Common Practices in Tabular Data Dissemination

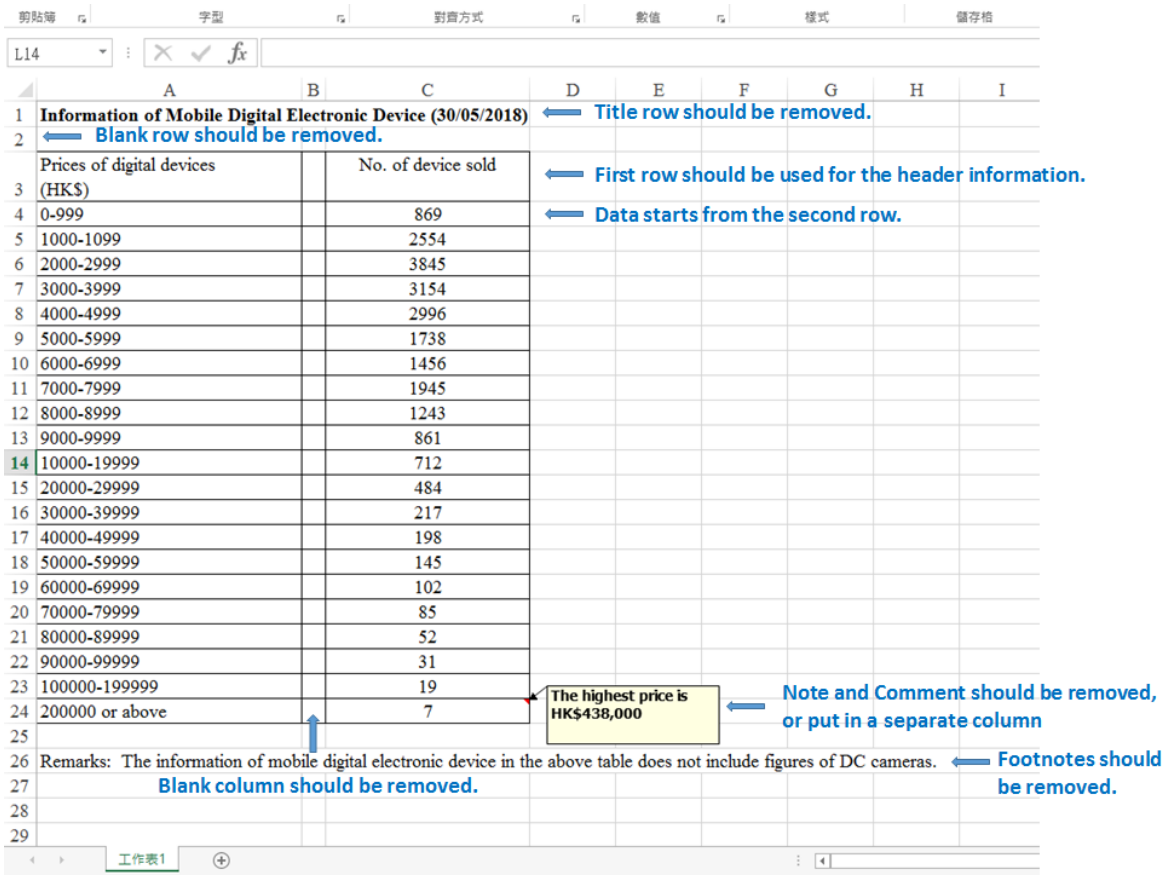
Common practices for all recommended tabular dataset formats

