

資料檔格式簡介

用於文化內容策進院之「文創產調視覺化」之資料檔說明文件

此文件說明如何建立、編輯使用於文創產調視覺化之資料檔案。首先，您可能會最關心的應該就是「檔案格式」跟「編輯軟體」了：

※ 檔案格式



文化內容策進院之「文創產調視覺化」(後稱產調視覺化)為一線上網站，使用 JavaScript 即時載入並繪製資料檔成互動圖表。

為了方便直接透過線上程式讀取，資料檔格式以 **CSV (逗號分格數值)** 格式為主，可以透過試算表軟體 (如 Google Spreadsheet 與 Microsoft Excel) 中編輯後匯出。此份文件將**著重在試算表中資料的編排方式**，不會涉及到檔案格式等較低階的細節。

※ 編輯軟體

要編輯此類檔案，您僅需要準備 Microsoft Excel 或 Google Spreadsheet 等軟體；若牽涉到軟體使用的部份，此文件將以 **Google Spreadsheet 做為示範工具**，但兩者的概念大多通用，若您希望使用 Microsoft Excel 或其它相似的軟體，您可將文件說明的部份類推至您的軟體上使用。



此文件分為以下章節：

- 簡介
- 試算表結構規範
- 多重資料表
- 屬性規劃方針
- NOIR 資料分類法

簡介

為了讓程式能夠順利讀取您的資料，您的資料必須要「結構化」。結構化不容易簡單說明，但大致上，資料必須要有一致性，避免混雜排版、自然語言等資訊，並且盡量將複合性的資訊分開記載。

※ 基本架構

為方便接下來的說明，我們簡單描繪一下表格的長相。如下圖：

- 資料以表格呈現
- 由上而下為逐筆資料 (data row)
- 由左至右則為各筆資料的不同欄位 (data column / field)

各筆資料			各個欄位		
第一筆資料			第一欄	第二欄	第三欄
第二筆資料					
第三筆資料					

※ 資料一致性

以下例來說，左表在數字中混雜了文字、縮寫跟單位，缺乏一致性；中表在數字中使用文字加註說明；右表在同一個欄位中記載兩組資料：

缺乏一致性	缺乏一致性	混合資料程式不能用
5k	102年	203 / 五折
256	103年	198 / 七折
五千元	104年 (僅 A 類)	200 / 六折

改進的方式並不難，首先統一格式，接著抽離單位，最後將混合欄位分開：

元	年	分類	元	折數
5000	102	全部	203	5
256	103	全部	198	7
5000	104	A	200	6

由於大多數的資料都有單位的概念，因此上表此類「標頭 (header) + 資料 (body)」的基本格式可說是我們表格的最基本格式。

又或者，下表某真實案例中，記載各類型廣播電臺的涵蓋電臺，等於將多組資料寫在單一儲存格中，並用頓號排板，前後文結構又不一致，程式無法正確分析：

類型	涵蓋電臺
公營電臺 (10)	內政部警政署警察廣播電臺（大功率）、國立教育廣播電臺（大功率）、漢聲廣播電臺（國防部心理作戰大隊）（大功率）、復興廣播電臺（大功率）、行政院農業委員會漁業署漁業廣播電臺（大功率）、臺北廣播電臺（大功率）、高雄廣播電臺（大功率）、財團法人中央廣播電臺（海外短波）、講客廣播電臺（客家委員會）（大功率）、財團法人原住民族文化事業基金會（大功率）
調幅電臺 (16)	中華、天南、天聲、民本、民立、先聲、成功、建國、電聲、益世、國聲、基隆、勝利之聲、華聲、鳳鳴、燕聲

以上例來說，我們可以做以下改進：

- 數字並不需要，程式即可做統計
- 各電臺分別陳列在不同筆資料之中

調整後的資料表如下：

電臺	類型
內政部警政署警察廣播電臺（大功率）	公營電臺
國立教育廣播電臺（大功率）	公營電臺
漢聲廣播電臺（國防部心理作戰大隊）（大功率）	公營電臺
中華	調幅電臺
天南	調幅電臺
天聲	調幅電臺
民本	調幅電臺
... 下略 ...	

