

## 政府統計のデータ提供について —政府統計の総合窓口（e-Stat）の紹介—

独立行政法人統計センター 統計情報・技術部 共同利用システム課  
総合戦略担当 統括統計職 西村 正貴  
にしむら しょうき

### 1. はじめに

日本の政府機関が実施した統計調査（政府統計）の結果は、政府統計の総合窓口（e-Stat）<sup>1</sup>というWeb サービスで一元的に提供している。e-Stat は、2008 年からサービスを開始し、2018 年 1 月にリニューアルを行い、約 600（「住宅・土地・建設」関係では 39）の政府統計の結果を提供している。本稿では、e-Stat における政府統計の結果データの提供の現状及び課題について紹介する。

### 2. 政府統計の総合窓口（e-Stat）の概要

e-Stat のサービスを開始する以前は、統計調査を実施している各府省のホームページなどで各府省の判断により統計結果を提供していたために、提供方法がバラバラだったり、PDF で提供する統計調査もあったりするなど統計結果を利用するうえで不便なところが多々あった。

この問題を解消するために、各府省の協力のもと政府統計共同利用システムを構築し、e-Stat において統計結果を一元的に提供することとなった。政府統計共同利用システムは各府省で共通化できるサービスをまとめるという考え方で構築され、統計結果を提供する機能である e-Stat だけでなく、インターネットから統計調査の回答ができるオンライン調査システムや統計調査を実施する際

に必要な企業系の母集団情報を調査実施機関に提供する事業所母集団データベースなどの機能も提供しており、独立行政法人統計センターが運用を行なっている。

現状の e-Stat では EXCEL ファイルや CSV でのデータ提供が中心となっており、約 600 の統計調査の 135 万ファイルが公開されている。

また、e-Stat では、EXCEL ファイル等だけでなく、データベース化された統計データから、必要な項目を選んで表示又はダウンロードする機能も提供している。EXCEL ファイル等のデータをデータベースに変換する必要があることから、2018 年 11 月現在、本機能の提供は全体の約 600 のうち 123 の統計調査に限られている。

その他、e-Stat では、統計で使用する都道府県・市区町村コードの検索ができる機能、統計の集計で用いる標準的な分類などが検索できる機能、調査項目を検索できる機能などを提供している。

これら e-Stat で提供している各種データは、政府共通ネットワークを通じてオンラインで登録する機能を提供しており、統計調査を所管する各府省がデータを登録する仕組みとなっている。

e-Stat で提供しているデータの利用条件は政府標準利用規約（2.0）に準拠しており、原則、出処明示さえすれば、データの加工、編集、商用利用等に制限はないオープンデータとして利用可能である。

<sup>1</sup> 政府統計の総合窓口（e-Stat）：<https://www.e-stat.go.jp>

### 3. e-Statによる統計データの利用

#### 3.1. 統計データの取得

e-Stat で公開されている統計データを利用する主な方法として、EXCEL ファイル等をダウンロードする方法とデータベースからデータを取得する方法がある。

EXCEL ファイル等は、約 600 統計の結果を公開しており政府統計のほとんどのデータが取得することができるが、ファイルをダウンロードした後で当該ファイルを開かないと統計結果がどのような情報で分類されて集計された結果なのかわからないというデメリットがある。また、EXCEL ファイルの場合は、従来の紙の報告書に近い形で作成されているものが多い。この形式は一覧性があることから人が見るには適しているかもしれないが、必要なデータを抜き出して分析をする場合や、他の分析ソフトなどにデータを取り込む場合などは、データを切り取ったり、配置し直したりする必要がある。さらに、セル接合や項目名にスペースが入っているものもあるため、機械的に変換することが困難な場合がある。

