880854

每戶（每人）每月水費及每戶家庭每年水費占消費支出比率

項目別	每戶每月水費（元）	每人每月水費（元）	每戶家庭每年水費占消費支出(%)
80年	206	50	0.54
85年	241	71	0.42
90年	262	84	0.44
95年	252	85	0.39
100年	233	83	0.35
102年	231	84	0.32
103年	233	86	0.32
104年	229	85	0.32
105年	231	86	(詳說明)

資料來源：1.依據本公司「用戶數」、「供水普及率」、「水費收入」表，及行政院主計總處「家庭收支調查平均消費支出」等資料彙編。
2.每戶家庭每年水費=本公司普通及軍眷用戶全年水費（不含代徵之各項處理、保育費用）。

說　明：①105年每戶家庭消費支出資料行政院主計總處預計106年8月下旬發布。
②查詢每戶（每人）每月水費相關歷年資料，請至本網站「計畫成果－台灣自來水事業統計年報」。

清理順序..

1. 刪除上方表頭和下方表尾（這些資訊記下來，把他放進資料集的詮釋資料備份中）
2. 移除合併儲存格的資訊（主要是單位，這個可以放入資料集的主要欄位描述中）
3. 移除掉資料內備註類的文字，也放入資料集的詮釋資料備份中

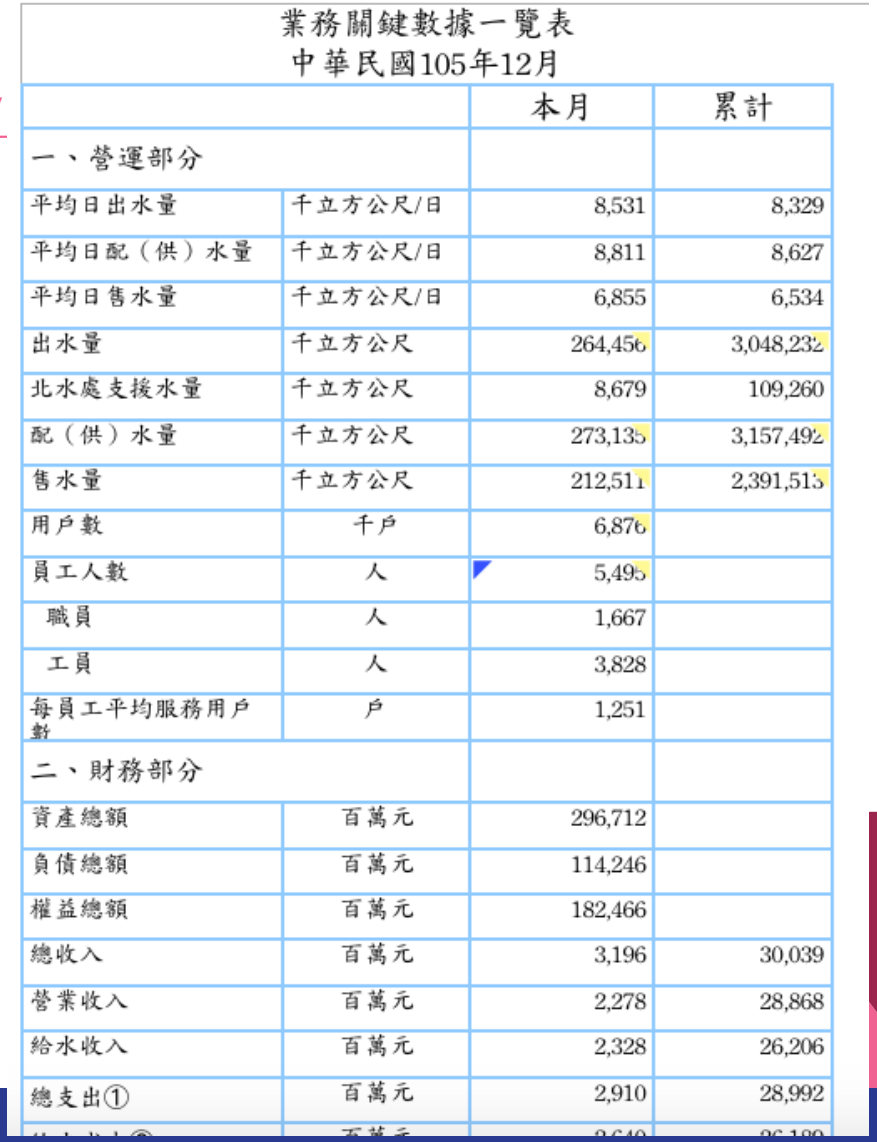
清理完成

https://docs.google.com/spreadsheets/d/1JfhSh5k5CKIVr2dF7lsudBBbIrOS8L3_mFpdKsySRoI/edit#gid=0

