新潟市オープンデータ

CSV ファイル変換・作成マニュアル

第1.1 版

新潟市

■目次

改版履歴

改版日	版数	内容
平成 26 年 12 月 24 日	1.0	初版
平成 27 年 12 月 15 日	1.1	・外部公開にあわせ、発行者を変更 （ｍ－ガバメント戦略タスクフォース ⇒ 新潟市） ・指針 8 を加筆。 （半角カナの除去、全角・半角混在の回避） ・指針 9 にイメージ画像を追加。

統計表などをオープンデータ化する処理方法

目的　→　機械判読に適した、指針に合ったオープンデータ用の
ＣＳＶファイルを作成する。

オープンデータ化したＣＳＶの概要

機械判読とは、コンピュータが自動的にデータの再利用（加工、編集等）ができることです。

コンピュータが自動的にデータを再利用するためには、コンピュータが、データの構造を識別でき、構造中の値（数値、文字等）が処理できるようになっている必要があります。

例えば、ホームページに統計表が画像データや PDF 形式のデータで公開されていたとします。このデータをコンピュータに解析させるには、事前に人間がその画像のデータを表計算ソフトに入力して保存または画像認識等の技術により、公開データから数値やテキストを得て、それをコンピュータに与える必要があります。

データをコンピュータに解析させる作業を効率化するには、情報提供者が、提供するデータについて、コンピュータが数値等を入手しやすい形式に変換し、コンピュータの解析に必要な利用者のコストをできるだけ軽減することが必要です。このような、コンピュータが数値や文字を抽出しやすい形式のデータを「機械判読に適したデータ」といいます。

実際の作業では、指針に適したＣＳＶファイルへの編集・変換を行います。

新潟市が毎月公表している統計データ（住民基本台帳人口など）について、毎月手作業で「機械判読に適した処理」を行うことは非常に困難です。

毎月または毎年同じ表形式で公表しているものであれば、エクセルなどで一度プログラム化してしまえば、次回から簡単にオープンデータ化の処理ができます。

表を**画像やＰＤＦファイル**で公開
↓
コンピュータは**画像など**で表された数値や文字を**認識できない**。
↓
人間が文字・数値データを入力してから、
コンピュータに認識させることができる。

表を**Excel ファイルなど**で公開
↓
コンピュータは、**特定のアプリケーション**がないと、数値や文字を認識できない。
↓
人間が汎用性のあるデータに変換してから、
コンピュータに認識させることができる。

表を**指針に適さない従来のＣＳＶ**で公開
↓
コンピュータは、**形式が整っていない**従来型のＣＳＶデータを、正確に認識できない。
↓
人間が共通の指針に沿った形式に整えてから、
コンピュータに認識させることができる。

表を**指針に適したＣＳＶ**で公開
↓
コンピュータは、**共通の形式・指針に沿ったＣＳＶ**データを正確に認識できる。
↓
指針に適したＣＳＶデータの作成が、
プログラムで自動化できたら便利！

次に、オープンデータに適したデータ処理の要点を挙げます。（指針 1～9）

（指針 1）　一つのファイルは、1 種類の表から構成されます。

図 1 に示すファイルは、複数の表を含んでいます。このようなファイルをコンピュータが判読するためには、表の切れ目を扱う必要があり、判読手順が複雑になります。このため、1 つのファイルは、1 種類の表からのみ構成されなければなりません。

ファイルに含まれる複数の表を分割し、それぞれ別のファイルに格納する処理をします。(図１－２)

図１－１　ファイルに複数の表がある　→　指針 1 を満たさない

ファイル X

月	A 区	B 区	C 区	D 区
1	−4.5	−0.5	1.6	11.3
2	−6.8	−2.1	0.4	8.4
3	−2.4	1.9	3.8	13.5
4	0.2	3.4	6.5	17.3

月	A 区	B 区	C 区	D 区
1	230	58	377	103
2	169	43	422	122
3	144	54	322	144
4	232	102	145	133

Excel なら、複数シートを分離する処理が必要。

図１－２　　図１－１を 2 つのファイルに分割　→　指針 1 を満たす

ファイル X

月	A 区	B 区	C 区	D 区
1	−4.5	−0.5	1.6	11.3
2	−6.8	−2.1	0.4	8.4
3	−2.4	1.9	3.8	13.5
4	0.2	3.4	6.5	17.3

ファイル Y

月	A 区	B 区	C 区	D 区
1	230	58	377	103
2	169	43	422	122
3	144	54	322	144
4	232	102	145	133

