オープンデータ化のためのCSV形式データ規格

資料3-6

オープンデータ流通推進コンソーシアム

2013.02.26.

目　次

[1 概要 2](#_Toc349472109)

[2 用語定義 3](#_Toc349472110)

[3 規定範囲 5](#_Toc349472111)

[3.1 表形式データ 5](#_Toc349472112)

[3.2 表形式データの構成要素 6](#_Toc349472113)

[3.3 レベル1表形式データ 8](#_Toc349472114)

[4 参照規定 9](#_Toc349472115)

[5 表記規格 10](#_Toc349472116)

[5.1 @Caption: データセットのキャプション 10](#_Toc349472117)

[5.2 @Creator: データセットの作成者 10](#_Toc349472118)

[5.3 @Date: データセットの公開日 10](#_Toc349472119)

[5.4 @Language: データセットの基本言語 11](#_Toc349472120)

[5.5 @@Title: タイトル行 11](#_Toc349472121)

[5.6 @@Unit: カラムの単位 11](#_Toc349472122)

[5.7 @@Baseval: カラムの記数単位 11](#_Toc349472123)

[5.8 @@Datatype: カラムのデータタイプ 11](#_Toc349472124)

[付録A 本規格によるデータ生成例 12](#_Toc349472125)

[A.1 基本的な整形 12](#_Toc349472126)

[A.2 ヘッダを追記 14](#_Toc349472127)

[A.3 CSV形式で出力 16](#_Toc349472128)

# 概要

本書は，政府自治体，企業等，さまざまな組織が保持する表形式データをオープンデータ化するための技術規格の1つである．

本書は，CSV形式のファイルを対象として，オープンデータ流通推進コンソーシアム「オープンデータ化のためのデータ作成に関する技術ガイド」[6]が規定する表形式データ・レベル1を満たすデータ（人間が、機械にそれを解読させるための方針を立てられるデータ）を，表形式データ・レベル2を満たすデータに整形するために，CSVファイルの冒頭に記述するヘッダを規定する．

本書が規定するヘッダは，データセットのキャプションや作成者，タイトル，単位に代表されるメタデータである．

# 用語定義

本書が使用する用語の定義を，表 1に示す．

表 1: 用語定義

|  |  |
| --- | --- |
| **用語** | **意味** |
| 二次利用 | 公開されているデータを引用・転載・コピーして利用すること． |
| 機械可読 | コンピュータプログラムに代表される機械が，データを自動的に解読し，技術的に二次利用できること． |
| オープンデータ化 | データを，機械可読な形式に加工すること． |
| 表形式データ | 行と列の，縦横2次元状に配列されたデータ． |
| メタデータ | あるデータ自身について記述した，抽象度の高い付加データ． |
| オープンな形式 | 仕様が公開されており，それが国際標準化団体によって標準化されている形式． |
| 表計算ソフト | 数値データの集計・分析に用いられるアプリケーションソフトウェア．画面上に格子状のマス目を表示し，そのマス目にデータを入れることにより表を作成する機能を有する． |
| キャプション（Caption・表題） | 表形式データ全体を表す短い説明． |
| カラム（Column） | 表形式データの，縦方向の列． |
| ロウ（Row） | 表形式データの，横方向の行． |
| セル（Cell） | 表形式データの各項目．表計算ソフトでは，個々のマス目として表現される． |
| データセル | 表形式データにおいて，データ本体の値が格納されるべきセル． |
| タイトル（Title・題目） | 表形式データの，各カラムの冒頭．カラムに含まれるデータセルの内容や単位を説明する． |
| タイトル行 | タイトルが配置された行． |
| データセット（Dataset） | 機械がセルを取得する対象となる，表形式データの基本単位．表計算ソフトでは，1シートにあたる．CSV形式ファイルでは，1ファイルにあたる． |
| テーブル（Table・表） | 1行以上からなるタイトル行，1行以上のデータセル，0行以上の脚注からなる，セルの集合． |
| 脚注 | 表形式データに付与する，タイトルやデータセルに対する補助説明． |
| 脚注番号 | タイトルやデータセルに付与する，脚注と結びつけるための番号． |
| 単位 | 数値の基準となる，約束された一定量．「m」「g」に代表される物理単位，または「円」「ドル」に代表される貨幣単位である． |
| 記数単位 | データセルの値の桁を示す数．たとえば，単位が「百万円」である場合，記数単位は「1,000,000」である．実際の値は，データセルの値に記数単位を乗じたものである． |
| データ型 | 機械が扱うデータの形式．文字列型・整数型・実数型・日付型などを指す． |
| CSV（Comma Separated Values）**エラー! 参照元が見つかりません。** | 表形式数値データの行を改行で区切り，セルを半角のカンマ「,」で区切る，テキストデータの表現形式．RFC 4180[2]により標準化されている． |

