2019/03/04
長野大学　前川道博

上田市の統計データのオープンデータ化（中間報告）

１．目的

上田市の統計データをオープンデータとして誰もが利用できるようにするため、データベースで管理できるよう正規化してデータベースに格納し、利用しやすい統計データにして公開する。そこに至るための中間段階として、すべての Excel 統計データを CSV 化する。

これと並行しながら、長野県内、全国の自治体が抱える統計のオープンデータ化に向けた課題提起、問題解決の啓発を図る。

２．対象

★上田市の統計　https://www.city.ueda.nagano.jp/shise/toke/

上田市の統計サイトに公開されているすべての Excel 統計ファイル全 837 ファイル（全テーブル数 9337）を対象とする。（これらはすべて CC-BY で公開されており、第三者（長野大学前川ゼミ）が CSV 化して公開することが可能である。

3. CSV 化の考え方

(1) 考え方の諸点

Excel の CSV 化は通常手作業で処理されているが、以下の点で大きな問題がある。

・手作業のため作業が極めて非効率である。加えて変換ミスに気付きにくく、見出しの付け替えなどは作業者の主観が介在して標準化、質の保証が全くなされない。

・オープンデータは機械判読可能なデータとして提供する必要がある。そのため、セマンティックウェブ化、RDF 形式での提供、API の公開が必要である。しかし CSV をその扱いにしても極めて中途半端な処置にすぎなくなるため、それらの対応はデータベース化後の対応とするのが順当である。

・Excel 統計表を CSV 化しても利用者にとっては断片化した年度ごとの統計データを参照するだけでは利用するメリットはほとんどない。

・統計ファイル名、統計テーブル名のネーミング等には正規化が必要である。これらの名称は、公開先のウェブページに記載された名称を初期値として用い、名称は統一的でかつわかりやすい体系にチューニングする。

・作業は常にさかのぼって行えるようにする。データの全面的差し替え、やり直しを可能にする。

・CSV 化は変換規則を明確にした上でプログラムにより自動処理する。これにより作業者

の恣意性やミスが介在しないようにする。

・データベース化が本命の作業である。CSV 化を極力早期に完了させ、データベースに格納するためのデータ構造の正規化を図るのがオープンデータ化の根本的解決につながる作業である。

・当面は前川ゼミで開発したツール（以下）を利用して CSV 化を行う。ただし学生がプログラミングの学習をしながら開発したプログラムで、プログラムは未完成（開発途上）である。そのため、一部のファイルが CSV 未変換のままであるが、当面はそれでよしと割り切る。

(2) 実現目標とする統計データのオープンデータ化（業務フローの革新）

一般に「神エクセル」と揶揄される文書主義・紙媒体依存のデータフローを根本から改め、統計データをデータベースによるマスタ管理に大きく変革させることが、根本的に求められる改善である。

