ISSN 0254–2757

QUÆSTIONES INFORMATICÆ

Volume 6 • Number 2

September 1988

The official journal of the Computer Society of South Africa and of the South African Institute of Computer Scientists

Die amptelike vaktydskrif van die Rekenaarvereniging van Suid-Afrika en van die Suid-Afrikaanse Instituut van Rekenaarwetenskaplikes

QUÆSTIONES INFORMATICÆ

The official journal of the Computer Society of South Africa and of the South African Institute of Computer Scientists

Die amptelike vaktydskrif van die Rekenaarvereniging van Suid-Afrika en van die Suid-Afrikaanse Instituut van Rekenaarwetenskaplikes

Subscriptions

The annual subscription is

	SA	US	UK
Individuals	R20	$ 7	£ 5
Institutions	R30	$14	£10

to be sent to:
Computer Society of South Africa
Box 1714 Halfway House 1685

Quæstiones Informaticæ is prepared by the Computer Science Department of the University of the Witwatersrand and printed by Printed Matter, for the Computer Society of South Africa and the South African Institute of Computer Scientists.

A Data Structure for Exchanging Geographical Information

A K Cooper
Centre for Advanced Computing and Decision Support, CSIR, P O Box 395, Pretoria, 0001

Abstract

Geographical information consists of non-spatial information (alphanumeric) and spatial information (vector and raster), the relationships between the non-spatial information and the spatial information, as well as the spatial relationships inherent in the spatial information, known as topology. It is undesirable for any exchange standard to lose, reduce or alter any information exchanged through the standard. For this reason, current alphanumeric and graphic exchange standards are insufficient for geographical information.

The project team drawing up a proposed South African standard for the exchange of geographical information has studied the proposals and standards of other countries and has held discussions with the users and potential users of computerised geographical information in this country. The project team feels that the best model for use with the data structures of the exchange standard is the relational one. This paper describes the nature of geographical information and the advantages of the relational model.

Keywords: *Geographical Information, Exchange Standard, Topology, Feature, Attribute, Relational Model.*

Received July 1987, Accepted July 1988

1. Introduction

Geographical information consists of all information that refers to the man-environment system and that can be localised in space and time. Computerisation of geographical information started with the statistical processing of thematic data and with automatic cartography. This has now developed to Geographical Information Systems (GIS). A GIS consists of a data base of spatially (geographically) referenced data and a collection of utilities for efficiently inputting, storing, retrieving, maintaining, manipulating, analysing and displaying the data.

A true GIS is a system that is orientated to the analysis of geographically referenced data to produce useful information [3]. To do this, the GIS data base must cater for the different types of geographical information (vector, raster and alphanumeric) and provide a topological structure for the spatial data – that is, a structure that explicitly encodes the spatial relationships inherent in the data.

A major problem with geographical information, and one that has delayed its computerisation, is the enormous volumes involved. Donald Light of the US Geological Survey (USGS) estimates to digitise the approximately 54 000 7.5-minute standard topographical quadrangle maps (at a scale of 1:24 000) that cover the lower 49 American states, the USGS will require a storage capacity of about 11 terabytes [10]. The scale of 1:24 000 would be acceptably large for most rural applications, but for urban applications, scales of 1:1000 or even larger become necessary. The enormous volumes are not restricted to topographical data – consider the amount of data captured in a population census.

It is not feasible to expect users of geographical information to capture all their own raw data because the time and costs involved are prohibitive. Users must be able to acquire data from other users. There are a number of different suppliers whose hardware and software are used for processing geographical information. Unfortunately, very few are compatible with each other to the extent that they can exchange geographical information. Generally, users have a great deal of freedom in structuring their systems, and consequently, there could even be significant problems in exchanging information between systems from the same supplier. Thus there is a need for a software- and hardware-independent exchange standard for geographical information. The data structures of the exchange standard should be a superset of the structures of the systems that might use the exchange – they must not lose, reduce or alter the information or its topology.