一方、データベースの場合は、統計結果の分類

情報等（メタ情報）と数値情報をデータベース化していることから、東京都の男性の人口のみのデータを取得するなど、メタ情報での絞り込みや行と列の配置を自由に設定でき、必要なデータを簡単に取得することが可能である。また、該当データセットでどのようなメタ情報を使用しているかを参照できるページ（データセット名をクリックすることで表示される）も用意している。データ表示には一定の時間がかかるため、このページを利用して目的のデータを探すことをお勧めする。

なお、目的のデータを e-Stat から探す手順として、まずは目的の統計データがどの統計調査で提供されているか特定することから始めることをお勧めする。なぜなら、その統計データが、何を目的に調査が実施され、どういった者を対象にした調査の結果であるのかを理解した上で統計データを利用しないと、誤った判断をする可能性があるからである。また、それぞれの統計データを表す用語（メタ情報）が同じでも統計調査によっては内容が異なることがあるので注意する必要がある。

各統計調査は「人口・世帯」、「労働・賃金」など 17 の分野で区分されている。e-Stat ではこの



図 1 統計分野での絞り込み例

分野による絞り込みが可能で、例えば「住宅・土地・建設」という条件で検索すると、図1のようにファイルで提供している調査が39、データベースで提供している調査が10という結果が返ってくる。この絞り込み機能は、キーワード検索の結果に対しても絞り込むことが可能なため、「土地」といったキーワードで検索したものを「住宅・土地・建設」という分野で絞り込むことにより精度が向上する。

目的のデータが含まれている統計調査がある程度絞れたら、その調査の詳細を各府省のホームページなどで確認して目的の内容と合致しているかを確認した上で、ファイルまたはデータベースからデータを取得するという流れとなる。なお、該当の統計調査がデータベースで公開されていたら、まずはデータベースからデータを探し、目的のデータが見つからなかった場合にファイルを探すといった流れをお勧めする。ファイルの方が提供しているデータが多くデータベースで提供していないデータもあるが、ファイルを開く手間を考えるとデータベースを検索の方が効率的だからである。

### 3.2. 都道府県・市区町村に関するデータ

都道府県や市区町村ごとの各種統計データを取得するのに便利な機能として「都道府県・市区町

村のすがた」（トップページの「地域から探す）」がある。この機能は、総務省統計局が編集している都道府県、市区町村単位に様々な統計結果をまとめた「社会・人口統計体系」を基にしたもので、都道府県は1975年から約4,300項目、市区町村は1980年から約760項目の統計結果を提供している。それぞれの項目は「A人口」、「H居住」などA～Hの13の分野に分けられ、さらにそれぞれの項目に「A1101人口総数」、「H1100総住宅数」というコードが振られており、「H居住」の分野では都道府県で約450、市区町村で約200の項目の統計データを取得できる。

「社会・人口統計体系」は年1回更新されるため、最新のデータは約1年前となるが、各種統計調査の主要な結果が都道府県及び市区町村でまとめてあり、さらに時系列でデータが取得できるため、各自治体の分析をするには非常に便利である。最新の結果や他の項目が必要な場合は、該当の統計調査の結果（ファイル又はデータベース）から取得すれば補完することができるため、必要とする統計データが見つからない場合は、まずはこのデータから探して見ることをお勧めする。

