**PlanetData**

**Network of Excellence**

**FP7 – 257641**

# D19.1 Call 2: Linked Map report on VGI data quality factors

## Coordinator: Francisco J Lopez-Pellicer (UNIZAR)
## With contributions from: Jesús Barrera (GEOSLAB)

**1st Quality reviewer: Giorgio Flouris (FORTH)**
**2nd Quality reviewer: Max Schmachtenberg (UMA)**

| | |
|---|---|
| Deliverable nature: | Report (R) |
| Dissemination level:<br>(Confidentiality) | Restricted to group (RE) |
| Contractual delivery date: | M44 |
| Actual delivery date: | M45 |
| Version: | 1.0 |
| Total number of pages: | 21 |
| Keywords: | VGI, databases, quality assessment |

## *Abstract*

This deliverable reviews how to assess the data quality of VGI databases by identifying quality dimensions and quality metrics. Quantitative quality concepts (completeness, resolution, logical consistency, positional accuracy, temporal accuracy, temporal quality, thematic accuracy and semantic consistency), non-quantitative quality concepts (lineage, purpose, usage and constraints) and quality concepts unique to VGI data (believability, compliance and convergence) are analysed. A few are selected as feasible to be tested with the Linked Map Platform.

# Executive summary

The objective of this deliverable is to review and to identify the most promising factors to improve the quality of existing VGI databases with the support of authoritative datasets (e.g. official national maps), and amending existing authoritative datasets with VGI data.

To do so, this deliverable presents an instantiation focused on VGI data of a conceptual model for data quality assessment developed in Work Package 2. The instantiation analyses several dimensions of data quality assessment for VGI datasets found in the literature including quantitative dimensions shared with GI (completeness, resolution, logical consistency, positional accuracy, temporal accuracy, temporal quality, thematic accuracy and semantic consistency), non-quantitative dimensions shared with GI (lineage, purpose, usage and constraints) and dimensions exclusive for VGI (believability, compliance and convergence).

Also, this deliverable identifies which of those quality dimensions can be assessed with the help of the Linked Map Platform that is being developed in Work Package 18 and defines a set of quality assessment metrics for these dimensions. These assessment metrics will be computed in the experiments that will be done in the Linked Map Platform.

# Document Information

| | | | |
|---|---|---|---|
| **IST Project Number** | FP7 - 257641 | **Acronym** | PlanetData |
| **Full Title** | PlanetData | | |
| **Project URL** | http://www.planet-data.eu/ | | |
| **Document URL** | http://wiki.planet-data.eu/web/D19.1 | | |
| **EU Project Officer** | Leonhard Maqua | | |

| | | | | |
|---|---|---|---|---|
| **Deliverable** | **Number** | D19.1 | **Title** | Call 2: Linked Map report on VGI data quality factors |
| **Work Package** | **Number** | WP19 | **Title** | Call 2: Linked Map Quality & Crowdsourcing experiments |

| | | | | |
|---|---|---|---|---|
| **Date of Delivery** | **Contractual** | M44 | **Actual** | M45 |
| **Status** | version 1.0 | | final ☒ | |
| **Nature** | prototype □ report ☒ demonstrator □ other □ | | | |
| **Dissemination level** | public □ restricted to group ☒ restricted to programme □ consortium □ | | | |

| | | | | |
|---|---|---|---|---|
| **Authors (Partner)** | Francisco J Lopez-Pellicer (UNIZAR), Jesús Barrera (GEOSLAB) | | | |
| **Responsible Author** | **Name** | Francisco J Lopez-Pellicer | **E-mail** | fjlopez@unizar.es |
| | **Partner** | UNIZAR | **Phone** | +34 876555552 |

| | |
|---|---|
| **Abstract**<br>**(for dissemination)** | This deliverable reviews how to assess the data quality of VGI databases by identifying quality dimensions and quality metrics. Quantitative quality concepts (completeness, resolution, logical consistency, positional accuracy, temporal accuracy, temporal quality, thematic accuracy and semantic consistency), non-quantitative quality concepts (lineage, purpose, usage and constraints) and quality concepts unique to VGI data (believability, compliance and convergence) are analysed. A few are selected as feasible to be tested with the Linked Map Platform. |
| **Keywords** | VGI, databases, quality assessment |

| | | | |
|---|---|---|---|
| **Version Log** | | | |
| **Issue Date** | **Rev. No.** | **Author** | **Change** |
| 2014/05/25 | 0.1 | Francisco J. Lopez-Pellicer | Document instantiation |
| 2014/06/12 | 0.5 | Francisco J. Lopez-Pellicer | Document submitted to QA |
| 2014/06/25 | 1.0 | Francisco J. Lopez-Pellicer | Document submitted to AL |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

# Table of Contents

# Abbreviations

| | |
|---|---|
| GI | Geographic Information |
| OSM | OpenStreetMap |
| VGI | Volunteer Geographic Information |

# List of Tables

# 1 Introduction

Volunteered geographic information (VGI) [1], also known as neogeography [2], user-generated geographic content [3] and crowdsourced geospatial data [4], is a term that describes a scenario in which everyone can contribute to the production of geographic information. Smartphones with GPS, open-source GIS, and Web sites such as GoogleMaps has reduced the entry cost for starting the production of geographic information by laymen. Since 2005, VGI projects, in particular collaborative mapping projects such as OpenStreetMap (OSM) [5], have attracted significant attention among researchers. For instance, this deliverable belongs to the Linked Map project, a subproject of the PlanetData project. The Linked Map project researches the integration of VGI databases, in particular the OSM database, with authoritative geographic datasets with the support of Linked Data best practices.

