
The Lineage of Data in Land and Geographic Information Systems (LIS/GIS)

Richard K. Grady

Richard K. Grady received his B.S., *cum laude*, in resource economics from the University of Massachusetts at Amherst. He began his career as a planner and cartographer with the Commonwealth of Massachusetts. He is currently executive manager for federal systems marketing at Intergraph in Reston, Virginia. He is interested in all aspects of LIS/GIS technology and its implementation, but particularly data quality.

Abstract: *The need to integrate data from many sources, often at different scales and projections with different contents and areas of coverage, is typical of building an LIS/GIS database. Source materials tend to be heterogenous. Applications of the data also tend to be heterogenous; and the data are likely to be rescaled, resymbolized, and reprojected to serve some specific purpose. This reprocessing of data often results in an output which feeds another process.*

The importance of recording the lineage of data becomes obvious when one considers the machinations mentioned above. To ascertain the validity and suitability of data to support a particular purpose, lineage information is needed by the user. Data producers must be responsive to this need. This paper will review what data lineage is and discuss ways in which it relates to the LIS/GIS community.

The National Committee for Digital Cartographic Data Standards (NCDCCDS), in particular, Working Group II on Data Set Quality, defined the lineage of data as follows:

The basis of any quality report is a narrative of the lineage of the data. Lineage includes the original source material and all the processes and transformations leading to the final product (digital database). This information is required for a user to evaluate fitness, and it is required by a producer to maintain and update the data. (1985, p. 115)

As the user of data accepts responsibility for determining fitness for use, the producer must accept responsibility for providing adequate information about data sources, and processing history. The simple statement that "This map complies with National Map Accuracy Standards" (as suggested by U.S.

Bureau of Budget in its 1947 document on U.S. National Map Accuracy Standards) no longer suffices; few would argue with this.

The producer of data should clearly label the product with information about its source and processing history; however, it is the user who ultimately must determine its fitness for a particular use. While "truth in labeling" should be demanded, it is too soon in the evolution of LIS/GIS databases to abandon the old axiom: "Let the buyer beware." In practice, neither producers nor users are blameless. Generally, professional ethics have prevailed on behalf of reasonable practices, but not always. Sometimes there is blind faith in the source of data; other times, there is acceptance or tolerance of error due to expediency.

Inevitably, there will be errors in measurement, processing, and representation of data in an LIS/GIS which result in uncertainty. We are faced with two basic choices for managing this uncertainty. They are: (1) uncertainty reduction, and (2) uncertainty absorption (Bedard 1987).

Techniques for reducing uncertainty include use of standards, testing for accuracy and consistency, and understanding the lineage of data. The absorption of uncertainty has to do with who pays if errors in data cause damages (Bedard 1987).

Clearly, data lineage is not all that a user needs to determine the data's fitness for a particular use. However, lineage is an important part of reducing the uncertainty associated with using specific data, and it can help a user decide if it is worth absorbing the risk of potential damages. As specified in the proposed

Spatial Data Transfer Specification (which has been submitted to the National Institute of Standards and Technology by U.S. Geological Survey), lineage is one of five sections that should be included in a Data Quality Report (DQR).

Data Quality Report

Traditionally, data producers may have filed a DQR as part of the external documentation associated with a project, perhaps. Some have even provided visual overlays in an attempt to diagram data quality and reliability. More often, it is neglected or not considered essential. The proposed SDTS breaks the DQR into the following sections (USGS 1988):

- Lineage
- Positional Accuracy
- Attribute Accuracy
- Logical Consistency
- Completeness

A description of methods for reporting on each of the above-mentioned sections of a DQR is beyond the scope of this paper. However, the importance of providing and monitoring this information should be recognized; yet, it is often overlooked as part of the operational considerations of an LIS/GIS project. Nyerges (1987) commented on this problem:

Many times databases are not useful beyond the original scope of a project because of the lack of documentation. Data are less useful as information when data describing database production conditions are forgotten and/or misplaced. (p. 320)

Typically it is true that the compilation and management of data for a DQR is of secondary interest to the main purpose of producing and/or exploiting an LIS/GIS database. Anything to make the task easier would improve the likelihood of getting it done. One area of promise is the utilization of on-line metadata.

Metadata, which is simply "data about other data," can be managed on-line as part of an operational LIS/GIS. Both the U.S. Forest Service and Bureau of Land Management are looking at making metadata a part of their information environments, as well as the U.S. Geological Survey and National Ocean Survey, among others. Once a format is established for the DQR, it can be populated with data as an extension to the database. A data server can store and disseminate the DQR to facilitate evaluation of a dataset's fitness for a particular purpose.

Data Lineage and the Audit Trail

There are a number of parallels between LIS/GIS and MIS (Management Information Systems). In fact, there has been considerable convergence in some areas. One area that could benefit from more convergence is database auditing. Audits, as well as checking for the absence or presence of lineage information, should become part of the lineage record.

