CHAPTER 10

Enhancing the management of quality of VGI: contributions from context and task modelling

Benedicte Bucher*, Gilles Falquet†, Claudine Metral† and Rob Lemmens‡

*Université Paris Est, IGN, France, Benedicte.Bucher@ign.fr
†University of Geneva, Switzerland, Gilles.Falquet@unige.ch, Claudine.Metral@unige.ch
‡University of Twente, The Netherlands, R.l.g.Lemmens@utwente.nl

Abstract

This chapter presents contributions to managing the quality of Volunteered Geographical Information (VGI) and of crowd sourced geographical information (CSGI) brought by the representation of specific knowledge items: task and context. Task and context modelling have been studied in different communities. We propose an approach for integrating their results with the perspective of improving the quality management of VGI and CSGI.

Keywords

Task modelling, context, quality, specification, user-generated

Introduction

The ENERGIC COST ACTION targets the usage of Volunteered Geographical Information (VGI) and of crowd sourced geographical information (CSGI) in

How to cite this book chapter:
Bucher, B, Falquet, G, Metral, C and Lemmens, R. 2016. Enhancing the management of quality of VGI: contributions from context and task modelling. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*, Pp. 131–142. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.j. License: CC-BY 4.0.

scientific applications. One challenge addressed in this action is data quality management.

Data quality management has been addressed for several years with respect to classical geographical data, by operational bodies like the National Mapping Agencies and by scientists mainly from the geographical information science community. The issues they tackle are also relevant in the context of volunteered and user generated data, for example managing the lack of a universal data model about geographical space and the unavoidable heterogeneities between geographical data. Section 1 lists findings about geographical data quality management from the literature and from the experience of the French National Mapping Agency, IGN. These findings relate to the management of data specifications, the definition and documentation of quality criteria, the assessment of which inherent characteristics of the data will impact the output of a given application result and the communication of quality to the user.

Section 2 specifically studies the potential of context and tasks modelling to implement these findings in the context or VGI and CSGI. Context can account for much heterogeneity in VGI and CSGI. Tasks are useful pieces of knowledge to plan the usage of relevant resources to achieve an objective. Context and tasks modelling are studied by communities tackling information management and exchange between implemented components and humans, like distributed architectures, interoperability, ubiquitous mapping, location based services and human–machine dialogue interfaces.

An approach to integrate the context and tasks models to address part of the research questions expressed in the beginning of this chapter is discussed at the end of this chapter.

Geographical data quality management

External quality

Quality is defined in ISO 9000 as the degree to which a set of inherent characteristics fulfils some requirements. This definition of quality is relative to an application. For example, important inherent characteristics of 3D data for visualization applications refers to accurate and realistic textures as well as consistency of visible shape elements and very low level of detail for elements non-visible in the current scene. Important characteristics of the same data for firemen access application requirements refers to the exhaustiveness and geometric accuracy of specific features like windows, electricity cables and tramway cables. This quality is referred to as 'fitness for use' or 'external quality' (Devillers & Jeansoulin 2006).

Most applications considered in ENERGIC action share some common data requirements:

- the ability to discover and reuse the data,
- the ability to combine the data with other data,

- the ability to ground a result on data, i.e. to control the gap between data and interpretation.

The two first criteria are recurrent and are promoted by initiatives like the INSPIRE directive or by the W3C vision for Linked Open Data where data should be produced once and made available (Bizer, Heath & Berners Lee 2009). The third criterion relates to having a documentation of uncertainty related to the data and how it propagates to the result. Besides, when it comes to volunteered content, it is also useful to consider requirements expressed by the Web community. W3C propose a ranking scheme for open data that list important quality criteria from their perspective: publication on the web (one star), in a machine readable format (two stars), in a non-proprietary format (three stars), compliant to RDF standards – using dereferenceable URIs to name things (four stars) and publish links to other URIs (five stars).

With respect to the above requirements, external quality of data will very much depend on metadata and documentation.

Besides, user requirements will eventually be met by an application involving software and data. Hence, geographical data quality assessment is closely related to geographical software quality assessment.