Authoritative geographic datasets (e.g. a national map) convey a sense of quality and rigour due to their professional origin. Quality is a major concern in the production of authoritative geographic datasets. It is the result of the application of standards and rigorous procedures in the geographic production process. However, quality in volunteer geographic datasets has a different origin. Quality relies on the collective intelligence argument. This argument suggests that a large number of people editing a database are likely to produce by consensus data with higher quality than data created by a single editor. This argument requires a sizeable community of users to become valid. For example, the OSM project has over 1 659 000 contributors. However, the size of its geographic contents is outstanding. The OSM database has over 4 billion GPS points, 2 387 million nodes, 238 million ways and 2 million relations [6]. Unless data is of high quality in the moment of its insertion into the database, there is a reasonable doubt that the quality of the OSM database relies on intensive and extensive user reviews.

The objective of this deliverable is to review and to identify the most promising factors to improve the quality of existing VGI databases with the support of authoritative datasets, and amending existing authoritative datasets with VGI data. To do so, this deliverable reviews how to assess the data quality of VGI databases by identifying factors and quality metrics. Also, this deliverable identifies which of those quality indicators could be used in experiments that will be conducted using the Linked Map Platform, which is described in [7], that is being developed in the Work Package 18.

This deliverable is structured as follows. Section 2 discusses different dimension of data quality assessment for VGI datasets found in the literature. Next, section 3 analyses which quality dimensions can be assessed and how with the Linked Map Platform during the planned experiments. This deliverable concludes summarizing the findings.

# 2 Quality Assessment of VGI datasets

The deliverable D2.1 [8] describes a conceptual model for assessing data quality that is based on the idea of quality as *fitness for use* [9]. The subsection 2.1 presents the core elements of such model: *dimensions*, *assessment metrics*, *data quality indicators*, and *scoring functions*. The remaining subsections discuss different aspects of data quality assessment for VGI datasets found in the literature. The subsections are grouped as follows:

- *Quantitative dimensions shared with GI* reviews the dimensions completeness, resolution, logical consistency, positional accuracy, temporal accuracy, temporal quality, thematic accuracy and semantic consistency.
- *Non-quantitative dimensions shared with GI* reviews the dimensions lineage, purpose, usage and constraints.
- *Dimensions exclusive for VGI* reviews the dimensions believability, compliance and convergence.

## 2.1 Conceptual Model for Data Quality

In this subsection, we summarize the core elements of the conceptual model for data quality introduced by the deliverable D2.1. This conceptual model is based on the idea of quality as *fitness for use*. The *fitness for use* of a dataset depends on several **dimensions** that may vary from an application domain to other. For example, the deliverable D2.1 discusses instantiations of this conceptual model for Linked Data and Data Streams and identifies quality dimensions such as *accessibility* (ability to get access to data), *openness* (freedom to use, reuse and distribute), and *completeness* (ideal size for the task). Below, we discuss an instantiation of the model for the case of VGI data based in the literature on GI and VGI data quality. *Completeness* is one of the quality dimensions of such model. However, *openness* is only an aspect of a wider dimension named *constraints*, and *accessibility* is out of the scope of the instantiation.

An **assessment metric** is a procedure for measuring a data quality dimension. Each assessment metric relies on a set of data quality indicators and calculates an assessment score from these indicators using a scoring function. For example, an assessment metric for *completeness* is a metric that reports if the dataset contains all the required attributes for a task.

A **data quality indicator** is an aspect of a data item that may give an indication to the user of the suitability of the data for some intended task. Classes, instances, properties and values are examples of data quality indicators that may be considered for the *completeness* dimension.

A **scoring function** is an assessment based on a data quality indicator that ranges from simple comparisons to complex statistical functions. For example, a scoring function that computes a proportion based on two data quality indicators may be used for calculating an assessment score on *completeness* using as input instances with and without a given property.

## 2.2 Quantitative dimensions shared with GI

Table 1 presents an inclusive list of quantitative quality dimensions for GI found in the literature [10]-[12]. Next, we analyse each dimension in the context of the evaluation of the quality of VGI databases.

**Table 1 – Quantitative dimensions shared with GI**

| **Dimension** | **Kresse** [10] | **Van Oort** [11] | **Veregin** [12] | **Definition** |
|---|---|---|---|---|
| Completeness | ✔ | ✔ | ✔ | A measure of the lack of data and excess data in the database |
| Resolution | | | ✔ | A measure of the amount of detail that can be discerned |
| Logical consistency | ✔ | ✔ | ✔ | A measure of the absence of logical contradictions in a database |

| Dimension | Kresse [10] | Van Oort [11] | Veregin [12] | Definition |
|---|---|---|---|---|
| Positional accuracy | ✔ | ✔ | ✔ | A measure of the deviation of the points in database from the actual location in the real world |
| Temporal accuracy | ✔ |  | ✔ | A measure of the agreement between encoded and actual temporal coordinates. |
| Temporal quality |  | ✔ | ✔ | A measure of the degree a database is up to date in relation to real-world changes |
| Thematic accuracy | ✔ |  | ✔ | A measure of the correctness of values and assignment to feature class |
| Semantic consistency |  | ✔ |  | A measure of the matching between the meaning of database objects and the meaning in the real word objects. |

### 2.2.1 Completeness and resolution

**Completeness** describes the comprehensiveness of a database. That is, the relationship between the objects in the database and the real objects that are expected to be represented in the database.

