COMMUNICATING THE RESULTS OF ACCURACY ASSESSMENT:

METADATA, DIGITAL LIBRARIES, AND ASSESSING FITNESS FOR USE

Michael F. Goodchild
National Center for Geographic Information and Analysis, and
Department of Geography
University of California
Santa Barbara, CA 93106-4060, USA
Phone: 805 893 8049; FAX: 805 893 7095
Email: good@ncgia.ucsb.edu

Abstract

The concept is proposed of a data life cycle extending from the initial collection of data in the field or by remote sensing to eventual archiving. Although GIS has been perceived to date largely as an analytic technology confined to a small part of the data life cycle, its influence is increasingly seen as extending throughout the cycle, in the form of newer technologies such as field GIS and digital spatial data libraries. The concept of a data model is introduced and examples are given of the role of data modeling along the life cycle. Accuracy assessment and concern for broader issues of data quality are critical if data are to be assessed as to fitness for use by the various actors in the life cycle. GIS data models must be extended to accommodate essential information on accuracy. The role of accuracy assessment in metadata is discussed, and a new view of metadata is proposed with extended functions and a hierarchical structure.

INTRODUCTION

Accuracy assessment of natural resource data can play a narrowly defined role in the quality control of a range of essential mapping activities, but it can also play a much larger role in supporting the collection, management, storage, use, and archiving of geospatial data. This paper explores the role of accuracy assessment throughout the *data life cycle*, a term that encompasses all of the stages through which data pass from original observation to eventual filing. There are several reasons for doing this. First, it is clear that any representation of the real world must be an approximation. Hence it is important that the degree of approximation, the amount of information lost when the representation was created, and the errors introduced along the way be maintained as an essential part of the data as they move through the life cycle. They should also be made available to whoever comes into contact with the data, for whatever purpose. Second, there has been a trend recently toward a broadening of interest in the applications of GIS from a fairly narrow stage in the life cycle, where GIS technology is applied to existing databases for the purposes of modeling, analysis, and decision-making, to a concern for the use of GIS throughout the entire life cycle. GIS, it is argued, is not only a desktop technology for the analyst, but also potentially a field technology (Kevany, 1996), and is also important in the later stages of the life cycle when data are shared (Onsrud and Rushton, 1995) and archived (Smith et al., 1996; and see http://alexandria.sdc.ucsb.edu). In this respect there is an increasing degree of linkage and integration between GIS and the related technologies of GPS, remote sensing, wide area networks, and digital libraries.

The paper is structured as follows. The next sections present an overview of the life cycle, the roles of technologies and stakeholders within it, and the trends that are currently influencing it. The following section discusses metadata, the key element that allows data to travel successfully and usefully through the life cycle and to be shared across a range of applications. The final section looks at the implications of this perspective for accuracy assessment in three contexts: in terms of functions that must be supported; in the need for a hierarchical approach to accuracy assessment; and in the extensions that will be required for existing GIS data models. The concern throughout is with geographic (geospatial) data, defined here as data about specific locations on the surface of the Earth.

THE DATA LIFE CYCLE

Geographic data begin with *direct observation*, either in the field or from above. Measurements are made and transferred to some appropriate medium in analog or digital form (in analog measurement, the results are represented as proportional electrical signals, sound waves, distances on paper, etc.; in digital measurement the representation is coded as a series of digits). Various stages of *interpretation* may occur, as the data are transformed through human interaction. The data may be reformatted, perhaps by conversion from analog to digital form (digitization). They may also be resampled by estimating the values that would have been observed had observations been made at different locations, or aggregated by merging observations. The data may also be converted to *map* form as a general synthesis of knowledge about a defined study area. A *database* is created to allow ready access to the data for the purposes of processing, analysis, and modeling. The data may be *visualized*, modified by various transformations, and used for a range of activities designed to support decisions. In the later stages of the life cycle the data may be stored in some form of *archive*, designed to preserve them for future use, or to allow them to be used for other purposes. *Sharing* of the data may occur between investigators, perhaps across the boundaries of disciplines; many cartographic data, for example, are collected for a wide range of purposes which may or may not be well-defined and well-understood. Eventually the data will become unusable, perhaps because the technology used to store them is no longer supported (punched cards, paper tape, 7-track magnetic tape, 5.25 inch floppy diskettes), or because the medium has deteriorated, or because information necessary for access is no longer available (lost, burnt, destroyed).