Transactions that change the database occur on a regular basis; this applies to both LIS/GIS and MIS. In a typical MIS environment, once a transaction is approved, it is entered into the database and is also recorded in

a transaction history file. This file serves as an audit trail. An audit trail is defined as follows:

... the presence of data processing media and procedures which allow a transaction to be traced through all stages of processing, beginning with its appearance on a source document and ending with its transformation into information on a final output document. (O'Brien 1979, p. 438)

In business management, the certified public accountant (CPA) lends credibility to financial reports through an audit function. He or she is paid for this service, but remains independent of the firm that is audited. The firm's transactions are examined, including the collection, classification, and presentation of financial data. This is done based on established professional standards, to verify that the financial reports of the firm adhere to "generally accepted accounting principles." This function helps to protect the public from misleading financial data. Many investors are fully capable of misleading themselves, without bad data. The same holds true in LIS/GIS, but there is no equivalent, generally speaking, to the CPA.

In staffing an LIS/GIS, the role of the auditor, to attest to data integrity, should not be overlooked. This particularly applies where very large databases and many users are involved. Many corporations with an MIS have a job function defined for a database administrator. This role is similar in some respects to the CPA, except it is internal to the company. If data producers do

not support this function internally, then in some cases the user might provide it (such as a utility company, to verify data from a conversion vendor).

One responsibility of the database administrator should be to define a data dictionary which can apply rules to transactions before any new data are entered into the database, preventing updates that would conflict with existing data. This front-end checking is the traditional means of protecting logical consistency. Another means to check logical consistency is deferred checking, in which the database is scanned after entries are made, checking for compliance with rules as stated in the data dictionary. Although the lineage of the data cannot easily be tested for validity, the above-mentioned techniques should be used for at least checking on the absence or presence of lineage records.

The argument against front-end checking is that it may slow down the data-collection process, which is already the biggest hurdle to timely LIS/GIS implementation. Front-end checking may be a hindrance at times, but it is the best way to assure that data lineage is carried into the digital domain. It is probably too harsh to reject all data that do not have suitable lineage, but such data must at least be flagged.

Deferred checking can supplement input checks. One advantage is that it checks the contents of the database, not the input process. Therefore, it is not a hindrance to data collection. When revision or auditing occurs on an LIS/GIS database, deferred checking can be a means of flagging data that do not have

associated lineage information.

These techniques, as basic as they are to MIS environments, have not been widely utilized in LIS/GIS environments, particularly where the lineage of graphic data is concerned. One positive example of long-standing practices of recording chart history is the National Ocean Survey, which is now fully automating these practices. However, more organizations need to institutionalize the use of periodic database surveys and/or continuous administrative processes for checking LIS/GIS databases.

Societal Mandate as Part of Data Lineage

A relatively unexplored area of data lineage is societal mandate. It is the mandate, issued by societal bodies, that enables organizations to build institutions to accomplish goals. As Chrisman (1987) pointed out, many institutions have been created pursuant to various mandates to collect data for some purpose:

The important data collection functions of society are not carried out for technical reasons. The creation of property maps, zoning maps, and all the other municipal functions are not driven by a benefit/cost ratio. Each record is collected and maintained in response to a social need as expressed by the legal and political system. The search should not be for the flow of data, but for the mandates that cause the flow.
(p. 1369)

This is an important point for users of LIS/GIS technology. The producer of data should in-

clude as part of its lineage the mandate, (or enabling legislation, where appropriate), that led to the collection effort. This information would be very useful to the user of such data in determining its fitness for use.

Cartography and Data Lineage

Anyone who is properly trained in cartography is fastidious about source materials. Decisions must be made about map contents, based on answers to the following questions:

- What is the name and nature of the authority responsible for the source materials?
- How familiar is the authority with the area of coverage?
- What is the date of the source, how and when was it originally compiled, and has it been revised?
- Was it derived from another source and, if so, is that source larger-scale?
- For what purpose was the data collected?
- What is the accuracy?
- What generalization and/or transformation procedures were used, if any?

These questions are representative of what must be determined by the cartographer who collects data. Over the past 25 years, there has been a proliferation of special-purpose mapping by non-cartographers. In part, this proliferation has been due to the advent of automated LIS/GIS technology, where a map is often just an output to show the results of a particular analysis. Such maps are often used themselves as input to another analytical process. When this is done without regard to the

fitness of the data, error propagation is highly probable.

Applications of LIS/GIS technology have increased the need for multi-discipline expertise. Cartography is fast becoming a set of communication tools and skills adopted by several disciplines outside of traditional map-making. While they are adopting some of the tools and skills, they are not necessarily adopting the principles.

One of the most basic cartographic principles is: "Thou shalt not derive a large-scale map from small-scale source materials." As we have moved into the digital domain, with many non-cartographers making maps to represent the results of spatial analysis in LIS/GIS, there is often blindness and/or ignorance related to this simple principle.

Clearly, the need for someone to make decisions about sources, content, and design of digital databases has intensified with the increased adoption of LIS/GIS technology. This role does not have to be played by a cartographer, but certainly by someone with the professional discipline to be fastidious about data lineage.

