A HIERARCHICAL SPATIAL CANONICAL DATA MODEL – TOWARDS FEDERATING HETEROGENEOUS GISs

Yaser A. Bishr, M.Sc.
Ph.D. Candidate
Department of Geoinformatics
International Institute for Aerospace Survey And Earth Sciences, ITC
P.O.Box 6, 7500 AA Enschede The Netherlands
email: Yaser@itc.nl

Inter-Commission Working Group III/IV

KEY WORDS: Database heterogeneity, federated GISs, semantic data sharing, spatial canonical data model

ABSTRACT

Decision making process is usually multi-level, e.g., local, regional and national. Taking a decision at any level requires the consideration of the other ones. Decision support systems which may support each level of decision making contain data, information, and knowledge pertaining to the problem domain. In the framework of developing a multi-level spatial decision support system for watershed management, it is required to share data, information, and knowledge between the involved disciplines. This requires the federation of the GISs which support such decision support system. Providing the co-operation between these autonomous heterogeneous GISs while at the same time maintain their autonomy has been an area of great interest in the past few years. This situation is called interoperability and the system which manages the interoperability is called federated database system. Most of the current research in the federated databases is tackling the problem of syntactic, schematic, and to a lesser extend, semantic heterogeneity of federated databases. This is because most of the publications are from non spatial database perspective, federating heterogeneous GISs will prose some complexities for the canonical data model.

A canonical data model, also called a unified data model, is a wrapper around the heterogeneous databases gives non-local users the illusion of an integrated view. It is a uniform interface to the underlying databases. The is an integral part of the client server architecture. it is a mediator between heterogeneous GISs which allows the exchange of data and services.

The objective of this paper is to represent a model for sharing geographic information. The concept is distinguished from other concepts by its high semantic contents. It presents a spatial canonical model and show that in order to provide reliable and full-fledged interoperability, semantic relationship and similarity between objects should be considered. The ideas which will be presented here are considered as an extension to GIS theory. Originally, the theory is focused on defining object hierarchies in a single database. Here the theory is extended in order to accommodate several hierarchies within different databases in order to provide co-operation between multiple heterogeneous GISs.

1. WHAT SHOULD BE SHARED

The need for shared information is a direct result of the multi-level nature of watershed management. There are many different views of a given data set by the three decision levels, local, regional, and national. Difference in views can also arise when management plans, designed at one level, are examined at a higher one. This is due to the diversity in the objectives of each management level. When various perspectives overlap, they necessitating the for sharing of information if management plans are to proceed concurrently and cooperatively. For more information to be shared, there must be a commonly understood representation and vocabulary [McGuire G.J., et al. 1993].

While computers are used extensively in product development, existing approaches do little to facilitate information sharing and coordination. The approaches are only focused on providing front end interface across the Internet. The interface allows users to browse a metadata directory and to locate data sources. Data are then transferred from the source either in a standard data format or users might be exposed to a set of available data standards from which they can select [Alaam M., 1994] and [Otoo, J.E., et al., 1994]. The main disadvantage in this approach is that users spend a substantial amount of time restructuring their data in order to comply to their data model.

Interoperability is defined as the ability of GIS users and developers to transparently access geographic data sets and processes available on the net from heterogeneous systems, provided that the autonomy of the members of the federation is maintained [Bordie M.L., 1993]. Accessibility of data, knowledge, or functions is guaranteed under full consistency, integrity constraints and concurrency specifications, i.e., full interoperability. Members of the federation are called component SDSSs, and their GISs are called component GISs.

Understanding what exactly is needed to be shared is a precursor for providing full interoperability. Spatial objects are traditionally described in geographic databases by their geometric and thematic attributes. This is known as the syntactic description of geographic objects. Object identification based on their syntax in a distributed heterogeneous database renders users to search for objects by their geometric and/or thematic attributes. The shared objects might have different meanings in both the data source and the receiver. For example, a user might query the federation for height information of a particular area which has aspect equal to 7, measured on a discrete scale from 1 to 9. On the other hand the data source might have aspects stored in degrees

measured on a continuous scale from 0-90. Our system must allow its users to query the federation by their own vocabulary. The system is then transparently handling and resolving the discrepancies. The approach introduced in this chapter is called semantic data sharing of spatial data.

