

Master Thesis

Development of an update procedure for authoritative spatial data by the combination with crowdsourced information

submitted by **Nita Maulia**
born on 24.11.1985 in Demak

submitted for the academic degree of
Master of Science (M.Sc.)

Submission on 10/09/2018

Supervisors Prof. Dr. Lars Bernard
TU Dresden
Dr. Rob Lemmens
University of Twente
Dr. Stefan Wiemann
TU Dresden

Task Specification



**TECHNISCHE
UNIVERSITÄT
DRESDEN**

Assignment of a Master thesis topic

Course of study: International Cartography
Student name: Nita Maulia

Topic: Development of an update procedure for authoritative spatial data by the combination with crowdsourced information.

Objectives:

In many applications, authoritative data is still considered the golden standard for geospatial data, especially in terms of data quality and liability. However, authorities and companies are increasingly looking for cost-effective ways to update and quality assure this data on a regular, best real-time, basis. For this purpose, the application of Volunteered Geographic Information (VGI) is often considered most promising.

The thesis shall design, develop and evaluate concepts to update authoritative spatial data by the combination with VGI data. The goal is to show how current update cycles could be enhanced by such a combination, e.g. in terms of data quality, update frequency and automation. For this purpose, a number of existing VGI data sources shall be identified and investigated with respect to their applicability to update authoritative data sources in a reasonable manner. Ultimately, a guideline shall be developed to indicate possible scenarios, recommendations and workflows for stakeholders to update authoritative data with VGI. The feasibility of the developed approach shall be demonstrated by prototypical implementation.

The following shall be submitted:


- Three printed versions of the thesis,
- Three CDs/DVDs with a digital version of the thesis (pdf) and the developed applications (including source code, documentation, etc.)
- A summary of the findings of the thesis for presentation on the Web (approx. 1000-2000 characters)

Supervisors: Prof. Dr. Lars Bernard, TU Dresden
Dr. Rob Lemmens, University of Twente
Dr.-Ing. Stefan Wiemann, TU Dresden

Handed out: 10.04.2018
Submission date: 10.09.2018



Prof. Dr. Michael Soffel
Examination Board



Prof. Dr. Lars Bernard
Academic Supervisor

Statement of Authorship

Herewith I declare that I am the sole author of the thesis named

„Development of an update procedure for authoritative spatial data by the combination with crowdsourced information“

which has been submitted to the study commission of geosciences today.

I have fully referenced the ideas and work of others, whether published or unpublished. Literal or analogous citations are clearly marked as such.

Dresden, 10/09/2018

Signature

Abstract

The demand for up-to-date spatial data is growing very high in this rapidly changing world. Updating the authoritative spatial data is not always associated with the updating of well-established existing spatial data, but it can be extended into the additional collection of new specific-themed spatial data. Unfortunately, authorities who produce 'golden standard' spatial data cannot fill those needs. Considering the merits that VGI can give to fill the gap between demand and supply in authoritative data update, a conceptual framework for authoritative data update using VGI data is proposed. The advantage of the proposed framework is its flexibilities that includes: customizable quality assessment modules; flexible assembly and placement of the quality modules; flexible workflow starting point; and the flexible weighting factor. Based on the framework, three scenarios are made; updating authoritative data with VGI, new theme adoption, and generating thematic information from VGI data. The updating of authoritative data with VGI is implemented to test the feasibility of the framework. The Implementation result shows that the current OSM data cannot surpass the superiority of older RBI data, especially in positional and thematic accuracy. Despite the usability output cannot be used for update, it can be use as the change detection indication that will help authorities to plan the updating processes effectively. Moreover, the implementation proves that automation of the framework is doable and consequently the update frequency could be improved.

Keywords: VGI, Authoritative Data, Quality Assessment, RBI

Acknowledgement

العالمين ربَّ الله الحمد

All praises to Allah the Almighty, for giving me the blessing and strength to finish this thesis.

I would like to express my sincere gratitude to my thesis supervisors: Prof. Dr. Lars Bernard, Dr. Rob Lemmens, and Dr.-Ing. Stefan Wiemann for their time, generous guidance and encouragement.

I would like to extend my deepest appreciation to Juliana Cron and all the staff in Cartography Master Programme for the opportunity and helpful guidance during my study.

I thank my supportive fellow #CartoSquad16 who makes my last 2 years more colourful and meaningful.

Last but not least, I would like to thank my beloved mother, my sister, my brother, and my hunnybear for the encouragements, supports, and continuous pray. Their love is my motivation and energy in completing this study.

Dresden, 10th September 2018

Contents

Task Specification	i
Statement of Authorship.....	ii
Abstract.....	iii
Acknowledgement.....	iv
Contents	v
List of Figures	vii
List of Tables.....	viii
List of Equations	viii
List of Acronyms.....	ix
1 Introduction.....	1
1.1 Research Objectives.....	2
1.2 Research Question	2
1.3 Research Innovation.....	3
1.4 Thesis Contents	3
2 Volunteered Geographic Information.....	4
2.1 VGI and Authoritative Data.....	4
2.2 VGI Data Repositories	5
2.3 VGI Data Categorisation	8
2.3.1 Vector Data.....	12
2.3.2 Textual Data.....	15
2.3.3 Media content	16
3 Spatial Data Quality Assessment	17
3.1 ISO Quality Assessment	17
3.2 VGI Quality Assessment	21
3.3 Quality Assessment Modules.....	23
3.4 VGI Usability	26
4 VGI Data Utilisation Framework and Scenarios.....	31

4.1	Conceptual Design of VGI Data Utilisation.....	31
4.1.1	User	32
4.1.2	Requirements	32
4.1.3	Contributor.....	33
4.1.4	VGI data	33
4.1.5	Data examiner	34
4.1.6	Pre-processing	34
4.1.7	Unqualified data	34
4.1.8	Quality assessment.....	34
4.1.9	Data reference.....	35
4.1.10	Confidence index	36
4.1.11	VGI usability	36
4.2	Framework Scenarios.....	36
4.2.1	Scenario 1: VGI data for updating	37
4.2.2	Scenario 2: New theme adoption	39
4.2.3	Scenario 3 : Generated thematic information from VGI data.....	42
5	Implementation.....	45
5.1	Updating RBI using OSM data	45
5.2	Implementation Result.....	54
5.3	Evaluation and Discussion	64
6	Conclusion and Recommendation.....	68
6.1	Conclusion	68
6.2	Recommendation.....	69
	References.....	71

List of Figures

Figure 1	VGI information flows (Haklay, et al., 2014).....	6
Figure 2	Types of volunteered Geographical Information (Cooper, et al., 2011).....	8
Figure 3	Detail of Categorisation of VGI data (See, et al., 2017)	11
Figure 4	Proposed classification of VGI data type	12
Figure 5	Point Geometry attributes	13
Figure 6	The use of VGI in the European NMAs (Olteanu-Raimond, et al., 2017a).....	27
Figure 7	Proposed framework for updating authoritative spatial data with VGI data	32
Figure 8	Workflow of scenario 1	38
Figure 9	Workflow of scenario 2.....	41
Figure 10	Workflow of scenario 3	44
Figure 11	Pre-processing model.....	46
Figure 12	Nearest object analysis model	50
Figure 13	Proximity model.....	51
Figure 14	Intra-theme logical consistency model.....	52
Figure 15	Inter-theme logical consistency model.....	52
Figure 16	Data updating scenario.....	53
Figure 17	Ratio of CA008020 to other RBI Road classes	55
Figure 18	Bridleway road overlaid in Bing satellite imagery	56
Figure 19	Comparison of RBI and OSM data distribution	57
Figure 20	Subset area of OSM and RBI buffer intersection.....	59
Figure 21	Intersection segments composition.....	60
Figure 22	Non-intersect segments composition.....	60
Figure 23	Non-intersect segments indication: a) additional road segments; b) wrong geometry	61
Figure 24	Intersection of OSM road with RBI building	61
Figure 25	Intersection of OSM road with RBI building overlaid in Bing satellite imagery ...	62
Figure 26	The intersect segments composition.....	62
Figure 27	Dangles identification during topology check.....	63

List of Tables

Table 1	Typology of VGI (Craglia, Ostermann, & Spinsanti, 2012).....	9
Table 2	Categorisation of VGI projects (Bodogna, et al., 2016).....	10
Table 3	Resume of quality assessment elements (Antoniou & Skopeliti, 2017)	19
Table 4	ISO quality elements, their requirements, and issues related to their use with VGI (Fonte, et al., 2017).....	21
Table 5	Summarized of proposed VGI quality assessment.....	22
Table 6	Example of a quality module for the line – vector data type.....	23
Table 7	Example of a quality module for the polygon – vector data type.....	24
Table 8	Example of a quality module for the geo-tagged posts – textual data type.....	26
Table 9	VGI usability summary.....	27
Table 10	Scoring of quality element	29
Table 11	Confidence index assignment.....	30
Table 12	RBI updating period (Perka BIG No.14 Th.2013).....	45
Table 13	Definition of RBI road classification (BIG, 2005).....	47
Table 14	Definition of OSM road classification (Wiki, 2018).....	48
Table 15	Classification conversion of OSM	49
Table 16	Quality element score and weighting factor	49
Table 17	Road length based on the classification	55
Table 18	Attribute completeness of OSM data.....	56
Table 19	Attribute comparison of RBI and OSM.....	56
Table 20	Thematic attribute analysis result.....	58
Table 21	Confusion matrix of OSM Classification.....	58
Table 22	Usability recommendation	64

List of Equations

Equation 1	Confidence Index equation.....	30
------------	--------------------------------	----

List of Acronyms

API	Application Programming Interface
BIG	<i>Badan Informasi Geospasial</i>
CGSC	Citizen-generated (geo)spatial content
DM4VGI	Dynamic Metadata for VGI
GIS	Geographic Information System
GPS	Global Positioning System
ISO	International Organization for Standardization
OSM	OpenStreetMap
POI	Point of Interest
ROI	Region of Interest
RBI	<i>Rupa Bumi Indonesia</i>
VGI	Volunteered Geographic Information
VSI	Volunteered (geo)spatial information

1 Introduction

The authoritative map was referred by Du, et al., (2017) as geospatial data that obtained from the survey and classified using formal quality assurance procedures. Producing the authoritative map needs special requirements related to survey data, tools, unique expertise, and specific time span, therefore it has a high production cost. This makes the authoritative map is referred as „the golden standard“ with respects to production standard. The term of authoritative spatial data is not always imbedded in the National Mapping Agency (NMA). In fact, these maps can also be produced by another government or authorized bodies. However, typically, NMA is entrusted to provide the base map (topographic map) to be used in every authoritative map products. Unfortunately, as the trend of downsizing the operation area increases, the ability to provide up-to-date map does not follow. Since NMA is considered to have the golden standard in mapping, down-sizing the operation will consequence in a higher budget and may take a longer time to complete the national coverage. Therefore, providing an up-to-date map becomes the new challenge for NMA, especially in this rapidly changing world.

The term of updating the authoritative spatial data is not always associated with the updating of well-established existing spatial data, but it can be broadened into the additional collection of new spatial data. In short, it is not just about upgrading and renewing the roads feature according to the changes that occurred, but also extending the specific-themed spatial data, such as migration of the animal, crime spots, and even further emotional mapping. Specific-themed spatial data is very important as the input for decision support especially to solve emerging problem. As the world develops, new problems will keep arising and therefore the solution and strategy to overcome the problems also have to be evolved. New thematic spatial data are keeps developing as a response to the emerging problem that needs to be solved by using a spatial approach. Thus, causing the pressure on spatial data update even greater for the authoritative body.

The expeditious development of Volunteered Geographic Information (VGI) recently has brought the idea of data integration with the authoritative spatial data. The idea is based on the merits that VGI have might possibly answer the challenge that NMA faced nowadays. As listed by Goodchild and Li (2012) the advantages of VGI are: free to access, a vast amount of data, timely, available in various data type, and can provide new data for mapping practices. Moreover, VGI also contains plentiful user-generated information, can overview the real world changes more quickly, and most importantly it has a lower acquisition cost (Du, et al., 2017). With all of the aforementioned merits, VGI can be the potential source for accelerating update cycle, new data collection, or as simple as the change detection indicator for the authoritative spatial data. Though VGI was considered as an inadequate source in the first decade of its development, the acceptance of VGI as a valuable and useful source of information for the government has been growing at all levels (Haklay, et al., 2014). However, despite the abundant research that devoted to the adoption of VGI data into authoritative data has been established, the implementation is still limited in practice (Haklay, et al., 2014).

There are some obstacles that might restrict the VGI to be adopted into the authoritative spatial data scheme. Olteanu-Raimond, et al., (2017a) identified that there are five major obstacles for VGI adoption; data quality and validation, legal issues, nature and motivation of the crowd, sustainability, and employment fears. Nevertheless, despite those barriers, the engagement of authoritative bodies to explore the potency of VGI data is still rising.

The engagement of authoritative bodies to adopt VGI data may vary in terms of data requirements. Some of them might specifically put strict restrictions while other might set more flexible stipulation for VGI data adoption. For example, the well-established NMA which has well-structured SDI and adequate data collection, the adoption of VGI data into their spatial data framework may require rigid specification regarding data quality because they already produced data using "the golden standard". In contrast, for the growing NMA, they might be willing to take lower data quality to optimize the use of VGI data in order to build their national spatial datasets. Furthermore, most NMAs have their own strategy in dealing with the VGI data adoption. However, since the demanding necessity of the up-to-date spatial data from the user arises, the user requirements for VGI adoption play an important role in determining the acceptance level of data quality to meet the user needs.

1.1 Research Objectives

Considering the merits that VGI can give to fill the gap between demand and supply in authoritative spatial data update, the conceptual framework for VGI data adoption into authoritative data need to be formulated. The ultimate goal of this research is to show how the current update cycle could be enhanced by such combination. Thus, this research points out some objectives as follows:

- Design, develop and evaluate concepts to update authoritative spatial data by the combination of VGI data.
- Develop framework scenarios, recommendation, and workflow for stakeholders to update authoritative data using VGI data.
- Show how current update cycles could be enhanced by such combination (data quality, update frequency, and automation)

1.2 Research Question

In order to meet the research objectives, some questions that need to be answered in this research are as follows;

- **What kind of VGI data that can be used?**

VGI data need to be explored and identified to find the potential contribution in the authoritative spatial data integration, especially for updating purpose. Based on the finding, those data need to be classified with the regard to data characteristics. Understanding the characteristics' differences among the data is essential in order to have a proper data handling procedure.

- **How can the VGI data be integrated into authoritative spatial data?**

A conceptual framework shall be designed and developed to integrate VGI with authoritative spatial data. The adoption should consider the input data, user requirements, data quality assessment, and output data as well as the stakeholder involved. Some possible scenario that may occur in the framework will also be designed. As the proof of concept, an implementation of the framework will be simulated.

- **How to conduct a quality assessment for VGI data?**

Considering that VGI data has a unique nature, a suitable accuracy assessment might be needed to be specifically designed based on the VGI characteristic. The quality assessment dimensions should be all explored to find a proper assessment method. For this purpose existing accuracy assessment methods are need to be identified. By doing so, the important aspects of quality assessment can be listed to be further use as the input in the formulating better quality assessment process.

- **What scenarios that might emerge in real-world regarding the proposed framework?**

Many possibilities could happen in the real-world implementation of the proposed framework. Therefore, some scenarios are simulated to unfold the possibility based on the different perspective of the framework (data input, stakeholders, quality assessment elements). Finally, a prototypical implementation from one of the scenario will be conducted to demonstrate the feasibility of the proposed framework in authoritative data updating.

1.3 Research Innovation

The innovation aimed at this research is to design and develop an alternative conceptual framework of an update procedure of authoritative spatial data by mean of VGI data. Data classification based on its type will be formulated to help data handling process. Furthermore, the confidence index of VGI data will be described and formulated so that it can be used as the reference for VGI usability.

1.4 Thesis Contents

Introduction and motivations of this thesis are discussed in chapter one. Meanwhile, the following chapter will discuss VGI data, especially the distinction of the data type. Chapter three will discuss and analyse the accuracy assessment that had been conducted in previous research as well as propose to use the confidence index as the reference in determining the usability of the VGI data. The design and development of a conceptual framework in updating procedure of authoritative spatial data with VGI data will be explored in chapter four, while the implementation of the framework will be discussed in chapter five. Finally, conclusions and recommendation of this research will be given in chapter six.

2 Volunteered Geographic Information

The year 2004 was a breaking point moment for the paradigm which believed that geographic information was only the domain of authoritative bodies. The establishment of OpenStreetMap¹ (OSM) has encouraged citizen to actively participate in producing geographic information data. Moreover, it also triggers another organization and communities to comprehend these interactions to create more specific geographic information data to meet their purpose. Since then, abundant data have been produced by such engagement.

There are many different terms to symbolized the involvement of the citizen to produce geographic information. See, et al., (2016) had listed existing terms that are widely used to define the involvement of citizen in geographical information production, along with definition and attribution. From the literature and media dated from 1990's to 2015, as many as 27 terminologies were compiled. A temporal analysis of the literature and google trend analysis was also conducted to examine the frequently used term (See, et al., 2016). Both analyses show that the term of crowdsourcing was the most significantly used term.

Though both crowdsourcing and VGI are used to describe the citizen involvement to produce geographic information, they have distinct definitions. Crowdsourcing is defined as the type of online participation to accomplish a task voluntarily by carrying out an open call for contribution which always entails mutual benefit (Estellés-Arolas & González-Ladrón-de-Guevara, 2012; See, et al., 2016), while VGI is defined as spatial information that voluntarily collected and made available (user-generated content) for public (Goodchild, 2007; (Coetzee, 2018); Elwood, et al., 2012; See, et al., 2016). In conclusion, crowdsourcing emphasises the method of information collection process, whereas VGI is the product of the process. Despite the fact that many literatures seems to consider both terms as the same (Fast & Rinner, 2014), the term of VGI will be used in this research because it serves better definition for this research.

2.1 VGI and Authoritative Data

In order to enable the integration of VGI data into authoritative spatial data, it needs to be ensured that the VGI data have characteristics of spatial data. Spatial data is considered as the representation of real world phenomena in terms of position, attributes, and interrelations (topology) (Burrough, McDonnell, & Lloyd, 2015). In respect of data structure, spatial data is usually stored with the coordinate information and its topology while in the context of the data model, any data that can be mapped will be considered as spatial data (ESRI, 2018). With this characteristics, most of the VGI data can be categorised as spatial data.

The abundance of VGI data that are available to access come in various forms and types of products. All those VGI data have innate nature which is described and divided by Capineri (2016) into three main elements; the geographical reference, the contents, and the attributes. The geographical reference refers to the location information of the data such as geotag,

¹ <https://www.openstreetmap.org/#map=19/-6.44692/106.84388> (accessed 04.09.2018)

coordinate, and geographic name. The data content may take different forms (i.e. image, text, symbols, maps, check-in, photo, video, drawing, etc) that make it possible to be transformed into information. The attribute describes various degrees of accuracy of content producers (participant as both the user and producer (Coleman, et al., 2009)) as well as the creation date of the content. Concerning issues and challenges, the nature of VGI that need to be paid attention to are the heterogeneity of the data and contributors, spatial bias, lack of specifications, and the dynamic nature in which the data are updated (Fonte, et al., 2017). More importantly, it needs to be kept in mind that the nature of VGI data is subject to specific VGI data type.

Compared to the authoritative spatial data, VGI data might be less accurate, less formally structured, and lack of associated metadata that allows it to be used as framework data (Jackson, et al., 2010; Du, et al., 2016). Framework data is the data collected by authoritative bodies, which also called as authoritative data. Framework data is characterized by seven themes; geodetic control, orthoimage, elevation, transportation, hydrography, governmental units, and the cadastre (Elwood, et al., 2012; See, et al., 2016; Sui & Cinnamon, 2017). Framework data is produced by mapping agency or companies that have specially trained staff, well-established standard, and advance technologies. Therefore, some of those themes are cannot be accomplished by VGI data as it requires a high level of expertise and specific equipment. Even if the VGI data were acquired from the collaboration between citizens and authorities, it still cannot meet all framework data aspects (Elwood, et al., 2012).

Though VGI cannot meet all those framework data characteristics, VGI data does have similar characteristics to be considered as a framework data; geographic reference, contents, and attributes. The combination of these components is powerful enough to be used to aggregate, synthesize and compare the information at different scales and time spans (Capineri, 2016). Despite the limitation of VGI data, considering that it has basic characteristics of spatial data and covering some themes of framework data, it can be integrated to authoritative spatial data. However, the quality assurance for VGI data is necessary for the integration.

2.2 VGI Data Repositories

As defined by Elwood, et al., (2012), the aim of VGI is to voluntarily provide information about the world to become available. The rapid and continuous growth of VGI data nowadays is resulting in a huge number of VGI products. Fortunately, the mass production of VGI data is also balanced with the easiness of data access. As a user generated content, a lot of VGI data are available online and can be publicly accessed. However, it is important to know the VGI data flow to understand the nature of the data before we decide to take advantage of it. By understanding the flow of information one can get the bigger picture of the contributor and user interaction through the information flow. Furthermore, how the data will be handled and utilised can also be determined.

In the context of VGI adoption into authoritative spatial data, Haklay, et al., (2014) summarized that there are different kinds of VGI information flow; public to government, government to

public, public to government to public, and government to public to government (Figure 1). Regarding interaction from the contributor to the user, these flows can be simply categorised as one way and two way information flows. In one way interaction, the information directly flows from the contributor to the user, such as from public to government or from government to the public. In contrast, two way interactions happen both from the contributor to the user, vice versa (e.g. public to government to public and government to public to government). Those interactions are the typical information flow in VGI practices. However, the government to public information flow is rarely implemented (Haklay, et al., 2014). Public to government is the typical information flow that will be used as the example in this research, in which the VGI data that provided by the public will be used to update the government authoritative data.

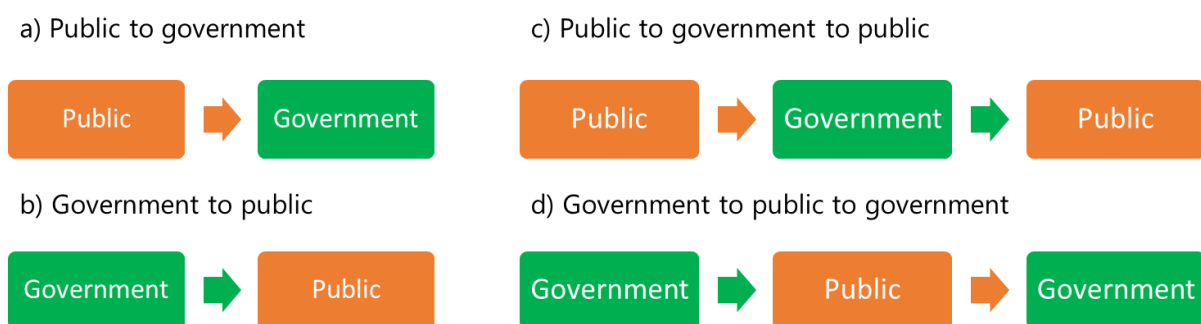


Figure 1 VGI information flows (Haklay, et al., 2014)

By understanding the information flow of VGI, it will be easier to find the VGI data sources. As known from the information flow, there are contributors who made the data available and the users who utilise the data. Generally, the contributor can be identified as an individual, organization, or government authorities. VGI data repositories usually can be traced based on its contributor. For example, the individual contributor tends to use the famous, familiar, and reputable platform to participate, while organization and government tend to use their own webpage for their VGI project. Considering this nature, the data produced from individual contributor usually is more recognisable rather than data produced by organisation or government. This is because usually, VGI data from the organization and government VGI data are specific-themed while data from individual contributors are open for more wide ranged theme and applications. To support this argument, OSM can be a strong evidence on how individually contributed data become more eminent.

It can't be denied that OSM has been widely known and become a leading example of VGI data on the internet (Mooney & Minghini, 2017). Even so, it is not the only source for VGI data. See, et al., (2017) had discussed and listed some possible VGI sources based on their categories. At least 44 sites, whether it was contributed by individual, organisation, or government, were listed as the potential VGI data sources. Furthermore, it also mentioned many famous social media and travelling sites that can be possibly used as VGI data source. Those sites are just small chunks of VGI data sources which might not represent all the possible VGI data sources both locally or worldwide. However, it is essential to keep in mind that those sites might not be

available to access after sometime (See, et al., 2017) due to rapid development and sustainability of the VGI project.

Furthermore, how to extract the data from its sources is also equally important to know the data repositories. Most of the well-established VGI sites provide the information, even a step by step guidance, on how their data can be downloaded. However, in order to collect good quality of VGI data, a generic protocol to download VGI data were proposed by Mooney, et al., (2016). The protocol was specifically designed for collecting vector data in three principal ways; manual vectorisation; field survey; and reuse of existing data sources. Those principles are used as the basic foundation to formulate a sequence data collection protocol that consists of five main steps; initialisation, data collection, self-assessment/quality control, data submission, and feedback to the community (Mooney, et al., 2016). The protocol is furthermore refined by Minghini, et al., (2017) as a generic and flexible protocol that can be used to collect all types of VGI data, both existing and new initiative data product. There are two examples of real-world application described in the refined protocol; updating and collecting new thematic information in a topographic building database and using geo-tagged photographs for Land Use/Land Cover Mapping.

