Supporting Multiple Representations in Spatio-Temporal databases

The MurMur Consortium
Contact: Christine Parent[1], Stefano Spaccapietra[2]
christine@lbd.epfl.ch, spaccapietra@epfl.ch
[1] HEC_INFORGE, University of Lausanne, Switzerland
[2] Swiss Federal Institute of Technology, Switzerland

The MurMur Consortium

- Ecole Polytechnique Fédérale de Lausanne, Switzerland
- CEMAGREF, Grenoble, France
- Institut Géographique National, Paris, France
- Star Informatique, Liège, Belgium (Project Coordinator)
- Université Libre de Bruxelles, Belgium
- Université de Lausanne, Switzerland

Abstract

Successful information management implies the ability to design accurate representations of the real world of interest to targeted applications. Current systems do not provide representation schemes supporting the diversity of user needs. In the context of interoperability, as in Web access to heterogeneous data sources, they cannot properly integrate the diversity of stored representations. The objective of the MurMur project is to enhance GIS (or DBMS) functionality so that, relying on more flexible representation schemes, users may easily manage information using multiple representations. The added functionality will support multiple coexisting representations of the same real-word phenomena (semantic flexibility), including representations of geographic data at multiple resolutions (cartographic flexibility). This will in particular make possible a semantically meaningful management of multi-scale, integrated, and temporal geo-databases.

What is multiple representation and why it is needed

Databases are intended to keep an integrated and consistent set of data that provides the information needed to support application requirements from one or several user communities. That data represent real-word phenomena that are of interest to its users. While the real world is supposed to be unique, its representation depends on the intended purpose. Thus, different applications that have overlapping concerns about real-word phenomena normally require different representations of the same phenomena. Differences may arise in all facets that make up a representation: what information is kept, how it is described, how it is organized (in terms of data structures), how it is coded, what constraints, processes and rules apply, how it is presented, what are the associated spatial and temporal framework, etc.

Current data management technology relies on a centralized representation paradigm, where all application requirements are integrated into a single stored representation (at least at the logical level). A view mechanism allows deriving, on demand, from the stored representation any other

representation that materializes the specific viewpoint of the requesting application. View mechanisms, however, are subject to strong limitations. In relational DBMSs it may not be possible to update the database using a view, because of the inherent ambiguity of updates on views that do not rely on a 1:1 mapping between tuples in the view and underlying tuples in the database. In object-oriented DBMSs view definition is further restricted to very simple views (e.g., defined by projection operations) because the rules governing data model constructs lead to inconsistencies in case of more complex view definitions.

What a view-based, centralized representation mechanism can definitely not support is the case where different application viewpoints are not derivable from each other (irreducible viewpoints). Assume a hospital information system, such that patients are identified by medical teams based on a patient number inscribed on a bracelet that the patient always carries, and the same patients are identified by the administrative staff based on a social security number. If the two viewpoints do not share other information (such as name and birthdate) that could provide a common identification scheme, when the patient leaves the hospital two different update operations have to be made for the medical and the administrative realms (no update propagation from one realm to the other is possible). This has evident drawbacks in financial terms (double cost for updates) and in terms of consistency of the database, that cannot be guaranteed by the system and has to rely on manual procedures.

From a traditional, centralized database system perspective, the coexistence of irreducible viewpoints in a database may be considered as a design error. From a user perspective it is not. In current DBMSs it is up to application designers and users to cope with the situation, whenever it arises, relying on primitive system functionality, such as foreign keys in relational DBMSs or is-a links in object-oriented DBMSs, to inter-relate different representations of the same phenomenon. It is again up to users and application designers to define and enforce the appropriate consistency rules that may constrain the set of representations.

The centralized representation paradigm is even more uncomfortable when a database results from the integration of data from different pre-existing data sets, as is the case in federated or cooperative information systems. Such systems are more and more frequently required to support interoperation among different organizations, as well as for a single organization that needs to coalesce data from different sources, including the Web, to support its enterprise strategy. When data from various sources come together into a single data store, the situation where different representations of the same phenomena coexist is likely to happen and cannot be considered as a design error.