In April 1986, a joint project between the National Research Institute for Mathematical Sciences (NRIMS) of the CSIR and the Chief Directorate of Surveys and Mapping (CDSM) was initiated to draw up a South African exchange standard for geographical information. The project team has studied the standards and proposals of a number of countries, including Australia [16], the United States [11,12,13,17,18] and the United Kingdom [8,9], as well as proposals submitted to the International Hydrographic Organisation, [1,14,20]. The project team distributed a questionnaire on GIS to over 130 organisations in South Africa and has also organised a number of workshops on GIS. This has enabled the GIS community to provide the project team with valuable inputs concerning the proposed exchange standard. A draft exchange standard was released in March 1987 for public comment [4], and in September 1987 the final report will be presented to the National Programme for Remote Sensing, the project's sponsors.

2. The Contents of a Geographical Information System

The following definitions are from [3,4,7].

A simple **feature** is a set of one or more uniquely identifiable objects in the real world where the defined characteristics of the objects are consistent throughout all the objects.

An **attribute** is a defined characteristic of any feature about which information is stored.

Non-spatial attributes are attributes that are independent of position – often termed the descriptive data of the feature.

A **spatial attribute** is an attribute whose value is a subset of any n-dimensional space. Most current GIS's work in two dimensions, and some work in three dimensions. There is no reason why future GIS's might not work in more dimensions. The three fundamental types of two-dimensional vector spatial attributes are nodes, chains and regions. The fundamental raster spatial attribute is the matrix.

A **node** is a 0-dimensional object with an n-tuple of coordinates specifying its position in n-dimensional space.

A **chain** is an ordered sequence of n-tuples of coordinates with a node at each end. The direction of the chain is defined when the chain is used in a specific feature.

An **arc** is any continuous part of the circumference of a circle with a node at each end. The arc is defined by giving the start and end nodes of the arc and either the centre of the circle or any other point on the circumference of the arc.

A **region** is the interior of a continuous and closed sequence of one or more chains, known as the region's outer boundary.

Examples of vector spatial attributes are shown in Figure 1.

A **matrix** consists of an n-tuple of coordinates (its origin) and an m-dimensional tesselation of data values encoded in a pre-defined format.

Three-dimensional vector spatial attributes are not yet well understood, and no exchange standard has attempted to cater for them. [7] refers to **solids** as the fundamental three-dimensional vector spatial attribute. As and when solids are formally defined and used in geographical information, they can easily be added to the proposed exchange standard. The advantage of the relational model for the exchange standard is that to add new types of data, all one has to do is formally define, and add to the standard, the relationships that should exist between the new data types and existing data types.

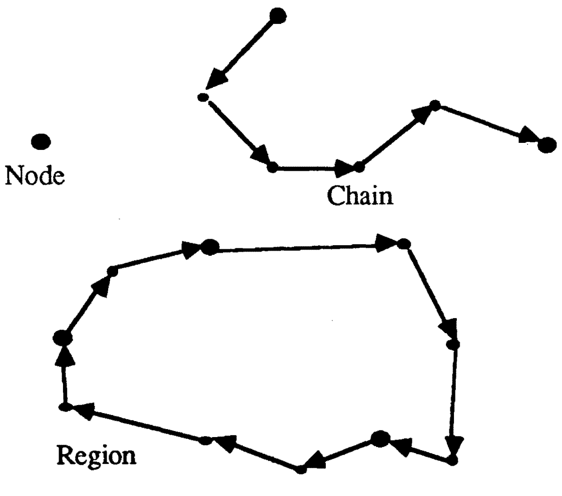

Figure 1

In a geographical data base, features are grouped on the basis of the nature of their spatial attributes. The three types of vector features are point, line and area, and the raster type of a feature is a grid.

A **point** feature is one whose position is described by a node.

A **line** feature is one whose position is described by a set of chains and/or arcs. These chains and arcs do not necessarily form a continuous object.

An **area** feature is one whose position is described by a set of regions. These regions do not necessarily form a continuous object.

A **grid** feature is one whose position is described by a set of matrices. These matrices do not necessarily form a continuous object.

Owing to the complex nature of geographical information, it is desirable to be able to combine a number of different features to form one feature. The proposed exchange standard allows **compound** features, which are collections of other features.