Beside the data collection protocol, the technical process to download and extract the VGI data from its repositories are also important. Application Programming Interface (API) key is often used to bridge the interaction between the VGI data provider and the user to download the data. The API keys allow the provider not to share all their secrecy regarding their source codes, data, and more importantly; their contributor's privacy. Thus, the data that we download sometime is limited in richness. Additionally, some tools and scripts might be needed to acquire the data. The paper by Juhász, et al., (2016) had provided and demonstrated a technical guide to extract and analyse VGI from different platforms. The standard output format of APIs was described as well as software requirement to formulate easy and generic ways to extract dataset. Flickr², Twitter³, Instagram⁴, Foursquare⁵, Wheelmap⁶, and Mapillary⁷ were taken as examples.

The utilisation of the SPAtial-TEmporal-teXTual Suite (Spatext) to fetch the data from social media platforms was demonstrated by Massa and Campagna (2016). Spatext is a Python based extension for the commercial software ESRI ArcMap© that enables the contextual social media data collection, management, geocoding, as well as the spatial, temporal and textual analysis of VGI data directly from GIS interface (Massa & Campagna, 2016). This tool will benefit the user by only using one interface to do the entire task needed on harnessing the VGI data. However, it is important to note that the method or tools that were demonstrated may no longer be sufficient to use after sometimes if not updated, due to the dynamic change of API services.

² <https://www.flickr.com/> (accessed 04.09.2018)

³ <https://twitter.com/?lang=en> (accessed 04.09.2018)

⁴ <https://www.instagram.com/challenge/?hl=en> (accessed 04.09.2018)

⁵ <https://foursquare.com/> (accessed 04.09.2018)

⁶ <https://wheelmap.org/map#/?zoom=14> (accessed 04.09.2018)

⁷ <https://www.mapillary.com/> (accessed 04.09.2018)

2.3 VGI Data Categorisation

Considering the abundance of VGI data, it is necessary to distinguish VGI into different classes in order to make it more understandable and manageable. There were some attempts to categorised VGI data based on different point of views. The taxonomy of VGI data in respect of contributors' motivation was developed by Coleman, et al., (2009). Contributors of geospatial information were divided into five categories (*Neophyte, Interested Amateur, Expert Amateur, Expert Professional, and Expert Authority*) and the manner of contributions then characterized into *Constructive* and *Damaging contribution*. Another point of view was put forward by Cooper, et al., (2011) by using the types of data and determination of specifications dimensions. Type of data was distinct into *Base* and *Point of Interest* (POI) while the specifications determination were divided into *User* and *Custodian*, regarding the of responsibility to determine the specification of data. As shown in Figure 2, the distinction of VGI data is also provided with the example of VGI data repositories.

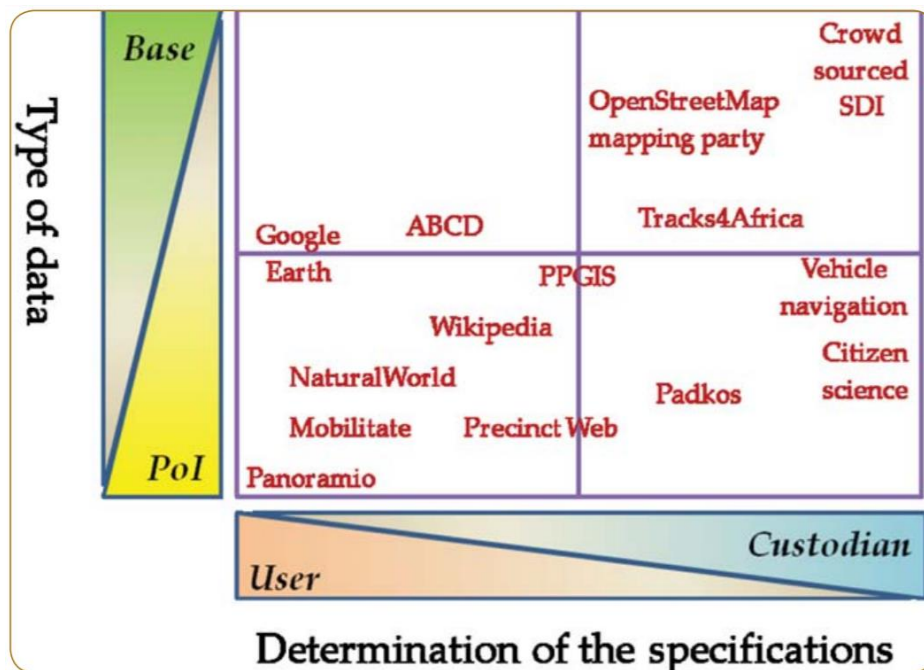


Figure 2 Types of volunteered Geographical Information (Cooper, et al., 2011)

Elwood, et al., (2012) distinguished the VGI based on its purpose and sharing into allocentric and egocentric. Allocentric information doesn't reflect the user's actual location; therefore the user can contribute information about any location. In other words, allocentric information is contributed for the benefit of others rather than personal purposes (Engler, Scassa, & Taylor, 2014). Wikimapia⁸ or OSM can be the exemplification of this type of VGI (Elwood, et al., 2012; Engler, Scassa, & Taylor, 2014). In contrast, egocentric information specifically refers to the

⁸ <http://wikimapia.org/#lang=de&lat=51.050000&lon=13.750000&z=12&m=b> (accessed 04.09.2018)

user's actual location that allowed the producers to share their real-time location or movements (Elwood, et al., 2012; Engler, Scassa, & Taylor, 2014).

Other efforts to develop VGI categorisation based on VGI nature are also proposed by differentiating the active and passive sensing, which correspond to explicitly volunteered and implicitly volunteered information (Craglia, Ostermann, & Spinsanti, 2012). In order to differentiate the VGI, two dimensions were considered: how the information was made available, and the way geographic information was formed. The division resulted in 4 different categories which each categories' description were shown in Table 1.

Table 1 Typology of VGI (Craglia, Ostermann, & Spinsanti, 2012)

	Geographic	
	Explicit	Implicit
Explicitly volunteered	This is 'True' VGI in the strictest sense. Examples include Open Street Map.	Volunteered (geo)spatial information (VSI). Examples would include Wikipedia articles about non-geographic topics, which contain place names.
Implicitly volunteered	Citizen-generated geographic content (CGGC). Examples would include any public Tweet referring to the properties of an identifiable place.	Citizen-generated (geo)spatial content (CGSC) such as a Tweet simply mentioning a place in the context of another (non-geographic) topic.

Bodogna, et al., (2016) developed a categorisation by dividing VGI projects into five categories; 1) scientific field, 2) volunteer's task, 3) way of VGI creation, 4) need for VGI, and 5) characteristics of the volunteer. Each categorisation is break down into sub-categories which represent the possible value of each category. The value of 'Need of VGI' and 'Characteristics of volunteer' categories are sequential while the other categories are not. Moreover, each category is also independent from each other. Expectedly, this categorisation can reflect the quality that influences the factors and characteristics of VGI data quality. The categorisation of VGI project can be seen in Table 2.

Recently, the categorisation of VGI data in respect to its fitness of use (framework or non-framework data) and data contribution (passive or active) has been proposed by See, et al., (2016). Framework data is the data that frequently needed by the user which collected by authoritative bodies. Regarding the contribution aspect, Harvey (2013, in See, et al., 2016) divided the process into active contribution and passive contribution. In the active process, the contributor is aware that the data they produce will be used for a specific purpose. In contrast with the active process, in a passive process, the contributor does not know that the data they produce will be used for specific purpose. Figure 3 shows the categorisation of VGI data (See, et al., 2017). The deliberation has resulted in four different possible categories that a VGI data can fall into; active framework data, active non-framework data, passive framework data, and passive

non-framework data. Some examples from different repositories for each category are also displayed on Figure 3. Furthermore, besides those categories, See, et al., (2017) also extended the VGI categorisation by adding 3D VGI as a special case which carries the information of height and elevation.

Furthermore, Sui and Cinnamon (2017) proposed a loosely grouping of VGI data based on multiple commonalities into; geospatial framework data, gazetteer/place name data, and miscellaneous geotagged content data (text, audio, photo, video, etc.). In general, geospatial framework data in this context have the same definition as referred by See, et al., (2016). This data type is the most common data that needed by the users (Sui & Cinnamon, 2017). Meanwhile, gazetteer/place name data are referred to the datasets that produced through VGI that contains place names and points of interest (Sui & Cinnamon, 2017). Another typology is proposed by Senaratne, et al., (2017) by describing three different forms of VGI data as: 1) map, 2) image, and 3) text. These three types of VGI are categorised based on the data capture methods, for example; a GPS data product will be categorised as a map, while a photo categorised as an image, and a plain text classified as text.

Table 2 Categorisation of VGI projects (Bodogna, et al., 2016)

Categories				
Scientific field	Volunteer's task	Way of VGI creation	Need for VGI	Characteristics of volunteer
Computer science	Massive computer time	Automatic and implicit	Low	Neophyte
Astronomy and space sciences	Specific human abilities	Manual and implicit Manual and explicit	Medium	Interested amateur
Weather and environment	Objects identification	Automatic and explicit	High	Expert amateur
Natural science	Observation measurement	Mixed strategy		Expert authority
Medicine	Transcription			Unaware volunteers
Biology	User indication			
Genetics	Complementary information			
Social science				
Urban mobility and planning				
Cultural heritage				

In line with the rapid development of VGI data in the era of big data, the boundary between VGI and other types of geographic information/spatial data is rapidly blurring (Sui & Cinnamon, 2017). However, in respect of data handling, categorisation of VGI data based on the product

types is also important. Understanding the product type of data can lead to a more efficient process in data handling; downloading, pre-processing, processing, and especially assessing the quality of the data. Each data product might need different handling procedure. For example, VGI data that come from OSM is relatively easier to download and process, but data from Twitter or Flickr might need different procedures which include more complicated requirement. This also applies to quality assessment, where those data might need different quality assessment elements since each data type cannot be treated exactly the same due to its nature.

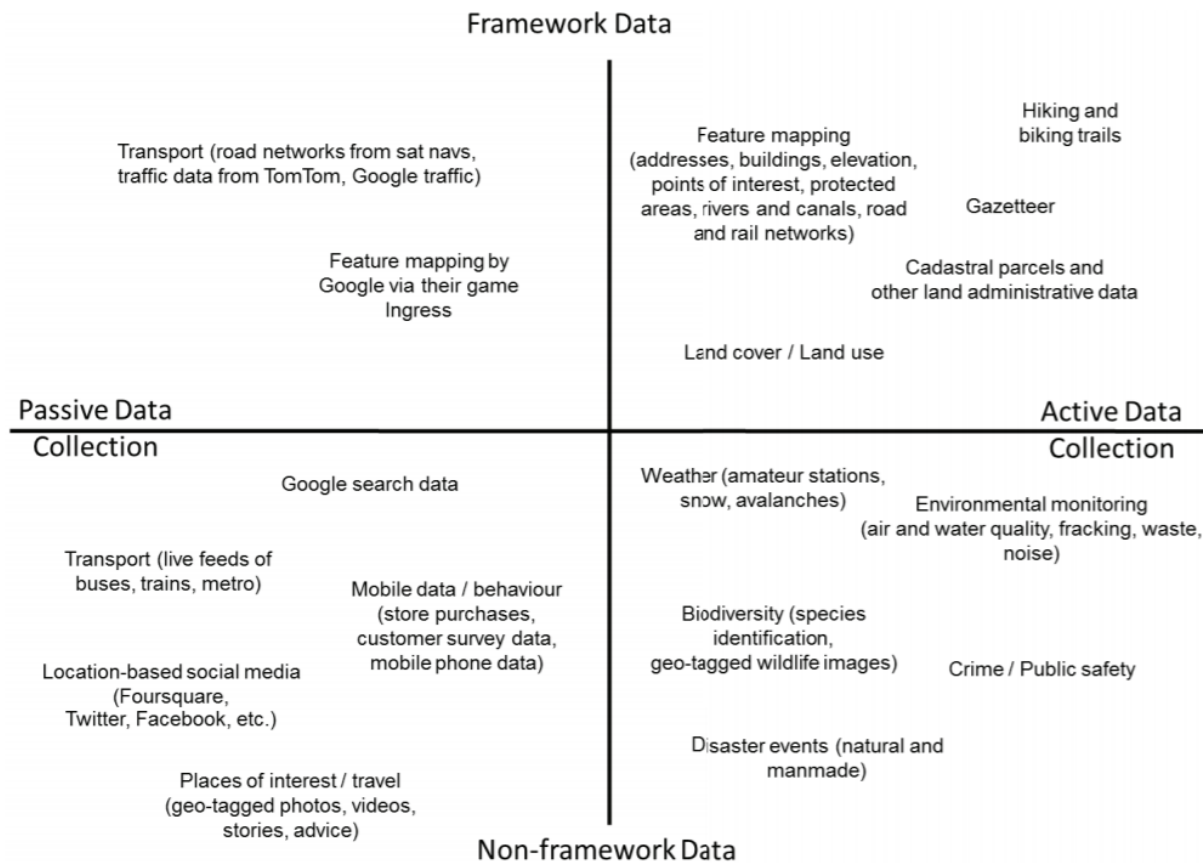


Figure 3 Detail of Categorisation of VGI data (See, et al., 2017)

In general, the categorisation that proposed by Sui and Cinnamon (2017) and Senaratne, et al., (2017) has the similar foundation since it tries to differentiate the most popular and most used VGI data. However, in regard to the data handling, the categorisation that is proposed by Sui and Cinnamon might not well represent the data, especially for the miscellaneous geotagged content data category. The data handling between text, audio, photo, and video will lead to different procedures. Whereas, a typology that is proposed by Senaratne, et al., (2017) might consider the data handling aspect, but due to the current VGI development, the concept needs to be reviewed.

In order to make a clearer distinction on VGI type with the consideration of the current state of VGI data, an adoption of those categorisations will be conducted. The categorisation will be

emphasized on data handling based on the current VGI data. Proposed typology will divide VGI data into three main categories; vector-based (point, line, and polygon), textual (geographic name and geo-tagged post), and media content (photo and video). Those three categories are chosen to represent the data that commonly used for Geographic Information System (GIS) analysis, where vector data represent the geometric aspect, textual data represent the attribute/valuable information, and media content could be the enrichment and comparison for the spatial data. Based on that consideration, a new data type that might emerge concomitant with the development of VGI can be accommodated within the categorisation scheme. Figure 4 shows the deliberation of each category. Description of each category and sub-category will be discussed in the following section. Detail of those distinctions is discussed in the following sub chapter.

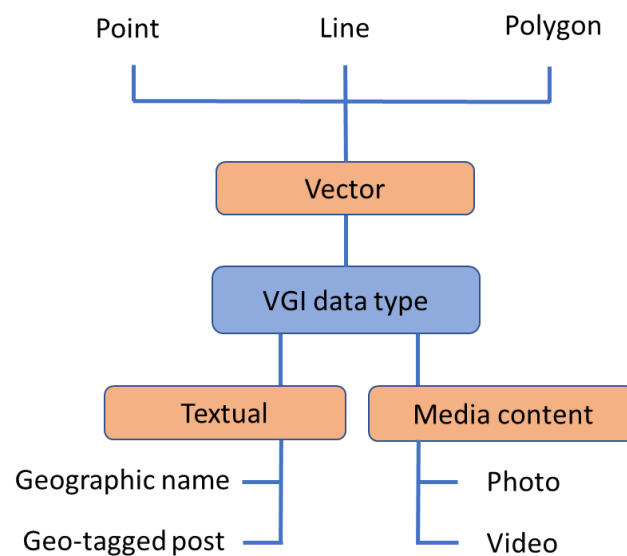


Figure 4 Proposed classification of VGI data type

2.3.1 Vector Data

Geometric primitive is the most commonly used data model to describe the phenomena in this world. This model can be further divided into three basic forms; point, line, and polygon. Though some reference mentioned that geometric primitive can be divided into more than three basic forms, this research will only focus on three basic forms. Those representations contained XY coordinate (location) as well as information related to temporal or spatial variability (Burrough, McDonnell, & Lloyd, 2015).

In terms of Geographic Information System (GIS), these geometric primitive forms can be considered as a vector data. Vector data enable the real world features to be presented in GIS environment (QGIS, 2018). Each geographic feature is represented in a coordinate based data which is featured as a series of one or more coordinate points (Escobar, et al., 2018; ESRI, 2018). A point feature is represented as a single coordinate pair; line feature is represented as a series of related points (vertices), while polygon feature is represented as the collection of related

lines (Escobar, et al., 2018). Related to OSM data, those primitive data types are called as; nodes (point), ways (line), closed ways or area (polygon) (OpenStreetMap, 2018).

This vector data type is not to be directly identified as the framework data, even though it might have some similar characteristics and quality of authoritative data. To bear in mind, the source of vector data is not only well established VGI platform such as OSM or Wikimapia, but also another platform that might have uncertainty on their data quality. However, regardless of the data quality, if the VGI data have the similar characteristic of vector data structure, they can be included into the category.

As listed by Burrough, McDonnell, and Llyod, (2015), the advantages of the vector data structure are; good and accurate representation of entity data model, compact data structure, good topology, easy coordinate transformation, and the possibility of data retrieval, updating, and generalization of graphics and attributes. While the disadvantages are: complex data structure, it might need big computational effort, and complex requirements for spatial analysis and modelling (Burrough, McDonnell, & Llyod, 2015). Nevertheless, vector-based data are the VGI product that mostly seized the attention of most researchers.

Point feature

A point is an abstraction of geographical extent which is represented by one set of XY coordinates that indicate its location (Burrough, McDonnell, & Llyod, 2015). The abstraction of geographic feature as a point is usually a matter of data producer preference or simply used when the feature cannot be displayed as a line or area (Escobar, et al., 2018). Moreover, point features representation is often dependent on the scale (QGIS, 2018). In a small and medium scale map, some line and polygon features from detailed scale map can be represented as a point.

The definition of a point entity can be extended by adding the Z value for the XY coordinate. If the XY value is related to the coordinate reference system being used, the Z value contains information about the height and elevation of an object. The typical XY point which have Z value can be a representation of new category of VGI data that is coined by See, et al., (2016); the 3D VGI. Figure 5 illustrate the dimension of a point with its XYZ attributes.

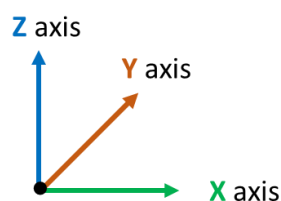


Figure 5 Point Geometry attributes

In terms of VGI data, a point vector data can represent many possibilities of geographic features. It could simply represent the location of an object, value, or information. Some examples of this VGI data type can be obtained from OSM and Wikimapia (i.e. school, restaurant) or simply

generated from social media posts (check-in) to extract POI such as hospitals, train stations, restaurants, and touristic places. Even more, a point can also contain information that could be further semantically analysed to address a problem.

Line feature

If a point feature is represented by a set of coordinate and defined as a single vertex, a line has a pair coordinate set and vertices (QGIS, 2018). Two connected vertices will form a line, while more than a pair of connected and ordered vertices will form "line of lines" or can be simply called as a polyline (QGIS, 2018; Burrough, McDonnell, & Llyod, 2015; Escobar, et al., 2018). A polyline is used to represent a path of linear features or geographic features through space that is too narrow to be displayed as an area at the given scale (Escobar, et al., 2018). For example, a polyline is frequently used to symbolize the geometry of linear features such as roads, railways, rivers, contours, and administrative boundary (QGIS, 2018). However, as a representation of a linear feature that might have width in the real world, the line doesn't represent the true width unless it is specified (Burrough, McDonnell, & Llyod, 2015).

Regarding the VGI data, there are some polyline potential issues that need to be aware of; the geometry, the attribute, and the topology. The geometry issues are related to how the data was retrieved because inconsideration on geometry when producing the data will result into a spikey or angular polylines instead of a smooth polyline that makes the data less accurate. The attributes of polyline are also important because they carry the properties and characteristics of polyline that can describe its semantic information (QGIS, 2018). Unfortunately, sometimes VGI data might not be equipped with complete attributes. No less important, the topology that describes the relation and rules of the polyline is necessary. A good topology will lead to a better accuracy. Nevertheless, some examples of VGI repositories of this data type are OSM, Wikimapia, and communities websites (i.e. Bikemap⁹, Bikely¹⁰, Alltrails¹¹, Mapmyrun¹², Navigasi¹³).

Polygon feature

A polygon can be simply defined as a homogeneous representation of two dimensional spaces that are used to be the 2D abstraction of an area (Burrough, McDonnell, & Llyod, 2015). Polygon is created from a series of vertices that are continuous just like polyline, but it is obliged to have the first vertex in the same position as the last vertex to create an enclosed boundary. A polygon data structure is described the topological properties of an area it represents, such as their shape, neighbours, and hierarchy (Burrough, McDonnell, & Llyod, 2015). Therefore, polygon often shares their boundary with the neighbouring polygons. This boundary sharing can also

⁹ <https://www.bikemap.net/> (accessed 04.09.2018)

¹⁰ <http://www.bikely.com/> (accessed 04.09.2018)

¹¹ <https://www.alltrails.com/> (accessed 04.09.2018)

¹² <https://www.mapmyrun.com/> (accessed 04.09.2018)

¹³ <http://www.navigasi.net/index.php> (accessed 04.09.2018)

determine the data quality since most of GIS analysis requires the boundaries of neighbouring polygons to be exactly coincided (QGIS, 2018). Some examples of VGI repositories of this data type are OSM, Wikimapia, and communities websites (i.e. brwa¹⁴).

2.3.2 Textual Data

The textual category is defined as all VGI data that is presented as text, phrase, sentence, or paragraph that is embedded with a geographical location. The geographical information might be directly or indirectly extracted from the data. There are two sub-categories of textual data; geographic name and geo-tagged post. The advantage of this data is the need for relatively small storage, while the disadvantages are; it might need some extra processing to extract information and typographical error can lead to misinformation.

In term of GIS, textual data can also be presented as points which are retrieved from its geographical location tag. However, the essential difference of this data to be not categorised as a point data (vector data) is that the textual information is the main highlight of the data instead of its location.

Geographic name

The geographic name is defined as label that identifies geographical objects which refers to a specific location (Ormeling, 2003). Geographic names mostly come as short text or phrase, and rarely as a sentence or paragraph. The data usually contains the administrative name (i.e. country name, capital city, city or town) or geographical features (i.e. river, mountain, lakes). Some examples of VGI repositories for these direct information geographic names are Geonames¹⁵, Wikimapia, and OSM, while the indirect information can be extracted from geo-tagged posts or photos in Twitter, Facebook¹⁶, Instagram, and Foursquare.

Geo-tagged post

The geo-tagged post can be simply defined as the social media status (post) that also embedded with the geo-location. In contrast to geographic names, geo-tagged posts mostly come in the long text (sentence or paragraph). The information from geo-tagged posts data type usually needs to be extracted before it can be useful information. Therefore, the usage of hashtag could be helpful to help to extract the information. In some cases, travel websites or blogs could also be categorised in this area as long as they associate their content with geographical location. However, the complication on extracting the information from the website or blog could be bigger and require more resources for the data handling.

The information that is mostly extracted from geo-tagged post could be highly varied. It also might share a disputed boundary with geographic name, but this category emphasises more on

¹⁴ <http://brwa.or.id/sig/> (accessed 04.09.2018)

¹⁵ <http://www.geonames.org/> (accessed 04.09.2018)

¹⁶ <https://www.facebook.com/> (accessed 04.09.2018)

the thematic information, for example political view, disaster response, or emotion mapping. Some example of the data repositories are; Twitter, Facebook, Instagram, and Foursquare.

2.3.3 Media content

Media content can be described as the information that is aimed to the end user which delivered through media. There are several types of media content; text, music, picture, and video. In this context, even though text might be categorised as a part of media content it is not categorised into this category since it has different perquisites in data handling. Therefore, based on the current status of VGI development, to narrow down the the context this category is meant to facilitate image content which includes photo (a digital visual representation of an object) and video (a digital recording of moving visual object) that is produced by social media. However, in consideration that there are various possible data types that are produced by media, this category can be expanded into a broader content. In the near future, there might be more VGI media content (i.e. audio, animation, graphics drawings) that can be used as an alternative to enrich the geographic information.

The geographic location of media content can be obtained explicitly or implicitly. Explicit location can be directly obtained from Global Positioning System (GPS) tagging or offline geotagging, while the implicit location can be retrieved from the hashtag. For this category, metadata is necessary since it contains important information for accuracy assessment.

3 Spatial Data Quality Assessment

As a crowdsourcing product, VGI has both potencies and weakness. Despite its advantages, VGI data also comes with drawbacks, such as; uncertain quality, undocumented quality, as well as disparities in data quality (Goodchild & Li, 2012; Minghini, et al., 2017). Moreover, the completeness of data coverage and the sustainability of the data also becomes the concern of the scientists and users (Goodchild & Li, 2012; Olteanu-Raimond, et al., 2017a). Those drawbacks, especially the uncertainty of VGI, have become the biggest obstacle for VGI data adoption into authoritative spatial data.