### 3.3. GISの利用

e-Statには、ファイルやデータベースから統計データを取得するだけでなく、地図と統計データ

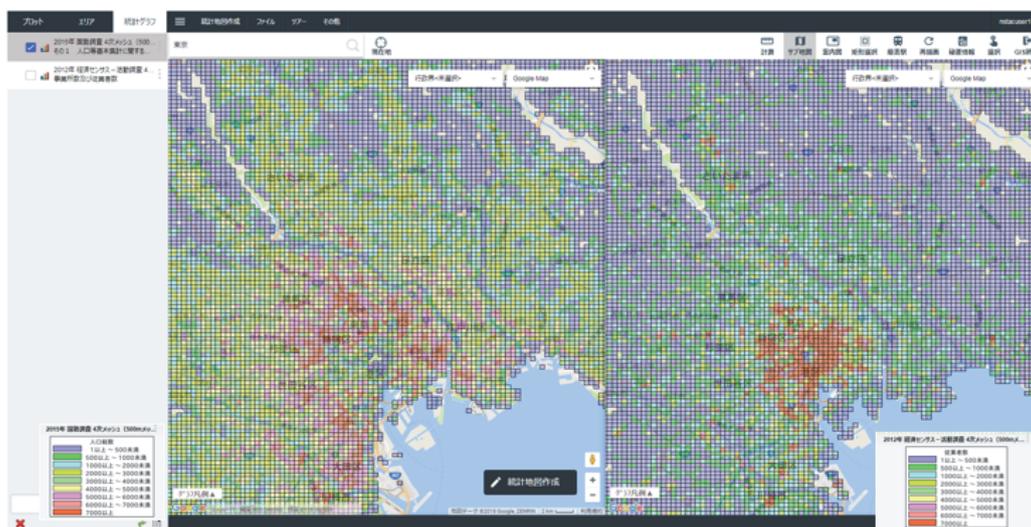


図2 jSTAT MAP画面（地域メッシュの事例）

を重ね合わせて分析できる GIS 機能 (jSTAT MAP) を提供している。重ね合わせることができる地域の単位としては、都道府県、市区町村、小地域 (町丁字等)、地域メッシュがある。特に地域メッシュは一辺の長さが 1Km、500m、250m の矩形を緯度経度で特定し、コード化したもので、時間でその範囲が変化しないことから時系列での比較に有用である。図 2 は 2015 年国勢調査の人口と 2012 年経済センサスの従業者数のデータをメッシュで表示したものである。国勢調査は住んでいる場所を表し、経済センサスは働いている場所を表すことになるため、東京の郊外から中心に人が移動していることが視覚的にわかる事例であるが時点が、メッシュを利用することで違う時点のデータを同じエリアで比較することができることを表している。

この地域メッシュは「JIS X0410」で標準化され、緯度経度から地域メッシュコード (逆も可能) が簡単に算出できることから、統計データだけでなく、気象情報 (降雨レーダーなど) や、標高情報、衛星データなどでも利用されているため、統計データと他のデータを重ね合わせて分析することも容易にできるという特徴ある。このように日本のメッシュは地理的なデータ分析の基盤として非常に有用であることから、日本のメッシュを世界でも利用できるように 2 桁のコードを追加した世界メッシュコードを作成するという取り組みも行われており、その仕様が Web サイト<sup>2</sup>で公開されている。

また、jSTAT MAP の大きな特徴として、地図上に国勢調査などの統計データを表示するだけでなく、利用者の保有するデータや他のオープンデータなどをシステム上に登録できるところである。例えば、自治体が公表している避難所の住所や収容人数などのデータを jSTAT MAP 上に登録し、避難所から半径 500m 以内に何人住んでいるのかを簡単に表示することが可能である。避難所などのポイントデータで緯度経度情報を持っていない場合は、処理件数に制限があるが、住所から緯度経度

に変換するジオコーディング機能も提供している。また、登録するデータは避難所等のような地点のデータだけでなく、エリアデータ (ポリゴン) も可能となっており、利用者独自のエリアで分析することも可能となっている。

さらに、jSTAT MAP では、任意の地点の国勢調査、経済センサスなどの基礎的な統計データを EXCEL ファイルに出力するリッチレポート機能というものを用意しており、GIS に慣れていない人でも、簡単に統計情報を得ることが可能となっている。

なお、jSTAT MAP は、利用者登録しなくても利用できるが、データの登録など全ての機能を利用する場合には、e-Stat に利用者登録 (メールアドレスのみで可能) し、e-Stat にログインして利用することをお勧めする。

