BEYOND METADATA: TOWARDS USER-CENTRIC DESCRIPTION OF DATA QUALITY

Michael F. Goodchild, Dept. of Geography, University of California, Santa Barbara, CA 93106-4060, USA – good@geog.ucsb.edu

KEY WORDS: Metadata, Data Quality, Standards, Autocorrelation, Collection-Level Metadata

ABSTRACT:

Data quality statements are now entrenched in metadata standards worldwide. I contrast the needs of the user with the production-control mechanisms of the producer, and argue that metadata standards are producer-centric. To the user, the ability of data sets to interoperate is of major concern, as is the experience of prior users, the accessibility of quality statements, and the ease with which quality information can be handled in local software. Many of the newer geospatial tools that are oriented to the general public provide no data quality statements whatever. I present a series of use cases of metadata, and use them to argue for a reexamination of metadata standards, and the beginnings of a second generation of standards development that addresses these issues.

1. INTRODUCTION

As a term of technical English *metadata* is still comparatively new, not yet appearing in the Shorter Oxford English Dictionary (Brown, 2002). Yet the concept of "data about data" is as old as language itself, and humans have struggled for millennia with the task it attempts to solve – the succinct description of the contents of a body of information. Metadata must serve several somewhat independent purposes, all of them related to the ability of a potential user to assess the suitability of a body of information for a specific use. It must provide sufficient information to allow the user to assess quality, the degree to which the contents match the requirements, the technical details of transporting or using the information, information about legal and ethical constraints on use, and sources of further information.

Over the past three decades an enormous amount of useful research has accumulated on the topic of data quality, much of it reported in this and previous meetings in this series. In principle that research should inform the ways in which data quality are described, notably in metadata. In this paper I examine current geospatial metadata standards from this and related perspectives, and ask whether it is time for a substantial rethinking. Unfortunately the standards process is inherently conservative. It proceeds by consensus, and may therefore overlook the most recent research; and once standards are in place there is clearly an incentive not to change them, since the cost of doing so can be high, both in the effort required to write new standards, and in the legacy of compliant descriptions that must be replaced. Ideally standards should be devised once research is complete and the domain is fully understood. But in areas such as this research is likely to continue almost indefinitely, whereas the need for effective means of data description and documentation is constant and immediate.

The first part of the paper examines current standards, emphasizing the ways in which they address the quality of a body of information, or *data set* for short. This is followed by a series of critiques, drawing in part on previously published discussions and extending them to reflect current practice and the results of recent research. The paper ends with an outline of steps that need to be taken to revise our approach to geospatial metadata to make it better able to solve the problems it was intended to address. The emphasis throughout is on geospatial metadata and efforts within the mainstream geospatial community. Thus related efforts in other overlapping domains, such as the Data Documentation Initiative (http://www.ddialliance.org) or Ecological Markup Language (http://knb.ecoinformatics.org/software/eml/), which address wider subject areas that include limited attention to geospatial data, are not discussed except in so far as they rely on similar concepts.

2. EXISTING STANDARDS

The need to document and describe geospatial data was well established by the 19th Century, when national mapping agencies began to dominate the process of acquiring, compiling, and publishing geospatial data in analog form. Metadata began to appear in the form of marginalia (map legends, scales, publication history, and other information printed around the edge of the map) or sometimes on the back. Metadata about entire map series, including the details of the series' specification and the series index, typically appeared in separately published documents. In the computing world, data documentation began informally with such simple approaches as tape labels, progressed to file names and file headers, and eventually became the structured, digital approaches we see in today's standards.

As our ability to share data through such media as CDs or the Internet has grown, so too has the necessary complexity of metadata. Sharing of data with a colleague in the same department is comparatively easy, since both the custodian and the potential user probably share a common discipline and language and common set of expectations. But sharing data over the Internet with potential users in other countries, cultures, and disciplines is vastly more problematic, and in the extreme may even be comparable to the problems faced by Columbus in communicating with the native inhabitants of the Americas, or NASA in deciding what message to send with the Voyager spacecraft (http://voyager.jpl.nasa.gov/spacecraft/goldenrec.html).

