Development of guidelines for publishing statistical data as linked open data

Mirosław Migacz
Chief GIS Specialist
Central Statistical Office of Poland

INSPIRE Conference, Strasbourg, 7 IX 2017
The project

• January 2016 – January 2018
• „Merging statistics and geospatial information” grant series
• main goal: prepare a background for linked open data implementation in official statistics
Now

powiat Łobeski (LAU 1)
lobeski
4326418

4.4.32.64.18

3218

14

5

618

29

139x133
Specific objectives

• identification of statistical units for which data can be published with harmonization of their geometries for respective years
• building standarized URIs for statistical units
• identification and analysis of potential data sources
• plan for transformation of existing data sources into open formats
• pilot for creation of RDF metadata for data sources
• feasibility analysis for publishing linked open data
Data sources

• Three major databases:
  • Local Data Bank
    • biggest set of statistical information available for a wide range of years
    • updated monthly
  • Demography Database
    • integrated data source for state and structure of population, vital statistics and migrations
  • Development monitoring system STRATEG
    • a system for facilitating and monitoring the development policy
    • key measures to monitor execution of strategies at local, regional, transregional and EU level.
Data sources

- Other data sources:
  - publications
  - tables
  - communiqués
  - announcements
  - articles
Preliminary analysis of data sources

• Key aspects:
  • openness
  • redundancy of information
  • popularity (based on view and download statistics)
Data sources – inventory

• Metadata:
  • thematic category,
  • format (PDF, DOC, XLS, CSV),
  • spatial reference (country, NUTS, LAU, functional areas, urban areas),
  • temporal reference (years)
  • presence of identifiers (TERYT, NTS, NUTS)
  • update cycle
Statistical units harmonization

• Basis:
  • NTS (Nomenclature of Territorial Units for Statistical Purposes)

<table>
<thead>
<tr>
<th>Name</th>
<th>NTS</th>
<th>NUTS/LAU</th>
<th>Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>1</td>
<td>NUTS 1</td>
<td>1.6</td>
</tr>
<tr>
<td>Voivodship</td>
<td>2</td>
<td>NUTS 2</td>
<td>2.6.22</td>
</tr>
<tr>
<td>Subregion</td>
<td>3</td>
<td>NUTS 3</td>
<td>3.6.22.40</td>
</tr>
<tr>
<td>Powiat</td>
<td>4</td>
<td>LAU 1</td>
<td>4.6.22.40.11</td>
</tr>
<tr>
<td>Gmina</td>
<td>5</td>
<td>LAU 2</td>
<td>5.6.22.40.11.01.1</td>
</tr>
</tbody>
</table>
Statistical units harmonization

• Input data:
  • administrative boundaries since 2002 for LAU 2 (gmina)

• Harmonization process:
  • structure standardization
  • standardization of identifiers (creating NTS identifiers)
  • aggregation to higher level units (LAU 1 -> NUTS 1)
Statistical units harmonization

- Non-standard statistical units:
  - functional areas
  - urban areas
  - Integrated Territorial Investment areas
- Groups of NTS units
- Derive mostly from strategic documents
- Geometries created
- Changes of geometries in time to be determined
# Statistical units URIs

- **NTS as basic classification**

<table>
<thead>
<tr>
<th>Name</th>
<th>NTS</th>
<th>NUTS/ LAU</th>
<th>Identifier</th>
<th>URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>1</td>
<td>NUTS 1</td>
<td>1.6</td>
<td>...1/6</td>
</tr>
<tr>
<td>Voivodship</td>
<td>2</td>
<td>NUTS 2</td>
<td>2.6.22</td>
<td>...2/6/22</td>
</tr>
<tr>
<td>Subregion</td>
<td>3</td>
<td>NUTS 3</td>
<td>3.6.22.40</td>
<td>...3/6/22/40</td>
</tr>
<tr>
<td>Powiat</td>
<td>4</td>
<td>LAU 1</td>
<td>4.6.22.40.11</td>
<td>...4/6/22/40/11</td>
</tr>
<tr>
<td>Gmina</td>
<td>5</td>
<td>LAU 2</td>
<td>5.6.22.40.11.01.1</td>
<td>...5/6/22/40/11/01/1</td>
</tr>
</tbody>
</table>