**Resolution** describes the amount of detail of a database. Resolution is limited by the precision of the methods used to establish position and affects the suitability of the database for a specific use. Some usages require data generalisation, which involves merging, smoothing and elimination of unnecessary details. Generalisation reduces data resolution.

**Analysis.** According to Veregin [12] the following metrics can assess the quality of the completeness of a GI dataset:

- *Data completeness*. Measurable error of omission or excess observed in the database with respect to the expected content.
- *Model completeness*. Measurable error of omission or excess observed in the database schema with respect to the expected universe of discourse.
- *Attribute completeness*. Measurable error of omission or excess observed in encoding the relevant attributes of the geographic features stored in the database.
- *Value completeness*. Measurable error of omission or excess observed in the degree to which values are present for all attributes of the geographic features stored in the database.

An assessment of the spatial resolution of a database should know in advance the specification of the smallest geographic features that can be represented in the database, and then, measure the features under the threshold.

The absolute completeness of a database can only be evaluated against a specification of such database, which includes details about its resolution. However, few VGI initiatives define clearly which is the final expected content or enforce adherence to strict specifications. Absolute completeness cannot be assessed without such specifications.

It is possible to evaluate the relative completeness of a VGI database against another database. For example, Haklay [13] compared the relative completeness of the urban area of London in the OSM database against a database produced by Ordnance Survey. However, its approach was a visual inspection of a dataset against other dataset. Girres and Touya [14] compared the relative completeness of some areas in France of the OSM database against a database produced by the Institut Géographique National of France. In that work, the assessment metric is the completeness ratio of a class of objects. The data quality indicators are the number of instances and the total length/area in each database of instances that can be classified as belonging to the same class. The scoring function was the ratio between the value in the OSM database for a class and the value in the other database for an equivalent class.

### 2.2.2 Logical consistency

Checks on **logical consistency** are a normal part of the production of geospatial data (e.g. checking if polygons are closed or addresses have a valid address pattern). Logical consistency measures the degree of adherence to the logical rules specified for a dataset. That is, logical consistency deals with the logical correctness of numerical values, text-based values and identifiers associated with a property of a point in space and time. Checking on the adherence to the rules can be automatized. However, there are rules that could require human verification.

**Analysis.** In the context of VGI datasets, logical consistency addresses the reliability of the topological and logical relationships encoded in the dataset against the mandatory schema agreed by the VGI community. Logical inconsistencies have impact on the visualizations of the database, and thus, it is a major concern on VGI projects [15]. It is widely acknowledged that there are VGI contributors reluctant to work according to specifications [14], [16], [17]. For example, Girres and Touya [12] analysed different cases of logical inconsistencies in the OMS dataset: roads not connected at an intersection, multiple overlapping lakes, and lines that must overlap do not overlap. Some of these errors are solved in the production of professional datasets with the help of auto-correction tools.

Assessment metrics can be based on best practices related to the fitness for use in an application scenario. For example, if a best practice for vehicle navigation applications says that a good network finishes each line at every intersection [18], then a possible data quality indicator is the amount of lines not connected at an intersection. A possible assessment metric is the average of such indicator.

Measuring the commitment of a VGI community to a shared schema, which is discussed in Section 2.4.2, is an alternative to this indicator.

### 2.2.3 Positional accuracy

**Positional accuracy** is related to the accuracy of the relative and absolute position of points in space in a database. Relative positional accuracy is the difference of the distance between two points in a database and the true distance between these points in an overall reference system. Absolute positional accuracy is the difference of the distance between a point in a database and its true position in an overall reference system. Lack of positional accuracy is acknowledged as a well-known source of errors in fields such as public health research [19] or transportation [20]. Measuring deviations is hard and requires travelling to the location and verifying coordinates using a GPS. However noise and bias can modify position estimates provided by the GPS. An alternative is the use of computer-based methods [21].

**Analysis.** The following are assessment metrics of the positional accuracy of a dataset:

- *Overall deviation*. Deviations of sampled points from a source of ground truth.
- *Thematic deviation*. Deviations of sampled points belonging to a theme from a source of ground truth.
- *Area deviation*. Deviations of sampled points belonging to an area from a source of ground truth.
- *Positioning method*. Identification and description of the method used to establish the location of an object.

The methodology for evaluating positional accuracy of the database can compute as assessment metric the root-mean square error of the deviations of sampled points or use the buffer-zone method proposed by Goodchild and Hunter [22] when the database contains linear features.

Understanding the absolute and relative deviation of VGI digitized points is resource-intensive and, therefore, literature often excludes explicitly positional accuracy from the data quality analysis [23]. The work of Haklay [13] constitutes an exception. Haklay proposed a comparison framework that evaluates relative positional accuracy. This framework is based on the selection of a database deemed more accurate as a consequence of using a more accurate positioning method. Rice and Brandon [24] claimed that the most common *positioning method* in VGI is digitizing points from maps followed by georeferencing locations by place name. If we assume this claim as true, the OSM database, for example, cannot be more accurate in average and ideal conditions than the resolution of the aerial imagery publicly available in their extent (e.g. 20 m). Thus, the comparison database for the OSM database must have a similar content and have the same ideal positional accuracy. Works such as Haklay [13] show that an assessment based on the relative positional accuracy is computationally feasible and provides information on the positional quality of VGI databases.