※ 嚴禁合併儲存格

為了將各筆資料明確分割，嚴禁使用合併儲存格。的確，由於統計上的困難，有時無法準確的將資料切分開來，然而這樣的資料就算您做成了資料檔，程式也是無法呈現。

為了盡量呈現您手中的資料，您可以選擇：

- 標註為「無資料」（留空白, 或使用橫線「-」）
- 將表格依欄位分割後，再合併的資料組重新編排

請不要這樣做

年度	金額	數量
2018	500	25
2019		16
2020	250	

寧可省略

年度	金額	數量
2018	-	25
2019	-	-
2020	250	-

切割表格後留下金額

年度	金額
2018 ~ 2019	500
2000	250

切割表格後留下數量

年度	數量
2018	25
2019 ~ 2000	16

進一步結構化

起始年度	結束年度	金額
2018	2019	500
2000	2000	250

起始年度	結束年度	數量
2018	2018	25
2019	2000	16

※ 摘要

依上述的結構化要求，一個最基本的資料表格如下：

欄位1	欄位2	欄位3	欄位4
第一筆資料 欄位1的值	第一筆資料 欄位2的值	第一筆資料 欄位3的值	第一筆資料 欄位4的值
第二筆資料 欄位1的值	第二筆資料 欄位2的值	第二筆資料 欄位3的值	第二筆資料 欄位4的值
... (後續其它筆資料)			
... (後續其它筆資料)			

最後，請確保單一檔案僅包含一個表格。如果您有多組表格，只要無法合理併成一個表格，就應該另外存成獨立的檔案，避免程式無法正確解析。

試算表結構規範

除了一般結構化的概念，我們還必須要考慮到程式的特定用途，在此處即為「產調視覺化」線上圖表繪製的用途。為了生成圖表，我們需要以下資訊：

- 說明性資料－資料集標題、註解等
- 欄位資訊－各欄位名稱、型態、單位等
- 資料內容－實際數字、文字等本身
- 保留欄位－格式版本等程式未來可能會需要的擴充機制

一般來說，這些資訊通常會使用程式常見的專用格式來儲存 (如 JSON 或 XML)，但由於此類文件通常較不容易讓一般人編輯，但一般較容易編輯的 Microsoft Excel 檔案卻又過於一般性，沒有這些資訊的規範，因此我們必須明確規定如何在試算表檔案中加入這些資訊，並請編輯人員遵照規定編輯試算表檔案。

首先，在一個普通的試算表中，我們在上面保留 7 行用以記載額外的資訊；第 8 行以後則為單純的資料區域。前 7 行資料又分為「表格」與「欄位」的中介資訊：

- 表格中介資訊：說明這個表格，包括標題、註解、程式用保留資料等，共 3 行。
- 欄位中介資料：說明表格中各個欄位的單位，意含跟資料類型等，共 4 行。

以圖像化的方式來呈現：

1	標題
2	保留列 (請留白)
3	註解
4	資料類型 (文字、數值等，以 NOIR 分類法表示，於後續章節說明)
5	單位 (例如：元、人、頁)
6	保留列 (請留白)
7	欄位名 (如年份、出口值、人次、頁數等)
8	資料1
9	資料2
10	...

以下以一個假想案例來做範例：（注意前兩行我們僅用到第一欄的儲存格。第二欄以後的儲存格因為未定義用途，應保持空白）

1	文創產業樣貌		
2			
3	自 104 年至 108 年，樣本 320 間公司，其中 105 年度樣本數較少，採正規化後推估。		
4	T	C	R
5	年		元新臺幣
6			
7	年度	產業別	產值
8	104	出版	2881
9	104	影視	3078
10	105	出版	2763
10	105	影視	3159