1. The dataset should contain the header and the data only.

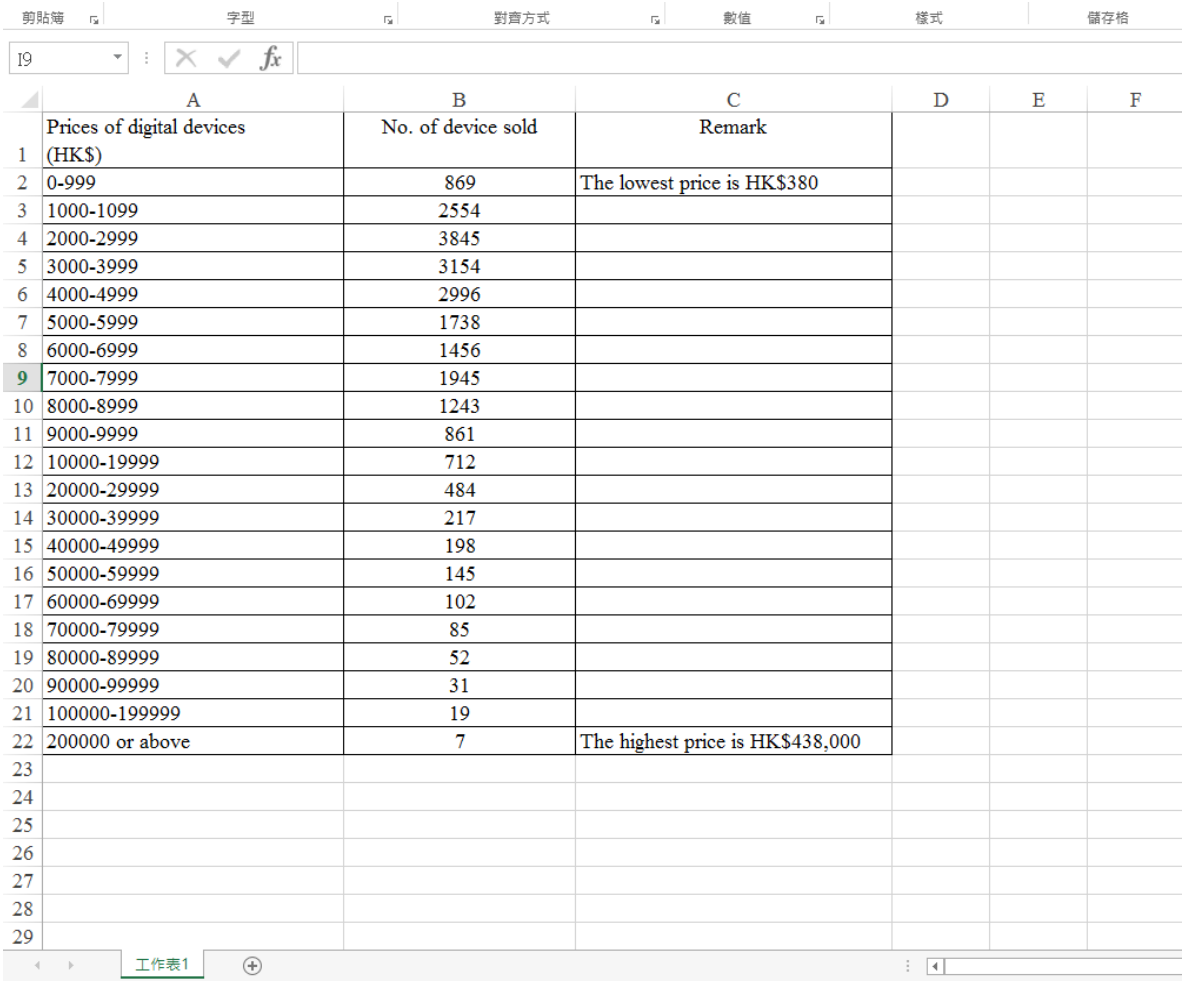
2. Each dataset should contain one and only one tabular data (table). In this way, the structure of the dataset should be unified as follows:
   - The first row of the table should be used for the header information
   - Data starts from the second row
   - Rows and columns should be continuous

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:



Upon applying the common practices above, the dataset should look like this:



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).

	B	C	D	E
1	No. of device sold	Remark	Device Code	
2	869	The lowest price is HK$380	H	
3	2554		H, P	
4	3845		H, P	
5	3154		H, P	
6	2996		H, P, M	
7	1738		H, P, M	
8	1456		H, P, M	
9	1945		H, P, M	
10	1243		H, P, M	
11	861		H, P, M	
12	712		H, P, M	
13	484		P, M	
14	217		P, M	
15	198		M	
16	145		M	
17	102		M	
18	85		M	
19	52		M	
20	31		M	
21	19		M	
22	7	The highest price is HK$438,000	M	
23				
24				
25				
26				
27				
28				
29				

工作表1

Data Dictionary Document

Device Code

H - Mobile Phone
P - Tablet
M – Notebook

.....
.....

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.

7. 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.

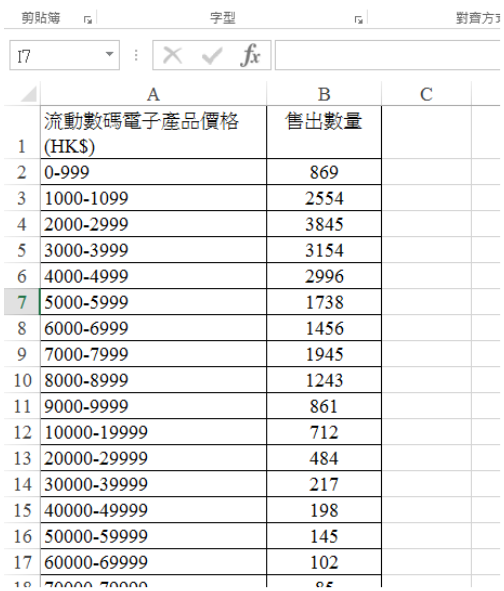
11. For data with address location and geo-location, the following practices are suggested:

- 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)

