Geographic Area Representations in Statistical Linked Open Data of Japan

Dan Yamamoto[1], Akira Ioku[1], Yoko Seki[1], Akie Mizutani[1] , Junichi Matsuda[1],
Hideaki Takeda[2] Ikki Ohmukai[2], Fumihiro Kato[2] ,Seiji Koide[2], and Shoki Nishimura[3]

[1] Hitachi, Ltd., Tokyo, Japan
{dan.yamamoto.vx, akira.ioku.dv, yoko.seki.df, akie.mizutani.kj, junichi.matsuda.ru}@hitachi.com
[2] National Institute of Informatics, Tokyo, Japan
{takeda, i2k, fumi, koide}@nii.ac.jp
[3] National Statistics Center, Tokyo, Japan
snishimura@nstac.go.jp

Abstract. The Japanese Statistics Center has been provided statistical LOD since 2016 and is still evolving them. This paper focuses on problems and solutions related to geographic area representations in statistical LOD. We describe ontologies for handling absorption and abolishment of municipalities and also confidentiality concerning small area as well as grid square statistics.

Keywords: Statistics, Linked Open Data, Area, Grid square, Concealment.

1 Introduction

The Japanese Statistics Bureau and National Statistics Center of Japan have published approximately 500 government statistics at the Portal Site of Official Statistics in Japan (e-Stat). The main seven statistics among them, including population censuses, economic censuses, and labor force surveys, have been provided as Linked Open Data (LOD) since 2016 [1]. The statistical LOD site was opened on June 30, 2016 and is still evolving. Dataset and classification criteria of LOD-based statistical data are now semantically clarified, which not only facilitates data retrieval, but also enables linkage with other domestic and overseas data. The published LOD consists of approximately 400 million triples and 40 million observations [2].

Throughout its development, we faced two main challenges related to geographic area representations in statistical LOD: (1) Properly expressing the temporal changes such as absorption and abolishment of municipal divisions and (2) constructing LOD-enabled small areas and grid squares with concealed statistical data.

The concept of municipalities is changing by absorption and abolishment. There have been existing works to handle these temporal changes to geographic data. For example, [3] proposed a logical model for changing taxonomic concepts. Furthermore, in Linking Geospatial Data Workshop (March 2014), the common problem of modelling and managing temporal changes to geographic datasets was discussed [4] in the context of UK Department for Communities and Local Government (DCLG). In this paper, we propose a solution to describe the changing municipality concept with the reason for the change, and also the method for linking statistical data and changing concept.

Statistical LOD publishes data for not only municipalities but also small areas and grid squares as geographic areas. We propose a way of constructing these types of LOD. Furthermore, when publishing small area or grid-square statistics, some data is often concealed because of privacy protection. We also propose a way of expressing LOD-enabled concealed statistical data.

Several statistical LODs have also been published by the national statistics institutes of Ireland [5] and Italy [6] as well as the Scottish government [7] and UK DCLG [8]. Each one of these institutions has its own solutions to geographic area representations in LOD. For example, the Scottish government uses ONS Boundary Change Ontology [9] to express boundary changes. Our results described in this paper can be an entry point to state concrete ontology matching between such existing ontologies and ours, toward semantic interoperability among international statistical LODs.

2 Geographic Areas Handled in Statistics

This section describes a brief summary of the types of geographic areas used in Japanese statistics, which we aim to represent as LOD. Each statistical data is related to the following three types of geographic areas:

(1) Administrative divisions:

The Japanese administrative divisions consist of two layers: prefectural divisions and municipal divisions. Most statistical data is organized per administrative division. The entire country is divided into 47 prefectural divisions, and each prefectural division is divided into municipal divisions. Approximately two thousand municipal divisions exist in Japan.

The Japanese Statistics Bureau has a standardized codelist, the Standard Area Codes for Statistical Use, which enables us to uniquely identify each division. Every code is represented as a five-digit number, where the first two digits identify a specific prefectural division and the remaining three digits represent a specific municipal division located in the prefectural one. For example, Shinjuku-ward, Tokyo is coded as “13104” comprising Tokyo, “13”, and Shinjuku-ward, “104.”

(2) Small areas:

Small areas are subdivided portions of a municipal division. Frequently-used statistical data such as the census are produced for small areas as well as administrative divisions. Several codelists, which are defined in each statistics and not standardized, exist to uniquely identify each small area.