The data life cycle is sometimes linear, but more often involves feedback loops of varying duration (Figure 1). Decisions made on the basis of analysis of data may modify the real world. Transformed data may replace earlier versions. Analysis may lead to further data collection.

[Figure 1 here]

The length of the data life cycle is highly variable. Historic data may be available in the form of early maps, and there are major efforts under way to preserve them in digital archives by organizations such as the Library of Congress. On the other hand it is already difficult if not impossible to access early data from remote sensing satellites, such as the early Landsat series, collected as little as 20 years ago. For many GIS projects the data life cycle may be a matter of months, as data are collected and assembled for some specific purpose. But even in these cases the life cycle is likely to be far longer than the duration of the GIS analysis itself. Analysis is increasingly a continuous activity, where GIS technology is applied almost instantaneously as data are collected. This kind of activity is now feasible in areas such as epidemiology, to provide early warning of disease clusters; in climatology; and in agriculture.

In the computer and information sciences, data modeling is defined as the selection of appropriate entities, attributes, and relationships in order to create a useful representation in a digital computer of some complex reality (e.g., Tsichritzis and Lochovsky, 1982). Data modeling is particularly important for geographic data because the gap between the complexity of the real world and the capacity of a digital database is so great; and because of the large number of options available, particularly for the digital representation of what is conceived as a geographically continuous surface (Goodchild, 1992). But data modeling is not confined to digital representations—rather, it occurs at multiple points in the data life cycle as information is reformatted, compressed, sampled, or otherwise transformed in ways that require a change of basic entities, attributes, or relationships.

For example, a process that a statistician would call point sampling can also be seen as a form of data modeling, in which irregularly spaced points capture complex variation that is conceived by the observer as spatially continuous. If the chosen data model is also to inform its potential users about the information lost when it was created, or the magnitude of the differences between it and the real world, then its entities, attributes, and relationships should include the structures needed to store parameters of error models, variograms, variance surfaces, or whatever is thought to be appropriate in the given instance. These error-related aspects of the data model might attempt to describe both the expected differences between recorded observations and the truth, due to measurement error. They should also describe the uncertainties anticipated when the full continuous surface is interpolated from the point observations. Note that in this example it is immaterial whether the representation is analog or digital—the data modeling issues exist in both cases.

Changes of data model, and decisions about data modeling, can occur at many stages in the data life cycle. They commonly occur:

- during observation;
- during interpretation of observations, such as the interpretation of vegetation boundaries on air photographs which creates linear entities;
- during digitization, such as the replacement of a smooth, analog line on a map with a polyline in a GIS database;
- during resampling associated with projection change or change of spatial resolution; generalization of data; and
- during assembly of results in support of decisions or for archiving.

Rarely does any one individual have a comprehensive view of the entire process of data modeling during the data life cycle. Some data model decisions are constrained by prior decisions made by the designers of observation instruments—the users of remotely sensed data are in this sense locked into a raster data model. Others are driven by communication mechanisms—the user of data expressed on a map, for example, is limited by the data models that can be expressed in map form. It is easy, for example, to create a map representation of a point or a line or a uniform bounded region, but more difficult to express continuous gradations of value using standard cartographic technique. The degree to which the technology of GIS imposes itself on the choice of representations is the subject of much recent discussion and research (Pickles, 1995).

GIS data models have evolved in a field driven largely by the need to represent information that historically has been portrayed in largely analog form on maps. As such, they are ideally suited to data that are static, two-dimensional, and planar, and that express information at a single level of geographic detail. The latter property is particularly problematic during the transition from analog to digital, since metric scale is normally defined from the analog properties of maps, and therefore has no well-defined meaning for digital data (Goodchild and Proctor, 1997). Instead, the stated metric scale of digital data refers to the scale of the map from which the data were digitized; to its thematic contents, since map scale determines the classes of features that can be shown; or to its positional accuracy, since map accuracy standards prescribe expected positional accuracies as a function of scale.