### 2.2.4 Temporal accuracy and temporal quality

Given a location, the dimension **temporal accuracy** is related to the accuracy of the relative and absolute position of points in time in a dataset, and the values associated with a point in space. That is, locations have an attribute that defines its temporal validity (e.g. "*this bridge existed between 1902 and 1944*").

A less stringent quality dimension related to time is **temporal quality**. Temporal quality is a measure of the validity of changes in a database in relation to real-world changes. A consultancy report that evaluated quality assurance practice in Ordnance Survey [25] provides a good example of a temporal quality indicator used in production: "*99.6% significant real-world features are represented in the database within six months of completion*".

**Analysis.** Measuring the temporal quality requires sampling the VGI dataset and sending to the ground people in order to detect differences between sampled points and real word points. This audit requires the adoption of new social agreements in VGI communities. For example, in the case of OSM, auditing temporal accuracy would require turning OSM Mapping parties into OSM Auditing parties.

Temporal quality may be analysed indirectly in the full history file of a VGI database as Arsanjani et al [15] mention. For example, Neis et al [23] investigated an aspect of the temporal quality of the OSM dataset in Germany between 2007 and 2011: when the data was introduced into the database. Similar research was performed by Girres and Touya [14] but restricted to a three-month period[1].

Nevertheless, knowing when the data was introduced in a VGI database does not provide reliable information on how long it takes a significant change in the real world to be added to the same database. Temporal grounding is required is always required for assessing temporal accuracy.

### 2.2.5 Thematic accuracy and semantic consistency

**Thematic accuracy** deals with the correctness of numerical or text-based values associated with an attribute.

**Semantic consistency** deals with the semantic correctness of the values of a property in a point in space in time. Both concepts are similar. In the geographic domain, thematic and semantic correctness requires a shared agreement on Geosemantics. The agreement defines why, what and when to use objective measurements (e.g. height), subjective judgments (e.g. building type), and identifiers (e.g. placenames) for codifying information about physical features on the ground.

**Analysis.** There are examples of such agreements in VGI communities. For example, the huge OSM community negotiates through the OSM wiki[2] agreements for codifying how OSM represents physical features on the ground. That is, a language formed by tags attached to basic data structures that can be understood by most OSM users and that the OSM community should use when creating data. The OSM language is an example of a social agreement on how to use certain terms related to geographic information. As Kuhn [26] points out, social agreements on geographic information are agreements on geosemantics that combine objective measurements (e.g. height), subjective judgments (e.g. building type), and identifiers of geospatial entities (e.g. placenames). But social agreements are a weak solution for interoperability. In large communities, the participants may have different perspectives or different requirements. As a result, the understanding of the meaning of the terms may vary across the community producing unclear, unreliable and non-homogeneous terms. Moreover, newcomers unaware of the social agreement may add additional vagueness, uncertainty and heterogeneity to the data if they update the data following their own conventions.

Thematic accuracy and semantic consistency assessment carries out sampling the VGI database. Captured objects should be analysed against a reference database deemed as semantically consistent. A possible data quality indicator is that the sampled objects and the homologs in the reference dataset have equivalent semantic classifications (e.g. both are classified as lakes). Using this approach Girres and Touya [14]

[1] Both Girres and Touya [14] and Neis et al [23] classify this part of their research as temporal accuracy wrongly.

[2] http://wiki.openstreetmap.org/wiki/Main_Page

discovered by analysing the OSM dataset that nearly 100% of primary roads but only 49% of secondary roads in the area of study had a semantically correct classification.

## 2.3 Non-quantitative dimensions shared with GI

Table 2 presents an inclusive list of non-quantitative dimensions found in the literature which is used for the compilation of Table 1 [10]-[12]. They differ from the dimensions discussed in section 2.2 because the conversion of data quality indicators to scores by the scoring functions is often subjective due to the nature of the information (e.g. "*assign true if the transformation method is well described*"). Next, we analyse each dimension in the context of the evaluation of the quality of VGI databases.

**Table 2 – Traditional non-quantitative quality concepts**

| Concept | Kresse [10] | Van Oort [11] | Veregin [12] | Definition |
|---|---|---|---|---|
| Lineage | ✔ | ✔ | ✔ | Describes the history of a dataset: source materials, methods of derivation and transformation applied to a database |
| Purpose | ✔ | ✔ | | Describes the rationale for creating a database and contains information about its intended use. |
| Usage | ✔ | ✔ | | Describes the application for which a database has been used |
| Constraints | | ✔ | | Describes the legal and financial constraints to access or particular use of the spatial data |

### 2.3.1 Lineage

The **lineage** of data is its entire processing history [27]. That is, lineage includes information about the origin of the data as well as all processes applied to it. Keeping a complete data linage record is an essential feature in the production of professional geographic information. For example, the widely adopted international standard ISO 19115 [28], a schema for describing geographic information and services, defines a set of metadata that allows the description of the data lineage for a resource in professional production environments. Data lineage is also important for quality control of primary and derived data products produced in service-oriented environments [29] and distributed environments [30].

**Analysis.** The following elements are typical indicators of the quality of the data lineage:

- *Description*. General information related to the processing history of the data.
- *Process description*. Identification and description on the process steps of the data.
- *Date and time*. The date and time when the process was completed.
- *Data source*. Identification and description of the sources used in the development of the data.
- *Responsible*. Identification and means to contact the person or parties that performed the process.
- *Software*. Identification and means to obtain the software that performed the process.