(3) Grid squares:

Grid squares are portions of the country divided into a grid by longitudinal and latitudinal lines. Frequently-used statistics like the census are also prepared for grid squares. As the basis of our grid-square statistical data, we use the world grid square system [10], a compatible extension of the Japanese grid-square coding system (JIS X0410) to worldwide. Table 1 shows three levels of grid squares used for Japanese statistics.

Table 1. Grid squares used for Japanese statistics

Third level grid squares	Divide Japan into equal parts measuring 2/3 degree of latitude by 1 degree of longitude. Divide one of the parts into 8 equal parts in latitude and longitude directions. Furthermore, divide one of the parts into 10 equal parts in latitude and longitude directions. Approximately 1-km square, 386,877 grid squares exist for Japan.
Fourth level grid squares	Divide third level grid square in half in latitude and longitude directions.
Fifth level grid squares	Divide fourth level grid square in half in latitude and longitude directions.

Other than the world grid square system we adopted, there are several worldwide grid-square systems such as Open Location Code [11] or “plus+codes.” Note that both the world grid square system and Open Location Code are derived from latitude and longitude coordinates. Hence, we can easily map one from the other, which potentially enables us to relate our grid-square statistical data with other datasets based on Open Location Code.

In our statistical LOD, these geographic areas can be referred to as objects of dimensions in terms of RDF Data Cube Vocabulary [12]. Fig. 1 depicts an instance of observation that refers to a certain administrative division, sac:C11203-20010401, as a dimension (in Line 3).

```
1:<http://data.e-stat.go.jp/lod/dataset/g00200521/d0003041389/obsKIBUUD4YPZBE7Q7P
6PNBWI7YYPG3RXLQ>
2:   a qb:Observation ;
3:   sdmx-dimension:refArea sac:C11203-20010401 ;
4:   cd-dimension:age  cd-code:age-0 ;
5:   cd-dimension:nationality  cd-code:nationality-japan ;
6:   cd-dimension:sex  cd-code:sex-female ;
7:   cd-dimension:timePeriod  "2010" ;
8:   estat-attribute:unitMeasure estat-attribute-code:unitMeasure-Person;
9:   estat-attribute:unitMult  estat-attribute-code:unitMult-0 ;
10:  estat-measure:population "2059"^^xs:decimal .
```

Fig. 1. RDF expression of the number of 0-year-old girls in a municipality

3 Expressing the Absorption and Abolishment of Municipal Divisions

3.1 Brief Description of Absorption and Abolishment of Municipal Divisions

The implication of each geographic area, in regard to attributes including city classification and borders, changes with time by passing through events such as: absorption, abolishment, separation, establishment of new municipalities, division into several municipalities, name change, boundary change, and shift to a designated city. Fig. 2 shows the case of Kawaguchi-City (standard area code: 11203) as an example of absorption and abolishment of city and district municipalities. The city absorbed the neighboring Hatogaya-City (standard area code: 11226) on October 11, 2011. The implication of “Kawaguchi-City” changed at the time. Specifically, attributes, such as borders, changed.

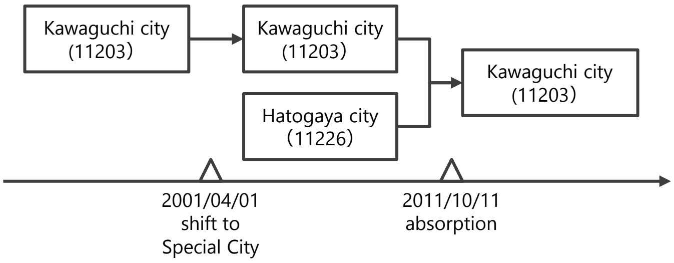


Fig. 2. Example of absorption and abolishment of municipal divisions

The areas referred to in statistical tables are those at a certain temporary point in the change or during a limited period, namely “areas as snapshots.” The population and the number of households of "Kawaguchi-City" are included in both the national censuses in 2010 and 2015. But “Kawaguchi-City” in 2010 and “Kawaguchi-City” in 2015 are not the same.

3.2 Standard Area Codes with the Notion of Time Period

Standard Area Codes for Statistical Use has included integration and abolition history (absorption, abolishment, new establishment, and so forth) since 1970 for the convenience of statistics users. We aim to include this historical information into our statistical LOD to increase its linkability to external LOD resources.