Compared to the authoritative spatial data, VGI data might be less accurate, less formally structured, and lack of associated metadata that allows it to be used as framework data (Jackson, et al., 2010; Du, et al., 2016). However, the investigation of VGI quality assurance that was conducted by Goodchild and Li (2012) infers that even though VGI quality is considered as highly varied and undocumented, it may play a useful role and could give great benefit if its quality could be improved and assured. Hence, the VGI quality assessment is essential. By conducting the data quality assessment, the potential usability of VGI data can be profoundly explored, especially related to its integration with authoritative data. Furthermore, the VGI improvement recommendation can be formulated based on the quality assessment result. Therefore establishing a data quality assessment framework has become a popular topic among researchers and practitioners.

3.1 ISO Quality Assessment

The International Organization for Standardization (ISO) published a standard reference for assessing the geographic information data quality, namely: *ISO 19157:2013 about Geographic Information - Data quality*. The standard describes the elements and procedure of data quality assessment. The elements that are specified in the standards including: positional accuracy, thematic accuracy, completeness, temporal quality, logical consistency, and usability (ISO, 2013).

Positional accuracy. The element focuses on evaluating the coordinate position of an object that is related to real location on the ground. This element might be the most essential aspect of quality assessment. Sometimes, the term of geometric accuracy is also used by the researcher to conduct positional accuracy assessment with the addition of geometries evaluation in it. Assessing the positional accuracy of vector data might involve a broad range of sub-element parameterizations rather than textual and media content. Each vector data type (point, line, and polygon) will need different parameterization due to different nature that they have. In contrast, textual data and media content might only need less parameterization to conduct the assessment.

Thematic accuracy. The conformity level between the object and its attribute are measure on this element. The attribute should reflect the semantic information of an object related to the condition of the real world. If the information cannot be interpreted in the way it should

represent the real world condition, it means that the thematic accuracy is low. Thematic accuracy has a similar purpose with the semantic accuracy. In terms of VGI project, the thematic accuracy will be assessed by considering the purpose of the VGI project.

Completeness. It shows the comprehensiveness of the data information, whether it is absence (omission) or excess (commission) of the information. The measurement can be conducted area-wise and attribute-wise or both approaches. Completeness has always been an issue for VGI data assessment since the distribution of data are uneven and has less attribute content if compared to authoritative spatial data.

Temporal quality. This element is used to validate the actuality of data in respect of the changes in the real world (Girres & Touya, 2010). Furthermore, it is also used to measure the rate of updates. One of the VGI advantages is its' currency and timeliness, thus making VGI as a highly dynamic data. However, the changes are usually well documented therefore it is traceable.

Logical consistency. The internal consistency of data, especially on topological correctness and relationships of the data are evaluated using this element (Haklay M., 2010; Girres & Touya, 2010).

Usability. The overview of non-quantitative information regarding the data quality is over-viewed using this element (Docan, 2013). The fitness of data to be used in the project is assessed, which can help potential users in deciding how the data should be used (Haklay M. , 2010). This element can be further detailed as: usage, purpose, and constraints that are also often used by the researcher to examine usability element.

ISO 19157:2013 covers two different perspectives of data quality assessment; internal quality and external quality. The internal quality (positional accuracy, thematic accuracy, completeness, temporal quality, logical consistency) are focussed on producers' point of view of a dataset, while the external quality (usability) is focussed on the user's needs and requirements (Docan, 2013; Meek, et al., 2014; Fonte, et al., 2017).

Similar components are proposed by Van Oort (2006) but with the addition of; data lineage, attribute accuracy, variation in quality, metaquality, and resolution. The lineage represents the mile stones of the data evolution; how and when it was obtained and how often and what kind of changes it had. All of these histories are recorded as the data lineage (Haklay, M, 2010; Girres & Touya, 2010). The attribute accuracy refers to the data attributes that does not represent the location and temporal aspect of the data (Van-Oort, 2006). The variation in quality occurs when there are diverse qualities within a data, while the metaquality describes the quality information of the data (Van-Oort, 2006). Both variations of quality and metaquality are considered as a part of other elements (Van-Oort, 2006). The resolution element is used to ensure that the user has the right data scale to meet the fitness-of-use of the assessment. However, the variation of quality, metaquality, and resolution are rarely used.

Those elements that mentioned above are widely used to conduct the quality assessment, either partially or completely. Each of the elements is break down into sub-elements that will help to describe the particular quality aspect of the elements. The sub-elements can be modified according to the user's needs. Even though the elements that will be used to assess the data are the same, but when it comes to different data types, some sub-elements need to be specifically customized according to the data type. For example in vector data type; point, line, and polygon will need different sub-elements assessment since they have different nature.

Antoniou and Skopeliti (2015) had summarized most of the elements and sub-elements that were used in previous research. A brief summary of the quality assessment elements are shown in Table 3. The table also includes some addition of the possible sub-element that can be used to assess the data. Although the sub-elements mentioned on the table are dominated by the sub-element that is intended for vector data type, it also mentions some sub-elements that are specifically use for another data type such as geo-tagged posts and photos. In respect of process automation, most of the quality assessment's elements and sub-elements are possible to be done automatically. The automation process can be conducted by using well established software or by creating plugins through programming environment.

Table 3 Resume of quality assessment elements (Antoniou & Skopeliti, 2017)

Elements	Sub-elements
Positional accuracy	<ul style="list-style-type: none"> • The buffer zone (Goodchild & Hunter, 1997; Haklay M., 2010; Kounadi, 2009; Koukoletsos, et al., 2011; Arsanjani, et al., 2013); • The distance between corresponding intersections of a road network (Antoniou, 2011); • The Euclidean distance for point features (Girres & Touya, 2010; Stark, 2011; Jackson, et al., 2013; Mashhadi, et al., 2014); • The average Euclidean distance for linear features (Girres & Touya, 2010; Fan, et al., 2014); • The Hausdorff distance for linear features (Girres & Touya, 2010; Wiemann & Bernard, 2015); • The surface distance, granularity, and compactness for area features (Girres and Touya, 2010); • The shape similarity (turning function) (Mooney, at al., 2010; Fan, et al., 2014; Kalantari & La, 2015); • X and Y error distance (Stark, 2011); • The grid based minimum bounding geometry and the directional distribution (Standard Deviatonal Ellipse) (Forghani & Delavar, 2014); • The number of vertices, the mean vertex distance, and distances between polygons centroids (Kalantari & La, 2015); • Spatial similarity in multi-representation considering directional and metric distance relationships (Hashemi, et al., 2015); • Comparison of coordinates (Stankute & Asche, 2009); • Searching zone, orientation, and length (Koukoletsos, et al., 2012).

Elements	Sub-elements
Thematic accuracy	<ul style="list-style-type: none"> • The percentage (%) of correct classification (Stark, 2011; Kounadi, 2009; Girres & Touya, 2010; Fan et al., 2014); • The percentage (%) of specific values existing in tags (Girres & Touya, 2010, Antoniou, 2011); • The Levenstein distance (Girres & Touya, 2010; Mashhadi, et al., 2014; Kalantari & La, 2015); • Damerau-Levenshtein (Wiemann & Bernard, 2016); text similarity constraint (Koukoletsos et al., 2012); • The number of features with specific attributes (Fan, et al., 2014; Arsanjani et al., 2013); • Confusion matrix and standard kappa index analysis (Arsanjani & Vaz, 2015; Arsanjani, et al., 2015; Docan, 2013); • User's and producer's accuracy (Arsanjani & Vaz, 2015).
Completeness	<ul style="list-style-type: none"> • Grid-based length comparison against authoritative data (Haklay M., 2010; Ludwig, et al., 2011; Zielstra & Zipf, 2010; Ciepluch, et al., 2011; Forghani & Delavar, 2014); • Comparison of the number of features (Girres & Touya, 2010; Jackson, et al., 2013); • Comparison of total length or total area (Girres & Touya, 2010; Kounadi, 2009; Koukoletsos, et al., 2011; Fan et al., 2014; Kalantari & La, 2015; Arsanjani & Vaz, 2015); • Completeness measure (Mashhadi, et al., 2014); and completeness index (Arsanjani, et al., 2015); • Misclassification matrix (Docan, 2013).
Temporal quality	<ul style="list-style-type: none"> • Date of publication, date of collection, update frequency, last update or temporal validity (Fonte, et al., 2017); • Evolution of VGI data record (Girres & Touya, 2010; Arsanjani, et al., 2013); • Time difference between photo capturing and uploading (Antoniou, et al., 2010).
Logical consistency	<ul style="list-style-type: none"> • Intra-theme consistency or inter-theme consistency (Girres & Touya, 2010); • Administrative data integrity (Ali & Schmid, 2014); • Topological consistency (Corcoran, et al., 2010); • Spatial similarity to assess topological relationships (Hashemi, et al., 2015); • Semantic similarity between the tags (Vandecasteele & Devillers, 2015); • The identification of entities with inappropriate classification (Ali, et al., 2014); • Ontologies in data tagging (Codescu, et al., 2011); • Tag recommendation system (Vandecasteele & Devillers, 2015).
Usability	Usability is depending on the end-user aims and usually expressed as "fitness for use".

Table 4 ISO quality elements, their requirements, and issues related to their use with VGI (Fonte, et al., 2017)

ISO Quality Elements		Requirements	Issues for the Application to VGI
Internal quality	Positional accuracy	<ul style="list-style-type: none"> • Data specification • Existence of reference data with similar characteristics and valid time frame 	<ul style="list-style-type: none"> • Lack of specifications • Dynamic nature of VGI • Inexistence of comparable reference data • Spatial and thematic heterogeneity
	Thematic accuracy		
	Completeness		
	Temporal quality		
	Logical consistency	<ul style="list-style-type: none"> • Other data of the same source or independent data 	<ul style="list-style-type: none"> • Applicable to VGI • May enable automatic validation checks
External quality	Usability	<ul style="list-style-type: none"> • Specification of user needs 	<ul style="list-style-type: none"> • May be assessed by combining quality measures and indicators

Regarding the VGI data, those six major elements can be used to explore the data quality. However, considering the nature of VGI data, not all elements are compatible to be used for VGI data. There are some issues regarding the application of ISO quality elements to be applied for assessing VGI data quality. Fonte, et al., (2017) had listed the related issues that might come up from the requirements of each element (

Table 4). In terms of VGI integration with authoritative spatial data in this research, to anticipate the issues that might occur the combination of ISO elements and additional elements will be further explored. The additional elements that will be used should accommodate the nature and type of VGI data, therefore the result of quality assessment can properly represent the characteristics of VGI data.

3.2 VGI Quality Assessment

To bridge the gap and improve the VGI data quality assessment, many scientists had proposed a "VGI friendly" framework which takes the nature of VGI data into consideration. Those frameworks came in different methods, either using a quantitative or qualitative approach. By doing so, they argue that it satisfactorily represents the nature of VGI which refer to be more suitable and fair method to assess VGI data quality. They also claimed that it could be the solution to improve the result of quality assessment.

Goodchild and Li (2012) described three approaches to determine the VGI quality; crowdsourcing revision, social measures, and geographic consistency. They claimed that these three approaches can be well compared to the quality assessment that is conducted by the authoritative body. Senaratne, et al., (2017) extended those three approaches by including data mining as an additional approach. Data mining approach is an independent factor that is gained through a computational process to discover patterns and data learning. However, the data mining might

involve extend effort of computational process that will prolong the quality assessment process. Another approach to assess VGI quality was proposed by Lush, Bastin, and Lumden (2012) by exploring the data elements, such as; metadata, metadata visualisation and comparison, community advice, reputation of data provider, and citation information.

Meek, et al., (2014) proposed to add stakeholder model between internal and external quality. The stakeholder model includes the quality elements such as; vagueness, ambiguity, judgement, reliability, validity, and trust (Meek, Jackson, & Leibovici, 2014). Bordogna, et al., (2015) introduced a user driven assessment to be combined with internal and external quality elements. The minimum acceptable quality level that is specified by the user will be used to select the corresponding VGI items (Bordogna, et al., 2015). Antoniou and Skopeliti (2015) examined the difficulties in using authoritative datasets for quality assessment based on the literature review. As the result, some quality indicators were proposed to overcome the problem which was formulated into four main categories; data, demographics, socio-economic situation and contributors. Forati and Karimipour (2016) presented a model for analysing characteristics of trustworthiness and reputation then creating and an automatic method to create trustworthiness and reputation scores in order to assess the quality of VGI features. Moreover, Fonte, et al., (2017) proposed additional indicators into the ISO 19157 quality components. The three main indicators proposed were: data-based indicators, demographic and socio-economic indicators, and contributors indicators. Table 5 shows the summary of proposed VGI quality elements.

Table 5 Summarized of proposed VGI quality assessment

Researcher	Proposed VGI Quality Elements
Goodchild and Li (2012)	Crowdsourcing revision, social measures, and geographic consistency
Lush, Bastin, & Lumsden (2012)	Metadata, metadata visualisation and comparison, community advice, reputation of data provider, citation information
Senaratne, et al., (2017)	Measure of quality, indicators of quality, data mining
Meek, et al., (2014)	Vagueness, ambiguity, judgement, reliability, validity, trust
Bordogna, et al., (2015)	User driven assessment
Antoniou and Skopeliti (2015)	Data, demographics, socio-economic situation and contributors
Forati and Karimipour (2016)	Trustworthiness and reputation
Fonte, et al., (2017)	Data-based indicators, demographic and socio-economic indicators, and contributors indicators

The automation of quality element always becomes the added value since it can accelerate the assessment process. Unfortunately, not all of the proposed additional VGI quality elements are proven can be done automatically. Moreover, some of the models proposed are remains theoretically explained without implementation (Fonte, et al., 2017). However, the latest

developments of additional VGI quality elements are always be proven with the ability to be performed automatically.

Proposed as the alternative method to assess VGI data quality, some of the elements might have an overlapping aspect with the ISO quality elements. For example, the data-based quality indicators (Fonte, et al., 2017) contain geometrical aspect (position accuracy) and logical consistency. Moreover, since usability is the main concern of quality assessment, most of the proposed methods always include it. Therefore, the combination of ISO and VGI quality elements should consider the overlapping aspect so that there will be no redundancy in the assessment process. Ultimately, the idea of integrating the quality elements and workflows design for quality assessment should address the potential issues and capable to improve the confidence level of VGI data effectively.

3.3 Quality Assessment Modules

Every different type of VGI data needs different accuracy assessment elements and sub-elements since they have different nature. If the quality assessment is conduct by using an inappropriate method it will mislead the assessment result and diminish the user's benefit. Therefore, it is very important to use the proper method to assess the VGI data. As always emphasised before, each VGI data type has specific nature that needs to be handle respectively. Therefore, this research proposes the usage of quality modules to conduct the quality assessment. The module consists of quality elements and sub-elements that are combined from ISO standard and VGI approach.

Some example of how this module assembled are shown in Table 6, Table 7, Table 8. Those modules consist of two different approaches; ISO standard and VGI approach. Each approach then breaks down into element and sub-elements. Specific VGI data type might require a specific sub-element for the accuracy assessment. Therefore, the selection of each sub-element has to consider the data type so that the data handling process can be conducted accordingly. Furthermore, in the implementation, the criterion for each sub-element is needed to determine the quality of the features. The criterion will be used to set the threshold that defines if the data meet the minimum requirement of quality assessment. Nevertheless, the selection of sub-elements in the quality assessment module has to refer the user requirement.

Table 6 Example of a quality module for the line – vector data type

Approach	Element	Sub-element
ISO Standard	Positional accuracy	<ul style="list-style-type: none"> • The buffer zone/Searching zone; • Orientation/Angular difference; • Length; • the average Euclidean distance; • The Hausdorff distance; • Comparison of coordinates;

Approach	Element	Sub-element
		<ul style="list-style-type: none"> • Topology relation;
	Thematic accuracy	<ul style="list-style-type: none"> • Confusion matrix; • user's and producer's accuracy; • Standard kappa index analysis; • The percentage of correct classification; • Damerau-Levenshtein distance; • The number of features with specific attributes;
	Completeness	<ul style="list-style-type: none"> • Comparison against reference data; • Comparison of number of features; • Comparison of total length or total area; • Completeness measure; • Completeness index;
	Temporal quality	<ul style="list-style-type: none"> • Date of publication; • Date of collection; • Update frequency; • Last update or temporal validity; • Evolution of VGI data record;
	Logical consistency	<ul style="list-style-type: none"> • Intra-theme or inter-theme consistency; • Administrative data integrity; • Topological consistency; • Spatial similarity (topological relationships); • Semantic similarity between the tags; • The identification of entities with inappropriate classification;
	Usability	<ul style="list-style-type: none"> • Usage • Purpose • Constrain
VGI Approach	Additional Module	<ul style="list-style-type: none"> • Metadata • Credibility • Reliability • Trust

Table 7 Example of a quality module for the polygon – vector data type

Approach	Element	Sub-element
ISO Standard	Positional accuracy	<ul style="list-style-type: none"> • The surface distance, granularity, and compactness for area features; • The shape similarity (turning function); • Standard Deviation Ellipse; • The number of vertices;

Approach	Element	Sub-element
		<ul style="list-style-type: none"> • The mean vertex distance and distances between polygons centroids; • Spatial similarity based on directional and metric distance relationships;
	Thematic accuracy	<ul style="list-style-type: none"> • Confusion matrix; • user's and producer's accuracy; • Standard kappa index analysis; • The percentage of correct classification; • The percentage of specific values existing in tags; • Damerau-Levenshtein distance; • The number of features with specific attributes;
	Completeness	<ul style="list-style-type: none"> • Comparison against reference data; • Comparison of number of features; • Comparison of total length or total area; • Completeness measure; • Completeness index;
	Temporal quality	<ul style="list-style-type: none"> • Date of publication; • Date of collection; • Update frequency; • Last update or temporal validity; • Evolution of VGI data record;
	Logical consistency	<ul style="list-style-type: none"> • Intra-theme or inter-theme consistency; • Administrative data integrity; • Topological consistency; • Spatial similarity (topological relationships); • Semantic similarity between the tags; • The identification of entities with inappropriate classification;
	Usability	<ul style="list-style-type: none"> • Usage • Purpose • Constrain
VGI Approach	Additional Module	<ul style="list-style-type: none"> • Metadata • Credibility • Reliability • Trust

Table 8 Example of a quality module for the geo-tagged posts – textual data type

Approach	Element	Sub-element
ISO Standard	Positional accuracy	<ul style="list-style-type: none"> • The buffer zone; • The Euclidean distance; • x and y error distance;
	Thematic accuracy	<ul style="list-style-type: none"> • The specific values existing in tags; • Damerau-Levenshtein distance; • Key words/tag suitability
	Completeness	<ul style="list-style-type: none"> • Spatial distribution of VGI data in research area; • Completeness measure;
	Temporal quality	<ul style="list-style-type: none"> • Date of publication; • Date of collection; • Update frequency;
	Logical consistency	<ul style="list-style-type: none"> • Intra-theme or inter-theme consistency; • Semantic similarity between the tags;
	Usability	<ul style="list-style-type: none"> • Usage • Purpose • Constrain
VGI Approach	Additional Module	<ul style="list-style-type: none"> • Contributors' indicator : interests, contribution history, recognition, location, behaviour, education, profile • Demographic and socio-economic indicators

Those proposed VGI quality modules are just some examples on how to use the sub-elements to perform accuracy assessment based on the data type. Following the concept of flexibility, each module can be modified by adding or eliminating the elements or sub-elements based on the user's requirement.

3.4 VGI Usability

VGI usability among European NMA had been inventoried by Olteanu-Raimond, et al. (2017a) as shown in Figure 6. The adoption of VGI into authoritative framework has been applied in a wide range of application, from new data collection, vernacular place name, change detection, photo interpretation, and report alerts. Furthermore, the paper also describes that the involvement of VGI in the authoritative framework is not limited to data adoption, but also used for development of tools and frameworks to promote, produce, use, and support the sustainability of VGI data.

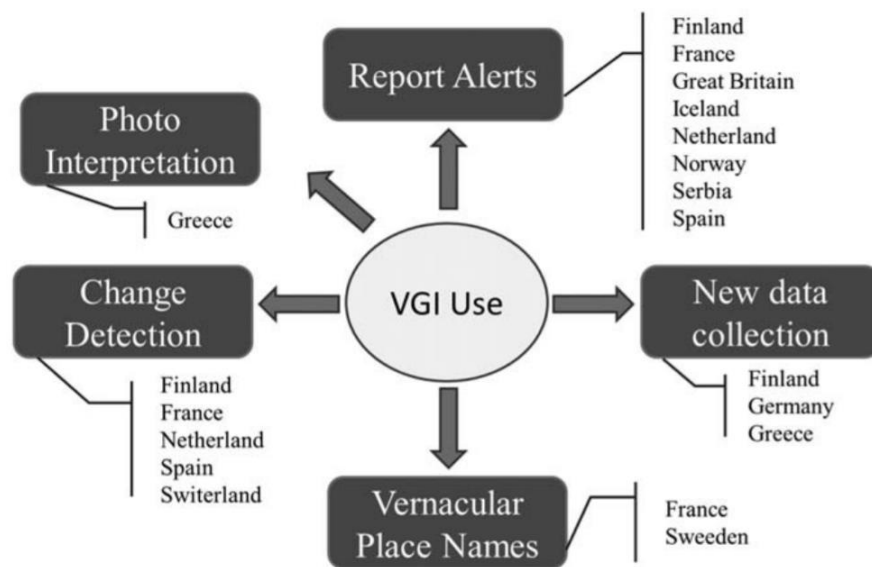


Figure 6 The use of VGI in the European NMAs (Olteanu-Raimond, et al., 2017a)

To be seen in the wider application of VGI data in the authoritative data scheme, there are more specific applications that can be found worldwide. The scale of involvement of VGI data to be used as authoritative map can be vary greatly, which mostly influenced by the condition of data availability in some area. For example, for the country or newly created nation that still lack of authoritative data, VGI initiative will be used by the government to collect the data. By doing so, they can obtain VGI data with controlled quality which are ready to be adopted as authoritative spatial data. In other case, the VGI can be used to support specific thematic map such as; disaster preparedness, animal migration, biodiversity, seismic activity, street bump, vernacular names, and slum settlement mapping (Haklay, et al., 2014). As investigated by Haklay, et al., (2014) from 29 case studies across the globe, the use of VGI are extended and specified into; basic mapping coverage, authoritative spatial datasets update, public sector services upgrade, policy development or reporting, natural disaster preparedness and crisis management. To summarize all possibilities of VGI usability, Table 9 was formulated to give general point of view of the commonly used data. All of the usability can give contribution in updating the authoritative spatial data. Even though it might not give direct contribution such as providing a new data for update but it can give an alert or indication if the update is needed.

Table 9 VGI usability summary

Usability	Description	Requirement
Data adoption	Adoption of a VGI dataset into an authoritative data scheme as a totally new feature or a part of existing authoritative data. The adoption could applied both for base map (topographic map) or thematic map. The datasets could be vector data, geographic name, or any other VGI type.	High accuracy

Change detection	Collation between VGI and authoritative data to spatially detect and analyse the change occurring in the real world. This category is mostly occupied with vector data since they usually have a good spatial aspect.	Intermediate accuracy
Visual interpretation	VGI data that can be used as the comparison material to help describe/characterize a phenomenon. This category is specifically related to the media content data.	Intermediate accuracy
Semantic Information	VGI is not directly integrate into authoritative spatial data but extracted into information that can be used to enrich the authoritative spatial data attribute. This category could cover all data type as long as they can demonstrate excellency in thematic aspect.	Intermediate accuracy
Rapid response	Emphasised in VGI data up-to-dateness to generated a quick response towards critical issues immediately (i.e. disaster, election).	Intermediate accuracy
Report alert	VGI data that is intended as an indicator of error or changes in authoritative data.	Low accuracy

There are three categories of accuracy requirement; high, intermediate, and low accuracy. High accuracy requires the VGI data to have, at least, similar quality as the authoritative data. Thus, the data that fall into this category are highly matched with the authoritative data and able to be fully adopted as a new or a part of existing authoritative spatial data. Some examples of data adoption application are; the mapping of South Sudan; FINTAN vernacular place names project in UK; Mapping School and health facilities in Nepal; and National Biodiversity Data Centre, Ireland (Haklay, et al., 2014). Based on those examples, arguably the high quality of VGI data usually resulted from VGI projects that are initiated together with the government.

In contrast, intermediate accuracy might not require the data to match all the 'golden standards', but still reliable and can be used as an information input to enrich the authoritative spatial data. However, to be considered as intermediate accuracy, besides an adequate geographic accuracy, the VGI data has to have unique or outstanding aspect (i.e. thematic, temporal, change records) that can contribute to data enrichment. As an example, some vector data that might not have adequate geometric accuracy, but have advanced thematic information in current time, are still considered as the valuable data since it can give change indication of an area.

For the low accuracy level, the VGI data can only be used as a report alert. Consequently, any kinds of VGI data with various condition and accuracy can be categorised into this level. Nevertheless, neither the contribution nor the usage to update the authoritative spatial data might be insignificant.

Taking into consideration that the usability of VGI data is driven by the quality assessment's result, a valuation process is needed in order to know the quality categorisation. The quality

valuation is conduct by using scoring and weighting method. The score will be given to each quality element as the representation of the proximity level and relevance of the element to the data quality. The score is given in ranged value from 0 to +1, where 0 indicate the incompatibility and +1 indicates the compatibility towards the quality element. The compatibility of a feature towards the quality element usually determined by a criterion that set by the user as threshold. For example, the criterion used for determining positional accuracy is 5-meters buffer area. If a feature fall within 5-meters buffer area it will be consider as positionally correct and get the value +1.