There has been much recent research devoted to extending GIS data models to three dimensions (Raper, 1989; Turner, 1992), time (Langran, 1992), multiple scales (Buttenfield and McMaster, 1991), and the curved surface of the Earth (Goodchild and Yang, 1992). On the other hand very little work has gone into extensions to accommodate knowledge of accuracy, except as additions to the standard documentation or metadata of a data set (http://fgdc.er.usgs.gov). This is unfortunate, since it means that there is nowhere in a GIS database to store the parameters and other aspects of accuracy that result from the kinds of research discussed in this volume. Moreover, it means that analysis of error propagation; visualization of error; and storage of error properties are difficult to achieve within current software without expensive and poorly linked additions.

Moreover, the frequent modification of data models that occurs along the data life cycle makes it difficult to trace the lineage of a given data item back to the original observations that supported it. For example, a single elevation value in a digital elevation model is likely to have been resampled from points at different densities; to be subject to errors of registration that affect all other points in its neighborhood similarly; and to share other sources of error with its neighbors as well. As a result, errors in geographic data sets tend to be strongly autocorrelated across space, and to have complex structures of spatial dependence. A single value in a geographic database is thus very far from a single, independent observation, and error models for geographic data tend therefore to be relatively complex. In most cases it is impractical to formulate or calibrate models of the various sources of errors that affect geographic databases—instead, error models must often be calibrated by direct comparison between database contents and reality. Much important and potentially useful information about data lineage, such as the locations of registration points, locations of original point observations, and details of interpolations are often lost, either because they are regarded as irrelevant or because of the cartographic tradition of portraying the world as an interpretation, rather than as the result of a set of scientific measurements. In summary, it is unusual to encounter cases in geographic data handling where traditional techniques of measurement error analysis can be applied.

Accuracy in the life cycle

Finally, Figure 1 shows the frequent changes of custodianship that occur in the data life cycle. Accuracy is defined as the difference between the value of a measurement recorded in the database, and the value of an equivalent measurement from a source of higher accuracy (in the case of geographic data, 'higher accuracy' can mean that the reference data are at a more detailed scale, or have been obtained by direct field observation). Unfortunately, a major source of weakness in this definition surrounds the meaning of the term 'equivalent'. When the custodian of a database assesses accuracy, it is the custodian who is the ultimate judge of whether a reference source is or is not equivalent—whether it is measuring 'the same thing'. In an extreme case, a custodian who mistook a database of vegetation for a database of geology might assess accuracy against entirely the wrong reference source. Thus it is possible for the accuracy of a database to change, either by improving or by deteriorating, when a database changes hands from one custodian to another. During the data life cycle such changes of custodianship can occur many times. Although the example of geology and vegetation may be extreme, it is relatively common for key elements of information about a data set to be lost in transmission, such as knowledge of the geodetic datum, or the details of a map projection, any of which can lead to inadvertent decrease in accuracy.

In summary, accuracy is a dynamic property of the data life cycle. With reference to Figure 1, changes of data model in the left-hand column, and changes of custodian in the right-hand column, produce the potential for significant changes of accuracy as the current representation is compared to the current reference source. Only effective transmission of metadata between custodians, and its effective updating after changes in data model, or repeated reassessment of accuracy, can ensure that adequate information on accuracy is always available to the data's users.

TRENDS AFFECTING THE LIFE CYCLE

This view of the data life cycle and the stages of data modeling is being profoundly affected by changes occurring in society, in computing, and in the application of geographic information technologies. This section examines four of these, and the impacts they are having.

First, GIS is becoming a much more portable technology than previously. In the past, the sheer bulk of the hardware needed for GIS ensured that it would be confined to those stages in the data life cycle that deal with analysis, modeling, and decision-making. GIS is still largely a desk technology, if not a desk-top technology, but recently technological developments have allowed it to move much closer to direct support of field observation. GPS is one such development, and the technologies that now support the direct integration of GPS observations with GIS databases. Another is the advent of the laptop computer, and the portable devices exemplified by the Apple Newton—the so-called palmtop computers or personal data assistants. These currently facilitate direct field collection of data, and its uplinking to databases, but will in the future also support more sophisticated analyses such as the direction of point sampling in the field based on statistical principles. They are being implemented in utility companies, forestry agencies, delivery companies, and a host of similar operations that involve acquisition of information on the ground. More efficient dissemination of remotely sensed data will allow them to be integrated with imagery in close to real time for purposes such as fighting wildfire. The set of field GIS applications now includes precision agriculture (Vandenheuvel, 1996), which comes close to locating a fully featured, GPS-linked GIS in the cab of a combine harvester.