However, because the above standard area codes themselves do not include the concept of time, they are insufficient to express areas as snapshots. In the example above, using only the code 11203, we cannot distinguish between “Kawaguchi-City” in 2010 and that in 2015 after absorbing “Hatogaya-City.” Such an absorption has a big influence on a change of population and the number of households. Therefore, being unable to distinguish between data before and after absorption will pose a problem when we use statistical data.

At first, in view of the above, we defined a system of standard area codes with the notion of time period (hereafter called “temporary standard area codes”) to identify an area at a certain period of time by expanding the conventional area codes. The temporary standard area codes are provided by connecting conventional standard area codes and the date when events such as absorption were enforced with a hyphen (-). The temporary standard area code is valid until the next event. In the example above, “Kawaguchi-City” at the time of the national census in 2010 is expressed as 11203-20010401. The city at the time of the national census in 2015, after the admission merger on October 11, 2011, is expressed as 11203-20111011.

To achieve our LOD-enabled statistical data format, we adopted the policy in which we consider the temporary standard area codes as core resources and link them to the relevant information regarding areas (as snapshots) during the period. Each observation in our dataset can refer to the temporary standard area code as a value of dimension such as sdmx-dimension:refArea.

In addition, we made conventional standard area codes without period (hereafter called “plain standard area codes”) LOD-enabled so that the users can use them as pointers to the standard data area codes with period. It enables users to refer to each area by one stop without considering period. Plain standard area codes are useful for the users who only know the existing standard area codes to obtain information from our statistical LOD. Fig. 3 shows an example of the LOD-enabled temporary standard area codes and plain standard area codes, both of which are described with namespace prefix “sac.”

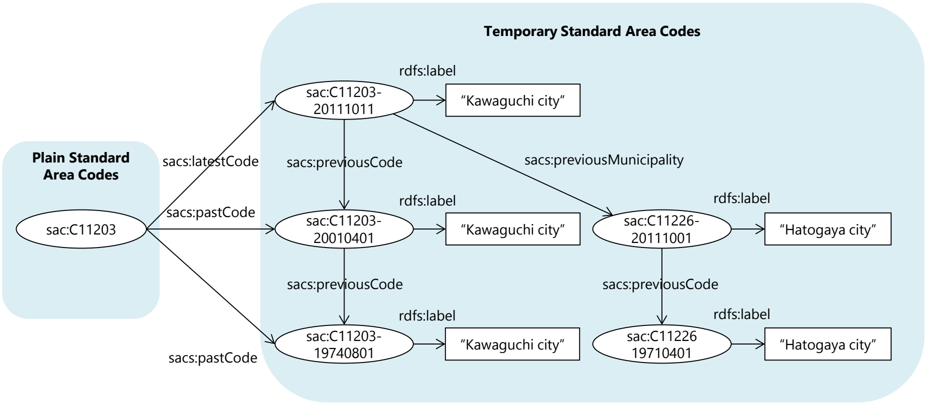


Fig. 3. Example of LOD-enabled temporary standard area code and plain standard area code

Because the causes of events such as absorption and abolishment, which change the implication of the area, are various and expected to be useful, we made them LOD-enabled as data expressing the reason of the change. As shown in Fig. 4, an event, described with namespace prefix “sace”, such as absorption and abolishment has a link to a change reason, described with namespace prefix “sacr”, and is linked from one or more temporary standard area codes related to the change event. Examples of links are as follows.

- Municipal organization enforcement from multiple towns to a single city: Link to each temporary standard area code of the abolished towns and the newly founded city.
- Change of the boundary line between a city and a town: Link to each temporary standard area code of the old and new city and town.