Internal quality

A specific intermediate quality concept is needed to document inherent characteristics of geographical data that will be useful for every user to evaluate their ability to fulfil their application requirements. This is the 'internal quality' (Devillers & Jeansoulin 2006). The data producer should distribute its data together with the description of this internal quality and the users can at the end use this description (and the data) to assess external quality of the data for their application. Indeed, many geographical data (base maps for instance) have seldom been acquired for one specific application but rather to be reused in several applications, and possibly by users who sometimes are far from the production of the data and who did not express their quality requirements, hence did not express which inherent characteristics are important for them.

Internal quality has been traditionally documented by national mapping agencies, and in current ISO/OGC metadata standards (ISO TC211 2014) based on three elements:

- the targeted description, called the data specifications,
- quality criteria describing some distance between the produced data and an imaginary flawless data sets compliant with these specifications,
- the lineage metadata.

In other words, data producers have considered that the characteristics that will help future users assess the 'external' quality of their data are globally the

assessment of how far the geographical data are from the reality they aim at describing. These three items are described more precisely hereafter.

The specifications are the scope of the observation process that will lead to the data: when, where, what objects, what level of detail. It is crucial to explicitly describe specifications because there is no natural such abstraction even though one may intuitively think so. Our world is heterogeneous in multiple ways whereas an abstraction will provide only one specific set of categories and classification schemes. The most satisfying solution so far is to provide abstractions that are relevant for a certain spatial and temporal scope and also for a given point of view on reality. Not only are specifications an important item for describing quality, but also they improve the homogeneity of the data during data production. It has always been a necessity and a challenge to share specifications among operators involved in the acquisition process for national mapping agencies producing topographical data over a national territory (Sheeren, Mustière & Zucker 2004). Indeed, geographical database specifications should refer to a common ontology of reality which does not exist (Abadie 2009).

Quality criteria are measures of distance between produced data and what is called 'terrain nominal', i.e. data that would have been produced strictly considering the specifications (and in real time). When a quality criteria is attached to a product (and not to a specific data set), it means a commitment of the producer to respect a certain thresholds during data production. Quality criteria describing 'uncertainties' and 'errors' possibly introduced during the actual production have been standardized after four fundamental dimensions: positional accuracy, attribute accuracy, logical consistency, completeness (Goodchild & Li 2012). Usually a product description includes some commitments of the producer about these criteria threshold. For example such metadata for a road data product can be: the product should describe every road longer than 50 m, thanks to a series of points acquired at the axis of the road, with 10 m precision, with time accuracy of 6 months and exhaustiveness of 98% on the national territory. Whereas these examples refer to explicit attributes and entities, it is also important in geographical data to consider some implicit spatial properties and relationships. An important paradigm of geographical data is that many relations and properties are not explicit in the data but can be computed based on the coordinates. Several authors study the evaluation of some spatial properties and relationships, usually referred to as spatial consistency rules (Servigne et al. 2000). A recurrent quality criteria referring to consistency is the topology.

Last, the lineage metadata refers to sources data and processes that led to the data. It is somehow comparable to the 'source code' of a software that will be useful to debug i.e. if something unexpected happens in the application to investigate if it can be explained by the geographical data production process.

Quality challenges intensified by VGI and CSGI context

Whereas by essence, VGI and CSGI production process can be seen as an opportunity to improve some dimensions of internal quality like the update frequency, this production process also makes it more difficult to handle certain dimensions of internal quality.

The main flaw in our opinion is the weak definition and documentation of expected specifications of the produced data. Besides, Linus's law 'given enough eyes, all bugs are shallow', referring to the ability of the crowd to converge on the truth, does not always work for VGI. Goodchild and Li (2012) and Haklay et al. (2010) have shown that users do not always agree on a value. This is also a motivation for explicitly stating quality specifications. Brando and Bucher (2010) focus on the definition of such specifications for user generated geographical content prior to the production of the data. They proposed a method to instantiate specifications based on OSM tags, Wikipedia infoboxes and the NMA product specifications (Brando, Bucher & Abadie 2011). Yet, their work does not address the issue of acceptability of these specifications by contributors and of evolution of such specifications.

