Quality in User Generated Spatial Content: A Matter of Specifications

Carmen Brando, Bénédicte Bucher
Université Paris Est, IGN-COGIT
73 avenue de Paris. 94160 Saint-Mandé, France
{carmen.brando-escobar, benedicte.bucher}@ign.fr

1. INTRODUCTION

In the last decade, the burst of the Web 2.0 has rapidly increased the amount of user generated content (UGC) on the Internet. Most of Web 2.0 applications allow users to interact with other users and edit website content. This collaborative edition is supposed to improve the content. Nowadays, the amount of people generating online content has dramatically increased. For instance, in the U.S., recent studies show that 35% of Internet users have created content (Flanagin, 2008). A widespread example is the online encyclopedia Wikipedia. It is a valuable source of information though it is not intended to users searching for verified information because it is often anonymous.

In this context, there has been a growing interest for online communities where users contribute to the creation of spatial content (SC). SC is any content with a spatial dimension. It may be created through annotating given resources with coordinates, such as pictures in Flickr, Wikipedia entries or crime activities in Ucrime, that is, geotagging. SC may also be directly created by editing geographical features like in Openstreet Map (OSM), Wikimapia and The People's Map. These projects belong to a movement, which has been baptized by Michael Goodchild as Volunteered Geographic Information (VGI) (Goodchild, 2007). Citizens are presented as observers and have no specific expertise in geomatics. The term neogeographers has been coined to describe them (Turner, 2006). Nevertheless, there have been several debates about the term *volunteered*, because it implies an altruistic gesture in the contributors' intentions. (Antoniou, 2009) designate the term User Generated Spatial Content (UGSC) as a more generic and appropriate designation.

The focus of this paper is data quality in UGSC. An important notion both to improve quality during production, and to help in providing quality metadata for users is that of specifications. They are the most detailed available source of knowledge about geographic databases content (Abadie, 2009). The rest of this paper will discuss important definitions and related work concerning quality in (2) User Generated Content (UGC) and (3) UGSC. Then, an approach to qualify UGSC is proposed in (4), based on specifications. Finally, (5) our conclusions and perspectives are presented.

2. QUALITY IN UGC

Quality is defined in ISO 9000, as the degree to which a set of inherent characteristics fulfills requirements. In other words, content quality is quite related to whether this content is useful or not to user's purposes. In UGC, quality is also associated to the user's trust in the content (a subjective concept) which leads to a connection between content quality and provider's authority.

There are several ways to improve quality during the content edition. Firstly, as in Wikipedia, citing external sources is an important quality criterion because it enhances readers' trust in the article. Another mechanism is the management of consistency during collaborative edition. A moderation mechanism is also used in which users can set privileges on their articles. It includes as well a revision control system for logging editing operations on articles. An important feature is that of conflicts edition, that is, when several users work concurrently on the same article. This is usually

detected by the system, but it must be manually resolved by those users, by comparing different revisions and documenting in discussion pages. More advanced collaborative systems attempt to resolve automatically these conflicts through timestamp (Tlili, 2008) or semantic (Preguiça, 2003) reconciliation.

More lately, Google has been involved in creating an on-line encyclopedia, named Knol (http://knol.google.com). They have made improvements to try to improve quality in their articles, mainly based on authority and comprehensiveness. Every knol is strictly associated with its (their) author(s); and the information on the author(s) profile is clearly shown, if available. Due to the non-anonymity condition, users are encouraged to make great efforts in redacting a knol. On the one hand, this is intended to increase articles quality through authors' well-known contributions. On the other hand, it might discourage others to contribute.

Documenting quality is another important aspect for UGC usability. Many works have proposed different procedures to qualify Wikipedia articles. Some authors have suggested metrics for measuring quality, such as number of edits, number of unique editors, intensity of cooperative behavior and analysis of featured articles (Wilkinson, 2007; Stvilia, 2005). Furthermore, in (Mcguinness, 2006), the authors present a trust tab which is associated with each Wikipedia article. The aim is to allow users to visually compare quality of fragments of an article based on the background color. The color corresponds to a certain extent, which is chosen by considering user activity in that article.