[http://nts.stat.gov.pl/5/6/22/40/11/01/1](http://nts.stat.gov.pl/5/6/22/40/11/01/1)
NUTS 2016

NUTS 2 = voivodeships

NUTS 2 ≠ voivodeships
NUTS 2016 - consequences

• **NTS classification repelled** due to difference between administrative and statistical division on level 2

• no information on potential similar changes in the administrative division following NUTS 2016

• problems in several dissemination systems:
  • INSPIRE services (statistical units published as NTS)
  • Geostatistics Portal (based on a 5-level statistical units system)
  • no identifier system ready to be adapted for LOD ...
KTS – a new classification for statistical units

- introduced August 24th 2017
- combines statistical and administrative units (again 😊)
- 14 digits

<table>
<thead>
<tr>
<th>description</th>
<th>source/value</th>
</tr>
</thead>
<tbody>
<tr>
<td>country code</td>
<td>10</td>
</tr>
<tr>
<td>macroregion code</td>
<td>NUTS1</td>
</tr>
<tr>
<td>voivodship code</td>
<td>TERYT</td>
</tr>
<tr>
<td>region # within a voivodeship</td>
<td>1 or 2 (NUTS2)</td>
</tr>
<tr>
<td>subregion code</td>
<td>NUTS3</td>
</tr>
<tr>
<td>powiat code</td>
<td>TERYT</td>
</tr>
<tr>
<td>gmina code</td>
<td>TERYT</td>
</tr>
<tr>
<td>gmina type</td>
<td>TERYT</td>
</tr>
</tbody>
</table>

http://
KTS – a new classification for statistical units

Before (NTS): 5.4.32.64.18.05.3

After (KTS): 10023216418053

<table>
<thead>
<tr>
<th>symbol</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000000000000000</td>
<td>Poland</td>
</tr>
<tr>
<td>1002000000000000</td>
<td>macroregion</td>
</tr>
<tr>
<td>1002320000000000</td>
<td>voivodship</td>
</tr>
<tr>
<td>1002321000000000</td>
<td>region</td>
</tr>
<tr>
<td>1002321640000000</td>
<td>subregion</td>
</tr>
<tr>
<td>1002321641800000</td>
<td>powiat</td>
</tr>
<tr>
<td>10023216418053</td>
<td>gmina</td>
</tr>
</tbody>
</table>
Statistical units harmonization (again 😊)

- datasets with administrative/statistical unit boundaries 2002-2016 recoded according to the new classification (KTS)
- several generalization scenarios tested for purposes of choosing an optimal one
- datasets with generalized and non-generalized geometries prepared for 2002-2016
Pilot transformation of three major databases into LOD

- selection of similar database fragments (basic demographic classifications),
- designing ontologies re-using existing vocabularies, e.g. Dublin Core, SKOS,
- setting up a software environment for pilot RDF creation (Ontop, Apache Jena Fuseki, Pubby)
Other project activities

- pilot transformation of geographic data into LOD,
- guidelines for implementing LOD in search services (for identified data sources),
- development of guidelines for a statistical linked open data web application.
Poland’s data opening strategy

- launched last year
- aimed at opening data resources of government institutions with respect to the 5-stars of linked open data goals
- the grant results (guidelines) in line with the strategy
- increased probability of acquiring financing for a fully fledged implementation
- LOD implementation planned as part of new developments in the Geostatistics Portal
INSPIRE Thematic Clusters

https://themes.jrc.ec.europa.eu – collaboration platform

**Statistical Cluster:**

- statistical units
- population distribution (demography)
- human health and safety

Cluster members meeting after this session (15:30-16:30) @ the INSPIRE booth
Development of guidelines for publishing statistical data as linked open data

Mirosław Migacz
Chief GIS Specialist
Central Statistical Office of Poland

@mireslav
www.linkedin.com/in/migacz
m.migacz@stat.gov.pl
www.slideshare.net/MirosawMigacz