	A	B	C	D
	項目別	每戶每月水費	每人每月水費	戶家庭每年水費占消費支出
	80年	206	50	0.54
	85年	241	71	0.42
	90年	262	84	0.44
	95年	252	85	0.39
	100年	233	83	0.35
	102年	231	84	0.32
	103年	233	86	0.32
	104年	229	85	0.32
	105年	231	86	

再複雜一點的表格

https://docs.google.com/spreadsheets/d/7jqPi5eHKR3ne-g/edit#gid=1662910623



業務關鍵數據一覽表
中華民國105年12月

		本月	累計
一、營運部分			
平均日出水量	千立方公尺/日	8,531	8,329
平均日配（供）水量	千立方公尺/日	8,811	8,627
平均日售水量	千立方公尺/日	6,855	6,534
出水量	千立方公尺	264,456	3,048,232
北水處支援水量	千立方公尺	8,679	109,260
配（供）水量	千立方公尺	273,135	3,157,492
售水量	千立方公尺	212,511	2,391,513
用戶數	千戶	6,876	
員工人數	人	5,495	
職員	人	1,667	
工員	人	3,828	
每員工平均服務用戶數	戶	1,251	
二、財務部分			
資產總額	百萬元	296,712	
負債總額	百萬元	114,246	
權益總額	百萬元	182,466	
總收入	百萬元	3,196	30,039
營業收入	百萬元	2,278	28,868
給水收入	百萬元	2,328	26,206
總支出①	百萬元	2,910	28,992

清理法

- 如果欄位已經幾乎確定不會再增加，可以把「出水量」、「北水處支援水量」等拉出來變成主要欄位
- 可用 Excel 的 TRANSPOSE 把表格轉 90 度
- 可以移除平均相關欄位

有分頁的案例

https://docs.google.com/spreadsheets/d/1igkhtN7ITG0c0Ynv9O1c91aPDIsbl9IaEfdY5vndf5k/ed

	A	B	C	D	E	F	G	H	I
1	燃料檢測服務收費標準								
2	序號	試驗項目	方法	收費標準NT$					
3	1	Distillation Temperature(蒸餾試驗	CNS 1218 A	2000					
4	2	Flash Point-PM Method(閃點測定)	CNS 3574 A	1500					
5	3	Water Content by Distillation(含水	CNS 3517 A	2400					
6	4	Pour Point(流動點測定)(TAF認證	CNS 3484 A	1800					
7	5	Corrosion, Copper strip(銅片腐蝕	CNS 1219 A	1600					
8	6	Existent Gum Content(含膠量測定	CNS 3382 A	3900					
9	7	Viscosity Kinematic at 40℃(動力黏	CNS 3390 A	1100					
10	8	Viscosity Kinematic at 50℃(動力黏	CNS 3390 A	3100					
11	9	Ash From Petroleum Products(灰份	CNS 3576 A	1900					
12	10	Ramsbottom Carbon Residue(10%Bottom) (蒸餘10%藍式殘碳量測定) (TAF 認證)	CNS 3776/1218 ASTM D524/D86	3800					
13	11	Oxidation Stability(氧化穩定性測	CNS 12014	5200					
14	12	Rust Preventing Characteristics(防鏽	ASTM D665	2000					
15	13	Rust Preventing Characteristics(防鏽	ASTM D665	2100					
			ASTM						

燃料檢測服務 | 技術服務檢測服務 | 環境檢測服務

清理法

- 不該使用分頁
  - 可選擇把三個分頁合併成一個
  - 或者是把三個分頁變成三個資料集

不過如果欄位都長一樣，比較建議是合併為一個

105年國產及進口其他釀造酒類數量表

多維度表格處理

來源: https://www.nta.gov.tw/_admin/_upload/Announce/8288/pic/files/各月別國產及進口其他釀造酒類數量表.xls

105年國產及進口其他釀造酒類數量表

單位：公石

月別/產品	其他釀造酒類					
	國產		進口		小計	
	數量	結構比%	數量	結構比%	數量	結構比%
1月	159.86	36.67	276.13	63.34	435.99	100.00
2月	37.73	20.32	147.93	79.68	185.66	100.00
3月	35.65	8.48	384.56	91.52	420.21	100.00
4月	34.53	1.31	2,604.18	98.69	2,638.71	100.00
5月	40.32	2.04	1,932.94	97.96	1,973.26	100.00
6月	35.80	7.48	442.62	92.52	478.42	100.00
7月	55.81	5.09	1,041.89	94.92	1,097.70	100.00
8月	74.94	13.57	477.19	86.43	552.13	100.00
9月	37.51	4.77	748.84	95.23	786.35	100.00
10月	30.71	7.10	401.53	92.90	432.24	100.00
11月	83.84	12.32	596.88	87.68	680.72	100.00
12月	88.46	4.05	2,094.98	95.95	2,183.44	100.00
合計	715.15	6.03	11,149.67	93.97	11,864.82	100.00
備　註：	資料來源：					
	(1)國產：依據財政部財政資訊中心提供之資料					
	(2)進口：依據財政部關務署提供之資料					