Building semantics onto the syntactic description of geographic objects can be considered as wrapping them with semantic descriptors. Users are then interacting with the federation through this semantic wrapper while the system is transparently resolving the syntactic differences. However, in order to establish such semantic descriptors, data sharing concept poses some problems in several disciplines which have to be resolved:

1. A common vocabulary must be defined that allows various decision levels to exchange information at the semantic level. This common vocabulary is know as
2. A set of protocols must be established that permit semantic-level exchange of information.
3. An architecture which implements the concept of semantic data sharing.
4. Applying this concept on a large scale will definitely increase the traffic on the network. A set of basic facilitation services is required that off-load functionality such as name service, buffering, routing of messages, and matching procedures and consumers of information.

In this paper a mechanism for building a canonical data model is introduced. Moreover a protocol for semantic-level exchange is established. The system architecture and the network aspects are outside the scope of this paper. An overview of the current technology and research activities in the field of data sharing is shown in section 2. The spatial canonical data model is explained in section 3. The proposed concept for semantic data sharing is shown in section 4.

2. LINKING HETEROGENEOUS SPATIAL DATABASES

[Saltor et al., 1993] provided a comprehensive classification of heterogeneity. The classification has three aspects: syntactic, schematic, and semantic.

2.1 Syntactic Heterogeneity

Each database may be implemented in a different DBMS with a different data model, e.g., relational model Vs object oriented model. Syntactic heterogeneity is also related to the geometric representation of geographic objects, e.g., raster and vector representations.

Current technology and research activities for sharing spatial information are tackling the above two aspects of syntactic heterogeneity. One of the most prominent and promising technologies which aim to provide connectivity between heterogeneous databases is the open database connectivity, ODBC [Kyle Geiger, 1995]. It can be plugged in most of the current platforms. The main objective of ODBC is to resolve the heterogeneity of the DBMSs, i.e., syntactic heterogeneity. Users are able to interact with different platforms regardless of their underlying operating system and DBMS. It is a standard application programming interface (API) for accessing data in both relational and non relational database management systems. Using ODBC's API, applications can access data stored in a variety of personal computer, minicomputer, and mainframe DBMSs, even when each DBMS uses a different data storage format and programming interface.

The open GIS consortium, OGC is responsible for improving the other aspect of syntactic heterogeneity, i.e., geometric representation of geographic objects, using its open Geoinformation specifications, OGIS [Schell D., 1995]. The specifications have two parts: **1)** *the Open Geodata Model*, OGM, which provides a common geodata model for all spatio-temporal data. The model supports both object and field based approaches. **2)** *Open Geoprocessing Services*, OGS. It defines a common consistent set of geoprocessing software interfaces. These interfaces define the behaviour of geoprocessing software services which access, interchange, manage, manipulate and present geospatial data specified in OGM.

The basic strategy of OGIS is to define a set of well known types and common aggregates as the basic building blocks. The well known types would include common programming types such as integers, real numbers, character strings. The aggregate types would include common programming database aggregate constructors such as list, set, mullet-set, and tuple. Moreover OGIS defines a basic level of spatial and temporal primitives which would allows systems to build their own internal representation. This includes point, line, areas, surfaces, curves, and simplexes.

Provided that in the near future all GISs are using these concepts in their basic definition, the transformation from one spatial domain to the other is straight foreword process.

2.2 Schematic Heterogeneity

Objects in one database are considered as properties in the other. Moreover, object classes of the same real world entity may have different hierarchies and descriptors in different databases. Unified data models are designed to handle this type of heterogeneity. The concept proposed in this paper is designed to handle the schematic and the semantic heterogeneity simultaneously, section 4.

2.3 Semantic Heterogeneity

A real world entity may have been represented in different ways by different designers in order to serve various applications, giving as a consequence semantic conflicts at the level of federation. For example a road network in a GIS for transportation has different semantics from that in a GIS for topographic mapping.