Classification is the arrangement of features into classes or groups and must be done on the basis of the qualitative characteristics of the objects, such as their function, and not on the basis of their quantitative characteristics.

There is a grey area between the classification and the non-spatial attributes of a feature, and different criteria apply for different users [5]. In fact, classification itself could be considered to be a non-spatial attribute.

For example, a tall office building could be classified as a building, as an area zoned for commercial use, or even as an aeronautical obstruction. Whichever classification is used, the other characteristics of the feature could be recorded as non-spatial attributes of the feature.

A feature's classification should be done on the basis of its characteristics that are least likely to change. In our example, it would be better to classify the feature as a building because if the building were destroyed then the feature would be as well, while the building could quite easily be rezoned. Conceptually, there is no reason why an object or event in the real world could not be represented by more than one feature, especially if the classifications of the different features are not closely connected. However, this could lead to problems such as redundant data and is not recommended. For details on how to set up a classification see [15].

3. Topology

Four types of topology are incidence, containment, intersection and adjacency.

Incidence means that more than one feature shares the same spatial attributes. If this topology is not explicitly captured with the spatial data, it is difficult to generate. In older GIS's, the same spatial data were often captured more than once. For example, the common boundary of two neighbouring regions would be captured twice, once for each region. As these data would be captured by hand digitisation, the two versions of the same boundary would not necessarily be the same. This would result in slivers in the data base – narrow areas of overlap (the intersection of two regions in the data base that do not intersect in the real world) and underlap (the area between two regions in the data base that in the real world are adjacent). The solution to the problem of slivers is to share spatial data so that when data are captured, the spatial data that are already captured can be used.

Incidence topology includes point features sharing the same node, line features and area boundaries sharing the same chains, area features sharing the same regions and grid features sharing the same matrices. An example follows.

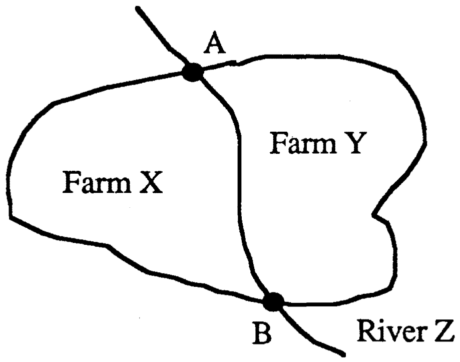

Figure 2

Thus, the chain between nodes A and B is shared by the line feature River Z and the boundaries of the area features Farm X and Farm Y.

Containment means that a spatial attribute lies entirely within the outer boundary of a region. The two containment topologies are **exclusion** and **inclusion**. If a region is completely contained within another region and does not form a part of the containing region, then the outer boundary of the contained region is an inner boundary of the containing region. We say that the containing region excludes the contained region. A region may exclude zero or more other regions which are within its interior. Inclusion means that the contained spatial attribute forms a part of the containing region. Inclusion can be determined from the coordinates of the spatial attributes. However, a user might wish to encode the relationship explicitly since transformations between different projections can alter the apparent relationships. Encoded inclusion relationships can speed up access to the data. Owing to the number of these relationships normally present in the data, it is unlikely that a user will explicitly encode all the inclusion relationships.

Intersection means that there is one coordinate common to more than one spatial attribute. In topologically structured data, one would expect a node at that point and that all chains passing through that point would be broken at that point into two chains, each of which has its start or end node at that point. To determine the intersections of line features, one would determine which of their constituent chains had terminal nodes in common.

Adjacency of regions can be determined through the chains which form boundaries common to more than one region.

Thus, topological connections on spatial attributes exist between a chain and its terminal nodes, a region and the chains that form its outer and inner boundaries, and a region and its excluded regions and included nodes, chains, regions and matrices.

In addition, between the features and the spatial attributes are the relations that take care of networks of chains that form a line feature, disjoint regions that form an area feature and different matrices that form a grid feature.

In the next example, chains A, B, C, D and E together form River X.

