VOLUNTEERED GEOGRAPHIC INFORMATION

MEASURING QUALITY, UNDERSTANDING THE VALUE

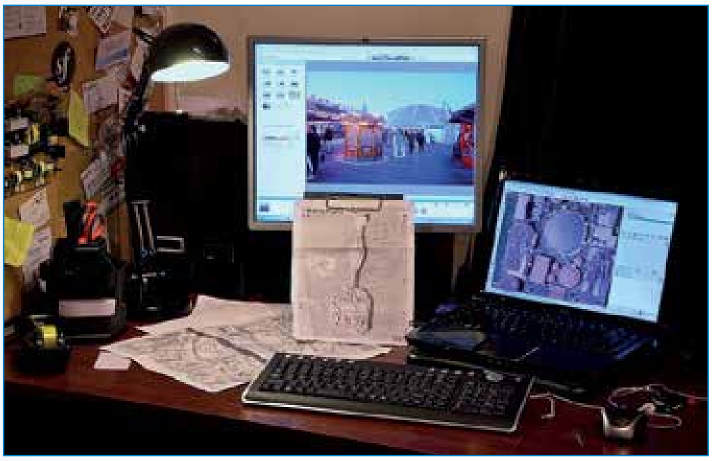


Fig. 1 – Adding data to OSM after mapping Brighton Pier. (Source: https://en.wikipedia.org/ Author: Alexander Kachkaev).

by Vyron Antoniou

The article explores the world of Volunteered Geographic Information from birth until today, observing the actors, the sources and the problems inherent the data quality according to ISO standards. The author underline the potential of this social phenomenon in constant evolution and the impact this could have, in the very near future, in the various fields of geospatial information.

The birth of VGI

In 2007, Mike Goodchild coined the term Volunteered Geographic Information (VGI). He was describing "*the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information*" (Goodchild, 2007: p.217). Many mark the birth of VGI with the birth of OpenStreetMap (OSM – www.openstreetmap.org) in 2004. While OSM has played a key role in the development of the phenomenon, the fact is that the crowdsourced and collaborative creation of spatial content was not something new.

The creation of VGI was sparked by a mixture of a different factors and it is difficult to understand the quality aspect of VGI without first examining the factors that lead to the appearance of this phenomenon. It is interesting to realise who these factors are, not only because some of them are new to the Geomatics domain and thus the professionals of Geographic Information (GI) need to extend their horizons so to study and understand them, but also because these factors are still the driving force behind the evolution of VGI. Thus, we need to have a clear view of their importance and role if we want to understand how VGI is evolving and what the quality caveats that come with it are.

One of them is the mentality of collaboration in order to achieve a goal. VGI mimicked, in a sense, the mechanisms of Open Source Software where a team of, otherwise unrelated, programmers joined forces to create a free and open piece of software. In the case of OSM for example, the equivalent was to create an open and free map of the world. This collaboration was further facilitated by: i) the proliferation of accurate and low cost GPS-enabled devices which turned technology savvy people into "neo-geographers" (Turner 2006) and citizen-sensors (Goodchild 2007), and ii) novel programming techniques which transformed Web into a bi-directional medium regarding content creation. Moreover, the turn to spatial applications, which were freely accessible to the public, by the technology giants (e.g. Google, Microsoft, Yahoo!) drew the attention around spatial data and applications. On the other hand, however, the National Mapping Agencies (NMAs) were, in effect, keeping Spatial Data Infrastructures (SDI) out of the reach of the general public with high pricing and complicated licensing terms. The intertwining of all these factors contributed to the appearance and the development of the VGI phenomenon. What has not been clearly stated as a contributing factor, but yet exists and affects VGI, is the social component. VGI, before and above all, is a social phenomenon and this factor will be further analysed when we turn the discussion to VGI quality.