Quality may sometimes be measured by comparison with a so-called reference resource, whose quality has been verified. In a study presented in the journal Nature, Wikipedia and Encyclopedia Britannica were compared to determine a credibility measure. Their results show that Wikipedia came close to the level of accuracy in Encyclopedia Britannica and had a similar rate of "serious errors". These claims were disputed by the latter.

3. QUALITY IN UGSC

Quality of Spatial Content

Quality of SC tends to be a concept quite complex probably because there are several points of view on this concept.

The user's point of view is fitness for use. It is based on user's requirements and intended use of the content. It is often called *external* quality (Devillers, 2006). In order to determine whether a geographic database really fit their needs, users must understand its content in detail. To do so, they cannot only rely on data. They also need metadata to understand what aspects of the real world the provider has represented (if a house is absent from the data, does it mean there is no house in the real world or does it mean the provider did not observe too small houses?).

The producer's point of view is the degree of similarity between the data and the representation of reality he has intended to build. It is often called *internal* quality (Devillers, 2006) and is documented both by:

- The product specifications which somehow depict the producer filter of the reality: what he means to represent and how (Mustière, 2003). As an example, the specifications of the topographic database (BDTopo®) of the French National Mapping Agency (Institut Géographique National –

IGN) indicate the valid domain values (in meters) of the attribute height for industrial buildings: (1: 1-3, 2: 3-5, 3: 5-10, 4: more than 10). They also describe complex relations between objects, such as a composition between a region and its counties or a conjunction between two cities. Detailed specifications ensure that data capture will be homogeneous even if the people are different, because they respect the same guidelines.

- Error criteria describing the gap between a given data set and its specifications, such as exhaustiveness of a specific type of feature. This gap is often summarized through specific variables according to the ISO19115 model: logical consistency, temporal accuracy, thematic accuracy, purpose, usage, lineage, positional accuracy and completeness.

Specifications are an important item with respect to SC quality. Firstly, specifications are useful during content creation because they ensure an homogeneous observation of the real world and may entail specific modeling elements to facilitate the design of a consistent content.

To illustrate this, Figure 1 displays an excerpt of OSM data (taken on December 2009) which could have been improved using specifications, by enhancing the spatial representation of these features. It corresponds to the administrative boundaries and waterways content around the French city of Grenoble. Evidently, there is a lack of geometrical consistency between those feature types. The data could be more consistent if the model supported sharing of geometry between features. A specification indicating the relation between these two features could place automatically the line boundary throughout the middle of the waterway. Considering this specification, if a user inserts a line boundary as shown in the figure, the system should be able to evaluate the correct placement of the line and propose the correction to the user; then he/she could agree or not.

Figure 1: OSM data corresponding to administrative boundaries and waterways around the French city of Grenoble.

Secondly, to assess a SC fitness for use, the user needs both the SC specifications and quality variables, (i.e. metadata). Non-formal organizations cannot always afford the cost of collecting, redacting and updating these specifications. Nor do they always know modeling techniques to ensure a SC consistency. For this reason, assisting users to provide specifications of their content is an important factor to consider. Obviously, many neogeographers would be reluctant to provide and work according specifications. Usually, their main concerns are not, for instance, logical consistency. But there are communities whose quality requirements are strong enough to accept some rules during

contribution. Usually, special interest groups (e.g. civic/governmental) have more focused requirements (Coote, 2008).

New Production Processes to Address Quality Issues in UGSC

UGSC is currently seen as a new production process that can enhance the quality of SC. It plays a strong part in updating the data. Navtech and some NMAs, rely on user generated alerts as a relevant source of alerts about content errors and data fix.