Second, the advent of object-oriented thinking has begun to permeate GIS, and to lead to a series of fundamental questions about traditional GIS architectures. One in particular is of importance to accuracy assessment. Most current GIS architectures store location in absolute terms as pairs of coordinates, either on a projected representation of the Earth's surface or in latitude and longitude. While the original observation may be that a given forest boundary occurs approximately 10m from the edge of a county road, the information actually stored gives measurements of the positions of both boundary and road in absolute terms with respect to the Earth frame, notably the Equator, Poles, and Greenwich Meridian. It has already been argued that many such GIS data items effectively hide their lineage, making error analysis difficult.

Object-oriented thinking favors a distinctly different approach, in which the original measurements are stored and absolute position is regarded as derivative, perhaps computed on the fly when necessary. Such *measurement-based* approaches to GIS are readily supported by object-oriented database designs and programming languages, which allow each location to *inherit* the lineage of its supporting measurements. If these change for some reason, as they often do when there is an opportunity to improve the positional accuracy of some aspect of a database, such as its geodetic control, the dependent data items can be made to update automatically. Similarly, analysis of error is now much easier because the lineage of each data item is not necessarily lost.

Third, the data life cycle as a whole can be seen not as a pipe for the flow of packets of data, but as a communication channel in which the medium of communication is a collection of measurements using some defined data model. In this context it is possible to ask questions regarding the channel's efficiency. Do the chosen data models result in effective communication between the various stakeholders in the data life cycle—do they allow the knowledge gained by the expert soil scientist in the field to be communicated with minimum loss to the eventual user of the data, such as a farmer planting crops? Is it possible to measure the channel's efficiency, and the loss of information resulting from various stages of data modeling? Do the media used at various stages in the cycle result in information loss or other constraints on communication? What needs to be passed through the information channel to make it as easy as possible to realize the benefits of the investment made in the data? How can the channel be broadened so that the data are as useful as possible to the largest number of disciplines?

The trend toward greater use of digital information technologies in the field has already been mentioned. At the other end of the data life cycle, concepts of digital libraries are beginning to influence the archiving of data, and browsing and searching of data resources by researchers and decision-makers. The Alexandria Digital Library project at UC Santa Barbara (Smith et al., 1996; http://alexandria.sdc.ucsb.edu) is one such project directed specifically at the issues associated with building a digital library for geographic data. Its objectives are 1) to develop digital library services for spatially referenced data accessible over the Internet, and 2) to exploit digital technologies in order to bring maps, images, and other forms of spatial data into the mainstream of future libraries.

METADATA AND ACCURACY ASSESSMENT

As data move along the data life cycle they encounter a number of actors who contribute to data's value, make use of data, or process and transform data in some way. The larger the number of actors, and the greater their physical and disciplinary separation, the greater is the challenge in making the data interpretable and useful. Moreover, the ability of the actor to make use of the data depends on the ability to search for suitable data, or more generally to browse.

In the traditional library, the processes of browse and search are supported both by the card catalog (and increasingly by its digital version), and by the order in which books are stacked on shelves. The digital equivalent of the card catalog is *metadata*, or data about data. In addition to the card catalog metaphor, it is useful to think of metadata as containing also the handling instructions for the data, such as details of its format; and also information needed by the user to determine fitness for use. Bretherton (quoted by Smith in Goodchild, 1996) has defined metadata as "information that makes data useful". An important component of metadata is information about the data's quality. Quality is something often taken for granted in the library, and in practice assured by the library's collection-building efforts; in the digital world, more explicit information on quality is often necessary, particularly in the case of geographic data, because of uncertainty about such factors as scale, positional accuracy, date of validity, and lineage.

