The impact of crowdsourcing on spatial data quality indicators

M. van Exel[1], E. Dias[1,2], S. Fruijtier[1]

[1]Geodan S&R, President Kennedylaan 1, 1079MB Amsterdam, the Netherlands
Email: [martijn, steven]@geodan.nl

[2]Vrije Universiteit – FEWEB/RE (room 4A-42), de Boelelaan 1105, 1081HV Amsterdam, the Netherlands
Email: edias@feweb.vu.nl

1. Introduction

Crowdsourced geospatial information has soared the last couple of years. For example the statistics of OpenStreetMap (OSM) show an accelerating growth (OSM 2010). At the same time, devices equipped with GPS became mainstream (e.g. iPhone), diminishing the threshold to participate in crowdsourced geospatial information projects.

Together with the growth in volume, the usage of crowdsourced geospatial information grew extensively as well. For example OSM maps are used in different commercial projects as background maps. This increased usage makes it important to identify quality indicators for crowdsourced geospatial information (Haklay and Weber 2008, Goodchild 2007, Flanagin and Metzger 2008) in order to:

1. compare and integrate crowdsourced data with institutional data (from e.g. National Mapping and Cadastral Agencies) and commercial data (e.g. TeleAtlas and NAVTEQ);
2. determine fitness for the intended purpose;
3. predict the quality developments for certain areas.

In this abstract, we introduce the concept 'Crowd Quality' (CQ) to describe and quantify the quality of crowdsourced geospatial information.

2. Quality elements

The term 'quality' has a meaning if we have a common understanding of its definition. According to ISO19113(2002), quality is the "totality of characteristics of a product that bear on its ability to satisfy stated and implied needs". For spatial quality elements, different definitions exist. Van Oort (2006) compiled five important sources of them and identified eleven elements of spatial data quality: Lineage, Positional accuracy, Attribute accuracy, Logical consistency, Completeness, Semantic accuracy, Usage/purpose/constraints, Temporal quality, Variation in quality, Meta-quality, and Resolution. These elements are used to describe the quality of geo-data collected and produced with a commissioned effort, which entails a specified and uniform method to gather and process the data. Therefore, the quality of such data is usually assumed homogenous and consistent.

However, volunteered geographic collections are characterised by heterogeneous and diverse quality, due to the fact that it is collected using different methods (e.g. GPS tracks, image tracing) and by different individuals with different motivations and preferences. Moreover, contributors and contributions are not distributed evenly over space. To address this problem, we introduce the concept of Crowd Quality (CQ) to describe and quantify the quality of crowdsourced geospatial information.

3. Crowd Quality

Crowd Quality (CQ) attempts to quantify the 'collective intelligence of the crowd generating data' in a spatio-temporal context. CQ is based on a two-dimensional approach: User-related quality aspects and Feature-related quality aspects. These aspects can comprise the existing quality elements, extended with quality elements specific for crowdsourced data.

The User dimension manifests the quality of information contributions from an individual contributor's perspective. This is a quintessential characteristic of crowdsourced data. Unlike institutional geospatial information collection, the individuals contributing to the product have no a priori established status or qualification. Nor is the individual's scope for contributing determined by organisational constructs.

The Feature dimension approaches Crowd Quality from the perspective of the spatial feature. Rather than looking solely at the established dimensions of spatial quality (e.g. accuracy and completeness), any quality indicator for user-generated spatial features should encompass the collective experience, knowledge and effort of the individuals who contributed to that feature.

We aim to establish operational indicators for both User Quality and Feature Quality

3.1 User Quality

We suggest three components to determine User Quality: Local knowledge, Experience and Recognition.

Individuals can have any number of personal motivations for contributing to crowdsourced geo-information projects. Ongoing research (Nedović-Budić 2010) suggests that the strongest motives have either an idealistic or a free-time nature, or are driven by personal place-based needs and local knowledge. This local knowledge enables the contributor to identify missing or incorrect information relatively easy. The following hypotheses, when successfully tested, help establish an operationalisation of the User's local knowledge: Familiarity to an area can be correlated to the spatio-temporal pattern of his contribution and the quality of a contribution is higher for areas he is most familiar with.

A second component of the User dimension is his experience in contributing to the project. We hypothesise that the quality of a user's contribution is correlated with his overall experience in contributing to the project.

Experience may be quantified using the amount of time the user has been registered with the project, the amount of GPS traces he registered with the project, the number of features added or edited, but also his activity in virtual as well as real-life forums within the context of the project. In participating in these forums, the User gains not only experience by learning from and generally interacting with his peers. He also gains recognition through this interaction both in real life and in virtual forums.