More recently, the focus has been put on facilitating the very edition of SC by users. Projects like OSM have had a great influence in the way SC is produced and updated nowadays. Moreover, Google are now providing their own SC for Google Maps in some parts of the world in order to crowdsource their data correction process (Batty, 2009). In moments of crisis, to rely on UGSC is very useful to provide, as soon as possible, updated content. For instance, Google have been encouraging local contributors to specify information on areas with difficulties after the earthquake in Haiti, using Google Map Maker. The idea is to aid in the logistics of delivering rescue through emergency maps (New York Times, 2010).

Besides, UGSC is also seen as a way to complement NMA's data, beyond simple alerts (Budhathoki, 2008). One of the main roles of UGSC could be to elaborate patchworks for NMAs (Elwood, 2008). These organizations should provide standards and protocols to create a composite coverage depending on users' needs. In this way, users could participate in updating their existing mapping products (Antoniou, 2009).

Qualification of contributors and contributions

Many approaches tackle quality on UGSC by qualifying contributors and contributions. There are several propositions which present a classification of users based on their purposes (Coleman, 2009), their geographic locations (Goodchild, 2009), or their trust relations with other users (Bishr, 2007). The aim is to distinguish between a high value and low value/fraudulent contribution. In this way, the former is embraced and the latter is discarded. Contributions can also be validated through rating systems, which implicitly assign reputation to contributors (Elwood, 2008). Users can also evaluate content by marking regions, which in their view lack of contributions and require more work (Maué, 2008).

Several studies have been done to evaluate UGSC quality with respect to a reference data set obtained from NMAs. Specifically, English OSM data have been compared to Ordnance Survey data (Haklay, 2009). These studies put special attention on positional accuracy and completeness as quality criteria. It showed a very good coverage in major cities, but poorly as you move far away. Even though these studies do not present a proposition to improve quality, they compile very valuable documentation related to OSM data quality. These studies could be seen as part of quality testing programs to increase credibility of UGSC (Goodchild, 2009).

4. PROPOSED APPROACH

Our approach aims at improving quality management in UGSC based on formal specifications and on external reference data. Quality management refers here both to improving quality during production and to providing quality metadata for the users. Formal specifications facilitate quality management during contribution in three ways: they entail integrity constraint (1) to support on the

fly consistency checking like in (Mäs, 2007), (2) to improve quality of UGSC through external reference data (i.e. IGN), and (3) to reconcile concurrent editions of data. More specifically, this approach relies on several components illustrated in Figure 2.

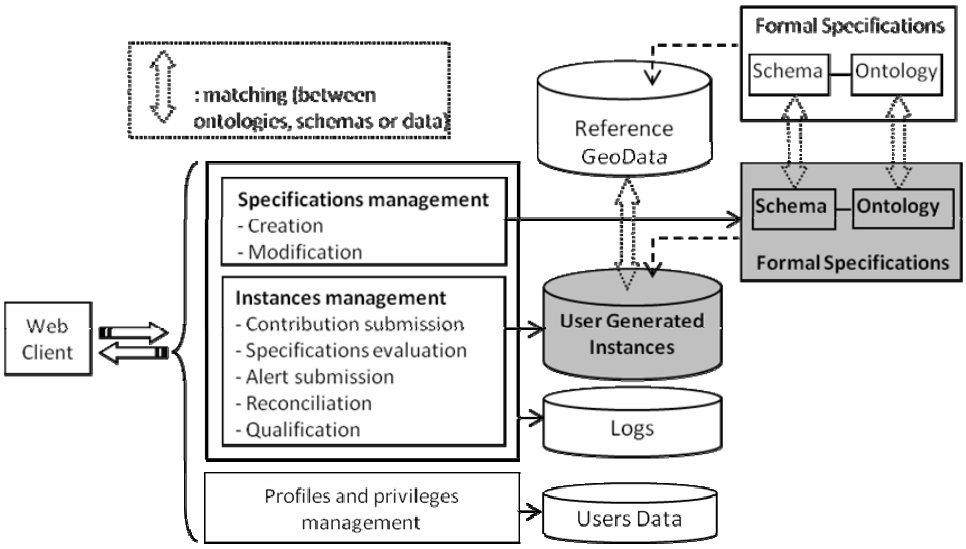

Figure 2: Component Diagram of the proposed approach to qualify UGSC.

