



# 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*

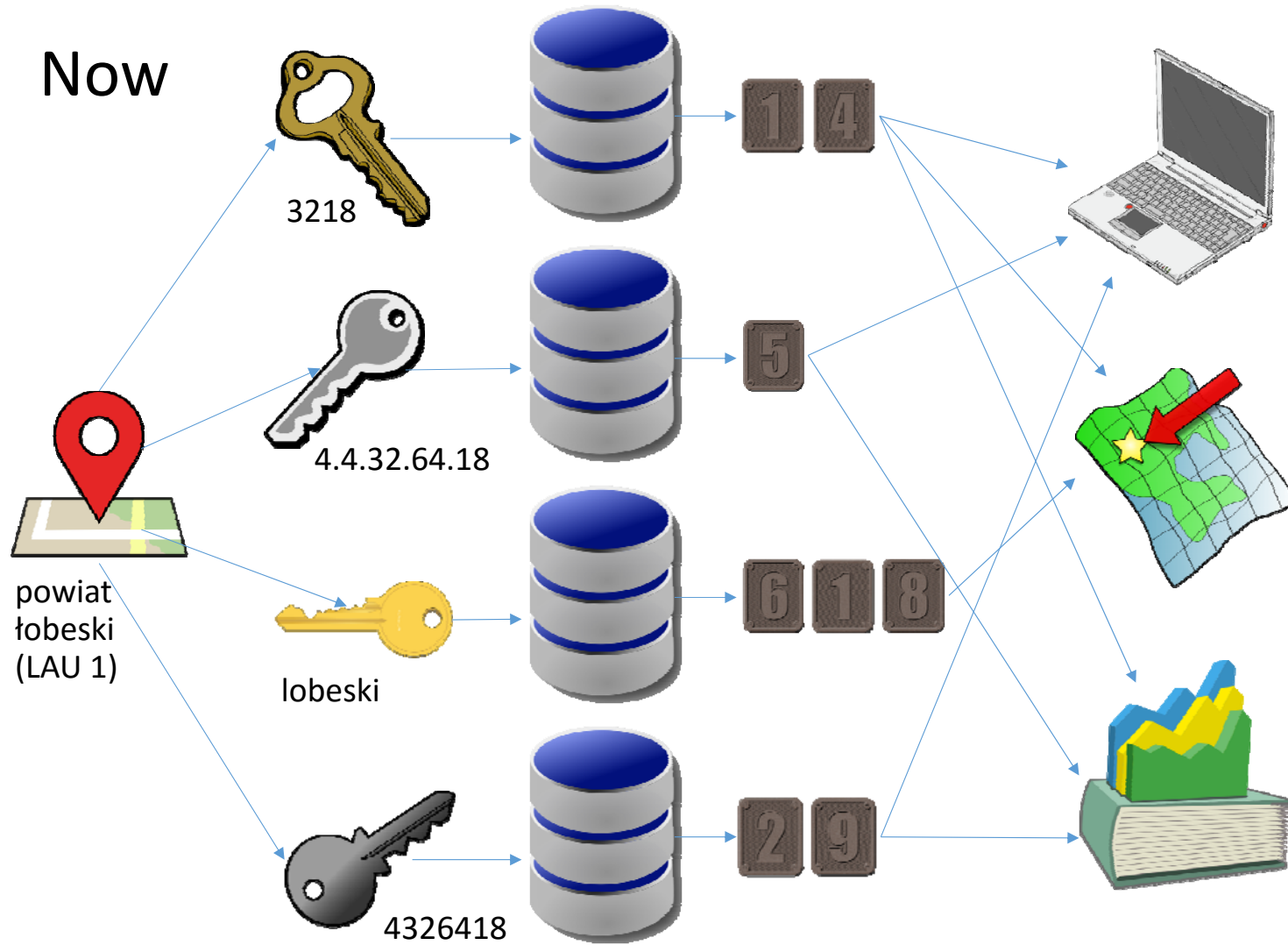


# The project

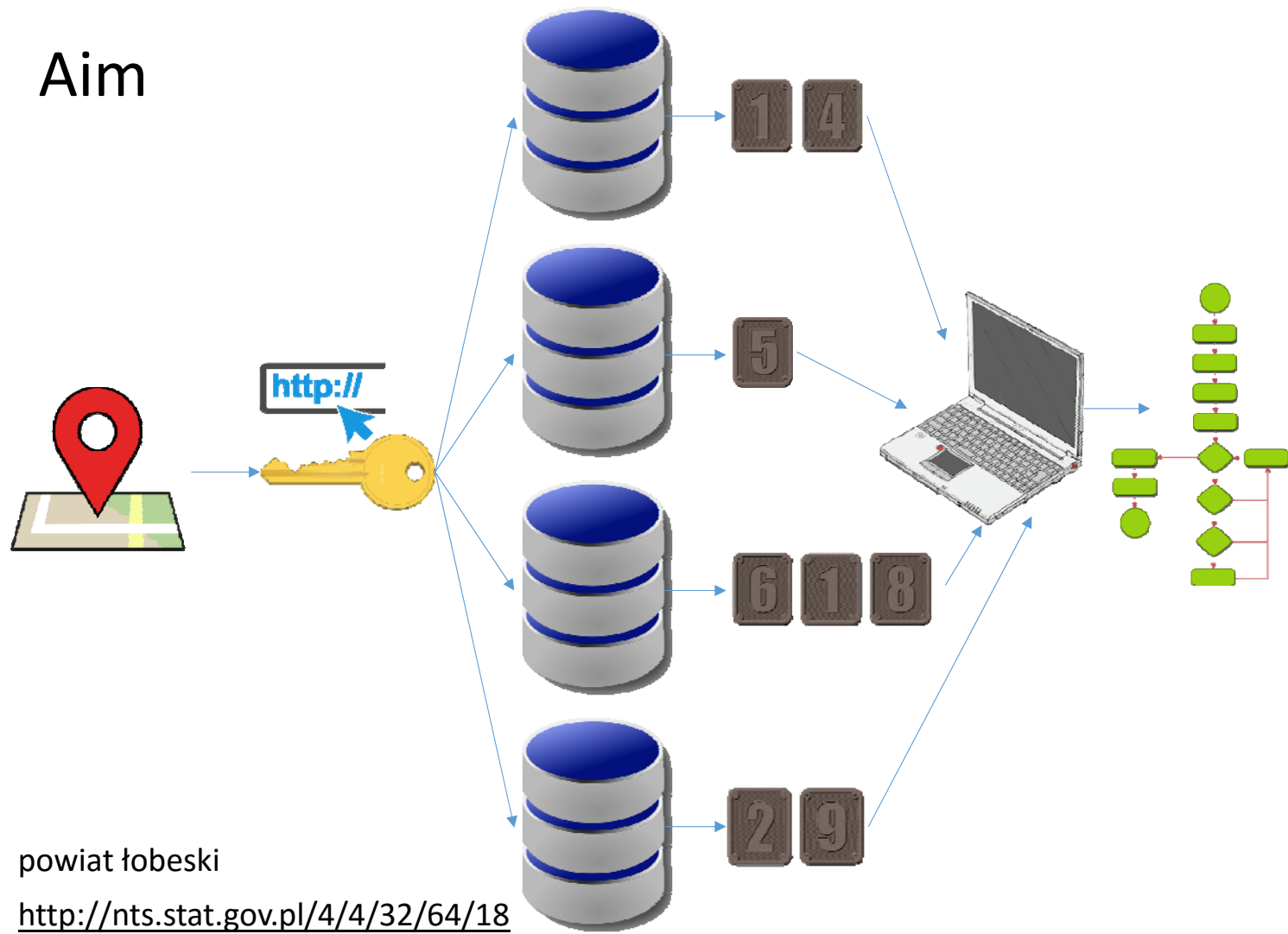
- January 2016 – January 2018
- „Merging statistics and geospatial information” grant series
- main goal: prepare a background for linked open data implementation in official statistics



Now










Aim





## Specific objectives

-  • identification of statistical units for which data can be published with harmonization of their geometries for respective years
-   
 • building standardized 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.