Recognition comprises the third and last proposed component of the User dimension of Crowd Quality. In online social networks and online contexts that allow for user contributions, often tokens are established. Examples include the feedback from Ebay.com and the reputation of StackOverflow.com. These are awarded by other users as recognition for specific contributions, or by the system when a certain quantitative or qualitative threshold is met. This type of User recognition is largely unknown in crowdsourced geospatial data. This puts a strain on our ability to assess Crowd Quality, as the peer reviewing of contributions lies at the core of internal

quality assurance of crowdsourced information repositories. Therefore, we need to devise implicit peer reviewing indicators.

Implicit peer reviewing indicators can be derived from subsequent contributions by individual users. If we consider a larger spatial context and find one feature that has received few improvements, whereas most features within the spatial extent under consideration have undergone many revisions, that lack of subsequent edits may be considered positive recognition. This recognition cannot be established a priori nor in an isolated case. Further study into the geo-social dynamics of crowdsourced geospatial information projects will reveal additional implicit peer reviewing patterns.

3.2 Feature Quality

Traditional quality assurance and assessment of geospatial information, departs from the spatial features that comprise the information entity. In crowdsourced datasets quality elements can be different for similar features, while in traditional datasets the quality elements will be uniform. For example a commissioned dataset about restaurants will include a defined and specified set of attributes, while a user contributing to a crowdsourced dataset can define his personal attributes (e.g. opening times, type of food, cosiness).

Feature Quality of crowdsourced geospatial data can be assessed by the same quality elements usually used for features from traditional, conformance-based spatial databases. Of particular interest are lineage, positional accuracy and semantic accuracy because these elements are not considered consistent for crowdedsourced data.

Crowdsourced data is typically generated using an array of different methodologies and tools. Some data is imported from other sources, if available under a compatible license. These imported features have a very clear lineage with regard to positional accuracy and precision. Another common tool used is derivation from GPS points collected in the field. Here the positional accuracy is harder to establish, mostly because the accuracy and precision metadata is usually stripped from the GPS data, and Users may not attribute their contributions to a GPS source. In a crowdsourced context, any single feature may have been affected by different methods. Moreover, the spatial accuracy and precision of neighbouring features may have affected the positioning of the feature under consideration. Further study is required to reveal the dynamics that determine positional accuracy and precision of any specific feature.

Another complex quality element is semantic accuracy, pertaining to the completeness and internal consistency of the attribute metadata. A predefined schema for attribute metadata is not common in crowdsourced geospatial data projects. Much trust is put in the selforganising capacity of crowdsourcing ecosystems. This lack of a priori organisation allows for the creative input of individuals and small groups with specific interests to benefit the project by generating a breadth of information that would not otherwise be feasible, but at the same time poses a threat to internal consistency.

Lastly, the diversity of feature types and feature attribution is hypothesised to have a positive correlation with the quality of the information.

3.3 Interdependency

User Quality and Feature Quality can often not be considered as independent, disparate entities. The user dimension manifests itself in the contributions that a user makes to the database, and must therefore be measured through the features. Feature Quality on

the other hand, is ultimately intertwined with the User Quality of the individuals that contributed to the feature under consideration.

The User and Feature Quality dimensions are intertwined forming a new indicator describing the spatio-temporal dynamism and persistence. This indicator deals with such characteristics as:

- How many different users contributed to a feature?
- How has a feature developed over time?

4. Further work

Future work aims at the operationalisation of the Crowd Quality concept. We propose a framework approach to determine the quality of crowdsourced geospatial information, using crowd dynamics. We will introduce different indicators that take into account Spatial Crowd Activity (surrogated by number of edits and editors) , Temporal Crowd Activity (number of edits per time periods) and Relative Crowd Activity (number of edits relative to an enclosing / neighbouring area).

Quantitive research on these indicators can only be carried out using historical information from crowdsourced data that includes temporal information about the features and the contributing users. The Openstreetmap database contains this information and will be used for validation of the indicators and assumptions.

References

Goodchild M.F. 2007. Citizens as Sensors: The World of Volunteered Geography, VGI Specialist Meeting Position Papers, Santa Barbara, CA. http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Goodchild_VGI2007.pdf

Flanagin, A.J. and Metzger, M.J., 2008, The credibility of volunteered geographic information. GeoJournal 72: 137–148

Haklay, M and Weber, P, 2008, OpenStreetMap: User-Generated Street Map. IEEE Pervasive Computing, 7(4): 12-18.

ISO 19113:2002, Geographic information - Quality principles.

Nedović-Budić, Z and Budhathoki, N.R., 2010. Motives for VGI Participants. Ongoing research presented at the workshop 'VGI for SDI', Wageningen University, NL, April 16th, 2010.

OSM, 2010, OpenStreetMap: Wiki, http://wiki.openstreetmap.org/wiki/Statistics

Van Oort, PAJ, 2006, Spatial data quality: from description to application, PhD Thesis, Wageningen University, NL, 132 p.