The digital documentation of geographic data inherent in the concept of metadata is a very active area of research and development in the geographic data community at this time. Two related standards are already in place in the US as a result of efforts by the Federal Geographic Data Committee (FGDC): the Spatial Data Transfer Standard [SDTS; also known as Federal Information Processing Standard (FIPS) 173; http://fgdc.er.usgs.gov], and the Content Standard for Digital Geospatial Metadata, commonly known as the FGDC standard. Both address issues of importance to the data life cycle, particularly in the area of data quality. SDTS provides extensive detail on the components of data quality, and potential procedures for measurement of appropriate statistics. The general approach aims at ensuring that providers of data have access to appropriate mechanisms for describing what is known about data quality—the so-called "truth in labeling" approach—rather than the establishment of thresholds of quality that must be met.

In attempting to prescribe how data quality should be expressed, however, the details of the standards illustrate a fundamental weakness in the standards enterprise—its inability to evolve in ways that are closely linked to the research frontier. The methods for measuring data quality discussed in the standard reflect the state of research that existed when the standard was written—in the mid-1980s—rather than the kinds of cutting-edge research being discussed in this volume. They are largely silent, for example, on the key issue of spatial structure of error. While it is recommended that accuracy of thematic data sets be described using the error (or confusion) matrix, there is no corresponding recommendation regarding the correlation structure of such errors. So although it is possible to rely on the standard to provide estimates of the uncertainty of classification at a point, it is impossible as a result to estimate the uncertainty associated with any analysis that requires simultaneous examination of more than one point, such as measurement of area, in a digital representation of a multinomial field. Similar problems exist in the suggested methods for describing error in interval/ratio fields, or in discrete point, line, or area objects. Although there is now abundant research on these issues, and although the standard allows for evolution in its specific profiles, there are no comparable mechanisms for modification of the standard itself, which has the effect of freezing progress in the practical implementation of recent data quality research. Of course, the architects of SDTS could hardly have anticipated the wealth of progress that would be made in geographic data quality research in the past decade.

Metadata are often conceptualized as the digital equivalent of a card catalog. But this is inherently limiting, for a number of reasons. First, it limits the digital world to being the analog of the traditional one, rather than encouraging the user to take advantage of the digital world's potential for doing new things, or doing old things in new ways. Second, it makes metadata the exclusive purview of the data producer, who must anticipate possible uses of the data in deciding how best to describe them. In the traditional library this was clearly reasonable, and librarians have developed highly sophisticated ways of cataloging materials in anticipation of the needs of users. But in the digital world there is the opportunity to think of metadata not as something produced by the data's originator or custodian, but as a function of both the producer and the user, or a tool for assessing the degree of fit provided by the data between the producer's assets and the user's needs. In this sense there is no single approach to metadata, but a range of approaches that tries to accommodate to the range of levels of expertise and requirements of the user community. In the case of data quality, for example, metadata might serve one type of user by expressing quality in largely descriptive terms, but might serve another type by providing parameters of a suitable error model, or even an embedded process that when run on the user's system would generate a sample of realizations of the error model. A third type of user might be best served by the creation of a suitable visualization of the effects of error.

From this perspective, it may be better to think of metadata not as the digital equivalent of the card catalog, but as a process of communication between producer and user, in a language understood by both. Since users can be expected to use a variety of languages, the successful producer must match them with a hierarchy of descriptions.

The card catalog metaphor fails in an additional respect which is likely to have profound implications for the geographic data community. In the traditional library, all content occurs within the covers of bound volumes, in convenient, discrete units. The cataloging system is also discrete, since there exists only a finite number of possible subjects, and authors, titles, and subjects can all be ordered along simple dimensions. However, this concept of information granularity fails to survive in the digital world. Data sets can be aggregated into larger units, for example by edgematching tiles to create a seamless view of the world, or by including maps and images in other, larger data sources such as CD encyclopedias. Data can also be usefully disaggregated, when the various digital layers of a simple topographic map are catalogued separately.