One aspect of the proposal concerns the establishment of explicit consistency rules between the user generated content and reference data provided by the French national mapping agency, a public funded professional organizations who commit to reach specific level of quality criteria for some image data and topographic themes. Acquiring external rules that can be used to evaluate consistency of geographical content now is more generally an important domain of research (Goodchild & Li 2012) and has led to the creation of the organization OSMGB in UK which aims at listing such rules and setting up a formal quality insurance model to improve the trust of local administration in collaborative geodata.

Another flaw is the lack of explicit commitment to follow these specifications and reach quality criteria thresholds, e.g. of any update frequency, and the lack of assessment of quality criteria to document the gap between acquired data and the specifications. So far, documentation of quality criteria is done in punctual studies, like for research about the quality of VGI data comparing OSM data with data whose quality already is documented, like Haklay (2010) in the UK and Girres and Touya (2010) in France. Goodchild and Li (2012) and Haklay et al. (2010) also showed evidence that there are not always enough people interested in a particular area or feature.

To conclude this first section of the chapter, managing quality of VGI and CSGI can benefit from knowledge gained about the management of quality of geographical data. The quality of a data set is documented either according to a dedicated application or in a more generic way as a distance between a flawless ideal representation of a geographical space conforming to a given abstract model and a data set produced by remote sensing, in situ sensing and symbolic

knowledge production. It is highly recommended to define explicitly a targeted abstraction of a geographical space and it is easier to try and provide some that is 'locally' relevant. A relevant abstract model should not only be composed of classification schemes but also of consistency rules. Documenting the distance between a targeted abstraction and a dataset cannot be done exactly but is approximated by: quality criteria (exhaustiveness of a feature or of an attribute, and so on), lineage information (also known as provenance metadata).

When it comes to VGI and CSGI specific stakes are:

- the actual description and maintenance of data specifications,
- the shareability of specifications among the contributors, among users and the possibility to compare and align the model it with other data models,
- producers' commitment to quality criteria.

Contributions brought from context and tasks modelling

This section lists some contributions to address the objectives of quality management listed just before.

Context modelling

Firstly, since there is no such thing as a universal widely shared abstract model of reality, we advocate it is better to keep the data as close as possible to their production process (typically to keep sensor data) with context information that explain the data (see Chapter *Enquiring VGI*) then trying to merge every contribution into a pivot model. In this perspective, context modelling is an important metadata to account for much heterogeneity in VGI and CSGI.

Some *context* elements are already studied in the literature about VGI to infer quality and trust metadata like the contributor profile, his status within the VGI system (normal/advanced user in Wikimapia, normal/sysop in Wikipedia, ordinary/Data-working-group in OSM), his motivation and level of quality requirements with respect to data (Coleman et al. 2009), the places they live in (Goodchild 2009) (Bishr & Kun 2007), their relationships with other contributors (Bishr & Kuhn 2007).

Other relevant elements are studied in the domain of location based services and ubiquitous mapping where *context* is an important element to understand how someone may mentally interact with an abstract representation – usually accessible through a visual representation- of his surrounding, which are the time of the day, the season, the user age, nationality, gender (Jakobsson 2002) and culture (Edsall 2007).

Another very important *context* element is the contributor intention. In collaborative content edition, it is described through the effects of the contribution

on the content at the moment when the contribution was defined (Sun et al. 1998), for example: refining a shape, fixing an alignment between two features, adding a missing building. Describing the intended effect on the representation requires some abstract model of reality that must be as close as possible to the model the contributor had in mind. This refers to the possibility for the contributor to annotate contributions with a shareable abstract model. To enhance the interoperability of abstract models, it is now encouraged to publish them as 'vocabularies' on the web of data, i.e. as RDF schemas available online thanks to dereferenceable URIs. RDF vocabularies to distribute and share geodata are studied in the geographical information domain and in the semantic web community (Goodwin, Dolbear & Hart 2008; Vilches-Blázquez et al. 2010; Atemezing et al. 2014).

Task modelling

Tasks models organize knowledge about the usage of relevant resources to achieve an objective. In the context of VGI and CSGI quality management, this is useful with respect to modelling three kinds of tasks:

- the usage of space by a citizen when he is producing data, for instance going to work – and producing a GPS track,
- the collaboration or cooperation between citizens to produce data, for instance the organization of edition during a mapping party,
- the user task that requires geographical data, for instance evaluating the impact of a new road on the local biodiversity.

