目的

仙台市が提供するオープンデータは Excel 形式で提供されており、表として読解することが容易な形式になっています。 しかし、データの利活用を促進するためには、より汎用的で機械可読性の高い形式での提供が望まれます。 特に CSV 形式でフラットなデータを提供することで、多くのデータ分析ツールやプログラミング言語で容易に取り扱うことが可能となります。 本ドキュメントでは、例として「区分住民基本台帳人口」データについて、Excel 形式のオープンデータを機械可読性の高い CSV 形式へ変換するための基本方針と変換ルールを定めます。

基本的なルール

- データは UTF-8 エンコードの CSV 形式で保存し、日本語が含むデータも文字化けなく利用できるようにします。
- 年号は西暦に統一し、日付・年の形式は ISO 8601 に準拠した形式で扱います。
- Excel の空白行、注記行、データ取得に不要なメタデータ行は削除します。
- 各列のヘッダー名は、元の Excel ファイルのヘッダー構造を踏まえつつ、CSV では階層構造を廃し、意味が明確になるよう平坦化します。
- ヘッダー名や各行の意味、算出方法などのメタデータは、CSV 本体とは別に定義書として提供します。

変換例

- 対象ファイル：2-11.区別住民基本台帳人口（Excel 形式）
- 対象期間：当該 Excel ファイルに収録されている前年度

変換方針

データ構造の正規化

Excel 形式の元データでは、以下のような特徴が見られます。

- 列見出しが複数行にわたる階層構造（入れ子構造）となっている
- 行方向に「総数」「区や支所」が混在している
- 年度が縦方向に並び、元号表記と省略表記が混在している

CSV への変換にあたっては、1 行が 1 つの観測単位を表すようにデータ構造を正規化します。

地域区分の整理

元データでは、区や支所ごとにデータがまとめられていますが、CSV 形式では「区名」列を設け、各行に対応する区名を明示的に記載します。 この構造は仙台市が提供するオープンデータの複数のシートに見られます。たとえば「2-4.人口移動」のデータでは、市町村とブロックが行方向に混在していますが、CSV 形式では「地域区分」列などを設けて各行に対応する地域区分を明示的に記載します。「2-4. 人口移動（続き）」であれば都道府県名と地方区分を明示的に記載します。

2-11. 区別住民基本台帳人口 の場合

- 区名
  - 区単位の名称を記載します。
- 地域名
  - 支所が存在する場合は支所名を記載し、支所が存在しない場合は区名をそのまま記載します。

なお、「総数」に該当する集計業は、CSV には含めません。総数は区分データから算出可能であり、データの重複や誤解を防ぐため削除します。

年度の取り扱い

- 元データに含まれる元号表記（平成・令和）は西暦に変換します。
- 数字のみで記載されている年度（例：30, 3）はも同様に西暦に変換します。
- 変換後は、西暦年（YYYY）のみを「年度」列に記載します。

列（カラム）の平坦化

Excel データにおける複数段の列見出しは廃止し、CSV では以下のように意味が明確な単一階層の列として定義します。

CSV の列名	元の Excel で対応する列名	説明
年度	年・区	データが属する年度（西暦）を記載します。
区名	年・区	区単位の名称を記載します。
地域名	（新たに作成）	支所が存在する場合は支所名を記載し、支所が存在しない場合は区名をそのまま記載します。
日本人住民で構成する世帯	世帯数_うち日本人住民で構成する世帯	日本人住民で構成する世帯数を記載します。
外国人住民で構成する世帯	世帯数_うち外国人住民で構成する世帯	外国人住民で構成する世帯数を記載します。
日本人住民と外国人住民で構成する世帯	世帯数_うち日本人住民と外国人住民で構成する世帯	日本人住民と外国人住民で構成する世帯数を記載します。
日本人住民（男）	人口_うち日本人住民_男	日本人男性人口を記載します。
日本人住民（女）	人口_うち日本人住民_女	日本人女性人口を記載します。
外国人住民（男）	人口_うち外国人住民_男	外国人男性人口を記載します。
外国人住民（女）	人口_うち外国人住民_女	外国人女性人口を記載します。

不要データの削除・変換

- Excel での空白行、注記行、メタデータ行は削除します。
- 「総数」に該当する集計行は削除します。
- データ取得に不要な補足情報や注釈は削除します。
- 数値データに含まれるカンマ（,）やその他の非数値文字は削除し、純粋な数値データとして保存します。