The granularity issue becomes more difficult again when discussed in the context of accuracy assessment. In some cases, accuracy may be an attribute of an entire data set, and thus compatible with the card catalog metaphor. In other cases, however, accuracy may be unique to one class of features in a given data set, or may vary by feature, or may be unique to specific subareas. In the latter case, metadata may be better modeled by analogy to the validity map that can be found on many topographic maps—metadata may be a map in its own right, rather than an attribute. In general, it is necessary to think in terms of a hierarchy of possible metadata descriptions of accuracy, ranging from the individual attribute or entity to the seamless database covering the entire Earth.

IMPLICATIONS

Thus far, the paper has raised a series of issues that arise when accuracy assessment is viewed as a task extending through the entire data life cycle, rather than a one-time process. This final section reviews the implications of this perspective, in three contexts: processes, hierarchical structures, and content.

Processes

From the data life cycle perspective, accuracy assessment is a form of metadata, and an essential part of the communication of information about a data set's fitness for use from one actor in the data life cycle to another, perhaps across substantial barriers of distance and discipline. The accuracy assessment of a given data set must change whenever its data model changes due to some transformation, or whenever other transformations are carried out which invalidate its current accuracy assessment. In some cases such updates can be automated, for example when scale or resolution changes; in other cases updates may require user intervention if there is no simple algorithm available.

Because the uses to which data may be put are likely to vary widely, accuracy assessments must be expressed in terms that are as comprehensive as possible. Not only is the producer or transformer of data under an obligation to describe whatever is known about accuracy that is likely to be of use to eventual users ("truth in labeling")—in addition, potential uses are likely to motivate or even force a more comprehensive assessment than might otherwise have been made. In particular, information on the spatial structure of errors is likely to be useful to a wide range of users, even though it may not be part of the customary apparatus of accuracy assessment.

In principle, the information provided as data move through the data life cycle should be sufficient to allow the user to estimate the uncertainty associated with any product of analysis or modeling. This will include information on the marginal distribution of error in each element of a data set, together with information on the joint distributions of errors at pairs of points. Hunter, Caetano, and Goodchild (1995) and others have described a general and robust strategy for evaluating the effects of error when propagated through GIS operations—the strategy consists of using a suitably calibrated error model to simulate (realize) a sample of possible and equally probable data sets; repeated analysis of each data set; and measurement of the variation across results. This Monte Carlo strategy is sufficiently robust and general to be applied across a full range of GIS operations, and more comprehensive than any analytic approach.

Besides suitable statistics included in metadata, it is desirable that information on data quality be passed between actors using techniques of visualization (Hearnshaw and Unwin, 1994), as the parameters of accuracy assessments and error models may be insufficiently informative to many users. The strategy of Hunter, Caetano, and Goodchild (1995) can be used in this sense also, if the sample of possible data sets is displayed in some appropriate fashion, either by animation or by display in multiple windows (Ehlschlaeger et al., 1997).

Finally, a comprehensive approach to data quality in the data life cycle must include facilities to report the effects of uncertainty in the form of confidence limits, and to update data quality information automatically as the data are passed through transformations and processes (Heuvelink, 1993). Full automation is not likely to be possible, at least in the near future, as the implications of many transformations are not well enough understood. Instead, data quality information may have to be reworked manually, or simulations may have to be performed to determine the necessary information.

Hierarchy

Although metadata has been seen primarily as the digital analog of the card catalog record, three reasons have been advanced for moving beyond this limited perspective. First, metadata must accommodate the varied experience, vocabulary, and skills of a range of actual and potential users, who range from sophisticated spatial statisticians, armed with geostatistical software, to those with very little familiarity with the deeper issues of geographic information. The statement "digital orthophoto quads have a scale of 1:12,000" may be meaningful to the spatially-aware professional, but may have to be significantly expanded to be meaningful to a grade 10 student. The process of accuracy assessment must lead to a corresponding range of levels of output if it is to be useful to a range of actors along the data life cycle.

Second, structures used to store the results of accuracy assessment in geographic databases must allow for a number of distinct levels of description, depending on the particular stratification of accuracy. It must be possible to describe uncertainty at the levels of attributes, features, feature classes, layers, data sets, and seamless coverages. This suggests that the extension of GIS data models to accommodate information on uncertainty may be far from straightforward. Until it is done, however, the transmission of quality information along the data life cycle will be seriously impeded.

