Merging statistics and geospatial information, 2015 projects - Poland

Statistics Explained

This article forms part of Eurostat 's statistical report on *Merging statistics and geospatial information: 2019 edition* .

Final report 3 January 2018

Problem

Statistics Poland possesses a vast amount of statistical data that is housed across a range of disparate databases and disseminated using various methods. End users would benefit considerably if this data could be georeferenced and provided in the form of linked open data.

Objectives

The overall objective of this project was to support the decision-making processes involving the provision of standardised, usable and open geo-referenced statistical data.

Specific objectives:

1. identification of territorial units for which data can be published, including identification of their spatial representation across different years;
2. standardisation of territorial unit identifiers — creating a basis for linking statistical information with geospatial data;
3. feasibility analysis for publishing official statistical resources as linked open data;
4. definition of actions needed to transform existing data to open data formats;
5. description of official statistical resources with metadata in RDF (resource description framework) standards;
6. feasibility analysis for publishing linked open data in the national Geostatistics Portal, including development of guidelines for a linked open data web application.

Method

Stocktaking — creation of a data source catalogue

The project involved a stocktaking exercise for various datasets and databases for a range of official statistics analysing these in terms of their content, geo-references and their degree of 'openness'. The stocktaking encompassed all statistical materials published by official statistical entities in whatever form and resulted in a set of metadata, explicitly including information about the use of territorial divisions, data structures and open data.

Territorial divisions

This step involved the identification, harmonisation and generalisation of units for spatial division. The most commonly used divisions were based on the NTS, the Polish equivalent of the NUTS regional classification, and local administrative units (LAUs). Data from the NTS and the geometries for gminas from the National Register of Boundaries (PRG) were combined with each other and then with data from the TERYT database in order to produce geometries for regions, voivodeships, subregions and poviats. Geometries for the units composing each of these territorial divisions were imported into a geo-database with harmonised attributes for specific years. A second version of the dataset was developed after
generalising (simplifying) shapes in order to reduce file sizes and thus ensure smoother presentation via the internet.

NTS level	NUTS / LAU	Name	Identifier	Unit type	KTS identifier
NTS 1	NUTS 1	Macroregion	1.6	Statistical	**XXXX**0000000000
NTS 2	NUTS 2	Voivodeship (TERYT)	2.6.22	Statistical and administrative	**XXXXXX**00000000
NTS 3	NUTS 3	Subregion	3.6.22.40	Statistical	**XXXXXXXXX**00000
NTS 4	LAU 1	Powiat/cities with powiat rights (TERYT)	4.6.22.40.11	Statistical and administrative	**XXXXXXXXXXX**000
NTS 5	LAU 2	Gmina/part of gmina (TERYT) 1 – urban gmina, 2 – rural gmina, 3 – urban-rural gmina, 4 – urban part of urban-rural gmina, 5 – rural part of urban-rural gmina, 8 – districts of Warsaw, 9 – delegacies of cities: Wrocław, Poznań, Łódź, Kraków.	5.6.22.40.11.01.1	Statistical and administrative	**XXXXXXXXXXXXX**

Table 1: Territorial and statistical unit coding system

Concerning other units for territorial divisions, STRATEG (see below) was a tool designed to facilitate the monitoring, development and evaluation of measures taken under cohesion policy. It presented data for non-standard units of territorial divisions, currently providing data for the following functional areas:

- four supraregional strategies;
- two levels of development;
- 20 functional areas related to voivodship development strategies;
- 24 integrated territorial investment areas of functional urban areas (ZIT).

Geometries were prepared for all units for these different territorial divisions.

