4 Quality of Data

WeGovNow aims on co-design between public service organisations and citizens designing better public services and effective community initiatives towards local solutions but also on co-management of urban services. It is obvious that WeGovNow deals with a large amount of data, mainly geographical and urban data, collected from public services and local stakeholder initiatives. Besides urban data, WeGovNow, as a crowd-source-based platform, also deals with user/citizen derived data. Users produce and manage data concerning urban entities such as everyday-life issues, proposals, projects, voting reports, local groups, etc. So, in order to assess the quality of data in WeGovNow we need to address:

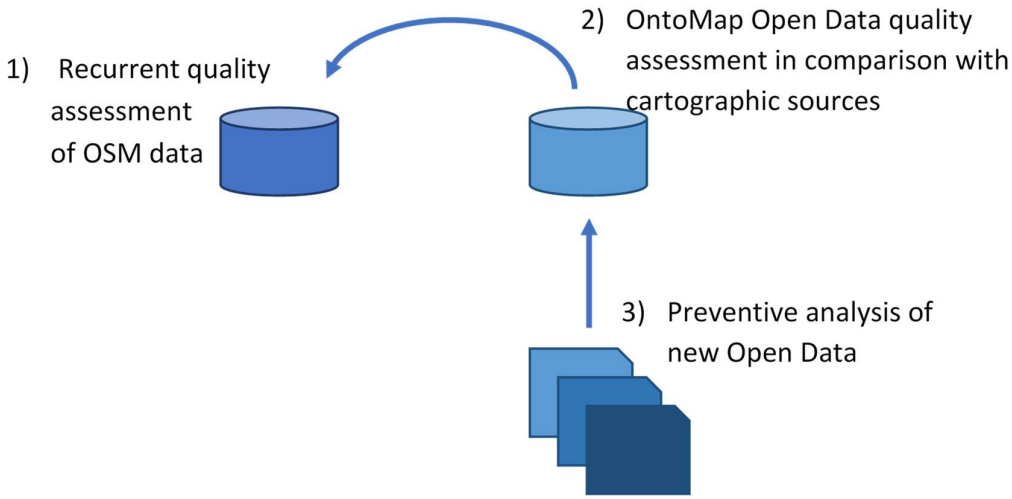
1) The context information provided to end users (urban data), cartographic source (OSM) and open data (OntoMap) about the city (including non-geospatial data), and
2) The user generated data (user activities and user actions within the WeGovNow ecosystem).

Thus, the general approach towards data quality within WeGovNow is therefore articulated considering two type of data sources, namely; **static** and **dynamic** sources:

Static sources

These are data sources that are not being updated very frequently such as urban data (e.g. schools, parks, traffic lights, etc). The approach to quality assessment of urban data in WeGovNow is depicted in Exhibit 12

Exhibit 12. Approach to quality assessment of Urban Data in WeGovNow



Data quality assessment of urban data is a batch activity procedure aiming at providing quality contextual data to WeGovNow users. The assessment should answer to the question **"is the source good enough to support users' activities?"**. In other words, the cartographic information and open data provided by WeGovNow should not mislead users in their

evaluations but should, instead, support the collaboration and cooperation within the platform.

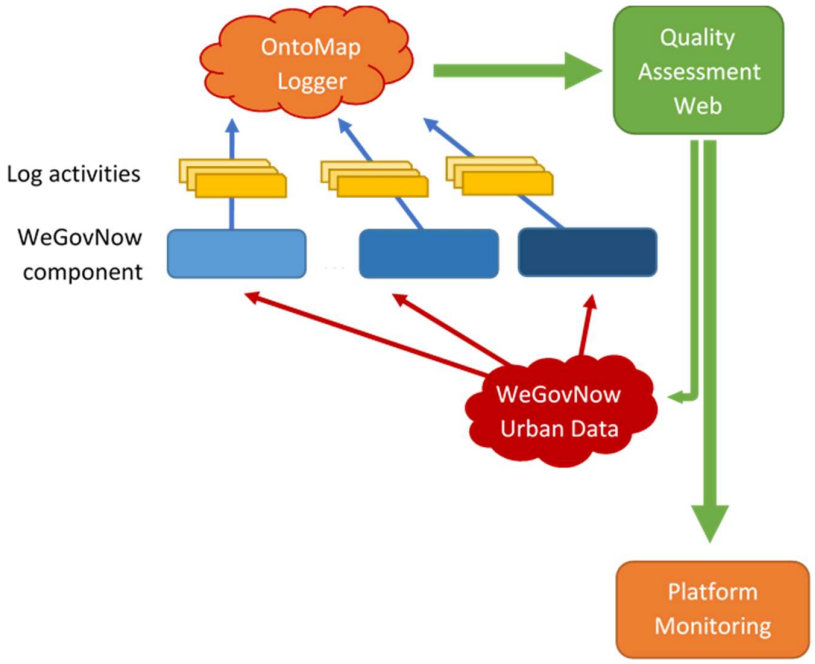
During the preparation phase of WeGovNow platform (1st revision period) we assessed data quality based on batch process, and more specifically; i) the OSM cartography completeness of the trial sites (see sections 4.1 and 4.2) and ii) the coherence and completeness of the open data available in OntoMap (in comparison with OSM data source and google maps), as explained in sections 4.3.