The first kind of task is an element of context that is useful to elicit the abstract model people have in mind when they produce data –the last *context* element mentioned in section above-. As demonstrated by (Gibson 1979), people see the landscape through his functional relevance to their goals. In other words, if a contributor rides a bike he will see the street from a different perspective than if a contributor is in a wheeling chair.

The second kind of task has been studied by (Das et al. 2014) who experimented with a task assignment model to organize the production of one content among several contributors to optimize exhaustiveness, cost and precision. The production is modelled as a task decomposed into subtasks that can be assigned to people. The system requires user profiles to make the assignment based on user expertness and availability, and define the reward they need. There exists relevant work in the literature to guide strategies for collaborative geographical data production. Wilkinson and Huberman (2007) study the nature of the collaboration that will impact the quality of the produced content. Maué and Schade (2008) propose a solution where contributors ask themselves for reviewers when they lack confidence in their own contributions.

In the domain of model collaborative edition, some authors have proposed a model were user contributions are directly expressed as operations and not as a new content, in order to be as close as possible to contributor intention. In Brando, Bucher and Abadie (2011), user edition can be expressed as the enforcement of relationships (i.e. implicit information) instead of geometries because the authors thought users may be more expert in assessing relationships between objects than geometries.

Rehrl et al. (2013) proposed a task/operation based model to analyse user contribution to a collaborative geographical content.

Last, task-based application design can be useful to express external quality criteria. The application is modelled as a task which has pre and post conditions, input and output data (Sun & al. 2012). A task also has a method to decompose high level tasks into elementary tasks, noting that these can be either machine tasks (computation) or user tasks (e.g. finding a geographic feature on a map). As an example let us consider the task 'find a restaurant'. This task is associated to subtasks such as (1) 'consult the list of all restaurants in a given area', which requires the completeness in the area, with an accuracy of 10 m, (2) 'find route to address', which requires a traffic network representation that is topologically correct and complete. The evaluation of fitness for use can benefit from the development of typologies and ontologies of tasks performed on spatial data. Several researchers have already worked in this direction. For instance, von Hunolstein and Zipf (2003) define a task typology in map-based mobile guides: high-level tasks have been associated to subtasks and a mapping between goals and tasks has also been defined. For example the task 'Navigation' is associated to subtasks such as 'routing from point A to B' and to goals/purposes 'navigation, exploring, planning, education. Park, Yoon and Kwon (2012) present a task ontology for intelligent tourist information service, based on travelers' needs and activities. Lemmens (2006) proposed an ontology to support the chaining of operations in geographical information architectures. Bucher and Jolivet (2008) demonstrated the difficulty to document pre and post-conditions of an elementary task (Bucher & Jolivet 2008). Beyond defining a vocabulary to express pre and post conditions, a major bottleneck is the acquisition of their value because it requires setting up benchmarks simulating all possible specific cases of geometrical configurations.

Discussion and conclusion

Quality management traditionally requires the documentation of specifications, the control of quality criteria value, and the description of lineage metadata.

An important challenge raised by VGI CSGI quality management is ambiguities, inconsistencies and heterogeneities due to different abstractions of the geographical space involved in production. These are not limited to features classifications; they should also include important relationships between

elements used in consistency management, affordances of features in contributor activities and rules to encode the perceived reality in data. Another challenge is to manage the quality of data products, hence to somehow commit to some thresholds for the quality criteria.

In section 2 we advocated that it is very relevant to tackle these issues from the perspective of knowledge engineering. The derivation of usable information from raw, heterogeneous and distributed acquisitions would greatly benefit from enhanced model of the context in which a contribution is produced. The modelling of information derivation from raw acquisition can be seen as a flexible process where the integration is done when it is needed and where the sources are preserved as much as possible in order not to lose any meaningful information. The notion of context comprehends many elements which have already been studied in various domains like VGI quality assessment, ubiquitous mapping and ecology. Task models can also contribute to this knowledge engineering project in several ways: to clarify how users perceive the space they will describe, to get external quality criteria, and to improve the coordination of citizen and their interactions towards the production of a common content.