One of the earliest attempts to devise a standard for geospatial metadata was made in the early 1990s by the U.S. Federal Geographic Data Committee (http://www.fgdc.gov), as part of a larger effort to establish the U.S. National Spatial Data Infrastructure. The first version of the standard, under the name Content Standard for Digital Geospatial Metadata (CSDGM) was adopted officially in 1994. Quality was recognized as a major part of the standard, and addressed in Section 2 using an earlier schema devised during the 1980s for the Spatial Data Transfer Standard (http://mcmcweb.er.usgs.gov/sdts/) and popularly known as the "five-fold way":

- Attribute accuracy, or the accuracy of the attributes by which geographic features are characterized;
- Logical consistency, or the degree to which the contents of the data set follow the rules by which they were specified;
- Completeness, or the degree to which the data set reports every relevant feature present on the landscape;
- Positional accuracy, or the degree to which locations reported in the data set match true locations in the real world; and
- Lineage, or details of the processes by which the data set was acquired and compiled.

In Version 2 of the standard, adopted in 1998, a sixth optional component was added to allow cloud cover to be described for remotely sensed data sets.

The International Standards Organization adopted its standard for geospatial metadata, ISO 19115, in 2003, with the intention that standards in member states would eventually be brought into compliance (the US compliance effort is being led by the Federal Geographic Data Committee; http://www.fgdc.gov). The ISO 19115 approach to geospatial data quality strongly resembles the earlier CSDGM, but following several commentaries (e.g., Guptill and Morrison, 1995) adds temporal accuracy, which had earlier been partially subsumed under completeness. Attribute accuracy was renamed thematic accuracy, but otherwise there is little effective difference between the two standards, and a simple cross-walk has been devised (http://www.fgdc.gov/metadata/documents/FGDC_Sections_v40.xls).

Substantial efforts have been made to accommodate the standards within GIS software and data formats, allowing metadata to be stored with the data set itself, ingested along with it, presented in different formats, and made available to users. For example, ESRI's ArcCatalog supports several variants of both FGDC and ISO standards, allowing metadata to be imported in a variety of formats. Automated update of metadata during GIS manipulation is clearly a desirable goal (Lanter, 1994) – it would be good if every new data set created through such GIS operations as overlay or join could be automatically documented. In practice, however, this remains a largely elusive goal, particularly in the area of data quality, because of the difficulties associated with processing metadata that are largely text-based and because of gaps in our knowledge of error propagation (1998) and more generally data-quality propagation.

The FGDC is composed of federal agencies with a long tradition of production and use of geographic data, and detailed knowledge of the processes used in production. On the other hand the average user of geographic data may know comparatively little about the process of production, and may be more concerned with the effects of data quality on the user's particular analyses. As a producer-centric view of metadata, the FGDC standard emphasizes:

- details of the production process, such as the measurement and compilation systems used;
- tests of data quality conducted under carefully controlled conditions; and
- formal specifications of data set contents.

By contrast, a user-centric view would emphasize:

- effects of uncertainties on specific uses of the data, ranging from simple queries to complex analyses;
- simple descriptions of quality that are readily understood by non-expert users; and
- tools to enable the user to determine the effects of quality on results.

Goodchild, Shortridge, and Fohl (1999) have examined the alternative ways of describing data quality, and have argued that simulation provides a general and readily understood option.

3. ISSUES

The following 7 sections discuss issues that I believe need to be addressed if geospatial metadata standards are

to adopt a user-centric approach, reflecting both the state of research knowledge and the practices of current technology.