Excel 形式の元データ

年齢別地区別人口統計（全市）

人口総数	男	女
804,570	387,360	417,210

年齢	男	女	合計	年齢	男	女	合計	年齢	男	女
0	3,099	3,041	6,140	30	4,541	4,525	9,066	60	5,440	5,422
1	3,201	3,005	6,206	31	4,612	4,627	9,239	61	5,501	5,620
2	3,246	3,065	6,311	32	4,614	4,710	9,324	62	5,844	6,192
3	3,394	3,220	6,614	33	4,895	4,755	9,650	63	6,018	6,473
4	3,307	3,263	6,570	34	4,882	4,991	9,873	64	6,445	6,874
5	3,387	3,283	6,670	35	5,285	5,067	10,352	65	7,029	7,649

市集計　区集計(北区)　区集計(東区)　区集計(中央区)　区集計(西区)　区集計(秋葉区)　区集計(南区)

シートを分割して指針に沿って編集

年	年	月	市区名	年齢[歳]	人口総数[人]	男[人]	女[人]
2014	平成26	10	新潟市	全年齢	804570	387360	417210
2014	平成26	10	新潟市	0～14	100588	51474	49114
2014	平成26	10	新潟市	15～64	494089	247062	247027
2014	平成26	10	新潟市	65以上	209893	88824	121069
2014	平成26	10	新潟市	0	6140	3099	3041
2014	平成26	10	新潟市	1	6206	3201	3005
2014	平成26	10	新潟市	2	6311	3246	3065
2014	平成26	10	新潟市	3	6614	3394	3220
2014	平成26	10	新潟市	4	6570	3307	3263

オープンデータの処理後の CSV ファイルを Excel で表示

```
年,年,月,市区名,年齢[歳],人口総数[人],男[人],女[人]
2014,平成26,10,新潟市,全年齢,804570,387360,417210
2014,平成26,10,新潟市,0～14,100588,51474,49114
2014,平成26,10,新潟市,15～64,494089,247062,247027
2014,平成26,10,新潟市,65以上,209893,88824,121069
2014,平成26,10,新潟市,0,6140,3099,3041
2014,平成26,10,新潟市,1,6206,3201,3005
2014,平成26,10,新潟市,2,6311,3246,3065
2014,平成26,10,新潟市,3,6614,3394,3220
2014,平成26,10,新潟市,4,6570,3307,3263
```

同ファイルをメモ帳で表示

（指針 2）　ヘッダは、1 行の構成にします。

図２－１に示すファイルのヘッダは、2 行からなっています。このようなファイルをコンピュータが判読するためには、ヘッダとデータの切れ目を解釈する必要があり、判読手順が複雑になります。

このため、ヘッダを１行で構成しなければなりません。

ヘッダの内容を統合して１行にまとめれば、指針を満たします。

図２－１
複数行から構成されるヘッダは、指針 2 を満たさない

月	気温			
	A 区	B 区	C 区	D 区
1	−4.5	−0.5	1.6	11.3
2	−6.8	−2.1	0.4	8.4
3	−2.4	1.9	3.8	13.5
4	0.2	3.4	6.5	17.3

「気温」がヘッダか、「A 区、B 区…」もヘッダか、コンピュータには判断できない

図２－２
ヘッダを 1 行に統合します

月	A 区の気温	B 区の気温	C 区の気温	D 区の気温
1	−4.5	−0.5	1.6	11.3
2	−6.8	−2.1	0.4	8.4
3	−2.4	1.9	3.8	13.5
4	0.2	3.4	6.5	17.3

←１行だと、コンピュータは、ヘッダだと認識できる

（指針 3）　データでない情報を、フィールドに含めません。

図３－１に示すファイルの C 区１月の値には、「1.6」という数値と「(＊1)」という注釈へのリンクが含まれています。ここで、注釈へのリンクである(＊1)と、その先にある注釈文は、コンピュータには判読できません。

機械判読性を高めるためには、注釈へのリンクである「(＊1)」を除き、数値「1.6」のみとしなければなりません。

図３－１　データでない情報がレコードに含まれている　→　指針 3 を満たさない

月	A 区	B 区	C 区	D 区
1	−4.5	−0.5	**1.6(*1)**	11.3
2	−6.8	−2.1	0.4	8.4
3	−2.4	1.9	3.8	13.5
4	0.2	3.4	6.5	17.3