There is still work to be done to integrate the different findings in context modelling and in tasks modelling. An interesting perspective is to improve the description of user intention when they contribute. Rehrl et al. (2013) paves the way for a relevant approach of the problem. Their low-level tasks categorization, such as create/update a geographic feature or a relation, could be extended to conceptualize higher level intentions, such as for instance to reflect a change of navigation restriction that occurred in the reality, to propose a more detailed description of the cross-road geometry and topology, to update an attribute value to reflect a change in the specifications, to fix an inconsistent misalignment of buildings in the data. Other typical VGI tasks need modelling such as selecting, evaluating, integrating existing data, assigning sensor task to contributors, evaluating user capacities with respect to quality criteria. The examination of data quality issues and the literature shows, in our mind, an opportunity to define an ontology of 'human sensing' tasks that would describes capacities to produce pieces of data by a given human agent or several human agents together, with explicit objectives assigned and in a given observation context.

References

Abadie, N. 2009. Schema Matching Based on Attribute Values and Background Ontology. In: *Proceedings of the 12th AGILE International Conference on Geographic Information Science (AGILE'09)*. Hanovre (Germany), June. Available at: http://www.agile-online.org/Conference_Paper/CDs/agile_2009/AGILE_CD/pdfs/138.pdf.

Atemezing, G., Abadie, N., Troncy, R., & Bucher, B. 2014. Publishing Reference Geodata on the Web: Opportunities and Challenges for IGN France.

In: *Proceedings of Terra Cognita, the 6th International Workshop on the Foundations, Technologies and Applications of the Geospatial Web in Conjunction with the 13th International Semantic Web Conference*, October, Trentino (Italy). Available at: http://event.cwi.nl/terracognita2014/terra2014_1.pdf.

Bishr, M., & Kuhn, W. 2007. Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In: Fabrikant, S., & Wachowicz, M. (Eds.) *The European Information Society – Leading the Way with Geo-information, Proceedings of the 10th AGILE International Conference in Geographic Information Science*. Aalborg (Denmark), Springer Verlag, LNGC, pp. 365–387.

Bizer, C., Heath, T., & Berners Lee, T. (2009). Linked data – the story so far. *International Journal on Semantic Web and Information Systems*, *5*: 1–22.

Brando, C., & Bucher, B. 2010. Quality in User Generated Spatial Content: A matter of specifications. In: Painho, Yasmina Santos, & Hardy (Eds.) *Geospatial Thinking, Proceedings of the 13th International Conference on Geographic Information Science (AGILE'10)*, Guimarães (Portugal). Available at: http://www.agile-online.org/Conference_Paper/CDs/agile_2010/ShortPapers_PDF/105_DOC.pdf.

Brando, C., Bucher, B., & Abadie, N. 2011. Specifications for User Generated Spatial Content. In: Geertman, S., Reinhardt, W., & Toppen, F. (Eds.) *Advancing Geoinformation Science for a Changing World, Springer-Verlag Lecture Notes in Geoinformation and Cartography*. Utrecht, Netherlands, CD.

Bucher, B., & Jolivet, L. 2008. Acquiring service oriented descriptions of {GI} processing software from experts. In: *Proceedings of 11th AGILE International Conference on Geographic Information Science*. Girona (Spain). Available at: http://plone.itc.nl/agile_old/Conference/2008-Girona/PDF/94_DOC.pdf.

Das, M., Thirumuruganathan, S., Amer-Yahia, S., Das, G., & Yu, C. 2014. An expressive framework and efficient algorithms for the analysis of collaborative tagging. *VLDB Journal, 23*(2): 201–226.

Devillers, R., & Jeansoulin, R. (2006). Spatial Data Quality: Concepts. In: Devillers & Jeansoulin (Eds.) *Fundamentals of Spatial Data Quality*. ISTE, London (UK), p. 312.

Edsall, R. 2007. Globalization and cartographic design: implications of the growing diversity of map users. In: *Proceedings of the 23rd International Cartographic Conference*. Moscow. Available at: http://www.academia.edu/2324974/Globalization_and_cartographic_design_Implications_of_the_growing_diversity_of_map_users.

Gibson, J. J. 1979. The Ecological Approach to Visual Perception. Boston: Lawrence Erlbaum Associates.