The Spatial Reference Component of Data Lineage

Most of the discussion so far has revolved around source materials as a component of data lineage. The other major component relates to "the processes and transformations that lead to the final product (digital database)." The processing history becomes particularly relevant with the transition from manual to automated LIS/GIS technology.

In the proposed Spatial Data Transfer Specifications (SDTS), spatial reference is divided into three modules:

- (1) Internal Spatial Reference
- (2) External Spatial Reference
- (3) Spatial Domain

The internal spatial reference module describes the translation and scaling parameters necessary to transform the internal coordinate system of the data into an external coordinate system. The external spatial reference module defines the geographic coordinate coding (spatial address) of objects, tied into a geodetic reference system. The spatial domain module specifies the geographic area of coverage by defining a series of spatial addresses that delineate the area (USGS 1988).

It is essential to tie spatial reference into ground control to make the coordinate system explicit. Many have called for better documentation of ground control, and for survey monuments to be included in digital coverage whenever possible. Not only does this allow for accuracy to be tested, but it allows for data from different sources to be registered in combination. Merging coverages is a common requirement, and problem, in LIS/GIS environments.

There is an ongoing debate over the application of geographic coding schemes that use coordinate locations alone to distinguish points. The problem arises when a given point has different coordinate locations on separate source materials. One approach to this problem is to distinguish between measuring a

point's location and naming it as a place. Dutton (1984) observed the following:

If survey monuments and other significant locations had systematically constructed geocodes [names], the problem of registering coverages would become that of identifying their common control points by name-matching, then fitting a transformation to common coordinates. (p. 278)

The spatial reference component of data lineage, including well-documented ground control and other significant point locations, can help ascertain data quality and reduce uncertainty.

The Temporal Aspect of Data Lineage

When it comes to data lineage, several "dates" are of obvious importance. These include: the compilation date of source materials; the publication date of source materials; the dates of original survey; and the date of revisions. Of course, all of these dates relate to single points in time. And yet, the phenomenon we try to model in an LIS/GIS is not static; it changes, continually.

The data structures being used today, which were designed to represent static phenomena, have limitations for modeling the time continuum. Chrisman and Langran (1988) have proposed a framework for handling temporal data, to facilitate queries such as:

- What was the previous state, or version, of this object?
- What has changed (during a period, or at a place)?

- What is the periodicity of change?
- What trends are evident?

Until the time dimension can be represented structurally, LIS/GIS can be used to store temporal attributes for each feature, thereby allowing some limited analysis based on change over time (Chrisman and Langran 1988).

There is great interest in change detection for many applications of LIS/GIS technology. One particularly useful technique is overlaying two images of the same area, taken at different times. This snapshot technique can be used to observe change, delineate it, and update vector map data. Obviously, the date associated with images and vector maps becomes very relevant to applying this technique. It is just one reason why the temporal aspect of data lineage is so important.

Summary

This paper has attempted to help build an understanding of

what lineage is comprised of, discuss some techniques for incorporating lineage into LIS/GIS, and raise some concerns about the risks of not having it. The inclusion of data lineage in LIS/GIS requires a conscious effort and application of sound procedures and principles. The lineage of data, including source materials and processing history, is an important part of the Data Quality Report (DQR) as defined in the proposed Spatial Data Transfer Specification (SDTS). Data producers must be fastidious about including lineage information, and it can be stored and disseminated as metadata—data about data. Users of data must be careful in determining the data's fitness for a particular purpose, and aware of the risk they absorb when they apply data without knowing their lineage. An important part of lineage is "when" data were acquired; but also important is "why" it was created to begin with.

References

- Bedard, Yvan. 1987. "Uncertainties in Land Information Systems." *Auto-Carto 8 Proceedings*, Baltimore, MD. pp. 175-183.
- Chrisman, N.R. and Langran, G. 1988. "A Framework for Temporal Geographic Information." *Cartographica*, No. 25, Vol. 4.
- Chrisman, N.R. 1987. "Design of Geographic Information Systems Based on Social and Cultural Goals." *Photogrammetric Engineering and Remote Sensing*, Vol. 53, No. 10, p. 1369.
- Dutton, G. 1984. "Truth and Its Consequences in Digital Cartography." *Technical Papers of the 44th Annual Meeting of the ACSM*, Washington, D.C. p. 278.
- National Committee for Digital Cartographic Data Standards. 1985. "Digital Cartographic Data Standards: An Interim Proposed Standard." *Issues in Digital Cartographic Data Standards, Report #6*, H. Moellerling, ed., p. 115.
- Nyerges, T. 1987. "GIS Research Needs Identified During a Cartographic Standards Process: Spatial Data Exchange." *International Geographic Information Systems: The Research Agenda (Proceedings)*, Arlington, VA, Vol. I, pp. 319-330.
- O'Brien, J. 1979. *Computers in Business Management*, Richard Irwin, Inc., Homewood, IL. p. 438.
- U.S. Geological Survey. 1988. "The Proposed Standard for Digital Cartographic Data." *The American Cartographer*, Vol. 15, No. 1, pp. 131-132. (An updated version of the proposed standard was made available in July 1990 by USGS, National Mapping Division, Office of Technical Management.)