There are two approaches for the development of assessment metrics for lineage. The simple approach is the evaluation of the completeness of the lineage data. For example, Girres and Touya [14] analysed the OSM dataset and evaluate the completeness of the indicators *data source* and *software*. They found that only 27.8% of the resources sampled contained information about the *data source* and only 6.0% contained information about the *software*. The complex approach involves the analysis of the content of lineage information.

### 2.3.2 Purpose, usage and constraints

**Purpose** is the rationale for creating a database and contains information about its intended use. **Usage** is the application for which a database has been used. **Constraints** are the legal and financial constraints to access or particular use of the spatial data. Van Oort [10] proposes grouping these dimensions as a single element with the purpose of assessing the fitness for use of a database. Van Oort's point of view differs from the mainstream approach in GI (e.g. Kresse and Fadaie [11]) that makes a distinction between the intended use (purpose) and the actual use (usage). Note that the mainstream approach does not include legal and financial constraints as part of the quality assessment.

**Analysis.** The community that maintains a VGI database defines its purpose. Its usage may be restricted a license (e.g. only for non commercial use). Often the discussion on usage restrictions is restricted to intellectual property violations but in a broader perspective it should include privacy, defamation and liability [31]. As Goodchild [32] highlighted, who is responsible for damages that result for the use of VGI data? Nevertheless, the relevant information about the constraints is in the metadata of the database.

An assessment metric on purpose can be implemented by checking if the purpose of the VGI database matches is aligned with the requirements of the intended usage. An assessment metric on usage can be implemented by listing projects that are using the VGI database. Finally, an assessment metric on constraints can check if the license of the VGI database has or not a set of features (e.g. attribution, share-alike, commercial use).

## 2.4 Dimensions exclusive for VGI

VGI has specific data quality aspects. This section presents dimensions that could be considered as unique to VGI databases when compared with geographic databases created in professional environments.

### 2.4.1 Believability

**Believability** is the quality of being believable or trustworthy [33]. Authoritative data sources are considered believable, that is, free of false content with the exception of copyright traps, generalizations, exaggerations, political propaganda, misleading symbols, etc. VGI databases are the result of users' assertions and, as a consequence, their content can be considered as less believable. The work of Monmonier [34] about how to lie with maps is a recommended reading for realising potential mischiefs in both professional and volunteer geographic information production.

**Analysis.** Assessing the level of believability of these assertions gives credibility to the VGI database. Data vandalism is a serious harm to the believability of a VGI database. Therefore, a potential assessment metric for believability is a metric that measures the amount of potentially malicious and mischievous content in a VGI database. Examples of actions that can be considered as indicators for such metric are:

- Deleting existing objects randomly.
- Generating fictional objects.
- Adding spam to attributes.

Researchers have recently begun the development of automatic tools for detecting vandalism in VGI databases. For example, Neis et al [35] have implemented a rule-based decision system, named OSMPatrol, able to detect vandalism in the OSM database as fast as possible. Their system was able to detect vandalism committed by new users. However, their system also classifies as vandalism commits made by users with a high reputation and legitimate deletes.

Neis et al identified as a potential improvement to their system the addition of a special tag to the OSM vocabulary to enable OSM users to tag edits as vandalism or non-vandalism and to maintain a white-list for their tool. That is, users continue to be the best whistle-blowers of vandal acts.

It is assumed that vandalism can be detected and corrected by other VGI contributors within a period of time (refer to Section 2.4.3). Thus, measuring the evolution of the vandalism can also assess if the collective intelligence works as expected.

### 2.4.2 Compliance

**Compliance** is the degree of constancy and accuracy which something implements a recommended set of guidelines. In the VGI context, compliance is a measure of the resilience of the users that contribute geographic information to adopt the social agreements that should rule the creation and the maintenance of the VGI database.

**Analysis.** Compliance is a relevant dimension because it reveals the balance in a VGI database between adherence to recommended specifications and freedom of VGI contributors. Examples of such freedom that can be used as indicators are:

- Adding a new object with no commonly used attributes.
- Updating existing objects with no commonly used attributes.
- Selecting a large set of objects and updating attributes.

The balance between adherence and freedom depends on the dynamics of participation. Some studies such as Budhathoki and Nedovic-Budic [36] have detected that only a fraction of VGI participants actually contribute geographic information. Among those contributors, many of them could have contributed only once. Girres and Touya [14], after analysing debates in the OSM contributors community, affirmed that a large amount of VGI contributors are occasional and mostly are interested amateurs that do not feel safe working with complex specifications. In this scenario, it is fair to assume that the content of a VGI database presents a heterogeneous level of compliance with the specification agreed by the VGI community.

A possible compliance metric is a metric that measures the amount of instances that totally ignore the social agreements that rule the creation and the maintenance of the VGI database. Characterising the VGI players that adhere or not to the specifications assesses not only the quality of the VGI database but also provides hints about its future evolution. Works such as Arsanjani et al [15] show that the characterization of VGI contributors into several categories using their edits in the VGI database as data source is possible.

### 2.4.3 Convergence

**Convergence** is the property of approaching a limit. In the context of VGI databases, convergence can be defined as the convergence of the database to a faithful representation of the real world. A convergence assessment may evaluate if the VGI database becomes increasingly a better representation of the real work without formal quality assurance procedures.