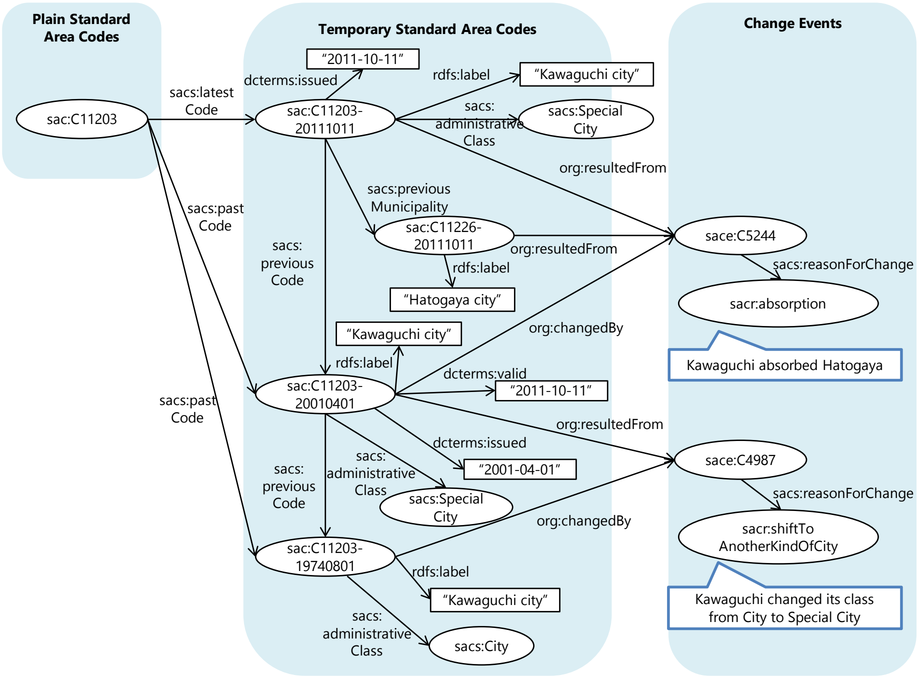


Fig. 4. Relations between temporary standard area codes and the change reason

Fig. 4 shows an example of the relations between the temporary standard area code and the change reason based on city classification and absorption change of "Kawaguchi-City." A temporary standard area code (e.g., sac:C11203-20010401), has links to: information such as name, type, date of issue and expiry date of the municipal division that the code identifies during the defined period, previous and subsequent temporary standard area codes (e.g., sac:C11203-19740801 and sac:C11203-20111011) before and after the defined period, and the change reason (e.g., sace:C4987) that led to the generation of the temporary standard area code. In addition, we also have a plain standard area code (e.g., sac:C11203) as a pointer to all the related temporary standard area codes via properties of sacs:latestCode and sacs:pastCode.

The average number of change events per municipal division is approximately 3.73 so that the data size of whole temporary standard area codes are about four times larger than the one of plain standard area codes, which is acceptable increment for our system.

Note that, we do not explicitly define the current or latest standard area codes in our LOD, instead we can identify the current or latest one according to sacs:latestCode property from a plain standard area code. We can also point any temporary standard area code at the point of survey based on issued and expired date described in each temporary standard area codes.

3.3 Comparison with an Existing Statistical LOD

We compare our ontology with an existing one, ONS Boundary Change Ontology [9], created by the Office for National Statistics (ONS) and used in ONS Geography Linked Data Portal as well as Scotland's statistics.gov.scot [7] to express boundary changes. In statistics.gov.scot, the following four properties defined by ONS Boundary Change Ontology are actually used:

- http://statistics.data.gov.uk/def/boundary-change/operativedate (Operative Date)
- http://statistics.data.gov.uk/def/boundary-change/terminateddate (Terminated Date)
- http://statistics.data.gov.uk/def/boundary-change/originatingChangeOrder (Originating Change Order)
- http://statistics.data.gov.uk/def/boundary-change/terminatingChangeOrder (Terminating Change Order)

The two date-related properties, i.e., Operative Date and Terminated Date, are corresponding to dcterms:issued and dcterms:valid, respectively. The main difference between them is the range of properties: the two properties in ONS ontology are object properties, which have values as resources, whereas the ones in our statistical LOD are dcterms properties that have values as literals. Scotland's statistics.gov.scot takes advantage of object property to link their operative date and terminated date to dereferencable URI, e.g., <http://reference.data.gov.uk/id/day/2006-07-21>, from which users can know further information related to the date from other linked datasets. By contrast, we adopted datatype properties to keep things simpler, while it is still possible for us to introduce additional object properties just like ONS ontology when similar rich linkable dataset of days/months/years is ready for our statistical LOD.

As for the other two properties related to change reasons, the situation is exactly opposite. The two properties used in statistic.gov.scot, i.e., Originating Change Order and Terminating Change Order, have literal values to express the order that caused the boundary change. In contrast with them, we use org:resultedFrom and org:changedBy to indicate change reasons as resources (e.g., sace:C5244) to represent various information about events causing boundary changes.