# 規定範囲

本文書は，[6]が規定する，レベル1表形式データを対象とする．

## 表形式データ

表形式データとは，行と列の縦横2次元状に配列されたデータである．統計の結果，あるいは計測の結果得られた数値データである．

表形式データの例を図 1に挙げる．

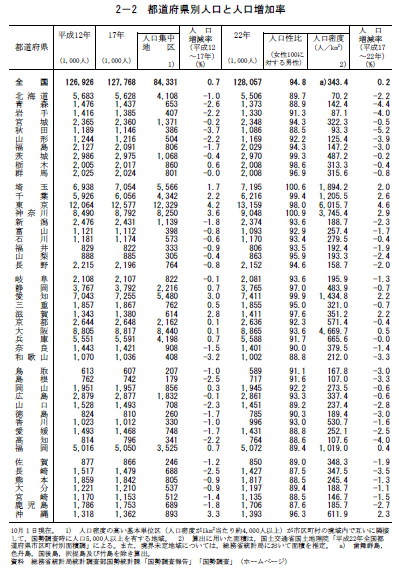


図 1: 表形式データの例[[1]](#footnote-1)

## 表形式データの構成要素

表形式データを構成する各要素の名称を，以下の通り定義する．

* キャプション（表題）:
  + 表形式データ全体を表す短い説明．
* カラム（Column）:
  + 表形式データの，縦方向の列．
* セル（Cell）:
  + 表形式データの各項目．表計算ソフトでは，個々のマス目として表現される．
* データセル（Data Cell）:
  + 表形式データにおいて，データ本体数値が格納されるべきセル．
* タイトル（Title・題目）:
  + 表形式データの，各カラムの冒頭．カラムに含まれるデータセルの内容や単位を説明する．
* タイトル行:
  + タイトルが配置された行．
* データセット（Dataset）:
  + 表形式データの，各カラムの冒頭．カラムに含まれるデータセルの内容や単位を説明する．
* テーブル（Table・表）:
  + 1行以上からなるタイトル行，1行以上のデータセル，0行以上の脚注からなる，セルの集合．
* 脚注:
  + 表形式データに付与する，タイトルやデータセルに対する補助説明．
* 脚注番号:
* タイトルやデータセルに付与する，脚注と結びつけるための番号．
* 単位:
  + 数値の基準となる，約束された一定量．「m」「g」に代表される物理単位，または「円」「ドル」に代表される貨幣単位である．
* 記数単位:
  + データセルの値の桁を示す数．たとえば，単位として「百万円」と書かれているカラムの記数単位は「1,000,000」である．実際の値は，データセルの値に記数単位を乗じたものである．



図 2: 表形式データの各要素の名称定義

## レベル1表形式データ

レベル1表形式データとは，オープンデータ流通推進コンソーシアム「オープンデータ化のためのデータ作成に関する技術ガイド」[6]が規定する，以下の要件を全て満たす表形式データである．

|  |  |
| --- | --- |
| 表全体に対する要件 | |
| (1) | 1つのデータセットには，1つの表のみを含むべきである．（複数個の表を含めるべきではない） |
| (2) | セルに，整形のためのスペース・改行，および位取りのためのカンマを含むべきではない． |
| (3) | 年の値には，西暦表記を備えるべきである． |
| (4) | 数値やタイトル・単位以外の情報を，セルに含めるべきではない． |
| セルに関する条件 | |
| (5) | すべてのセルが，他のセルと結合されているべきではない． |
| (6) | 値が存在しない場合を除き，データセルを空白にするべきではない（データ表記を省略すべきではない）． |
| タイトルに関する条件 | |
| (7) | データセルの内容を示すタイトルは，1行で構成すべきである． |
| (8) | データの単位を明記すべきである． |
| (9) | データセルの内容，単位，記数単位を示すタイトルを，それぞれ別の行に記載すべきである． |