**Analysis.** A possible assessment is analysing how the crowd fixes errors inadvertently introduced by a contributor. Software developers face a similar problem: the detection of software bugs in software. There is a mantra in the open source community: given a large enough developer base, almost every problem should be characterized quickly by consensus and its fix should be obvious to someone. This is known as the *Linus's Law* [37]. *Linus's Law* applied to a VGI dataset means that the more contributors are working in an area, the less errors inadvertently introduced by a contributor and not detected by the remaining contributors. A possible assessment metric is a metric that measures the statistical correlation between errors and the number of users in an area can assess the convergence dimension.

Haklay et al [38] tested the application of *Linus' Law* for different aspects of spatial data quality in the OSM dataset in this way. For example, their study concludes that with more than 15 contributors per square kilometre, the positional accuracy is very good (below 6 m). But more relevant is their consideration that *Linus's Law* is an indicator of the intrinsic quality of a VGI dataset without the use of a reference dataset. However, as Goodchild and Li [39] point out this should be restricted to prominent geographic facts. Locations in sparsely populated areas, ephemeral locations, and locations that interest few people are examples of facts that may not attract sufficient users. Vandalism, disputes, abnormal user behaviour and the use of automated bots to edit a VGI database could also invalidate Linus's Law.

There is an inherent risk in the *Linus's Law* applied to software. It is the problem of *unfocused contributions* [40], that is, developers that contribute to many places without focusing in an area. For example, Meneely and Williams [40] had found that source code files changed by clusters of developers with different focus of interest is more likely to present bugs than source code files changed by a single cluster, and source code files changed by many developers that have made many changes to other files are likely to contain vulnerabilities.

In the context of VGI databases, measuring how a community is reviewing can assess this risk. A possible assessment metric is a metric that measures the statistical correlation between errors and the number of users' edits in an area. For example, the work of Mashhadi et al [41] analysed an area of the OSM database and found that there is no statistical correlation between error and the number of user edits. This result suggests that the OSM dataset is the result of VGI contributors focused in an area rather than unfocused VGI contributors.

# 3 Quality Assessment in the Linked Map project

This section describes which dimensions of the VGI instantiation of the conceptual model for assessing data quality will be analysed in the experiments that will be done in the Linked Map project.

## 3.1 Context

The Linked Map Platform, which is described in the deliverable D18.1 [7], will present users a National Map data and VGI data as a map, and as machine processable-data, RDF mappings between resources of both datasets, and other users' reviews on such mappings. The RDF mappings and the process for creating them are described in the deliverable D16.3 [42]. Users will be enabled to add their own reviews on the perceived quality of the National Map data, the VGI data and the mappings. Figure 1 presents an overview of the information flow in the Linked Map Platform.

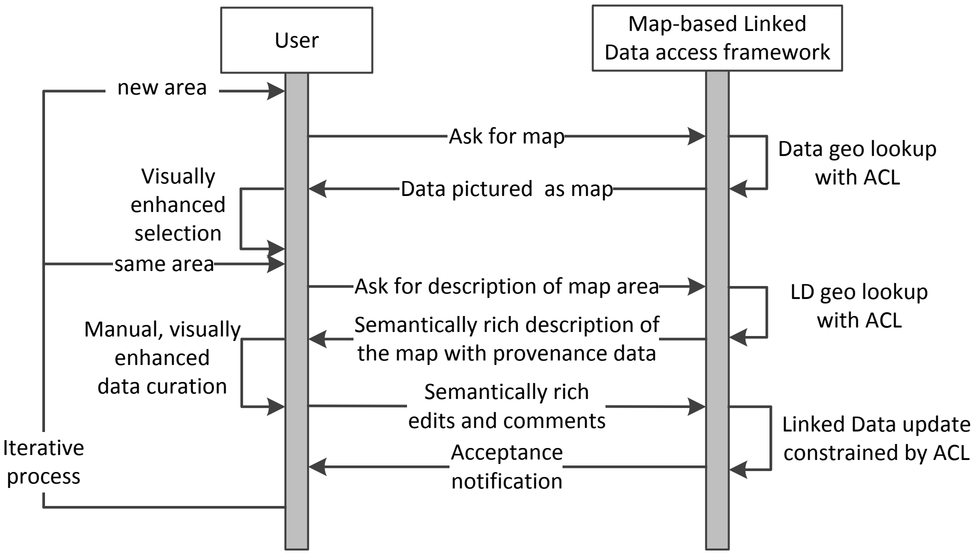
User
Map-based Linked Data access framework
new area
Ask for map
Data geo lookup with ACL
Visually enhanced selection
Data pictured as map
same area
Ask for description of map area
LD geo lookup with ACL
Manual, visually enhanced data curation
Semantically rich description of the map with provenance data
Semantically rich edits and comments
Linked Data update constrained by ACL
Iterative process
Acceptance notification

**Figure 1 – Linked Map Platform information flow**

Users can assert in their reviews if a geographic object is a good representation of a real world object, or if a mapping between two geographic objects from different databases is correct. In addition, they can add a rationale of their reviews. Users can read previous reviews made to an item so they can back or reject previous reviews by adding their own review.

The deliverable D20.2 [43] introduces the experiments that are planned to be done in this platform with the help of volunteers:

- *Link quality assessment*. Participants should assess the quality of RDF mappings between an official geographic dataset and a VGI dataset.
- *Data vandalism detection*. Participants should find wrong RDF links or ignore mischievous reviews that have been introduced knowingly in the aforementioned dataset as part of the experiments.

