The Need for Metadata for GIS Data Layers and Products

Stephen Leisz

Email: sleisz@care.org.vn
M&E/GIS Advisor, CARE International in Vietnam
P.O. Box 20
Hanoi

This paper focuses on the need for metadata for products derived from geographic information systems (GIS). As GIS technology is increasingly integrated into Agriculture and Natural Resource projects and programs in SE Asia, there is a growing need for products derived from these technologies to have metadata attached to them so that users can determine the quality of the products and the products' appropriate uses. This is especially the case when the products are used as "scientific" support for proposed policy changes or interventions. It is also important that the conclusions drawn from the analysis used to create these products is not overstated. In other words product's metadata should have information stating the limitations encountered in making the product and the limitations of interpretations that may be derived from the product. Without these caveats there is the danger that the products will be looked at uncritically as infallible "scientific" data, by users who do not understand the limitations of the different resource information technologies, and the products may be used to influence policies that they should not be applied to.

I. INTRODUCTION

The first class I had relating to Resource Information Technologies was a class at The University of Wisconsin – Madison called "Principles of Land Information Systems" taught by Professor Jim Clapp, a Civil Engineer who spent his life working with GIS, LIS, GPS, and Remote Sensing. On the first day of class Professor Clapp made two important points. First, he suggested that GIS's and LIS's are not new. Rather they have been around since the first cartographer drew a map and kept files related to the information found on the maps. Professor Clapp suggested that the only new aspect of a GIS or an LIS was the introduction of the computer and the increased speed with which the information that is mapped and attributed can be accessed and used. The second point was that even with the introduction of the computer, the system can still fail if the user does not understand the quality of the map product. Furthermore, with the GIS/LIS in digital form those failures can be on a scale not dreamed of previously.

To illustrate the second point, Professor Clapp told an anecdotal story about a large irrigation project that was being constructed in the State of Arizona in the U.S.A. The engineers working on the project had very carefully laid out their plans and referenced them to all the pertinent maps and data which were stored on an LIS. Construction had begun. Then one of them happened to check some of the data and realized that if they continued as the work had been specified the project would be built so that the water would flow in the opposite direction from the desired flow direction. Why? Because even though the maps that had been digitized and put into the LIS the project was using were in the same UTM coordinate system, they were based on two different datums. At that time in the U.S.A. a new datum, NAD83 was being adopted to replace NAD27 and some of the maps put into the LIS were based on NAD83, while others were based on NAD27. No one had caught the mistake earlier, because the data's metadata was not properly recorded.

Since that day I've tried to be aware of errors that can be caused by reliance on maps that are of poor quality or do not have the proper metadata, so that use of the maps is made more difficult. As I've been aware of this I've recognized some tragedies that have occurred because of these types of errors. Examples that I've catalogued include: the recent shooting of an Indonesian border guard by International troops near the East Timor (some reports indicate that the two sides may have misinterpreted where the border was because of poor quality or contradictory maps); the death of a New York City construction worker in 1995 who was digging up a sewer line, his map located a gas line a few meters from where it really was and when he hit it he was incinerated; and countless border conflicts caused by people's land boundaries being either mis-mapped or misrepresented on maps.

These examples, and Professor Clapp's dictum, suggest that in this day of the computerized GIS there is still a need for basic quality control in how maps are made and attribute data recorded. Furthermore, given the ease with which digital map data is combined to form new information products, there is an even greater need for accurate metadata being attached to each of the digital layers and attached to the "new information product" so that users understand the limitations of the product, i.e. of the computer generated map, they are using.

II. METADATA AND QUALITY CONTROL WITHIN A GIS

There are two aspects of the data within a GIS that need to be "quality controlled" and have metadata attached to them. The first is the actual data put into the GIS (the individual map and attribute layers) and the second is the information products that are made by carrying out spatial analysis on the GIS map and attribute layers.

Data Input into GIS

One reason why the types of errors that are mentioned above happen is that, although paper maps have always been looked at as authoritative renderings of "reality," people who use and consult digital data often implicitly expect it to be of higher quality than conventional map data. The reason for this is that there seems to be an inherent belief that if something is in a computer it is superior, due to a certain technological advantage, over earlier, paper maps (Bernhardsen 1992). It is because of this belief that producers of digital data, and those who "use" this digital data, have to pay extra attention to documenting the lineage of the data, e.g. publishing metadata for the data within a GIS.