# 参照規定

1. Crocker, D. and P. Overell. Augmented BNF for Syntax Specifications: ABNF. RFC 5234, 2008.  
   http://www.ietf.org/rfc/rfc5234.txt
2. Y. Shafranovich. Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC 4180, 2005.   
   http://www.ietf.org/rfc/rfc4180.txt
3. Codes for the representation of names of languages – Part 1: Alpha-2 code. ISO 639-1:2002, 2002.
4. Data elements and interchange formats – Information interchange – Representation of dates and times. ISO 8601:2004, 2004.
5. Paul V. Biron, Kaiser Permanente, and Ashok Malhotra. XML Schema Part 2: Datatypes Second Edition. W3C Recommendation, 2004.  
   http://www.w3.org/TR/xmlschema-2/
6. オープンデータ流通推進コンソーシアム．データ公開に関する技術ガイド．2013．

# 表記規格

本章は，表形式データのキャプション，タイトル，単位などのメタデータを，CSVデータに付与する表記法を規定する．表形式データのメタデータは，データセルの先頭に，表 2に示すヘッダを利用して付与する．

ヘッダは”@”または”@@”から始める．”@”で始まるヘッダに対する値は，その行に記述する．”@@”で始めるヘッダに対する値は，次の行に記述する．

表 2: 本文書が規定するヘッダ

|  |  |
| --- | --- |
| **ヘッダ** | **意味** |
| @Caption | データセットのキャプション |
| @Creator | データセットの作成者 |
| @Date | データセットの公開日 |
| @Language | データセットの基本言語 |
| @@Title | タイトル行 |
| @@Unit | カラムの単位 |
| @@Baseval | カラムの記数単位 |
| @@Datatype | カラムのデータタイプ |

それぞれのヘッダについての詳細を，以下に記す．

## @Caption: データセットのキャプション

@Captionは，データセットのキャプションを記述するヘッダである．@Caption，キャプション名，言語コードの3つのセルからなる．

言語コードは省略可能であり，省略した場合は，@Languageヘッダが指定する言語コードが指定されたものとする．言語コードはISO639-1[3]に基づく値である．

## @Creator: データセットの作成者

@Creatorは，データセットの作成者を記述するヘッダである．@Creator，作成者名，言語コードの3つのセルからなる．

言語コードは省略可能であり，省略した場合は，@Languageヘッダが指定する言語コードが指定されたものとする．言語コードはISO639-1[3]に基づく値である．

## @Date: データセットの公開日

@Dateは，データセットの公開日を記述するヘッダである．@Date，公開日の2つのセルからなる．公開日はISO 8610[4]に基づく値である．

## @Language: データセットの基本言語

@Dateは，データセットの言語を記述するヘッダである．@Language，言語コードの2つのセルからなる．言語コードはISO639-1[3]に基づく値である．

## @@Title: タイトル行

@@Titleは，タイトル行を記述するヘッダであり，2行で構成される．

このヘッダの1行目は@@Title，言語コードの2つのセルからなる．

言語コードは省略可能であり，省略した場合は，@Languageヘッダが指定する言語コードが指定されたものとする．言語コードはISO639-1[3]に基づく値である．

このヘッダの2行目は，各タイトル名である．

## @@Unit: カラムの単位

@@Unitは，カラムの単位を記述するヘッダであり，2行で構成される．

このヘッダの1行目は@@Unit，言語コードの2つのセルからなる．言語コードは省略可能であり，省略した場合は，@Languageヘッダが指定する言語コードが指定されたものとする．言語コードはISO639-1[3]に基づく値である．

このヘッダの2行目は，各カラムの単位である．単位に記数単位を含めてはならない．物理単位のべき乗数は，そのままテキストで記述する，たとえば加速度の単位「m/s2」は，「m/s2」と記述する．