Dynamic sources

Dynamic sources refer to data generated by users while using/interacting the WeGovNow platform. For that case, the batch process, followed on static sources case, is not suitable. Instead we are following the real-time approach. In this case, user data need to be analysed frequently (at least on daily basis) to assess the internal coherence of users' activities, the revision of urban data and the monitoring of the platform.

To approach the analysis and monitoring of users' data, in WeGovNow we introduced a common logger and translation system. Each component generates different data in their own schema, then OntoMap logger translates the schema into a common format and it provides a homogeneous source data focused on users' activities (see Exhibit 13, log activities). The assessment and monitoring will be provided by software component, accessible as web service, as explained in section 4.4.

Exhibit 13. Quality assessment of user generated data in WeGovNow platform.



The quality assessment of user generated data in WeGovNow OntoMap is therefore the analysis of components logs in a OntoMap-compatible common format, and based on a metaphor in common among current and future components.

The implementation of the quality assessment web service will follow the development lifecycle of WeGovNow prototype, and will be tailored on the applicative scenarios collected during the engagement activities on the trial sites.

In the following sections are presented the theoretical framework for the future quality assessment web service, the general architecture of the web service, the conceptual approach and the results of the quality assessment of the urban data provided by WeGovNow.

4.1 Data in WeGovNow

As mentioned previously, WeGovNow platform handles the following data sources:

Static:

- OpenStreetMap (OSM) data
- Public Sector Information (PSI) and open data

Dynamic:

- User provided data (including non-geospatial data)

Each of the different type of data, is further analysed below:

OpenStreetMap (OSM) is utilized mainly as a base map for location based services of WeGovNow. OpenStreetMap is the most popular VGI (Volunteered Geographic Information) project. OSM allows users to digitize new map features and modify existing objects. 3.8 million of users have been registered since 2005. Today OSM's database stores more than 3.5 billion of nodes covering the whole planet. The project is very active and competes with traditional map providers (e.g., state mapping agencies) and popular commercial internet map providers (e.g., Google Maps). OSM allows anyone to download, modify and distribute their data under a liberal open license for any purpose (including commercial). OSM provides the following groups features: roads (including objects related to road infrastructure (e.g., sidewalks), buildings, water objects (e.g., rivers, lakes), land cover or land use data, amenity points, ferry routes, power lines, railways, railway stations. Most of the data are buildings, roads and land cover data. OSM uses a topological data format. Thus, shared borders and nodes of multiple features are stored only once. It allows users to easily input geographical features and prevents topological errors. OSM provides metadata (e.g., usernames, change sets' information, time of creation and modification of an object). Attribute data are provided by tags in a "key-value" format. This approach is flexible and allows users to apply any number of characteristics and any datatypes.

Public Sector Information (PSI) and open data are provided through OnToMap. Point, polyline and polygon features related to a city infrastructure (e.g., bike lanes, urban parks, hospitals, etc.) are provided as PSI datasets. WeGovNow's Location Based Services and Applications will widely utilize PSI data. It allows users to select geospatial objects on an interactive map, derive and manipulate with attribute information (e.g., address of a cinema stored as an attribute), search (e.g., find a shop by its name), etc. PSI data could be retrieved by OnToMap's API. An application of WeGovNow may derive PSI data by this API using different parameters (e.g., bounding box, language, etc.). The datasets are provided in GeoJSON format. Higher level manipulations may be implemented on a side of a specific application. Objects of city infrastructure are provided as point, polygon and polyline features. Currently, the following data types are provided: urban parks, schools, stores, markets, bike lanes, restaurants, museums, places of worship, monuments, libraries, health social services, hospitals, drug stores, law enforcement objects, art galleries, clubs, etc.

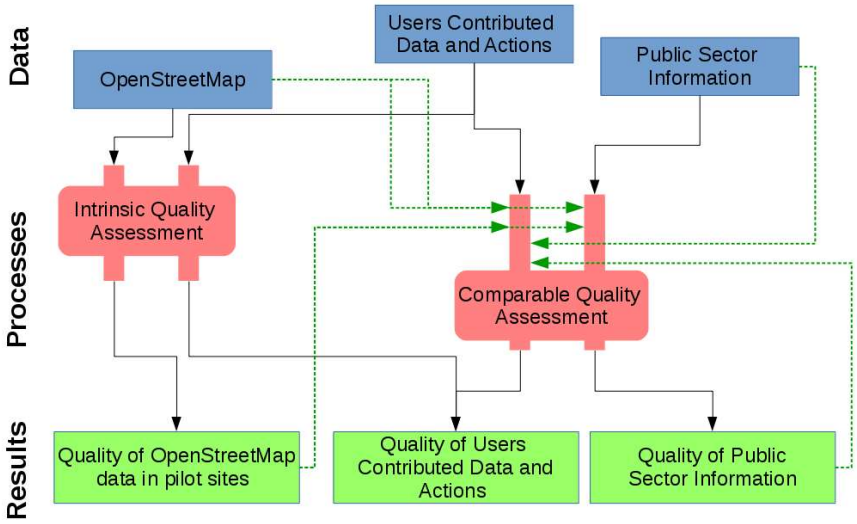
Users contributed data and actions are registered by OnToMap's Logger to achieve homogeneity and easier querying from the various WeGovNow components. It enables the developers to track activity of users, detect possible problems and possible low-quality input. OnToMap's Logger stores events sent by applications of WeGovNow. Similar to PSI data, OnToMap uses the GeoJSON data format for Logger. An event sample may look as follows:

```
{
  "type": "Feature",
  "properties": {
    "hasType": "School",
    "external_url": "http://xxxxxxxxx",
    "hasName": "NUOVO ACTIVITY OBJECT",
    "additionalProperties": {
      "description": "Descrizione aggiunta in fase di update"
    }
  }
}
```