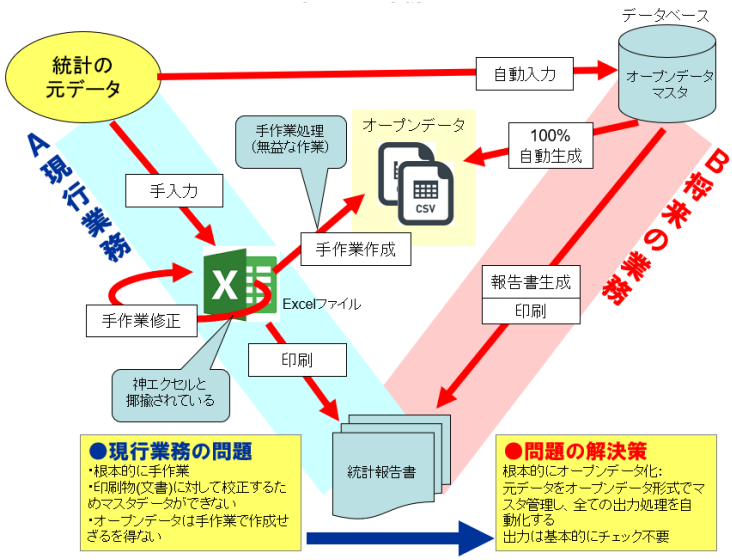

図 1　統計データの業務フロー改善の概念図

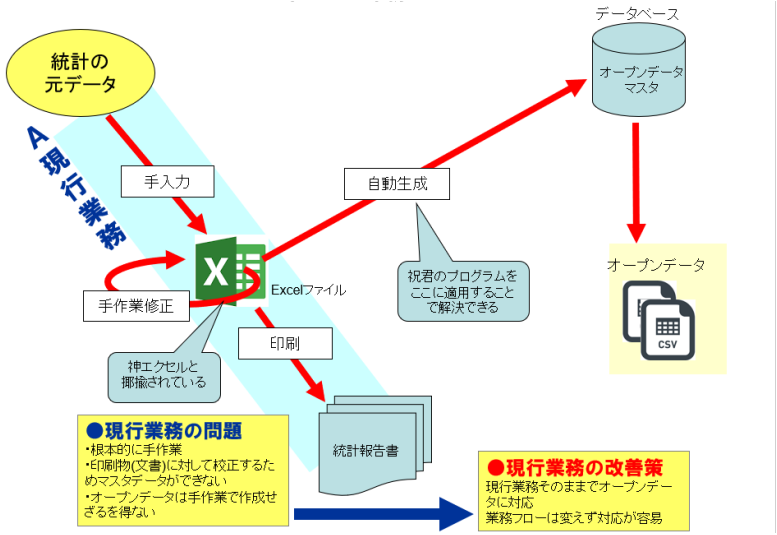

図２　当面の改善（Excel を CSV 化しその公開を図る）

(3) 神エクセル解決の課題調査と解決ツール設計

以下の資料を参照されたい。

★「神エクセル解決アプリ『Excel→CSV 直行便』単なる変換に非ず。オープンデータを促進し業務革新に誘う導火線ツール」解説資料

https://www.mmdb.net/sdc/tokei/files/udc2018app.pdf

4. 上田市の統計 CSV 化の手順

①統計サイトに公開されたすべての Excel ファイルを対象とする。

・Excel ファイルのリストアップ（ファイル名＋タイトル）はプログラムにより自動処理する。

・Excel ファイルリスト（メタデータ）は手作業により編集を施す。

（プログラムによる自動処理も不可能ではないが、手作業で問題点を摘出しながら編集の精度を上げることが最初の段階では必要である。）

・Excel ファイルは②の処理のため最寄りのコンピュータに自動で一括転送する。

②　全ての Excel ファイル＆テーブルを対象に CSV に変換する。

③　処理後の CSV データが正しく変換されているかをチェックし部分修正する。

（修正要のものについては事後、プログラムの処理で処置することを検討する。）

④Excel ファイル、CSV ファイルはデータベースで管理運用する。

・恒常的に公開するサイトとする。

⑤ほどほどに使いやすい公開サービスを構築する。（検索、抽出が行えるもの）

5. 今後の課題

①各テーブルを時系列で編成されたデータ群に正規化する。
②利用者にとって使いやすく役立つ統計データの公開モデルを構築する。
③こうした統計データ活用文化、オープンデータ促進の機運を醸成していく。

6. 行政の統計データの課題

(1) 長野県内における上田市統計データの優越性

・行政の統計データのオープンデータ化の状況をざっくりと調査した。
参考サイト「長野県市町村統計サイト一覧
https://www.mmdb.net/sdc/tokei/app/sites.php
本調査ではオープンデータ化達成度を 8 段階（A～H）で評価した。

A:データベースでマスタ管理　B:ほぼ CSV でも公開　　C:一部 CSV で公開
D:Excel で公開　E:PDF で公開　　F:公開しているが不十分
G:統計サイトはあるが件数が少ない　　H:統計サイトなし