BAZA DEMOGRAFIA





# Data sources

- Other data sources:
  - publications
  - tables
  - communiqués
  - announcements
  - articles

No.	Type ^ v	Document title ^ v	Date of publication ^ v
1		A guide on surveys and studies concerning cross-border areas	27.06.2016
2		Accidents at work and work-related health problems	21.03.2014
3		Accidents at work in 2014	27.10.2015
4		Activities Of Associations And Similar Organizations, Foundations, Faith-based charities, Business and Professional Associations In 2014 - preliminary results	04.04.2016
5		Activities of non-profit organizations in 2013: Management, cooperation and social services	17.03.2016
6		Activity of enterprises having foreign entities in the years 2008-2014	29.04.2016
7		Activity of enterprises with up to 9 persons employed in 2014	18.02.2016

Specification	The implementation of each month by calendar days											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
Price index of consumer goods and services (month as related previous month)	15	12	15	11	12	13	11	12	12	11	14	12
Price index of consumer goods and service (January 2016 as related to December 2015 - final data)	--	--	15	--	--	--	--	--	--	--	--	--
Price index of consumer goods and service (quarter as related previous quarter)	15	--	--	11	--	--	11	--	--	11	--	--
Price index of consumer goods and services for the first three quarters of a year	--	--	--	--	--	--	--	--	--	11	--	--
Price index of consumer goods and services in first half (as related to the first half of previous year)	--	--	--	--	--	--	11	--	--	--	--	--





# Preliminary analysis of data sources

- Key aspects:
  - openness
  - redundance 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)

Name	NTS	NUTS/LAU	Identifier
Region	1	NUTS 1	1.6
Voivodship	2	NUTS 2	2.6.22
Subregion	3	NUTS 3	3.6.22.40
Powiat	4	LAU 1	4.6.22.40.11
Gmina	5	LAU 2	5.6.22.40.11.01.1





# 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

Name	NTS	NUTS/ LAU	Identifier	URI
Region	1	NUTS 1	1.6	...1/6
Voivodship	2	NUTS 2	2.6.22	...2/6/22
Subregion	3	NUTS 3	3.6.22.40	...3/6/22/40
Powiat	4	LAU 1	4.6.22.40.11	...4/6/22/40/11
Gmina	5	LAU 2	5.6.22.40.11.01.1	...5/6/22/40/11/01/1

<http://nts.stat.gov.pl/5/6/22/40/11/01/1>

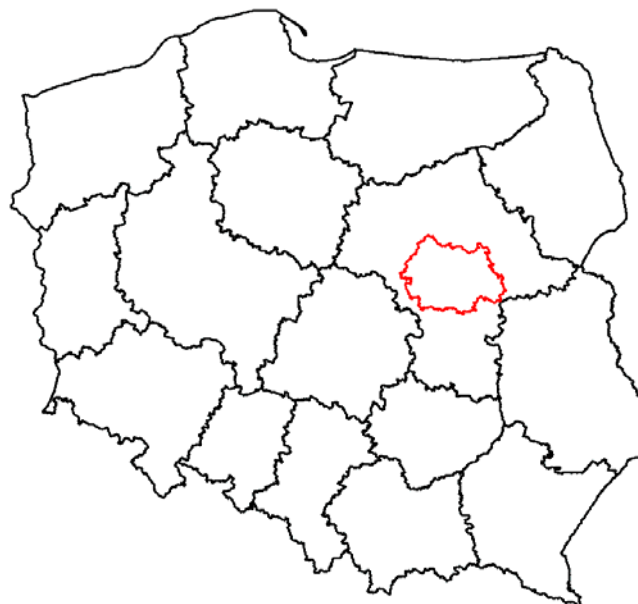




# NUTS 2016



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

	description	source/value
XX	country code	10
XX	macroregion code	NUTS1
XX	voivodship code	TERYT
X	region # within a voivodeship	1 or 2 (NUTS2)
XX	subregion code	NUTS3
XX	powiat code	TERYT
XX	gmina code	TERYT
X	gmina type	TERYT







# KTS – a new classification for statistical units

Before (NTS):

5.4.32.64.18.05.3

After (KTS):

10023216418053

symbol	name
1000000000000000	Poland
1002000000000000	macroregion
1002320000000000	voivodship
1002321000000000	region
1002321640000000	subregion
1002321641800000	powiat
10023216418053	gmina





# 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



BAZA DEMOGRAFIA



- 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](http://www.linkedin.com/in/migacz)



[m.migacz@stat.gov.pl](mailto:m.migacz@stat.gov.pl)



[www.slideshare.net/MirosawMigacz](http://www.slideshare.net/MirosawMigacz)