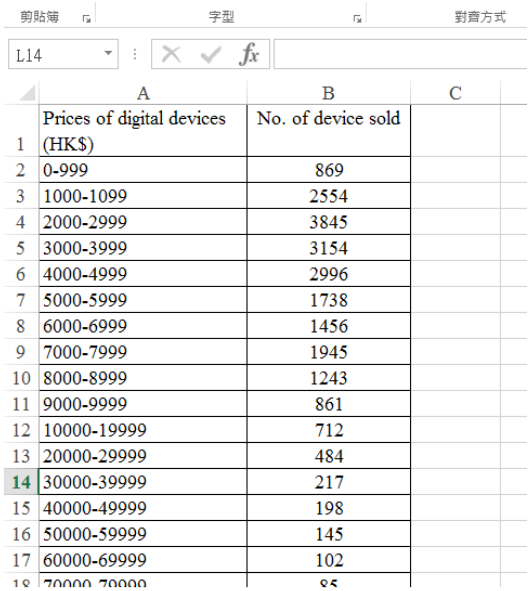
If there are multilingual versions of datasets (e.g. English, Traditional Chinese and Simplified Chinese), data providers should use separate files for each version.



	A	B	C
1	流動數碼電子產品價格 (HK$)	售出數量	
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	

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)



	A	B	C
1	Prices of digital devices (HK$)	No. of device sold	
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	

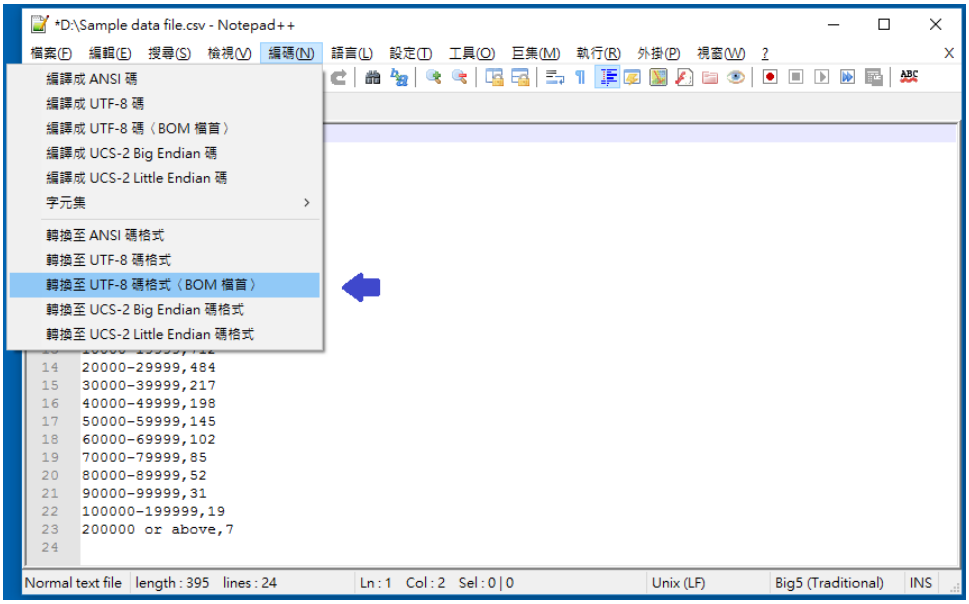
Resource Description defined in psi-data.json

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)

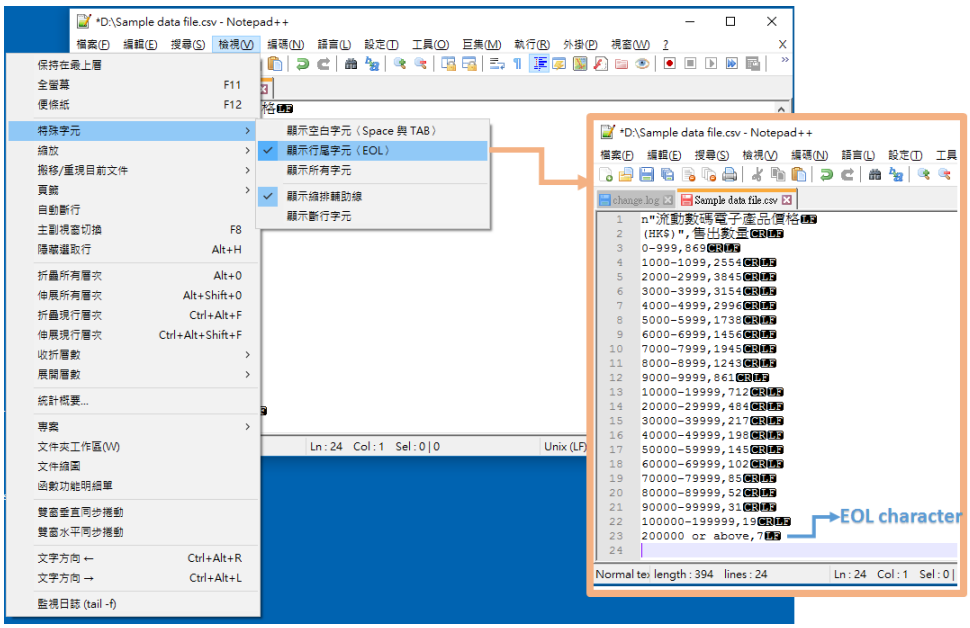
Common practices for dataset in CSV open format

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.



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.