Stable identifiers for historical topographic objects

Grzegorz Myrda[*1] and Tomasz Panecki[*2]

[1]Institute of History, Polish Academy of Sciences – Poland
[2]Institute of History, Polish Academy of Sciences – Poland

Abstract

During the works on developing spatial database in the "Ontohgis.pl" (https://ontohgis.pl/) project, the subtask of harmonizing gathered data on settlements and administrative units was fulfilled by using external databases maintained by Polish state geodetic service (gugik.gov.pl). The problem emerged however, how to deal with regions outside modern Poland which are not covered by Polish state spatial databases. The solution would be to use worldwide databases which can be used during the process of harmonization of historical spatial data, such as: Getty Thesaurus of Geographic Names (TGN), World Historical Gazetteer (WHG) or Wikidata. One of the most important criterion of selecting a suitable database as a reference dataset for harmonization is that it should use stable identifiers.

What is a stable (permanent) identifier (PID) though? It is the identifier which provides a long lasting and stable reference to the defined digital resource. It is used in citing, analysing, integrating and harmonizing data. Features of a good PID encompass: uniqueness, stability (both original and copies), possibility to work as URL, compatibility with URI standards, access to metadata (as well its fragments and versions), no semantics and safety. Besides, for harmonizing purposes, additional functionalities of external, reference databases are: full or partially full data coverage for settlements and – preferably – for administrative units, possibility of adding own settlements (single entries) in any moment by anyone (including adding administrative units), possibility of adding features without coordinates (when we do not know feature's location). During harmonization process, features identity is crucial and we can indicate three criteria which constitute the same feature over time: proper name, location and type.

It is also important to bear in mind that coordinates do not form clear identifiers as features' location (coordinates) are one of their attributes which can change over time. Just like features' proper names shall not be used as identifiers, neither should coordinates. In other words: we would like to keep the identifier of the feature, not the place.

An example of "Adamowo" village which is located within contemporary Poland, identifier in various databases in given below:

GUGIK: https://pzgik.geoportal.gov.pl/prng/Miejscowosc/PL.PZGiK.204.PRNG.00000000-0000-0000-0000-000000000100-67

Geonames: https://www.geonames.org/776782/adamowice.html

GUS: https://bdl.stat.gov.pl/BDL/metadane/teryt/miejscowosci/2186

---

[*]Speaker

Wikidata: https://www.wikidata.org/wiki/Q4680202

IHPAN: https://atlas.ihpan.edu.pl/geoserver/ontohgis/wfs?service=wfs&version=1.1.1&request=getfeature&typen

For the area outside contemporary Poland other than official databases and identifiers have to be used to data harmonization, such as: Open Street Map (OSM), Geonames, WHG, TGN or Wikidata. However, in case of OSM a question arises if we can depend on the features' identifiers in this database? The answer provided by one of the OSM contributors on StackOverflow was negative as OSM changes features' identifiers from time to time (e.g. when feature is deleted and then re-created)[1]. It is unknown if the problem is still actual. Preliminary studies show that the Wikidata may prove to be the most suitable solution as it provides stable identifiers as well as the possibility do add new features. During the presentation various approaches will be tested, analysed and presented. The work will contribute to the problem of historical spatial data harmonization.

https://gis.stackexchange.com/questions/279755/can-i-depend-on-the-country-city-ids-of-osm?rq=1