In the context of providing data sharing at the semantic level [Daruwala A., et al., 1995] proposed a strategy based on the notion of context interchange. In the context interchange framework, assumptions underlying the interpretations attributed to data are explicitly represented in the form of data contexts with respect to a shared ontology [Goh C., et al., 1994, 1995]. Ontology is a specification of a conceptualisation. That is, an ontology is a description of the concepts and relationships that can exist for a component GIS or a set of interrelated components.

The ontology of certain application domain is implemented in a component called mediator. A mediator is a paradigm which provides a link between data sources and receivers [Siegel M.,

et al., 1991]. This paradigm described how those features can be realised by showing:

1. How domain and context specific knowledge can be represented and organised for maximal sharing.
2. How these bodies of knowledge can be used to facilitate the detection and resolution of semantic conflicts between different systems.

A context mediator does a number of things each time it receives a query referencing multiple data sources. First it compares the contexts of the data sources and receivers to determine if semantic conflicts exist, and if so, what concessions need to take place to resolve them. This is referred to as conflict detection and the information detailing the conflicts are presented in a conflict table. This query then undergoes an optimisation process. Optimisation takes number of forms: a subquery can be reformulated so that constants referred to in a query are concerted to the context of the data sources executing that query. Finally, the intermediate answers obtained form component systems must be merged together and converted to the context understood by the receiver.

There are two pre-requests for mediation of semantic conflicts:

1. All sources and receivers must describe their contexts explicitly with respect to a collection of shared ontology, i.e., expert schema
2. All queries must routed through the context mediator mentioned above.

It is worth mentioning here that the context interchange concept is only suited for non spatial databases. However, merging the OGIS concept and the context interchange concept will result a proper semantic data sharing mechanism. In the next section a proposal with such an idea is presented. Furthermore, a different approach will be followed for modelling ontology of spatial objects. A canonical model is presented. This model is based on using OGIS concepts for defining the syntactic structure and spatial representation of geographic objects, and metadata and relationship between objects for building semantics. A rule base stored in the mediator server accesses the metadata and the knowledge base for identifying objects and resolving semantic conflicts.

3. A SPATIAL CANONICAL DATA MODEL

The proposed theory for semantic data sharing is based on building layers of semantics onto the syntactic definition of geographic objects, Figure 1. At the lowest level of the syntactic definition we find the classic data structure, i.e., field and object based approaches. The GIS theory formalises the topologic relationships amongst objects, uncertainty aspects, and the handling of geometry and topology of fuzzy objects [Molenaar M., et al., 1993 a, b and 1995]. Finally the theory introduces a consistent framework for object hierarchies, i.e., generalisation, aggregation, etc. This theory is however focused on object representation is a single database. OGIS introduced an elaborate set of common syntactic vocabulary for spatial object representation for sharing and exchange. The syntactic part of the canonical data model is based on OGIS specification mentioned in section 2.1.

A node is a collection of interrelated contexts. Nodes can be divided into sub-nodes. A sub-node can have one or more context. A Context refers to the assumptions underlying the way in which an interoperating agent represents or interprets data. A context is defined by one and only one set of semantic specifications. Contexts can be structured in a hierarchical way. Hence, semantic specifications of a lower level context are used as building blocks for those at a higher level. A context can have one or more sub-contexts. Each context corresponds to one and only one database. A database in turn is corresponding to one and only one data model. A data model consists of one or more hierarchies. A hierarchy is formed by

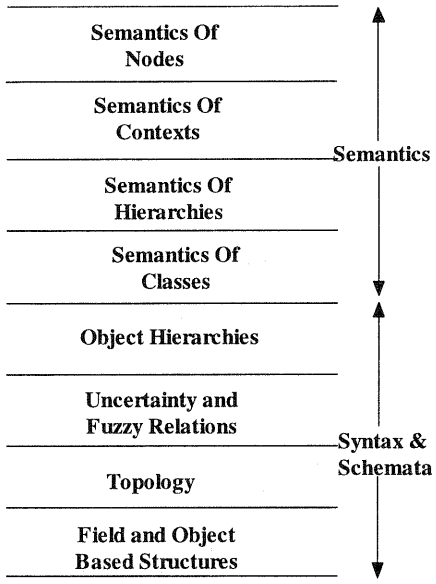

Figure 1 Syntactic and Semantic Definition

one or more object classes, i.e., intension. A class can have one or more instances, i.e., extension.