Third, any move to measurement-based GIS requires that the lineage of data be stored explicitly. Instead of latitude/longitude coordinates, the position of a feature in a measurement-based GIS is determined by measurements with respect to some parent feature; absolute coordinates may appear only at the highest level in the hierarchy, which is often associated with geodetic control. Thus data models for measurement-based GIS must also be based in hierarchical structures.

Content

Finally, this paper has identified various changes that are needed in the content of GIS databases to support an explicit consideration of data quality throughout the data life cycle. First, measurement-based content is clearly more effective than coordinate-based content in allowing us to use standard methods of error analysis, and in dramatically reducing the costs associated with increases in positional accuracy. Second, a comprehensive approach must include not only statistics of the marginal (independent) distributions of data items, but also the joint distributions. Without such information it is impossible to determine the uncertainty associated with any GIS product that requires examination of data items taken more than one at a time—the list of such products includes slope and aspect estimation, determination of polygon area, and a host of other GIS operations, both basic and sophisticated. To support assessment of the accuracy of spatial interpolation, for example, access is needed in GIS data structures to variograms and correlograms, statistics for which there is currently no place in GIS databases.

CONCLUSIONS

This paper has proposed that accuracy assessment be placed in the broader context of the entire geographic data life cycle. Although there is still much work to be done in refining methods of assessment, and models of error in databases, the value of such work lies ultimately in the degree to which it affects the activities of users who may be very far removed from the originator of the data. Many different actors may be involved in the data life cycle, and many different transformations involving change of data model. At this time, we have only limited understanding of the effects of transformations on accuracy assessments, or the effects of accuracy on the products of GIS operations. On the other hand we do have general strategies which are capable of resolving these issues in robust ways.

The field of GIS is now moving toward such an integrated view of the data life cycle, through the development of search engines for the Web, digital libraries, metadata standards, and related activities. It is important that accuracy assessment be recognized at each stage in the life cycle, and that the results of accuracy assessment be available to all of the actors, in suitable forms.

Acknowledgment

The National Center for Geographic Information and Analysis and the Alexandria Digital Library project are supported by the National Science Foundation, NASA, and the Advanced Research Projects Agency.

References

Buttenfield, B.P., and R.B. McMaster, editors, *Map Generalization: Making rules for knowledge representation*, Longman Scientific and Technical, Harlow, UK, 1991.

Ehlschlaeger, C.R., A.M. Shortridge, and M.F. Goodchild, Visualizing spatial data uncertainty using animation, *Computers and Geosciences*, 23(4), pp. 387–395, 1997.

Goodchild, M.F., Geographical data modeling, *Computers and Geosciences*, 18(4), pp. 401–408, 1992.

Goodchild, M.F., *Report on a Workshop on Metadata held in Santa Barbara, California, November 8, 1995*, http://www.alexandria.ucsb.edu/public-documents/metadata/metadata_ws.html, 1995

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

Goodchild, M.F., and S. Yang, A hierarchical spatial data structure for global geographic information systems, *Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing*, 54(1), pp. 31–44, 1992.

Hearnshaw, H.M., and D.J. Unwin, editors, *Visualization in Geographical Information Systems*, Wiley, Chichester, UK, 1994.

Heuvelink, G.B.M., *Error Propagation in Quantitative Spatial Modelling Applications in geographical information systems*, Faculteit Ruimtelijke Wetenschappen, Universiteit Utrecht, Utrecht, 1993.

Hunter, G.J., M. Caetano, and M.F. Goodchild, A methodology for reporting uncertainty in spatial database products, *Journal of the Urban and Regional Information Systems Association*, 7(2), pp. 11–21, 1995.

Kevany, M., The role of field computers in GIS, Paper presented at the 1996 North Carolina Geographic Information Systems Conference, Winston-Salem, February 7–9, 1996.

Langran, G., *Time in Geographic Information Systems*, Taylor and Francis, New York, 1992.

Onsrud, H.J., and G. Rushton, editors, *Sharing Geographic Information*, Center for Urban Policy Research, New Brunswick, NJ, 1995.