WeGovNow's logged relevant events could be displayed by a specific application. The quality of the mentioned data types is assessed in a frame of the WeGovNow platform. A concept of quality evaluation is further presented;

The overall concept of data quality assessment approach on WeGovNow platform is illustrated in the following Exhibit 14.

Exhibit 14: General concept of quality assessment of WeGovNow data



In the above exhibit, data flows are depicted by black arrows. Data types are presented by blue. Quality assessment processes are shown by red. Results are marked by green. Green dashed arrows depict utilized reference datasets for comparable quality assessment. OSM data are assessed intrinsically. PSI data are evaluated comparably, using OSM and results of its quality assessment as reference datasets. User Contributed Data and Actions are assessed both intrinsically and comparably. OSM and PSI data, as well as results of their assessment, are utilized as reference datasets for comparable assessment of User Contributed Data and Actions quality.

Data quality is defined in ISO 9000:2015 as the degree to which a set of characteristics (e.g., completeness, validity, accuracy, consistency, availability and timeliness) of data fulfils requirements. In other words, it could be expressed as degree of correspondence of information to user's expectations. Data quality indicators could be defined in the various contexts: completeness, accuracy, precision, trustworthiness, credibility, etc. In order to obtain a comprehensive evaluation of quality, a set of approaches covering wide range of contexts should be utilized.

When making use of data with a geospatial aspect, it is important that users are aware of the underlying quality of the data so that they can make an informed decision that the information is "fit for use". With regards to this data quality, standards are available (such as the ISO 19157 standard on geospatial data quality) that define a number of metrics that can be used to assess this. These metrics include aspects such as geographic accuracy, logical consistency, temporal consistency and data omission/commission. Within the WeGovNow project, research has been conducted as a means of identifying what metrics are important for different types of users (i.e. a layman user looking up local initiatives, or a council employee assessing the density of requests for assistance in a local community), determining how these metrics can be derived from the data, and what methods are appropriate for

portraying these quality metrics to the end user. Overall, 82 main methods are presented in the review. The first page of document is depicted below. The full document is available in Appendix B.

Exhibit 15: Data quality methods

Id	MId	Metric	Evaluator	Primary Attributes	Method	Source	Intrinsic/Extrinsic	Priority
1	Cl1	**Completeness** It is the presence and the absence of features, their attributes and relationships. **Commission** (Excess data present in a dataset) **Omission** (data absent from a dataset)	**Commission/Omission**	-Geographic coordinates of features -Road data sets	Road length comparison of OSM data with authoritative data in cell grids.	How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordance Survey datasets.	Extrinsic	2
2	Cl2		**Commission/Omission**	-Geographic coordinates of features -Area data sets	Visual inspection of urban areas for incompleteness in OSM in comparison with authoritative data. The examination of the raster tiles could show the percentage of completeness.	How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordance Survey datasets.	Extrinsic	2
3	Cl3		**Commission/Omission**	-Geographic coordinates of features -Road, lake and river data sets	OSM data was compared with authoritative data for the number of objects, the total length and the area of objects. Roads zones, lakes and rivers was compared in two datasets and the percentage of OSM completeness comes out.	Quality assessment of the French OpenStreetMap dataset	Extrinsic	2
4	Cl4		**Commission/Omission**	-Geographic coordinates of features -Road data sets	Assess how complete are the OSM data -separated in different feature categories- in comparison with Google and Bing data. Google and Bing datasets need conversions and transportations before	Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps	Extrinsic	3

The table contains the following columns:

1. methods' ids (numeric and alphanumeric),
2. metrics' names with short descriptions,
3. correspondent evaluators,
4. primary attributes,
5. short description of method itself,
6. source of a method