Table 10 shows the score range of each element and its definition. The total score ranges from 0 – 7, where the perfect score indicates that the data have very high quality while lowest score indicates bad data quality. The high quality data mean that the entire quality element standards are met and the data can be adopted into authoritative spatial data. It also can be seen from Table 10 that all of those seven elements are considered to have the same score, which indicates that all of them have equal role in determining the data quality. However, dealing with VGI data in real-world cases, each element cannot be considered to have equal role. Thus, a supplementary method will be used to overcome the problem.

Taking into account that each element has different influence on data quality, a weighting factor is need to be included. With the unpredictable nature of VGI data, the weighting factor cannot be presented as a fixed value for each criterion. Fixed weighting value will reduce the added value of VGI data. For example, vector data type that have more advantages on spatial aspect cannot fairly be compared to the geo-tagged posts data that are excellent on providing the thematic content. Therefore, the weighting value are suggested to be specified by the user and tailored according to their needs. Moreover, the flexible weighting can extend the flexibility of method application in wider range of cases.

Table 10 Scoring of quality element

Quality Element	Score	Description
Positional accuracy	0 – 1	0 = inaccurate position; +1 = positionally correct
Thematic accuracy	0 – 1	0 = thematically unsuitable; +1 = thematically suitable
Completeness	0 – 1	0 = incomplete data; +1 = complete data
Temporal quality	0 – 1	0 = incorrect time; +1 = temporally correct
Logical consistency	0 – 1	0 = unlikely; +1 = very likely
Usability	0 – 1	0 = unsuitable; +1 = suitable
Additional Module	0 – 1	0 = inaccurate position; +1 = positionally correct
Total Score	0 – 7	0 = low quality data; +7 = high quality data

The valuation process (scoring and weighting) will be further calculate to obtain the confidence score. Since the weighting is left flexible for the user, the determination of confidence index

category will be based on the percentage of total value obtained from the total ideal score (Equation 1). By doing so, the various value resulted from different usage of weighting factor can be covered by the confidence score. Confidence score are expressed in percentage which represent data correctness. The values that are obtained from the calculation will be translated into confidence index. The confidence index in this research defined as the index that suggest the proper usability of VGI data. By using the assumption that VGI can give potential contribution in every quality level, the confidence index shall represent all potential usability of VGI. Table 11 display the confidence index assignment that shows the translation of confidence score and quality level.

$$\text{confidence score} = \frac{\text{total obtained score}}{\text{total ideal score}} \times 100\%$$

Equation 1 Confidence Index equation

Table 11 Confidence index assignment

Confidence index	Confidence Score	Quality Level
1	>80%	High accuracy
2	35% - 80%	Intermediate accuracy
3	<35%	Low accuracy

4 VGI Data Utilisation Framework and Scenarios

Haklay, et al., (2014) state that "The acceptance of volunteered geographic information (VGI) as a valued and useful source of information for governments is growing at all levels" this statement has indicated the powerful potency and increase reliability of VGI data for official use. High enthusiasm from government agency to embrace the VGI data was also indicated by Brabham (2013). By applying the right method for data processing, the VGI data can be an effective approach to problem solving (Haklay, et al., 2014). Though the utilisation of VGI data for authoritative purposes is not a new topic for research, the further development of framework for updating authoritative spatial data is very important.

Before designing the framework, it's good if the recommendations from previous research are taken into consideration. Brabham (2013) suggested ten best practices for crowdsourcing in government: 1) Clearly define the parameters of problem and solution; 2) Determine the level of commitment for the outcomes, commit to communicate to the online community on exactly how much impact user-submitted ideas and labour will have on the organization; 3) Know the online community and their motivations; 4) Invest in usable, stimulating, well-designed tools; 5) Craft policies that consider the legal needs of the organization and the online community; 6) Launch a promotional plan and a plan to grow and sustain the community; 7) Be honest, transparent and responsive; 8) Be involved, but share control; 9) Acknowledge users and follow through on obligations; and 10) Assess the project from different angles.

Furthermore, Haklay, et al., (2014) delineates seven main components of VGI projects for government use: 1) Incentives/drivers to start a project (mostly from the government perspective); 2) Scope and aims; 3) Participants, stakeholders and relationships, identifying the roles that different participants play; 4) Modes of engagement; 5) Technical aspects; 6) Success factors; and 7) Problems encountered.

Referring to those recommendations, important aspects in creating a framework can be mapped. Moreover, the conceptual design of the framework should be able to address and encounter the issues that might emerge. Finally, flexible application of the framework should be taken into account.

4.1 Conceptual Design of VGI Data Utilisation

The conceptual design of VGI data utilisation involve four major components; stakeholders, inputs, technical process, and outputs. Those elements will be formulated into a framework that illustrated the relationships and roles of each element. But above all, to identify what elements should be included in the components are necessary.

Identifying the stakeholders that involved in the framework is necessary since the stakeholder plays an important role. There are three groups of stakeholders that will be involved in the conceptual design; user, contributor, and data examiner. Meanwhile, the input needed for the framework are; user requirements, VGI data, and data reference. Technical process includes;

data pre-processing, quality assessment process, and determination of confidence index. Ultimately, from all the processes two types of output data are produced; unqualified data and VGI usability. The framework is illustrated in Figure 7.

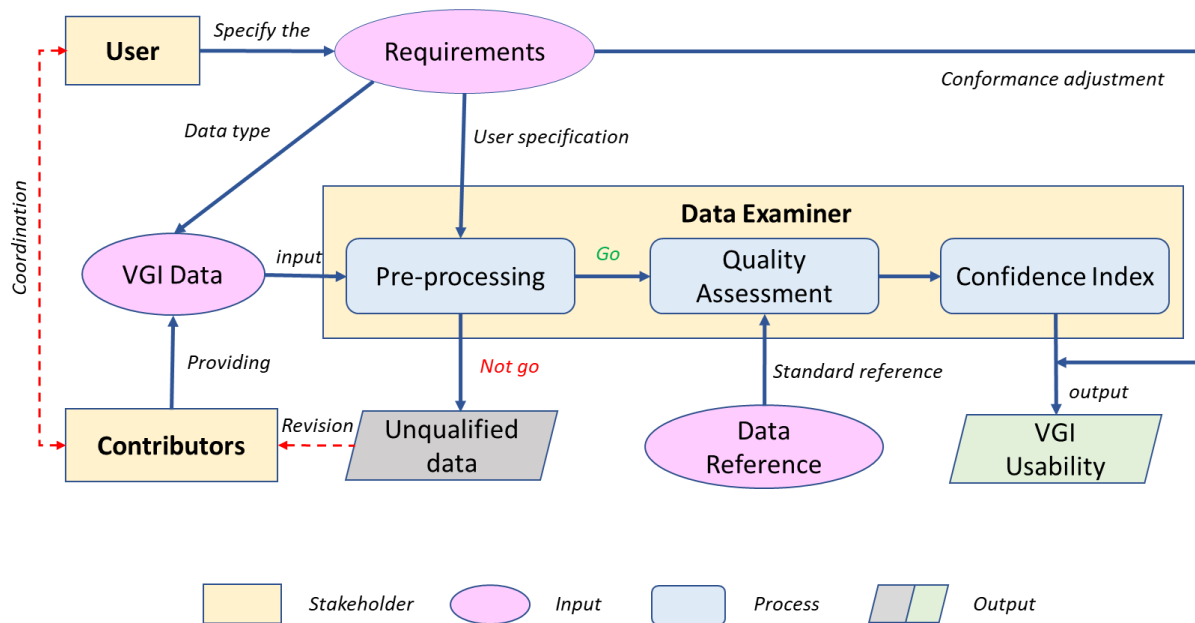


Figure 7 Proposed framework for updating authoritative spatial data with VGI data

4.1.1 User

Since this framework is designed for updating the authoritative spatial data by mean of VGI data, the user in this framework is described as an authoritative body that needs the VGI data for their project. By referring the user as authoritative bodies, it covers various level of administration including: national scale agencies (i.e. National Forest Agency, National Disaster Management Agency, National Mapping Agency), provincial government, local government, or any authoritative level. The user in this conceptual framework does not always become the end-user of the output product. In some cases, the user in this framework will further disseminate their output to be use by the end user (referring to "*public-government-public*" information flow). The user is the key player on this framework since all the process customized based on their requirements and purposes.

4.1.2 Requirements

The requirements are specified by the user based on their project's goal. It will be used to determine the inputs and tasks for the next stages in the framework. In general, the requirements can be divided into three different parts; 1) data specification, 2) user specification, and 3) conformance adjustment.

The data specification requirements are determined by the user according to their project needs, which includes: data type, data repositories, and in some cases data contributors. The

VGI categorisation scheme that is discussed in section 2.3 will be referred to determine and classify the desirable data type. By selecting a data type, the potential data repositories and data handling methods can be determined. Data repository is chosen by the user, which later downloaded and used as the data input. Choosing the data contributors is possible by selecting the contributor's reputation or similar factor, but this step will be applied during the filtering process.

The detail of user requirements are expressed in user specification. In this part, the user determines the region of interest (project area/coverage), desirable thematic information, time constraint of the data, and other project-specific criteria such as quality module, elements, sub-element and threshold that will be employed on the process. All those specifications will be used as the input to run the quality assessment.

The third part of user requirements is conformance adjustment where the user provides the information of tolerable aspect of VGI that can be accepted by the users. This conformance adjustment is not always specified in the beginning of the process, but it also can be specified later to be confronted with the usability recommendation derived from confidence index before defining the final VGI usability. Therefore, categorisation of VGI data usability according to the purpose of the project is important.

4.1.3 Contributor

The contributors are defined as the individual, community, non-profit organisation, or commercial sector that produced any kind of VGI data. The contributors do not just acts as the VGI data provider but in specific case they also can request or offer their data to the user (authority) so that their data can be integrated with authoritative data. Thus, the communication between user and contributor are established. Another case of communication type happens in the VGI project that is initiate by the government. The communication is established when the government, as the user, gives the protocol on how the data should be created to the contributors. Even though the communication between contributor and user do not always happen, there are many possibilities that this kind of communication could develop and evolve into a broader extent.

4.1.4 VGI data

Referring to the data specification that described by the user, the VGI data are collected from the preferred VGI repository. Using the categorisation that was explained in section 2.3, type of the data is determined. The chosen VGI data will be used as the input of the process. The following stages are very dependent to the chosen VGI data type since all the processes are adapted based on it. Some projects might need more than one data type as the input. In that case, the users can utilise more than one data type. However, because the processing module is customized based on the data type, the process for each data is conducted separately.

4.1.5 Data examiner

The data examiner is the stakeholder who is in charge of performing the data processing. They are responsible in communicating and translating the user requirements into an examination method for VGI quality assessment. They also play a role in detailing and setting the criteria for each sub-element. It's very likely that the user and the data examiner can be different or the same stakeholder.

4.1.6 Pre-processing

Pre-processing process is designed to avoid possible issues that might emerge as well as to minimize the effort on data processing. By eliminating the unnecessary data, the following process can run efficiently. This stage can be divided into two phases: data preparation and data filtering. In the first phase, the data will be prepared before it could proceed any further. This includes the projection checking, area cropping, segmentation, or any other necessary process. The second phase is data filtering, where the unnecessary data will be eliminated by using some criteria set by the user. The filtering criteria for this stage can be derived from user requirements and also can be combined with the elements from quality modules to increase the filter's effectiveness. The quality module's element that is deployed on this stage will not be scored or weighted. However, the threshold for each quality elements should be defined.

4.1.7 Unqualified data

The unqualified data are resulted from the pre-processing because the data didn't meet the filtering criteria. This data could be send back to the contributors by enclosing some feedback comment for improvement. From the given feedback from the data examiner, the contributors are free to decide if they want to revised the data or not. Since not all the contributors of VGI data can be reached, this relation can only occasionally occur. This output is not always produced during this process, a good input data that pass all the pre-processing phase will have no unqualified data and will be led to the next process directly.

4.1.8 Quality assessment

The quality assessment is conducted to know the quality insurance of the data. In this phase, quality element and sub-element will be assembled according to the user requirements. The modules that are discussed in section 3.3 can be used depending on the data type and the case that needs to be solved. To make an easy distinction of what approaches are used in this process, the modules will be called as ISO module (the quality element taken from ISO standard) and VGI module (the quality element taken from the proposed method by researchers). These modules (section 3.3) do not have to be taken as a whole extent, but can be added or eliminated or even completely removed. Besides customizing the quality module, the users are allowed to determine the rank and importance of each element. The rank and importance then will be taken as the consideration to determine the weighting factor. Since the score for each element are the same, the rank determination will only be applied as the weighting factor. The values resulted from scoring and weighting from each element is summed and then calculated. The

calculation method is refer to the Equation 1 that mentioned in section 3.4. Result of the calculation is considered to be the confidence score.

There are also some cases that data quality assessment modules are combined with the software that is specifically dedicated for this purpose. SpaText (Massa & Campagna, 2016) and GRASS GIS modules (Brovelli, Minghini, Molinari, & Mooney, 2017) are some examples of quality assessment software that can be utilised. The implementation of the software within the framework is plausible; however some problems regarding the scoring and weighting calculation need to be taken into account because the software might not be designed to accommodate this matter. Hence, the scoring and weighting can be calculated separately or simply consider that each element has the same weighting factor. Nevertheless, the calculation of confidence score is necessary because it will be refer to determine usability recommendation.

In this quality assessment stage, data reference are necessary. Not all the quality elements are required in the data reference to run the assessment process. However by using the ISO module, this data reference is crucial for most of the quality elements, for example, to conduct the positional accuracy, thematic accuracy, logical consistency, and completeness.

4.1.9 Data reference

To control the quality of VGI data, data reference are used as the standard reference. The data can be functioned as the data comparison or simply give the brief spatial overview. In case of comparison, the data should demonstrate a top quality on all elements because it will be referred as the benchmark against the VGI data. For this purpose, besides it has to show the excellent spatial dimensions the data should also equipped with a very good data attribute. The spatial overview is used to appraise the data distribution and also to test spatial logical consistency of the data. It can be obtained if the data reference shows a good representation of seven themes of the framework data.

Types of data reference can be distinguished as; authoritative map, satellite imagery, or geo-referenced raster data. Authoritative map is the map that published officially by the government. It is the data reference that mostly used in many research as the comparison against VGI data. The satellite images that can be used as the data reference can vary in resolutions (spatial, spectral, radiometric, temporal) and vendors. To take into consideration, the usage of satellite imagery has to fit the theme, scope, and scale of the assessed VGI data. Different from satellite images, geo-referenced data is defined as the raster data obtained from scanning process and then geo-referenced. The data could be sourced from a paper map or image that converted into digital format. Though satellite imagery and geo-referenced data can be acquired from a non-authoritative body, in this case since the quality assessment is meant for updating authoritative spatial data, it should be taken from authoritative body or at least acknowledge by the authoritative body.

There are many potential data reference that can be used for quality assessment, however not all VGI data can be equally cross-referenced with the data mentioned above. For example, the VGI data generated from geo-tagged posts that have particular theme will not have comparable

authoritative data with a similar theme. Hence, the data reference still can be used to give spatial overview.

4.1.10 Confidence index

Confidence score that obtained from scoring and weighting calculation is then converted into confidence index. Referring to the Table 11 about confidence index assignment, the usability of VGI data is determined. However, the usability that resulted from this indexing process is in the form of recommendation of the appropriate usability that refers to the maximum utilisation of the data. In the end, even though the suitable usability is recommended, the user conformance adjustment can be used to finalise the usability.

4.1.11 VGI usability

The VGI usability in the authoritative scope of work was discussed on the section 3.4. The summary is as follows; data adoption, change detection, visual interpretation, semantic information, rapid response, and report alert. The usability is derived from the recommendation obtained from confidence index combined with the user conformance. Usability recommendation that was resulted from process might not meet the expectation of the user. For example, the project purpose are to adopt VGI data, but the recommendation suggests that the data can only be used as change detection. Hence, the conformance adjustment is needed to direct the usability to fit the user requirement. Based on the conformance, the user can decide if they will use the data as it is, use selected data that have high quality, or dismiss all the data. However, if it is decided to take all the data despite of its accuracy, the information about the quality must be stated in metadata.

4.2 Framework Scenarios

The advantage of the proposed framework is its flexibility. To recapitulate, the offered flexibility including; 1) quality assessment modules that can be customize; 2) flexible assembly and placement of the quality modules; 3) flexible workflow starting point, both from user as well as contributor; and 4) flexible weighting in quality assessment elements. With those flexibilities, this framework expectedly could be applied into all-type real-world cases instead of case specific framework cases.

Based on the framework described in previous sub-chapter, some scenarios are designed with the consideration of real-world cases. Three scenarios are made based on cases that often occur in the real world. The scenario will describe the workflow of the framework in a detailed manner which is customized specifically according to a particular case. Since the ideal condition is rarely found in real-world, the scenarios will simulate different circumstances that possibly happen with different perspective of framework which includes different workflow, different stakeholder, different VGI data, and different theme. Besides demonstrating the flexibility of the framework, it also shows its capability for real-world problem solving.

4.2.1 Scenario 1: VGI data for updating

The first scenario is designed based on the typical case of VGI data integration; update the existing theme of authoritative spatial data. This updating scenario can be implemented in many cases such as the road network update, landcover update, or building and infrastructure. As the illustration, in this scenario the user and the data examiner is the same and since the purpose is for updating the existing theme, the similar theme from authoritative data will be used as the data reference. There is no communication between contributor and user, consequently data revision cannot be applied in this scenario. The workflow of the scenario is shown on Figure 8.

Starting from the user that set their project goal: to update authoritative spatial data. The update in this term refers to renew the existing data as the response of dynamic change in the real world. Inferring the goal, the purpose of the project is to adopt the VGI data to get the update. The user will specify the requirements including; desirable data type, user specification, conformance adjustment, as well as the quality module they want to use for the assessment process. There are many source of VGI data related to the project's theme, the user is free to choose from one or more repository. However, if the user decides to use different repositories, the process has to be run separately from each data. By choosing the data repository, the user has built indirect relation with the contributor who provides the VGI data.

VGI data that is chosen by the user then will be proceeded to the pre-processing stage. If the project is held in specific area, the data cropping can be done in this stage as well as the projection check. Another process that can be conducted in this phase is data segmentation (if the data type is line). Since the main purpose is for updating, the current data is very important. The VGI data have to be produced during the specific period (i.e. six months) and newer than the authoritative data. Therefore the temporal accuracy is set as the filter factor. The data that do not meet the time constraint will be categorised as unqualified data.

Next stage is quality assessment process by using ISO module and VGI module. Only five ISO elements are used in this stage because one element had been use as the filter factor. However, the same element can also be used in both pre-processing and quality assessment process as long as the sub-element and criteria are clearly distinct. Authoritative data will be use as the data reference to assess the quality elements: positional accuracy, thematic accuracy, logical consistency, and completeness. The options for sub-elements that can be used in ISO module elements can refer to Table 6. Since the data will be adopted into authoritative data, the selection of sub-elements and its criteria are very important because it can affect the quality assessment result. VGI module elements that will be used include contribution index (Arsanjani, et al., 2015) and trustworthiness (Forati & Karimipour, 2016). The scoring and weighting will be calculated based on the threshold criteria from every element both in ISO and VGI modules. The intermediate result is the confidence score which then converted into confidence index to determine the recommended usability. If it turns out that the recommendation doesn't fit the purpose, the user conformance will be used to adjust the final usability.

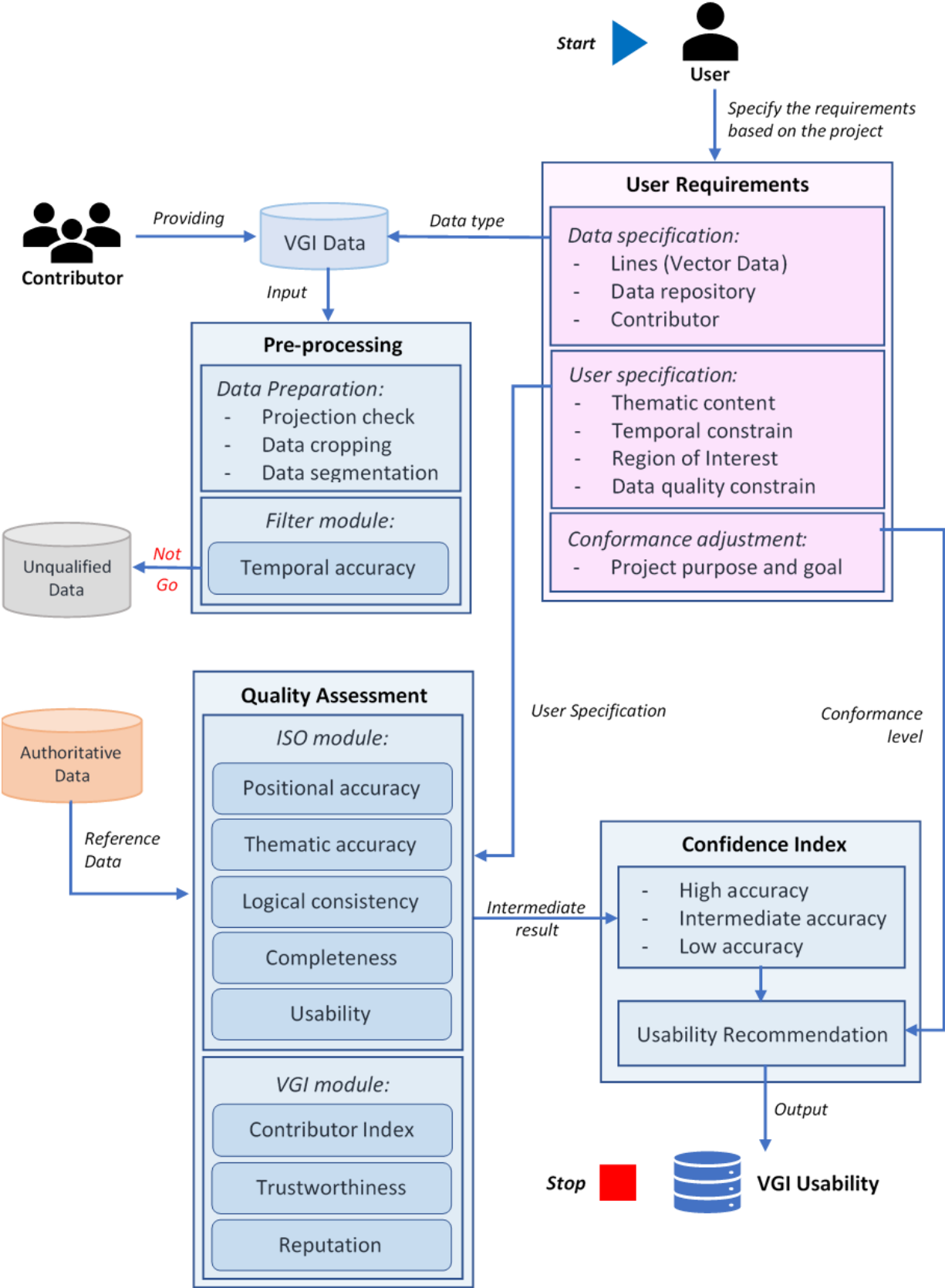


Figure 8 Workflow of scenario 1

4.2.2 Scenario 2: New theme adoption

The second scenario will simulate the condition where all relation in the framework occurs, especially the coordination between contributor and user as well as data revision. Moreover, this scenario will show that workflow of the framework doesn't always start from the user, but also can start from the contributor. In this case, the user doesn't need to specify the VGI data type since the contributor already offers a specific data type. Vector data will be used as the example to simulate the quality assessment modules. The workflow of this scenario, relation of each component, stakeholders involved, and quality element modules are shown in Figure 9. This scenario is highly possible for the VGI data that is produced through participatory mapping.

To start the scenario, contributor that carries out a VGI project sees the opportunities that their data can be utilised by the government since there is no similar authoritative data yet whilst the data is needed for the decision support. The contributor then establishes the communication with the potential user (authoritative body) to assess the possibilities of their data to be integrated with authoritative data. The user will analyse and consider the benefit and urgency of proposed data for their decision support input. If they find the merits of data utilisation, the user will notify the contributor for the preliminary requirements and necessary adoption procedures. From that notification, the contributor then provides the data and all the necessary documents to be proceed.

At the same time, the user specifies their requirements details for adopting the data. In this case, the requirements are break down in two parts: user specification and conformance adjustment. The data type is not to be specified by the user since the contributor already offers a particular type of data. Defining the user specification is subject to the user interest and need. It means that the user can specify their requirement as much as they need to. For example, this scenario divides the specification into four main aspects: thematic content, temporal constraint, Region of Interest (ROI), and data quality constrain. Thematic content is related to the specific information that the user wants to retrieve from the proposed VGI data. How detail is the desirable information can be specified in this specification. Temporal constraint refers to the specific time period. It could refer to the time when the data was produced or published. Desirable data coverage is specified by the user through a region of interest. The region of interest could be represented as administrative boundary, bounding box, or another unit area. Meanwhile, the quality constraint that specified in this step is not about the level, but more about in which aspect the quality should standout. For this scenario, the data should have excellent thematic accuracy as well as the positional accuracy. Another aspect could be added or eliminated according to user preference. The quality level is not included in this step because naturally, the user wants the highest level of data quality. However, if the user has reference for tolerable quality levels it could be included in user conformance component together with another aspect such as project purpose and goal as well as necessity of the data for the user. Based on this requirement, the criteria and quality module will be customised. The user not only can explicitly specify the quality module and criteria they want to use in this stage, but also can let the data examiner decide what quality module and criterion that is better for the assessment.

