# Investigation of linked open data technologies for purposes of publishing georeferenced statistical data

Mirosław Migacz

Central Statistical Office of Poland, m.migacz@stat.gov.pl

**Abstract:** Polish official statistics possesses a vast amount of statistical data dispersed among different databases and disseminated using various publication methods. While there is a significant increase in openness of the data, there is still a lot of work to be done in terms of integrating different data sources. That is why the Central Statistical Office of Poland decided to look into the linked open data technology. An ongoing project has resulted in an inventory of databases and data sources currently published by official statistics and is investigating linked data technologies in order to prepare a "cook-book" for a linked open data implementation.

## 1 The "Development of guidelines for publishing statistical data as linked open data" project

In January 2016 Central Statistical Office of Poland (CSO) launched the "Development of guidelines for publishing statistical data as linked open data" project under Eurostat's "Merging statistics and geospatial information in Member States" grant series. The overall objective of the project is to support decision-making processes involving provision of standardized, usable and open georeferenced statistical data. Polish official statistics possesses a vast amount of statistical data dispersed among different databases and disseminated using various publication methods. Providing access to statistical data as linked open data would tremendously increase value of the data for end users.

The aim of the project can be achieved through identification of datasets and databases existing in official statistics and their thorough analysis in terms of content, georeference and the degree of "openness". The next step will be the development of a strategy for publication of official statistics' resources as linked open data with respect to the following specific objectives:

- identifying units of territorial division of the country, for which data can be published, including identification of their spatial representations for respective years,
- standarization of territorial division units identifiers – creating basis for linking statistical information with geospatial data,

- feasibility analysis for publishing official statistics' resources as linked open data,
- defining actions that need to be taken in order to transform existing data to open formats,
- description of official statistics' resources with metadata in RDF (Resource Description Framework) standards.

## 2 Identification of statistical data sources

The first stage of the project was identification of data sources possessed by polish official statistics and selection of those, which could be published in open formats. The following data sources were primarily taken into consideration:

- *Local Data Bank* – the biggest set of information on socio-economic 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* – provices access to statistical information on the demographic situation. It is an integrated data source for the state and structure of population, vital statistics and migrations. The database allows combining data on population dynamics and enables multidimensional statistical analysis,
- *Development monitoring system STRATEG* – a system designed to facilitate programming and monitoring of the development policy. The database contains a comprehensive set of key measures to monitor execution of strategies at the national, transregional and voivodship level as well as in the European Union (Europe 2020 strategy). Additionally, the system provides access to statistical data significant for cohesion policy. Along with an extensive database, STRATEG offers tools enabling statistical analysis based on graphs and maps.

Data sources indicated above and other sources identified in the first phase of the project were analysed for their openness. All sources will be confronted with the principles of "5 star open data" by Tim Berners-Lee:

- make your data available on the Web under an open license (★),
- make it available as structured data (to allow processing in software) (★★),
- use non-proprietary formats (★★★),
- use URIs to point to data and provide metadata according to the RDF standard (★★★★),

- link your data to other data to provide context (★★★★★).

The three major databases (Local Data Bank, Demography Database, Development Monitoring System STRATEG) were chosen for a pilot linked open data implementation, which will be described further on.

Aside from those databases, all statistical data sources published on CSO information portal have been catalogued and assessed in terms of openness. These include: publications, communiques and announcements. All data sources have been described with metadata including:

- thematic domain (e.g. society),
- thematic category (e.g. national accounts),
- type (e.g. a communique),
- name (of the resource),
- location of the source (link),
- format (PDF, DOC, XLS, CSV),
- geographic scope of the data (Poland / Europe),
- territorial detail level of the data (Poland, macroregions, voivodships, regions, subregions, powiats, gminas, districts),
- presence of territorial division units' identifiers (NUTS, TERYT – territorial unit register maintained by official statistics),
- years for which data is available,
- frequency of the data (year, half year, quarter year),
- update cycle (i.e. monthly, quarterly, semi-annually, annually),
- linked open data category.

The main conclusion from this stage of the project is that most of the data published in CSO's information portal is in some way structured (most of the datasets are published in XLS). However the table layout of XLS files is similar to the one in official publications and does not facilitate easy reuse (batch copying, processing etc.).