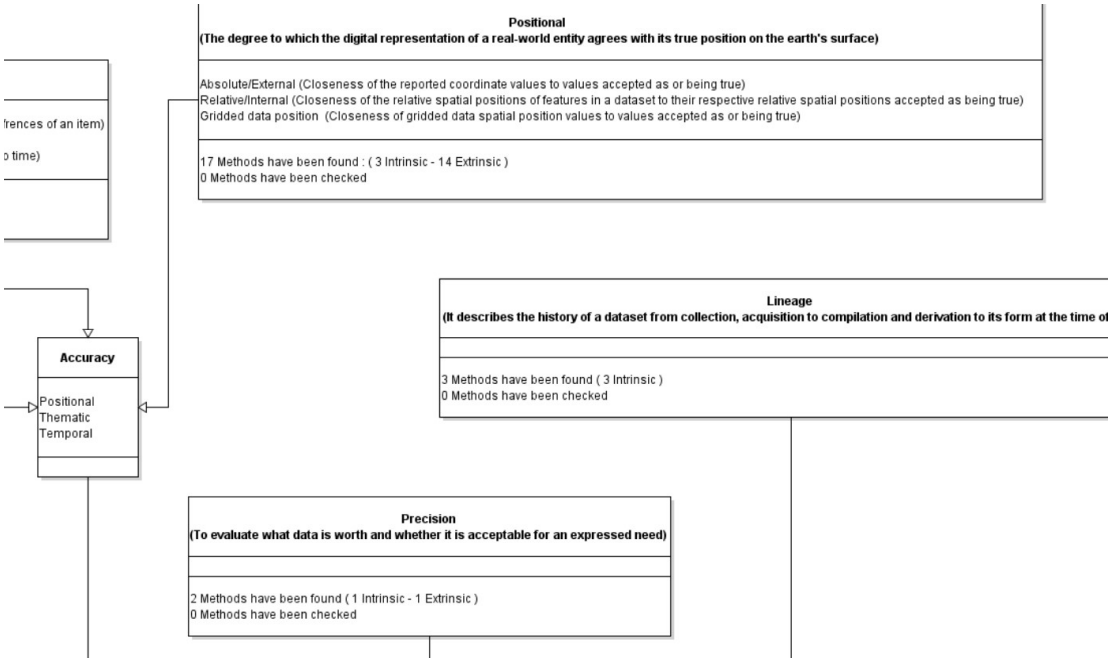
It should be noted that notations and terminology are not well established in spatial data quality assessment area (especially when intrinsic parameters are considered), as well as consensus between authors of methods.

Ids (1) are utilized for internal use, they do not play a meaningful role. Method are classified by the Metric column (2). It is partially based on ISO recommendations. Evaluator column (3) states particular name of quality evaluator. In Method column (4), the implementation is shortly described. Name of relevant paper is described in Source column (5).

Since it is quite problematic to obtain quality data for using as reference sources, it was decided to concentrate attention mainly on intrinsic analysis. In order to distinguish the approaches, Intrinsic/Extrinsic (6) column has been added to the table. The subjective estimation of applicability of methods is provided by Priority column (7): 1 (green) - high priority, 2 (yellow) - medium, and 3 - low.

In order to describe relations of the reviewed methods and present review graphically, UML diagram has been prepared. For illustrative purposes, a part of the diagram is depicted in Exhibit 16. The full version is available in Appendix B.

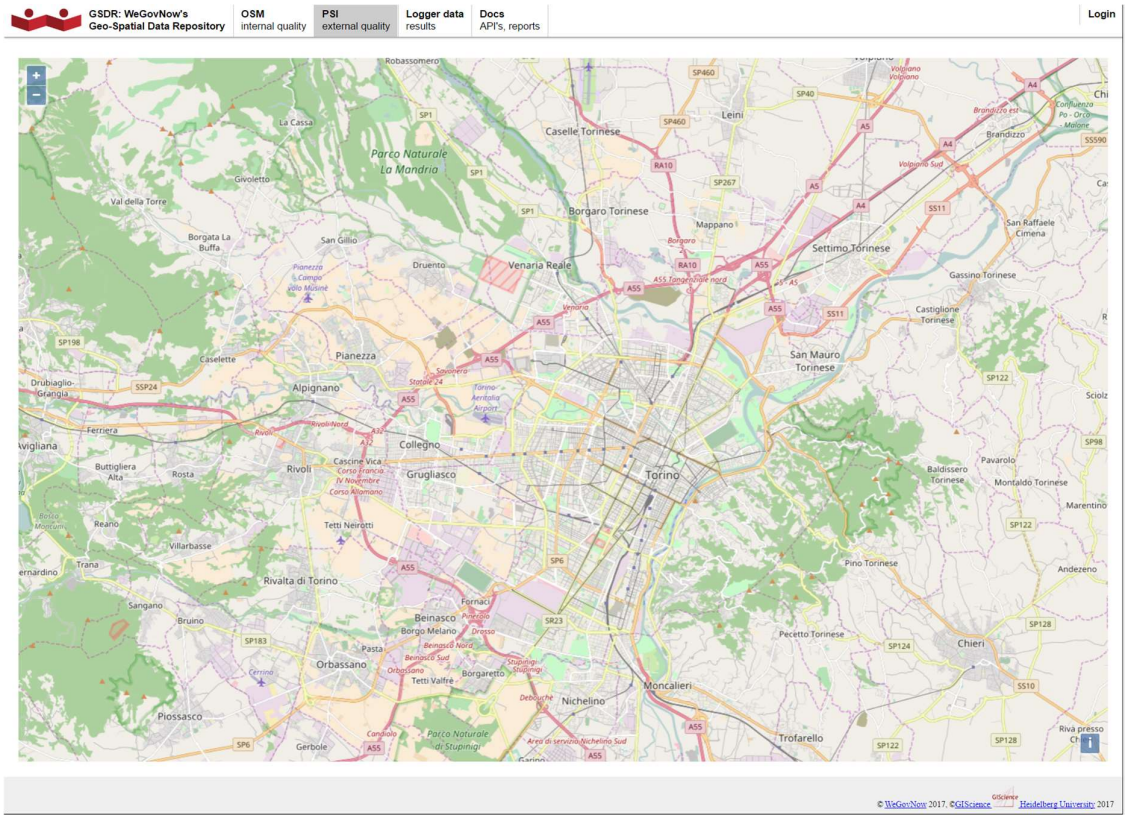
Exhibit 16: Data quality methods relations in UML