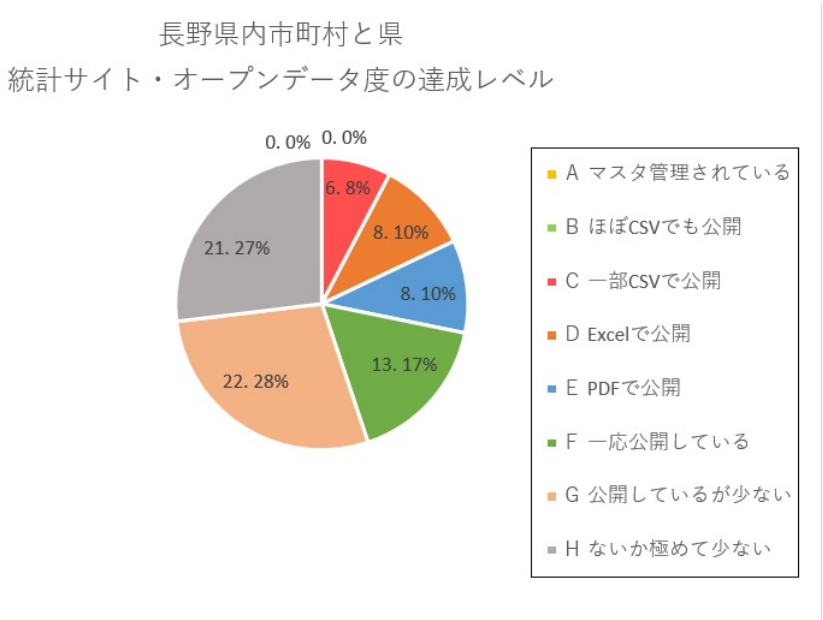

図 3　統計サイトオープンデータ化の達成レベル評価

表 1　統計サイトオープンデータ化のレベル基準と長野県内の状況

達成レベル	数	対象市町村と県
A マスタ管理されている	0	（該当なし）
B ほぼCSVでも公開	0	（該当なし）
C 一部CSVで公開	6	上田市 須坂市 中野市 長野市 松本市 長野県
D Excelで公開	8	安曇野市 阿智村 飯田市 小諸市 坂城町 佐久市 塩尻市 茅野市
E PDFで公開	8	高山村 大町市 岡谷市 下諏訪町 諏訪市 千曲市 東御市 南木曽町
F 一応公開している	13	阿南町 飯山市 伊那市 大桑村 小布施町 軽井沢町 中川村 原村 富士見町 南相木村 南牧村 箕輪町 御代田町
G 公開しているが少ない	22	青木村 朝日村 飯島町 飯綱町 池田町 売木村 王滝村 小川村 川上村 木島平村 木曽町 北相木村 佐久穂町 高森町 辰野町 天龍村 豊丘村 野沢温泉村 白馬村 宮田村 山形村 山ノ内町
H ないか極めて少ない	21	栄村 信濃町 上松町 生坂村 大鹿村 小谷村 麻績村 木祖村 小海町 駒ヶ根市 下條村 喬木村 立科町 筑北村 長和町 根羽村 平谷村 松川町 松川村 南箕輪村 泰阜村
（合計）	78	

この調査（図３、表１）により、長野県内市町村、長野県の統計データは不十分なオープンデータ化の段階に留まっていることがわかった。全体でレベル A、B は皆無の状況である。

長野県統計データは新しい統計サイト「統計ステーションながの」（https://tokei.pref.nagano.lg.jp/）に再編されたが、Excel ファイルの取り出しが極めてしにくい点は大きな課題である。分野別、年度別などの条件を組み合わせ検索する方式のため、必要とするデータをマッチングさせることが難しい。また一度に必要とするデータを取りだすことができない。不自由この上なく、これでは使い物にならない。CC ライセンスも利用者からは極めて確認がしにくいのも難点である。

