Common Practices in Tabular Data Dissemination

Common practices for all recommended tabular dataset formats

- 1. The dataset should contain the header and the data only.
- 2. Each dataset should contain one and only one tabular data (table). In this way, the structure of the dataset should be unified as follows:
 - The first row of the table should be used for the header information
 - Data starts from the second row
 - Rows and columns should be continuous
- 3. Any title, footnote and comment should be removed. The explanation of information on the table/spreadsheet can be achieved in the following way(s):
 - Include the information in a documentation (e.g. data dictionary, specification document, etc.).
 - Use separate column(s) for the note(s) and comment(s) when necessary. For example, if an annotation is attached to a particular cell for explanation purpose, such information can be added to a new column.
 - Put the common note(s) in the "description" field of the dataset or data resource.

The following sample tabular dataset is used to illustrate how to apply the common practices from (1) to (3) above:

剪贴簿 га	字型	G.	對音方式	G.	數值	G.	樣式		語存枱
L14 -	$\therefore \checkmark f_x$								
	А	В	С	D	Е	F	G	н	Ι
1 Informa	tion of Mobile Digital	Electroni	c Device (30/05/2018)	—	Title row sl	hould be	removed.		
2 📛 B	lank row should be	remove	ed.						
Prices of	digital devices		No. of device sold			ould be	used for th	o hoodo	information
3 (HK\$)					FIISTIOWS		used for th	le neaue	mormation.
4 0-999			869	-	Data starts	from the	e second ro	w.	
5 1000-109	99		2554						
6 2000-299	99		3845						
7 3000-399	99		3154						
8 4000-499	99		2996						
9 5000-599	99		1738						
0 6000-699	99		1456						
1 7000-799	99		1945						
2 8000-899	99		1243						
3 9000-999	99		861						
4 10000-19	9999		712						
5 20000-29	9999		484						
6 30000-39	9999		217						
40000-49	9999		198						
8 50000-59	9999		145						
9 60000-69	9999		102						
20 70000-79	9999		85						
21 80000-89	9999		52						
22 90000-99	9999		31						
100000-1	99999		19	The bi	abost price is		Note and	Commen	t should be removed
24 200000 c	r above		7	HK\$4	38,000	-	or put in a	conarat	e column
25							or put in a	separat	ecolumn
6 Remarks	The information of m	obile digita	l electronic device in th	e above	table does no	t include fig	ures of DC o	ameras. 🔹	— Footnotes shou
27	Blank column	should	be removed.						be removed.
28									
9									
	工作表1 (+)								

55		12月11月11日	13 RX EL 13	135, - ²⁴ V		100 127 100
I9	\cdot : $\times \checkmark f_x$					
	A	В	С	D	Е	F
	Prices of digital devices	No. of device sold	Remark			
1	(HK\$)					
2	0-999	869	The lowest price is HK\$380			
3	1000-1099	2554	_			
4	2000-2999	3845				
5	3000-3999	3154				
6	4000-4999	2996				
7	5000-5999	1738				
8	6000-6999	1456				
9	7000-7999	1945				
10	8000-8999	1243				
11	9000-9999	861				
12	10000-19999	712				
13	20000-29999	484				
14	30000-39999	217				
15	40000-49999	198		1		
16	50000-59999	145				
17	60000-69999	102				
18	70000-79999	85				
19	80000-89999	52				
20	90000-99999	31				
21	100000-199999	19				
22	200000 or above	7	The highest price is HK\$438,000			
23						
24						
25						
26						
27						
28						
29						
	□ ▶ 工作表1 (+)					

4. If code value(s) is being used in the data, data dictionary document should be published together with the dataset to explain the meaning of the code value(s).

	В	С	D	E	
	No. of device sold	Remark	Device Code		
1					
2	869	The lowest price is HK\$380	H		
3	2554		H, P		
4	3845		H, P		Data Dictionary Document
5	3154		H, P		Data Dictionary Document
6	2996		H, P, M		
7	1738		H, P, M		Device Code
8	1456		H, P, M		
9	1945		H, P, M		H - Mobile Phone
10	1243		H, P, M		P - Tablet
11	861		H, P, M		M – Notebook
12	712		H, P, M		
13	484		P, M		
14	217		P, M		
15	198		М		
16	145		М		
17	102		М		
18	85		М		
19	52		М		
20	31		М		
21	19		М		
22	7	The highest price is HK\$438,000	М		
23					
24					
25					
26					
27					
28					
29					
4) 工作表1 (+)			