多重資料表

如同在前文中所提到的，我們禁止在單一檔案放多個表格。

每個資料檔僅能含有一個表格，再強調一次，不得將多組資料記載在同一份資料表中。

下圖為一錯誤的示範，將多個表格拼貼在一個試算表中，其中包括「電影放映業」、「電影後製特效業」等兩產業，又分別將兩產業各年度之資料分不同表格記載：

	A	B	C	D	E	F	G	H
1	電影映演業	2019年	電影後製特效業	2018年	電影後製特效業	2017年	電影後製特效業	2016年
2	電影放映收入	90.02%	電影後製收入	23.32%	電影後製收入	20.90%	剪輯	22.78%
3	廣告收入	0.64%	電影特效製作收入	25.51%	電影特效製作收入	20.88%	音樂製作	3.45%
4	食物飲料販售	8.41%	其他內容製作收入	2.67%	其他內容製作收入	7.81%	後製錄音	9.43%
5	衍生商品販售	0.34%	其他內容後製收入	13.03%	其他內容後製收入	11.87%	視覺特效	35.07%
6	其他電影商務收入	0.09%	其他內容特效收入	6.48%	其他內容特效收入	10.75%	字幕、拷貝	2.99%
7	其他收入	0.51%	影視相關設備器材或場地租賃收入	15.09%	影視相關設備器材或場地租賃收入	16.10%	SR Dolby	2.44%
8			藝人經紀	0.39%	置入性行銷、贊助收入	1.47%	器材及設備提供	19.93%
9			其他	4.58%	其他	5.35%	其他	3.91%
10			電影版權銷售收入	4.31%	電影版權銷售收入	4.66%		
11			電影於網路數位發行之相關收入	0.22%	電影於網路數位發行之相關收入	0.15%		
12			電影衍生商品授權收入	3.36%	電影票房分成	0.05%		
13			電影代理發行收入	1.03%				
14								
15			電影映演業	2018年	2017年	2016年		
16			電影放映收入	94.66%	93.47%	89.47%		

當多重表格混合在一起時，程式完全無法知道該如何判讀資料，也因此這樣的資料只能透過人工使用，無法輸入程式使用，這也是為什麼我們不應該這樣整理資料。

然而，分成多檔案有時會讓整理資料變得相當不容易。為方便管理，我們可以將多個表格以 **不同工作簿** 的形式儲存在同一個試算表檔案中；只要將相關的資料放在同一個試算表中，資料整理或檔案管理都會變得相對容易。下圖展示了使用 Google Spreadsheet 做多重表格管理的範例，各個工作簿以表格標題命名，內容則依上述資料表結構編排：

年度	參與國際市場展	輔導參與國際市	入圍國際影展國	榮獲國際影展獎	輔導參與國際影展國片人次		
97	208	-	91	16	69		
98	211	-	125	12	110		
99	239	82	124	32	77		
100	206	80	122	23	51		
101	244	122	96	35	60		
102	294	138	86	31	53		
103	289	122	86	16	56		
104	222	153	88	17	49		
105	214	14	60	12	44		
106	217	67	66	10	78		
107	215	70	20	83			
108	213	76	8	113			
109	219	41	12	15			

- 國片製作兼發行情形
- 跨國合作影片數量
- 國片導演兼任編劇情形
- 國際影展活動參與度
- 收益結構
- 全台電影票房
- 台北市電影票房
- 人均觀影次數

+

三

情形

跨國合作影片數量

國片導演兼任編劇情形

國際影展活動參與度

收益結構

全台電影

由於程式使用資料時是以單一 CSV 檔為單位，而 CSV 檔並不支援工作簿的概念，我們仍需要將工作簿輸出成多個 CSV 檔，才能供給程式使用。資料表數量眾多，手工逐一輸出資料較費時，因此這時我們可以使用輔助程式如 VBA 或 Script Tool 來協助我們自動生成 CSV 檔。

輔助程式屬編輯軟體操作的範疇，並非此文件所涵蓋的範圍，但相關教學可以在網路上找到，這裡附上一個用於 Google Spreadsheet 的工具程式，可協助我們生成所有工作簿的 CSV 壓縮檔：

程式碼出處：[Michael DeRazon](#)

此程式需透過 Google Spreadsheet 的 Script Editor (選單 → 工具/Tool → 指令碼編輯器/Script Editor)，關於使用 Script Editor 的教學，請另參考 Google Spreadsheet 的使用文件。