3.1 Decoupling

As analog representations, paper maps are characterized by a scaling ratio, or *representative fraction* (RF), which is defined as the ratio of distances on the representation to their corresponding distances in the real world. The fact that no flat paper model of the Earth's curved surface could ever have a truly constant RF is a minor issue for maps representing areas of small extent, but significant for maps of substantial fractions of the Earth's surface. The representative fraction acts as a surrogate for the map's contents, as formalized in the map's specification, such that maps with a coarser RF portray only the larger features of the Earth's surface. It also acts as a surrogate for spatial resolution, since there is a lower limit to the sizes of symbols that can be drawn and read on a map, and thus to the real-world sizes of features that are likely to be portrayed. Finally it also acts as a surrogate for positional accuracy, since national map accuracy standards commonly prescribe the positional accuracy of features on maps of a given RF.

In the 1980s comparatively few geographic data sets were available in digital form, and virtually all that were had been created by digitizing or scanning paper documents. Table digitizers and large-format scanners were regarded as an indispensable part of any GIS lab, and their use was included as a substantial part of any training program. The digitizing process introduces errors and uncertainties that add to those present in the paper document. For example, digitizing is found to introduce positional errors on the order of fractions of a mm in addition to those already present in the map, which are themselves of similar order. As a result positional errors in data sets derived by digitizing paper maps are themselves directly related to RF, and the RF of the original map is an effective measure of positional accuracy for such data sets.

Unfortunately this strategy fails for other data sets that have not been obtained by digitizing or scanning paper maps. In principle RF is not defined for digital data, since there are no distances in digital media to compare to distances in the real world. Goodchild and Proctor (1997) argue that by the 1990s the coupling of content, spatial resolution, and positional accuracy under a single surrogate measure had broken down. Data sets created with newer technologies, such as digital orthophotos, were never in analog form and never had an RF to inherit. Spatial resolution is well defined for raster data, but its definition for vector data is often problematic, since for area-class maps it is related both to the minimum patch size (minimum mapping unit) and to the level of detail with which boundaries are drawn.

However little of this is evident in the current metadata standards. The FGDC standard mentions RF once, as a parameter defining source documents in its section on lineage, but it is not mentioned in the ISO standard. Positional accuracy is mentioned as one of the five components of data quality in the FGDC standard, and the ISO standard allows for both absolute and relative positional accuracy. Spatial resolution is not mentioned in the FGDC standard; in the ISO standard it is mentioned once as an optional element of "Core Metadata" but no further detail is provided. In summary, the authors of the standards appear to be aware of the difficulties associated with RF as a surrogate for several aspects of data quality, but have not fully adopted the decoupled approach that now seems needed in its place.

3.2 Uncertainty

Early discussions of the quality of digital geographic data sets focused on concepts of accuracy and error, perhaps reflecting the roots of GIS in area measurement (Foresman, 1998) and the earlier work of Maling (1989). Efforts were made to apply the theory of errors to the compilation of digital representations of features (Goodchild and Gopal, 1989), a practice that was already well established in surveying. By the 1990s, however, it had become clear that there was much more to data quality than error and accuracy, because for many types of geographic data it was unreasonable to assume that observations reflected some real-world truth modified by the process of measurement. The principle that an observed measurement x^* could be modeled as a true value x plus an error δx clearly could not apply to the nominal data of soil, land-use, or vegetation-cover classifications, but neither could one define a probability $p_i(i^*)$ that the true class found at a point would be i if the observed class was i^*. Moreover the definitions of the classes used in many mapping programs were clearly open to varying interpretation, and the maps compiled by two equally trained observers could not be expected to be equal.

Some progress was made using concepts of probability, but theoretical frameworks more compatible with vagueness, such as fuzzy and rough sets (Fisher and Unwin, 2005), proved very attractive for many forms of geographic data. The terms *error* and *accuracy* are now generally avoided in the research community, which tends instead to favor *uncertainty* as the umbrella term, along with *imprecision*, *vagueness*, and terms more closely related to the theories of evidence and non-Boolean sets. Yet neither standard shows any evidence of this significant change of thinking. The term *accuracy* occurs 7 times in the ISO standard and 85 times in the FGDC standard, while *uncertainty* occurs in neither.