VGI data that is prepared by the contributor will be proceeded to pre-processing stage by the data examiner. The data examiner in this scenario is the same as the user. In the first phase, the projection and coordinate system of data will be checked. If they match the national standard or reference the data could go on to next phase, but if not the projection and coordinate system will be converted according to the standard. After the first phase is cleared, the filtering process will be conducted by using VGI modules; credibility, mapping protocol, and metadata. The credibility of the contributor is very important as the first impression of the quality of produced data. The credibility can be determined from the trustworthiness, reputation, or the contribution index. Beside the credibility, the mapping protocol is also very important to give brief overview of data quality. How the mapping was conducted, how they do the quality control and how they manage the data. No less important, metadata is necessary for VGI since authoritative spatial data always have standardized metadata while not all of VGI data enclosed metadata. Containing the "data about data", metadata might hold the quality assurance that provide by the contributor (De Longueville, et al., 2010; Souza, et al., 2016) as well as other important information needed for the quality assessment.

The credibility in the pre-processing process can be done automatically by measuring the trustworthiness and reputation (Forati & Karimipour, 2016) or using contribution index for OSM data (Arsanjani, et al., 2015). Meanwhile, the metadata can be assessed both manually or automatically. The manual checking will be conducted by reading all the information enclosed while the automatic process can be conducted if the VGI data have a standardized format. As proposed by Souza, et al. (2016) Dynamic Metadata for VGI - DM4VGI can be used to documented VGI metadata. DM4VGI can be automatically generated when the data are downloaded. With standardized metadata, it would be easier to conduct the assessment automatically. The mapping protocol need to be assessed manually. Since there is no well-established standard for VGI mapping, the protocol could vary for each project.

The threshold for each element will be set for filtering purpose. If the data meet the threshold criteria, it will be considered as eligible data that can be further proceeded. In contrast, the data that do not meet the criteria will be categorised as unqualified data. After conducting all the data filtering phase, the result will be overviewed. If the result of filtering phase is not satisfactory and the data ends as unqualified data, the user (or in this phase data examiner) will give the feedback to improve the data. The contributor could improve and correct their data and resubmit it or stop the process by not taking any action regarding the given feedback. However, there is a possibility that within the same process some of the data are rejected while the other accepted. In the optimistic case, the process will keep continuing only for the qualified data. In the optimistic case where all of the data pass the filtering phase, all of them will go to the quality assessment.

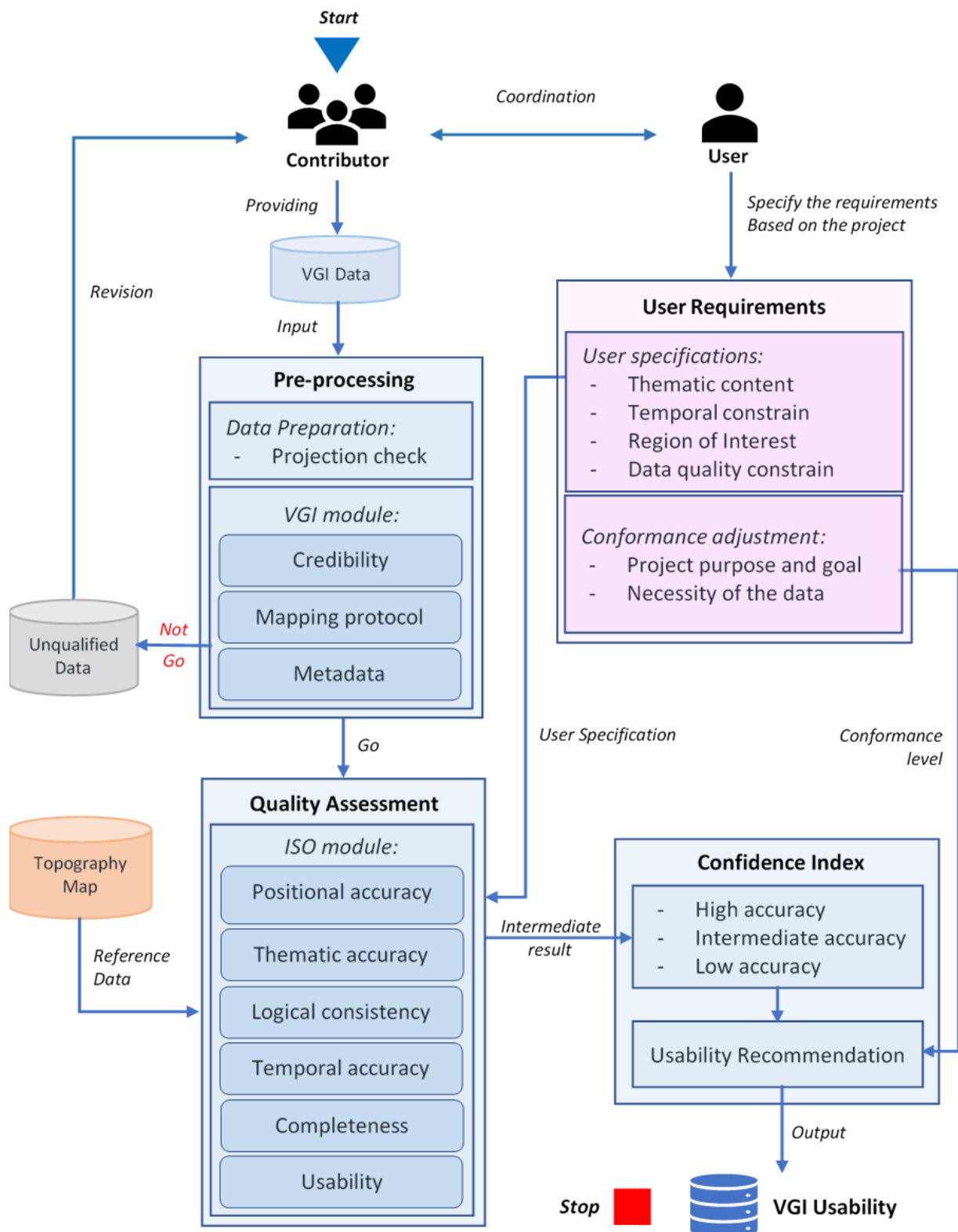


Figure 9

Workflow of scenario 2

The quality assessment in this scenario is dedicated for the vector data case, therefore the quality assessment will be conducted by emphasising the ISO elements. The module that will be used in this process can be directly specified by the user or chosen by the data examiner to meet the user specification. All of six elements will be explored in this scenario to know the quality of the data. For this purpose, quality assessment element modules that suggested in Table 6 and Table 7 can be used. Again, the number of sub-element that will be used is decided in accordance with the user requirement. To execute the process, the threshold for each sub-element in the quality modules needs to be described. For this purpose, some of the sub-elements also need a reference data as the comparison and set the threshold criteria. Since the scenario simulates the adoption of new data, there is no authoritative spatial data with the same theme that can be compared. In that case, topographical map will be chosen because it can give the overview of positional accuracy and logical consistency even though the thematic aspect could not be retrieved. After setting the threshold, the object is then scored and weighted to calculate the confidence score. The data feature that contains the confidence score then considered as the intermediate result, since the main process is done.

Based on the confidence score, the confidence index will be assigned to inform the data quality as well as recommend the usability based on it. The quality assessment process might not give the result as the user expected. In that situation the user conformance adjustment is used to decide the final usability of the VGI data. The data still can be adopted or just use it as the usability recommendation suggests.

4.2.3 Scenario 3 : Generated thematic information from VGI data

To give different overview on how the framework can be applied, the third scenario will explore different type of data for the input. The use of textual VGI, especially geo-tagged posts will be the model on how the specific thematic information can be generated from VGI data. There are many potential implementations of this scenario such as: mapping the image of public transportation from the user point of view so that the service can be improved; mapping the electability of president candidate to predict the result of the election; and mapping the affected area from a disaster event. All of those examples can be used as the decision support for the government. Scenario 3 are illustrated in Figure 10.

The user initiates a project related to specific case which needs strong aspect not just spatially but also thematically. Sometimes, the thematic content could not be synthesized from the existing authoritative data, hence the VGI are considered as the potential source for this purpose. Based on the purpose of the project, the user requirement can be described. Regarding the data type, geotagged posts are chosen as the data type since the thematic aspect is highlighted in this project. Defining the data repositories is also important, since each data repositories produce unique data. The user specification will put more constraints in the thematic aspect. This includes identifying the possible words that are used in hashtag that later on will be used as the thematic filter factor. Events happen in a specific time window, therefore setting the time constraint is also important. Besides thematic and time constraint, defining the region of

interest is also important to narrow down the analysis coverage. Finally, the user conformance is also specified by the user in this stage.

After setting all the user requirements, the selected VGI data that is produced by the contributors proceeds to the pre-processing stage. Data preparation phase for this case includes projection checking and data cropping based on the region of interest. Following the data preparation, filtering phase are conducted by using three factors: positional accuracy, thematic accuracy and temporal accuracy. Positional accuracy will separate the geotagged data and non-geotagged data. The data that are geo-tagged then can continued to the next filtering process. As mentioned before, the thematic accuracy will be generated from the hashtag and specific words contained in the posted status. Determination of time period as the temporal accuracy aspect is depends on the project, some might need longer time period while another demands the current data. The longer the data generation period the larger the data size of VGI data resulted. Therefore data filtering is essential to reduce the amount of data. Because this type of data tends to have a large amount, the unqualified data resulted might also larger than the data that can go to the next process. In this scenario, the unqualified data are produced from each filtering steps. Filtering process in this scenario is strongly recommended to be conducted sequentially to help lessen the processing effort. The qualified data then proceeds to the quality assessment process.

Three of the ISO elements are used as the filter in the previous stage, therefore the ISO module in this stage only contains; logical consistency, completeness, and usability. The topographic map will be used as the data reference to give the spatial overview related to logical consistency and completeness aspect of the data. Spatial aspect is not the only consideration in determining the logical consistency and completeness, but another sub-element related to the thematic content and social-economic condition is also possible. Nevertheless, when VGI module is added, choosing the sub-elements for each module should be considered very carefully so that there are no overlapping element criteria between ISO module and VGI module. However, since the VGI module that is used in this scenario are the contributors' indicator as well as demographic and socio-economic, the clear distinction on criteria have to be set to avoid deficiency. The scoring and weighting that are also carried out in this stage are based on the predefined threshold.

Confidence score resulted from the scoring and weighting calculation will be translated to confidence index to determine the usability recommendation. Since the thematic aspect is the bigger consent for the project in this scenario, the user might set the conformance level to accept the intermediate data quality as long as it contains the desirable thematic aspect. Moreover, the data could be combined with the authoritative data but only as the data attribute.

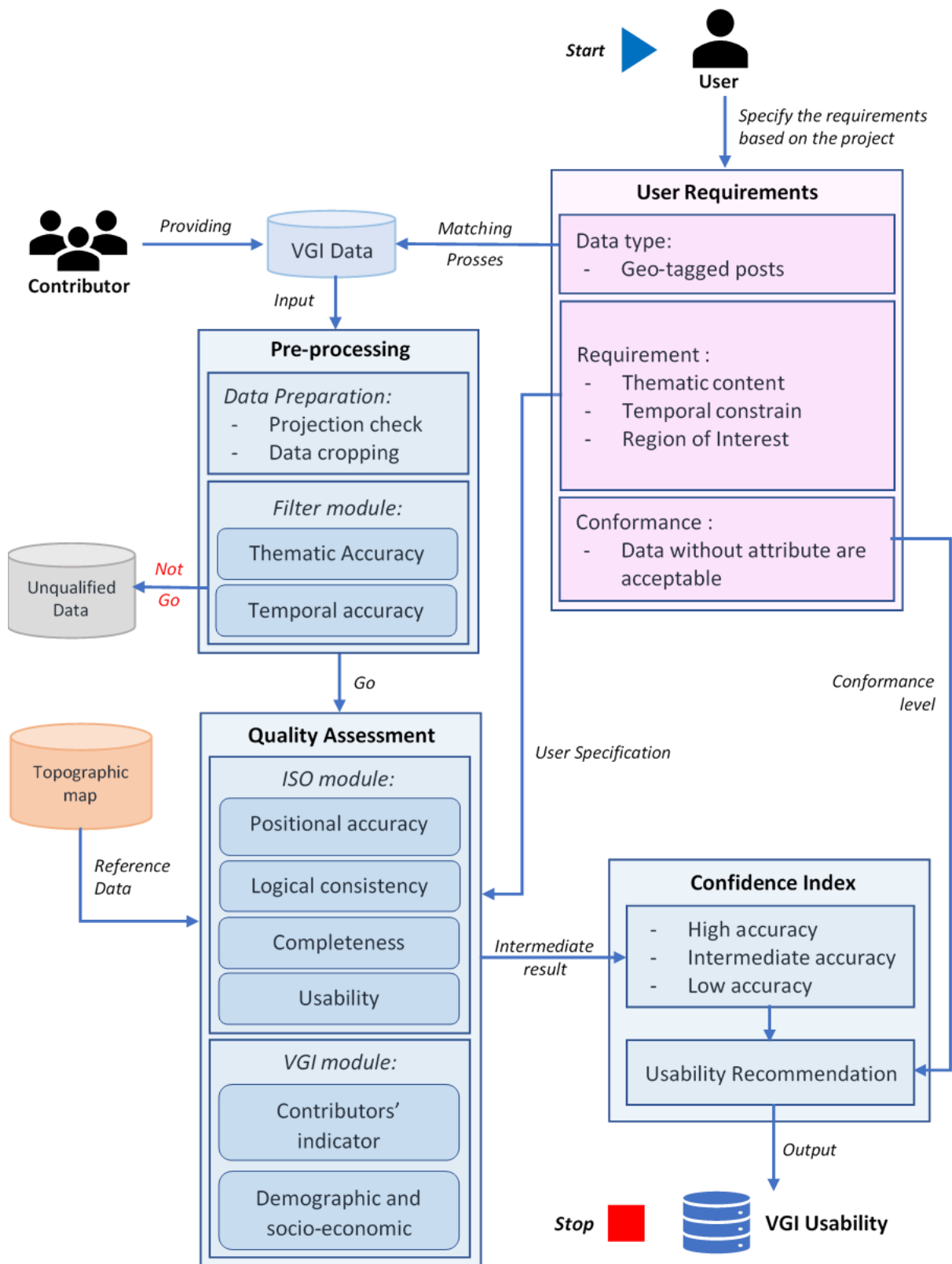


Figure 10 Workflow of scenario 3

5 Implementation

The implementation will be conducted to test the feasibility of the framework scenario. As the most typical case of authoritative and VGI data combination, the first scenario will be taken as the exemplification to demonstrate the applicability of the proposed framework scenario. Updating Indonesian authoritative topographic map will be carried out to represent the real case in the real world. *Peta Rupa Bumi Indonesia*, or usually shortened to RBI, is an authoritative topographic map that published by Indonesian NMA, *Badan Informasi Geospasial* (BIG). It covers all Indonesian territory and is available in various scales. However, as a growing country with the massive national scale development, the necessity of updated spatial data to support this development are in high demand. Unfortunately, the delay of data updating is inevitable with the limitation of the government on resource and funding. Although there are periodic updating schedules for RBI (Table 12), since Indonesian government is more focussed on providing complete coverage and multiscale map rather than updating the existing map, it creates even a bigger gap to provide an updated data. Therefore, the combination of RBI and VGI data can be a potential solution to fill the gap. This implementation will examine such combination especially in how the updating procedure could enhance the updating cycle in terms of automation, data quality assessment, and frequency.

Table 12 RBI updating period (Perka BIG No.14 Th.2013)

RBI Category	Scale	Update Period (year)	
		Minimum	Maximum
Small scale	1:1.000.000; 1:500.000; 1:250.000	10	25
Medium scale	1:100.000; 1:50.000; 1:25.000	5	15
Big scale	1:10.000; 1:5.000; 1:2.500; 1:1.000	1	10

The implementation is conduct by taking some part (around 27 km²) of Depok city as the case study area. Located in the southern border of Jakarta, the development of Depok city are massively influenced by this megacity. Many new roads, settlement area, and other infrastructures are growing significantly. The latest topographic map with scale 1:5000 for this area were launched in 2016, but the data generated from satellite imagery that obtained in 2014. Technically, in 2018, the map is already 4 years old. According to the updating period table, the map had passed the minimum period (1 year) but has not yet reached the maximum period (10 years). However, with the rapid development that affects the land use and land cover change the updating is necessary to be executed.

5.1 Updating RBI using OSM data

The goal of the project is to update the RBI map, specifically the road feature. Based on this purpose, NMA as the user needs to specify the necessary requirements related to the updating process. First of all, the data specifications are described by the user (NMA). Since the updating

process will be conducted for road feature, the desirable data type is lines vector data so that it can be easily integrated with the authoritative data. Normally, these kinds of data are available and can be obtained online from various data repositories. However, unlike developed country, the option for data repositories in Indonesia might be limited. Thus, the data repository is preferably a well-established VGI project site such as OSM. Second step is to define the user specification that related to the usage, purpose, and constrain of the project. The purpose of the project is to partially update RBI data, which includes the adoption of new road features as well as its name from VGI data. Data updating process requires the latest released data, hence the time constraint related to the data release shall be at least in early 2018. The selected area (ROI) to perform the updating process is set by the user by giving the bounding box. The conformance adjustment is set by specifying the alternatives usability of the VGI data instead of updating it (i.e. change detection).

Referring to the data specification, the VGI data are fetched from the data repository. For OSM data, besides downloading the data directly from the official OSM sites it also possible to do so in the mirroring sites that offer the similar service. But, the downloaded data might be slightly different because of different data pre-processing and API that is they used. In this case, the OSM data are downloaded from BBBike¹⁷ OSM extract service. Unfortunately, the 'last-change' information is not included in the downloaded data. However, the time constraint is still relevant based on the time of the data download: April, 16th 2018. The advantage of downloading data via BBBike is that the data is already processed into shapefile format so that it can be directly used as the data input in the framework. All the processing activity of the framework will be conducted in ArcGIS environment.

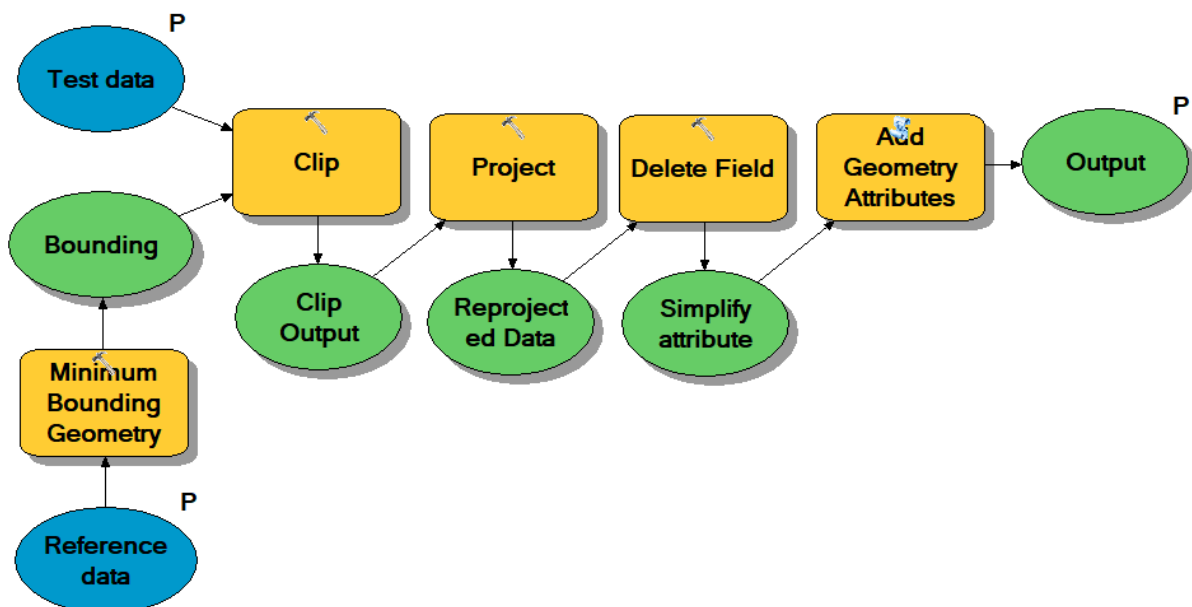


Figure 11 Pre-processing model

¹⁷ <https://extract.bbbike.org/> (accessed 04.09.2018)

Pre-processing stage includes two phases: the data preparation and filtering. Data preparation conducted in this phase are; reprojection, data cropping, and length calculation. The reprojection is aimed to make the OSM data has the same projection with RBI so that the projection conflict can be avoided in the following process. To fit the processing area, OSM data are cropped with the given bounding box. As the final step in this data preparation phase, geometric length for each segment are calculated. The length calculation is done by assuming that the processed data are not stored as geodatabase, if it is stored as geodatabase the length will be automatically calculated when the data is imported. Data cropping and length calculation is subjected to the input data condition therefore the process is not mandatory. Figure 11 shows the pre-processing model, where it can be modified based on tested data condition.

The next phase is filtering process by using the filtering module. The filtering process is intended to sort out the relevant and irrelevant information in this updating purpose. Thematic filter module in this step includes synchronisation of road classification and deletion of unnecessary information. The RBI and OSM data obviously have different classification system, therefore synchronisation and reclassification of OSM and RBI road type are essential. The synchronisation process is carried out by finding the equivalent description of OSM class in the RBI standard document, then the OSM will be reclassified based on RBI classification. Reclassification is only conducted for the class that is found in the test area. Description of RBI classification can be seen in Table 13, while description of OSM classification can be seen in Table 14.

Based on each class definition, the OSM class is assigned to RBI class that has the equivalent meaning. The reclassification result can be seen in Next step is deletion of unnecessary information which includes the removal of irrelevant feature or unnecessary attribute. The excess information fields that will not be used in the next stage such as 'oneway', 'bridge', and 'maxspeed' will be deleted. Deletion of feature is not necessary since the OSM data are well classified and can be used as input in the next process.

Table 13 Definition of RBI road classification (BIG, 2005)

RBI Classification	Definition
Arterial Road (<i>Jalan Arteri</i>)	Public roads that serve the main transportation which have characteristics of long distance travel and high average speed. (UU No. 23, 2004; BIG, 2005)
Collector Road (<i>Jalan Kolektor</i>)	Public roads that serve transportation which have characteristics of moderate distance travel and moderate average speed. (UU No. 23, 2004; BIG, 2005)
Local Road (<i>Jalan Lokal</i>)	Public roads that serve local transportation which have characteristics of short distance travel and low average speed. (UU No. 23, 2004; BIG, 2005)
Unclassified (<i>Jalan Lain</i>)	Roads that does not include in any existing classifications of roads. (BIG, 2005)

RBI Classification	Definition
Path (<i>Jalan Setapak</i>)	Pedestrian roads that link villages to another areas or paths in mountainous areas. (BIG, 2005)
Embankment (<i>Pematang</i>)	Small elevated path that is usually found in rice fields or marshy fields (<i>Kamus Besar Bahasa Indonesia</i> (KBBI), 2018).

Table 14 Definition of OSM road classification (Wiki, 2018)

OSM Classification	Definition
Trunk	The most important roads in a country's system that aren't motorways. (Not necessarily a divided highway.)
Trunk-link	The link roads (sliproads/ramps) leading to/from a trunk road from/to a trunk road or lower class highway.
Secondary	The next most important roads in a country's system. (Often link larger towns.)
Tertiary	The next most important roads in a country's system. (Often link smaller towns and villages)
Unclassified	The least important through roads in a country's system – i.e. minor roads of a lower classification than tertiary, but which serve a purpose other than access to properties. (Often link villages and hamlets.)
Residential	Roads which serve as an access to housing, without function of connecting settlements. Often lined with housing.
Service	For access roads to, or within an industrial estate, camp site, business park, car park etc.
Track	A track provides a route that is separated from traffic. It could be referred to bike lanes that are separated from lanes for cars by pavement buffers, bollards, parking lanes, and curbs.
Cycleway	Road that is designated for cycleways.
Bridleway	Road for horses.
Path	A non-specific path. It's mainly for walkers, also used for cyclists, horses, as well as walkers and passable by agriculture or similar vehicles.

The OSM data that pass the pre-processing stage will be further proceeded to the quality assessment process. The quality assessment process will only use ISO module since the downloaded OSM data attribute is limited. Completeness, thematic accuracy, position accuracy, logical consistency, and usability are used as the quality elements for this process. Those elements will be divided into sub-elements and weighted to determine the data usability recommendation. Table 16 describes the detail of scoring and weighting for each sub element that will be used in quality assessment process. The score of element will always count as +1, so if one

elements contained several sub-elements the value will be divided accordingly for each sub-element. In contrast with the score, the weight value is different for each element depend on the user specification. Referring to the user specification, the position and thematic aspect are valued more than other elements consequently it weighs more than other elements. Finally, all the weighted score from each element will be calculated using Equation 1. The result will be stored in the new fields as the usability recommendation per segments.

