

データクレンジング（オープンデータ公開に向けたワーク）

機械判読に適したデータとは、コンピュータが自動的に処理しやすい形式のデータを指し、その代表例が「CSV形式」です¹。CSV形式のデータは、様々なソフトウェアで開き、編集することができます。また、多様なシステムで情報を読み込み、利用することが可能ですが、それには読み込みやすい形に整っている必要があります²。

本ツールは、「機械判読性のあるデータ」とは何かを理解し、データを機械判読可能な形式に置き換える方法を学ぶためのものです。

[illegible]

¹ データをカンマ (,) で区切って保存するシンプルなテキスト形式のファイル。1 行が 1 レコード、各列がカンマで区切られたフィールドになっており、構造がシンプルで機械処理が容易。

² [5starOpendata](https://5stardata.info/ja/)(<https://5stardata.info/ja/>)では、CSV 形式以外のデータ形式とその特徴が紹介されています。

■関連フレームワーク

オープンデータ

■関連スキル

データマネジメント

■ツールレベル

初級

■意義・特徴

上図の ①と②のどちらが機械判読性のあるデータでしょうか？正解は②です。①は表題が付与され、セル結合が施されるなど、人が読みやすいように整理されていますが、主に資料として閲覧するためのものです。一方、②は資料としては必ずしも見やすいとは言えませんが、コンピュータにとっては処理しやすい形式となっています。このようなデータは、データベース構造を持つデータとも表現されます³。

(人が読みやすいデータと機械判読性のあるデータの違い)

- ① 人が読みやすいデータ： 標題や小見出しがあり、情報のグループ化が明確であり、セル結合や色分けでデータ間やグループ間の差異が強調されている、すなわち視覚的に整理されているデータ
- ② 機械判読性のあるデータ： 1行1レコード、1列1項目であり、セル結合や装飾がない、いわば統一された構造を持つデータ

また、①の形式のままでは Excel などですぐにグラフを作成できないことがあります。そのため、②のような形式にデータを整理してからグラフを作成した経験がある方も多いのではないのでしょうか。つまり、②のようなデータが「機械判読性のあるデータ」であり、より活用しやすい形式だといえます。

このワークでは、「機械判読性のあるデータ」とは何かを理解するために、資料として整理された分析しにくいデータを、分析しやすいデータに変換する作業＝データクレンジングを行います。機械判読性のあるデータを理解し、分析用のデータとして適切に準備できるようになることで、そのデータの利用者にとって、より活用しやすいデータとなります。

³ 機械判読性のあるデータについての詳細は、総務省より公開されている「統計表における機械判読可能なデータ作成に関する表記方法」を御覧ください。
(https://www.soumu.go.jp/menu_news/s-news/01toukatsu01_02000186.html)

行政機関等、特に自治体のデータ利活用では、「データを使うこと」が重視されがちですが、データを活用するためには、行政機関等自体もまた「データを提供する側」であることを意識することが重要です。

なお、②のデータは機械にとって扱いやすい形式ですが、①のデータが不要というわけではありません。①のような形式は、人が理解しやすく、資料向けとして適しています。

目的に応じて、データの形式を意識的に使い分けていくことが求められます。

■活用シーン

データを利活用するには、まず利活用可能なデータが存在することが前提となります。行政機関等におけるデータ利活用では、利用・閲覧が庁内に限定されるデータをどのように有効活用するかが重要なポイントとなります。そのため、単にデータを活用するだけでなく、庁内でデータを提供し合うこととデータ利用を促進することの両方が求められます。

機械判読性のあるデータが提供されることによって、利用側でデータクレンジングの手間が省け、データ分析やデータ連携がスムーズになるほか、変換に伴うミスが少なくなり、より正確な活用が可能になります。データの利便性や有用性が向上することで、オープンデータの活用が一層促進されることになります。

本ワークを通じて得られる知識は、データの利用者だけでなく、将来的にデータを提供する立場になり得る全ての行政機関等の職員にとって有益です。

■前提・留意事項

なし

■使い方

（ワークの流れ）

教材①のデータを Excel 等で編集し、②のようなデータベース形式に整え、csv 形式で保存します。

（参考時間配分）

15 分～ ※用意するデータによります。添付データを使用する場合は、この時間が目安になります。

（利用する教材）

①編集ツール（Excel など）

②資料として見せるために作成された Excel データや、Word などのテキストデータで作成

された一覧表など

例：

- ・ つくば市投票所一覧表（7 5 投票区投票所）（html）
[https://www.city.tsukuba.lg.jp/soshikikarasagasu/senkyokanriinkaijimukyoku/gyomua
nnai/1/1/1002455.html](https://www.city.tsukuba.lg.jp/soshikikarasagasu/senkyokanriinkaijimukyoku/gyomua
nnai/1/1/1002455.html)
- ・ つくば市投票所一覧表（7 5 投票区投票所）（Excel）
https://www.city.tsukuba.lg.jp/material/files/group/173/tohyozyo_list20240702.xlsx
- ・ つくば市投票所一覧表（7 5 投票区投票所）（doc）

③正解例データ

- ・ つくば市投票所一覧表（7 5 投票区投票所）（1）（csv）
- ・ つくば市投票所一覧表（7 5 投票区投票所）（2）（csv）

■実績・有用性

機械判読性のあるデータを作成する機会は多くありません。実際に取り組みながら理解を深めていくケースがほとんどです。

このツールは、つくば市のデータ利活用研修で長年の実績があり、繰り返し改善を重ねてきたものです。このツールを活用することで、機械判読性のあるデータ作成の重要性を実感し、実践的に身に付けることができます。

機械判読性のあるデータは、国が定めるオープンデータの要件の一つです。オープンデータを推進するためには、機械判読性の概念を理解し、適切にデータを作成できることが重要となります。

また、行政機関等におけるデータ利活用においては、行政機関等の職員が「データを活用する側」とすると同時に、「データを提供する側」でもあることを意識することが求められます。職員が互いにより利活用しやすいデータを提供することは、庁内での円滑なデータ活用を進める上で欠かせません。本ワークを通じて、その重要性を認識し、実践につなげることを目指します。

■次のステップ

（アドバンス研修・実習）

データの公開レベル分類：データには様々な公開レベルがあることを理解した上で、機械判読性のあるデータを分類・整理し、安全に公開できるようになることを目指します。

（参考文献とその概要）

なし：オリジナルワークのため

（関連フレームワーク等）

なし

■著作者・連絡先

つくば市 家中賢作

朝日航洋株式会社 新井千乃

■ダウンロード資料

- ・ワークシート：以下のシートを格納した Excel ファイル
 - ・人にとって見やすく作られたデータ（例：つくば市投票所一覧）
 - ・正解例データ(1)
 - ・正解例データ(2)
- ・ユーザー向けガイド：データクレンジング

■掲載日

2025 年 3 月 31 日