A traditional way of thinking about quality in mapping is to suggest that it has several properties, which include positional accuracy, attribute accuracy, logical consistency, and completeness (Schmidley 1997). Positional accuracy refers to how closely a feature’s position on a map represents its position on the earth’s surface. For example, the U.S. national map accuracy standard states that not more than 10% of well-defined points tested shall be in error by more than 1/30th of an inch (0.85 mm) measured on the publication scale map for maps larger than 1:20000 scale, and for maps with a scale smaller than 1:20000, the measurable error is 1/50th of an inch [0.51 mm] (USGS). Attribute accuracy refers to how accurate the descriptive information that is associated with the map is. Logical consistency refers to whether the position of the geographic data represented on the map makes sense (e.g. rivers are in the correct location and do not appear to flow uphill). Completeness refers to whether the map actually contains all the features that it should. This last is often defined by the type of map it is (e.g. a land-cover map can be complete, yet not have soil information for the area that it covers). Many of these measures of map quality reflect the methods used to collect data for inclusion in the map and the care with which the map was made.

GIS Products

The conception of map quality described above is appropriate for hard-copy paper maps, since those maps are tied to the scale at which they are made. However, when one moves the data into a GIS the concept of quality needs to change. In a GIS metadata needs to be attached to each layer in order to help ensure the quality of the products that may be produced using the stored information. This is the case as users need to know about each layer so they can choose the correct layers for the analysis that they wish to do and for the products they wish to produce.

In a GIS, maps are stored in digital form. Therefore, their display scale can change. However, the accuracy of the maps does not change, as scale restricts type, quantity and quality of data (Star and Estes 1990). Therefore, when using data to analyze a problem, it is necessary to match the appropriate scale to the level of detail of the problem being investigated or the project being carried out (Burrough 1996). Enlarging a small-scale map does not increase its level of accuracy and using this enlarged map for analytical purposes may lead to misanalysis and false results.

This last point touches on the fact that within a GIS maps are usually used in conjunction with other maps (e.g. they are represented as layers of data that can be overlain, intersected, buffered, etc.). Each of the layers comes with error attached to it. Because of this there is the danger of cascading error and propagation of error through the GIS analytical process. Any new layers created from the original layers will have all the errors from the previous layers within it (propagated error) and possibly these errors will be compounded in the final product (cascading error). Due to these errors, solutions derived to a GIS problem may be inaccurate, imprecise or erroneous. This raises the point that inaccuracies, imprecision and error may be compounded within a GIS that employs many data sources (Foote and Huebner 1995). Furthermore, a new form of error, the spurious polygon, may enter into the data product (Burrough 1996).

In a GIS analysis, error can only be guarded against, or minimized, and quality of the product controlled, if each data layer is documented with information about it (e.g. its source, accuracy, the methods used to collect the data in it, etc.). This information allows the user to make informed decisions regarding what data they should use within their analysis in order to get the best quality results at the scale they need. Such information, or metadata, may include: map source, its original scale, methods used to gather the data, the accuracy standards it meets, its projection, datum, ellipsoid, etc. Thus, for quality control within a GIS, it is necessary not only to know the accuracy of the individual data layers, but also to make sure that the layers are used for analysis in a fashion that produces accurate results (Foote and Huebner 1995).

III. METADATA IN VIETNAM AND QUALITY CONTROL OF GIS PRODUCTS

Data within GIS

With these issues in mind, what is the current situation in Vietnam? Geographic data in hard copy and digital form comes in variable quality. In hard-copy form, the largest scale topographic maps that are available for the whole country are at a scale of 1:50000 (Christ and Kloss 1998). It is hard to assess the quality of the data represented in these map sheets, because methods used to collect the data are not readily known and the presence of metadata for the maps vary in quality from sheet to sheet. On some of the map sheets metadata is present, but there is no display of estimated map accuracy. On others, there is nothing other than the title of the map sheet. In the latter case, if the user wishes to integrate the map sheets into a GIS for use with other data, they must make a best guess regarding the map's metadata. At the least, this means determining the projection of the map (there are at least two, UTM or Gauss) and the datum the map is based on (at last count I could determine four datums used at different times in Vietnam: Indian 1960, Pulkovo 1942[?], Hanoi 1972, and WGS 1984).