Table 15 Classification conversion of OSM

Feature Code	RBI Classification	OSM Classification
CA008008	Arterial Road (<i>Jalan Arteri</i>)	Trunk, trunk-link
CA008010	Collector Road (<i>Jalan Kolektor</i>)	Secondary
CA008012	Local Road (<i>Jalan Lokal</i>)	Tertiary
CA008014	Unclassified (<i>Jalan Lain</i>)	Unclassified, residential, cycleway, service, track, bridleway
CA008016	Path (<i>Jalan Setapak</i>)	Path
CA008020	Embankment (<i>Pematang</i>)	No comparable class

Table 16 Quality element score and weighting factor

Quality Elements	Sub-elements	Criteria	Score	Weight
Completeness	Attributes:			1
	Type	Existence	0,35	
	Name	Existence	0,35	
	Total length difference	\geq RBI length	0,30	
Thematic accuracy	Classification similarity	Similar	0,25	2
	Classification accuracy	Overall accuracy $\geq 85\%$	0,25	
	Road name	Existence	0,25	
	Additional road name	Existence	0,25	
Position accuracy	Proximity	Intersect in buffer area	1	2
Logical consistency	Intra-theme consistency	Topology check	0,50	1
	Inter-theme consistency	Non-intersect segments	0,50	
Usability	Usage, purpose, constraint	High positional and thematic accuracy	1	1

Completeness is divide into two sub-elements; the completeness of attribute and the feature completeness. The reference data are needed to conduct the assessment. The completeness of attribute is intended to check whether there is empty attribute in the OSM data and then compare it with the reference data. These comparisons are only conducted for the 'type' and 'name' attribute of the feature. Meanwhile, the feature completeness will compare the RBI and OSM

data in terms of total length difference. Statistic calculation are used to define the total length difference between two data. The valuation of completeness is based on the three sub-parameters: name attribute (+0,35), type attribute (+0,35), and total length (+0,30). From the attribute parameter the value will be added based on the attribute existence. Meanwhile, since the length is compared in total instead of segment pair, all the segments will be given the same value that concluded from the total length comparison. If the total length is equal or exceeds the RBI total length, +30 will be given to each segment.

Although thematic element is already used in the pre-processing stage, but it is also used again in the processing stage. Nevertheless, the accuracy sub-elements used and data treatments are different in each stage. This is one of the advantages from using a flexible framework, so that the parameters can be implemented according to the user needs. The thematic accuracy elements will be examined in two main aspects; road classification (classification similarity & classification accuracy) and road name (existence and additional name). Before conducting the thematic accuracy, nearest object analysis is conducted to find the comparable segment in RBI dan OSM. The 'Near' tool from ArcGIS is used for this analysis. The algorithm calculates the nearest distance from each of the end vertices of OSM to the near RBI segment (ArcGIS for Desktop, 2018). The nearest object analysis were identified RBI segments' ID that close to OSM. It is important to note, that the near algorithm in ArcGIS calculate the nearest distance from each of the end vertices of the input segment to the near segment. After the nearest segments were found, based on the RBI segments' ID, the table join is conducted especially for the feature code and object name fields. Both feature code and object name are then compared and analysed, when the difference is identified, the value is given and stored in the new attribute field. The nearest object analysis process is shown in Figure 12.

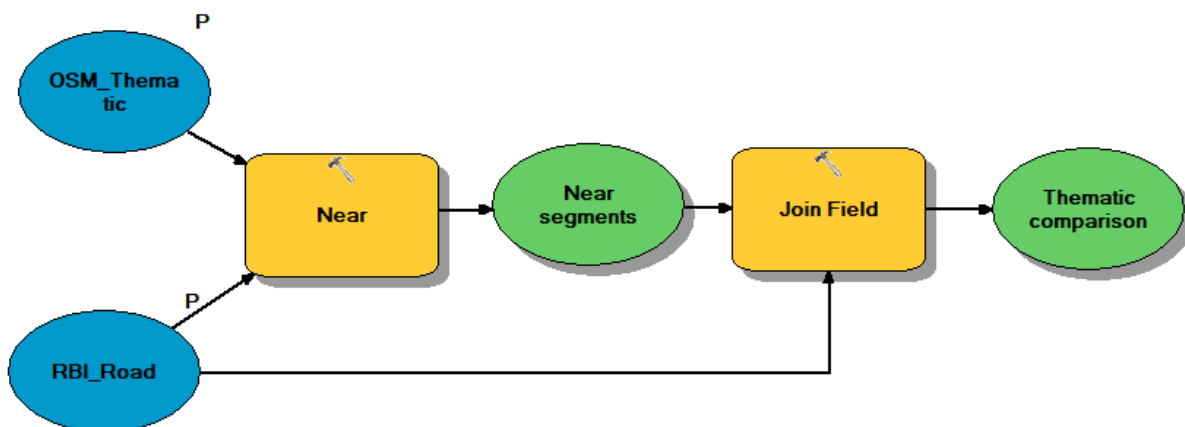


Figure 12 Nearest object analysis model

Comparing the road classification between OSM and RBI is meant to test how reliable the road type information from OSM is. Classification similarity is identified by comparing RBI and OSM feature code per segments, in which matched feature code valued +0.25. As a complement, the

confusion matrix is built to know the overall classification accuracy. If the overall accuracy result reaches 80%, all the segment will be valued +0.25. The road name is only identified based on its existence despite the information correctness, therefore every segment with name attribute will be given value of +0.25. Not all of RBI road feature is provided with the name attribute, in this case it would be an extra advantage if OSM can be used to fill it. Thus, additional value of +0.25 will be given to OSM segment that can give contribution to RBI name attribute.

Position accuracy is calculated based on the intersection between RBI and OSM. Considering the accuracy of handheld GPS and as one of the data collection method in OSM, 5-meters buffer area is created from RBI. The 'buffer', 'intersect', and 'symmetrical difference' tools are used in this process. The symmetrical difference tool is used to reconstruct the OSM data as the original data before intersection process. The proximity model can be seen in Figure 13. The proximity between OSM feature and RBI will be decided based on the intersection of OSM feature within RBI buffer area. Based on the proximity, the value will be given to each feature; 0 for non-intersection objects and +1 for intersect objects.

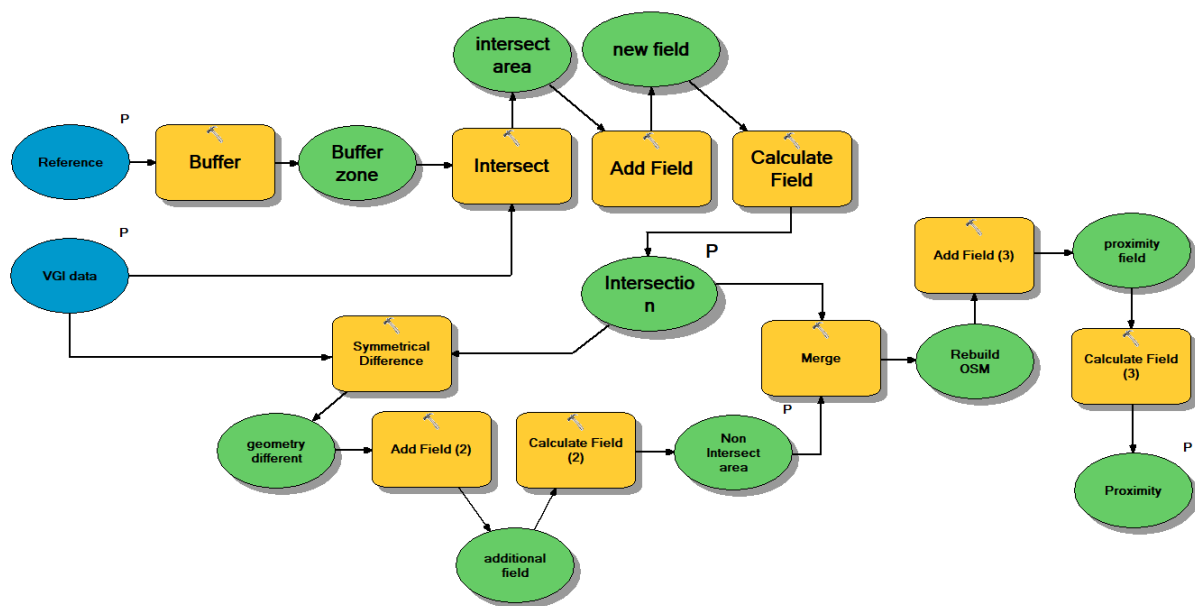


Figure 13 Proximity model

Logical consistency of a data can be examined using four sub-elements: conceptual, domain, format, and topological accuracy. For this scenario, only the topological consistency sub-element will be used. Spatial relationship of each object within the database (intra-theme consistency) and another theme (inter-theme consistency) will be tested in this step. Intra-theme consistency is tested by using topology tools with two rules; must not have dangles and must not overlap among the objects (Figure 14). Meanwhile to perform the inter-theme consistency, additional thematic layer from RBI will be used. Land cover features, especially the buildings are used to test the inter-theme consistency. Spatial integrity constrain between those two features

is that the road should not cross the buildings. Intersect tool will be used to perform this inter-theme consistency.

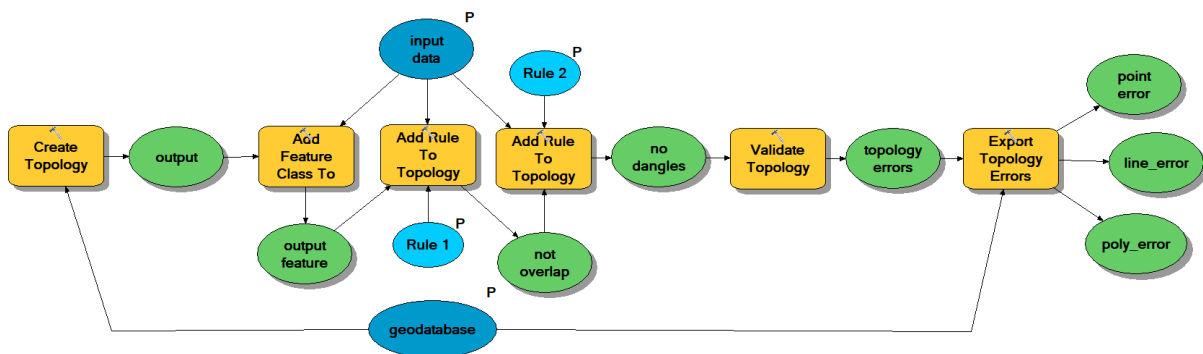


Figure 14 Intra-theme logical consistency model

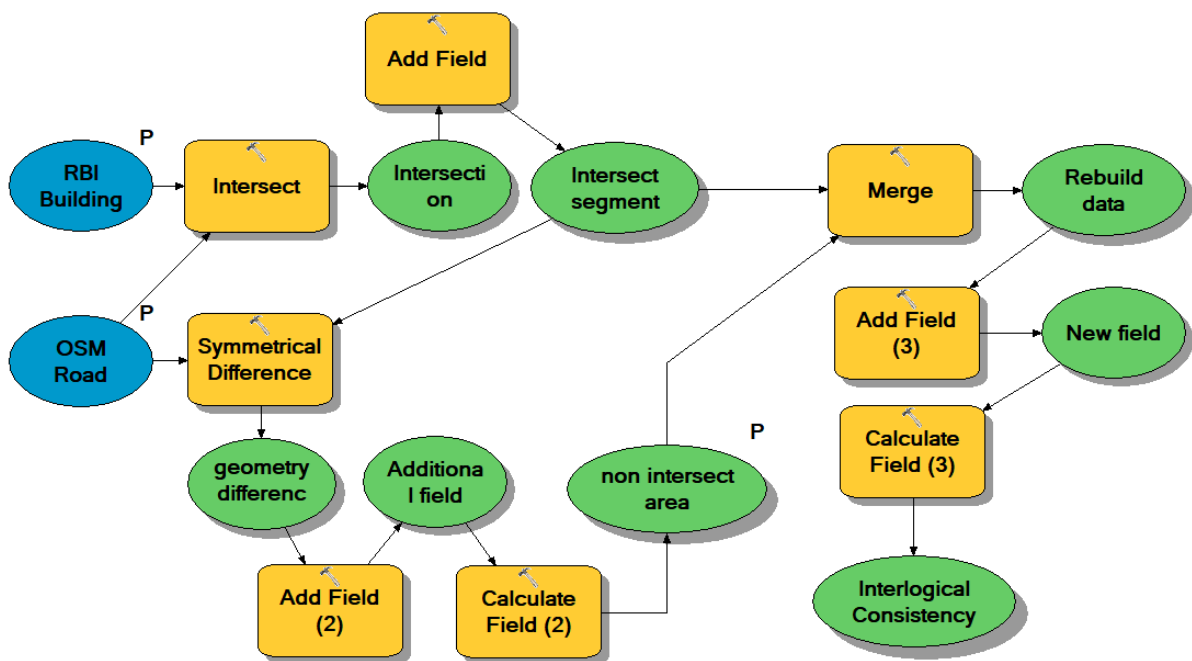


Figure 15 Inter-theme logical consistency model

Finally, as the summary of all quality assessment process above, the value obtained from previous process is compared with the usage, purpose, and constraint of the project. The data will be used to update the authoritative map. Hence, the data must have high accuracy especially in position accuracy and thematic information so that the data can be adopted into authoritative spatial data. Since the constraint set by the user is only related to release time and is already used as a parameter when looking for data, this sub-element will not include in this stage anymore. The abstraction of the whole updating scenario can be seen in Figure 16.

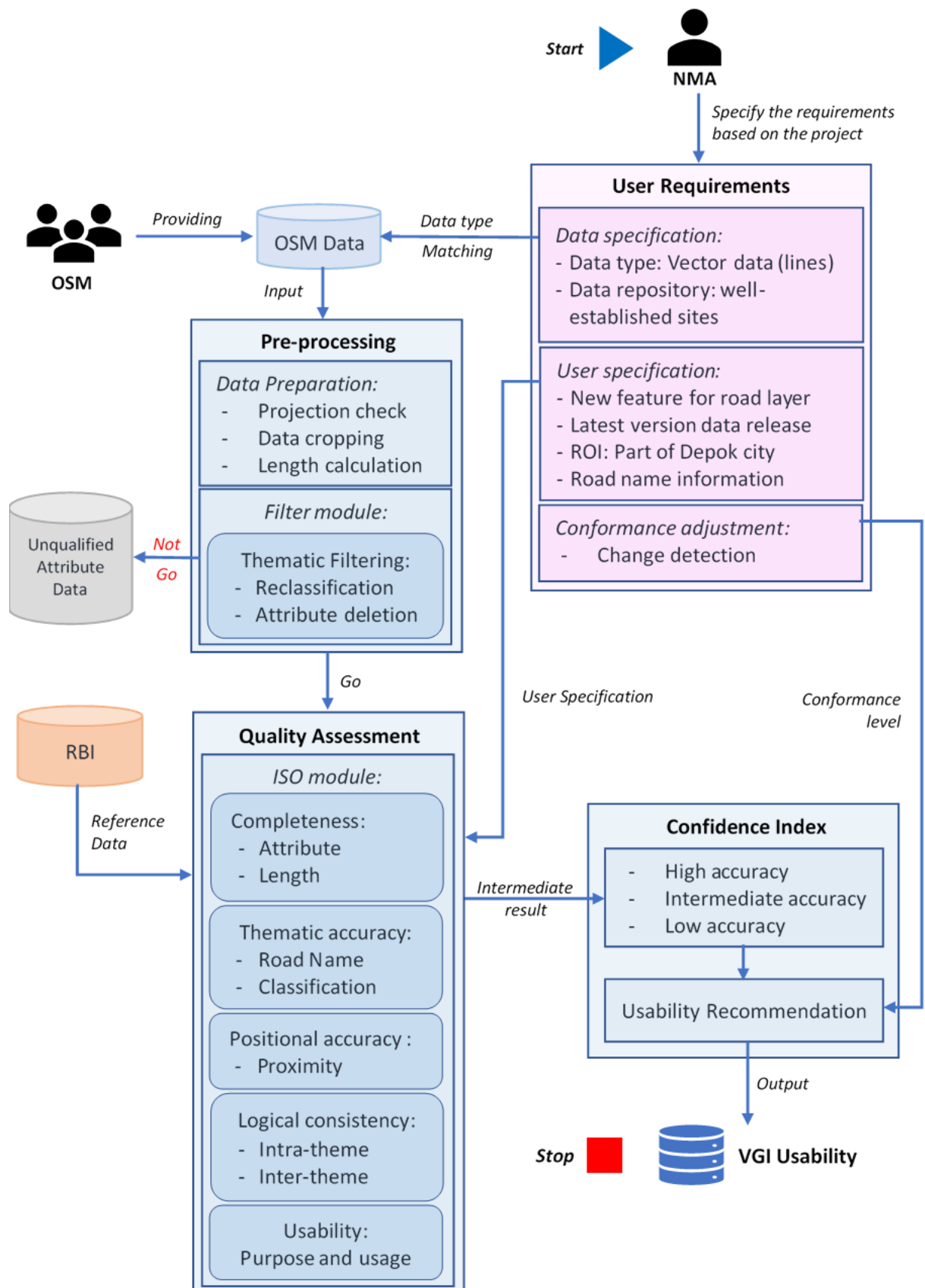


Figure 16 Data updating scenario

5.2 Implementation Result

The most important parts in pre-processing data in this updating case are data reprojection and reclassification of attribute. Those two processes are the key to conduct the next processing. Data projection is vital since all the analysis will be conducted in lengthwise approach. If the projection between OSM and RBI is different, the analysis will become invalid. Meanwhile, class reclassification is essential for checking the thematic accuracy. The reclassification is written as feature code instead of feature name for the attribute simplicity and then stored in new field. In the sample area, there is a RBI class (CA008020) that do not have any equivalent description in OSM type. Whereas, about 30% of RBI data are classified in this class (Figure 17) and most of it is not mapped in the OSM data. Although it does not have equal comparison, this class will still be included in the following process since it is assumed as reference for the correct classification. The result of reclassification statistic is presented in Table 17. Before the reclassification, the dominant road type is residential road which takes 87,9% of total OSM road length. The high number might be caused by condition of the sample area which is apparently dominated by residential area. Meanwhile, the shortest road is bridleway with only one segment. Typically, bridleway in urban area is rarely found. Moreover, its only has one segment and the length are too short making the suspicion of misclassification becomes more intense. Therefore, the bridleway is traced in the Bing satellite imagery to clarify the suspicion. Apparently, the anomaly is found since the bridleway road is laid among the resident area, which is not commonly found in Indonesia (Figure 18). This could be the indication of misclassification. Compared to RBI data, there is no similar segment found in those position, while the surrounding area were classified as local road (CA008012). However, it could not yet be decided whether the classification is wrong because the reference data has no comparable feature, ground truthing need to be conduct to prove the hypothesis. There is no action taken regarding this finding, but it will be taken as the consideration to the future research especially to validate the classification. The OSM data that already projected and reclassified then proceed to the quality assessment process.

Completeness of OSM data is observed from two aspects: attribute completeness (name and type) and total length difference compared to RBI data. Table 18 shows the assessment result of OSM data without comparing it to RBI data yet. In terms of road classification, all features in OSM data are attributed completely. In contrast, the road name only covers 3,7% of OSM data if it is calculated based on the numbers of feature records. The percentage is increased up to 12,6% if calculated by lengthwise. However, those numbers are still considered very low and even lower if compared to the RBI data. The comparison of RBI and OSM data can be seen in Table 19. The comparison is conducted by lengthwise instead of segment-wise because segment in RBI and OSM cannot be compared equally. While digitizing a road segment, the contributor might create it as two different segments or even combine two road segments as one segment. Thus, the analysis is conducted based on the length. However, the valuation of completeness will still be embedded in each feature records (segment-wise).

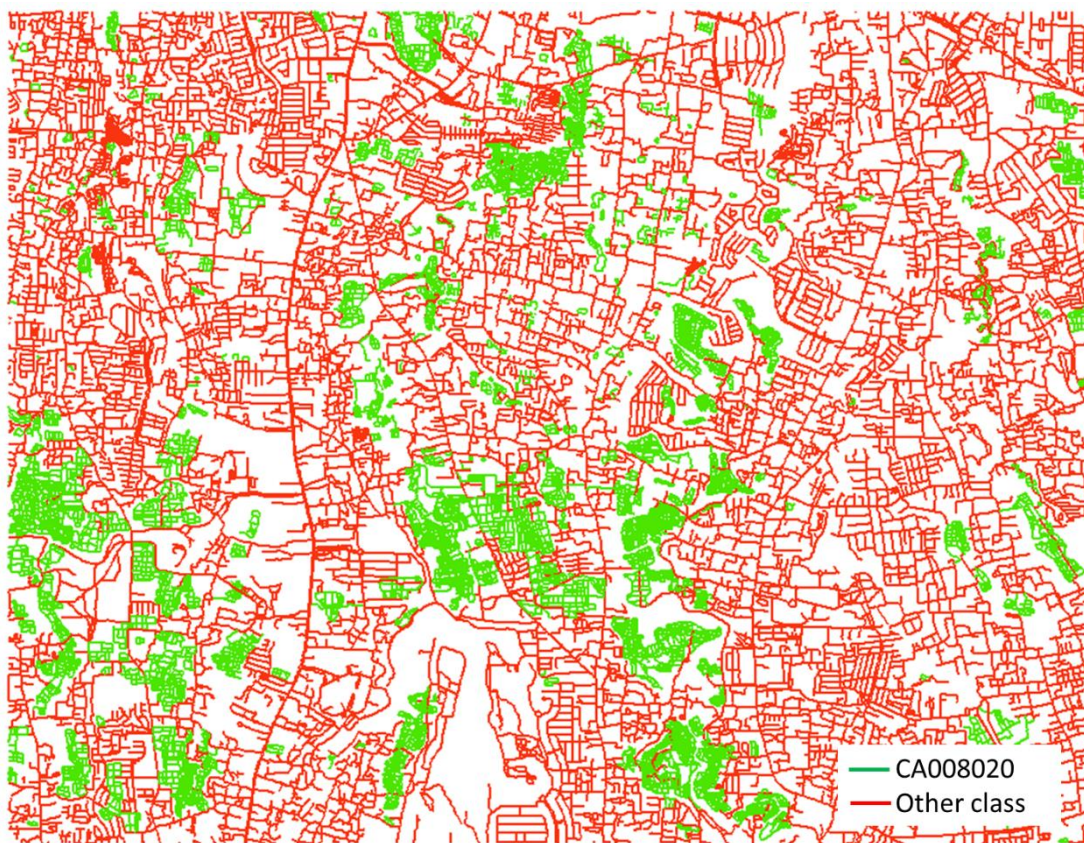


Figure 17 Ratio of CA008020 to other RBI Road classes

Table 17 Road length based on the classification

Feature Code	OSM Reclassification	Length (Km)	OSM Classification	Length (Km)
CA008008	Arterial Road	9,56	Trunk	9,55
			Trunk-link	0,014
CA008010	Collector Road	9,29	Secondary	9,29
CA008012	Local Road	4,34	Tertiary	4,34
CA008014	Unclassified	508,48	Unclassified	15,68
			Residential	468,44
			Service	16,84
			Cycleway	2,88
			Track	4,54
			Bridleway	0,095
CA008016	Path	1,16	Path	1,16
Total		532,83	Total	532,83



Figure 18 Bridleway road overlaid in Bing satellite imagery

Table 18 Attribute completeness of OSM data

Attribute	Empty Attribute	Filled Attribute	OSM Completeness
Type/class	0	4854	100%
Road name (segments)	4673	181	3,7%
Road name (km)	465,46	67,37	12,6%

Table 19 Attribute comparison of RBI and OSM

Parameter	RBI	OSM	Difference	OSM Completeness
Segments	19662	4854	14808	-
Total length (Km)	953,79	532,83	420,96	55,86%
Type/class Attribute (Km)	953,79	532,83	420,96	55,86%
Name Attribute (Km)	171,20	67,37	103,83	39,35%

Based on the completeness analysis, RBI is still superior to OSM data in all aspect. The total length of OSM is only 55.86% from RBI. Its means that even though RBI is created with 2014 data, the data completeness of OSM that is downloaded in 2018 is still behind the authoritative data. Nevertheless, if observed from the data distribution, some area that are not covered by RBI are covered by OSM. The difference can be seen from Figure 19 where the blue line represents the OSM data and red line represents RBI data. The blue line shows the area that is mapped by OSM but not yet mapped by RBI. Since the data is only overlaid each other, there

are possibilities that some data might only show the wrong geometry instead of the new road feature that is not mapped yet. However, it still can be used as the indicator of occurring changes in certain area.



Figure 19 Comparison of RBI and OSM data distribution

Regarding the road name, only 17,9% of RBI roads are accompanied with the name attribution. Most of the unnamed road were come from the unclassified class downward, while all the arterial, collector road, and half of the local road are well defined. Compared to OSM data, RBI limitation seems still superior in total length of named road. However, there are possibilities that there is unnamed road in RBI but is named in OSM. Furthermore, those possibilities will be considered in the thematic accuracy assessment. Overall, none of the segments get the perfect value of +1 and only 181 segments get +0.70 while the rest of segments get value of +0.30 in completeness element.