#### 4. APIによるデータ取得

e-Stat では、プログラムや他のシステムなどから呼び出して統計データを取得できる API (Application Programming Interface) サービスを 2014 年から提供している。この API サービスは、インターネットを通じてアクセスする REST 方式の API となっており、「<http://api.e-stat.go.jp/rest/2.1/app/getStatsData?>…」といった Web ページのアドレスで利用されている URL (URI) でリクエストすると指定された統計データが XML、JSON、CSV のいずれかのフォーマットで取得できる機能である。

API で利用可能な統計データは、データベースで提供されている統計調査となっており、2018 年 10 月現在で約 120 調査、12 万データセットが利用可能である。

API では 7 つの機能を提供しているが、主要な機能は次の 3 つである。

##### ・「統計表情報取得」

API 機能で取得できる統計表 (データセット) の一覧を取得する機能。統計調査名、府省、表題等の情報とともに他の機能で必要となる統計表 ID (統計表を特定するコード) の情報が

<sup>2</sup> 世界メッシュ研究所 : <https://www.fttsus.jp/worldgrids/ja/top-ja/>

取得できる。検索パラメータとして、政府統計コード、調査年月、公表年月、検索キーワードなどがある。

- ・「メタ情報取得」

「統計表情報取得」で得られた統計表 ID に含まれるメタ情報(時間軸、地域、分類項目等)が取得できる機能。それぞれのデータセットにどのようなデータがあるのか詳細に調べたい場合に利用する。

- ・「統計データ取得」

「統計表情報取得」で得られた統計表 ID に対応したデータを取得する機能。時間軸、地域、分類事項等を指定することにより取得するデータを絞り込むことが可能。なお、一度に取得できるデータは 10 万に制限されているため、10 万以上のデータを取得する場合は、複数回にわけてリクエストする必要がある。

e-Stat からデータを取得する場合と比較すると、どのようなデータセットがあるのかを探し(「統計表情報取得」)、そのデータセットにどのようなデータが含まれているのかを確認し(「メタ情報取得」)、必要なデータを表示又はダウンロードする(「統計データ取得」となる。なお、REST 方式の API のため、それぞれの機能は単独で使用でき、同じパラメータ (URI) でのリクエストは必ず同じデータが返ってくるという仕組みである。

API で取得できる統計データは、一つのデータ(セル)に対し、メタ情報を付与している。つまり、1セル1レコードというイメージとなる。EXCEL や CSV では 1レコードに複数のデータ(セル)があるのが通常であるが、API では 1セル1レコードとすることにより組み換えが自由にできるという特徴と存在しないデータは作成されないという特徴がある。

API の利用ケースとしては、オンラインで統計データを取得して Web ページなどで情報提供するケースと、統計データを一括で取得して利用者のシステムに取り込んで利用するケースの 2つがある。

オンラインで統計データを取得するケースは、

API を利用してオンラインでデータを取得するため、常に最新のデータを表示することが可能である。ただし、取得するデータ量が多い場合は、取得時間が長くなるため、その都度 Web ページを表示するようなサービスの場合は、表示時間に注意する必要がある。

一括して取得するケースは、EXCEL 等のデータをダウンロードして必要なデータを抽出する手間がかからないため、効率的に利用者のシステムなどに取り込むことが可能である。オンラインでデータを取得していないため、定期的にデータを更新する必要があるが、統計調査の結果の更新頻度は短くても 1ヶ月単位なので、それほど頻繁に更新する必要はないと思われる。API で取得するデータには更新日時の情報があるので、更新処理の際には活用されたい。

なお、API の利用するためには、e-Stat に利用登録をし、アプリケーション ID を発行する必要がある。アプリケーション ID は各 API をリクエストする際の appId パラメータに必ず指定する必要がある。API のさらに詳しい仕様については、専用のサイト<sup>3</sup>を用意しているのでそちらを参照願いたい。

