



Spatial Accuracy 2.0

Michael F. Goodchild

University of California, Santa Barbara

Abstract. Research on spatial accuracy assessment occurs within a broader context that provides its motivation. That broader context is dynamic, and has been changing at an accelerating rate. The concept of the Geospatial Web imagines a world of distributed, interoperable, georeferenced information in which it is possible to know where everything of importance is located in real time. It assumes an ability to conflate that is far beyond today's capabilities. Web 2.0 describes a substantial involvement of the user in creating the content of the Web, and has particular relevance to geospatial information. Metadata 2.0 shifts the onus for metadata production to the user, and addresses some of the growing issues surrounding existing standards. There is a growing need to address the accuracy assessment of the vast quantities of geospatial data being contributed by individual Web users.

Keywords: Geospatial Web, user-generated content, volunteered geographic information, Web 2.0, metadata

1. Introduction

In May 1994 a steering committee of Russell Congalton, Timothy Gregoire, Stephen Prisley, James Smith, and Robert Weih Jr organized an International Symposium on the Spatial Accuracy of Natural Resource Data Bases in Williamsburg, Virginia, the first in what is to date a fourteen-year series of biennial conferences. Of the 64 listed pre-registrants most were American, but an impressive total of 8 other countries were represented. The topics addressed were very much as they are at the eighth conference: experimental determinations of the accuracy of various data sets and analyses, theoretical models of uncertainty in spatial data, and papers on the visualization of uncertainty. Also represented were several papers on the broader context of the accuracy issue: on why forest resource information managers should care about accuracy, on the consequences of error, and on prospects for improved accuracy in the future.

That concern for the broader context has waxed and waned over the eight conferences in the series. At the second conference in 1996 in Fort Collins, Colorado, my own keynote was concerned with communicating the results of accuracy assessment. In it I argued that the spatial statistical framework of accuracy assessment was inherently difficult and inaccessible to a large proportion of the user community, and that special attention needed to be given to communication if the research enterprise as a whole was to be judged a success.

Over the 14 years since the first conference the geospatial landscape has changed almost beyond recognition. Web-based services such as Google Earth and Google Maps have popularized access to geospatial data, and today citizens routinely rely on such services, and the specialized services that leverage them, for driving directions, buying houses, and planning foreign travel. New data acquisition technologies such as LiDAR, high-resolution Earth imaging, and radar interferometry now supply vast resources of geospatial data, through a variety of data warehouses, digital libraries, and geoportals. Today's Web now makes everyone a potential creator of geographic data through services such as Google MyMaps and Wikimapia.

In this paper I explore some of these changes and their implications for spatial accuracy research. The title refers to Web 2.0, a phrase coined by Tim O'Reilly which has become popular in the past two or three years and which contrasts the early days of the Web, in which all information flowed from server to user,

with the recent rapid multiplication of services that allow the user to become a significant source of information. It encompasses blogs, wikis, social sites such as Facebook, and increasingly sites that emphasize the assembling of geographic information. Web 2.0 has a post-modern feel with its emphasis on engaging the individual, on personalization of content, and on the subjective side by side with the objective.

The paper is structured in three sections. In the first I discuss the Geospatial Web [1], a vision for a distributed world of interoperable data acquisition, storage, processing, and archiving. The Geospatial Web is rapidly becoming a reality, and raises some significant issues for accuracy assessment. The strongly personal, user orientation of Web 2.0 suggests a more user-centric approach to accuracy which is discussed in the second section. Finally the third section looks at user-generated content and its accuracy assessment, and makes some preliminary conclusions about the quality of this rapidly growing data source.

2. The Geospatial Web

Definitions of the Geospatial Web or GeoWeb typically emphasize the power of geographic location as a key for integrating knowledge, and for providing context. Both have long been strongly motivating arguments for geographic information systems (GIS), but have gained new momentum because of the growth of the Web and because of ready access to mechanisms for adding location to data. For example, it has become trivially easy to take a street address and convert it to latitude and longitude. This allows any information associated with the street address to be linked to any other information associated with that latitude and longitude, or with its surroundings or spatial context. The Global Positioning System (GPS) has also made it trivially easy to measure latitude and longitude directly, and to create vast databases from the tracks of vehicles, animals, and pedestrians.