- 5. The numeric data should be presented using a number. Any formatted number should be avoided. (e.g. use 1234 instead of 1,234 or 1 234)
- 6. For the text data, any heading or trailing space(s) should be removed.
- Any redundant space(s) in between a word should be removed, especially for the space(s) between each Chinese word. (e.g. use 中文字 instead of 中文字)
- 8. For Simplified Chinese (SC) and Traditional Chinese (TC) character, standard unicode should be used. Data providers are advised to use a browser (e.g. Chrome) to check if the dataset containing Chinese character(s) can displayed before the dataset is published.
- 9. Full-width symbol should be used for SC/TC characters (i.e. 全形).

- 10. Aggregated data (i.e. summaries of data) should be disaggregated as far as possible. Generally speaking, open data that is disaggregated allows more different ways of use and thus gives more values on research and statistical analysis. For example, breaking student data down into grade level within school aged students, district of origin, or gender among student populations.
- 11. For data with address location and geo-location, the following practices are suggested:
 - For address data presentation, data providers are advised to adopt the Address Data Infrastructure (ADI) as far as possible.
 - The Latitude and Longitude (Lat/Lon) coordinates should be encoded in WGS-84 standard with the following column header names:

English header name	TC header name	SC header name				
Latitude	緯度	纬度				
Longitude	經度	经度				

- Separate columns should be used if there exists geo-location in other geo-coded format (e.g. HK80)

If there are multilingual versions of datasets (e.g. English, Traditional Chinese and Simplified Chinese), data providers should use separate files for each version.

剪	貼簿 г,	字型	G	對齊方3
I7	- · · · ·	$\times \checkmark f_x$		
	А		В	С
	流動數碼電子。	蚤品價格	售出數量	
1	(HK\$)			
2	0-999		869	
3	1000-1099		2554	
4	2000-2999		3845	
5	3000-3999		3154	
6	4000-4999		2996	
7	5000-5999		1738	
8	6000-6999		1456	
9	7000-7999		1945	
10	8000-8999		1243	
11	9000-9999		861	
12	10000-19999		712	
13	20000-29999		484	
14	30000-39999		217	
15	40000-49999		198	
16	50000-59999		145	
17	60000-69999		102	
10	70000 70000		05	

朝	站簿 гу 字型	Gr	對齊方式
L14	1 · · · · · · · · · · · · · · · · · · ·	fx	
	А	В	С
	Prices of digital devices	No. of device sold	
1	(HK\$)		
2	0-999	869	
3	1000-1099	2554	
4	2000-2999	3845	
5	3000-3999	3154	
6	4000-4999	2996	
7	5000-5999	1738	
8	6000-6999	1456	
9	7000-7999	1945	
10	8000-8999	1243	
11	9000-9999	861	
12	10000-19999	712	
13	20000-29999	484	
14	30000-39999	217	
15	40000-49999	198	
16	50000-59999	145	
17	60000-69999	102	
1 Q	70000 70000	25	

Resource Description defined in psi-data.json

Prices of digital devices (Traditional Chinese) 流動數碼電子產品價格 (繁體中文) 流动数码电子产品价格 (繁体中文) Prices of digital devices (English)

流動數碼電子產品價格(英文) 流动数码电子产品价格(英文)

Common practices for dataset in CSV open format

- 1. The international standard "RFC 4180 Common Format and MIME Type for CSV Files" published by IETF (https://www.ietf.org/rfc/rfc4180.txt) should be followed.
- 2. CSV and TSV dataset should be encoded in UTF-8. The dataset in UTF-8 encoding with byte order mark (BOM) format is suggested if the content contains non-English word (e.g. Simplified Chinese) such that it can be directly opened by common office tools (e.g. Microsoft Excel). The following example uses Notepad++ for encoding.