Types and Sources of VGI

Today VGI is omnipresent. It comes from various sources and it can be found in many flavours including toponyms, GPS tracks, geo-tagged photos, synchronous micro-blogging, social networking applications, blogs, sensor measurements, complete topographic maps, etc. Topographic VGI can come as a result of field work or bulk data import of authoritative datasets that are now freely available. It is obvious that all these sources cannot just fall under one category. There are many aspects that can be examined here, but of particular interest when we examine the sources and types of VGI is the focus, the origin, the motivation and the scope of the VGI contributors; in a sense if they are generating VGI in an implicit or explicit manner and in what context. Implicit contribution takes place via websites or applications where their main focus is on activities not related to the geospatial domain. This does not preclude the presence of a geospatial aspect as one of the many interesting features that such applications could have but spatial information is neither one of the core features nor the main motivation of their contributors. Often, contributors are not aware of the fact that their digital presence leaves also a spatial footprint. On the other hand, spatially explicit sources, urge their users to use geography and location as a motivational and organisational factor. The narrative behind these sources asks contributors to interact directly with spatial features and consciously focus their attention into capturing spatial elements.

Both the source and the type of VGI play a role in the quality and value of VGI. However, before turning to this, we briefly review the basics of spatial data quality.

Spatial Data Quality

In general, according to ISO 9000 (ISO 2005), quality is the "*degree to which a set of inherent characteristics fulfils requirements*". *Characteristics* (or quality elements) are defined as distinguishing features of a product that can be either inherent or assigned, and can be either qualitative or quantitative. *Requirement* is defined as a need or an expectation that is stated, obligatory or generally implied. Thus, understanding and measuring quality boils down to defining the elements of a product and how these elements serve the usages expected; in one word: fitness-for-purpose. While this might seem as oversimplification, it is not. Most of the times it is very difficult to analyse and measure correctly these inherent characteristics, and the same applies in unequivocally defining the requirements to be met. Spatial data is no different and the same rules, and problems, apply when it comes to understanding and measuring spatial data quality; either quality pertains to authoritative data or VGI.

The discussion about quality becomes even more intriguing when product specifications are included in the equation.

This is because a product might adhere to the existing specifications but fails to fulfil requirements. In quality terms, this product has high *internal* quality (i.e. is produced according to specifications), but it has poor *external* quality (i.e. it does not fulfil its purpose). Again, this is the case also with spatial data. In other words, the fact that a VGI dataset (implicitly or explicitly created) is created according to some initial specifications does not necessarily mean that it can be used to cover all or any requirements stated by potential end-users.

Spatial data quality has long been an interesting topic for academics and GI professionals alike. There are obvious reasons for that. GI is the basic ingredient for all mapping and geo-spatial products and applications. If this ingredient is of poor quality, it just dooms any other effort. This explains the special interest shown by NMAs and corporations for the standardization of the terms and procedures used in spatial data quality evaluation. A prime example towards this end is the specifications issued by the International Standards Organization (ISO) and the Technical Committee 211 (ISO/TC211) responsible for the geographic data. In 2013, a new international standard was issued, ISO 19157 (ISO, 2013), which provides a holistic approach for spatial data quality (see fig. 2).

"This International Standard establishes the principles for describing the quality of geographic data. It:

- *defines components for describing data quality;*
- *specifies components and content structure of a register for data quality measures;*
- *describes general procedures for evaluating the quality of geographic data;*
- *establishes principles for reporting data quality.*

This International Standard also defines a set of data quality measures for use in evaluating and reporting data quality. It is applicable to data producers providing quality information to describe and assess how well a dataset conforms to its product specification and to data users attempting to determine whether or not specific geographic data is of sufficient quality for their particular application". ISO 19157

Fig.2 - The scope of ISO 19157 international standard.