Indeed it is now possible to envision a world in which it is possible to know where *everything* is. The widespread use of RFID (radio-frequency identification) tags for commodities in stores, farm animals, passports, and mobile phones, and the development of accurate technology for determining mobile phone positions, are already demonstrating the potential and some of its more negative social implications. In future some areas of human activity may come to resemble airspace, in the sense that the location of every aircraft flying through an airspace is always known. All of this new information is geospatial, and all of it therefore a legitimate topic for accuracy assessment. Thus the world of spatial accuracy assessment has moved in fourteen years far beyond the constraints of the paper maps that initially provided much of its motivation. At the same time it would be hard to find any explicit concern for spatial accuracy in any of the more popular sites. In Google Earth, for example, even the most obvious elements of data quality, the date and time at which the base imagery was acquired and its spatial resolution, are not available to the user.

The term *mashup* has become the preferred way of referring to the linking of Web services through common references, and georeferences are clearly one of the most powerful and ubiquitous bases for what is essentially a generalization of the concept of a relational join. There are now thousands of published mashups, many of them created by people with very little knowledge of the principles of cartography, geographic information science, or spatial accuracy assessment, using simple and readily available software. Members of this community often refer to themselves as *neogeographers*, signalling that they discovered the importance of geographic location without any formal training in the geographic sciences.

Accuracy is an essential component of any integration, because the measurement of location cannot be perfect, and hence two independently determined estimates of the location of any feature on the Earth's surface will not agree. Thus a central issue in mashup is whether separate references to two locations x_1 and x_2 are actually referring to the same place. Like the early world of GIS, developments in the Geospatial Web have leapt far ahead of any concern for confidence limits or metadata, so information on the uncertainty associated with locations is almost certainly unavailable, and unlikely to be inferrable from the precision with which the locations are specified.

However it is important to note that in most cases mashups do not require the establishment of a logical linkage between features. In a typical Google Maps mashup, locations of houses for sale are extracted from one Web service in the form of street addresses, geocoded to latitude and longitude, and then visually superimposed on a map display. Any associations with other features on the map are the responsibility of the user's eye and brain, and small inaccuracies in position can be ignored.

Figure 1 shows an example using Google Earth. Here a high-resolution image of part of the university campus has been carefully registered by taking repeated GPS observations of a number of control points to provide a positional accuracy in the 1m range. This has been mashed with the Google Earth base imagery by visual superimposition. Note the misregistration by approximately 15m. Also visually superimposed is Google's street data, supplied by Navteq and known to be positionally accurate to much better than 15m. Because the street data and the high-resolution data are in substantial positional agreement, we conclude that it is the Google Earth imagery that is the source of most of the misregistration. Microsoft's Virtual Earth for the same location showed a misregistration in the opposite direction by approximately 3m.

Visual superimposition of data sets can often reveal shared lineage through inheritance of the same positional errors. Figure 2 shows the 7 available street centerline databases of part of Goleta, California, and it is clear that two of the data sets have been obtained from the same source.

The existence of data sets depicting multiple versions of the same geographic features has led to renewed interest in conflation, particularly when any type of formal analysis of data is intended. Hastings [2] has studied the conflation of gazetteer data, which is defined as a collection of triples linking named features, their geometry, and their feature type. He uses the example of Lake Tahoe, which may appear under different names, with different approximations to its detailed geometry, and classified in some cases as a lake and in some cases as a reservoir. He develops metrics of similarity of all three components (name, geometry, and type), and shows that similarity of geometry should always be given highest priority, followed by similarity of type and then similarity of name. In other work, members of my research group are examining the conflation of multiple versions of street centerlines in a generalization of the familiar problem of matching GPS tracks to existing databases.



Figure 1: Mashup of a high-resolution image with Google Earth. Note the approximately 15m offset of the Google Earth base imagery, and the close agreement between the Google Earth roads layer and the high-resolution image.

Unfortunately the problem of uncertainty in position is likely to grow as interest in the Geospatial Web leads to the exploitation of many different sources of locational data. Using such services as Google Earth to determine location is especially troublesome, since locations determined in this way will inherit any misregistration of the service's base mapping. Replacing the base mapping, particularly replacing the imagery, will lead to further uncertainty analogous to that which results through the occasional replacement of a geodetic datum.

3. User-generated quality assessment

In the world of Web 2.0 the term *user-generated content* refers to the ability of Web users to create content

that is then integrated and made available through Web sites. In the geospatial domain a number of sites have been established for the purpose of inviting and assembling map data, and such sites are proliferating rapidly. A leading example is Wikimapia (Figure 3), modeled on Wikipedia and dedicated to “describing the whole world”. Any Web user is able to focus on any part of the world at any scale using a Google Maps interface, identify a feature by outlining its footprint, and provide descriptive information that may include a name, links to other information sources, text, and imagery. At time of writing Wikimapia provided almost 7 million feature descriptions, ranging in size from entire continents to individual houses. Another is OpenStreetMap, dedicated to providing a free, open digital map of the planet as a patchwork of contributions by individual volunteers. Collectively this activity has been termed *volunteered geographic information* (VGI).