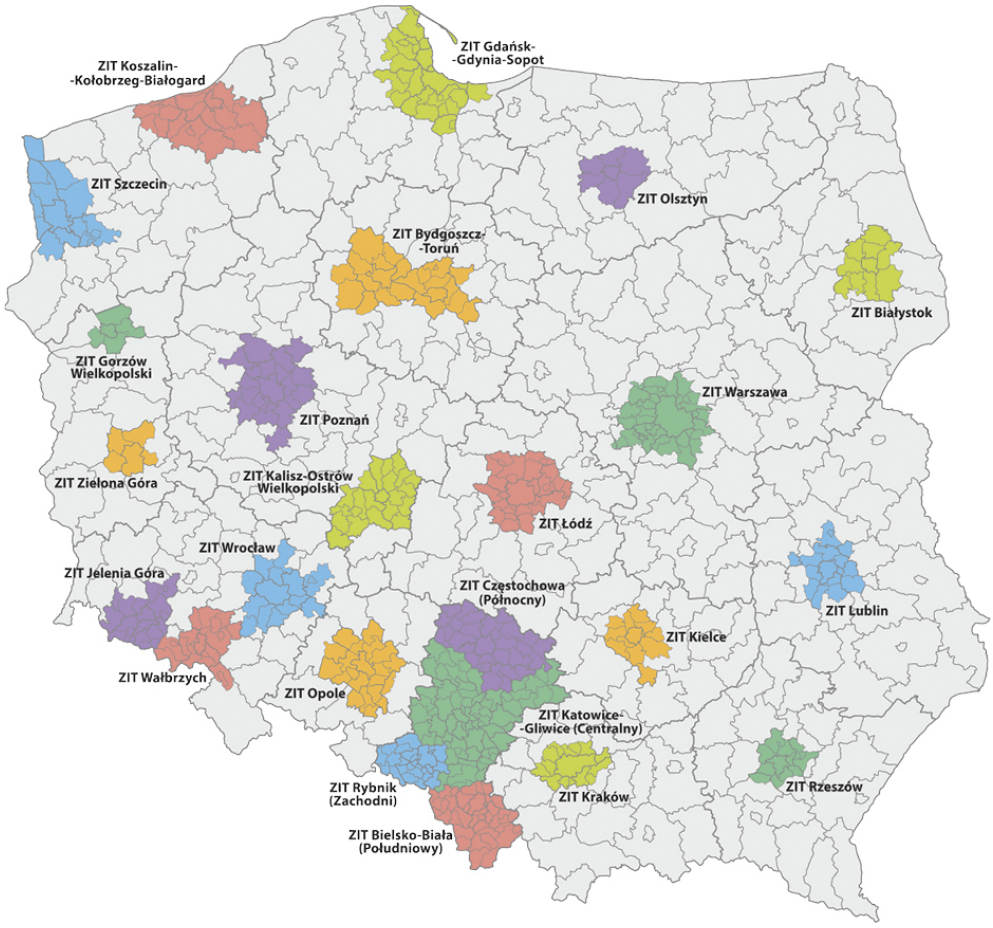

Figure 1: Integrated Territorial Investment Areas of Functional Urban Areas (ZIT)

Open data technology research

The following conversion process, from data to the development of an ontology to the provision of a service, was developed and researched:

- determine the scope of data publication and methods of searching;
- establish an ontology;
- map the ontology onto existing databases;
- export data to the resource description framework (RDF) format;
- load data to the RDF data store;
- publish data on a linked data server.

Appropriate open source tools and technologies for each of the specific steps of this conversion process were sought after designing the test ontology and then analysed in order to identify the optimum solution for the complete implementation. The following tools were tested:

- the Ontop platform — modelling, mapping and exporting data in the RDF format;
- Apache Jena Fuseki — a SPARQL server;
- Pubby — a linked data front-end (user-friendly interface) for SPARQL end-points;
- OpenCube Toolkit — a set of integrated open source components.

Two proposals for an open data web application for official statistics were developed. The first concerned the development of an open data portal for the statistics office using the SPARQL end-point technology (Apache Fuseki) and an extensive user interface. Browsing query results on a webpage interface could be provided by the

Pubby software. Data for an RDF data store (Apache Jena) would be prepared using one of the available solutions (Ontop or Python RDFLib) and then imported to the RDF data store. The second option concerned developing the national open data portal (danepubliczne.gov.pl). At the time of writing, the website did not have any solution for SPARQL end-points, but it did provide access to data via an application programming interface (API) . The Central Statistical Office had a webpage established with hyperlinks to various datasets and services.

