Good Practice Guidelines for Assessing VGI Data Quality

EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Introduction

Volunteered Geographic Information (VGI) may be potentially of value for many areas, such as validation of Land Cover Maps and to complement authoritative data. However, questions around the quality of the data and the credibility of the volunteers arise. The assessment of VGI quality is fundamental for determining its fitness-for-use in specific applications. Several aspects of data quality can be assessed, such as positional quality (precision and accuracy), thematic quality (level of detail and accuracy), credibility of the data and of the volunteer, completeness, currency and logical consistency.

Criteria for the assessment of VGI quality

Traditional aspects of geographic information quality assessment considered:

Positional accuracy - Is usually associated with data georeferenced as points, lines or areas. The positional accuracy of points representing geotagged photographs may also be considered and analysed, particularly since the location recorded by the device that took the photograph and the subject depicted in the photograph may be offset by a certain distance.

Thematic quality - Assesses the accuracy of classes or thematic tags associated with specific locations or objects placed in geographical space, such as classes assigned to pixels in a land cover map, a tag assigned to a linear entity or a polygon, as for example a highway, river, building or green area. Some research has examined how well volunteers can classify satellite imagery from the Geo-Wiki application but more research is needed in this area.

Additional aspects related to VGI quality assessment:

Credibility of the data - Credibility of the data is used to denote the quality of the information contained in that observation.

Credibility of the volunteer - The credibility of a volunteer is the degree to which the information he /she provides can be trusted – may be an indicator of the reliability of the data provided.

Keywords:

Volunteered Geographic Information
Quality
Positional Accuracy
Thematic Quality
Credibility

Cidália C. Fonte	Department of Mathematics, University of Coimbra, Coimbra, Portugal - cfonte@mat.uc.pt
	Institute for Systems Engineering and Computers at Coimbra, (INESC Coimbra) , Coimbra, Portugal
Lucy Bastin	School of Engineering and Applied Science, Aston University, Birmingham, UK - l.bastin@aston.ac.uk
Linda See	International Institute for Applied Systems Analysis (IIASA), Laxenburg Austria - see@iiasa.ac.at
Giles Foody	School of Geography, University of Nottingham, Nottingham, UK - giles.foody@nottingham.ac.uk
Jacinto Estima	Information Management School (IMS), Universidade Nova de Lisboa, Lisbon, Portugal - D2011086@novaims.unl.pt

Good practice guidelines for quality assessment

Additional data that should be recorded or used, which can add value to the assessment of VGI quality, or additional procedures that could be implemented are listed:

Positional quality

Collect information from multiple contributors	• Enables the assessment of positional consistency • This information may be used to conflate data It has been shown that the positional accuracy of roads in OSM improved with an increase in the number of contributors, illustrating that Linus' Law applies to this source of VGI
Store historical information	• Would enable the identification of change and assess the stability of positional information over time
Store data about the methodology used to determine the position Different positioning procedures, may have different degrees of reliability, precision and accuracy	• Positioning over a geo-referenced image? The location of a phenomenon may be assessed by positioning it on top of a satellite or aerial image.
	• The nature of the image - spatial resolution and spectral composition — Will provide information about the volunteer' difficulty in identifying the features, and the accuracy of the obtained position
	• The date of image collection - year, month, day and even time of day — May give information about the season (which may influence the vegetation, phenology, degree of human occupancy in touristic regions, amount of traffic depending on, e.g., whether it was rush hour or not, the amount of light, the direction of shade in the image, etc.
	• Whether specific instructions were given to the volunteers about where to locate the features. — Ideally, information regarding where to place the photographs should be provided, including the location from which the photograph was taken and its orientation. Other information regarding points of interest should also be specified, e.g. whether the points correspond to the building centroid or to the entrance of the building that gives access to the point of interest
	• Positioning over a map? The following may provide information about the reliability of the base map used:
	• The type of map used — A map made by volunteers, a thematic map, a topographic map created by a national mapping agency?
	• The map scale and/or the minimum mapping unit — May provide a measure of the maximum precision attainable
	• The date of the map — Important for determining the currency of the data
	• The level of generalization of the features and classes present in the map
	• Positioning using GNSS measurements?
	• The measurement is made automatically when the data are collected and uploaded - e.g. when taking a picture and uploading an EXIF file • The positional data is collected separately and uploaded later - less reliable • Additional information to assess quality, which would be useful if recorded, includes: a) The type of GNSS receiver used; b) The number of measurements used to determine the location; c) The date and time of the measurement, which enables the determination of the Dilution of Precision associated with the measurements; d) The number of satellites used for positioning.
	• Conflating data provided by volunteers When the position of the feature is obtained through the conflation of data provided by several volunteers (or from data involuntarily obtained, e.g. from mobile phones), it would be useful to know:
	a) The amount of data for a given feature that has been used to obtain the indicated location; b) The degree of variability of the data used to determine the most probable value; c) The dates and times associated with the collection of data about a particular feature. d) Alternatively, interested users can be provided with access to the raw data.