Spatial Data Quality Elements

When it comes to the evaluation of spatial data quality, a basic component is the characteristics or elements that compose this quality. These elements are factors that can be measured and the conformance of a dataset can be documented and reported to any interesting party. Thus, *spatial data quality elements* provide a tangible façade of a dataset's quality, irrespectively of whether it is an authoritative or VGI one. First the understanding and then the assessment of these elements is fundamental when it comes to measuring GI quality. A brief description is provided (ISO, 2013): i) *Completeness,* refers to the presence or absence of features, their attributes and relationships compared to the product's specification; ii) *Logical consistency,* refers the degree of adherence to logical rules of data structure, attribution and relationships as described in product's specifications; iii) *Positional accuracy,* refers to the accuracy of the position of features within a spatial reference system; iv) *Thematic accuracy,* refers to the accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships; v) *Temporal quality,* refers to the quality of the temporal attributes and temporal relationships of features; vi) *Usability***,** refers to how a given dataset can meet specific user requirements that cannot be described using the quality elements described above.

All the spatial quality elements (with the exception of Usability) can be further analysed into quality sub-elements so to better assess and measure the quality of a dataset.

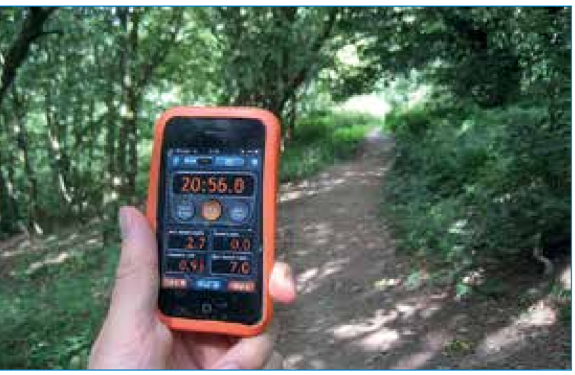


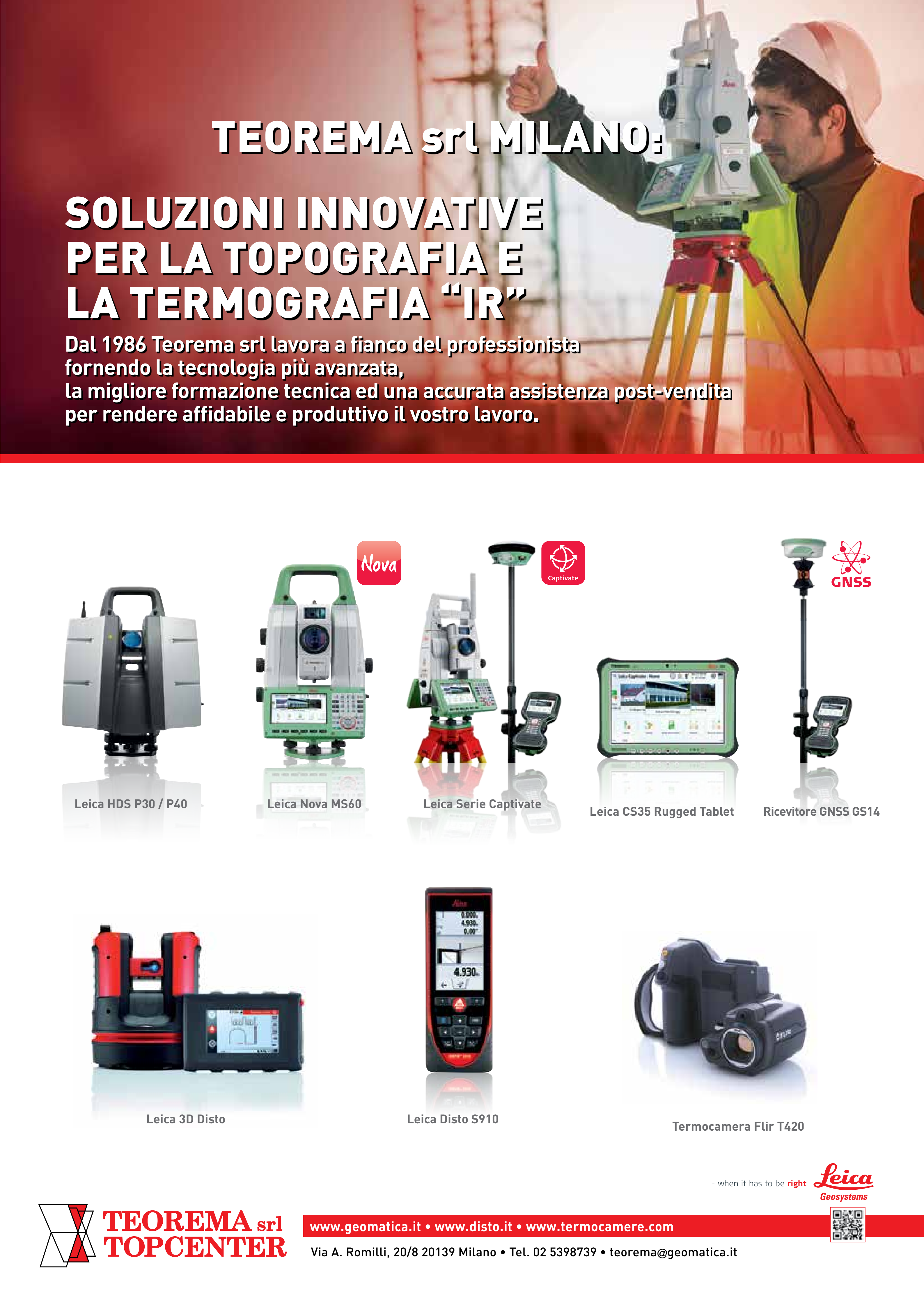
Fig. 3 – Motion X GPS and OSM. (Source: https://en.wikipedia.org/ Author: Harry Wood).