It was decided to implement a **web service** for data quality assessment and improvement. Geo-Spatial Data Repository (GSDR) will play this role. Currently, the developing version is accessible by the following link: https://gsdr.gq (see Exhibit 17). The service is being developed for WeGovNow project needs and will provide the following functionality:

1. Intrinsic data quality assessment,
2. Comparable data quality assessment,
3. Presenting data quality assessment results in interactive maps and auto-generated report,
4. Public APIs for quick data check and retrieval.

Exhibit 17: WeGovNow GSDR data quality assessment (available also as web service)



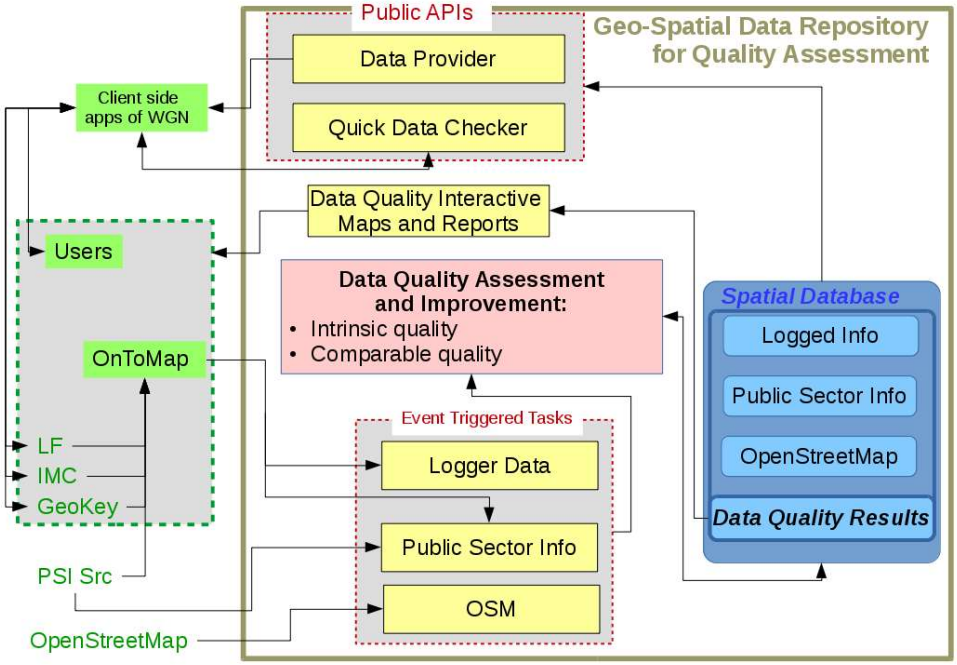
Ground truth reference datasets are not available in frame of the project. Despite this, data quality could be evaluated and improved by the methods mentioned above. Intrinsic quality assessment allows defining data quality of OSM data (e.g., road network completeness). Additionally, the service is designed for detecting errors in PSI and WGN apps' data. These errors will be reflected in interactive maps and auto-generated reports. Quality and imperfections of PSI and WGN apps' data will be defined by a comparison with OSM data with known quality calculated intrinsically.

It should be mentioned that the service is not developed for geo-spatial data only. **Many elements could be utilized for non-spatial data** as well: completeness, logical consistency (excluding topological consistency), thematic accuracy and temporal quality. Since WeGovNow provides mainly location based services though, our attention is concentrated on geo-spatial data quality.

The core component of GSDR is a database. The main harvesting data types are as follows: (1) OpenStreetMap (OSM) covering pilot sites, (2) PSI data provided by OnToMap and (3) OnToMap Logger data. OSM is a very dynamic map, thus OSM data need to be updated periodically. This functionality will be supported by the aforementioned web service.

In Exhibit 18, the conceptual architecture of GSDR is presented. Spatial database's components are marked by blue. GSDR's components responsible for interaction with external applications are depicted on yellow. External components are distinguished with green colour. Black arrows represent data flows.