Figure 1 shows the semantics that have to be built onto the syntax. Object classes, intensions, hierarchies are considered as syntactic problem while the functional relationships between classes, within and across hierarchies, are considered as semantic problem. This is similar to the association concept between objects [Date C.J., 1995]. The relationship between hierarchies is the second semantic level. The relation between contexts, i.e., different databases, is the third semantic level. Relationship between contexts is defining the relationship between different GIS applications. The highest level of semantics is the relationship between nodes.

In the subsequent sections the semantics of classes, hierarchies, contexts, and nodes, are explained. FGDC metadata standards and add-hoc knowledge bases will be used simultaneously to model the two parts of semantics, i.e., metadata and relationships, respectively.

4. SEMANTIC DATA SHARING: A PROPOSAL

The semantic domain sem (D) is defined as the set of attributes used to define classes, hierarchies, contexts, and nodes

sem $(D) = <Y_1, Y_2, Y_3, \ldots, Y_n>$ where each Y_i is an attribute.

For each value d in the domain of D the semantics of that value can be defined in terms of the semantic domain as

sem $(d) = <y_1, y_2, y_3, \ldots, y_n>$ where $y_i \in$ domain (Y_i)

Throughout the explanation examples will be drawn from the field of watershed management which is the application used for implementing the concept.

4.1 Semantics of nodes

The metadata standards are structured in layers. Those parts which are needed to uniquely identify nodes, contexts, hierarchies, and classes, will be introduced in their respective layer. Figure 2 shows a hierarchy of contexts classified into nodes. Each node corresponds to a particular theme of applications.

At node 0 the common syntactic vocabulary is defined, as stated by OGIS. In addition to the common syntactic definition, the semantics of nodes are also defined at this node. Nor example node 1 is for earth resources and node 2 is related to mapping. A node is a defined by the 7-tuple given by

sem <node> = <name, Address, sub-nodes, super-node, context-thesaurus, role, contexts>

Name uniquely identifies nodes in the whole federation.
Address is the URL Internet address of the node.
Sub-nodes the immediate children of a node.
Super-node is the immediate parent of a node. Both super and sub nodes will allow scanning the node/context hierarchy.
Context-thesaurus is a set of alternative names of contexts within a node. This facility is added in order to resolve the naming conflicts that can arise from having different remote users, i.e., from different contexts, who have various perceptions and understanding of another context.
Role is a list of data types provided by contexts which belong to the underlying node.

As an example the semantics of node 1 and node 2 can be as follows
sem (node 1) = < Earth resources, itc.nl, none, federated GIS, {natural resources, environmental data}, {watershed data, water quality}, {environment, basin management}>

sem (node 2) = < mapping, ma.nl, none, federated GIS, {topography, surveying}, {topographic information, land parcel>, <cadastral, maps, triangulation information>

4.2 Semantics Of contexts

Contexts refers to the assumptions underlying the way in which an interoperating member represents or interprets data. Syntax and semantics at node 1 are used as building blocks for defining the immediate higher level contexts, i.e., WSM and water quality in Figure 2. Hence syntactic and semantic definitions are recursively defined from the lower level contexts to the higher ones. The semantic relationships are defined between each context and its super context. In this case the semantic level for data exchange between two contexts is the one defined at the common super context. For example the semantic level for sharing data between analysis and monitoring contexts is the regional context. A context is defined by 14-tuple
sem (context) = <node, name, address, sub-contexts, super-context, list of classes, class thesaurus, identification information, distribution information, quality, role, association, spatial reference information, spatial data organisation information>

