Evaluating Quality of Spatial Data Coming from Multiple Suppliers. Case Finnish National Topographic Database.

Mari Isomäki

National Land Survey of Finland, Application Services, Helsinki, FI-00521, Finland
mari.isomaki@maanmittauslaitos.fi

The Finnish National Topographic Database (NTDB) is a centralized database for topographic data of Finland. It consists of buildings (2.5D and 3D), road links, other manmade structures and addresses. The NTDB consists of data coming from multiple suppliers: governmental organizations and municipalities. Different suppliers use varying systems in gathering and storing spatial data among each other's (Lundvall 2019). Because the systems vary greatly, also the data supplied comes in many different forms in regard of their format and quality. In effect, large diversity of data is the biggest challenge in collecting national data into a centralized database.

To overcome the versatility of data, schema transformation and quality check must be implemented on every data import. Data validation is based on quality specifications formatted as *quality rules*. Quality rules are derived from INSPIRE and current national topographic database norms, quality rules produced by European Location Framework project (ELF) and international and national standards. Quality rules are divided into three ISO 19157 quality elements: format consistency, domain consistency and topological consistency (ISO/TS 19157:2014). This paper focuses on implementing quality rules on spatial data: what is being tested before importing data into the NTDB and how quality rules are implemented.

Before a data supplier can import data into the NTDB, a corresponding schema transformation document is created into the automated data import system called *Quality Guard and Data Upload Service*, of which schema transformation is an integrated part of. During schema transformation, attribute names and values are converter to the ones used in the NTDB. After the schema transformation, data is compatible with the data model used in the NTDB and with the quality rules.

In Quality Guard and Data Upload Service, there are fifteen individual rule types (fig 1) and 351 different quality rules. Besides rule type, each rule consists of rule identifier, attribute it is targeted on, feature type that is being validated, severity, description and rule parameters. Rule parameters contain detail-leveled information of what is being validated, whereas rule type guides how a rule is tested. Based on feature type and attribute, correct quality rules can be targeted to appropriate features. The most crucial rules, such as

geometry validity, empty geometry and attribute data types are tested on all features passing Quality Guard and Data Upload Service.

Rule type	What is tested?
Not null	Attribute has a value
Character length	Value consists of certain amount of characters
Geometry type	Geometry is the right type (area, point, or line)
Value range	Value belongs in a predefined range of values
Belongs in a set	Value belongs in a predefined set of values
Data type	Data type is correct (integer, double, numeric, boolean, string, timestamp or date)
Distance	Distance between features (features that are linked to each other's can only have certain distance between them)
Compare	Value must be bigger, smaller or equal compared to another value
Overlaps	Features must not overlap more than defined ratio
Geometry validity	Geometry meets OGC SFSQL standards
Empty geometry	Feature has a geometry
RegEx	Value is consistent with a given regular expression
Within	Feature is within a given area
Name list	Value is found on a list (used to find out misspellings in address names)

Fig 1. Quality rule types implemented in building imports.

Depending on a quality rule, not passing a quality check causes either a warning or an error. All rules related to geometry validity cause an error. Instead, attribute rules usually only cause an error when data type is not correct. A feature causing error will not be inserted into the NTDB, but a feature causing warning will. On the both cases, failed feature will be added in automatically constructed quality rapport data supplier receives in case warnings or error are discovered. Quality rapport is a shapefile where every failed feature is represented as a point marking failed feature's center point. However, if error relates to geometry invalidity, the point is in the invalid spot. Besides error location, quality rapport also includes rule description, identifier, severity and original value. In the case a rule compares a value to another, also compared value is represented in the quality rapport.

Having correct results in data validation depends fully on correct schema transformation. Because quality rules expect input data to be in a certain schema, schema transformation is actually the most crucial part of data validation in Quality Guard and Data Upload Service. Moreover, schema transformation document is created manually, which makes it prone to mistakes. In the case schema transformation is not done correctly, in addition to having false errors, data upload process usually fails when starting to write features into database because of mismatches between NTDB and input data.

Both, schema transformation and data validation, are being implemented in FME based application. FME is a versatile and effective platform for building ETL (extract, transform & load) workflows. Workflows used by Quality Guard and Data Upload Service are parametrized, making the application dynamic and automated. FME offers a great number of built-in tools to test, route and manipulate data. Also, many spatial data tools, such as spatial filtering and geometry validation, are available. In case FME's built-in tools do not offer solution, Python, TCL and SQL can be used. For example, overlap and distance rules are executed by Postgis functions instead of built-in tools, because queries are more efficient compared to built-in functions (fig 2).

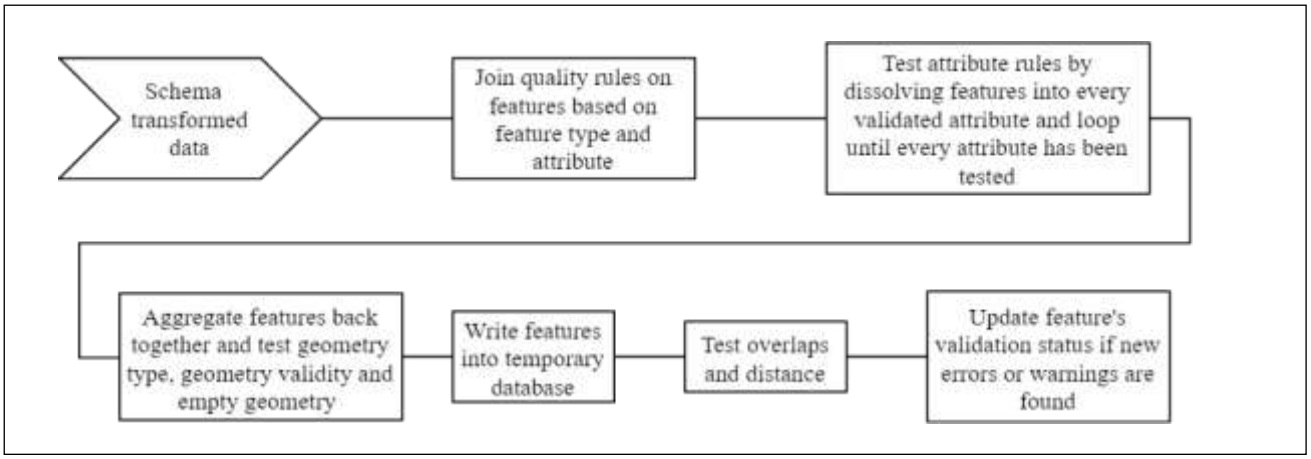

Fig 2. Data validation process in Quality Guard and Data Upload Service

To conclude, NTDB can be built on data that is versatile in format and in quality. However, schema transformation and data validation are crucial parts in the process. Either cannot be skipped, when building a centralized database that uses diverse data sources and only includes high-quality data. Furthermore, such a process should be as dynamic as possible, since dynamic process often lead to easy maintainability.

References

ISO/TS 19158:2012, *Geographic information − Quality assurance of data supply.*
ISO/TS 19157:2014, *Geographic information − Data Quality.*
Lundvall, A. (2019), The topographic database of the future is being built right now. Positio.