Why this is not enough for VGI

The framework suggested by ISO, and now followed by many authoritative sources of GI, has been rigorously developed by the Geomatics community, and is serving very well the efforts to provide a tangible description of GI quality. However, these guidelines have been developed in a totally different context compared to what we face today. Quality evaluation guidelines have been created for authoritative datasets. Authoritative datasets come from an ecosystem composed of trained personnel that follow tested protocols and procedures, rigor product specifications, certified equipment and software, organizational structures and processes that work towards a high quality result, multiple quality control levels, and of course the absence of social, spatial or other biases as most of the authoritative data come from NMAs. For this kind of data, ISO standards (or similar quality evaluation procedures) will continue to be the basic reference point. What is not clear, however, is how to handle VGI data. First, the evaluation process cannot easily be implemented. Evaluating VGI against a reference dataset (i.e. authoritative data) is not always possible, due to limited data availability, contradictory licensing restrictions or high procurement costs of the authoritative data. Moreover, internal or external quality cannot be easily assessed as the wiki-based nature of VGI data results in the absence of data specifications (Antoniou, 2011). Then it is the nature of VGI which paints a completely different picture from the one described earlier. In this front, the first element to consider is biases, both social and spatial ones: knowledge of language, users' available time, their technical capability, origin or cultural differences are all factors that introduce subtle or important biases in VGI datasets. Then is the digital divide that should make us very careful about the coverage and representativeness of the data that is being collected. A third element is the GI itself: lack of metadata, heterogeneity, patch work and fragmented contributions should be expected when using VGI. This includes also high volatility as frequent changes made by contributors in important attributes can deteriorate the overall quality and the usability of VGI datasets.

New methods for quality measures in VGI

In this context, researchers need to explore ways to determine VGI quality using existing methods and, in parallel, find new ways that will suit better the nature of VGI. The former group of efforts includes efforts that adapt the existing measures of spatial quality elements, discussed above, to VGI datasets. The latter refers mainly to research aiming to reveal intrinsic to VGI quality indicators, sometimes new to Geomatics domain, so to facilitate the understanding of such data. Here, we turn our focus to the novel evaluation efforts that use intrinsic VGI quality indicators. These indicators can be grouped in four different groups: i) data indicators, ii) demographic indicators, iii) socio-economic indicators and iv) indicators about contributors (Antoniou and Skopeliti 2015).