Topological information is implicit in the real world and is explicitly captured to reduce data redundancy, enhance data consistency and completeness, and speed up access to the data. The topology of geographical information is not yet well understood, and while the proposed exchange standard [4] has attempted to cater for topology that is known and used, it might be necessary to revise the topological relationships catered for in the standard as the use of topology in geographical information matures.

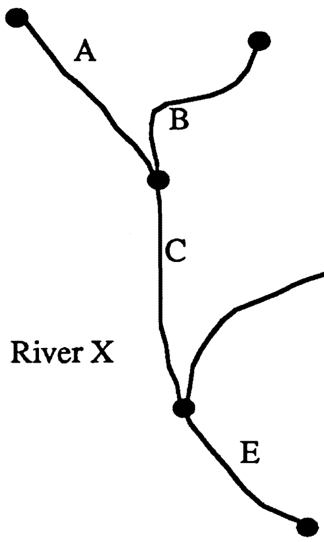

Figure 3

4. Advantages of a Relational Structure

There are three common models for data structures, namely the hierarchical, the network and the relational. Current exchange standards use hierarchical (for example, the Australian standard [16]), network or hybrid models (for example, the British standard [9] uses a combination of network and relational models) for their data structures. Two major disadvantages of these models are first, that users have to enter data into a number of fields that they do not wish to use, and secondly, that the structures are rigid and can only be changed with difficulty to cater for new types of data, such as the third dimension.

With a relational structure, users merely omit those relations for which they have no data, and it is almost trivial to add new relations to an existing exchange standard. In fact, data that can be exchanged through a relational exchange structure can always be exchanged through the structure, no matter how many new relations are added to cater for new types of data.

Normalisation of geographical data in relations minimises the redundancy and removes anomalies in updating the data [19]. All the relations in the proposed exchange standard [4] are in at least the second normal form. Once the contents have been finalised, further normalisation will be investigated, bearing in mind the danger that over normalisation of geographical information could slow down the retrieval of the information [19].

The project team drawing up the proposed exchange standard has studied the hierarchical, network and relational data models in the context of geographical information and believes that the relational model is the only one that provides the power and flexibility required.

The following are a few examples of the relations provided in the exchange standard. The relations are given in the boxes with rounded corners, and the attributes of the relations are given in the boxes with square corners. The keys of the relations are given above the relations.

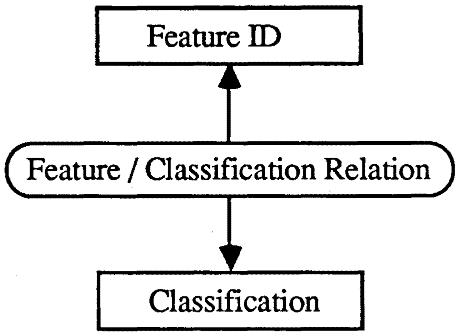

Figure 4

This is a one-to-many relation – each feature may have only one classification, while there may be any number of features with the same classification.

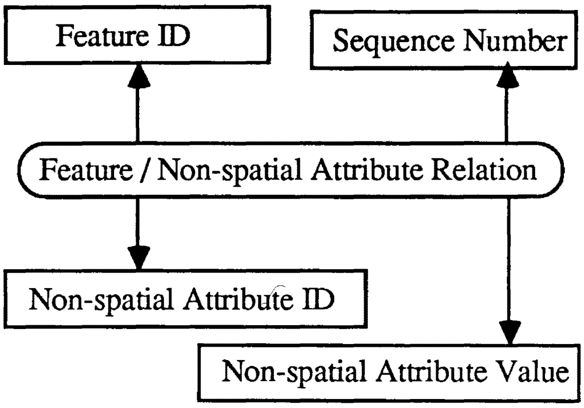

Figure 5

Because a feature can have more than one non-spatial attribute, a sequence number has to be added to the key to ensure that the key is unique.