3.3 Separability

Both standards distinguish clearly between the accuracies of attributes and positions. It has long been known, however, that in the case of geographic variation conceptualized as a continuous field the two concepts are not readily separable. Consider, for example, a continuous field

of topographic elevation, in other words a mapping from location **x** to a single-valued function $z = f(\mathbf{x})$. Suppose that at some location $\mathbf{x}_0$ a value z_0 has been measured. Except under special circumstances it will be impossible to separate the errors in these two parameters – for example, to distinguish between one case in which the correct elevation has been measured at an incorrect location, and another in which an incorrect elevation has been measured at a correct location; or any combination of the two. Only when some independent means exists to specify location, such as the existence of a survey monument or a sharp peak, or if the measuring instruments have known error metrics, is it possible to separate the two sources. The first case implies a shift from a continuous-field to a discrete-object conceptualization, while the second implies that each measurement inherits levels of error that were known *a priori*. Similar arguments exist for area-class maps, where it is impossible to separate errors of boundary positioning from errors of class determination, except when boundaries follow well-defined features such as roads or rivers, again implying a shift of conceptualization. Given these arguments, it makes little sense to attempt to specify the positional accuracy of isolines, or to separate attribute and positional accuracies for area-class maps, as both the FGDC and ISO standards do.

3.4 Granularity

In the earlier world dominated by paper maps the body of information described by metadata was a single map, and an intimate association existed between a map's contents and its marginalia. In the digital world, however, the concept of a data set is much more fluid. The digitized contents of maps can be separated into layers, based perhaps on print color or thematic divisions. They can be partitioned spatially as well into separate tiles, or integrated (*edgematched*) into larger, apparently seamless data sets. Agencies such as the Ordnance Survey of Great Britain have invested heavily in *on-demand* mapping, allowing the user to extract data for uniquely defined areas from seamless databases. To some extent the metadata for such sub- or super-sets can be inferred automatically. But it is clearly difficult to define a single positional accuracy for a combination of two layers of data.

In principle it is possible to define quality at a hierarchy of levels, from the single attribute or measurement of position through entire features, entire layers, and entire seamless databases. A topographic map is typically compiled from many sources using a complex process of analysis and inference, yet very little of this lineage is preserved when the finished product is made available for use. It is possible, for example, that the buildings have been obtained by photogrammetry; that the roads have been obtained by tracking vehicles; and that the rivers have been extracted from an existing digital source, and ultimately from topographic mapping at a different RF. Each of these sets of features has very different data quality characteristics that are difficult to capture in a single data-quality statement prepared according to a metadata standard.

In the 1990s a significant shift occurred in the dominant paradigm of geographic data modeling. Today, projects are likely to be supported by integrated databases containing representations of many different types of features, linked by the full range of relationship types defined by object-oriented principles: specialization, association, aggregation, and composition. Existing arrangements for handling metadata attempt to describe the database at the level of the class or collection of features (the *feature data set* in the terminology of ESRI's Geodatabase). But one could equally well argue that metadata are needed at the level of the entire database, and that the quality of the information about relationships between classes needs to be described.

3.5 Collection-Level Metadata

Metadata that describe the properties of individual data sets are sometimes termed *object-level metadata*, since they focus on a single information object within the larger framework of an entire collection. Many such collections exist in the form of geospatial data warehouses, digital libraries, or geolibraries (Goodchild, 1998), each containing potentially thousands of separate data sets. The U.S. Geospatial One-Stop (GOS; http://www.geodata.gov) is an effort by the federal government to provide a single point of entry into this distributed resource, with a union catalog that describes each available data set and points to its host server. GOS currently provides access to more than a thousand such collections, and its catalog includes tens of thousands of entries.

Consider a user faced with searching this distributed world for a data set meeting specific requirements. While GOS attempts to be a single point of entry, inevitably the collections available through it are only a subset of the collections available in the entire universe of servers. Thus the user requires some form of guidance as to which collections to search: GOS or one of numerous alternatives, any of which may contain terabytes and even petabytes of information. Collection-level metadata (CLM; Goodchild and Zhou, 2003) is defined as data about the contents of an entire collection, describing such characteristics as geographic and temporal coverage, the set of themes that dominate the collection, and the general level of data quality. Efforts have been made to develop content standards for CLM (http://www.alexandria.ucsb.edu/~lhill/alex-imp/Metadiversity_narrative.html), but the task of describing collections is far more complex than the task of describing individual data sets.