Node the name of the node where the underlying context belongs to.
Name is uniquely identifying the context in the whole federation.
Address the URL Internet address of the context.
Sub-contexts are the immediate children of a context.
Super-context is the immediate parent of a context. Both super and sub contexts will allow scanning the node/context hierarchy.
List of classes is a set of class names which are in the database of the underlying context.
Class thesaurus is the list of the alternative names of classes within the underlying context.
Identification information basic information about the data set as defined by FGDC. A Boolean value which indicates whether the metadata exist or not.
Distribution information it is about the distribution options for obtaining the data set, as defined by FGDC. A Boolean value which indicates whether the metadata exist or not.
Quality is the general quality parameters of the database contents as defined by FGDC. A Boolean value which indicates whether the metadata exist or not.
Role is a set which represents the data types offered by the underlying context.

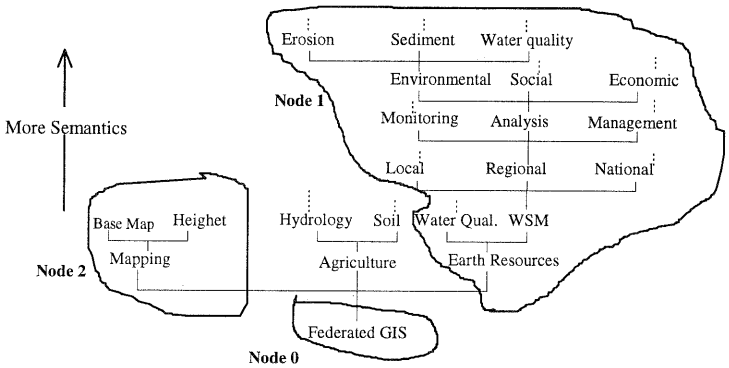

Figure 2 Node And Context Hierarchy

Association A Boolean value which indicates if the underlying class has a functional relationship with other classes.
Spatial reference information a Boolean value which indicates whether there exist metadata.
spatial data organisation information is the mechanism used to represent the spatial information in the data set. It is a Boolean value which indicates whether there exist metadata

As an example consider the context of watershed management, decision making for watersheds involves three types of activities: **1)** *monitoring:* where the aim is to prepare a concise data inventory of the status of the watershed; **2)** *analysis:* where the aim is to analyse watersheds in order to assess their vulnerability for degradation, and quantify the causes of degradation; **3)** *management:* where the intention is to introduce and implement new management plans. These activities are performed by different groups which can be considered as different contexts. Corresponding to these contexts are three object hierarchies Figure 3. At the lowest level of these hierarchies are elementary objects. Elementary objects are those at the lowest level of an abstraction hierarchy

of a particular database. At the highest level of abstraction and corresponding to the three main activities of watershed management there are three different views. The views are further abstracted to other decision levels, e.g., local, regional and national

- At the monitoring hierarchy elementary objects are interpreted as hydrologic response units in the context of land degradation analysis. They will be abstracted to subcatchments, catchments, and basins at the local, regional, and national levels respectively.
- At the analysis hierarchy elementary objects will be abstracted to tessellation. These are the processing units for the simulation models which are used for analysing watersheds' degradation and also the impact of the new management practices. The tessellation units can be cells, polygons, or triangles.
- At the management hierarchy elementary objects will be abstracted into management units, farms, districts, provinces, etc.

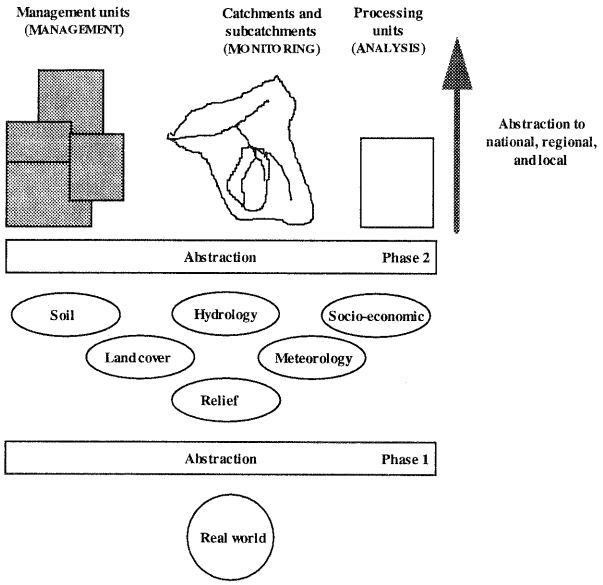

Figure 3 Abstraction of elementary objects to the three hierarchies

Semantics of the analysis and management context hierarchies are defined respectively as follows:

sem (analysis) = < earth resources, management, analysis.itc.nl, regional, none, {soil, land cover, hydrology, relief}, none, True, True, True, {water quality, Best management, erosion, sediment, water level},True, True, True>

Following is an example of a rule that defines the relationship between analysis context and the supper context, earth resources. The rule is written in Pseudo language for illustration.

data type = "soil erosion" **and** measurement unit = "kg/ m^2" **and** target = "earth resources" → data type = "soil degradation" **and** measurement unit = "pounds/in^2" **and get function** unitconv

The above rule means that if the context receives information for soil erosion and the requester is earth resources, then rename the data type to soil degradation and trigger the function unitconv in order to transform the units of measurements from kg/m^2 to $pounds/in^2$. The unit of measurements and other information are obtained from the metadata associated to the underlying context. At the super-context, earth resource in our example, a typical rule can be defined as follows:

data type = "soil degradation" **and** target = "management" then data type = "upstream erosion" **and** get function project.

The above rule means that if the data type to be transferred is soil degradation and the target context is management then rename to type into upstream erosion and apply function project in order to propagate the values stored per cell to that for the whole catchement.

The association relationship ensures a consistency in the retrieved data. For example if a user requests a land cover and soil information which are functionally related, preference will be given to the context which has both types rather than retrieving them from two different contexts.

4.3 Semantics of Hierarchies and Classes

A context, or a database, contains one or more aggregation hierarchies. Hierarchies in turn are formed by classes. The reason for introducing semantics of hierarchies and classes as one unit is that the underlying hierarchies exist in one database and hence are introduced in one data model. Semantics of hierarchies and classes are defined by the 3-tuple.
sem (hierarchy class) = <context, list of classes, class hierarchy, entity and attribute information>

Context the name of the underlying database.
list of classes is a set of class names which are in the database.
entity and attribute information is about the information content of the data set, including the entity types, their attribute, accuracy and domains from which attribute values may be assigned., as defined by FGDC.

Following the example in section 4.2, semantics of hierarchies and classes are shown as follows:

Sem (analysis) = <analysis, {cells, chemical output, sediment output,}, True>

class = "chemical output" and nitrogen > XXX then change land cover and get function AGNPS

This rule means that if the nitrogen value in the class chemical out is larger than a certain amount then change land cover type and run a simulation model, AGNPS, in order to calculate the new value of the nitrogen.

5. Conclusions

Information sharing has become an active area of research in the last decade. Exchanging information is currently achieved by providing files in a standards format. In the last five years several research activities were focused on providing a high level of semantic data sharing. This is achieved by providing mechanism for resolving the classic three aspects of

heterogeneity: syntactic, schematic, and semantic. In this paper, an overview of the technology showed that research efforts have to streamlined in order to achieve our objective. A novel concept for semantic data sharing is introduced. In this context, the difference between syntax and semantics is shown. Then a theory for building semantics onto syntax is explained. The concept is based on using hierarchy theory in building semantics onto syntax. The next phase in this research to build a prototype that implements the introduced concept. The prototype will be built using an expert system for introducing rules for controlling the access of nodes, context, hierarchies and classes. Information required for executing the rules will be stored in an object oriented metadata.