For digital data, the situation is the same. Almost all the digital data in Vietnam originates from the aforementioned maps or comes from analysis of satellite imagery (in the case of vegetation and land use maps). In almost all cases complete metadata is not distributed with the data and no estimates of the accuracy of the data are done. For digital topographic maps, this means that the lineage of the map is unknown and that there is no estimate of positional accuracy. For thematic maps derived from satellite imagery, this means that there are no accuracy assessments provided of the analysis that has been done (vegetation cover, land-use patterns, etc.). Another shortcoming to the digital data is that in the cases I have examined there are digitizing errors (undershoots, overshoots, erroneous nodes, etc.) that add another dimension to the error already found within the original hard-copy maps.

Data Products

Products derived from these data layers also often do not have metadata. Therefore, it is difficult to estimate the accuracy of the products and also difficult to place a confidence factor on the use of these products. Another result is that often the results of GIS analysis are used at a scale for which they were not intended and inappropriately applied to policy questions. This misapplication is carried out equally by international organizations and by national organizations and is not a phenomenon unique to Vietnam.

While international organizations do not produce large-scale topographic maps of Vietnam, they often do produce interpretations of satellite imagery (such as SPOT and Landsat) for estimating vegetation coverage and land-cover change and this analysis of geographic data has the capacity to influence project design and government policies. In these areas international organizations often have not done a very good job of providing accuracy assessments of their work, and in some cases, they have not done a good job of providing metadata.

In most cases, image interpretations are at small-scale and use either unsupervised classification techniques, or interpretations of NDVI as surrogates for different types of land-cover, in their vegetation or land-cover analysis (see, for example the Pathfinder project www.bsrsi.msu.edu). These methods and the fact that international groups usually do not have ground data to work with, make it hard for them to provide accuracy assessments of their results. This provides the rationale for their products lacking complete metadata. To further complicate matters, some organizations and researchers have used these derived products to analyze phenomena at scales which are inappropriate given the data sets. Fox et al. (1999) give an explanation of how this happens with regards to land-cover in Southeast Asia. First, it is noted that most of the land-cover analysis done using satellite imagery for the area classify around 30% of the land-cover as an "other" or "shrub" class. This is done because the researchers doing the classification do not necessarily understand the landscape dynamics and farming systems in use and the scale of the classification (1:250000) is to small to permit an analysis of the mosaic of land-cover types found within this class. Other researchers then use this data and analyze it at a larger scale to suggest that large amounts of deforestation are taking place, an argument that might not be supported if the land-cover were analyzed at a scale appropriate to the scale of the data or if a land-cover analysis that captured the dynamics of the landscape at large scale were done.

Another example is found by examining a presentation on forest cover change given at a workshop in Vinh in 1998. This presentation used small scale land-cover data derived from the Pathfinder project. Pathfinder separates land-cover into four classes: forest, non-forest, water and cloud. Using this data Brunner and Nielsen (1998) suggested that the fallow system in the Ca River area is a 1-2 year cycle. However, field studies in the area suggest that this is not the case. Rather, field sizes are probably too small to be adequately analyzed using 1:250000 data derived from Landsat images and the land-cover is in a mosaic that is much more complicated than forest and non-forest. What may have been seen by the analysis is a cycle that has swidden rice for one-year, maize for one-year, then a longer-term cover of cassava. Since cassava provides a full ground cover after a few months, analysis of the type used by Pathfinder would classify the cassava areas as "forest." After two to five years of growing, the cassava is harvested and the Pathfinder analysis would interpret it as "non-forest" again, thus providing the estimate of 1-2 years of fallow.

These two examples show how data products can be derived inappropriately from the analysis of data within a GIS and in both cases, the products could be entered back into a GIS. If this is the case, then, unless the process by which the analysis is done and the products are made is appropriately documented, there is a danger that the products could be unquestionably used in policy making decisions.

IV. WHY DOES THIS MATTER?

A very basic reason why the inclusion of metadata within a GIS matters is that maps, and the GIS itself, are supposed to be used, they are not end results of a project, rather they are tools that are supposed to be used after the project implementing them has ended. The systems that are set up are supposed to be used for planning and managing a multitude of projects ranging from economic development projects to environmental projects. If there are questions regarding the quality of the data that can not be resolved because metadata is not present, then either questionable quality maps will be used or new maps will need to be made each time a new project is started. In either case, the money spent collecting and inputting the original data and building the GIS will have been wasted. Thus, the inclusion of metadata within a GIS becomes an economic consideration.