Exhibit 18: Data quality methods relations in UML



In general, a process of quality evaluation may be described as follows. First, quality of OSM data will be defined intrinsically at the first step. Next, quality of PSI data will be evaluated using shared objects of OSM data with known quality. It should be mentioned that many objects provided by the PSI datasets of OnToMap are presented in OSM. These shared objects will be used for external quality evaluation of the PSI data. OSM is utilized as a reference dataset in this case. It allows us to apply approaches to data quality assessment described in the ISO 19157-2013. Logger data will be assessed using both OSM and PSI datasets.

4.2 Intrinsic data quality assessment

OpenStreetMap project provides different types of data: OSM Weekly Planet XML file, full history dump, tiles logs, Wiki documentation. This allows us to evaluate quality of data intrinsically without reference dataset. Many different approaches could be utilized. Concrete references are presented in Appendix B (methods for quality assessment of non-spatial data are provided in the appendix as well). Intrinsic assessment based on an idea that

data may be assessed without external datasets. E.g., degree of road network completeness may be evaluated as follows. If historical data show that no one modifies major roads when neighbour minor roads are modified actively, one can conclude that major roads are complete. Logical consistency is evaluated intrinsically as well. In contrast, to the previous example historical data is not required for logical errors detection. E.g., topological errors may be found without external knowledge.

One can notice that logger data provided by OnToMap could be processed as OSM full history dump. Thus, the mentioned intrinsic approaches could be applied to logger data as well. Logger and OSM data will be periodically updated by event triggered tasks. OSM data will be periodically updated. If it will be defined that data in pilot sites were significantly changed after the last update, new data will be added to the database. Before import to database, data pass Quality Assessment and Improvement module. It calculates intrinsically data quality related measures. Data and quality assessment results are stored in database. Logger data pass the same procedure, it requires more frequent updates. As shown in Exhibit 15, OnToMap's Logger collects data from WGN apps (e.g., LF, IMC, GeoKey, etc). Then, GSDR obtains and analyse the Logger data. Quality assessment results are stored in the GSDR's spatial database.

4.3 Comparable data quality assessment

Comparable data quality assessment will be utilized for the following tasks:

- Comparison of popular raster tiles,
- Import correctness evaluation of PSI data to OnToMap,
- Comparison of PSI derived from different sources (OpenData and OSM).

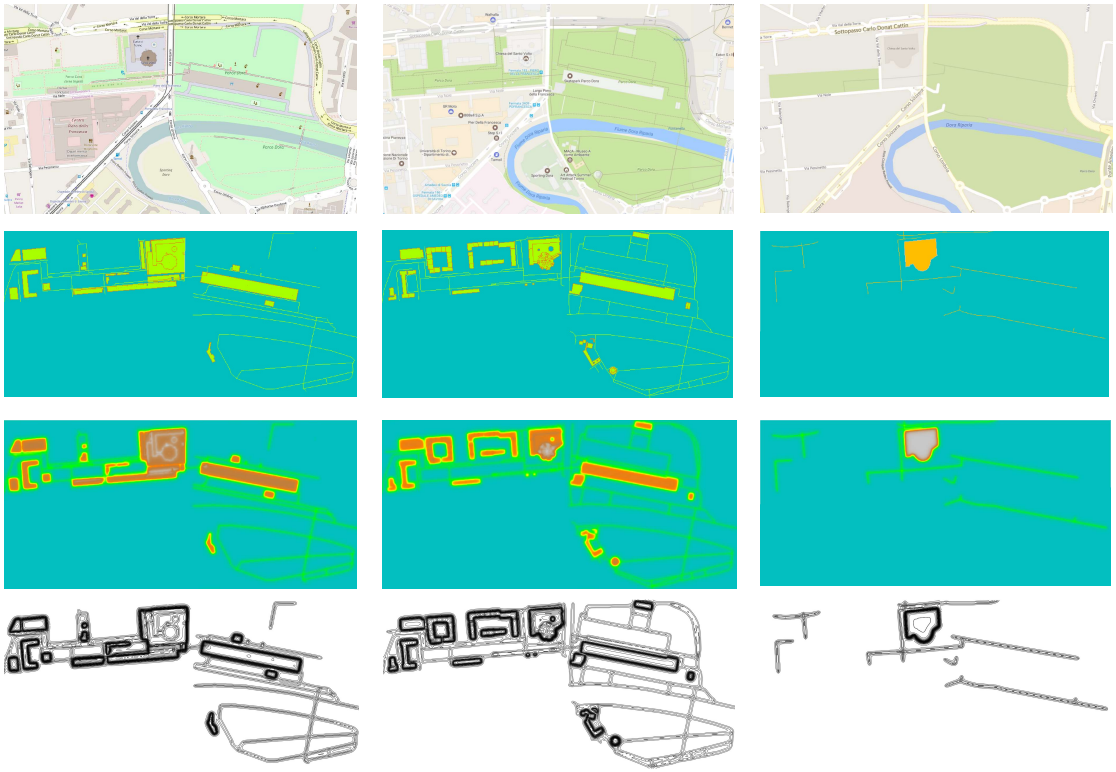
OpenStreetMap, Google Maps, and Bing Maps are three popular raster tile map providers. We can compare raster tiles for small key areas. Public Sector Information (PSI) is not capable for intrinsic assessment. It does not provide history of changes or metadata suitable for such analysis. Thus, a comparable data quality assessment strategy will be applied in this case. Quality of PSI and Logger data will be assessed using OSM (with known quality derived intrinsically) as reference dataset. PSI data contain features provided by OSM as well. We will use this fact to calculate a wide range of data quality established extrinsic measures.

