# Classifying Volunteered Geographic Information: A Model for Retrieving Spatial Data from the Internet

**Scott Bell**[1,2]

1 Spatial Structures in the Social Sciences, Brown University, Scott_bell@brown.edu
2 Geography, University of Saskatchewan, Saskatoon, SK

## Abstract

While Internet Mapping services abound on the web there is another potentially more important movement taking place. Community mappers, including individuals, groups, governments, and other organizations are collecting and delivering the raw materials for mapping at an alarming rate. Web 2.0 is central to this and other emerging social, political, economic, and geographic phenomena. While there is still an important place for traditional web mapping services the participatory nature of blogs, wikis, and bottom-up web development is providing a new way of thinking about spatial data.

## Background and Relevance

The availability of spatial data at publicly accessible locations via the Internet is having an important impact on research, social exchange, and social justice (Goodchild, 2007). While mapping services are well known, what is potentially more important are mappable bodies of spatial data that are currently unmapped. Mapping Volunteered Geographic Information (VGI) requires two parties of users, the *internal* users who provide the data and *external* users who map the data and make it accessible in mapped form on the Internet. Bell and Logan (2008, forthcoming) have introduced the concept of *internal* and *external* users in the context of research-focused Internet mapping systems. With respect to Internet mapping sites Inside and Outside user categories are suggested; *Inside* users having conceived of and created the research focused internet mapping systems and *Outside* user being those people who access the map services and spatial data to answer new research questions. In the broader context of VGI, data providers operate with fewer constraints in terms of what they must offer to the eventual mapper of their data. The data provider might offer a dataset that brings together data from disparate sources in a form that can be downloaded and immediately mapped or geocoded. At the other extreme they might offer a single piece of qualitative spatial information (a place name for instance) that the user must validate, integrate, and hopefully geocode in order to map and use. The following classification scheme is one attempt to help clarify the variety of spatial information that might be classified as VGI and discusses some of the implications of each category.

## Methods and Data

Experience gained from the development and construction of the US School Matters mapping system led to the following classification scheme. The following simple classification scheme includes three categories based on the type of geographic information being provided its level of integration with similar information. While the

data sources for our project were of a specific type (presented both Type 1 and Type 2 information) we were challenged to consider the continuum on which such CGI exists. Our project goal was to map all US public schools (all schools governed under No Child Left Behind policies) and present that information with information regarding populated places, demographics, and some physical and infrastructure. While the latter data were all available form reliable sources (Census, American Factfinder, USGS, Geography Network, etc.) the school data is just now becoming available and is characteristic of much of the VGI available on the Internet. Our experience has been that a great deal of VGI providers aren't yet familiar with how the spatial component of their data will be used or the needs of the end-users (researchers, cartographers, etc.) of this data. Furthermore, when VGI is provided in a global coordinate system (latitude/longitude, UTM, state plane, etc.) the necessary metadata does not always accompany the coordinates. This was our experience when using SchoolMatters to access both geographic (school addresses and later, latitude and longitude coordinates) and non-geographic (enrolment, test scores, school demographics, etc.) information. SchoolMatters integrates data from each state to provide a single location for school performance (and location) for the whole country. This service comes at some cost, as the origins of the data are the individual states who are responsible for testing, setting standards, collecting data, and making it publicly available.

## Results

### A VGI Classification

**Specific or Narrow Data/Low Value-added:** Internal User Provides VGI; External User Validates, Integrates, and Maps
FORMAT: Information comes in a form that requires intervention to map, this might include street addresses, city names, place names, etc.
COLLECTION: The collection is in no way comprehensive and the developer (*internal* user) has made no effort to integrate their data with similar data. While such information might be in an Internet location associated with a community such integration does not facilitate the retrieval of the geographic information along with large amounts of similar data. Examples might include wikipedia entries, photo archives (flickr, facebook, etc.), blogs, etc. and have the potential to provide rich and novel information. The spatial component (place name, coordinate, address, etc.) can be in almost any form and will likely require intervention to map. Interventions include geocoding, metadata search, coordinate system calculation/conversion and are the sole responsibility of the *external* user of the data.

**Topical Data/Low Value-added:** Internal User Provides and Integrates VGI; External User Geocodes and Maps (and may Integrate with other data)
FORMAT: Information comes in a form that requires intervention to map, this might include street addresses, city names, place names, etc.
COLLECTION: The data source has been integrated in some fashion with similar data. The integration might be the work on a third party or by the developer/owner. Such integration includes search companies such as City Search that collects useful information about businesses and services in urban areas. The primary value added for

the potential end-user of such data sources is that data can be retrieved efficiently based on the logical and consistent structure of the data (using a computer/database program, either commercial or proprietary). We have had success using "webscraping" programs that are commercially available as well as producing our own programs and scripts for pulling together data with some common and formal structure. Both commercial and self-produced approaches have the capacity to pull spatial and non-spatial information from websites and can perform searches through hierarchically structured websites, as long as such sites' structure is formalized.

Both Type 1 and Type 2 data sources involve more work by the *external* end-user of the data but should increase data reliability and reduce data uncertainty as the application of discrete geographic coordinates is the job of the *external* user.

**Topical or General/High Value-added:** Internal User Provides Georeferenced VGI; External User Validates (and Integrates)
FORMAT: Information comes in a form that requires no intervention to map, data is tagged with a latitude and longitude with a reasonably easy to assess coordinate system making reliable mapping straightforward
COLLECTION: Data provider is likely actively interested in having the spatial component of their data used by external users. Data sources of this type have the greatest potential for providing extensive datasets that have internally consistent characteristics. *Internal* data providers are likely high up on a hierarchy of similar organizations (state and federal government, national non-profit or international NGOs, foundations, etc.); this is a direct result of their own access to such data. Our experience is that without comprehensive metadata geographically referenced data has a greater potential for spatial uncertainty than data that requires the *external* user to assign such coordinates. Take for example a website such as SchoolMatters that provides the geographic coordinates and test scores for individual schools in the USA. Since this site provides data that is generally the responsibility of individual states it is difficult to assume that all the data was collected or the geographic coordinates recorded in a common frame of reference (datum, reference system, etc.). Without explicit metadata concerning the origins of the spatial data the user is presented with questions that are among the first things students of geography, cartography, and GIS are warned about with respect to the importance of map projections, geodesy, and frames of reference.

## Conclusions

This classification scheme is useful for determining the state of a potential VGI source and both the work that will be involved in mapping and analyzing data from such sources. Furthermore, it the *external* user can more quickly evaluate the type of work and how much work will be involved in using data from a particular source.

## References

Bell, S., & Logan, J. (2008, forthcoming). Distributed research and scientific creativity: Accessible data for the social sciences. In M. Peterson (Ed.), *Internet and Mapping II*.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography [Electronic Version], 15. Retrieved September 23, 2007 from http://www.ncgia.ucsb.edu/projects/vgi/docs/Goodchild_VGI2007.pdf.