清理法

1. 一一把各合併儲存格的維度拆解出來
2. 拿掉被運算出來的值
3. 把時間獨立拉出來變成一個欄位

資料易於理解

主要欄位說明 - 105年1月全國賦稅收入實徵淨額

COUNTY（縣市名稱）、FLD01（總計）、FLD02（稅課收入）、FLD03（關稅）、FLD04（所得稅小計）、FLD05（營利事業所得稅）、FLD06（綜合所得稅）、FLD07（遺產及贈與稅）、FLD08（遺產稅）...

COUNTY_ID	COUNTY	FLD01	FLD02	FLD03	FLD04	FLD05
65000	新北市	13829532	12992035	0	2303638	714624
63000	臺北市	48265667	45641689	0	12017654	236100
68000	桃園市	11999876	10908946	0	1438568	130235
66000	臺中市	11766450	10968525	0	1451550	254240
67000	臺南市	4355934	4268873	0	515447	24970
64000	高雄市	11247415	11063048	0	1274758	327299
10002	宜蘭縣	1012644	993920	0	125651	5730
10004	新竹縣	2614622	2590730	0	599126	149009

回顧一下為什麼需要壓縮檔

- 原始資料過大筆數過大，想節省下載速度
- 雖然資料都長一樣，但是有很多不同年份（或不同縣市）的資料，不同年份（或不同縣市）的資料想各別一個檔案
- 有些不同性質的資料，但常常同時被一起使用，希望能方便使用者同時一起取得，所以用壓縮檔包在一起

想節省檔案空間...

- 如果壓縮檔內只有一個檔案
  - 恭喜你，什麼都不用做就是合格的了
- 除了主角檔案外還有其他非資料本體檔案(Ex: readme.txt)
  - 加上 manifest.csv ，把主角檔案加進去

假設情境：台灣公司資料的 company.csv 因為資料量太多，所以壓縮起來節省空間

壓縮檔內有 company.csv 和 readme.txt 兩個檔案，因此檢測不合格

修理方式：再手動加一個 manifest.csv

檔案內容如下：

```
name,description
company.csv,台灣公司資料
```

資料都長一樣，但是有很多不同年份（或不同縣市）的資料，不同年份（或不同縣市）的資料想各別一個檔案

- 壓縮檔內除了資料檔案外沒其他檔案，而且每個資料檔的主要欄位都一模一樣
  - 恭喜你，什麼都不用做，已經是合法的了
- 除了資料檔案以外，還包含 README.txt 之類的非資料檔案
  - 需要加上 manifest.csv 把主角們列出來

假設情境：台灣歷年村里人口數資料放在同一個壓縮檔中

壓縮檔內有 population-101.csv, population-102.csv, population-103.csv, population-104.csv, 和 readme.txt 等檔案，因此檢測不合格

修理方式：再手動加一個 manifest.csv

檔案內容如下：

```
name,description
population-101.csv,101年村里人口數
population-102.csv,102年村里人口數
population-103.csv,103年村里人口數
population-104.csv,104年村里人口數
```

有些不同性質的資料，需要被放一起使用

- 因為有不同性質的資料，所以這個壓縮檔內可能有很多組不同的主要欄位，因此需要 schema.csv 來描述這些主要欄位

（如果這些不同性質的資料並不一定或不是同時被一起使用，也可以考慮不要壓縮在一起，而是選擇拆成多個資料集上傳）

案例：實價登錄

https://data.gov.tw/dataset/18703 縮減下載版

實價登錄包含「買賣」、「建物」、「土地」、「車位」四個資料集

而四種資料集各有不同欄位，
買賣有鄉鎮市區、交易金額、交易日期、交易建物數、交易土地數、交易車位數...
建物有屋齡、面積、樓層...
土地有面積、使用分區、區段位置...
車位有類別、面積...

所以會需要用到 manifest.csv 描述各別檔案功用
以及需要 schema.csv 描述不同資料的主要欄位

其他壓縮檔注意事項

- 檔名應避免使用中文字
- 不該在壓縮檔內再放壓縮檔

RSS 或 XML 相關問題

Ex: http://data.gov.tw/node/15307

六月底平台會增加針對 XML, JSON 可以指定資料所在路徑的功能，屆時這些資料集就可以通過檢測了