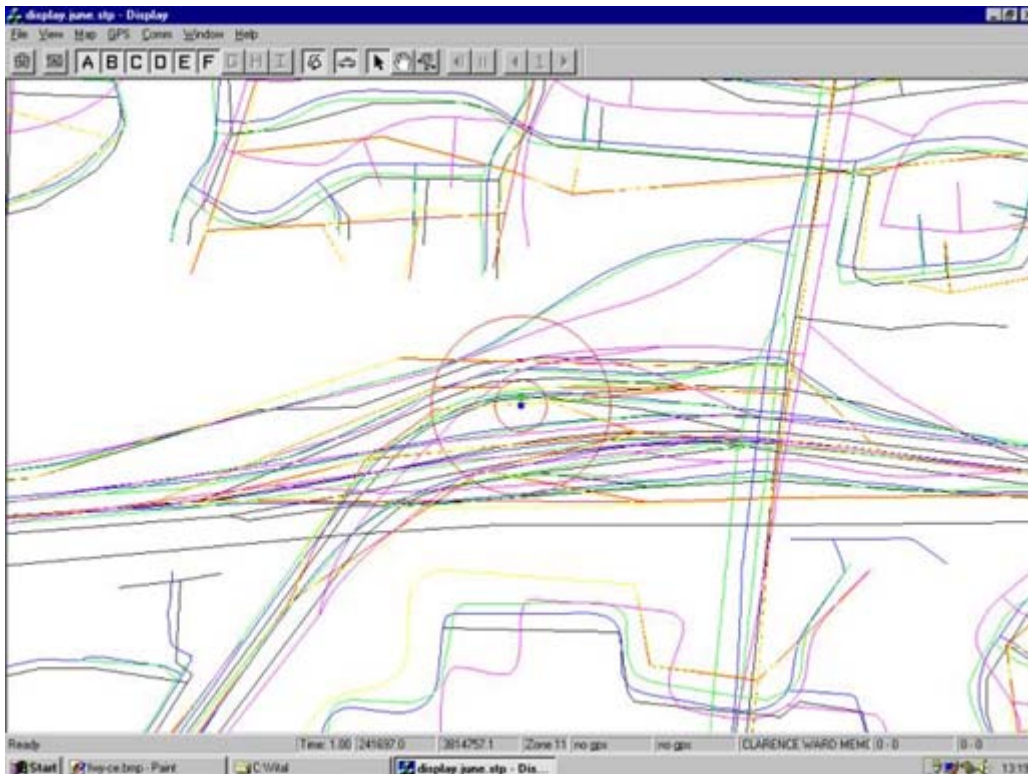


Figure 2: Overlay of the seven available street centerline databases of part of Goleta, California

Questions of accuracy arise in many ways with respect to VGI. Some sites allow editing, again by individual volunteers, based on the principle of *collective intelligence* or *crowdsourcing*, that information provided through a group consensus tends to be more accurate than information provided by a single individual. This principle has a sound theoretical basis in statistics, since the mean of a sample will converge on the mean of the population as the sample size increases. But in many geospatial contexts there is no concept of a true value, so the analogy between a group consensus and a sample mean becomes tenuous at best.

UGC can potentially provide a powerful mechanism for error correction. Google has recently allowed users to edit the locations of landmarks, such as their own houses, in Google Maps. It is well known, however, that strong spatial dependences exist in the positional errors of geospatial data, such that if point x is displaced, it is likely that the displacement of a nearby point at $x+\delta x$ will be similar in magnitude and direction. Spatial dependences make it virtually impossible to insert partial, independent corrections in geospatial data because they create the potential for unacceptable topological errors. For example, moving the house at 909 West Campus Lane 70m to the south to correct an apparent positional error will place it on the wrong side of the neighboring house at 908 West Campus Lane, unless the latter is also repositioned.