## 5. 新たなデータ提供(統計LOD)

API サービスの提供により、自動的なデータ取得などが可能となったが、データセットの名前や分類項目などのメタ情報が統一又は明確化されていないため、目的の統計データを見つけるためには、人の目で確認する必要がある。また、API 機能で提供する XML や CSV は、WWW(World Wide Web)の考案者であるティム・バーナーズ＝リーが提唱するオープンデータの 5つ星レベル(図 3)では、3つ星になる。

このため、新たな統計データの提供方法として、最上位の 5つ星である LOD(Linked Open Data)での提供を 2016 年から開始した。統計データを LOD で提供することにより、メタ情報が明確化される

<sup>3</sup> API 専用サイト : <https://www.e-stat.go.jp/api/>

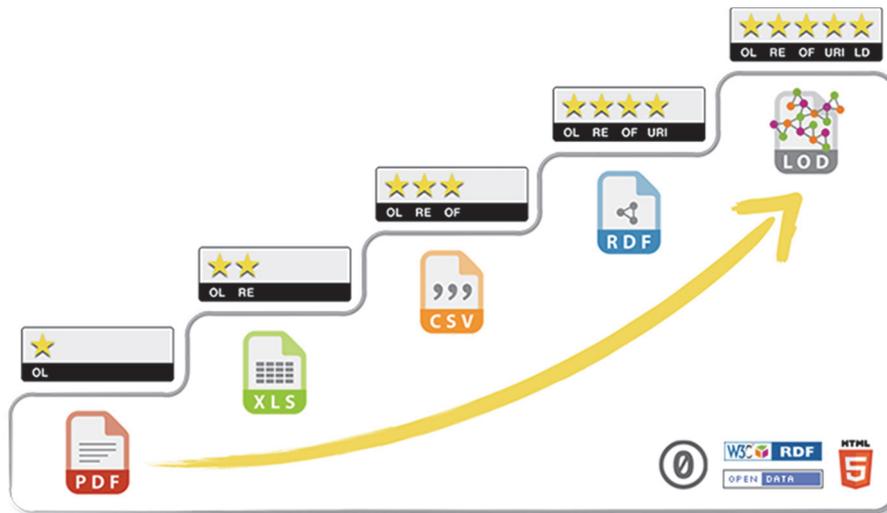


図 3 5スターオープンデータ<sup>4</sup>

ため、目的のデータの検索が用意になるだけでなく、国内外の他のデータとの連携も可能となることが期待される。

統計データを LOD で提供するためには、RDF (Resource Description Framework)<sup>5</sup>で各種データを表現する必要がある。RDF は Web の標準化を行なっている W3C(World Wide Web Consortium)で規格化されている Web 上のデータを表すための統一化された枠組みである。RDF は、各データに URI を付与し、それぞれのデータを「主語」、「述語」、「目的語」の3つ(トリプル)で表現する。今までの Web はページやファイルに付与されている URI に対してのリンクだったが、RDF にすることによりページへのリンクからデータへのリンクに変わっていくこととなる。LOD はリンクするオープンデータという言葉から Web 上で各データがリンクされ Web 上で各データが見つけやすくなることが期待されているものである。

統計データを RDF 化する方法として、RDF データキューブ(RDF Data Cube Vocabulary)<sup>6</sup>を採用している。RDF データキューブは RDF と同じく W3C

で規格化されている。RDF データキューブは、主にヨーロッパの統計機関で広く利用されている統計データとメタデータの交換規約である SDMX(Statistical Data and Metadata eXchange)<sup>7</sup>がベースとなっていることから既存データとの互換性もある。

RDF データキューブでは、統計データを表すために、次元(Dimension)、測度(Measure)、観測値(Observation)、属性(Attribute)の4つの要素で構成されている。次元は、統計データを何で分類したかを表すもので、地域、時間、性別、年齢などが該当する。測度は、統計データを何で集計等したものか表したもので、人、価格、指数などが該当する。観測値は、統計データの数値そのものを表すものである。属性は、統計データに付随する属性を表すもので、単位、状態(暫定値かなど)などを表すものである。実際の統計表で説明すると以下ようになる。