In summary, modern data management requires a new representation paradigm, such that multiple representations of the same phenomenon may coexist in a database, and this should be explicitly described and made known to the system so that it may manage the situation accordingly. In other words, existing data models need to be extended with new concepts such as a multi-representation link, with a well-defined semantics (which says "this representation is about the same real-word phenomenon that this other representation"), and associated constraints and operators. Expected benefits include better real-word modeling, enhanced understanding of data structure diagrams and database content, improved consistency management, automatic update propagation, and data cleaning facilities (when two representations are used to check one against the other and determine if there has been some erroneous data acquisition).

What is multiple resolution and why it is needed

Geographic data is a major provider in terms of systems and services for the citizen, as one of the most common concerns in everyday life is locating something you are looking for. Maps are the most natural way to provide location information, and also serve as an excellent mean of visualizing analytical data about phenomena that have a geographical correlation. Hence, map production and display is an economically significant activity, nowadays supported by GIS/DBMS databases. This includes geography-compliant maps, that show items of interest as faithfully as possible with respect to their real-word location and shape, as well as schematic maps (e.g., city transport systems, airline connections diagrams, train networks, facility management networks), where the focus is on correct connections and readability rather than on precisely locating lines and nodes.

A map is drawn according to a given scale. At different scales, the same information is usually drawn differently, a physical zoom-in or zoom-out is simply inadequate. This is because drawing standards may change from one scale to another one, items may (dis)appear or be (de-)aggregated because their size make them (in)visible depending on the scale, their shape may be modified (made simpler or more precise), or simply the information is not available at the requested scale. Unfortunately, there is no complete set of algorithms that automatically derive a map at some scale from a map at a more precise scale. Some algorithms exist and more are being investigated (see, for instance, the AGENT project), and the process they implement is called cartographic generalization. Given this situation, map production systems tend to keep a separate database per scale, leading to already mentioned problems such as lack of consistency and uncertain update propagation.

The ideal setting would be of course to maintain a database where geometry information is kept at the most precise scale, and all geometries at less precise scales are automatically computed through cartographic generalization. Because this derivation cannot be fully automated, and also because cartographic generalization may be a long and costly process, the alternative is to perform cartographic generalization off-line and to store its result for direct reusability. Eventually, that means that a spatial object in a database may be associated to a variety of geometric representations that are scale-dependent. Databases with such a facility are called multi-scale databases. Actually, while the scale concept is perfectly understood and relevant when talking about maps, it is not relevant anymore when talking about representation of an object in a database (scale has to do with map drawing, not with object representation). When focusing on databases, it is more correct to use the term "resolution", usually defined as the minimum dimension of an object to be represented. The resolution of information in the database is the resolution used at data acquisition. If different geometries are acquired for the same object, we can talk about multi-resolution objects. In geographic databases, multiple resolutions pertain to different factors, such as multi-scale geometry, hierarchical value domains and object aggregation. From the geodata application perspective, multiple resolution is certainly the most urgent multi-representation problem to solve.

Objectives of the MurMur project

The overall objective of the MurMur proposal is to demonstrate that current functionality provided by commercial data management software (DBMS or GIS) can be extended to support more flexible representation schemes, such as multiple coexisting representations of the same real-word phenomena (semantic flexibility), in particular representations of georeferenced data at multiple resolutions or scales (cartographic flexibility).

Supporting multiple representations means that both users and the system are aware that two (or more) stored representations are related to the same real-word phenomenon. For instance, in a relational or object-relational database the same car may be represented once as an instance of a Vehicle table and once as an instance of the Collectible table, where the two tables may have no common attribute. In a geographical multiscale database, the same building may have a rather precise geometric representation of its shape at 1/10'000 scale and a less accurate representation at a 1/50'000 scale.

Current DBMS provide limited facilities to support multiple representations. Relational systems use views to support representations that are derivable from another one, but for representations that are not derivable from each other all the burden is on the users. Users are responsible for adding the necessary artificial keys and foreign keys to link related tables, and for providing the procedures to enforce the desired consistency rules. Object-oriented or object-relational DBMS provide some additional support for multiple representations through generalization/specialization hierarchies and the use of system- generated object identifiers. However, this is known not to be sufficient (in terms of expressive power, user-friendliness, and practicality) to provide full flexibility in multiple representation support.

For geographical data, the situation is even worst. In these applications the variety of representations extends over different facets, such as:
- multiple geometries (within the same geometric type or not) for the same object,
- multiple abstraction levels that make a set of objects coexist with the object(s) that represents their aggregation (whether the aggregation is based on geometric, temporal or semantic criteria),
- multiple abstraction levels that result in hierarchical value domains for attributes, and
- multiple representations as for traditional databases.