Goodchild, M., & Li, L. 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*: pp. 110–120.

Goodwin, J., Dolbear, C., & Hart, G. 2008. Geographical Linked Data: the Administrative Geography of Great Britain on the Semantic Web. *Transactions in GIS*, *12*(1): 19–30.

Haklay, M. 2010. How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England. *Environment and Planning*, *37*(4): 682–703.

Haklay, M., Basiouka, S., Antoniou, V., & Ather, A. 2010. How many volunteers does it take to map an area well? *The validity of Linus's Law to volunteered geographic information, 4*(4): 315–322.

ISO TC211 2014 19115-1:2014. Geographic information — Metadata — Part 1: Fundamentals.

Jakobsson, A. 2002. User requirements for mobile topographic maps, GiMoDig deliverable IST-2000-30090 D2.1.1. Available at: http://lib.tkk.fi/Diss/2006/isbn9512282062/article5.pdf.

Lemmens, R. 2006. Semantic interoperability in distributed geo-service, PhD thesis, ITC, Enschede, Netherlands. Available at: http://repository.tudelft.nl/view/ir/uuid:31b0eae6-c411-4bbd-a631-153498889671/.

Maué, P., & Schade, S. 2008. Quality of Geographic Information Patchworks. In: *Proceedings of 11th AGILE International Conference on Geographic Information Science*. Girona (Spain). Available at: http://ww.w.agile-online.org/Conference_Paper/CDs/agile_2008/PDF/111_DOC.pdf.

Park, H., Yoon, A., & Kwon, H-C. 2012. Task model and task ontology for intelligent tourist information service. *International Journal of U-and E-Service, Science and Technology*, *5*(2): 43–58.

Rehrl, K., Gröechenig, S., Hochmair, H., Leitinger, S., Steinmann, R., & Wagner, A. 2013. A Conceptual Model for Analyzing Contribution Patterns in the Context of VGI. In: Krisp (Ed.) *Progress in Location-Based Services, Lecture Notes in Geoinformation and Cartography*. Springer-Verlag, Berlin. Available at: http://flrec.ifas.ufl.edu/hochmair/pubs/Rehrl_LBS2012_analyzingContributionPatterns.pdf.

Servigne, S., Ubeda, T., Puricelli, A., & Laurini, R. 2000. A Methodology for Spatial Consistency Improvement of GeographicDatabases. *Geoinformatica*, The Netherlands, *4*(1): 7–34.

Sheeren, D., Mustière, S., & Zucker J-D. 2004. Consistency Assessment Between Multiple Representations of Geographical Databases: a Specification-Based Approach, In: Proceedings of the 11th International Symposium on Spatial Data Handling (SDH'04), Leicester (UK)

Sun, C., Jia, X., Zhang, Y., Yang, Y., & Chen, D. 1998. Achieving convergence, causality preservation, and intention preservation in real-time cooperative editing systems. *ACM Transaction CHI*, *5*(1): 63–108.

Sun, Z., Yue, P., Lu, X, Zhai, X., & Hu, L. 2012. A Task Ontology Driven Approach for Live Geoprocessing in a Service-Oriented Environment: A Task Ontology Driven Approach for Live Geoprocessing. *Transactions in GIS*, *16*(6): 867–884.

Vilches-Blázquez, L., Villazón-Terrazas, B., Saquicela, V., de Leon, A., Corcho, O., & Gómez-Pérez, A. 2010. GeoLinked Data and INSPIRE through an Application Case. In: *Proceedings of the 18th ACM SIGSPATIAL*

International Conference on Advances in Geographic Information Systems. ACM SIGSPATIAL GIS 2010, San Jose, California, USA.

Von Hunolstein, S., & Zipf, A. 2003. Towards task oriented map-based mobile guides. In: *Proceedings of the International Workshop "HCI in Mobile Guides"*. 5th International Symposium on Human Computer Interaction with Mobile Devices and Services, Udine (Italy), pp. 8–11

Wilkinson, D., & Huberman, B. 2007. Assessing the value of cooperation in Wikipedia. *First Monday, 12*(4). Available at: http://firstmonday.org/article/view/1763/1643