# 3 Harmonization and generalization of territorial units used for publishing statistical data

Another stage of the project comprised tasks on identification, harmonization and generalization of territorial division units used for publishing statistical data. Initially the core set of those units was the Nomenclature of Territorial Units for Statistical

Purposes (NTS). NTS comprised three NUTS levels and two LAU levels into one 5-level classification with its own set of identifiers.

CSO had in its possession non-harmonized datasets with geometries of the gmina (LAU2) units from the National Register of Boundaries (PRG) for years 2002-2006, 2008, 2009, 2015, 2016 and harmonized geometries of all territorial division units (region, voivodship, subregion, powiat, gmina) for years 2010-2014. For all years mentioned above, NTS unit lists with identifiers were present. The table below presents the NTS unit breakdown with examples of identifiers:

| **NTS level** | **NUTS / LAU** | **name** | **type of unit** | **identifier** |
|---|---|---|---|---|
| NTS 1 | NUTS 1 | region | statistical | 1.6 |
| NTS 2 | NUTS 2 | voivodship | statistical and administrative | 2.6.22 |
| NTS 3 | NUTS 3 | subregion | statistical | 3.6.22.40 |
| NTS 4 | LAU 1 | powiat | administrative | 4.6.22.40.11 |
| NTS 5 | LAU 2 | gmina | administrative | 5.6.22.40.11.01.1 |

Table 1. Nomenclature of Territorial Units for Statistical Purposes (NTS)

For years 2002-2016 geometries of gmina units (NTS level 5) were harmonized with the NTS identifiers and geometries of higher level units were built from gmina units for each year.

The NTS classification was thought to be a convenient way to unite the administrative division and the statistical division in a single classification. However, a new NUTS 2016 division is to be introduced in 2018. One of the major changes is the division of mazowieckie voivodship, which is now a single NUTS2 unit, into two NUTS2 units: the capital city of Warsaw with surrounding powiat units and the rest of the mazowieckie voivodship. This change will not be reflected in the administrative division of the country, therefore an incoherence between the statistical and administrative division will emerge at NTS level 2.

During the course of the project it has been decided that the NTS classification will cease to exist and the regulation establishing NTS was repelled. A new coding system for statistical units has been introduced to unite the administrative and statistical division

in a single classification: KTS (coding system for territorial and statistical units). Each unit, regardless of its level is described with a 14-digit code built as follows:

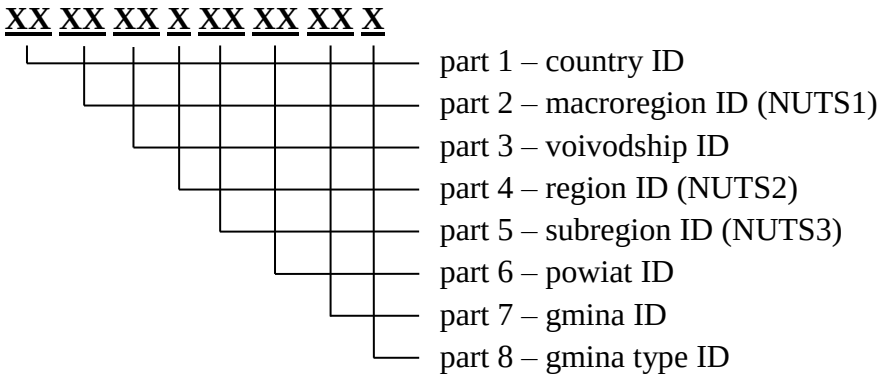


Fig. 1. KTS code breakdown

Country ID (part 1) for Poland is 10. Macroregions (part 2) and subregions (part 5) are numbered sequently within Poland. Regions (part 4) are numbered sequently within a voivodship. Voivodship, powiat and gmina IDs (parts 3, 6, 7 and 8) are taken from the TERYT register (a register of identifiers for territorial division units).