Few GISs can explicitly represent objects with multiple geometry, a feature needed in particular to keep representations at different resolutions. This is a major technological limitation, as it forces, for instance, organizations that produce maps at different scales to keep one database per scale, which of course opens up to severe consistency problems, requiring manual procedures. When maintaining multiple databases is not possible, e.g., for economical reasons, the need to represent data at a scale different from the one recorded in the database (through simple zooming, for instance) triggers error-prone processes and may lead to wrong decisions due to misleading representations.

In the database and GIS research communities the state of the art is more advanced, as both communities have been active in developing proposals for new object identification schemes. Database researchers proposed concepts such as roles, prototypical objects, deputy objects, or aspects. GIS researchers focused on inter-level connectivity, scale transition relationships, or stratified map space. Academic partners in the MurMur consortium have actively participated into this research stream, for both traditional and geographical data. Their work provides a theoretical background to the MurMur proposal.

MurMur is specifically interested in providing a support to multiple representations that is adequate for geographical data. One of the motivations is the societal importance of geographical data: 80% of all business and policy problems involved in creating livable communities include geographical data as a key element in their decision-making process (source: NCGIA 1999 National GeoData Forum, Washington, DC, June 1999). Another motivation is that the diversity of user profiles is much broader in geodata applications (where it spans from environmentalists to city planners, cadastral administrations, sociologists, etc.) than in conventional database applications, where a given database usually serves a community of users belonging to the same organization. Targeted applications include environment management, natural risk assessment, multiscale map production systems, and embedded navigation applications. Because consistency of multiple representations can only be achieved if the system is capable of explicitly interrelating them and dealing with their differences, an appropriate data model is first needed. The MurMur proposal aims at developing such a data model and framework.

This consortium has been formed to include geodata providers, geodata users, a GIS supplier and academics with R&D experience in geographical databases. From the scientific perspective, its main focus is on addressing the problems due to the coexistence of various representations of the same piece of land using different resolutions and different viewpoints. The approach will be to deal with the problems in a way that is consistent with a generic approach to multiple representations in traditional databases.

MurMur shall take into account three different perspectives:
- multiscale databases, where representations at various resolution levels are stored in a single database, typically correlated via cartographic generalization functions,
- integrated databases, where representations coming from different existing databases are correlated and virtually integrated in a federated framework, and
- temporal databases, where multiple representations correspond to representations taken at different points in time.

The scientific outcome will be the formal definition of an enriched data model, and associated definition and manipulation languages, to allow the management of real-word objects at different spatial and semantic resolutions. The enriched model and languages shall be able to: 1) associate several representations to an object, 2) define topological relationships that differ with regard to spatial resolution, 3) link objects corresponding to each other at different resolutions, and finally, 4) describe and query objects and their attributes at different levels of semantic resolution.

From a development perspective, MurMur intends to develop, test, and validate a software layer (middleware) on top of an existing GIS/DBMS. This layer will interface with users on one side and with the underlying GIS/DBMS on the other side.
Functionality provided by the MurMur layer shall allow users to:
- define and edit a schema of a spatio-temporal database using the enriched data model,
- input data into the database and let it evolve through insert, update, and delete operations, while preserving the integrity constraints on related multiple representations,
- query the database using thematic as well as spatio-temporal criteria.

Innovation

A first innovative aspect is a new representation paradigm, the multiple representation approach, which is the core of the MurMur proposal. Multiple representation is the necessary consequence of both the subjectivity of perception and the diversity of interests over shared data. To enhance the limited multi-representation functionality of current data management systems, researchers have investigated multi-instantiation mechanisms, i.e., allowing an object to be simultaneously represented in several classes. Proposals exist for multi-instantiation at the object level (objects may be dynamically re-grouped to form new classes), or at the schema level (an analysis of the object life cycle determines the set of relevant possible representations, or roles, for an object). MurMur will follow the latter stream and take it to the implementation phase. The innovative scientific contribution is a comprehensive approach that also takes into account inheritance and polymorphism mechanisms as well as data manipulation facilities. Moreover, our approach to roles does not require that a primary role is identified, a constraint that may not be consistent with the real world to be represented but that is nevertheless quite common in the proposed approaches as it makes design and implementation of the approach easier.