Pickles, J., editor, *Ground Truth: The social implications of geographic information systems*, Guilford Press, New York, 1995.

Raper, J., editor, *Three-Dimensional Applications in Geographical Information Systems*, Taylor and Francis, New York, 1989.

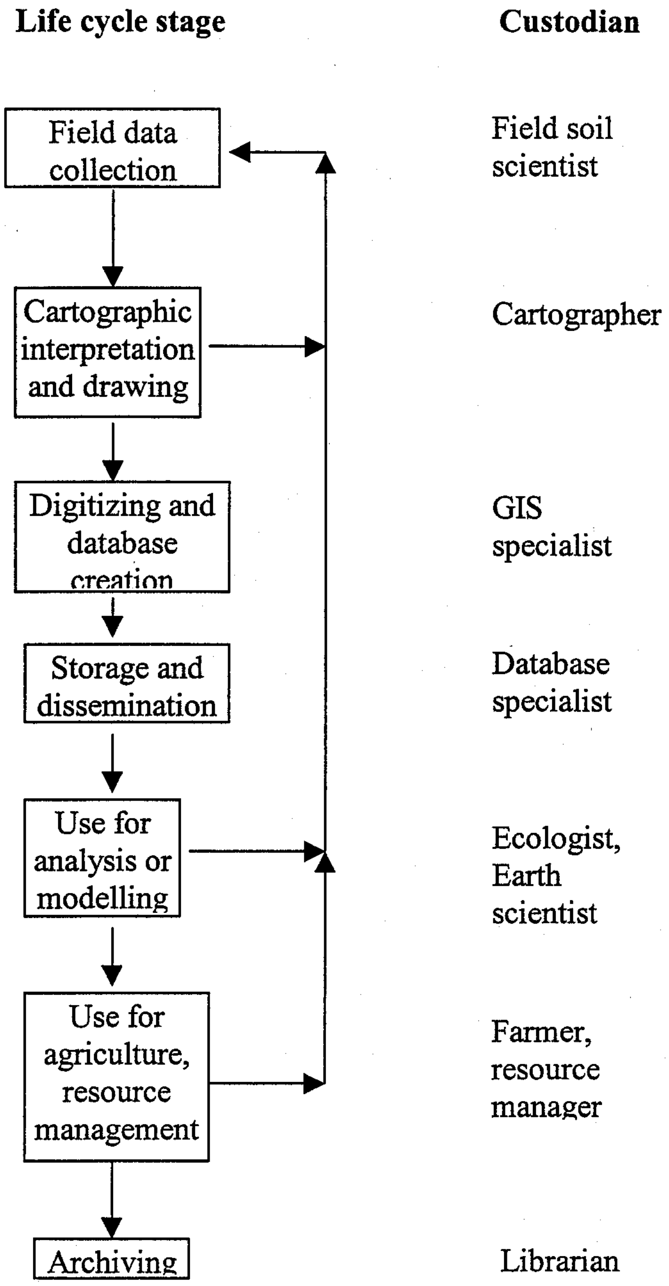
Smith, T.R., D. Andresen, L. Carver, R. Dolin, and others, A digital library for geographically referenced materials, *Computer*, 29(7), p. 54, 1996.

Tsichritzis, D.C. and F.H. Lochovsky, *Data Models*, Prentice-Hall, Englewood Cliffs, N.J., 1982.

Turner, A.K., editor, *Three-Dimensional Modeling with Geoscientific Information Systems*, Kluwer Academic Publishers, Dordrecht, 1992.

Vandenheuvel, R.M., The promise of precision agriculture, *Journal of Soil and Water Conservation*, 51(1), pp. 38–40, 1996.

Figure 1: The life cycle of a soil database: an example of the complex patterns of custodianship, transfer, and feedback now common for many types of spatial data.


Life cycle stage
Custodian
Field data collection
Field soil scientist
Cartographic interpretation and drawing
Cartographer
Digitizing and database creation
GIS specialist
Storage and dissemination
Database specialist
Use for analysis or modelling
Ecologist, Earth scientist
Use for agriculture, resource management
Farmer, resource manager
Archiving
Librarian