## 3.2 Dimensions

We believe that the following dimensions can be assessed with the help of the Linked Map Platform:

- *Semantic consistency*. Participants can explicitly assess the semantic consistency of RDF mappings.
- *Believability*. Participants can explicitly assess the degree of trueness of the contents of the National Map and the VGI database. It is also possible to assess if participants' opinions are affected by mischievous reviews about the quality of the National Map, the VGI database and the RDF mappings.
- *Convergence*. Participants' behaviour can be analysed to determine if the quality of their assessments increases when the density of participants increase.

Next we describe the assessment metrics that at the moment of writing will be used for asessing the above quality dimensions.

### 3.2.1 Semantic consistency

**Perceived semantic consistency**. This assessment metric is only for assessing RDF mappings. Participants can explicitly assess the semantic consistency of each RDF mapping by validating or dismissing the mapping. The measure is the proportion of links validated by the users in relation to the overall of links reviewed by users.

### 3.2.2 Believability

**Believability**. This assessment metric is only for assessing the quality of VGI data against a reference database (the National Map). Participants can explicitly express their belief on if a geographic object is a good representation of a real world object. The measure is the proportion geographic objects from a dataset validated by the users in relation to the overall of geographic objects from a dataset reviewed by the users. The metric is computed for both the VGI database and the National Map and the higher value marks the source that is considered more believable by participants.

**Resilience to vandalism**. This assessment metric measures if participants' reviews can be influenced by mischievous reviews. The measure is the proportion of the difference between items that contain a mischievous review on which participants disagree and similar items in which users agree in relation to the overall of items with mischievous reviews. If participants cannot be influenced by mischievous reviews, the value of this metric should be near to 1. Otherwise, if participants are influenced, the value of this metric should be near to -1.

### 3.2.3 Convergence

**Linus' Law**. This assessment metric is the correlation between the number of users assessing in an area and the correctness of their assessments for such area. This metric can be computed after sampling users reviews and analysing their correctness. An alternative is the use of mischievous reviews as indicators and, then, computing three metrics: reviews fixed, reviews ignored, and wrong reviews (participants agree with the review).

# 4 Conclusions

This deliverable proposes an instantiation focused on VGI of a conceptual model for data quality assessment. The instantiation covers the following quality dimensions:

- *Quantitative dimensions shared with GI*: completeness, resolution, logical consistency, positional accuracy, temporal accuracy, temporal quality, thematic accuracy and semantic consistency.
- *Non-quantitative dimensions shared with GI*: lineage, purpose, usage and constraints.
- *Dimensions exclusive for VGI*: believability, compliance and convergence.

The Linked Map Platform, which is described in [7], will present users a National Map data and VGI data as a map, and as machine processable-data, RDF mappings between resources of both datasets, and other users' reviews on such mappings. Users will be enabled to add their own reviews on the presented RDF mappings.

We consider that the quality dimensions below can be assessed with the help of the Linked Map Platform:

- *Semantic consistency*. Participants can explicitly assess the semantic consistency of RDF mappings.
- *Believability*. Participants can explicitly assess the degree of trueness of the contents of the National Map and the VGI database. It is also possible to assess if participants' opinions are affected by mischievous reviews about the quality of the National Map, the VGI database and the RDF mappings.
- *Convergence*. Participants' behaviour can be analysed to determine if the quality of their assessments increases when the density of participants increase.

The following assessment metrics have been also identified and their operationalization have been defined.

- *Perceived semantic consistency* for assessing semantic convergence.
- *Believability* for assessing the homonymous dimension.
- *Resilience to vandalism* for assessing the believability.
- *Linus' Law* for assessing the convergence.

# References

[1] M. F. Goodchild, “Citizens as sensors: the world of volunteered geography,” *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.

[2] A. J. Turner, “Introduction to Neogeography,” pp. 1–54, Dec. 2006.

[3] M. F. Goodchild, “Assertion and authority: the science of user-generated geographic content,” presented at the roceedings of the Colloquium for Andrew U. Franks th Birthday, 2008.

[4] M. T. Rice, F. I. Paez, A. P. Mulhollen, B. M. Shore, and D. R. Caldwell, “Crowdsourced Geospatial Data,” U.S. Army Corps of Engineers, BAA #AA10-4733, Contract #W9132V-11-P-0011, 2012.

[5] *OpenStreetMap*. [Online]. Available: http://www.openstreetmap.org. [Accessed: 09-Jun-2014].

[6] “OpenStreetMap Statistics,” *OpenStreetMap*. [Online]. Available: http://www.openstreetmap.org/stats/data_stats.html. [Accessed: 09-Jun-2014].

[7] J. Barrera and F. J. Lopez-Pellicer, “D18.1 Call 2: Linked Data Platform Alpha version ,” PlanetData, 2014.

[8] P. N. Mendes, C. Bizer, Z. Miklos, J. P. Calbimonte, A. Moraru, and G. Flouris, “D2.1 Conceptual model and best practices for high-quality metadata publishing,” PlanetData, 2012.

[9] J. M. Juran, *Juran's Quality Handbook*, 6 ed. McGraw-Hill Professional, 2010.

[10] W. Kresse and K. Fadaie, *ISO Standards for Geographic Information*. Springer, Berlin, 2004.

[11] P. van Oort, “Spatial data quality,” 2006.

[12] H. Veregin, “Data quality parameters,” in *Geographical Information Systems*, no. 12, P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds. 1999, pp. 177–189.

[13] M. Haklay, “How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets,” *Environment and Planning B*, 2010.