＊１　注釈○○○○○○○○○○○○○

図３－２　データでない情報を除去する　→　指針 3 を満たす

月	A 区	B 区	C 区	D 区
1	−4.5	−0.5	1.6	11.3
2	−6.8	−2.1	0.4	8.4
3	−2.4	1.9	3.8	13.5
4	0.2	3.4	6.5	17.3

なお、図３－１のような、注釈を含むファイルは、人がデータを解釈するためには必要です。このため、＊１のような注釈文は、機械判読性の高いファイルとは別に提供する必要があります。

例えば、オープンデータが掲載されているホームページ中の「メタ情報」などに記載します。

（指針 4）　全てのフィールドは、他のフィールドと結合されない。

図４－１に示すファイルの「年」の各フィールドが結合されています。人が見れば、この 4 か月のデータが 2013 年のものであることは分かりますが、コンピュータはそれを判読できません。

機械判読性を高めるためには、フィールドの結合を解除し、それぞれ値を記載します。(図４－２　)

図４－１　フィールドが結合されている　→　指針 4 を満たさない

セルを結合 → すると、コンピュータが判読できない。

年	月	A 区	B 区	C 区	D 区
2013	1	−4.5	−0.5	1.6	11.3
	2	−6.8	−2.1	0.4	8.4
	3	−2.4	1.9	3.8	13.5
	4	0.2	3.4	6.5	17.3

図４－２　フィールドの結合を解除する　→　指針 4 を満たす

それぞれの → 行が何年の何月のデータか理解できる

年	月	A 区	B 区	C 区	D 区
2013	1	−4.5	−0.5	1.6	11.3
2013	2	−6.8	−2.1	0.4	8.4
2013	3	−2.4	1.9	3.8	13.5
2013	4	0.2	3.4	6.5	17.3

（指針 5）　値がない場合を除き、フィールドを空白にしない（省略しない）。

図５－１に示すファイルでは、「年」フィールドは行ごとに分割されていますが、第 2 行目以降の記述が省略されています。人が見ればこの 4 か月のデータが 2013 年のものであることは分かりますが、コンピュータはそれを判読できません。

機械判読性を高めるためには、省略されている値を補完しなければなりません。(図５－２)

図５－１　フィールドの値が省略されている　→　指針 5 を満たさない

値を省略　→ すると、コンピュータが判読できない

年	月	A 区	B 区	C 区	D 区
2013	1	−4.5	−0.5	1.6	11.3
	2	−6.8	−2.1	0.4	8.4
	3	−2.4	1.9	3.8	13.5
	4	0.2	3.4	6.5	17.3

図５－２省略されている値を補完する　→　指針 5 を満たす

それぞれの行が何年の何月のデータか理解できる

年	月	A 区	B 区	C 区	D 区
2013	1	−4.5	−0.5	1.6	11.3
2013	2	−6.8	−2.1	0.4	8.4
2013	3	−2.4	1.9	3.8	13.5
2013	4	0.2	3.4	6.5	17.3

（指針 6）　年の値には、西暦表記を備える。

図６－１に示すファイルの「年」の各フィールドは、和暦で記載されています。コンピュータは、数値の大小で年を比較できる方が処理しやすいため、年の値が単調に増加する西暦の方が扱いやすい。

このため、西暦値を追記する。（図６－２）

Ｈ２５、Ｓ２５の表記も和暦の一種になるので、西暦を入れます。

図６－１　和暦で年が記載されている　→　指針 6 を満たさない

年	月	A 区	B 区	C 区	D 区
平成 25	1	−4.5	−0.5	1.6	11.3
平成 25	2	−6.8	−2.1	0.4	8.4
平成 25	3	−2.4	1.9	3.8	13.5
平成 25	4	0.2	3.4	6.5	17.3

図６－２　西暦を付加する　→　指針 6 を満たす

年[西暦]	年[和暦]	月	A 区	B 区	C 区	D 区
2013	平成 25	1	−4.5	−0.5	1.6	11.3
2013	平成 25	2	−6.8	−2.1	0.4	8.4
2013	平成 25	3	−2.4	1.9	3.8	13.5
2013	平成 25	4	0.2	3.4	6.5	17.3