TEOREMA srl MILANO:
SOLUZIONI INNOVATIVE PER LA TOPOGRAFIA E LA TERMOGRAFIA "IR"
Dal 1986 Teorema srl lavora a fianco del professionista fornendo la tecnologia più avanzata, la migliore formazione tecnica ed una accurata assistenza post-vendita per rendere affidabile e produttivo il vostro lavoro.
Nova
Captivate
GNSS
Leica HDS P30 / P40
Leica Nova MS60
Leica Serie Captivate
Leica CS35 Rugged Tablet
Ricevitore GNSS GS14
Leica 3D Disto
Leica Disto S910
Termocamera Flir T420
- when it has to be right
Leica Geosystems
TEOREMA srl TOPCENTER
www.geomatica.it • www.disto.it • www.termocamere.com
Via A. Romilli, 20/8 20139 Milano • Tel. 02 5398739 • teorema@geomatica.it

Data Indicators. The direct quality evaluation can be problematic for VGI. This is because usually there are no detailed specifications or the evaluation against authoritative data might not be possible, not least because there is no access to reference data. Hence, the focus is on indicators that could reveal VGI quality by solely examining VGI data. Such indicators include features' length and point density in a square-based grids or feature-level attributes such as the number of versions, the stability against changes and the corrections and rollbacks of features, the provenance of contributed features

Demographic Indicators.
As VGI is user generated content, it is expected that a correlation between data quality and demographic data might exist. Empirical studies revealed the correlation between the demographics of an area and the completeness and positional accuracy of the data. Also, it has been shown that the low population density areas (i.e. rural areas) negatively affect the completeness of VGI data. On the contrary, population density positively correlates with the number of contributions, thus affecting data completeness or positional accuracy (see for example Zielstra and Zipf, 2010; Haklay et al, 2010).

Fig. 4 – OpenStreetMap GPS trace density. (Source: https://en.wikipedia.org/ Author: Eric Fischer).

Socio-economic Indicators.
Closely related to the demographics is the existing socio-economic factors. The grassroots engineering and the bottom-up process of VGI turned the research focus in socio-economic factors and indeed, it has been shown that social deprivation and the underlying socio-economic reality of an area considerably affects completeness and positional accuracy (Haklay et al, 2010; Antoniou, 2011). Similarly, other factors such as high income and low population age result into a higher number of contributions, a positive factor of VGI quality (Girres and Touya 2010).

Contributors' Indicators.
This group of indicators focuses on revealing the contributor's motivation drivers as this can give a better insight into user generated data. To this end, quality indicators can include the history and the profiling of contributors or the experience, recognition and local knowledge of the individual (Van Exel et al., 2010). Moreover, the number of contributors on certain areas or features has been examined and it has been positively correlated with data completeness and positional accuracy (Keßler and Groot, 2013).
VGI is a new development for the Geomatics domain. As such, some of the existing tools used so far for the quality evaluation of GI can be applicable here as well. However, is evident that the very nature of VGI imposes a broader thinking of how to be more inclusive so to better analyse the quality of VGI dataset.
As there are still ongoing efforts to build a solid framework that will efficiently assess VGI quality, there is active research around novel quality indicators.

Understanding VGI value

When we solely focus on measuring the quality of VGI data, we run the risk of missing the bigger picture that this phenomenon paints: the true value of VGI. Before VGI, spatial data was a privilege in the hands of governments or few corporations. Datasets where stored in silos and the vision of functional and public-serving SDIs was strangling to stay alive. What VGI did was to introduce geography to the general public, increase awareness of its value and consequently the demand for up-to-date spatial products; in a sense VGI managed to spatially enable our societies. Moreover, VGI sparked the creation of a virtuous circle around the linkage between society and spatial information. The technological advances facilitated spatial data collection and online diffusion, and this made people familiar with spatial content, cartographic products and location based services. This in turn, created the need for more, freely available, spatial content of high quality and thus VGI sources were better placed to cover this need resulting to more crowdsourced spatial content to become available on the Web. This positive spiral was also fuelled by the intrinsic characteristics of VGI data. First, is the fact that now we can record how people value and understand space. Now, for the first time, the user's perception of space is tangible through the volunteered recording of spatial features or phenomena they consider important to have on a map. Moreover, as daily life is local by its nature, VGI supports the recording of issues that range from health to entertainment, to education, or other local-scale activities. Closely related to this is the fact that VGI encapsulates the local knowledge that contributors have. Following