The definitions for terms and conditions of use (information on licences) for published data were researched for both options, looking at issues such as the type and scope of licences (for example, public domain or open data licenses) and the compatibility of any licences was assessed when linking to external data sources (for example, the scope of any license, re-licensing conditions, or conditions on the publication of derived work). Equally, certification by the Open Data Institute was researched.

Pilot implementation

Three main public statistics databases were reviewed:

- local data bank — the biggest set of information on the socioeconomic situation, demography and the state of environment in Poland; it provides access to up-to-date statistical information and enables multidimensional regional and local statistical analysis;
- demography database — provides access to statistical information on demography; an integrated data source for the state and structure of the population, vital statistics and migration; enables multidimensional statistical analysis;
- development monitoring system, STRATEG — a system designed to facilitate programming and monitoring of development policy; contains a comprehensive set of key measures to monitor the execution of strategies at national, transregional and voivodship level, as well as in the European Union (Europe 2020 strategy); provides access to statistical data for cohesion policy; along with an extensive database, STRATEG offers tools facilitating statistical analysis based on graphs and maps.

From each of these databases a selection of data concerning population by sex and by age was extracted for Poland and its voivodships. Equally, the geometries of Poland and its voivodships were selected for transformation. The data source catalogue developed as part of this project was used as the source of metadata.

A URL structure was established made up of a local server address (for the pilot project) and separate folders for the statistical databases and each of the territorial divisions.

Prior to establishing and implementing an ontology, existing vocabularies were researched, without finding one that could be considered as a reference. As a result, an existing SDMX codelist (CL_SEX) was used for the sex dimension, whereas a new dimension was created for age groups, a codelist was created to specify the population and new dimensions were created for the territorial and statistical units. Namespaces were defined to provide the vocabularies created as part of this project.

After designing the ontologies they were transformed into RDF metadata. Initially, Ontop was used, but this gave unsatisfactory results and so a switch was made to use the Python RDFlib package instead.

Two datasets were created from spatial data, one for Poland as a whole and one for voivodships, both were based on 2016 territorial boundaries. The Python RDFlib package was again used, in this case to convert data from the shapefile format to a coordinate based reference system.

Finally, the three datasets were encoded using the codelists already prepared, with the following structure: `http://[server_name]/[database_name]/2016/POP_SEX_AGE/` . The Uniform Resource Identifier (URI) for each value within a dataset was a combination of the geographic dimension (a 14-digit KTS unit code), sex (based on the SDMX codelist) and age (based on a new codelist). The time dimension was set in the URI at the dataset level. A URI example for the total male population in 2016 for the area of Poland from the local data bank (BDL) is: `http://[server_name]/BDL/2016/POP_SEX_AGE/` 10000000000000/Sex-T/TOTAL. In other words, the URI (**in bold**) is followed by a 14-digit KTS code, then by the code for total (from the SDMX CL_SEX codelist) and finally by the code for all age groups (from the Local Data Bank age ontology).

An example of encoding for the total male population in 2016 for Poland from the local data bank (BDL) in Turtle (TTL) format is:

<http://10.51.20.122:8090/BDL/2016/POP_SEX_AGE/10000000000000/Sex-M/TOTAL> a **qb:Observation** ; **qb:dataSet** <http://10.51.20.122:8090/BDL/2016/POP_SEX_AGE> **sdmx-dimension:age** statpl-bdl-age:TOTAL ; **sdmx-dimension:refArea** <http://10.51.20.122:8090/KTS/2016/10000000000000> ; **sdmx-dimension:refPeriod** <http://reference.data.gov.uk/id/gregorian-year/2016> ; **sdmx-dimension:sex** sdmx-code:Sex-M ; **sdmx-measure:obsValue** "18593166"^^xsd:longint .