それぞれの数値(観測値)については、API と同様に一つの数値ごと(セル単位)に定義する方法を採用している。セル単位に情報を記述できるため、細かな記述が必要となる複雑な構造の統計表にも対応できるからである。実際の統計表を RDF にした時のイメージは以下ようになる。

<sup>4</sup> 5-star Open Data : <http://5stardata.info/ja/>

<sup>5</sup> Resource Description Framework (RDF) Model and Syntax Specification : <https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>

<sup>6</sup> The RDF Data Cube Vocabulary : <https://www.w3.org/TR/vocab-data-cube/>

<sup>7</sup> The official site for the SDMX community : <https://sdmx.org/>

例：2010年国勢調査の人口

次元	総数（男女別）		男		女
	44歳 【人】	45歳 【人】	44歳 【人】	45歳 【人】	...
...	...	...	...	...	...
さいたま市	16,130	19,245	8,293	9,938	...
川口市	6,582	8,022	3,526	4,289	...
...	...	...	...	...	...

測度: 人口  
属性: 年齢, 性別  
観測値: 人口数

図 4 RDFデータキューブモデル

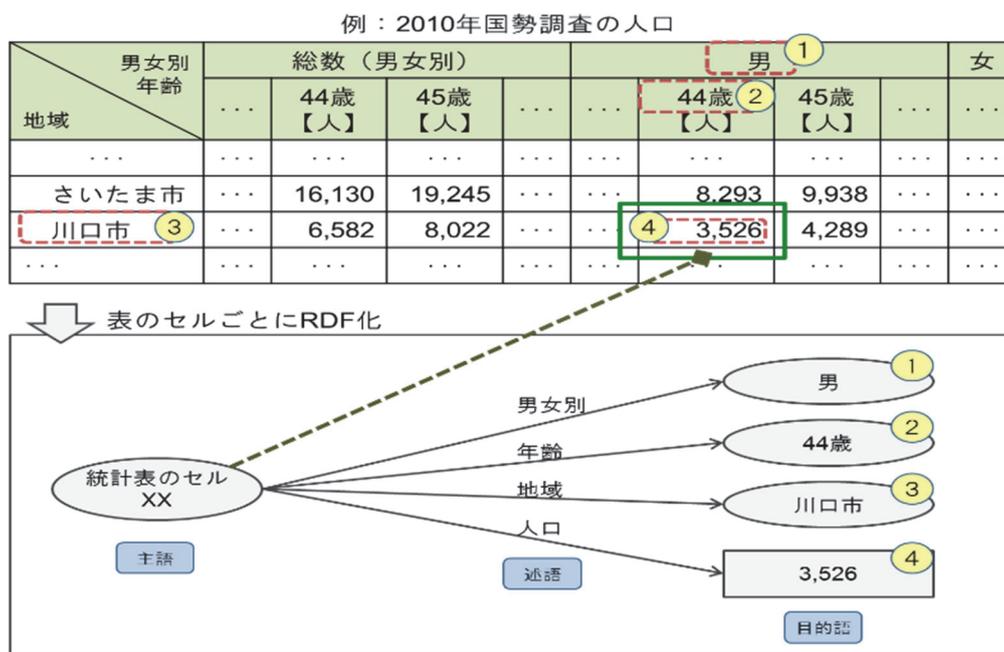


図 5 統計表のRDF化のイメージ

これにより各セルに URI が付与されることとなり、特定の数値に対してリンクを貼ることも可能となる。

LOD のメリットを活かすには、それぞれのデータの意味を表す情報（語彙）の定義が重要となる。統計データの場合、次元、測度、属性の定義となるが、統計調査で横断的に利用される項目などは、共通的な語彙として定義し、統計調査ごとに利用が限定される又は定義が異なる項目などは統計調査ごとに語彙を定義するという考え方で整備している。共通的な語彙としては、性別、年齢、日本標準産業分類、都道府県・市区町村などが挙げられる。特に地域に関する情報（次元）である都道