（指針 7）　フィールドの値の単位を明記する。

図７－１に示すファイルには、値の単位が記載されていない。

データの単位（物理単位、貨幣単位等）は、データ処理に必須であるので、ヘッダに単位を付記することによって、フィールドの単位を明記できる。（図７－２　）

図７－１　フィールドの単位が記載されていない　→　指針 7 を満たさない

月	A 区	B 区	C 区	D 区
1	−0.5	−1.5	1.6	2.3
2	−1.8	−2.1	−0.4	1.4
3	2.4	1.9	3.8	3.5
4	4.2	3.4	5.5	7.3

図７－２　ヘッダに単位を付記する　→　指針 7 を満たす

月	A 区[℃]	B 区[℃]	C 区[℃]	D 区[℃]
1	−0.5	−1.5	1.6	2.3
2	−1.8	−2.1	−0.4	1.4
3	2.4	1.9	3.8	3.5
4	4.2	3.4	5.5	7.3

単位は[]ではさむ。
例
[℃][メートル]
[人][%]など

（指針 8）　使ってはいけない文字・記号などを処理する。

①余計な空白を削除する

エクセルなどでは、文字のバランスをとるために文字と文字の間に空白を入れますが、オープンデータ化する際には、余計な空白を除きます。

②コンマ「,」を削除する

CSV 形式では、数値と数値、文字と文字などの間をコンマで区切ります。数値にコンマがあると、コンピュータが「そのコンマがデータの区切りなのか、桁の区切り表示なのか」判断できません。

エクセルなどによる桁区切りの処理で、コンマを削除します。

③機種依存文字は使用しない

機種依存文字は、パソコンやＯＳなどの違いにより、表示が異なります。機種依存文字はあらかじめ別の文字に変えておきます。半角カタカナも全角にしてください。

④全角・半角は混在しない

アルファベット・数字には「全角文字」「半角文字」がありますが、コンピュータ上は「別の文字」として扱われます。そのため、混在していると、データの並び替えや抽出などで誤動作を起こしてしまいます。

少なくとも、どちらか一方の文字に統一してください。半角文字に統一するのがベターです。

（指針 9）　利用している文字コードを明記する。
また、国際的に広く利用されている文字コードを利用する。

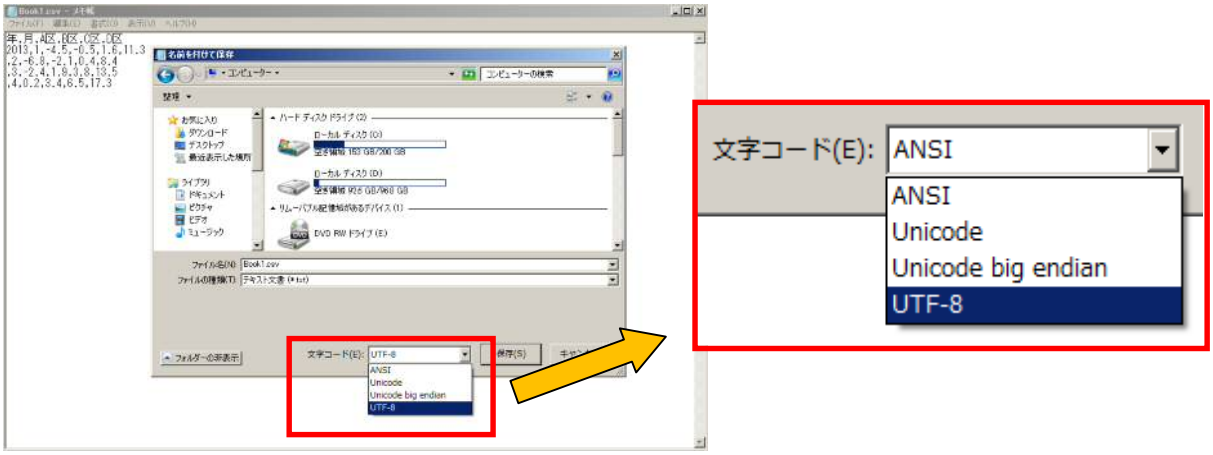
日本語を記述する文字コードには、JIS、Shift-JIS、EUC、UTF-8 等、複数あります。記述されている文字コードが明記されていなければ、コンピュータが読み取ることは困難です。さらに、データの国際的な展開や他の規格との整合を考慮し、「UTF-8」の文字コードを利用します。

現在広く利用されている Microsoft Excel の日本語版は、Shift-JIS で CSV 形式のデータを出力しています。これを UTF-8 に変換する代表的な方法は次のとおり。