Open data catalogue and dataset metadata

The data source inventory (see above) was transformed into a linked open data catalogue using the DCAT application profile for data portals in Europe. The URI was http://[server_name]/KATALOG/.

Equally, each dataset from the data source inventory received a random universally unique identifier (UUID) generated using the uuid Python module. This UUID combined with the URI of the catalogue created the dataset URI.

RDF data store setup and usage

The final stage of the pilot exercise was the creation of a test SPARQL end-point and the creation of a front-end to view (in a web browser) the data store. Apache Jena Fuseki software was used as a SPARQL server. All RDF graphs created within the project were serialised and exported in RDF-XML and Turtle (TTL) formats. All data was loaded as a single Fuseki dataset into the RDF data store using the RDF-XML files.

Finally, a linked open data front-end was set-up using Pubby, which was used to create webpages for each local URI defined in the datasets uploaded to the Apache Jena Fuseki SPARQL server. Each URI created can be viewed with its associated properties as a webpage.

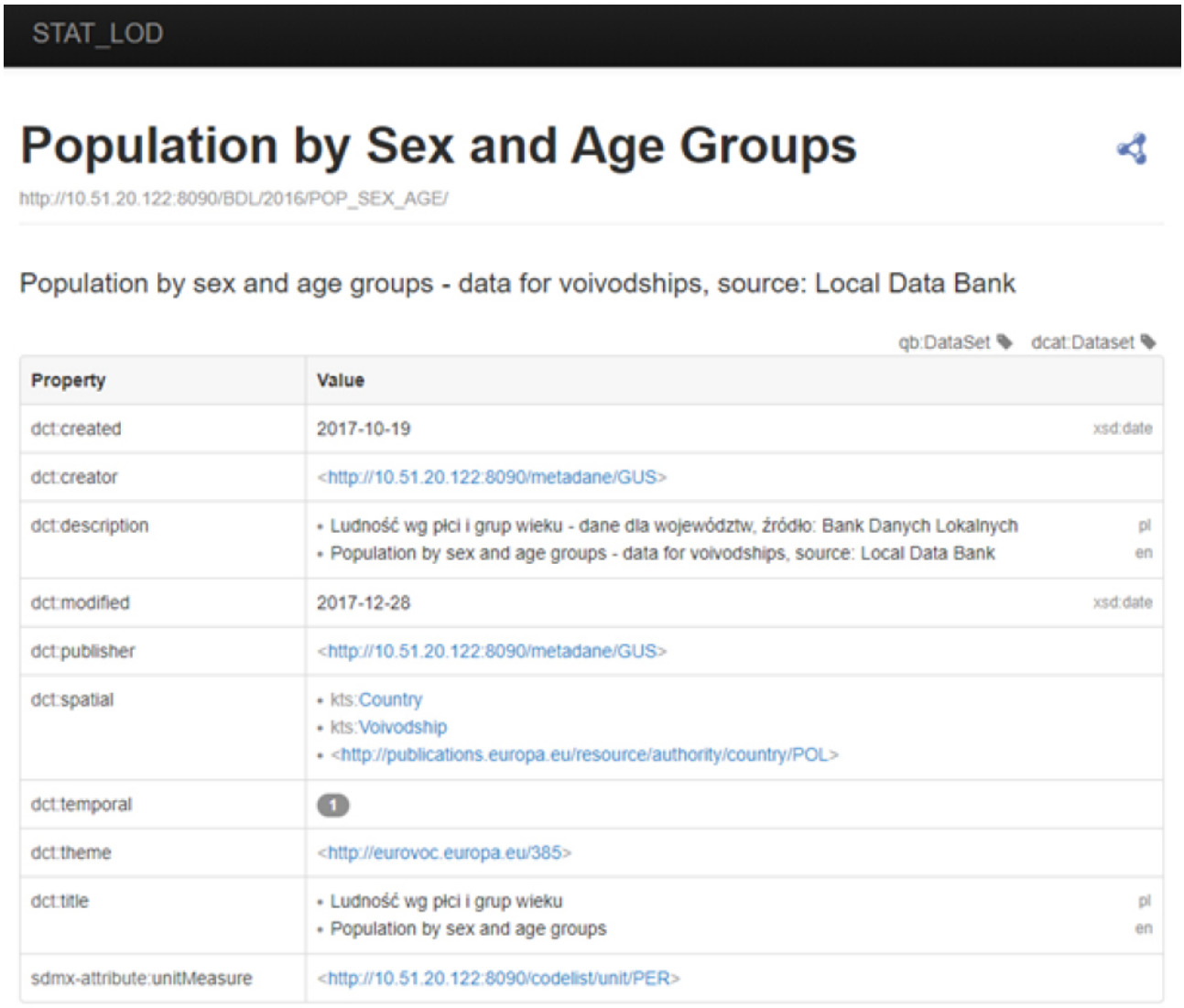

Figure 2: Example webpage for a dataset

Results

The pilot project provided valuable knowledge on linked open data technologies and vocabularies; several conclusions emerged.

There was no existing reference implementation for statistical linked open data that could be considered as fully appropriate. Existing implementations had some of the following issues:

- lack of integrity between RDF metadata sets published by one authority, probably due to different software or programming components used;
- links to non-existing entities (for example, old ontologies that were not online anymore);
- lack of maintenance (for example, containing data only for a specific reference year).

At the time of writing, there were no pan-European guidelines for statistical linked open data (for example, which vocabularies or software components to use), although there were several initiatives being run under Eurostat's DIGICOM project.

Some of the tools tested within this project (for example, Ontop or Pubby) became redundant during the course of the project (they were no longer developed), so any implementations based on these might become unstable over time. The Python RDFlib package was considered as sustainable, but it also ceased to be developed during the course of the project.

Linked open data makes most sense if it is connected with as many other data sources as possible. This project used several existing vocabularies and published datasets but a reference statistical linked open data implementation would be a more desired resource.