The first component is dedicated to *users' profiles and privileges management*. In this way, a user may connect through a login component. This will be necessary to qualify contributions and integrate them into the content. The *UsersData* database contains users' information and their privileges over features in the *UserGenerated Instances* database (UGSC).

The user may then use the *specifications management component*, in charge of the edition of formal specifications for the UGSC. These specifications (Abadie, 2009) should contain the community ontology of concepts of the real world, and how these concepts are encoded. Formal models for expressing geographic database specifications have been proposed (Gesbert 2004, Christensen 2006). Importantly, we mean to explicit, whenever it is possible, relationships between these specifications and the specifications of *reference data*: to relate similar concepts or express integrity relationships between UGSC and reference data. Therefore, in the matter of quality and authority, using specifications on UGSC and reference data may bring the best of both worlds.

The last component is in charge of the edition of the data themselves: the *Instances management*. A user may submit a contribution, which is any proposition to create or modify a feature. Then, the system is able to evaluate this contribution considering specifications and possible logical relationships with the reference data (thanks to the alignment between both specifications). If it detects an inconsistency he may suggest a modification of the contribution. These steps will rely on spatial analysis algorithms for data matching developed in (Bucher, 2009), but also for proposing reparations. This component also manages true collaboration in a wiki manner. Some wikis (http://concerto.xwiki.com) implement a cooperative mechanism to reconcile conflicting operations performed concurrently on related items. For instance, let us consider a user who inserts a new segment to a roadway R. At the same time, another user changes the name of R to R'. The system can optimally execute these two operations in the right order. In this case, it would be appropriate to fill

the attribute name to R'. Afterward, when executing the other operation, the new segments inherit the new name R', instead of R.

Users can also simply submit alerts if they detect mistakes according to their knowledge. When considering several contributions over a feature or related features (through a constraint, for instance), the system performs a reconciliation process in order to merge these contributions and propose a common one. To do so, it may rely on a log of all operations performed by the concerned contributors.

Lastly, this process supports the qualification of the UGSC with respect to the corresponding formal specifications. Qualification relies on the *Logs* information but also on a comparison with reference data for which the quality has been assessed by an official NMA. When it comes to documenting SC quality, metadata standards are available for traditional SC (ISO 19115) but there is no agreement on what metadata should be provided for UGSC. A priori, our interest is not to set these guidelines; otherwise, to eventually elucidate which elements should be considered in UGSC metadata.

5. CONCLUSION AND PERSPECTIVES

UGSC refers to a new paradigm for producing spatial content. Managing quality of UGSC is being more and more crucial as this content is growing. In this paper, we presented an on-going work which aims at extrapolating the way a NMA pretends to manage quality of spatial data in the context of community based collaborative edition. Our approach relies on the explicit and formal definition of content specifications. Specifications are the most detailed available source of knowledge about geographic database content. In our approach, users improve UGSC quality by contributing with SC in a collaborative manner, and also by describing this content in the form of specifications to make it fit their purposes. Besides, in a context where reference data are available, our approach supports the evaluation of the UGSC by comparison with these data.

Finally, the perspectives are to implement a prototype, which aims to integrate several works and tools conceived and developed at COGIT laboratory. Evidently, the main purpose is to examine the feasibility of our approach. An experimental study to evaluate the results obtained is contemplated, considering IGN data.

BIBLIOGRAPHY

Abadie, N., 2009. Formalisation of Geographical Database Specifications, in Proceedings of AGILE workshop.

Antoniou, V., Morley, J. and Haklay, M., 2009. The role of user generated spatial content in mapping agencies, in GISRUK.