4 Expressing LOD-Enabled Small Areas and Grid Squares

4.1 Brief Description of Small Areas

Most public statistical information is shown in every municipal division defined as an administrative unit. However, based on the municipal division, we cannot grasp the various and detailed statuses of each geographic area, such as where exactly in the area population and stores are concentrated. To better understand the distribution patterns of population and households or statuses of private establishments and stores in specific areas such as elementary school districts or city centers, we have to prepare statistical information in a smaller unit.

Therefore, statistics such as population censuses have been established in smaller units than municipalities. In particular, we have two smaller units, namely small areas and grid squares mentioned in Section 2.

4.2 LOD-Enabled Grid Square Statistical Data

Small areas are subdivided portions of municipal divisions so that we can treat them equally with municipal divisions. However, the grid-square system is different from an administrative district and may be shared in the international system. Therefore, we defined additional attribute information such as latitude and longitude as LOD for grid squares. Fig. 5 shows an example of RDF expression of the grid-square used in our Statistical LOD.

```
grid:G2047306771 a gridCode:GridCode3 ;
    dcterms:identifier "2047306771" ;
    gridCode:lat-NW  34.925000 ;
    gridCode:long-NW  136.887500 ;
    gridCode:lat-SE  34.916667 ;
    gridCode:long-SE  136.900000 ;
    gridCode:span-EWN  1.142142 ;
    gridCode:span-EWS  1.142258 ;
    gridCode:span-NS  0.923454 ;
    gridCode:area  1.054769 .
```

Fig. 5. RDF expression of the grid square

Each grid square can be uniquely identified using the world grid square codes. For example, “2047306771” shown in Fig. 5 as a value of dcterms:identifier is a code of the world grid square code system. In addition, “grid:G2047306771” is an URI derived from the same code, which can be referred to from observed values in grid-square statistical dataset via dimension properties such as sdmx-dimension:refArea.

Expressed as RDF, both small areas and grid squares are now linkable to other existing geographic data such as Open Location Code [11] and GeoNames [13].

By describing grid-square statistical data and grid-square codes as LOD, the exact geographic location and semantics of statistical data are clarified. Furthermore, even non-statistical data can be easily linked using the grid-square codes defined as URI. The grid-square codes can also be used as unique keys to acquire statistical data belonging to specific geographic areas without regard to administrative divisions, which enables us to visualize statistical data as heatmaps.

4.3 LOD-Enabled Concealed Statistical Data

Target areas include small areas and grid squares with a very small population. If all the counted results regarding such areas are released, the private information of survey subjects might be open.

Therefore, concealment is performed in the count process if the private information of survey subjects can be open. For example, when there is an extremely small population or number of households in a certain small area, results should be added into the results of neighboring small areas and released. The basic idea of concealment is that items that should be concealed have to be clear in advance, and the results regarding areas smaller than a certain size should be concealed.

Our proposed LOD-enabled method will make the public statistical table LOD-enabled as is. In other words, this method follows the expression method of the statistical table of Table 2 and defines the adding-up destination as a specific area. An example of this method is shown in Fig. 6.

Table 2. Example of dataset with concealment

Code	City	Town	Block	Concealment	Concealed by	Conceals	Population
S103201189 001	Morioka	Sunakozawa	First	concealed	189005		X
S103201189 002	Morioka	Sunakozawa	Second				-
S103201189 003	Morioka	Sunakozawa	Third	concealed	189005		X
S103201189 004	Morioka	Sunakozawa	Fourth				-
S103201189 005	Morioka	Sunakozawa	Fifth	concealing		189001; 189003	35