3.6 Autocorrelation

Tobler's First Law (Tobler, 1970; Sui, 2004) describes the tendency for "nearby things to be more similar than

distant things". There is now abundant evidence that this principle applies to errors and uncertainties in geographic data sets. For example, we know that errors in elevation in a digital elevation model are strongly autocorrelated, such that nearby errors tend to be similar, and indeed if this were not so our ability to estimate such properties as slope and curvature would be severely impaired (Hunter and Goodchild, 1997). Recently such errors have been analyzed within the framework of geostatistics, which formalizes Tobler's First Law as the theory of regionalized variables.

There are many common causes of this pattern of autocorrelation. Any geographic data set inherits errors and uncertainties from many parts of its compilation process. For example, misregistration of an image affects the positions of all of the features extracted from that image; and misclassification of an agricultural field from a rasterized aerial photograph affects the classes assigned to every pixel intersecting that field. Goodchild (2002) has discussed the implications of the common practice of storing every coordinate in a GIS in absolute form, and has proposed a radically different design which he terms a *measurement-based* GIS. By storing the uncertainties associated with each measurement and the process by which the database is obtained from the measurements, it would be possible to update automatically when improved measurements become available.

In summary, it is known that the correlations or covariances of errors of attributes and positions is as important, if not more important, than their variances – that the joint properties are at least as important as the marginal properties. Such covariances account for the widely observed tendency for *relative* errors in spatial databases to be less than *absolute* errors. For example, even though a road segment may be substantially out of place in absolute terms, its representation in a spatial database is likely to record its shape with a much higher degree of accuracy. While it is difficult to measure position on the Earth's surface in absolute terms to much better than a meter due to Earth tides, wobbling of the axis, poor approximation of the geoid, and tectonic movement, it is possible to measure relative position to mm over substantial distances.

Knowledge of covariance of errors may be of only limited significance to the visualization of geographic data in maps, but it is critical to any analysis of the propagation of uncertainties during manipulation of spatial data sets. Virtually all interesting products of GIS analysis, from simple measures of slope or area to complex analyses, respond directly to the covariances of errors and uncertainties. Thus appropriately defined parameters should be an essential part of any attempt to describe data quality in metadata. Yet current standards focus entirely on marginal properties such as mean positional error.

3.7 Cross-Correlation

This discussion of autocorrelation leads directly to the final issue, which is in many ways the most problematic. Although the ability to overlay disparate layers is often presented as a major advantage of a GIS approach, many users will have experienced the problems of misfit that almost always occur. If the positional uncertainties in two layers are other than perfectly correlated, and if both layers contain representations of the same features, then the result of overlay will be a large number of small slivers, formed by the two versions of each feature. On the other hand if the two layers were both obtained from the same root, then uncertainties may be perfectly correlated and no misfits will occur.

While it is possible to describe the uncertainties of each data set independently, the results of overlay cannot be obtained from this information – misfit is a *joint* property of a pair of data sets, rather than a *marginal* property of either of them. More broadly, one might define *binary* metadata as metadata describing the ability of two data sets to interoperate, and note that such metadata cannot be obtained from the separate metadata descriptions of each data set (although perfect correlation of uncertainties might be inferred in some circumstances by comparing information about lineage).

Such information seems essential to the entire GIS enterprise, in so far as it is based on the ability to overlay, and to extract layers of data from widely disparate sources. Great effort has been expended over the past decade at making geographic technologies and data sets interoperable. Yet data quality has received very little attention in this drive to open, interoperable GIS (http://www.opengeospatial.org), and the approach to metadata reflected in the standards is uniformly unary.