欄位規劃方針

有時候，即便我們已經依照上述方式來整理資料，仍不免將資訊藏進欄位名稱中，減損了資料可再利用性。

比方說，下表利用多個欄位分「年度」來呈現「產值」這個數據，由於「年度」之值儲存在欄位中，並沒有對應的描述性資料 (年度的資料分類、單位等資訊都缺失了)，我們在利用資訊時就變得更加侷限；此外，「產值」的概念也被混進年度之中，更難正確的解析：

產業別	105年產值	106年產值	107年產值
出版	2517	1923	1498
影視	3011	3262	3119

這裡我們應考慮將欄位重新匯整成單一欄位，並將資料做對應的轉換。此動作一般稱之為「反向樞紐 (Unpivot)」。

以上表為例，我們先將「年度」獨立成一欄，產業別則會因年份而需要重覆輸入，如下表左。接著，再增加一欄「產值」，將「產業別 x 年份」原本對應到的數值填回此欄，如下表右。

先將年份獨立成一欄

產業別	年份
出版	105
影視	105
出版	106
影視	106
出版	107
影視	107

再依對應位置合併原本多欄的產值

產業別	年份	產值
出版	105	2517
影視	105	3011
出版	106	1923
影視	106	3262
出版	107	1498
影視	107	3119

反向樞紐後的資料表仍能透過正向樞紐轉換變成原本的形式，卻多了額外欄位的資訊，同時程式也更容易依範圍過濾、使用額外的欄位做更多樣化的資料呈現，因此較建議在編撰資料時使用此種形式的欄位規劃。

NOIR 資料分類法

掌握正確良好的資料結構後，讓我們來看看資料數據本身。

數據有不同的型態，最常見的就是「文字」與「數字」的分野：

我是文字	5447
------	------

數字跟文字可以做的事情不太一樣。數字可以加總、平均，文字可以排序、做分類，因此若能妥善的判斷數據的類型，我們才能正確的實現資料的視覺化。

程式基本上可以自動判斷數據的類型；然而問題是當我們有多組數據時，往往會因為一些例外狀況而使得判斷數據類型變得困難。比方說，以下兩個表格中的數字完全不應該做為數字使用，但程式很難自行判讀：

公司名稱	公車路線
1111	508
104	260
591	紅15

也因此，我們在整理資料時，最好同時將各欄資料的類型標註在表格中，這也就是我們在「試算表結構規範」一節中，表格第四行「資料類型」的用途。然而除了「文字」、「數字」外，資料還可細分成相當多不同的種類；為了避免讓資料梳理的過程變得過度複雜，我們僅依照資料視覺化的需求，使用「NOIR 資料分類法」將資料大致上分成以下幾種：

N Nominal / 記名式

代表標示、名稱，如代號型資料，如產品編號，國家、產業別、姓名等。

O Ordinal / 有序式

可以排序的資料都可視為有序式，但一般以直觀認定其是否有序來標註即可。如月份、日期、流水號。

I Interval / 區間式

以數字為主，若加總無意義（如年齡相加），或數值實為相對值（如攝式溫度），則視為區間式資料。

R Ratio / 比例式

非區間式的數字即為比例式。如金額、人數、面積、降雨量等等。

此外, 為更精準進行視覺化, 我們進一步將 N 與 O 類資料分為下列兩種子類：

C

Categorical / 分類式

當一組資料中此欄位中的數值僅有少數幾種數值, 可視為分類用途, 即可進一步標註為 C.

T

Temporal / 時序式

時序式資料. 當有序式資料為時間的概念, 可另標註為 T

了解資料分類方式後，我們在編撰資料表時，僅需將各種資料的類型以其對應代號 (**N O I R C T**) 填入第四行之「資料類型」欄位即可。下表即為一範例，依各欄位類型填入了對應的各代碼：

文創產業樣貌					
無註解					
T	C	N	O	I	R
年				%	元新臺幣
年度	產業別	公司名	資本額排名	員工男女比	產值
104	出版	雜誌A	23	25	1237
104	出版	雜誌B	15	28	672
104	影視	電視臺A	2	48	7826
104	影視	電視臺B	9	46	6271