府県・市区町村については、「統計で用いる標準地域コード」として標準化されており、統計以外にも利用されていることから、優先して LOD として公開した。「統計で用いる標準地域コード」は都道府県を 2 桁、市区町村を 3 桁で表し、「JISX0401」及び「JISX0402」でそれぞれ規格化されている。同様なコードとして「全国自治体コード」があるが、こちらは 6 桁で構成されており、先頭の 5 桁は「統計で用いる標準地域コード」と同じで、6 桁目はチェックディジットとなっている。また、「統計で用いる標準地域コード」は 1970 年からの廃置分合の情報（新設、廃止、合併など）も蓄積されていることから、これらの情報も含めて LOD

として整備し、時点ごとの情報も取得できるようにしているため、統計以外のデータで当該データを使用する（リンク）することで LOD のメリットである統計データと他のデータが繋がるといったことが期待される。

統計 LOD を検索するには、W3C で規格化されている SPARQL (SPARQL Protocol and RDF Query Language) を使用する。SPARQL はデータベースを検索する際に用いられる SQL に似た記述で RDF データの検索が可能となっている。SPARQL でデータを取得するには SPARQL エンドポイントに対し SPARQL で記述されたクエリをリクエストする。e-Stat ではこのエンドポイントを提供しており、Web 画面上から SPARQL を記述してリクエストする方法と API としてリクエストする方法の二つがある。SPARQL の記述方法などの説明は本稿では省略するが、SPARQL を利用した API のメリットは必要なデータの抽出条件や並び替えなどが自由にできるため、通常の API に比べて利用しやすい。通常の API は検索条件や並び替えなどの機能をサーバー側で予め用意しておく必要があるためである。

LOD はオープンデータの最高位として位置付けられ、標準化された技術を使用していることから、欧州を中心にその活用について検討されており、すでに統計 LOD を公開しているイタリア、スコットランドを初めとし、ユーロスタットでは 2020 年までに統計 LOD を公開するプロジェクトが進められている。これにより統計データが日本だけでなく世界中のデータとリンクで繋がることが期待される。

## 6. 提供地域

e-Stat で提供されている統計データの地域は、そのほとんどが全国、地方、都道府県単位といった広域な地域での提供となっており、一部の統計調査で市区町村単位まで、町丁字やメッシュといった小地域なると限られた統計調査のみとなっている。土地や建物に関する統計調査では建築着工統計調査などが市区町村単位での提供となっている。広域な地域で集計しているほとんど理由は、

該当の統計調査が標本調査であることから小さな地域では誤差が大きくなることと、小さな地域で結果を表すと対象者や対象企業が特定されてしまうことがあるためであり、統計の特性上致し方ないものである。

土地や不動産に関連する他のデータと比べると地域範囲が広すぎるかもしれないが、都道府県などの広域な地域の統計結果は全体的な傾向を見る指標とし、特定する地域での分析は国勢調査や経済センサスなど小地域で提供されているデータを利用することにより、土地関連の他データとの分析は可能であると考えられる。

小地域での分析では前述した GIS 機能 (jSTAT MAP) を活用することにより比較的簡単に始めることが可能である。また、すべてのデータには対応できてはいないが、小地域のデータは API 及び LOD でも提供しているため、独自のシステムに取り込んでの活用も可能である。

## 7. 現状の課題

e-Stat で統計データを取得する方法について、いくつかの方法を紹介したが、e-Stat でデータを取得するための大きな課題としては、次の二つがあると考えられる。一つは、EXCEL ファイルなどで公開されているデータが多く、データベース化されているデータが限られていることである。もう一つは、各統計データがどのように集計されたものであるかという情報（地域、時間軸、分類（次元）など）であるメタ情報が体系的に整備されていないことである。