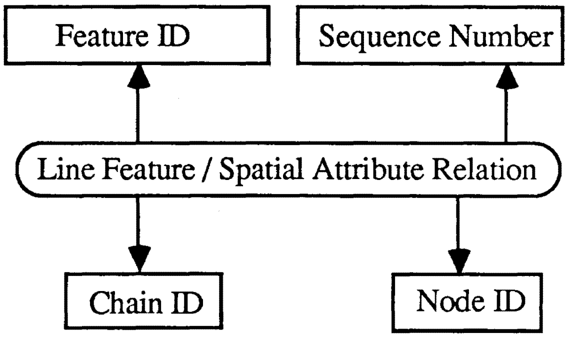

Figure 6

A line feature consists of a set of chains. In the relation, the node ID indicates which of the chain's two terminal nodes is the start node. The relationships in geographical information are complex, and it is not possible to define all the relations so that they might easily be inverted. For example, the inverse of the previous relation is the following (note that the node ID is not in the relation).

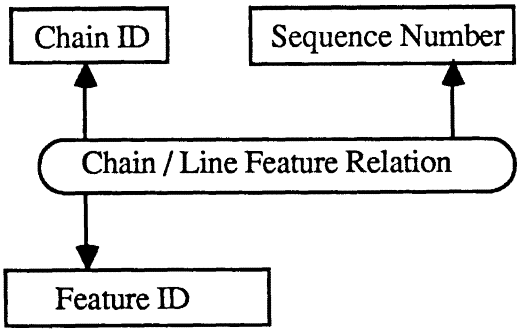

Figure 7

5. Conclusion

To be useful, an exchange standard should be a superset of any system that might use the exchange. It should also be expandible to cater for any advances that might take place. Thus, while the standard should cater for the topology required by GIS's, a user should be able to use it in a stripped down form if he does not require any topology.

Acknowledgements

I should like to thank the other members of the project team, Derek Clarke, Hester van Rooyen and Elri Liebenberg, as well as my colleagues at NRIMS and the GIS community in South Africa for all their help. I should like to thank the anonymous referees for their helpful comments.

References

[1] Canadian Hydrographic Service, [1985], Evolving Communications Standards in the Mapping and Charting World: a Report and a Proposal, submitted to the IHO's Committee on the Exchange of Digital Data.

[2] D.G. Clarke, A.K. Cooper, E.C. Liebenberg, M.H. van Rooyen, [1986], Questionnaire on GIS Exchange Format Proposals, National Standards for GIS Exchange, May 1986.

[3] D.G. Clarke, A.K. Cooper, E.C. Liebenberg, M.H. van Rooyen, [1986], Glossary of GIS terms and English/Afrikaans Translations, National Standards for GIS Exchange, May 1986.

[4] D.G. Clarke, A.K. Cooper, E.C. Liebenberg, M.H. van Rooyen, [1987], A proposed national standard for the exchange of digital geo-referenced information: Draft, NRIMS CSIR Technical Report TWISK 517, March 1987.

[5] A.K. Cooper, [1986], Report on the GIS project planning discussions, NRIMS CSIR Internal Report I668, January 1986.

[6] A.K. Cooper, [1986], An Introduction to Geographical Information Systems, NRIMS CSIR Internal Report I718, September 1986.

[7] A.K. Cooper, [1986], Thoughts on exchanging geographical information, Proceedings 1987 ASPRS-ACSM Annual Convention, **5**, 1-9, December 1986.

[8] P. Haywood, chair., [1986], The National Transfer Format Final Draft (Issue 1.3), Ordnance Survey, September 1986.

[9] P. Haywood, chair., [1987], The National Transfer Format Release 1.0, Ordnance Survey, January 1987.

[10] D.L. Light, [1986], Mass Storage Estimates for the Digital Mapping Era, *Photogrammetric Engineering and Remote Sensing*, **52** (3), 419-425, March 1986.

[11] H. Moellering, ed., [1985], Digital Cartographic Data Standards: An Interim Proposed Standard, Report No 6, National Committee for Digital Cartographic Data Standards, January 1985.

[12] H. Moellering, ed., [1986], Digital Cartographic Data Standards: A Report on Evaluation and Empirical Testing, Report No 7, National Committee for Digital Cartographic Data Standards, April 1986.