## @@Baseval: カラムの記数単位

@@Unitは，カラムの記数単位を記述するヘッダであり，2行で構成される．

このヘッダの1行目は@@Basevalである．

このヘッダの2行目は，各カラムの記数単位である．値を省略した場合，「1」が指定されたものと見なす．

## @@Datatype: カラムのデータタイプ

@@Unitは，カラムのデータタイプを記述するヘッダであり，2行で構成される．

このヘッダの1行目は@@Datatypeである．

このヘッダの2行目は，XML Schema[5]に基づくデータタイプ値である．

# 付録A 本規格によるデータ生成例

本章では，図 3を例に，本規格によるCSVデータの生成手順を示す．



図 3: オリジナルデータ[[2]](#footnote-2)

## A.1 基本的な整形

まず，「3.3 レベル1表形式データ」節に記した要件を満たすように，図 3を整形する．具体的には，以下の作業を行う．

* 不必要なスペース・改行・カンマを除く．
* 年の値を西暦表記にする．
* キャプション・脚注・脚注番号を除く．
* タイトルを1行化する．
* データの単位や記法単位を，タイトルの後に移行する．

整形の結果を，図 4に示す．

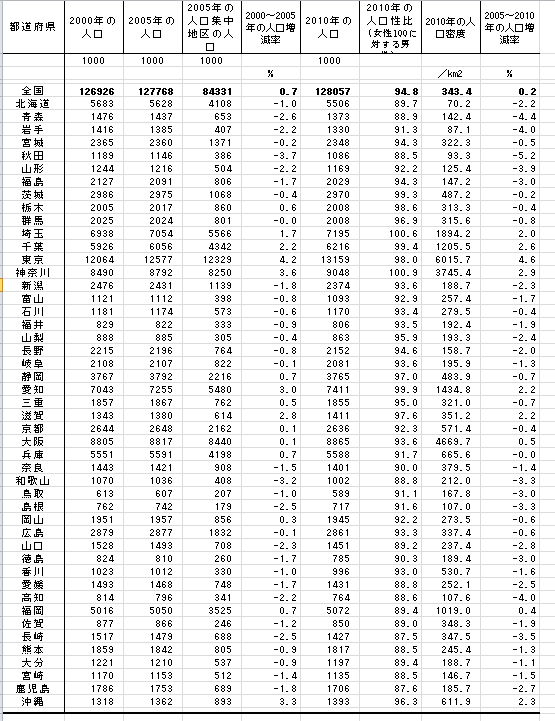


図 4: 図 3を整形した結果

## A.2 ヘッダを追記

図 4のデータセルの上部に，本文書が規定するヘッダを追記する．

今回の例では，表 3に示すヘッダを追記する．

表 3: 追記するヘッダ

|  |  |
| --- | --- |
| **追記するヘッダ** | **値** |
| @Caption | 都道府県別人口と人口増加率 |
| @Creator | 総務省統計局 |
| @Date | 2010-10-01 |
| @Language | ja |
| @@Title | （カラムのタイトルの上に「@@Title,ja」という行を追加） |
| @@Baseval | （カラムのタイトルの下に「@@Baseval」という行を追加） |
| @@Unit | （記数部分を記した行の上に「@@Unit」という行を追加） |
| @@Datatype | * 第1列はxsd:string * 整数値の列はxsd:integer * それ以外の列はxsd:double |

Microsoft Excelでヘッダを追記する際には、以下の点に留意する。

1. ヘッダ名の先頭に、アポストロフィ「’」をつける。  
   これは、@で始まる文字列をMicrosoft Excelが関数と判断するためである。
2. 日付値「2010-10-01」の先頭にアポストロフィ「’」をつける。  
   これは、Microsoft Excelが日付値を自動的に「2010/10/01」等の形式に変換してしまうためである。