Thematic accuracy assessment result also shows the similar trend to completeness assessment. Based on the nearest analysis (Table 20), only 26% (1245 segments or 138,55 km) has similar classification as RBI, while road name attribute existence is only 12,6% (181 segments or 67,37 km) and additional road name that can be potentially added to RBI is only 5.8% (83 segments or 31,41 km). The result shows that apparently thematic aspect from OSM data are also weak. Furthermore, to complete the thematic accuracy assessment, how accurate the OSM contributor classified the data are examined through the confusion matrix (Table 21). According to the

result, the overall accuracy for OSM data is very low (25,8%) which can be inferred that the misclassification often occurs within the data. Though there is a road classification protocol in OSM, it might not well be understood by the local contributor since the description and picture provided in the protocol refer to the European road condition which might not well represent the Indonesian road condition. Consequently, it can lead to the misclassification of road object. No equivalent class for CA008020 (embankment/*Pematang*) in OSM data could also be the reason why the overall classification is low. However, the highest similar classification is found in the unclassified class. Almost most of the OSM segments belong to the CA008014 class (unclassified/*Jalan lain*) because the RBI class does not provide the specific description as the OSM does. Nevertheless, the highest producer accuracy falls in this class with the percentage of 99,4%, but for the user accuracy the highest accuracy is achieved by CA008012 (Local road/*Jalan lokal*) with the percentage of 75%.

Table 20 Thematic attribute analysis result

Attribute	Number of Segments	Length (Km)	Percentage
Classification similarity	1245	138,55	26%
Road name existence	181	67,37	12,6%
Additional road name	83	31,41	5,8%

Table 21 Confusion matrix of OSM Classification

		RBI Data						Row total	User Accuracy
		...8008	...8010	...8012	...8014	...8016	...8020		
OSM Data	...8008	5	0	2	1	0	0	8	62.5%
	...8010	0	3	4	4	0	0	11	27.3%
	...8012	1	0	3	0	0	0	4	75%
	...8014	39	59	3223	1240	72	189	4822	25.7%
	...8016	0	0	6	3	0	0	9	0%
	...8020	0	0	0	0	0	0	0	0%
Column total		45	62	3238	1248	72	189	4854	
Producer Accuracy		11%	4.8%	0,1%	99,4%	0%	0%	Overall Accuracy: 25,8%	

The lack of OSM data completeness, both in the geometry and attribute aspect, might be caused by the absence of an RBI classification class in OSM. With the unequal classification, in this case CA008020, the mappable features are skipped because of no OSM classification can accommodate the feature. Whereas, the missing feature class is important to Indonesian government for management purpose. Nevertheless, this classification issue needs to be further

explored to find the proper solution. The options could be to adjust the RBI class and OSM class or to propose suitable OSM classification according to the Indonesian specific condition.

During the positional accuracy assessment process, the OSM segments will be break down based on the intersection area. If one segment only partly intersects, then it will be divided into separate segments. Total new segments resulted from this process are 7696 segments. Although there are 2842 additional segments, the total length of OSM data remain the same. The intersect part of the segments gets the value of +1 while the other part get 0 value. Because of the segmentation reason, positional accuracy assessment is conducted after the completeness and thematic assessment. Using 5 meters as the buffer area, the result of intersecting segments is quite good with the total length of 427,15 km or about 80% of OSM data. A subset of tested area in Figure 20 shows the buffering area of RBI and its intersection with OSM data. Though most of the OSM data are intersected within the buffer radius, overall if compared to the total length of RBI the intersect segments are only 44,78%.

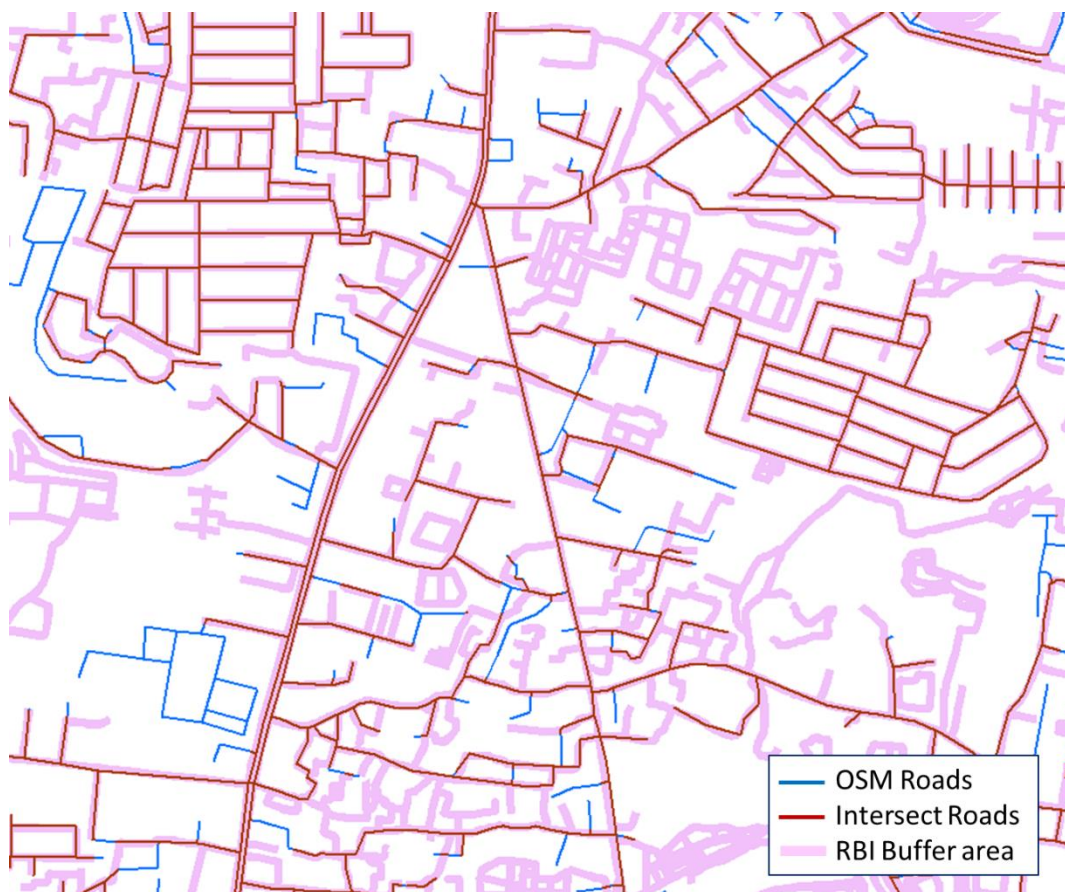


Figure 20 Subset area of OSM and RBI buffer intersection

To break down the intersection and non-intersection segments, Figure 21 and Figure 22 show the length proportion of each classification. Residential road always shows striking difference towards another class since in the beginning it already takes bigger proportion in total OSM data. Focussing in the non-intersection segments of residential road, there are two possibilities

applied in the condition: it could be indication of new road occurrence or the segment has wrong geometry. Those condition are shown in Figure 23 where the yellow circles are highlighted exemplifications of the additional road segments (a) and geometry error (b).

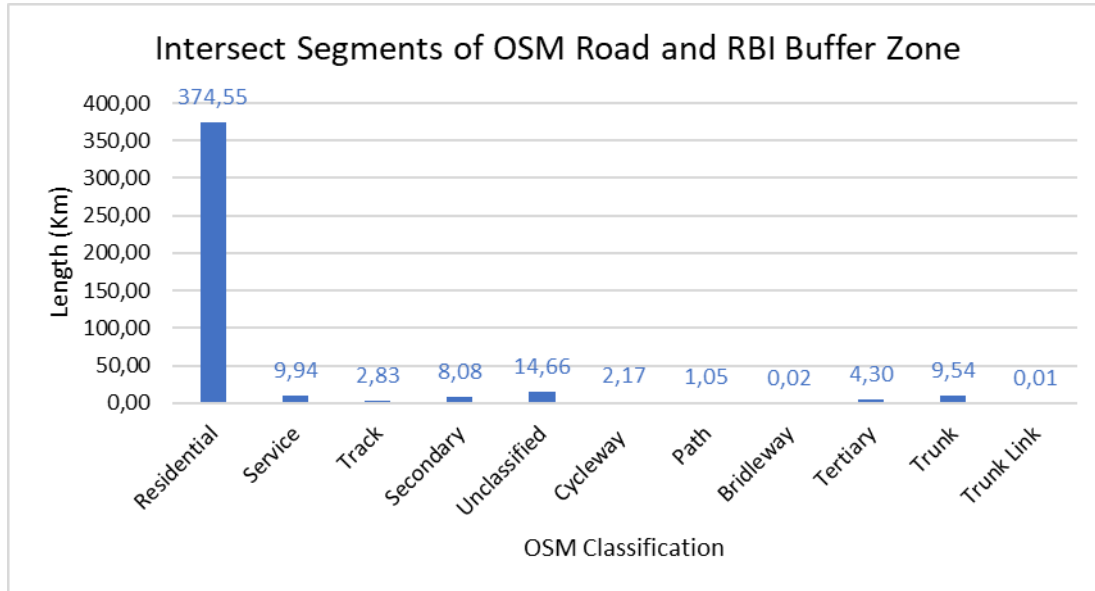


Figure 21 Intersection segments composition

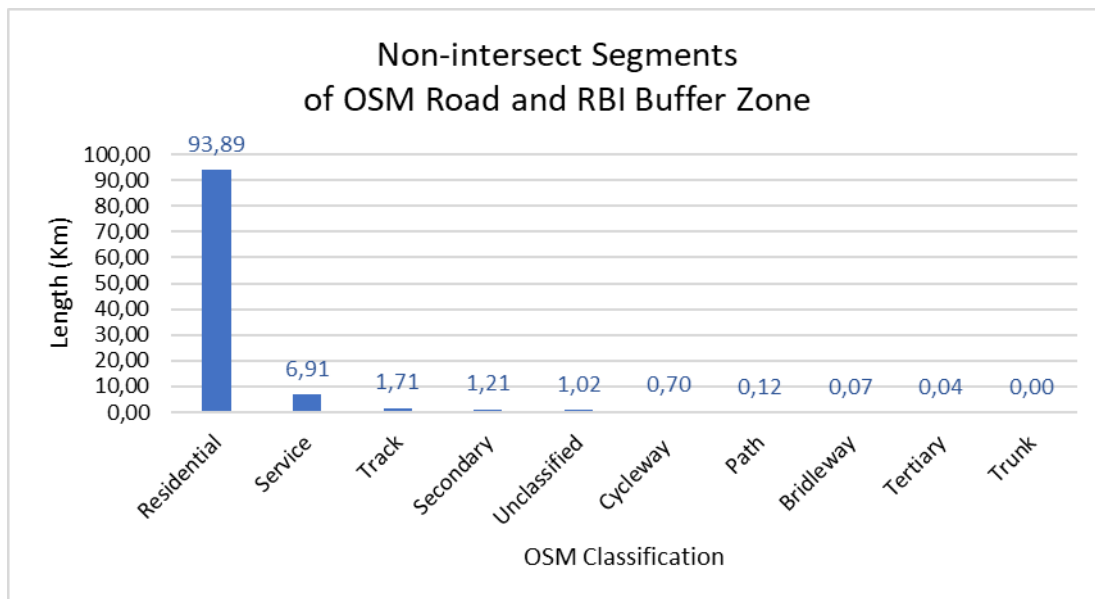


Figure 22 Non-intersect segments composition

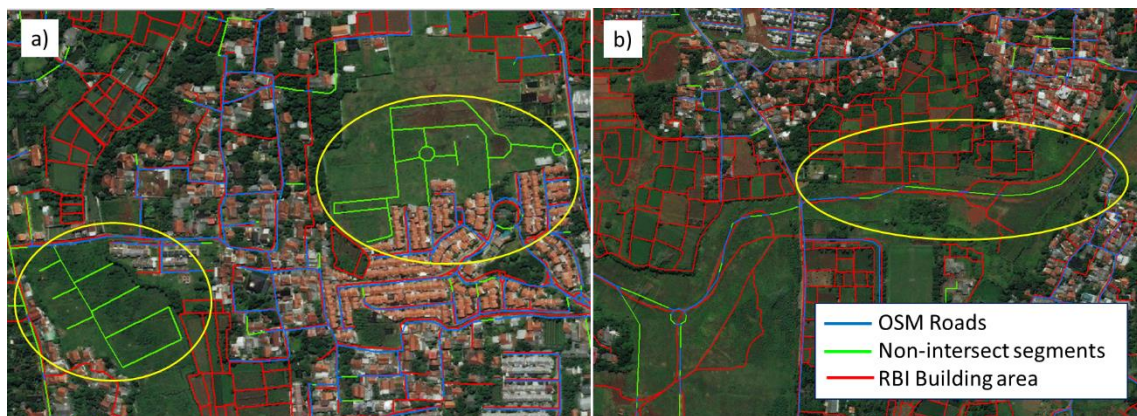


Figure 23 Non-intersect segments indication: a) additional road segments; b) wrong geometry

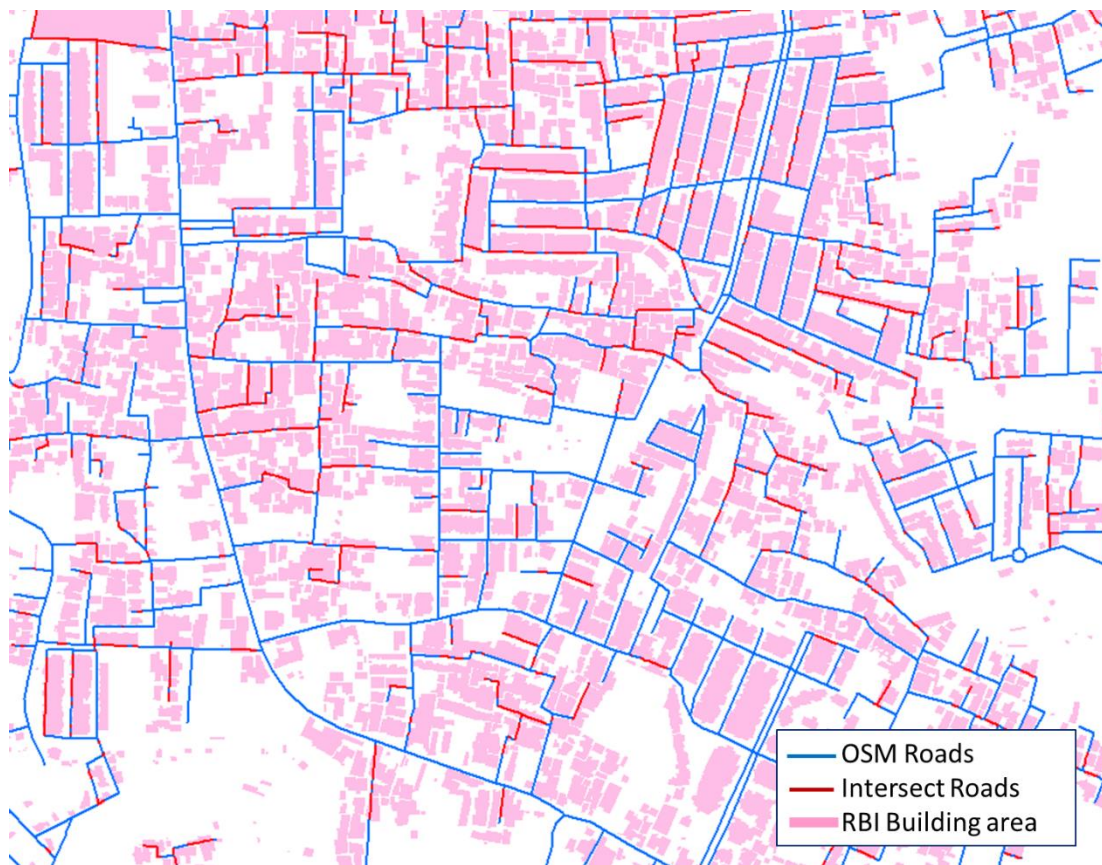


Figure 24 Intersection of OSM road with RBI building

Inter-theme logical consistency assessment is conducted before the intra-theme assessment because during this process new segments formed. By doing so, it can assure that after topology error check there will be no more changes in the data that might cause another error. After the positional accuracy assessment the number of segments becomes 7696 and during this process the segments grow as much as 16362 segments. The result shows that around 53,23 km or around 10% of OSM roads are intersect with the RBI buildings. In this case, the intersection represents the wrong geometry of OSM roads against RBI buildings. Figure 24 shows the

intersect and non-intersect segments overlaid in RBI buildings, while Figure 25 shows the zoomed intersection area with satellite imagery to give the abstraction of ground condition. Both of the figure represent the geometry errors that are found in OSM data if compared to another data. If it is closely observed, in Figure 25 the OSM road is not drawn accordingly above the image. Whereas, Bing satellite imagery is one of the satellite imageries sources that provided by OSM. Unfortunately, the information on how the data generated is not included in the OSM data that used in this prototype. Nevertheless, the result of this process were quite good with only 10% inconsistency theme topology. Statistically, the wrong topology is dominated by the residential roads. The summary of intersect road composition based on the classification can be seen in Figure 26.



Figure 25 Intersection of OSM road with RBI building overlaid in Bing satellite imagery

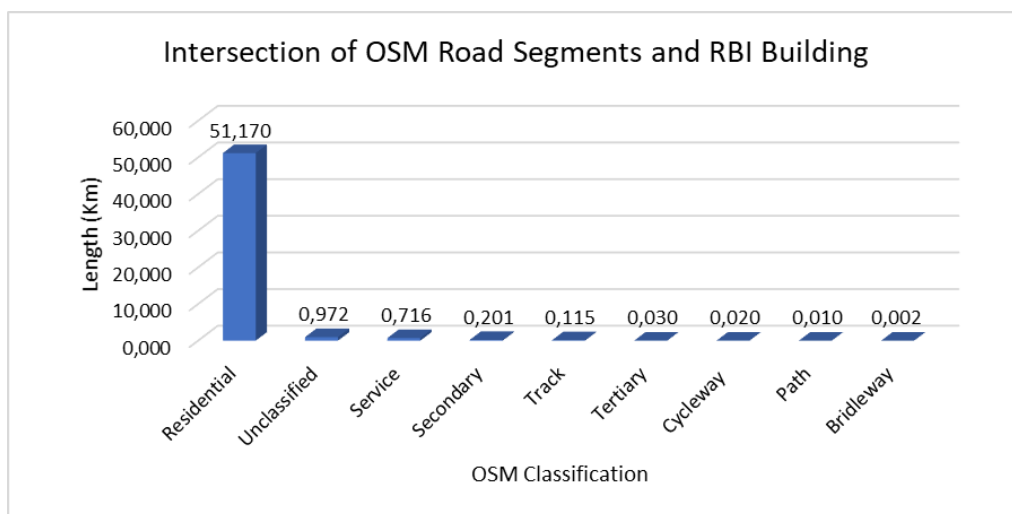


Figure 26 The intersect segments composition

The intra-theme logical consistency that is checked using topology error model shows that there is no overlapping segment found. However, there are many dangles identified during the process. When the highlighted error is checked, all of them are located in the end node of the segments which should not be identified as the geometry error. The result of dangles error identification can be seen Figure 27. The subset area clearly exposes that the dangles are identified in the end nodes, and since the road network in this context are not closed network the errors can be ignored. With no topology error found in the OSM data, all the segments will be given value of +0,50.

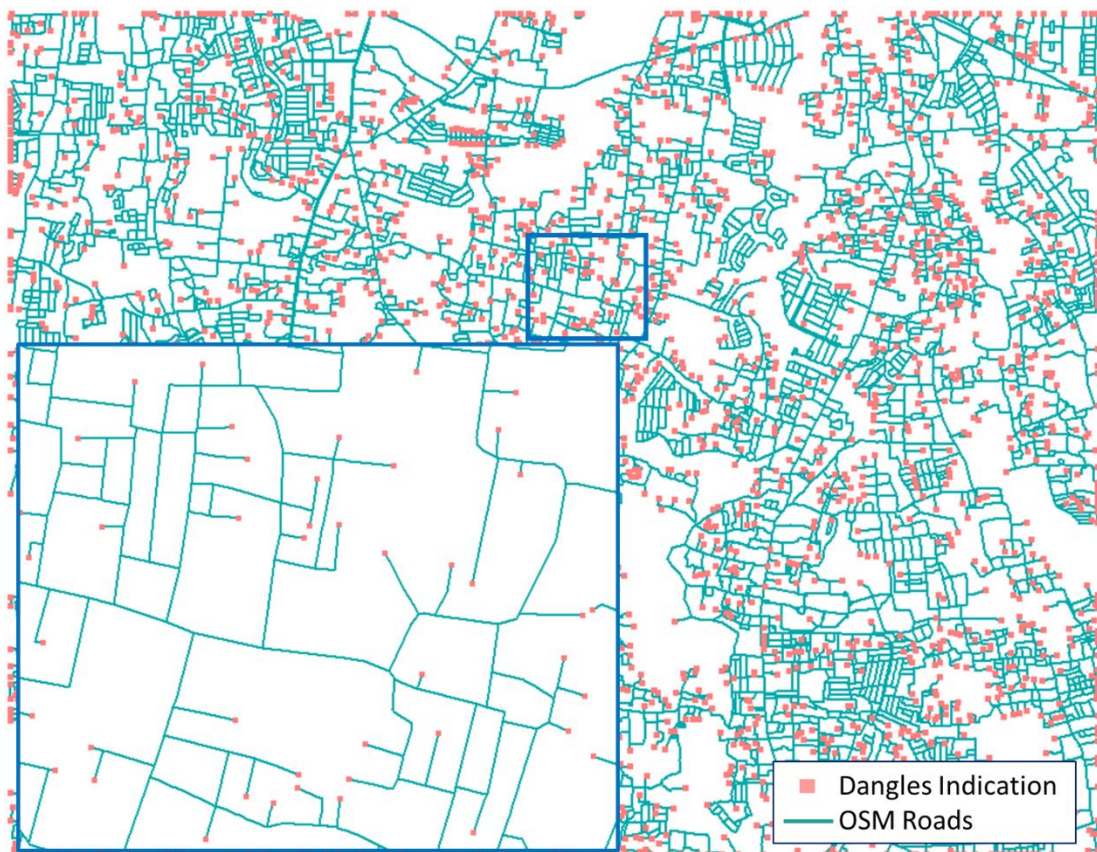


Figure 27 Dangles identification during topology check

Final assessment element in this process is usability which very much depends on the previous result and project goal. As mentioned before, the project goal is to update the RBI map especially for the road features. Since the ultimate purpose is to update authoritative data, all the quality assessment results are required to be excellent, especially for positional accuracy and thematic accuracy. Authoritative data usually is referred as the reliable data in terms of geometry and information. Thus, the OSM data must have excellent positional and thematic accuracy in order to fulfil the requirements. Therefore, to calculate the usability score those two indicators are used. A segment has to had perfect score both for positional accuracy and thematic accuracy, then the segment will be given +1 as the value. If only one criterion is fulfilled or both criteria are not fulfilled, the segment will be valued as 0. After the confidence index calculation,

the usability recommendation is assigned. Table 22 shows the result of usability recommendation derived from confidence index. Unfortunately, based on the quality assessment most of the OSM road segments are suggested to be used as the reference for change detection. There is no segment that has proper quality to be adopted into authoritative data. However, based on the assessment the data still can be used as the change detection.

Table 22 Usability recommendation

Confidence Index	Usability Recommendation	Length (Km)	Percentage
1	Data Adoption	0	0
2	Change Detection	428,88	80,5%
3	Report Alert	103,94	19,5%
Total Score		532,83	100,0%

Nevertheless, even though the usability recommendation is suggested, the final decision for data usability will be decided by the user. The user conformance will be used to determine the final usability of OSM data. In the first phase of this framework, the user conformance is set by the user by defining the usability alternative as change detection. Therefore, the data will be used as change detection for the authoritative data as suggested by the framework which is also supported by the user conformance. Indications of change can be taken from OSM segments that do not intersect with the RBI segments (Figure 23a). However, before the OSM data is used accordingly to its recommendation, some additional process might be needed to optimize the OSM usability.

Technically, the result of this implementation cannot be used to update authoritative data because the important data dimension (positional and thematic accuracy) is not fulfilled by the OSM data. However, the OSM data is still can be used in change detection process which can recommend the priority area for data updating. With this recommendation, the authoritative body can make update plan more effectively and efficiently to accelerate the updating process. Therefore, even the OSM data cannot be used for authoritative data updating directly it is still can give indirect effect in updating process acceleration.

5.3 Evaluation and Discussion

Implementation of the framework is successfully performed even though the result doesn't meet the requirements for updating authoritative data. The result obtained in this process can be influenced by many factors, but the evaluation will highlight data input and quality assessment module use in this implementation.

Regarding the data input, in some actual event, a user might have no clue about what kind of data they should take for their project. Therefore, sometimes data matching to find the prospective data that can be use in the project is needed. The data examiner is responsible to find the matched data based on the user requirements. In this case, constructing a data base that

contains data repository list, data type they provide, and brief description about the VGI data, is very useful. Regular updating is compulsory for the database because of dynamic nature of VGI data. Communication between user and data examiner in deciding data type and repository is very important to help the user in finding proper data source for their project. By doing so, expectedly the usage of improper data as input can be minimized and can increase the efficiency of the whole process.