KTS codes for each unit are completed with zeroes to 14-digits. The following table presents an example of KTS coding for each level down to the same gmina unit that has been presented in Table 1.

| KTS code | territorial / statistical unit name |
|---|---|
| **10**000000000000 | Poland |
| 10**04**0000000000 | Macroregion |
| 1004**22**00000000 | Voivodship (TERYT) |
| 100422**1**0000000 | Region |
| 10042214**0**00000 | Subregion |
| 1004221401**1**000 | Powiat (TERYT) |
| 100422140111**011** | Gmina (TERYT) |

Table 2. KTS code construction for each level

Based on this new classification gmina geometries for 2002-2016 have been recoded with KTS codes and for each year 6 datasets have been created representing all levels of the new classification (macroregion to gmina). The KTS classification will serve as a basis for a Uniform Resource Identifier (URI) system for statistical units, which will be designed at a later stage of the project.

Gmina geometries for 2002-2016 have been generalized for purposes of presentation and dissemination (as INSPIRE statistical units). Geometries of higher level units have been built from generalized gminas. Generalization has been performed in GRASS GIS software.

During the analysis of data sources, including the development monitoring system STRATEG, it turned out that official statistics also disseminates data for functional areas and metropolitan areas. These areas are groups of NTS units, e.g. several gminas or several voivodships and are defined in local or regional strategies. Geometries for those areas have also been created in order to facilitate georeferenced data dissemination. There is a plan to construct identifiers for those areas within the KTS coding system to unify coding for all statistical units.

# 4 Data transformation plan into linked open data formats

One of the aims of the "Development of guidelines for publishing statistical data as linked open data" project is to create guidelines for publishing data from three major databases of Polish official statistics as linked open data. Those databases include: Local Data Bank, Demography Database and Development Monitoring System STRATEG.

From all three databases an excerpt of similar data has been selected for a pilot transformation including:

- Defining the scope of published data and their browsing methods,
- Creating ontology,
- Carrying out ontology mapping on the existing databases,
- Exporting data to RDF format,
- Uploading data to RDF data store,
- Publishing data on the linked data server.

The Ontop platform was selected for ontology modelling, creating mapping and exporting data to RDF format. It enables performing queries on the virtual RDF graph

established on the existing relational database. Furthermore, the platform is characterised by:

- Support for SPARQL 1.0,
- Interface for ontology modelling,
- Intuitive and advanced mapping language,
- Support for free of charge and commercial databases,
- SPARQL endpoint.

Despite the fact that the Ontop platform provides access to the RDF data store and a SPARQL query endpoint, for efficiency purposes it was decided to use a separate technology for this.

The selected tools for linked open data dissemination:

- Apache Jena Fuseki
    - Fuseki is a SPARQL server, it enables performing SPARQL queries with the REST API,
    - Data exported from the Ontop platform are imported into the Apache Jena platform.
- Pubby
    - Interface for linked data made available by SPARQL endpoints,
    - Enables creating a linked data server on the basis of a SPARQL endpoint. It is implemented as a Java web application,
    - Provides access to linked open data under the URI address.

The linked open data pilot implementation for the database excerpts with the above mentioned tools is still ongoing. Results are expected at the end of 2017.

# 5 Summary

The “Development of guidelines for publishing statistical data as linked open data” project is a valuable exercise for Central Statistical Office of Poland.

During the course of the project CSO had the opportunity to identify and catalogue the statistical data resources published on the information portal. The resources have been assessed in terms of their openness and described with metadata.

The project was also a great opportunity to identify and harmonize the geometries of territorial units used for statistical data dissemination. Data on administrative division boundaries acquired from the mapping agency proved to be very useful at this stage. A coherent identifier system and harmonized geometries are a key aspect in open data dissemination. The new NUTS 2016 classification introduced an incoherence between

the statistical and administrative division at NUTS2 level and forced CSO to create a completely new coding system for statistical units. For years 2002-2016 geometries of statistical units have been generalized and attributed with the new coding system.

Subsequently, excerpts from three major statistical databases: Local Data Bank, Demography Database and Development Monitoring System STRATEG have been selected for a pilot linked open data implementation, which includes testing various software tools. The project results, expected at the end of 2017, will provide a foundation for a full-on linked open data implementation by official statistics in the near future.