4. THE WAY FORWARD

While the focus of this discussion has been on data quality, several other authors have commented on the need to reopen the metadata question. Schuurman (http://www.sfu.ca/gis/schuurman/research/onto.html), for example, has identified several ways in which changing practices, particularly the increasing sharing of data across widely disparate cultural and disciplinary divides, is prompting a demand for more comprehensive and thus more complex metadata. It has long been known that metadata have the potential to exceed data in sheer volume, and it is not unreasonable to expect that as much effort be spent documenting data as in compiling them.

That said, however, the willingness of data custodians to document and describe data is clearly an issue, and there is no doubt that the generation of metadata lags behind in many domains (National Research Council, 2001). Thus one can expect resistance to any effort to reexamine metadata standards if the result is likely to be greater complexity. Yet it is clearly naïve to expect that the costs

of metadata creation should be borne entirely by the custodian, and models of geographic data dissemination that recover at least part of the costs of creation from users are attractive in this regard.

I believe that the responsibility for improving the description of data quality in metadata lies firmly with the research community, who must decide whether the results of research are sufficiently stable and conclusive to merit being embedded in standards. What is needed is a concerted effort on the part of this community to define a more enlightened and research-based approach; and if successful, to lobby the standards community for its adoption. This seems to me to be one of the most important things we can do to bring the results of our research into practical use, and to demonstrate its benefits.

REFERENCES

Brown, L., editor, 2002. *Shorter Oxford English Dictionary*. Fifth Edition. Oxford University Press, Oxford.

Fisher, P.F. and D.J. Unwin, editors, 2005. *Re-Presenting GIS*. Wiley, New York.

Foresman, T.W., editor, 1998. *The History of Geographic Information Systems: Perspectives from the Pioneers*. Prentice Hall PTR, Upper Saddle River, NJ.

Goodchild, M.F., 1998. The geolibrary. In S. Carver, editor, *Innovations in GIS 5*. Taylor and Francis, London, pp. 59-68.

Goodchild, M.F., 2002. Measurement-based GIS. In W. Shi, P.F. Fisher, and M.F. Goodchild, editors, *Spatial Data Quality*. Taylor and Francis, New York, pp. 5-17.

Goodchild, M.F. and S. Gopal, editors, 1989. *Accuracy of Spatial Databases*. Taylor and Francis, Basingstoke.

Goodchild, M.F. and J. Proctor, 1997. Scale in a digital geographic world. *Geographical and Environmental Modelling*, 1(1), pp. 5-23.

Goodchild, M.F., A.M. Shortridge, and P. Fohl, 1999. Encapsulating simulation models with geospatial data sets. In K. Lowell and A. Jaton, editors, *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*. Ann Arbor Press, Chelsea, Michigan, pp. 123-130.

Goodchild, M.F. and J. Zhou, 2003. Finding geographic information: collection-level metadata. *GeoInformatica*, 7(2), pp. 95-112.

Guptill, S.C. and J.L. Morrison, editors, 1995. *Elements of Spatial Data Quality*. Elsevier, Oxford.

Heuvelink, G.B.M., 1998. *Error Propagation in Environmental Modelling with GIS*. Taylor and Francis, Bristol, PA.

Hunter, G.J. and M.F. Goodchild, 1997. Modeling the uncertainty in slope and aspect estimates derived from spatial databases. *Geographical Analysis*, 29(1), pp. 35-49.

Lanter, D.P., 1994. A lineage metadata approach to removing redundancy and propagating updates in a GIS database. *Cartography and Geographic Information Systems*, 21(2), pp. 91-98.

Maling, D.H., 1989. *Measurement from Maps: Principles and Methods of Cartometry*. Pergamon, Oxford.

National Research Council, 2001. *National Spatial Data Infrastructure Partnership Programs: Rethinking the Focus.* National Academy Press, Washington, DC.

Sui, D.Z., 2004. Tobler's First Law of Geography: a big idea for a small world? *Annals of the Association of American Geographers*, 94(2), pp. 269-277.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2), pp. 234-240.

ACKNOWLEDGMENTS

Support from the National Science Foundation (Award 0417131), the Army Research Office, and the National Geospatial-Intelligence Agency is gratefully acknowledged.