Barbara Poore
US Geological Survey

Toward Quality Assessment for Hybrid Databases of Official and Crowdsourced Spatial Data

"What should spatial accuracy assessment mean in a world in which everyone is a potential user of geospatial data?" --Goodchild (2008)

Spatial data quality is increasingly difficult to evaluate for digital data on the Internet. When Goodchild (2008) wrote his seminal paper on this issue, the problem was that the Internet allowed data producers to freely mix data from different sources making it difficult to determine the origins or quality of the data. Since then, the situation has been exacerbated by the rise of online databases such as OpenStreetMap (OSM) that are entirely produced through crowdsourcing. Now the question becomes "what should spatial data accuracy mean in a world in which everyone is a potential user and producer of geospatial data?" Studies of the quality of volunteered geographic information (VGI) have principally focused on positional accuracy of OSM road layers in European countries where OSM is well developed (for example Haklay 2010). Other traditional aspects of spatial data quality such as lineage, completeness, attribute accuracy, and logical consistency are just beginning to be examined--for example Kouloletsos et al. (in press) have suggested a technique for evaluating the completeness of the road network in Great Britain and Mooney and Corcoran (in press) have looked at tag editing behavior as a way to assess attribute accuracy. Van Exel et al. (2010) proposed trust among contributors as a proxy for data quality.

National mapping agencies and spatial data infrastructures rest their reputations on the production of authoritative data. That is, data that is of documented quality, assembled from the "best available" sources (Federal Geographic Data Committee 2011). It is commonly asserted by professional data producers that crowdsourced data or VGI is of poor or unknown quality (Goodchild 2007). Even so, many of these organizations are looking into the use of crowdsourcing to enhance national mapping databases and spatial data infrastructures (Elwood 2008, Coleman 2010). Such databases can be called hybrids because they incorporate data from so many different sources including VGI. The denial of data quality from non-professional sources and the simultaneous desire to use this data to bolster official databases would seem to be a contradiction, and one of the questions I would like to explore further at the GIScience workshop is how this hybrid nature changes both the volunteer contributions and the sponsoring organiztion. Drawing on an ongoing project investigating the use of crowdsourcing at the US Geological Survey (USGS), I plan to discuss the results of an extensive quality analysis of the data produced by this project and question whether traditional methods of data quality analysis remain appropriate and if not, what should replace them.

In 2011, following a specialist meeting on volunteered geographic information (http://cegis.usgs.gov/vgi/), the USGS launched a project to investigate whether it was feasible to use volunteers to collect data suitable for incorporation into *The National Map*, the digital equivalent of the USGS topographic maps. The first phase of the project investigated whether

the open source software used by OSM could be repurposed to support collaborative editing of transportation data between USGS and the Kansas Data Access and Support Center (Wolf et al. 2011). The project proved that collaborative editing was possible, but that transportation data might be too complex for unsophisticated users. The second phase of the project which concluded in the fall of 2011 shifted focus to the collection of structures data (points of interest), using a USGS-branded instance of OSM's Potlatch 2 software. Student volunteers were recruited from GIS classes at Denver area universities. Data on structures were collected over a limited time period in four quadrangles covering metropolitan Denver. Students edited existing points from the USGS's Geographic Names Information System (GNIS), correcting positions and adding attributes. A Wikipedia-style editing routine was established by which students quality checked the data produced during editing. In addition, because the OSM software generated a planet file every evening that maintained a record of each edit to every structures point, USGS was subsequently able to perform a detailed in-house quality check.

The quality information on over 1,000 points added by volunteers demonstrates that student volunteers were capable of collecting data that met horizontal positional accuracy standards suitable for incorporation into *The National Map.* However we found it difficult to apply traditional measures of positional accuracy based, as they were, on paper maps. We were hindered by unclear documentation of the source data the students had used for editing. In addition, as Corcoran and Moody (in press) have shown, attribute accuracy are difficult for volunteers to achieve due to issues of ontology. Even though the USGS provided strict editorial guidelines for how names and addresses were to be formatted, volunteers had trouble with spelling and the ordering of terms. An analysis of completeness was not attempted due to the absence of a database of higher accuracy with which to compare the volunteered data, however, the USGS quality check discovered that the students missed many structures. A report on this phase of the project will likely be available at the time of the workshop. I will also be able to discuss insights that have caused changes for the next phase of the project. This phase is open ended. It encompasses the whole state of Colorado, allows anyone to edit structures data, and will employ an "adopt-a-quad" quality check routine by volunteers.

It may be argued that this project is not truly VGI in the sense that unlike OpenStreetMap, it is not bottom up. A great deal of attention was paid to channeling the student effort using system documentation, class presentations, and customized drop-down menus in the interface. It is a hybrid method that is also employed by many citizen science efforts, for example Audubon's Christmas Bird Count. I would like to stimulate some discussion at the workshop of how software routines, user documentation, user interfaces, and work practices affect volunteer conformance and how volunteered data might in turn conform spatial data infrastructures (Sieber 2000).

References

Coleman D.J. (2010). The potential and early limitations of volunteered geographic information, Geomatica, 64(2): 209-219

Ellwood, Sarah (2008). Grassroot groups as stakeholders in spatial data infrastructures: challenges and opportunities. *International Journal of Geographical Information Science*, 22 (1): 71-90.

Federal Geographic Data Committe (2011). Annual Report. Reston VA. Available at http://www.fgdc.gov/library/whitepapers-reports/annual%20reports/2011. Accessed 10 May, 2012.

Goodchild, M. F. (2007).. Editorial : Citizens as Voluntary Sensors : Spatial Data Infrastructure in the World of Web 2 . 0. *International Journal of Spatial Data Infrastructures*, *2*, 24-32.

Goodchild, M. F. (2008). Spatial Accuracy 2.0. In J. Zhang & M. F. Goodchild (Eds.), *Eighth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 1-7). Shanghai: World Academic Union.

Haklay, M. M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682-703.

Kouloletsos, T., Haklay, M., Ellul, C (in press) Assessing data completenes of VGI through an automated matching procedure for linear data. *Transactions in GIS*

Sieber, Renee (2000). Conforming (to) the opposition: the social construction of geographical information systems in social movements. *International Journal of Geographical Information Science* 14 (8): 775-793.

van Exel, M., Dias, E. Fruijter, S. (2010). The impact of crowdsourcing on spatial data quality indicators GIScience 2010: Proceedings of the 6th International Conference on Geographical Information Science. University of Zurich.

Wolf, E.., Matthews, G. McNinch, K, Poore, B.(2011). OpenStreetMap Collaborative Prototype. Open-File Report 1136. Available at http://pubs.usgs.gov/of/2011/1136/pdf/OF11-1136.pdf. Accessed 10 May, 2012.