To achieve this, semantic harmonisation of statistical classifications would be needed. This was considered not only a pan-European issue as it may also impact national data providers, if various datasets have different uses for apparently identical classification elements.

In terms of technology, GeoSPARQL was considered to be an appropriate way to publish spatial data as linked open data. In terms of the temporal aspects, it was considered more complicated.

Separate statistical unit geometries were published for each year, regardless of changes over time. The URIs were constructed based on meaningful identifiers (KTS unit codes). A more appropriate situation may have been to analyse an inventory of statistical unit boundary changes over time and to provide separate geometry instances with non-meaningful identifiers (UUIDs): this would provide a single geometry for a defined period of validity for a unit whose boundary did not change. A prerequisite to do so is information on boundary changes (rather than on the boundaries themselves).

By using existing software and/or programming components it was nearly impossible to produce incorrect RDF metadata files. However, most linked open data producing components allow almost anything to be encoded, so the implementations may not always make sense semantically.

Linked open data implementations based on Python scripts were considered easy to amend, providing flexibility for the future.

RDF vocabulary specifications were considered easier to interpret with a unified modelling language (UML) model. The DCAT-AP specification provided a full UML model of all used classes and properties with a clear indication which of these classes and properties were mandatory, recommended or optional. The RDF Data Cube Vocabulary specification also had a simple graphical representation of some of its classes and their relations.

The pilot project identified that the policy for creating links to the statistical office's information portal needed to be redesigned to allow a linked open data implementation. The landing page and download URLs were based on links to specific webpages (ending in *.html) or links to specific files. To enable a correct linked open data implementation, it was recommended that they should be constructed as URIs.

Methodology

- Final report
- Annex I Python scripts
- Annex II LOD scripts
- Annex III Python scripts source data
- Annex IV LOD graphs RDF XML
- Annex V LOD graphs TTL

External links

- Statistics Poland
- Open data portal/mapping portal
- TERYT register

View this article online at https://ec.europa.eu/eurostat/statistics-explained/index.php/Merging_statistics_and_geospatial_information,_2015_projects_-_Poland