[13] H. Moellering, ed., [1987], A Draft Proposed Standard for Digital Cartographic Data, Report No 8, National Committee for Digital Cartographic Data Standards, January 1987.

[14] North American Work Group Committee on the Exchange of Digital Data, [1985], Proposed Format for the Exchange of Digital Hydrographic Data, submitted to the IHO's committee on the Exchange of Digital Data.

[15] C.F. Scheepers, W.R. Van Biljon, A.K. Cooper, [1986], Guidelines to Set Up a Classification for Geographical Information, NRIMS CSIR Internal Report I723, September 1986.

[16] Standards Association of Australia, [1981], Interchange of Feature Coded Digital Mapping Data, Australian standard 2482-1981, August 1981.

[17] G. TeSelle, chair., [1986], Federal Geographic Exchange Format – A Uniform Format for the Exchange of Spatial Data Among Federal Agencies, Federal Interagency Coordinating Committee on Digital Cartography, August 1986.

[18] U S Geological Survey, [1984], USGS Digital Cartographic Data Standards, Geological Survey Circulars 895-A to 895-G.

[19] J.W. Van Roessel, [1987], Design of a spatial data structure using the relational normal forms, *International Journal of Geographical Information Systems*, **1** (1), 33-50.

[20] J.A. Yeager, chair., [1986], Format for the Exchange of Digital Hydrographic Data, Committee on the Exchange of Digital Data, November 1986.

NOTES FOR CONTRIBUTORS

The purpose of the journal will be to publish original papers in any field of computing. Papers submitted may be research articles, review articles and exploratory articles of general interest to readers of the journal. The preferred languages of the journal will be the congress languages of IFIP although papers in other languages will not be precluded.

Manuscripts should be submitted in triplicate to:

Professor J M Bishop
Department of Computer Science
University of the Witwatersrand
Johannesburg
Wits
2050

Form of manuscript

Manuscripts should be in double-space typing on one side only of sheets of A4 size with wide margins.

The first page should include the article title (which should be brief), the author's name and affiliation and address. Each paper must be accompanied by an abstract less than 200 words which will be printed at the beginning of the paper, together with an appropriate key word list and a list of relevant Computing Review categories.

Manuscripts may be provided on disc using any Apple Macintosh package or in ASCII format.

For authors wishing to provide camera-ready copy, a page specification is freely available on request from the Editor.

Tables and figures

Tables and figures should not be included in the text, although tables and figures should be referred to in the printed text. Tables should be typed on separate sheets and should be numbered consecutively and titled.

Figures should also be supplied on separate sheets, and each should be clearly identified on the back in pencil and the author's name and figure number. Original line drawings (not photocopies) should be submitted and should include all the relevant details. Photographs used as illustrations should be avoided if possible. If this cannot be avoided, glossy bromide prints are required.

Symbols

Mathematical and other symbols may be either handwritten or typewritten. Greek letters and unusual symbols should be identified in the margin. Distinction should be made between capital and lower case letters; between the letter O and zero; between the letter I, the number one and prime; between K and kappa.

References

References should be listed at the end of the manuscript in alphabetic order of the author's name, and cited in the text in square brackets. Journal references should be arranged thus:

[1] E. Ashcroft and Z. Manna, [1972], The Translation of 'GOTO' Programs to 'WHILE' programs, *Proceedings of IFIP Congress 71*, North-Holland, Amsterdam, 250-255.

[2] C. Bohm and G. Jacopini, [1966], Flow Diagrams, Turing Machines and Languages with only Two Formation Rules, *Comm. ACM*, **9**, 366-371.

[3] S. Ginsburg, [1966], *Mathematical Theory of Context-free Languages* McGraw Hill, NewYork.

Proofs

Proofs will be sent to the author to ensure that the papers have been correctly typeset and *not* for the addition of new material or major amendment to the texts. Excessive alterations may be disallowed. Corrected proofs must be returned to the production manager within three days to minimise the risk of the author's contribution having to be held over to a later issue.

Only orginal papers will be accepted, and copyright in published papers will be vested in the publisher.

Letters

A section of "Letters to the Editor" (each limited to about 500 words) will provide a forum for discussion of recent problems