

Statistics Poland linked open data pilot – short summary of conclusions and recommendations

The following questions had to be tackled

1. How can the data integration method be described?
2. What has been the motivation to execute the method? (advantages, benefits)
3. What are the needs / the lacks identified which have still to be tackled?
4. What might be the main message / recommendations for the management level from the example?

Topic	Level	Stakeholders	BKG
A common approach needed for statistical linked open data – NSIs look to Eurostat for pan-European guidelines, incl. common vocabularies, data scope recommendations, reference implementations. A ESSNet project on Linked Open Statistical Data (ESSNet LOSD) conducted after Statistics Poland’s pilot came up with the same realizations.	European	NSIs, Eurostat	ref. 2: Linked open data (LOD) can be highly interoperable method for data publishing, as it is machine-readable. It allows automatic processing of large amounts of data across various sources, provided they are linked together and use the same vocabularies. ref. 4: A common approach led by EU-level institutions/working groups should be proposed in order to maximize the benefit of publishing data as LOD
Semantic harmonization of statistical classifications needed – notations of classification elements can look the same but have different meanings, e.g. age groups 0-5 can mean “0 to 5” or “0 to less than five”. This should also be addressed in pan-European guidelines.	Internal / European	NSIs, Eurostat	ref. 3: Clear definitons of statistical classifications need to emerge in order to make sure that data, which is linked from different sources, refers to the same phenomenon
Reference geometries for statistical data – for publishing statistical data it is essential to be able to reference correct geometries for relevant years. Most often statistics are published with reference to administrative units, which change over time. An ideal situation would be to publish new geometry instances for units only when the boundary actually changes and define their period of	European	NSIs, NMAs	ref. 4: Statistical world needs historical boundary data to properly reference time series data

<p>validity in the metadata. This however would require a thorough inventory of boundary changes done in NMAs. Therefore: NMAs should take into consideration that NSIs need historical boundaries to reference statistical data to.</p>			
<p>Statistical linked open data is not a technical challenge, it's a data challenge – while this conclusion was formulated by the ESSNet LOSD project, Statistics Poland came to the same conclusion. The biggest challenge to publish correct and linkable linked open data is to encode it correctly, use appropriate and preferably existing vocabularies. In simple words: to publish data you need to know your data well.</p>	Internal	NSIs	<p>ref. 1: Linked open data (LOD) is publishing data as resolvable web addresses, which can be read automatically by computers, scripts etc. LOD can be used to publish all kinds of data incl. spatial data, statistical data etc. Using the same vocabularies – describing the data in the same way – by different data producers allows linking data from various sources.</p> <p>ref. 4: Even though LOD seems complicated at first, the technical difficulties can be overcome quite quickly (e.g. using dedicated applications with a graphic user interface). The main challenge lies in knowledge about data, which is published, as it needs to be properly described with metadata to make it most usable.</p>