Preliminary practical results demonstrating advantages of comparable data quality assessment are further considered. Blurred features' entropy is utilized for comparable assessment which enables to calculate amount of information provided by a map and compare maps. Results of calculations are described further. The approach is implemented as a part of GSDR's Data Quality Assessment and Improvement module (the pink rectangle in Exhibit 18).

Parco Dora is a pilot area related to several scenarios. Thus, this area should be considered in detail. Using blurred features' entropy, we can calculate raster tile comparable

completeness provided by popular map services with OSM map for Dara Park's area. OpenStreetMap was compared with Google and Bing maps. One can easily conclude, that Bing maps provides less information related to Dora Park than OSM and Google map. At the same time, differences between OSM and Google Maps are not clear. The process of manipulation with raster tiles for an evaluation of the entropy is presented in Exhibit 19.

Exhibit 19: Calculation of the blurred features' entropy of raster tiles (Dora Park, Turin)



In Exhibit 19, OSM, Google and Bing Maps are analysed (from left to right). Upper row presents original raster tiles, row 2 – recognized geometric features, row 3 – blurred map and lower row – calculated contour lines. Length of derived contour lines is amount of information provided by raster tiles. Results are presented in Exhibit 20.

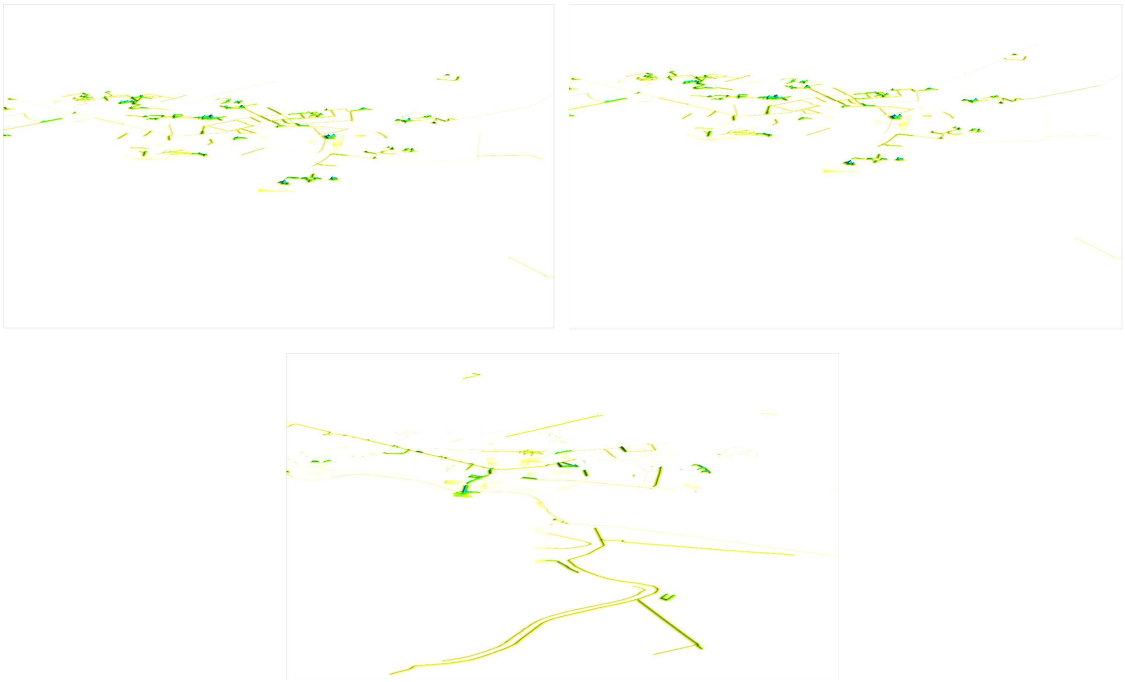
Exhibit 20: Amount of information provided by the raster tiles (Dora Park, Turin).

Parameter	OSM	Google	Bing
Volume of the aggregated raster map	**31244 (100%)**	30305 (96.99%)	6383 (20.43%)
Volume of the blurred raster map	**31281.73 (100%)**	30334.43 (96.97%)	6382.18 (20.40%)
Length of the contour lines	**150371.71 (100%)**	146617.65 (97.50%)	39825.49 (26.48%)

Exhibit 20 shows that OSM and Google provide almost same amount (quantity) of information (OSM slightly more). Difference of the main indicators (length of contour line) is about 2.5%. Bings Maps provides much less information (about quarter of OSM data).

Blurred features' entropy is useful for assessment of PSI data provided by OnToMap. Using this approach, correctness of PSI data import to OnToMap could be evaluated. The entropy was calculated for PSI original data files and data provided by OnToMap. In the demonstration San Dona Di Piave bike lanes are used. Additionally, bike lanes were extracted from OSM and compared with PSI bike lanes provided by OnToMap. In Exhibit 21, 3D blurred models of original PSI data files, data provided by OnToMap and OSM are shown.