上記の指針 1～8 に沿ってエクセルでデータを編集した後、CSV 形式で保存。

↓

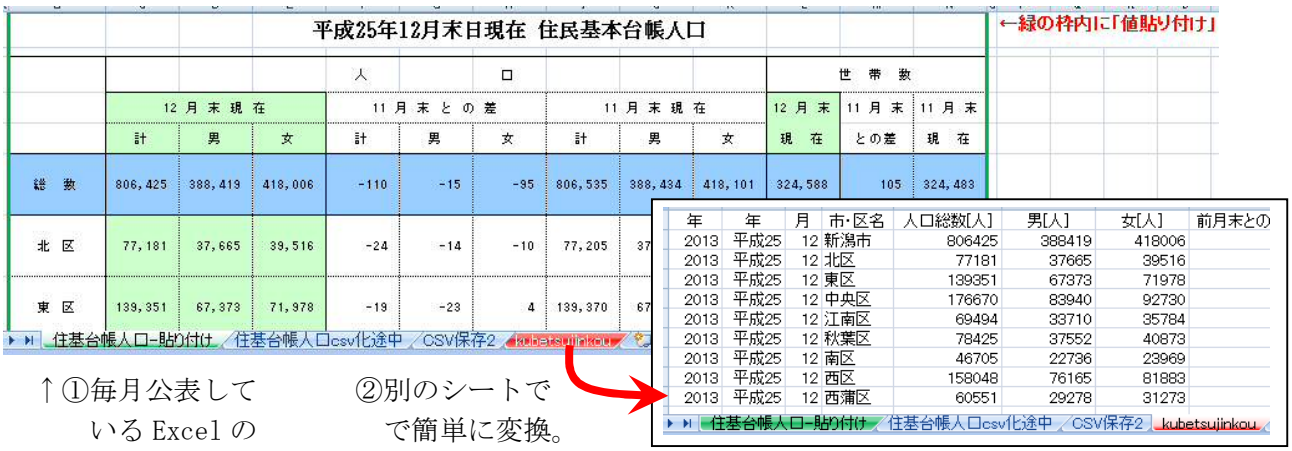
メモ帳でその CSV データを開き、UTF-8 形式で保存する。



（住民基本台帳の例）　一度プログラム化すれば、次回から「コピペ」でデータ変換。

これまで示してきた各指針を反映させる「関数」を組み込んだエクセルファイルを作成・利用すれば、毎月更新が必要なデータを簡単に変換・オープンデータ化できます。

例えば、表形式が定型で、毎月公表している「住民基本台帳人口・世帯数」や「推計人口」のエクセルファイルは、下図のようにコピー・ペーストしてオープンデータ化しています。



平成25年12月末日現在 住民基本台帳人口

←緑の枠内に「値貼り付け」

	人口									世帯数		
	12月末現在			11月末との差			11月末現在			12月末現在	11月末との差	11月末現在
	計	男	女	計	男	女	計	男	女			
総数	806,425	388,419	418,006	-110	-15	-95	806,535	388,434	418,101	324,588	105	324,483
北区	77,181	37,665	39,516	-24	-14	-10	77,205	37				
東区	139,351	67,373	71,978	-19	-23	4	139,370	67				

住基台帳人口-貼り付け　住基台帳人口csv化途中　CSV保存2　kubetsujinkou

年	年	月	市・区名	人口総数[人]	男[人]	女[人]	前月末との
2013	平成25	12	新潟市	806425	388419	418006	
2013	平成25	12	北区	77181	37665	39516	
2013	平成25	12	東区	139351	67373	71978	
2013	平成25	12	中央区	176670	83940	92730	
2013	平成25	12	江南区	69494	33710	35784	
2013	平成25	12	秋葉区	78425	37552	40873	
2013	平成25	12	南区	46705	22736	23969	
2013	平成25	12	西区	158048	76165	81883	
2013	平成25	12	西蒲区	60551	29278	31273	

住基台帳人口-貼り付け　住基台帳人口csv化途中　CSV保存2　kubetsujinkou

↑①毎月公表しているExcelのデータを値貼り付けする。

②別のシートでで簡単に変換。
あとはCSVファイルとして保存。

③メモ帳でcsvファイルを開き、文字コードをUTF-8形式で保存。

……総務課統計係での変換方法

この資料は「オープンデータ流通推進コンソーシアム　オープンデータガイド第1版　2014年7月31日」および「一般社団法人オープン＆ビッグデータ活用・地方創生推進機構　オープンデータガイド第 2 版 2015 年 7 月 30 日」などを参考に作成しています。