Tobler's law which states that "*everything is related to everything else but near things are more related than distant things*" (Tobler, 1970: 234) it is not strange that contributions in VGI tend to be more accurate in places the contributor knows best. Another issue is the extended field of scope. While VGI become mostly known from a handful of champion projects such as OpenStreetMap, Wikimapia and Geonames, examples include also data gathering for air pollution, urban noise, traffic and congestion maps, cycle maps, gpx-trail maps or soil mapping. Most of these topics were usually under the radar of the NMAs as their focus was on few well defined mapping products. This leads to the fact that now we can open our horizons to new geospatial products and applications. Examples can be found in the field of urban sensing and smart cities. Today, with ubiquitous sensor networks our living environments are being transformed into smart cities where the flow of VGI in terms of volume and currency opens the opportunity to monitor and understand, in an unprecedented way, what exactly takes place in every corner of the urban fabric. Apart from new products, VGI can also play an important role in correcting, enriching, and updating existing datasets. Furthermore, VGI shortens the time horizons of geographic data update as in most cases the time gap between data capture and data consumption is minimal. Finally, most of the times, all these come with no cost and without sophisticated and restrictive licensing agreements.

What is next?

The evolution and possibilities of VGI in the Geospatial domain have attracted the interest of academics and professionals alike with a growing number of governments and corporations leveraging this kind of GI information. However, it is clear that while, at this point, VGI cannot replace proprietary and authoritative data, it can play a crucial role in correcting, enriching and updating existing datasets or provide the basic information layer for new products.
VGI has been a growing phenomenon for over a decade now. Notwithstanding the acceptance that it has received so far, the most important factor that hinders a more widespread diffusion is the lack of a stable and standardized way to evaluate data quality. Existing and well-established methods and processes for spatial data quality evaluation, while still valid, are not always applicable to VGI datasets. This drives researchers and academics into the study of new methods so to eloquently answer the pressing question about "how good is VGI data?". The nature and the creation mechanisms of VGI led to the analysis of a number of factors. However, research is still far from providing concrete answers and methods regarding the evaluation of VGI quality. Here, we just scratched the surface of the ongoing research on VGI quality evaluation.
Now, if we had to provide a prediction for the future, which is always a challenging task, it would be that the times ahead will get extremely interesting in this field. This optimistic view is based on the trends which more or less serve as the driving forces of VGI: technology and society. In the technological front, the evolution in Information and communications technology (ICT) will not leave VGI unaffected: bandwidth will keep increasing, the cost of hardware will keep dropping and the number of people online will keep growing and thus the pool of contributors will become larger and better equipped. This alone is great news for the future of VGI. However, the most crucial role is expected by the spatial data capturing devices that will proliferate or be introduced in the future. On the one hand is the ubiquity of sensors that passively collect spatial data, mostly in urban context. The transformation of our living environment into smart cities inevitably passes through a better understanding and a more detailed recording of space and human activity. This development is based on the consideration that location and spatial information are common goods and promotes their availability in order to stimulate innovation (Roche et al. 2012). Then, is the individually controlled devices. The spread of drones, for which we are still exploring their abilities to contribute in systematic data gathering, is expected to bring VGI in a whole new level. Moreover, the evolution of the wearable technology, while still in its early days, is expected to contribute to the evolution of VGI. The omnipresence of wearable sensors is expected to multiply the availability of spatial data on the Web. Similar impact is expected by the development of indoor positioning and mapping systems (e.g. Google's Tango project) which will extend VGI into new fields. So, in short, GI capturing devices, on top of what it is today available, will cover also the area of aerial surveying, of everyday activities and of indoor mapping, and this is just a sneak preview of the near future.
In the societal front, the future could be even more exciting. Crowdsourcing, volunteerism, citizen science and social enterprises are just some of the early formations which the increased online connectivity has brought.