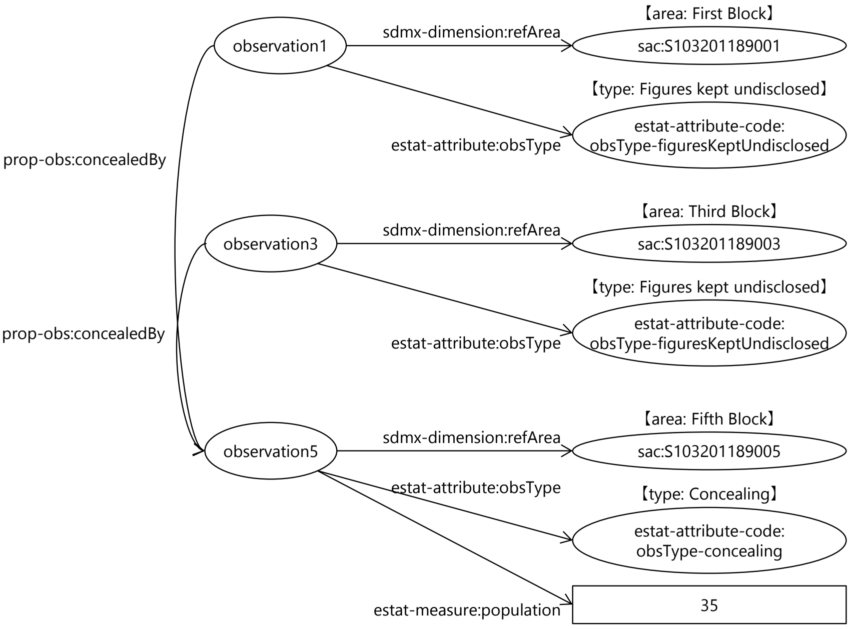
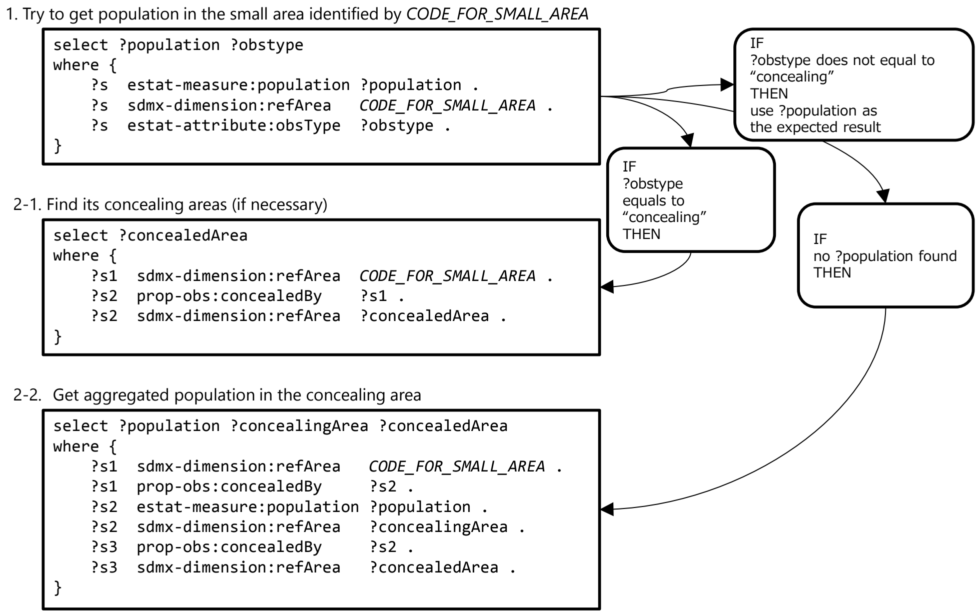


Fig. 6. LOD-enabled small area codes with concealment

Fig. 7 shows an example flow of data acquisition using SPARQL from LOD built using our method.



1. Try to get population in the small area identified by *CODE_FOR_SMALL_AREA*

```
select ?population ?obstype
where {
    ?s  estat-measure:population ?population .
    ?s  sdmx-dimension:refArea   CODE_FOR_SMALL_AREA .
    ?s  estat-attribute:obsType  ?obstype .
}
```

2-1. Find its concealing areas (if necessary)

```
select ?concealedArea
where {
    ?s1  sdmx-dimension:refArea  CODE_FOR_SMALL_AREA .
    ?s2  prop-obs:concealedBy    ?s1 .
    ?s2  sdmx-dimension:refArea  ?concealedArea .
}
```

2-2. Get aggregated population in the concealing area

```
select ?population ?concealingArea ?concealedArea
where {
    ?s1  sdmx-dimension:refArea   CODE_FOR_SMALL_AREA .
    ?s1  prop-obs:concealedBy     ?s2 .
    ?s2  estat-measure:population ?population .
    ?s2  sdmx-dimension:refArea   ?concealingArea .
    ?s3  prop-obs:concealedBy     ?s2 .
    ?s3  sdmx-dimension:refArea   ?concealedArea .
}
```

Fig. 7. Data acquisition from LOD-enabled small area codes with concealment

5 Conclusion

In this paper, we described the problems and solutions about an extension method of statistical LOD regarding geographic areas. The data is increasing sequentially, but performance will deteriorate along with it. Further speedup solutions will be necessary in the future. We also plan to introduce geometric information based on GeoSPARQL [14].