Perhaps the most significant area of geospatial data quality for VGI is currency, or the degree to which the database is complete and up-to-date. Consider, for example, the problem faced by a vendor of street

centerline data in attempting to keep its data current, and assume that the standard of currency is a few weeks at most—that changes in the street network should appear in the database within a few weeks of their appearance in the real world. Standard methods of data collection, including feature extraction from imagery or ground-based survey, have high fixed costs and are therefore unreasonably expensive for the capture of scarce, randomly located changes. Instead it makes much more sense to enlist local citizens, each of whom is an expert in their local area, to acquire and upload data as UGC. Their involvement may be entirely voluntary, driven by altruistic motives, or may be rewarded in various ways. The potential of local citizens to act as intelligent sensors of changes in their environment has recently been explored in many contexts [3].



Figure 3: Wikimapia coverage of the campus of the University of California, Santa Barbara. Each rectangle denotes an area with a volunteered description.

The primary mechanism for communicating information about data quality is metadata, and most current metadata standards use some form of the *five-fold way*, the five components of data quality identified in the 1980s. In a recent paper [4] I argued that these standards, of which the Federal Geographic Data Committee’s Content Standard for Digital Spatial Data (<http://www.fgdc.gov/metadata/csdgm/>) is the oldest and best known, are increasingly flawed. I argued that 20 years of research has moved far beyond the approach to data quality of the 1980s, and that changes in standard practices in data modeling had also created new needs that were not envisioned at that time. Of particular importance for the Geospatial Web is the insistence that data quality is an attribute of a single data set, whereas it is increasingly important to know the *relative* data quality of *pairs* of data sets that are being integrated through mashups and other means. For example, a misregistered Google Earth image may be perfectly acceptable in an application involving features that have been georegistered using the same image, but may cause unacceptable problems if combined with features whose locations have been determined independently, as Figure 1 illustrates. I argued for a concept of *binary* metadata that described the ability of two data sets to work together, since such information cannot be deduced from the *unary* metadata records.

The paper argued for a more user-centric approach to metadata, a concept that is eminently compatible with the notion of user-generated quality assessment. Define as *metadata 2.0* an approach that collects the experiences of users who have attempted to access and exploit a data set in one of a number of possible applications. By being application-specific the approach addresses some of the concerns raised in the previous paragraph about the need for a binary form of metadata. It would rely on mechanisms for collection and integration that are now available on the Web, such as wikis. It could be managed by the custodian of the data set, such as a data warehouse, and each collection could be closely linked to its data set and made accessible to potential users. While metadata 1.0 has always struggled with the apparent unwillingness of

custodians and producers of geospatial data to invest the time and effort required to create effective metadata, the experience of VGI suggests that many users who have spent time struggling with the problems of accessing and using a data set would be willing to contribute their war stories to such a repository. It also raises an interesting research question regarding the tools that the spatial accuracy assessment community might be interested in developing to help users make and contribute their own assessments.

4. The quality of VGI

As the product of volunteers who are often untrained and unqualified, VGI clearly raises its own issues of spatial accuracy. Users have clear expectations about the quality of geospatial data produced in the traditional manner by national mapping agencies and corporations, based on experience, the standards published by the producers, or simply the reputation of the producer's brand. Collectively one might term such sources *authoritative*. By contrast VGI is simply *asserted*, by individuals with no brand, no experience or training, and no standards. There are, for example, no standards concerning the relationship between a feature's footprint as entered into Wikimapia and the feature's actual footprint; whereas there are often detailed standards regarding the quality of authoritative gazetteers.

This contrast between authoritative and asserted is somewhat artificial, however, since numerous intermediate cases exist. The term *citizen science* is often used to refer to the activities of communities of volunteers who nevertheless provide data that can meet the standards needed for scientific research. The Christmas Bird Count is a longstanding example. Its volunteers are amateur ornithologists, many of whom have a high degree of experience and skill, and the program has detailed protocols designed to ensure reliability. In other cases individuals may have little training but a contractual relationship with the sponsor of their data collection. However there is plenty of experience now to show that virtually any Web activity will attract its share of *griefers*, spammers, and other disruptive users.