REFERENCES

1. Adil Daruwala, Cheng Hian Goh, Scott Hofmeister, Karim Hussein, Stuart Madnick and Michael Siegel, 1995. The Context Interchange Network. IFIP WG2.6 Sixth Working Conference on Database Semantics (DS-6), Atlanta, Georgia, May 30 to June 2.
2. Alaam M., 1994 *"management perspective of an infrastructure for GIS interoperability - the delta-x project"*. Proceedings of ISPRS commission II symposium on System for data processing analysis and presentation, Ottawa, Canada. Mosaad Alaam, and Gordon Plunkett (eds) Vol 30 No. 2. The survey, mapping and remote sensing sector, natural resource Canada.
3. Bordie M.L., 1992 *"The Promise of Distributed Computing And The Challenge Of Legacy Information Systems"*. In Proceedings of the IFIP WG2.6 Database Semantics Conference on Interoperable Database Systems (DS-5), Lorne, Victoria, Australia, 16-20 November, David K. Hsiao, Erich J. Neuhold, and Ron Sacks-Davis (eds), pp. 1-31.
4. Cheng Hian Goh, Stuart Madnick and Michael Siegel. Ontologies, Contexts, and Mediation: Representing and Reasoning about Semantic Conflicts in Heterogeneous and Autonomous Systems. Sloan School of Management Working Paper #3848; also CISL Working Paper CISL 95-04. Under review for publication. Get postscript. <Picture>
5. Date C.J., 1995 *"An Introduction to Database Systems"*, Sixth edition, Addison-Wesley Publishing Company, Inc., 839 pages, ISBN 0-201-54329-X.
6. David Schell, 1995 *"The Open Geodata Interoperability Information Package"*. Open GIS Consortium, Inc. 35 Main Street, Suite 5 Wayland, MA 01778.
7. FGDC, *Content Standards For Digital Geospatial Metadata*, Federal Geographic Data Committee, June 8, 1994. US Geological Survey, 590 National Center, Reston, Virginia 22092.
8. Goh H.C., Stuart Madnick, and Michael Siegel. Context Interchange: Overcoming the Challenges of Large-Scale Interoperable Database Systems in a Dynamic Environment. In Proceedings of the Third Int'l Conf on Information and Knowledge Management, pages 337--346. Gaithersburg, MD, Nov 1994. Get postscript or view the htmlized version of the paper. An expanded version of the paper is available as CISL Working Paper CISL-94-01. Get postscript .
9. Kyle Geiger *"Inside ODBC"*, 1995, Microsoft Press, Redmond Washington 98052-6399. 482 pages. ISBN 1-55615-815-7.
10. McGuire G.J., Kuokaa D.R.,Weber J.C., Tenenbaum J.M., Gruber R.T., Olson G.R.. Journal of Concurrent Engineering: Applications and Research (CRERA), 1 (2), September 1993.
11. Molenaar M., 1993 (a) *"Object Hierarchies and Uncertainty in GIS or why is Standardisation so difficult"*. Geoinformation systems, Vol6, No.3, pp22-28.
12. Molenaar M., 1993 (b) *"Object Hierarchies and Uncertainty in GIS or Why is Standardisation so Difficult?"*. Geo-Information system, vol. 6, No 3.
13. Molenaar M., 1995 *"An Introduction Into The Theory Of Topologic and hierarchical Object Modelling In Geo-Information Systems"*. Lecture Notes, Department of Land-Surveying & Remote Sensing, Wageningen Agricultural University, The Netherlands.
14. Otoo J.E., and Mamhikoff A., 1994 "Delta-X Federated Spatial Information Management System". Proceedings of ISPRS commission II symposium on System for data processing analysis and presentation, Ottawa, Canada. Mosaad Allam, and Gordon Plunkett (eds) Vol 30 No. 2. The survey, mapping and remote sensing sector, natural resource Canada.
15. Saltor F., Castellanos M.G., and Gracia-solaco M., 1993 *"Overcoming Schematic Discripancies in Interoperable Databases"*. In Proceedings of the IFIP WG2.6 Database Semantics Conference on Interoperable Database Systems (DS-5), Lorne, Victoria, Australia, 16-20 November, David K. Hsiao, Erich J. Neuhold, and Ron Sacks-Davis (eds), pp. 191-206.
16. Siegel M., and Madnick S.E., 1991 *"A Metadata Approach to Resolve Semantic Conflict"*. In Proceedings of the 17th international conference on Very Large Databases, pp. 133-145.