[14] J.-F. Girres and G. Touya, “Quality Assessment of the French OpenStreetMap Dataset,” *Transactions in GIS*, vol. 14, no. 4, pp. 435–459, Oct. 2010.

[15] J. J. Arsanjani, C. Barron, M. Nakillah, and M. Helbich, “Assessing the Quality of OpenStreetMap Contributors together with their Contributions,” presented at the AGILE 2013, 2013.

[16] C. Brando and B. Bucher, presented at the 13th AGILE International Conference on Geographic Information Science, 2010.

[17] M. W. Dobson, “VGI as a Compilation Tool for Navigation Map Databases,” in *Crowdsourcing Geographic Knowledge*, no. 17, D. Sui, S. Elwood, and M. F. Goodchild, Eds. Dordrecht: Springer Netherlands, 2013, pp. 307–327.

[18] M. J. Egenhofer, “What's special about spatial?: database requirements for vehicle navigation in geographic space,” *SIGMOD Record*, vol. 22, no. 2, pp. 398–402, Jun. 1993.

[19] M. R. Bonner, D. Han, J. Nie, P. Rogerson, J. E. Vena, and J. L. Freudenheim, “Positional Accuracy of Geocoded Addresses in Epidemiologic Research,” *Epidemiology*, vol. 14, no. 4, pp. 408–412, Jul. 2003.

[20] M. F. Goodchild, “GIS and transportation: status and challenges,” *Geoinformatica*, 2000.

[21] M. J. Strickland, C. Siffel, B. R. Gardner, A. K. Berzen, and A. Correa, “Quantifying geocode location error using GIS methods,” *Environmental Health*, vol. 6, no. 1, p. 10, Apr. 2007.

[22] M. F. Goodchild and G. J. Hunter, “A simple positional accuracy measure for linear features,” *IJGIS*, vol. 11, no. 3, pp. 299–306, Apr. 1997.

[23] P. Neis, D. Zielstra, and A. Zipf, “The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011,” *Future Internet*, vol. 4, no. 1, pp. 1–21, Mar. 2012.

[24] M. T. Rice and B. M. Shore, “VGI Research Review,” presented at the oint Army Geospatial Center George Mason University Research Meeting, April th,, 2012.

[25] P. Cross, M. M. Haklay, and A. Slingsby, “Agency Performance Monitor Audit - Consultancy Report,” UCL, London, 2005.

[26] W. Kuhn, “Geospatial Semantics: Why, of What, and How?,” in *Journal on Data Semantics III*, vol. 3534, no. 1, S. Spaccapietra and E. Zimányi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 1–24.

[27] A. Woodruff and M. Stonebraker, “Supporting fine-grained data lineage in a database visualization environment,” *ICDE-97*, pp. 91–102, 1997.

[28] ISO/TC 211, “ISO 19115:2003: Geographic information -- Metadata,” International Organization for Standardization, Geneva, Switzerland, 2003.

[29] I. Foster, “Service-Oriented Science: Scaling eScience Impact,” presented at the IEEE/WIC/ACM International Conference on Web Intelligence, 2006, pp. 9–10.

[30] P. Yue, Y. Wei, L. Di, L. He, J. Gong, and L. Zhang, “Sharing geospatial provenance in a service-oriented environment,” *Computers*, 2011.

[31] T. Scassa, “Legal issues with volunteered geographic information,” *The Canadian Geographer / Le Géographe canadien*, vol. 57, no. 1, pp. 1–10, Sep. 2012.

[32] M. F. Goodchild, “Spatial Accuracy 2.0,” presented at the th international symposium on spatial accuracy assessment in natural resources and environmental sciences, 2008.

[33] A. J. Flanagin and M. J. Metzger, “The credibility of volunteered geographic information,” *GeoJournal*, vol. 72, no. 3, pp. 137–148, Jul. 2008.

[34] M. S. Monmonier, *How to Lie with Maps*. University of Chicago Press, 1996.

[35] P. Neis, M. Goetz, and A. Zipf, “Towards Automatic Vandalism Detection in OpenStreetMap,” *IJGI*, vol. 1, no. 3, pp. 315–332, Nov. 2012.

[36] N. R. Nudhathoki and Z. Nedovic-Budic, “How to motivate different players in VGI?,” presented at the GIScience, Zurich,, Zurich, 2010.

[37] E. S. Raymond, *The Cathedral & the Bazaar*. O'Reilly Media, Inc., 2001.

[38] M. M. Haklay, S. Basiouka, V. Antoniou, and A. Ather, “How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus’ Law to Volunteered Geographic Information,” *The Cartographic Journal*, vol. 47, no. 4, pp. 315–322, Nov. 2010.

[39] M. F. Goodchild and L. Li, “Assuring the quality of volunteered geographic information,” *Spatial Statistics*, vol. 1, pp. 110–120, May 2012.

[40] A. Meneely and L. Williams, “Secure open source collaboration,” presented at the the 16th ACM conference, New York, New York, USA, 2009, p. 453.

[41] A. Mashhadi, G. Quattrone, L. Capra, and P. Mooney, “On the accuracy of urban crowd-sourcing for maintaining large-scale geospatial databases,” presented at the the Eighth Annual International Symposium, New York, New York, USA, 2012, p. 1.

[42] F. J. Lopez-Pellicer and J. Barrera, “D16.3 Call 2 : Linked Map authoritative dataset,” PlanetData, 2014.

[43] F. J. Lopez-Pellicer and J. Barrera, “D20 2 Call 2: Linked Map Community Awareness plan,” PlanetData, 2014.