Antoniou, V.; Morley, J. and Haklay, M., 2009, Do photo sharing websites represent a sufficient database to aid in national map updating or change detection?, in EuroSDR Workshop on Crowd Sourcing for Updating National Databases.

Batty, P., 2009, Google shakes up the geospatial data industry, October 7th. http://geothought.blogspot.com/2009/10/google-shakes-up-geospatial-data.html.

Bishr, M. and Kuhn, W., 2007, Geospatial Information Bottom-Up: A Matter of Trust and Semantics, in Proceedings of AGILE: 365--387.

Bucher, B., Brasebin, M., Buard, E., Grosso, E. and Mustière, S., 2009, GeOxygene: built on top of the expertness of the French NMA to host and share advanced GI Science research results, in International Opensource Geospatial Research Symposium.

Budhathoki, N., Bruce, B. and Nedovic-Budic, Z., 2008, Reconceptualizing the role of the user of spatial data infrastructure, GeoJournal 72(3), 149--160.

Christensen, J. (2006), Formalizing Specifications for Geographic Information, in Proceedings of AGILE: 186--194.

Coleman, D., Georgiadou, Y. and Labonte, J., 2009, VGI: the nature and motivation of produsers, International Journal of Spatial Data Infrastructures Research 4, 332--358.

Coote, A. and Rackham, L., 2008, Neogeographic data quality – is it an issue?, in AGI Geocommunity conference, ConsultingWhere Ltd.

Devillers, R. and Jeansoulin, R., 2006, Fundamentals of Spatial Data Quality, ISTE.

Elwood, S., 2008, VGI: future research directions motivated by critical, participatory, and feminist GIS, Geojournal 72(3-4).

Flanagin, A. and Metzger, M., 2008, The credibility of VGI, Geojournal 72(3-4), 137--148.

Gesbert, N., 2004, Formalisation of Geographical Database Specifications, in Proceedings of ADBIS.

Goodchild, M., 2009, NeoGeography and the nature of geographic expertise, Journal of Location Based Services 3(2), 82--96.

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

Haklay, M., 2009, How good is OpenStreetMap information? (to be published), Environment & Planning B.

Mäs, Stephan: Reasoning on Spatial Semantic Integrity Constraints. In: Conference on Spatial Information Theory, Lecture Notes in Computer Science 4736, S. Winter et al. (Eds.), pp. 285–302. Springer, Heidelberg (2007)

Maué, P. and Schade, S., 2008, Quality Of Geographic Information Patchworks, in Proceedings of AGILE.

Mcguinness, D., Pinheiro Da Silva, P., Zeng, H., Ding, L., Narayanan, D. and Bhaowal, M., 2006, Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study, in Proceedings of workshop on Models of Trust for the Web.

Mustière, S., Gesbert, N. and Sheeren, D., 2003, A formal model for the specifications of geographic databases, in Proceedings of GeoPro workshop.

New York Times, 2010, The Haiti Earthquake By Dan Fletcher, January 14th. http://www.time.com/time/specials/packages/article/0,28804,1953379_1953494_1953677,00.html.

O'Reilly, T., 2005, What Is Web 2.0?, http://oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

Preguiça, N., Shapiro, M., Matheson, C., 2003, Semantics-Based Reconciliation for Collaborative and Mobile Environments, in CoopIS/DOA/ODBASE: 38-55

Stvilia, B., Twidale, M. B., Smith, L. C. and Gasser, L., 2005, Assessing Information Quality of A Community-Based Encyclopedia, in Proceedings of the International Conference on Information Quality.

Tlili, M., Kokou Dedzoe W., Pacitti, E., Valduriez P., Akbarinia R., Molli, P., Canals G., Laurière, S., 2008, in Proceedings of VLDB 1(2): 1420-1423.

Turner, A., 2006, Introduction to Neogeography, O'Reilly Media.

Wilkinson, D. M. and Huberman, B. A., 2007, Cooperation and quality in Wikipedia, in Proceedings of WikiSym, ACM, New York, USA, pp. 157--164.