Exhibit 21: 3D blurred models and the result table (bike lanes, San Dona Di Piave).



Parameter	OSM	OnToMap	Original Shape File
Volume of the blurred raster map	309803.62 (95.84%)	320015.98 (99.00%)	**323248.39 (100%)**
Length of the contour lines	**221561.90 (100%)**	164987.25 (74.66%)	163224.57 (73.67%)

In Exhibit 21, 3D models are presented. Upper row left is a model reflecting data provided by the original ESRI-shape file, right is OnToMap data, lower is OSM bike lanes. Volume of blurred map allows us to evaluate correctness of the data import. In our case, we can evaluate correctness of shape file import to OnToMap. No significant changes occurred (1%

according to Exhibit 21) during the import according to this parameter, thus we can conclude that import was implemented correctly.

Slight differences could be observed by the following animation: https://gsdr.gq/media/data/onm_shp.gif.

Amount of information could be evaluated by lengths of the contour lines; it reflects quantity of information provided by a data layer. We see that OSM provides +25% of information in comparison to OnToMap data set. OnToMap slightly increases amount of information provided by the original shape file. Probably, it occurred because OnToMap service added unique index to features, when in original shape file index is not unique for few features.

One can mention that blurred features' entropy is powerful measure. It will be utilized by GSDR web service as well as others relevant approached reflected by Appendix B.

4.4 Data quality assessment module

Data Quality Interactive Maps and Reports (yellow rectangle in Exhibit 15) is GSDR's module allows to presenting data quality assessment results in interactive maps and also allows autogenerated reports. It is designed to report data quality results and show data errors. The module can be used by WeGovNow components to improve data and algorithms. Any user concerning data quality of WeGovNow will be able to obtain results using this module.

Relevant concepts and advantages of OSMatrix and iOSMAnalizer will be taken into account for presenting results in interactive maps and auto-generated reports. OSMatrix[10] is a tool for visualising mapping progress/quality on various metrics. Developed by the Geography department of the University of Heidelberg, OSMatrix provides an overlay of hexagonal cells visualizing a range of metrics for the data in each cell. iOSMAnalyzer[11] is a tool for intrinsic OpenStreetMap data quality analyses. This command line tool generates a PDF document containing statistics, maps and diagrams which can be used for assessing the quality of a selected OSM area. As an input only an OSM-Full-History-Dump is needed.

Public APIs for quick data check and retrieval.

Two APIs will be provided by GSDR:

- Data Provider,
- Quick Data Checker.

Both APIs are designed to improve and evaluate users' input. Data Provider API allows us to evaluate data of client side applications. For instance, user reports a traffic light problem. Client side application downloads a small piece of OSM data covering the current map

[10] http://koenigstuhl.geog.uni-heidelberg.de/osmatrix/
[11] https://github.com/zehpunktbarron/iOSMAnalyzer

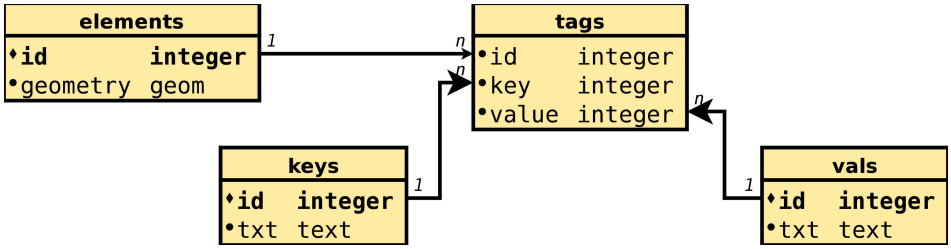
extent. It allows the developer to implement a snapping functionality. User's input may be snapped to the nearest road intersection using a defined threshold.

Quick Data Checker allows us to evaluate users' input directly on the server and immediately return a result. For instance, user filled a form with a word "Schol". Client side application sends a request to Quick Data Checker. It detects that in the considering context there is a term in Logger data with Levenshtein distance equals 1 (it identifies possible error in user's input). The term is "School". The client side application may recommend changing text in a form field.

Many relevant spatial and non-spatial data quality indicators considered in Appendix B will be utilized by the public APIs.

A universal model of spatial database was developed. It allows us to store any type of spatial database. The model is presented in Exhibit 22.

Exhibit 22: GSDR's spatial database model.



One can mention that attribute information is presented as key-value relations. This model was successfully tested for storing and processing data imported from GeoJSON, ESRI-Shape, GML, OSM XML data files.

As was shown, the developing data quality assessment web service is designed to evaluate quality of data provided by WeGovNow platform. OSM, PSI and Logger data will be assessed by event triggered tasks. Fast and dynamic assessment is provided by public APIs. It enables WeGovNow developers to evaluate users' contribution. GSDR is developed with respect to relevant ISO standards (e.g., ISO 19157:2013 and ISO families 8000 and 9000).