Burrough (1996) gives a second reason. He suggests that, "it is implicit in the whole business of geographical resource information processing that the collection and processing of environmental data leads to improvements in environmental management and control. This can only be so if the data that are collected, entered, stored, and processed are sufficiently reliable and error-free for the purposes for which they are required." If the data and data products are not error free or do not have metadata attached to them to inform the user about their relative degree of error, scale, etc., there is the danger that the data and products will be applied to problems for which they are poorly suited. This can lead to poor analysis being done, inappropriate solutions being proposed, and poor management decisions being made.

A third reason is that products from a GIS can be, and are, used to influence policy decisions. If the GIS products have errors in them, the resulting policies may be flawed. Also, if the product’s limitations are not explained in attached metadata, the analysis may be misapplied, also resulting in flawed policy decisions. Flawed policy decisions have the potential to cause a large amount of harm and waste large amounts of money.

V. DISCUSSION AND SUGGESTIONS

This paper is an attempt to move the discussion of quality control beyond the data that goes into a GIS and towards the question of how to understand the quality of the products derived from the analysis of data within a GIS. It appears that a fundamental aspect to solving this problem is the inclusion of metadata for each layer of data within the GIS and also for the analytical products produced using the GIS. The centrality of metadata to quality control reflects a change in the concept of quality control for geographic data. This shift in the way of thinking about map quality reflects what Cartwright (1993) suggests with regards to spreading GIS technology into developing countries. He suggests that users have to internalize newly acquired technical knowledge to the point where it begins to shape their conception of their job and how to do it. In this new view, maps are no longer the central concept, rather the GIS database is the central concept and information products are spun off as maps or as analysis of data to fit special needs. In either case, if the products are going to be useful, then attaching metadata to the data that goes into a GIS and to the analysis and products produced from it is central to the GIS function. Otherwise the result could reflect the famous computer dictum: GIGO – garbage in, garbage out, **and nobody would know the difference.**

Given that metadata should be an integral part of all GIS data and products produced in Vietnam, how can the geographic information users start to address this need? Following are three suggested steps:

1. Publicize current map accuracy and quality standards already in place and encourage their use for any mapping projects undertaken.
2. Agree to metadata standards for GIS information (both for digitized vector maps and for land-cover / land-use products derived from remotely sensed information)
3. Develop and use metadata forms and require that the metadata is distributed with data when it is transferred from one user to the next (including international groups who produce geographic information and analytical products for use in Vietnam)

In order to encourage the adoption of these types of practices, other groups of geographic information users in areas where I have worked have formed users groups whose main task is to encourage the recording and distribution of metadata for geographical data and analytical products derived from GIS. This could also be done within the context of Vietnam and would start the geographic data user community on the road towards increasing the quality of geographic data used throughout the country

REFERENCES

Bernhardsen, T. Geographic Information Systems. 1992. VIAK IT. Arendal, Norway.

Brunner, J. and Nielsen, D. 1998. "Ca River Basin Forest Cover Analysis: Preliminary Results." World Resources Institute.

Burrough, P.A. Principles of Geographic Information Systems for Land Resources Assessment. 1996. Oxford University Press. New York.

Cartwright, T. "Geographic Information Technology as Appropriate Technology for Development," in Diffusion and use of Geographic Information Technology. 1993. Klumer Academic Publishers. Boston.

Christ, H. and Kloss, D. 1998. "Land Use Planning & Land Allocation in Vietnam with Particular Reference to Improvement of its Process in the Social Forestry Development Project Song Da (SFDP)." Consultancy Report No. 16. April/May 1998. ADB Forestry Sector Project. Hanoi, Vietnam.

Foote, K. and Huebner, D. 1995. "The Geographer’s Craft Project, Department of Geography, University of Texas at Austin."

Fox, J., Dao Minh Truong, Rambo, A. T., Nghiem Phuong Tuyen,Le Trong Cuc, and Leisz, S. 1999 (in press). Shifting Cultivation without Deforestation: A Case Study in the Mountains of Northwestern Vietnam.

Schmidley, R. "Quality Control in Mapping: Some Fundamental Concepts." Surveying and Land Information Systems. Vol. 57, No. 1, 1997, pages 31 –36.

Star, J. and Estes, J. 1990. Geographic Information Systems: an Introduction. Prentice Hall. Englewood Cliffs.