At present, general users face a big hurdle in using statistical LOD because they have to write code in SPARQL. To address this problem, we will not only expand the statistical LOD, but also provide usage samples and use cases.

When convenience improves with the improvement of speedup solutions and sufficient use cases, and LOD-enabled statistical data is aggregated, we will be able to create indexes from multiple databases and received insights not provided from a single dataset. Following the trend of European countries, the number of countries releasing government statistics as LOD are increasing. We will try to unify the data structure of LOD so that we can use federated queries more effectively.

References

1. Statistical LOD of Japan, http://data.e-stat.go.jp/ (accessed 2017-09-15).
2. Yu Asano, Yusuke Takeyoshi, Junichi Matsuda, and Shoki Nishimura: Publication of Statistical Linked Open Data in Japan. In: 4th International Workshop on Semantic Statistics, http://ceur-ws.org/Vol-1654/article-01.pdf (2016).
3. Chawuthai, R., Takeda, H., Wuwongse, V., & Jinbo, U.: A logical model for taxonomic concepts for expanding knowledge using Linked Open Data. Semantics for Biodiversity (S4BioDiv 2013), pp9-16.
4. Steve Peters and Bill Roberts: Modelling and managing temporal changes to geographic datasets. W3C Linking Geospatial Data Workshop, March 2014, available at https://www.w3.org/2014/03/lgd/papers/lgd14_submission_12
5. data.cso.ie: A Linked Data Service for the Census 2011 Results, http://data.cso.ie/ (accessed 2017-09-15).
6. Linked Open Data – PIATTAFORMA SPERIMENTALE PER I LOD DELL’ISTITUTO NAZIONALE DI STATISTICA, http://datiopen.istat.it/ (accessed 2017-09-15).
7. STATISTICS.GOV.SCOT – Open access to Scotland’s official statistics, http://statistics.gov.scot/ (accessed 2017-09-15).
8. OpenDataCommunities – Open access to local data, http://opendatacommunities.org/data (accessed 2017-09-15).
9. Office for National Statistics: ONS Boundary Change Ontology, http://statistics.data.gov.uk/def/boundary-change (accessed 2017-09-15).
10. Research Institute for World Grid Squares, http://www.fttsus.jp/worldgrids/en/top/ (accessed 2017-09-15).
11. Open Location Code, http://openlocationcode.com/ (accessed 2017-09-15).
12. Cyganiak, R., Reynolds, D. (eds.): The RDF Data Cube Vocabulary. W3C Recommendation 16 January 2014, http://www.w3.org/TR/vocab-data-cube/, World Wide Web Consortium, (accessed 2017-09-15).
13. GeoNames, http://www.geonames.org/ (accessed 2017-09-15)

14. Open Geospatial Consortium: OGC GeoSPARQL - A Geographic Query Language for RDF Data. version 1.0, 2012. available at http://www.opengis.net/doc/IS/geosparql/1.0

Appendix

A. List of prefixes and namespaces

prefix	namespace
qb	http://purl.org/linked-data/cube#
sdmx-dimension	http://purl.org/linked-data/sdmx/2009/dimension#
dcterms	http://purl.org/dc/terms/
geo	http://www.opengis.net/ont/geosparql#
sf	http://www.opengis.net/ont/sf#
org	http://www.w3.org/ns/org#
xs	http://www.w3.org/2001/XMLSchema#
sac	http://data.e-stat.go.jp/lod/sac/
sacs	http://data.e-stat.go.jp/lod/terms/sacs#
sace	http://data.e-stat.go.jp/lod/sace/
sacr	http://data.e-stat.go.jp/lod/sacr/
estat-attribute	http://data.e-stat.go.jp/lod/ontology/attribute/
estat-attribute-code	http://data.e-stat.go.jp/lod/ontology/attribute/code/
estat-measure	http://data.e-stat.go.jp/lod/ontology/measure/
cd-dimension	http://data.e-stat.go.jp/lod/ontology/crossDomain/dimension/
cd-code	http://data.e-stat.go.jp/lod/ontology/crossDomain/code/
grid	http://data.e-stat.go.jp/lod/gridCode/
gridCode	http://data.e-stat.go.jp/lod/terms/gridCode/
prop-obs	http://data.e-stat.go.jp/lod/ontology/property/observation/