Downloading the VGI data from a certain data repository also needs to be decided carefully because they come with advantages as well as disadvantages. In the implementation, the OSM data is taken from mirroring sites (BBBike) therefore there are some information limitation in this data, but as the advantage it has already been processed into shapefile format. Another mirroring site for OSM data offers pre-processed shapefile with more complete data attribute but only for paid data. The complete OSM data can be downloaded directly from OSM site, but it demands some additional pre-processing steps which can prolong the processing time. However, more complete data attribute might improve the quality of data.

VGI data availability in developing country also becomes the major concern. As a well-established VGI project, OSM evidently cannot surpass the RBI especially in providing the data update. The sample area is taken in the border of Indonesian capital city, which logically has more potential contributor than smaller cities. Assuming that the citizens have better education and familiar with mapping technology and gadget, the contribution rate should be higher. However, as well recognised by Coleman, et al., (2009) and Olteanu-Raimond, et al. (2017a) the contributor motivation issue strongly influences the VGI data quality. Thus, it is very recommended to use the VGI module for quality assessment since the VGI data quality are strongly related to its nature.

As mentioned before, the data type is divided into three main categories: vector data, textual data, and media contents. All of those data have their own uniqueness regarding the data handling. Therefore, specific quality assessment module can be tailored depends on the requirement. In this implementation, the vector data can be consider as 'ready to use' data for mapping purpose, hence it relatively easy to handle. Furthermore, many references regarding the VGI data utilisation are widely available. However, dealing with textual data and media contents might require more efforts since this kind of data is less explored compare to the vector data. Nevertheless, encouraging the exploration on textual data and media contents is necessary to do.

Beside the data input aspect, the evaluation also concerns in the quality assessment modules that are used in the implementation. As demonstrated, the quality elements that are used in this implementation are very simple and limited to ISO quality elements only. However, the combination of ISO and VGI module are highly recommended to setup the quality assessment module because it can improved the accuracy result for VGI data. Nevertheless, the utilisation of sub-elements quality without considering the relevancies towards tested data will only lead to opposite direction. Therefore, knowing the data type and its nature are very important

because it can be the basic consideration in determining the sub-elements and criteria that will be used in the quality assessment.

Defining the quality assessment module (quality elements, sub-elements, criteria, score, and weighting factor) is the crucial part in the quality assessment process because it serve the user requirement as well as maintains the high standard of data quality. This process might take longer time than running the process itself. Once the quality assessment module is set up, with the specific data condition and quality element used in this implementation, the total time required to run all the process only takes few hours. For more complicated data and quality assessment module, the execution process might take longer but feasible to do with the proper use of quality assessment module. Most of the implementation process are conducted automatically using models, but some process, especially the valuation, are still conducted manually. However, in principal, fully automated process is doable whether using well established software or creating the new tools by using specific programming language. The implementation also demonstrates the overlapping usage of a quality assessment element in pre-processing and processing stages. Overlapping of an element is feasible, but criteria that are used in each element/sub-element need to be clearly distinct to avoid redundancy process.

Referring to the result, for the future implementation of this frameworks the quality assessment module can be modified to improve the result. For example, the selection of sub-element for positional accuracy can include the Hausdorff distance, average distance, angular tolerance which executed using segment pairs approach. In addition, Damerau-Levenshtein distance also can be added as the sub-element for thematic accuracy. The algorithm can detect the road name difference between OSM and RBI data. Furthermore, the VGI elements that can be included in the module includes metadata, contributors' indicators, trustworthiness and reputation.

There are no significant difficulties in valuating each quality assessment elements, including calculating the confidence score and index. This method is feasible to be implemented and can give strong emphasise to important elements for the user. Since the user specification is referred when defining the score and weight, the confidence index resulted from the calculation already accommodates the user requirement. The usability recommendation that is suggested in the end of the process is not obligatory because the user conformance can change the data usability. There is a possibility that the final usability is not in accordance with the usability recommendation. In such case, it is strongly recommended to mention the result of accuracy assessment, thus the end user know the quality assurance of the data they use. Furthermore, some additional processes need to utilise assessed VGI data. As an example, in data adoption for authoritative data update case, data conflation process especially in geometry adjustment is needed to integrate both data. Regarding the geometry adjustment, algorithm to join the linear structures based on coordinates comparison (Stankutė & Asche, 2009) can be further explored.

Two major issues that emerge from the implementation are the lack of VGI data completeness and VGI adoptability in authoritative data. Based on the implementation results, it seems that even a giant VGI project such OSM cannot give significant contribution in authoritative data updating. The data currency that becomes the ultimate advantage of VGI is irrelevant since the older RBI data shows superiority above VGI data. The low participation rate of the contributors might be caused by the lack of project promotion and failure in mapping the contributor motivation. However, generalising the VGI role in authoritative data update from particular area and data source is not fair. Therefore further research that involve the local VGI project are strongly encouraged. Speaking of roads feature update, an example of local VGI project that collects and compiles roads data is Navigasi. With the slogan “every place is unique”, the communities aim to collect navigation route across Indonesia (navigasi.net, 2018). They have active participation of their members that enable them to provide regular map update. This VGI project can be a valuable and promising source for data updating. Furthermore, the VGI explorations also can use another well-established data such as Google map¹⁸ and Waze¹⁹. Regarding the test area, the coverage could be broadened to the rural area, therefore better point of view can be obtained as well as mapping the potential issues. The second concern is VGI adoptability, especially in data quality and classification/tag issues. To address the data quality issues, the combination of ISO standard and VGI approach can be used. Whereas, to address the classification or data tagging a conversion table is needed to be formulated to bridge the classification gap between authoritative data and VGI data.

Generally, developed country, developing country, and under develop country are facing the same problem regarding VGI adoption. However, focus on addressing the problem might vary among the countries. Reviewed form the research finding, the lack of VGI data is one of the problems that need to be prioritised. To overcome the problem, the engagement of government into VGI project needs to be increased and strengthened. Collaboration of government and VGI community is also encouraged to promote participation in VGI data creation and utilisation.

Finally, to summarize the implementation process and result there are some important aspects to be highlighted. First of all, correctly defining the data type and repository can make the process run efficiently. Subsequently, defining the quality assessment module which includes quality elements, sub-elements, criteria, score, and weighting factor is crucial because it can lead to the data quality result. Combining the ISO module and VGI module is highly recommended to explore the VGI data potency to the fullest. Not less important, as a key of the framework, the user specification has to be clearly defined so it can be referred by the following process within the framework.

¹⁸ <https://www.google.com/maps> (accessed 04.09.2018)

¹⁹ <https://www.waze.com/> (accessed 04.09.2018)

6 Conclusion and Recommendation

6.1 Conclusion

This research proposes a conceptual framework in authoritative data update with combination of VGI data. The framework is formulated by adopting the flexibility concept in the usage of quality assessment elements as well as promoting the scoring and weighting factor to define the confidence index that can be referred for VGI usability. Moreover, the user requirement is used as the key concept in developing the framework. Existing VGI data are identified and then classified based on the data handling. The VGI data type are classified in three classes; vector data, textual data, and media contents. Considering the unique nature of VGI data, this classification can be helpful in the data integration process of VGI dan authoritative data, especially in data quality assessment process. Hence, VGI data types are taken into consideration in determining the of quality assessment module.

Three scenarios are simulated to give abstraction on how the proposed framework can be generically used to solve real-world problems with different point of views. First scenario simulates the typical case of authoritative data updating by using vector data type of VGI. In this scenario, the equivalent theme for data reference is available to be used to assess the VGI data quality. The second scenario simulates the extended term of data update; new theme data collection. Since there is no equivalent authoritative data that can be used as the reference, the flexibility of framework is demonstrated in this scenario by utilising alternative data as the reference. The third scenario simulates more complicated case where the VGI data used are textual data which need special treatment before it can be integrated to the map. The equivalent data are also not available for this data type. The main difference in each scenario are mainly determined by the quality assessment module they used. Quality assessment module that used in each scenario is promoting the combination of ISO standard and VGI approach. However, each quality assessment elements is customized specifically in regards of data type.

The first scenario is implemented to test the feasibility of the proposed framework. The implementation demonstrates the use of OSM data to update the road features in RBI data. The scenario implementation successfully demonstrates the feasibility of the framework. The scenario is mostly run automatically in ArcGIS environment by using a model without any significant difficulties. This process proves that automation of the framework is doable. Only ISO modules are applied in the assessment process without includes VGI module. Therefore some improvements in the quality assessment module are recommended, especially in selection of elements, sub-elements, and criteria that will be used in the module. Unfortunately, the data output from this process cannot be used for updating RBI map because the quality is not sufficient. However, it can still be use as the change detection indication that will help authoritative body to plan the updating process effectively. In other words, it still can contribute indirect effect on data updating acceleration by providing the change detection information.

The research finding highlights two issues for OSM data utilisation in sample area; the lack of VGI data completeness in urban area and classification or tagging issues. Current OSM data apparently cannot surpass the superiority of older RBI data, especially in positional and thematic accuracy aspects. Furthermore, unequal classification between OSM dan RBI also inflicts some problems in quality assessment. Those issues have become the obstacle in VGI data adoption into authoritative data. However, since the implementation is conducted in particular area with specific data type, the issues cannot be generalised for every case. Thus, further exploration of various data sources and types as well as sample area are needed to be done.

6.2 Recommendation

Based on the research findings, some recommendations are formulated to help improving the VGI utilisation, especially the integration procedure of VGI and authoritative data for data updating. The recommendations are postulate some actions as follows:

- **Encourage the involvement of government in VGI projects.**

The involvement can be manifested by creating a VGI collection platform (Olteanu-Raimond, et al., 2017b) that is directly managed by the authoritative bodies so that the data can be easily controlled. Sharing some parts of the government data to existing VGI project (Olteanu-Raimond, et al., 2017a) can stimulate the VGI data development. Furthermore, it will also benefit the government to make VGI data more adoptable. Promoting the VGI project and maintaining the contributor involvement are also very important in sustaining VGI Project (Brabham, 2013; Haklay, et al., 2014).

- **Investment on VGI**

Investment on VGI includes: developing of VGI quality assessment tools (Brabham, 2013); establishing authoritative VGI platform (Olteanu-Raimond, et al., 2017b); training the human resources to study VGI data (Johnson & Sieber, 2013); encouraging the VGI research; investigating legal aspect of VGI data and crafting policies related to VGI data (Brabham, 2013).

- **Understanding the VGI contributors and communities**

Knowing the contributors and the communities that are involved in VGI data production, especially their motivation, can be used as inputs to setup the strategy on VGI project initiation and promotes participation among the contributors (Brabham, 2013). Furthermore, understanding the VGI stakeholder's role and relationship is necessary to maintain the project sustainability and successfulness (Haklay, et al., 2014). Ultimately, to encourage establishment of VGI project collaboration

- **Further research direction**

Most of the VGI accuracy assessment research are conducted in developed countries which have good establishment of VGI data. However, the applicability of those tools in developing country needs to be further explored. The further development of proper VGI quality

assessment elements is still relevant since the VGI data are very dynamic. Moreover, Johnson & Sieber (2013) found out that VGI data could affect the political distance between state and citizen. Therefore, the impact of VGI data utilisation within the government needs to be investigated in broader aspects.

Generally, those recommendations apply globally and relevant to the current conditions worldwide. However, to be specific in addressing the research finding from implementation, some recommendations that need to be highlighted are; authoritative data sharing that can stimulate the VGI data development and the research of VGI adoptability in Indonesian authoritative data.

References

- Antoniou, V., & Skopeliti, A. (2015). Measures and Indicators of VGI Quality: an Overview. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences Volume II-3/W5.*, 345-351. Doi:10.5194/isprsannals-II-3-W5-345-2015.
- ArcGIS for Desktop. (2018, August 29). Tratto da <http://desktop.arcgis.com:>
<http://desktop.arcgis.com/en/arcmap/10.3/tools/coverage-toolbox/near.htm>
- Arsanjani, J., Mooney, P., Helbich, M., & Zipf, A. (2015). An Exploration of Future Patterns of the Contributions to OpenStreetMap and Development of a Contribution Index. *Transactions in GIS* 19, 896–914. DOI: <https://doi.org/10.1111/tgis.12139>.
- BIG. (2005). *Spesifikasi Pemetaan Rupabumi 55 - Spesifikasi Kode Unsur*. Bogor, Indonesia: Badan Informasi Geospasial.
- Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., & Rampini, A. (2015). A User-driven Selection of VGI Based on Minimum Acceptable Quality Levels. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences Volume II-3/W5.*, 277-284. Doi:10.5194/isprsannals-II-3-W5-277-2015.
- Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., & Rampini, A. (2016). On predicting and Improving the Quality of Volunteer Geographic Information Projects. *International Journal of Digital Earth*, 9:2., 134-155, DOI: 10.1080/17538947.2014.976774.
- Brovelli, A., Minghini, M., Molinari, M., & Mooney, P. (2017). Towards an Automated Comparison of OpenStreetMap with Authoritative Road Datasets. *Transactions in GIS*, 21(2), 191–206. Doi: 10.1111/tgis.12182.
- Burrough, P. A., McDonnell, R. A., & Lloyd, C. D. (2015). *Principle of Geospatial Information Systems. Third Edition*. Glasgow: Oxford University Press.
- Capineri, C. (2016). The Nature of Volunteered Geographic Information. In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, . . . (eds.), *European Handbook of Crowdsourced Geographic Information* (p. Pp. 15-33.). London: Ubiquity Press. DOI: <http://dx.doi.org/10.5334/bax.b>. License: CC-BY 4.0.
- Coetzee, S. (2018, July 13). *icaci.org*. Tratto da International Cartographic Association: https://icaci.org/files/documents/wom/16_IMY_WoM_en.pdf
- Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered Geographic Information: The Nature and Motivation of Producers. *International Journal of Spatial Data Infrastructures Research*, Vol. 4, 332-358.

- Cooper, A., Coetzee, S., & Kourie, D. (2011). Volunteered Geographical Information – The Challenges. *AfricaGEO*, (p. 34-38). Capetown, South Africa.
- Craglia, M., Ostermann, F., & Spinsanti, L. (2012). Digital Earth from Vision to Practice: Making Sense of Citizen-Generated Content. *International Journal of Digital Earth*, 5(5), 398–416.
- De Longueville, B., Ostländer, N., & Keskitalo, C. (2010). Addressing vagueness in Volunteered Geographic Information (VGI) - A case study. *Int. Journal of SDI Research*, v.5, 1725-0463.
- Docan, D. (2013). Spatial Data Quality Assessment in GIS. *Proceedings of the 1st European Conference of Geodesy & Geomatics Engineering (GENG '13); Recent Advances in Geodesy and Geomatics Engineering* (p. 105-112). Antalya, Turkey, 8-10 October 2013.: WSEAS Press. ISBN: 978-960-474-335-3.
- Du, H., Alechina, N., Jackson, M., & Hart, G. (2017). A Method for Matching Crowd-sourced and Authoritative Geospatial Data. *Transaction in GIS*, 21(2): 406-427. doi: 10.1111/tgis.11210.
- Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, 102:3, 571-590, DOI: 10.1080/00045608.2011.595657.
- Engler, N., Scassa, T., & Taylor, D. (2014). Cybercartography and Volunteered Geographic Information. In D. editor: Fraser, & T. Lauriault, *In Developments in the Theory and Practice of Cybercartography: Applications and Indigenous Mapping 2nd edition* (p. Taylor, vol. 5, 2nd edn, 43–57.). New York: Elsevier.
- Escobar, F., Hunter, G., Bishop, I., & Zerger, A. (2018, July 14). <https://geogra.uah.es>. Tratto da Departamento de Geología, Geografía y Medio Ambiente Unidad Docente de Geografía: https://geogra.uah.es/patxi/gisweb/GISModule/GIST_Vector.htm
- ESRI. (2018, July 14). <https://support.esri.com>. Tratto da ESRI Technical Support: <https://support.esri.com/en/other-resources/gis-dictionary/term/7cbd3f7c-e17f-4bb0-a51a-318ccf5b68f1>
- ESRI. (2018, July 14). <https://support.esri.com>. Tratto da ESRI Technical Support: <https://support.esri.com/en/other-resources/gis-dictionary/term/d70395b4-f5c4-4375-b053-c2d3b24a3bce>
- Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an Integrated Crowdsourcing Definitions. *Journal of Information Science*, 32(2):189-200. doi: 10.1177/0165551512437638.

- Fast, V., & Rinner, C. (2014). A System Perspective on Volunteered Geographic Information. *ISPRS International Journal of Geo-Information*, 3: 1278-1292. doi: 10.3390/ijgi3041278.
- Fonte, C. C., Antoniou, V., Bastin, L., Estima, J., Arsanjani, J. J., Bayas, J.-C. L., . . . Vatsava, R. (2017). Assessing VGI Data Quality. In G. Foody, L. See, S. Fritz, P. Mooney, A.-M. Olteanu-Raimond, C. C. Fonte, . . . (eds.), *Mapping and the Citizen Sensor*. . London: Ubiquity Press. DOI: <https://doi.org/10.5334/bbf.g>. License: CC-BY 4.0.
- Forati, A. M., & Karimipour, F. (2016). A VGI Quality Assessment Method for VGI based on Trustworthiness. *GI_Forum Vol. 1; Advances in GIScience*, 3 -11. Doi: 10.1553/giscience2016_01_s3.
- Girres, J., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. . *Transactions in GIS 14*, 435–459. Doi: 10.1111/j.1467-9671.2010.01203.x.
- Goodchild, M. (2007). Citizens as Sensor: The World of Volunteered Geography. *GeoJournal*, 69, 211-221.
- Goodchild, M., & Li, L. (2012). Assuring the Quality of Volunteered Geographic Information. *Elsevier, Spatial Statistics 1*; 110-120. doi:10.106/j.spasta.2012.03.002.
- Haklay, M. (2010). How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design 2010, volume 37*, 682-703. Doi:10.1068/b35097.
- Haklay, M., Antoniou, V., Basiouka, S., Soden, R., & Mooney, P. (2014). *Crowdsourced Geographic Information Use in Government*. London: Report to GFDRR (World Bank).
- ISO. (2013). *ISO 19157:2013 Geographic Information - Data Quality*.
- Jackson, M. J., Rahemtulla, H., & Morley, J. (2010). The Synergistic Use of Authenticated and Crowd-Sourced Data for Emergency Response. *Proceedings of the Second International Workshop on Validation of GeoInformation Products for Crisis Management*, 91-9.
- Johnson, P., & Sieber, R. (2013). Situating the Adoption of VGI by Government. In D. Sui, S. Elwood, & M. (. Goodchild, *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (p. 65-81). Netherland: Springer. Doi:10.1007/978-94-007-4587-2_5.
- Juhász, L., Rousell, A., & Arsanjani, J. J. (2016). Technical Guidelines to Extract and Analyze VGI from Different Platforms. *Data*, 1, 15; doi:10.3390/data1030015. Liscence: CC - BY 4.0.

- Kamus Besar Bahasa Indonesia (KBBI)*. (2018, August 29). Tratto da <https://kbbi.web.id>:
<https://kbbi.web.id/pematang>
- Koukoletsos, T., Haklay, M., & Ellul, C. (2012). Assessing Data Completeness of VGI Through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, 477–498. Doi: <https://doi.org/10.1111/j.1467-9671.2012.01304.x>.
- Lush, V., Bastin, L., & Lumsden, J. (2012). Geospatial data quality indicators. *Proceedings of the 10 International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Florianopolis-SC, Brazil.
- Massa, P., & Campagna, M. (2016). Integrating Authoritative and Volunteered Geographic Information for spatial planning. . In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, . . . (eds.), *European Handbook of Crowdsourced Geographic Information* (p. Pp. 401–418.). London: Ubiquity Press. DOI: <http://dx.doi.org/10.5334/bax.ac>. License: CC-BY 4.0.
- Meek, S., Jackson, M., & Leibovici, D. (2014). A Flexible Framework for Assessing the Quality of Crowdsourced Data. *Proceedings of the 17th AGILE International Conference on Geographi Information Science: Connecting a Digital Europe through Location and Place*. Castellón, Spain, 3-6 June 2014: Available at: https://agile-online.org/conference_paper/cds/agile_2014/agile2014_112.pdf [Last accessed 21 July 2018].
- Minghini, M., Antoniou, V., Fonte, C. C., Estima, J., Olteanu-Raimond, A., See, L., . . . Lupia, F. (2017). The Relevance of Protocols for VGI Collection. In G. Foody, L. See, S. Fritz, P. Mooney, A. Olteanu-Raimond, C. C. Fonte, & V. Antonious, *Mapping and the Citizen Sensor* (p. Pp. 223-247.). London: Ubiquity Press. Doi: <https://doi.org/10.5334/bbf.j>. Liscence: CC-BY 4.0.
- Mooney, P., & Minghini, M. (2017). A Review of OpenStreetMap Data. In G. Foody, L. See, S. Fritz, O. Mooney, A.-M. Olteanu-Raimond, C. Fonte, & V. Antoniou, *Mapping and The Citizen Sensor* (p. Pp. 37-59). London: Ubiquity Press. DOI: <https://doi.org/10.5334/bbf.c>. Liscence: CC-BY 4.0.
- Mooney, P., Minghini, M., Laakso, M., & Antoniou, V. (2016). Towards a Protocol for the Collection of VGI Vector Data. *ISPRS Int. J. Geo-Inf.*, 5, 217; Doi:10.3390/ijgi5110217.
- navigasi.net*. (2018, September 01). Tratto da [navigasi.net](http://www.navigasi.net): <http://www.navigasi.net/index.php>
- Olteanu-Raimond, A., Laakso, M., Antoniou, V., Fonte, C., Fonseca, A., Grus, M., . . . Skopeliti, A. (2017). VGI in National Mapping Agencies: Experiences and Recommendations. In G. Foody, L. See, S. Fritz, P. Mooney, A. Olteanu-Raimond, C. Fonte, . . . (eds.), *Mapping*

- and the Citizen Sensor*. (p. Pp. 299–326.). London.: Ubiquity Press. DOI: <https://doi.org/10.5334/bbf.m>. License: CC-BY 4.0.
- Olteanu-Raimond, A.-M., Hart, G., Foody, G., Touya, G., Kellenberger, T., & Demetriou, D. (2017). The Scale of VGI in Map Production: A Perspective on European National Mapping Agencies. *Transactions in GIS*, 21(1): 74–90. doi: 10.1111/tgis.12189.
- OpenStreetMap. (2018, July 16). *Beginners Guide 1.3*. Tratto da <https://wiki.openstreetmap.org>: https://wiki.openstreetmap.org/wiki/Beginners_Guide_1.3
- Ormeling, F. (2003). Functions of Geographical Names for Cartographic and Non-cartographic Purposes. In F. Editor: Ormeling, K. Stabe, & J. Sievers, *Mitteilungen des Bundesamtes für Kartographie und Geodäsie, Band 28: Training Course on Toponymy* (p. 29-37). Köln: Moeker Merkur GmbH.
- QGIS. (2018, 07 14). <https://docs.qgis.org>. Tratto da Documentation QGIS 2.8: https://docs.qgis.org/2.8/en/docs/gentle_gis_introduction/vector_data.html
- See, L., Estima, J., Pöör, A., Arsanjani, J. J., Bayas, J.-C. L., & Vatsava, R. (2017). Sources of VGI for Mapping. In G. Foody, L. See, S. Fritz, P. Mooney, A.-M. Olteanu-Raimond, C. C. Fonte, & V. Antoniou, *Mapping and The Citizen Sensor* (p. Pp. 13-35). London: Ubiquity Press. DOI: <https://doi.org/10.5334/bbf.b>. License: CC-BY 4.0.
- See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., . . . Rutzinger, M. (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS International Journal of Geo-Information*, 5, 55; doi:10.3390/ijgi5050055.
- Senaratne, H., Mobasher, A., Ali, A., Capineri, C., & Haklay, M. (2017). A Review of Volunteered Geographic Information Quality Assessment Methods. *International Journal of Geographical Information Science*, 31:1., 139-167. Doi: 10.1080/13658816.2016.1189556.
- Souza, W., Lisboa-Filho, J., Filho, J., & Câmara, J. (2016). DM4VGI: A template with dynamic metadata for documenting and validating the quality of Volunteered.
- Stankutė, S., & Asche, H. (2009). An Integrative Approach to Geospatial Data Fusion. *Computational Science and Its Applications – ICCSA 2009, vol. 5592.*, pp. 490–504.
- Sui, D., & Cinnamon, J. (2017). Volunteered Geographic Information. In D. Richardson, N. Castree, M. Goodchild, A. Kobayashi, W. Liu, R. Marston, & (eds), *The International Encyclopedia of Geography* (p. 1-13). John Wiley & Sons, Ltd. Doi: 10.1002/9781118786352.wbieg0913.

- Van-Oort, P. (2006). *Spatial data quality: from description to application*. Rotterdam: Optima. ISBN 90-8504-339-5.
- Wiemann, S., & Bernard, L. (2016). Spatial Data Fusion in Spatial Data Infrastructures using Linked Data. *International Journal of Geographical Information Science Vol. 30, No.4*, 613-636. Doi:10.1080/13658816.2015.1084420.
- Wiki. (2018, August 29). Tratto da <https://wiki.openstreetmap.org>:
<https://wiki.openstreetmap.org/wiki/Key:highway>