一つ目の問題である EXCEL ファイルで公開されているデータが多いという問題であるが、データ加工しやすい EXCEL ファイルであればそれほど大きな問題ではないと考えられる。しかしながら、e-Stat で公開されている EXCEL ファイルには、項目のセルが結合されていたり、項目にブランクが入っていたり、項目が複数行に渡って記述されていたりするという項目表現の問題だけでなく、印刷を綺麗にするために、複数の表が横に繰り返す（行で表が繰り返す）又は縦に繰り返す（列で表

が繰り返す) といった神 EXCEL と呼ばれる EXCEL ファイルがある。このような形で作成されているとデータを抽出して分析するための前作業の時間が多く取られてしまう。

これを解消するためにデータベースに変換する作業を進めているところであるが、いわゆる神 EXCEL は機械的な作業だけでは変換できないため、変換作業に時間がかかっているのが現状である。根本的に解消するためには神 EXCEL を作成しないことであるが、すでに各府省で稼働している多数の統計集計システムをすぐに変更するのは困難であるため、この見直しにも時間がかかるのが実情である。2018 年 11 月現在でデータベース化されている住宅・土地関連の統計調査は 39 調査のうち 10 調査となっており、かつデータベース化されている統計調査でも最新の結果はデータベース化されていないといった状況となっていることから、データベース化への取組みを推進する必要があると考えられる。

データベース化しなければ API での提供だけでなく、二つ目の課題であるメタ情報の整備にも繋がっていないため、EXCEL ファイル等で作成されている統計データをデータベース化するという取り組みが最も優先すべき課題であると考えられる。

二つ目の課題であるメタ情報の整備であるが、残念ながら日本ではメタ情報の整備がほとんど進んでいないのが現状である。欧州などでは、EU の関係もあることからメタ情報の整備が進んでおり、メタデータレジストリの構築や SDMX でのデータ整備などが積極的に行われている。

欧州などでは、統計データを作成する際に既存のメタ情報を利用して作成する方法がある程度確立されているようであるが、日本では、標準産業分類など、一部のメタ情報しか整備されていないため、統計調査や担当府省ごとに作成している状態となっている。さらに、メタ情報に統一化されたコードが振られているわけではないため、e-Stat での検索の仕組みも単純なキーワード検索といった手法となっており、目的のデータが見

つからない要因にもなっている。

しかし、すでに大量にあるメタ情報をすべて統一し整備することは、多大な労力と時間がかかることが想定される。この問題を解決するためには、メタ情報整備の対象を主要な統計に絞り込み (欧州でもすべてのメタ情報が整備されていない)、既存のデータをなるべく再利用することが重要であると考えられる。現在、既存データを再利用してメタ情報を整備する方法として LOD の活用を検討しているところである。

メタ情報の整備はデータベース化以上に時間がかかることが想定されるが、統計データをさらに活用してもらうためには避けては通れない課題であると言える。

## 8. おわりに

政府が行っている統計調査の結果は、我が国の社会経済を表す重要な情報であり、EBPM などの基となるデータである。このため、e-Stat を通じて広く公開を行ってきたが、公開するデータ量が増えるに伴い、目的のデータが見つからない、データの取得に手間がかかるなどの問題が発生した。その問題の解決のためにデータベース化を行いつつ、プログラムから取得しやすい API サービスの提供を開始し、各データが持つ情報を明確化する統計 LOD の提供を開始するなど少しずつではあるが発展してきている。しかしながら、データベース化されているデータが限定されており、さらにそのデータを表すメタ情報が体型的に整備されていないなどの課題が残っている。

国民の共有財産である政府統計の結果が、さらに利用しやすくなるように課題を解決しつつ、今後も改善等を実施していく予定である。