A second innovative aspect is that MurMur specifically contributes in advancing the state of the art in multi-resolution geographical databases. The resolution concept applies to spatial, temporal, and thematic characteristics of information. In the spatial dimension, the multi-resolution aspect relates to cartographic data presentation at different scales. In the temporal dimension it is usually understood as related to temporal granularity. In the thematic dimension, multiplicity of resolution levels may materialize through definition of classification hierarchies, aggregation of objects (based on thematic, spatial, or temporal criteria), and definition of hierarchical domains for attribute values. Most of the work in multi-resolution has been done in the GIS domain, where researchers have focused on multi-scale aspects.
MurMur intends to develop a more comprehensive approach, building on similarities between the different facets of multi-resolution. In particular, thanks to the experience of the consortium in cartographic generalization, MurMur will be able to provide more intelligent links among representations. Another specific target will be an elaborated scheme for aggregations, to apprehend $n_x m$ links where a group of n objects in one resolution level corresponds to a group of m objects in another resolution level, an area where little research results are known. Both geometric and semantic resolution models will be dealt with.

The new paradigm will be embedded in a conceptual model for spatio-temporal data, whose foundations have been developed by consortium partners. The conceptual model builds on the orthogonality principle, i.e., the thematic, spatial, and temporal dimensions are independently organized so that choices in one dimension do not constrain choices in another dimension. Orthogonality provides design flexibility, hence an expressive power that is beyond those of current data models. Inclusion of multi-representation functionality will add in making this model the most advanced conceptual model for spatio-temporal data. The model is expected to play a major role in fulfilling the current gap between user expectations and requirements for user-friendly systems, and the tools that are on the marketplace.

Several tools will be implemented in order to assess the adequacy of the new multi-representation paradigm with respect to the users' need.

First, a visual schema editor will allow the definition of spatio-temporal schemas through direct manipulation of the modeling constructs on the screen. While visual schema editing is a well-established technology, its usage for spatio-temporal databases is limited to very few tools. However, these tools inherit the limitations of the underlying model, such as lack of orthogonality and limited data structuring capabilities.

Next, a visual query tool will combine direct manipulation with textual specifications and menu-driven query formulation. Developing a fully multi-paradigm query tool (e.g., including sketches, voice and gesture interactions) for spatio-temporal databases is beyond the scope of MurMur, as it would be a full project in itself due to the inherent complexity of spatio-temporal queries, in particular for disambiguation. Still, the MurMur query facility is expected to provide significant advances to the state of the art in query tools.

Furthermore, the tools developed in the MurMur project will constitute an initial answer for supporting the process of integration of existing information sources in distributed and interoperable environments. Although MurMur does not intend to develop a full data integration approach (again, this would be a research project in itself), MurMur plans to support acquisition of information about correlated data objects in different databases. This kind of information is typically produced by object-matching processes applied to separate geographical data sources. MurMur will consider this information as a possible input to create a multiple-representation database (i.e., an instance-driven approach rather than a schema-driven approach). From this input MurMur will build an integrated view of the correlated objects and databases. This may be seen as a minimal mediation facility, whose purpose is to match user requirements on a simple, incremental basis. Such an approach is innovative for geographical databases.

Bibliography

Parent C., Spaccapietra S., Zimanyi E., "MurMur: Database Management of Multiple Representations", AAAI-2000 Workshop on Spatial and Temporal Granularity, Austin, Texas, July 30, 2000

Spaccapietra S., Parent C., Vangenot C., "GIS Databases: From Multiscale to MultiRepresentation", SARA 2000, Symposium on Abstraction, Reformulation and Approximation (SARA-2000), Lago Vista, Texas, July 26-29, 2000, Workshop

Vangenot C., Parent C., Spaccapietra S., "Multi-representations and multiple resolutions in geographic databases", Advanced Database Symposium'99 (ADBS 99), IPSJ Symposium Series Vol.99, N°19, Tokyo, Japan, December 6-7, 1999

Spaccapietra S., Parent C., Zimanyi E., "MurMur: A Research Agenda on Multiple Representations", 1999 International Symposium on Database Applications in Non-Traditional Environments (DANTE'99), Kyoto, Japan, November 28-30, 1999

Parent C., Spaccapietra S., Zimanyi E., "Spatio-Temporal Conceptual Models: Data Structures + Space + Time", 7th ACM Symposium on Advances in GIS, Kansas City, Kansas, November 5-6, 1999