Thematic quality

• The quality control of thematic data may be facilitated if some procedures are implemented during the data collection process, such as:	
• Collecting information from multiple contributors	• Enables checking the consistency of the results • Assigning a label through the conflation of data, whenever divergent data are provided, using, for example, latent class analysis
• Asking volunteers for a confidence rating with the tags	
• Keeping historical information for the same reasons as outlined above for positional quality	
• Indirect information about the confidence of the volunteer in the assignment of the tag or label may be obtained by:	
• Collecting additional information such as the amount of time taken to assign a label	
• Whether the volunteer used instructions or consulted training materials between assessing a point and providing a label	
• Other metadata might be useful, such as:	
• The prevailing atmospheric condition at the time of data collection, which may be relevant for the collection of biological or environmental data	
• Additional data useful for photographs that may be useful for applications in land cover /use mapping:	• The orientation of the photograph • A description of whether the surrounding area is homogeneous or heterogeneous the date when the photograph was taken, • Data about the exposure of the photograph and the type of camera used

Guidelines for assessing volunteer credibility

Volunteer credibility = volunteer expertise + volunteer trustworthiness	
• Volunteer's expertise may be assed using metadata about the volunteer:	• Education • Profession • Interests
• Volunteer's trustworthiness may be assed using:	• Use control information, such as test sites where information provided by experts or selected volunteers is available, which can be used to assess the contributions of each volunteer. • Use historical data provided by the volunteer, such as the number of times their contributions were corrected by other volunteers, selected volunteers or experts • Use information about the where the volunteer is located. Here we assume that the closer a volunteer is to the location of the data that was uploaded by them, the more reliable the data will be.

Generic good practice guidelines

Some general practices may be implemented that can contribute to the production of more reliable information, such as:	
•Implement automatic means to check the data.	May use additional data or metadata and make an automatic check of whether the data provided are likely to be correct.
•Enable volunteers to identify erroneous contributions (regarding positions or attributes). May provide valuable information about:	•the contributors
	•difficulties in assigning classes
	•the credibility of locations of phenomena.
•Enable discussions among the volunteers whenever difficulties are found, such as the best class to assign to a particular location. This may enable the sharing of locally relevant information; improve the understanding of ontologies; self-correction; and quality control.	

The role of protocols

Strict protocols for data collection are used in some areas of citizen science projects. However, in other projects the contributers have freedom in what they map, how and which tags they use. This means the quality of data is more difficult to assess.

The establishment of protocols may, on the one hand, provide valuable information which may make the data useful for additional applications. However, too demanding protocols might demotivate the volunteers to contribute.

A balance needs to be identified so that protocols are not seen as restrictive but rather a way to help users in providing higher quality data.

Conclusions

Quality assessment of VGI remains one of the most important issues for determining the fitness-of-use of VGI for different applications. Some good practice guidelines are presented here that may provide valuable information to assess the positional and thematic quality as well as the volunteers' and data credibility, enabling its potential utilization for a wider range of applications. However, the implementation of some of these guidelines might require the definition of protocols for data collection, which has advantages and disadvantages, and therefore needs to be defined with caution.