It is really amazing how online communities address real world problems and even more impressive how this grassroots collaboration overcomes societal barriers and enables citizens to participate in the management and improvement of quality of life. The social transformation shaped by online communities will prove equally important factor in the evolution of VGI as the technological advances.

How this ecosystem affects the understanding of VGI quality? We need to understand that this area is highly interdisciplinary in that intertwines the advances of many domains. VGI is the grafting of the underlying social, economic and technological situation with the geospatial domain. It is incarnated with the tangible recording of citizen's perception for space and phenomena they consider important to have on a map. However, despite the work and empirical research available on the subject of VGI quality, a solid framework for assessing the quality of crowdsourced spatial data is far from being established for all the reasons explained here. This should be the next goal for VGI on our way towards Digital Earth.

REFERENCES

Antoniou, V. (2011). *User generated spatial content: an analysis of the phenomenon and its challenges for mapping agencies*. Ph.D. Thesis, University College London.

Antoniou, V. and Skopeliti, A. (2015). Measures and Indicators of Vgi Quality: An Overview. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, II-3/W5, pp.345-351.

Girres, J. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4), pp.435-459.

Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), pp.211-221.

Haklay, M., Basiouka, S., Antoniou, V. and Ather, A. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47(4), pp.315-322.

International Organisation for Standardisation 2005. *ISO 9000: Quality management systems - Fundamentals and Vocabulary*, Geneva: ISO.

International Organisation for Standardisation, 2013. *ISO19157:2013 Geographic information - Data quality*, Geneva: ISO.

Keßler, C., de Groot, R. T. A., 2013. Trust as a proxy measure for the quality of Volunteered Geographic Information in the case of OpenStreetMap. In: *Geographic information science at the heart of Europe*, Springer International Publishing, pp. 21-37.

Roche S, Nabian N, Kloeckl K and Ratti C (2012). Are 'smart cities' smart enough? In: *Rajabifard A and Coleman D (eds) Spatially Enabling Government, Industry and Citizens: Research Development and Perspectives*.

Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, p.234.

Turner, A. (2006). *Introduction to Neogeography*. Sebastopol, Calif.: O'Reilly.

Van Exel, M., Dias, E., Fruijtier, S., 2010. The impact of crowdsourcing on spatial data quality indicators. In: *Proceedings of GiScience 2010,* Zurich, Switzerland.

Zielstra, D., Zipf, A., 2010. A comparative study of proprietary geodata and volunteered geographic information for Germany. In: *Proceedings of the Thirteenth AGILE International Conference on Geographic Information Science*, Guimarães, Portugal.

PAROLE CHIAVE

VGI; DATASETS, GI; SPATIAL DATA QUALITY

ABSTRACT

Oggi la Volunteer Geographic Information (VGI) è onnipresente. Proviene da varie fonti e può essere trovata in molti aspetti tra cui toponimi, tracce GPS, foto geo-tag, applicazioni di social networking, blog, misurazioni dei sensori, mappe topografiche ecc. Può essere il frutto di un lavoro sul campo o di bulk data importati da un dataset autorevole, disponibile gratuitamente. E' ovvio che tutte queste fonti non possono solo cadere in un'unica categoria. Ci sono molti aspetti che possono essere esaminati qui, ma di particolare interesse, quando si esaminano le fonti e le tipologie di VGI, sono l'origine, la motivazione e lo scopo di chi contribuisce alla VGI.

AUTORE

VYRON ANTONIOU
V.ANTONIOU@UCL.AC.UK
HELLENIC MILITARY ACADEMY, GREECE