上田市は長野県、他市町村に比べ、データ、とりわけ Excel ファイルの公開数が格段に多いことがわかった。1972 年からのデータをすべて Excel で公開しているのは上田市のみである。今回、数多くの統計データを CSV 化したことにより、上田市は長野県内ではレベル B に一市のみがレベルアップすることになる。

これらは歴代の統計担当のデータ公開の取り組みがそれだけの蓄積を行った背景にある。今後のデータベース化（レベル A）に向けた方向性に進めることができるデータの十全性がある。長野大学前川ゼミが関わったことにより CSV 化が一気に進んだ。

長野県内の立ち遅れた統計データのオープンデータ化のグッドプラクティス（鑑となる良き実践）となるものであろう。上田市の取り組みはさらに先に進めていくとともに、長野県および県内市町村でも同様に統計データのオープンデータ化を進める契機となることを期待したい。

(2) その他の視点から

・統計データのオープンデータ化は、あらゆるデータが対象のオープンデータ化の one of them である。たとえ小さな取り組みでもこの実績（経験知）があらゆる分野のオープンデータの確実な第一歩となる。

・内閣官房、総務省の要請に応えるだけでは真のオープンデータ利用促進には進み切れない現実を、各自治体、担当セクションで認識してもらう必要がある。
・提供するデータをいかに利用してもらうかを第一に考えた対応を図ってもらいたい。
・小規模自治体、あるいは対応がわからないでいる自治体であっても、その対応が図れるよう最低限の公開をしてもらいたい。（そのための派遣は別途考えたい。）

7. 上田市統計データの CSV 化（実績とデータ）
①対象：上田市の統計　https://www.city.ueda.nagano.jp/shise/toke/
②そのメタデータ化
★統計ファイル一覧：上田市統計
https://www.mmdb.net/sdc/tokei/app/tokei.php
★上田市統計：CSV 変換ファイル：Excel→CSV 直行便による変換検証データ
https://www.mmdb.net/sdc/tokei/app/csvlist.php
③参考：長野県市町村統計サイト一覧とオープンデータ化達成度評価
https://www.mmdb.net/sdc/tokei/app/sites.php
④CSV 化支援ツール群　（前川研究室・ゼミで開発）
行政サイトから Excel ファイルリストを抽出（非公開）　excel_list.pl
Excel ファイルを一括ダウンロード（非公開）　excel_copy.pl
Excel→CSV 直行便
ツール概要（https://www.mmdb.net/sdc/tokei/files/e2c.pdf）
Excel→CSV 変換　excel2csv.py（gitHub からソース公開）
Excel→CSV 一括変換　txt.py（gitHub からソース公開）
その他②の公開用ツール
公開用サイト『統計オープンデータへのいざない』
https://www.mmdb.net/sdc/tokei/
⑤データベース
長野県内統計データベース（兼オープンデータ公開サービス）
信州デジタルコモンズサービス mmdb.net/sdc の DBMS で運用
（mysql を利用　データベース tokei）

7. アーバンデータチャレンジへのチャレンジ
地域課題解決のために公共データ活用をするコンテスト「アーバンデータチャレンジ」で前川ゼミ学生が開発した上記プログラム「Excel→CSV 直行便」がアプリケーション部門の一次選考を通過した。3/16 に開催される最終審査会ではその社会的意義の大きさと実現可能性の高さをＰＲしたい。同時に以下の呼びかけを行い、根本的な問題解決につながるように社会に働きかけていきたい。

8. より根本的な問題解決のために

開発中のプログラムは学生がプログラミングの学習をしながら開発した未完のプログラムである。主開発者が卒業することから、その後のエンハンスはむずかしい。私たちは「行政の統計データのオープンデータ化」という課題にチャレンジし、未熟なプログラニングでもその問題解決に当たれること、また、統計データのオープンデータ化の課題は全国の自治体に共有の課題であることが見えている。その課題解決は全国規模で求められているものである。プログラムのエンハンスないし新規開発に関わっていただける方のご支援をいただきたい。また、こうした課題解決を図りたい自治体の方々と協働ないしは研修などの面での支援を合わせて検討していきたい。