3. The last character in the dataset should be an end-of-line (i.e. CRLF or LF). The following example uses Notepad++ to verify the EOL character.

		e data file.csv - Note	pad++						- 0	ı x	7			
	檔案(E) 編輯(E)	〕 搜尋(S) 檢視(V)	編碼(N) 語音(L)	設定① エ	具(O) 巨集(M)	執行(R) 外掛(P) 視窗	(W) <u>2</u>)	<			
保持有	最上層) D C M	hg 🔍 🔍	: 📭 🖬 🚍	e 11 厓 🗷 🕅	🔊 🖿	۲			»»			
숮쭘힁	1	F11												
価格が	r f	F12	24 Million							_	-1			
Dis Divers	*	112	Petra				-							_
特殊导	■元	>	顯示空白字元	〈Space 與 TA	B >		🛒 *D	:\Sample	data file.cs	w - Notep	ad++			
縮放		>	✓ 顯示行屬字元	(EOL)			檔案(E) 編輯(E)	搜尋(S)	檢視Ѹ	編碼(N)	語言(L)	設定(T) 工具
搬移/	重現目前文件	>	顧示所有字元						🗟 📭 🚔	l k B	n 💼 l a		hg (2 2
頁鏡		>	✓ 顧示編排輔助約	<u>a</u>		1	- char	non la gr Eil	Sample d	ata fila cou				_
自動	行		顯示斷行字元				1	- 11 Sources	the second se	こ志口/画				
主副視	電切換	F8					2	(HK\$)"	」「「「「「「「「「「」」「「」」「「」」「「」」「「」」「「」」「」」「」」	문데에	1200			
講義者	≇ B▽ 2〒	Alt+H					3	0-999,	869	0				
1.00.000.00							4	1000-1	1099,255	40000				
折壘月	有層次	Alt+0					5	2000-2	2999,384	5 CRIE				
伸展所	有層次	Alt+Shift+0					6	3000-3	3999,315	4 CRIE				
折壘明	見行層次	Ctrl+Alt+F						4000-4	1999,299					
伯里利	行軍力	Ctrl+Alt+Shift+F					i i	6000-6	5999 145	60303				
14/202		CurrAicronnerr					10	7000-7	7999,194	5 GRUG				
以377/目	Ξ.Ψ.	,					11	8000-8	9999,124	3 CRIE				
展開層	數	>					12	9000-9	9999,861	CREF				
(2) 5+ (5	Ŧ						13	10000-	-19999,7	12 CRUE				
AVE 1	4.3K		9				14	20000-	-29999,4	840305				
専案		>					15	30000-	40000 1					
文件す	る工作區(W)		Ln:24	Col:1 Sel:0	0	Unix (LF)	17	50000-	-599999.1	4500000				
文件編							18	60000-	69999,1	02				
13.917	helen kang						19	70000-	-79999,8	5 CRIE				
1990 A	加州加州平		-				20	80000-	-89999,5	2 CRUE				
雙窗雪	自同步捲動						21	90000-	-999999,3	1 CRUE	_	FOI	char	acter
雙窗기	<平同步援動						22	100000)-199999) or sho	,19 08 0		LOL	critari	accer
			-				2.4	200000	OI abo	ve, / e	_			
文字方	5向 ←	Ctrl+Alt+R					1							
文字方	5向→	Ctrl+Alt+L					Norma	l te> lengt	h:394 lin	nes : 24		Ln : 24	Col:1	Sel:0
監視日	甜誌 (tail-f)													

Convert XLS/XLSX Spreadsheet into CSV open format

You may use Microsoft Excel to convert a worksheet with content containing non-English word (e.g. Traditional Chinese, Simplified Chinese) into CSV format. It is suggested to follow the steps below:

- 1. Open the file in Excel, click **File/Save As**. In the **Save As** pop up window for **Save as type**, change **Excel Workbook** to **Unicode Text**.
- 2. Click **Save**. Now you have a text file in which your non English language characters are properly displayed.
- 3. Open the file you saved with a plain text editor, for example, WordPad or Notepad. Note that the file is tab delimited and you need to change it to comma delimited.
 - Highlight any one "tab" character, which is the entire space between 2 columns, and copy that space (i.e. tab).
 - Click **Edit** choose **Replace**. Paste the tab character into Find what. Enter "," (without quotes) in the Replace with box.
 - Click **Replace All**, and exit the dialog.
 - Go to **File** click **Save**. The text file is now comma delimited.
- 4. Go to **File Save As**, change the **Encoding** from **Unicode** to **UTF-8**. For **Save as type**, change from Text Document to **All Files**. For the **File name**, give an extension **.csv**.
- 5. Click **Save**. You have successfully saved the file in .csv with UTF-8 encoding.