追記した結果を図 5に示す．

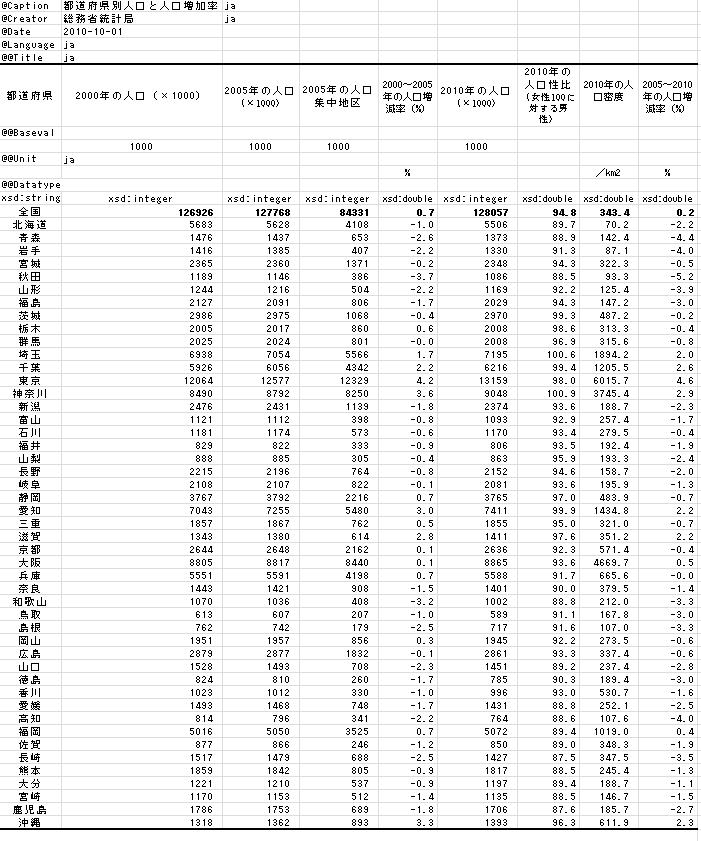


図 5: ヘッダを追記

## A.3 CSV形式で出力

図 5のファイルをCSV形式で出力する．その結果は図 6の通りである．

|  |
| --- |
| @Caption,都道府県別人口と人口増加率,ja,,,,,,  @Creator,総務省統計局,ja,,,,,,  @Date,2010-10-01,,,,,,,  @Language,ja,,,,,,,  @@Title,ja,,,,,,,  都道府県,2000年の人口,2005年の人口,2005年の人口集中地区の人口,2000～2005年の人口増減率,2010年の人口,2010年の人口性比（女性100に対する男性）,2010年の人口密度,2005～2010年の人口増減率  @@Baseval,,,,,,,,  ,1000 ,1000 ,1000 ,,1000 ,,,  @@Unit,ja,,,,,,,  ,,,,%,,,／km2,%  @@Datatype,,,,,,,,  xsd:string,xsd:integer,xsd:integer,xsd:integer,xsd:double,xsd:integer,xsd:double,xsd:doule,xsd:double  ,,,,,,,,  全国,126926 ,127768 ,84331 ,0.7,128057 ,94.8,343.4 ,0.2  北海道,5683 ,5628 ,4108 ,-1.0,5506 ,89.7,70.2 ,-2.2  青森,1476 ,1437 ,653 ,-2.6,1373 ,88.9,142.4 ,-4.4  岩手,1416 ,1385 ,407 ,-2.2,1330 ,91.3,87.1 ,-4.0  宮城,2365 ,2360 ,1371 ,-0.2,2348 ,94.3,322.3 ,-0.5  秋田,1189 ,1146 ,386 ,-3.7,1086 ,88.5,93.3 ,-5.2  山形,1244 ,1216 ,504 ,-2.2,1169 ,92.2,125.4 ,-3.9  福島,2127 ,2091 ,806 ,-1.7,2029 ,94.3,147.2 ,-3.0  茨城,2986 ,2975 ,1068 ,-0.4,2970 ,99.3,487.2 ,-0.2  栃木,2005 ,2017 ,860 ,0.6,2008 ,98.6,313.3 ,-0.4  群馬,2025 ,2024 ,801 ,-0.0,2008 ,96.9,315.6 ,-0.8  埼玉,6938 ,7054 ,5566 ,1.7,7195 ,100.6,1894.2 ,2.0  千葉,5926 ,6056 ,4342 ,2.2,6216 ,99.4,1205.5 ,2.6  東京,12064 ,12577 ,12329 ,4.2,13159 ,98.0,6015.7 ,4.6  神奈川,8490 ,8792 ,8250 ,3.6,9048 ,100.9,3745.4 ,2.9  新潟,2476 ,2431 ,1139 ,-1.8,2374 ,93.6,188.7 ,-2.3  富山,1121 ,1112 ,398 ,-0.8,1093 ,92.9,257.4 ,-1.7  石川,1181 ,1174 ,573 ,-0.6,1170 ,93.4,279.5 ,-0.4  福井,829 ,822 ,333 ,-0.9,806 ,93.5,192.4 ,-1.9  山梨,888 ,885 ,305 ,-0.4,863 ,95.9,193.3 ,-2.4  長野,2215 ,2196 ,764 ,-0.8,2152 ,94.6,158.7 ,-2.0  岐阜,2108 ,2107 ,822 ,-0.1,2081 ,93.6,195.9 ,-1.3  静岡,3767 ,3792 ,2216 ,0.7,3765 ,97.0,483.9 ,-0.7  愛知,7043 ,7255 ,5480 ,3.0,7411 ,99.9,1434.8 ,2.2  三重,1857 ,1867 ,762 ,0.5,1855 ,95.0,321.0 ,-0.7  滋賀,1343 ,1380 ,614 ,2.8,1411 ,97.6,351.2 ,2.2  京都,2644 ,2648 ,2162 ,0.1,2636 ,92.3,571.4 ,-0.4  大阪,8805 ,8817 ,8440 ,0.1,8865 ,93.6,4669.7 ,0.5  兵庫,5551 ,5591 ,4198 ,0.7,5588 ,91.7,665.6 ,-0.0  奈良,1443 ,1421 ,908 ,-1.5,1401 ,90.0,379.5 ,-1.4  和歌山,1070 ,1036 ,408 ,-3.2,1002 ,88.8,212.0 ,-3.3  鳥取,613 ,607 ,207 ,-1.0,589 ,91.1,167.8 ,-3.0  島根,762 ,742 ,179 ,-2.5,717 ,91.6,107.0 ,-3.3  岡山,1951 ,1957 ,856 ,0.3,1945 ,92.2,273.5 ,-0.6  広島,2879 ,2877 ,1832 ,-0.1,2861 ,93.3,337.4 ,-0.6  山口,1528 ,1493 ,708 ,-2.3,1451 ,89.2,237.4 ,-2.8  徳島,824 ,810 ,260 ,-1.7,785 ,90.3,189.4 ,-3.0  香川,1023 ,1012 ,330 ,-1.0,996 ,93.0,530.7 ,-1.6  愛媛,1493 ,1468 ,748 ,-1.7,1431 ,88.8,252.1 ,-2.5  高知,814 ,796 ,341 ,-2.2,764 ,88.6,107.6 ,-4.0  福岡,5016 ,5050 ,3525 ,0.7,5072 ,89.4,1019.0 ,0.4  佐賀,877 ,866 ,246 ,-1.2,850 ,89.0,348.3 ,-1.9  長崎,1517 ,1479 ,688 ,-2.5,1427 ,87.5,347.5 ,-3.5  熊本,1859 ,1842 ,805 ,-0.9,1817 ,88.5,245.4 ,-1.3  大分,1221 ,1210 ,537 ,-0.9,1197 ,89.4,188.7 ,-1.1  宮崎,1170 ,1153 ,512 ,-1.4,1135 ,88.5,146.7 ,-1.5  鹿児島,1786 ,1753 ,689 ,-1.8,1706 ,87.6,185.7 ,-2.7  沖縄,1318 ,1362 ,893 ,3.3,1393 ,96.3,611.9 ,2.3 |

図 6: CSV形式で出力

1. 総務省統計局刊行，総務省統計研修所編集「日本の統計 2012」内「2-2 都道府県別人口と人口増加率」による．http://www.stat.go.jp/data/nihon/02.htm [↑](#footnote-ref-1)
2. 総務省統計局刊行，総務省統計研修所編集「日本の統計 2012」内「都道府県別人口と人口増加率」による．http://www.stat.go.jp/data/nihon/02.htm [↑](#footnote-ref-2)