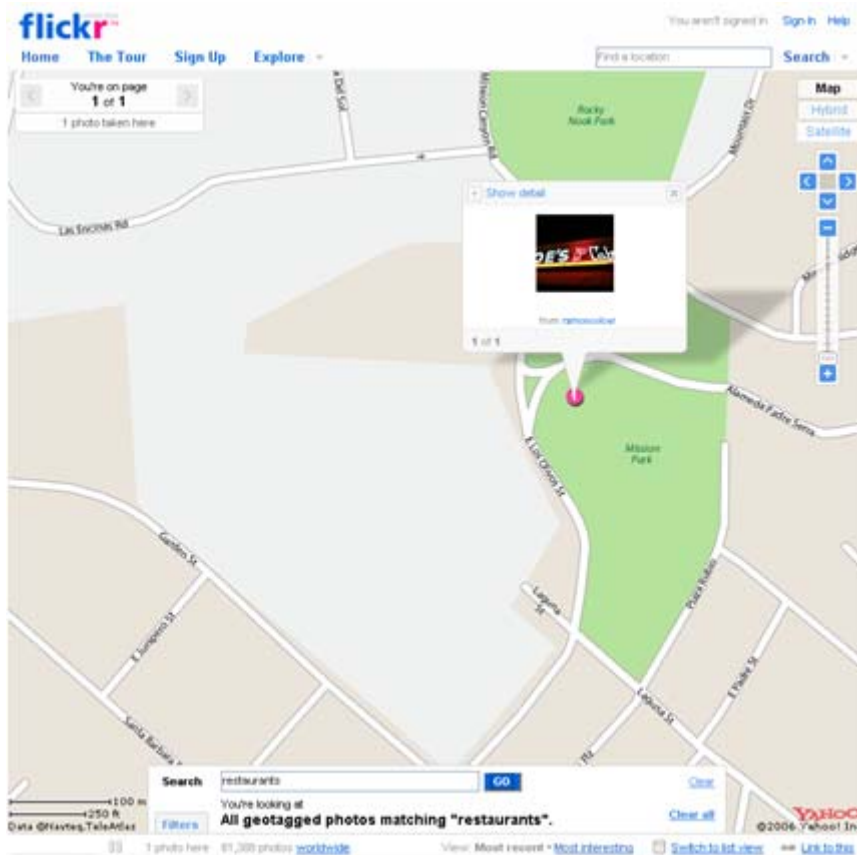


Figure 4: An obvious error of georegistration in a photograph contributed to Flickr.

Logical consistency checks are a normal part of the production process for many forms of geospatial data. For example, it is often possible to check various topologic properties against geometric properties, to ensure for example that all polygons close or that addresses are sequentially numbered along streets. Such checks

can often expose errors, since errors in geospatial data are often particularly glaring. It would be easy, for example, to detect and remove the error shown in Figure 4, where a user of Flickr has georegistered a photograph of a restaurant in the middle of a park. Nevertheless it is certain that VGI sites will attract their share of malicious behavior, and given the experience of the past fifteen years it would be interesting to attempt to anticipate how this will happen.

The question of legal liability also raises its head. Can a user who contributes geospatial information be eventually held liable for damages that result from its use? Is a contributor of street data to OpenStreetMap adequately protected against unintended consequences?

5. Conclusion

I began by arguing that spatial accuracy assessment occurs within a broad context that is defined by the acquisition, production, and use of geospatial data. While there is plenty of attention to the longstanding topics of accuracy assessment, visualization, and applications in this conference, the number of papers that place this work in the broader context of the data life cycle can be counted on the fingers of one hand. Yet that context has been changing dramatically in recent years. Google Earth has been called “the democratization of GIS” [5], and anyone with an Internet connection can now acquire an expert familiarity with many types of geospatial data and with simple GIS functions in a few minutes.

I believe that we as a community need to ask a series of questions, beginning with “What should spatial accuracy assessment mean in a world in which everyone is a potential user of geospatial data?” This is a very different perspective from the traditional one of the past 14 years, when it was possible to believe that the results of spatial accuracy assessment were of concern only to a small elite of geospatial professionals. Over time that community has been growing, and as it has grown it has engaged with new communities that lack virtually all of the background and experience of the geospatial professional. It now includes the neogeographers, a group that almost certainly outnumbers the geospatial professional, and will shortly include virtually everyone with an Internet connection, if it does not already.

6. References

- [1] A. Scharl and K. Tochtermann, editors, *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 Are Shaping the Network Society*, Springer, 2007.
- [2] J. Hastings, Automated conflation of digital gazetteer data, *International Journal of Geographical Information Science* (in press).
- [3] M.F. Goodchild, Citizens as sensors: the world of volunteered geography, *GeoJournal*, 2007, **69**: 211-221.
- [4] M.F. Goodchild, Towards user-centric description of data quality, Keynote presentation, *International Symposium on Spatial Data Quality*, Eenschede, Netherlands, June 2007.
- [5] D. Butler, Virtual globes: the web-wide world, *Nature*, 2006, **439**: 776-778.