

DESIGN AND IMPLEMENTATION OF DIGITAL DATABASE AND GIS SOLUTIONS IN MARINE GEOSCIENCES – AN INTEGRATED APPROACH

Tapan K. Jana and Debasis Sengupta

Geodata Division, Marine Wing, Geological Survey of India, DK6, Sector-II, Salt Lake City, Kolkata, India – (tkjana, dsengupta)@gsimarine.org

Keywords: *Geospatial database, Entity-Relationship diagram, Map viewer, GIS solutions Normalisation Three-tier architecture*

ABSTRACT :

The marine geosciences data management practice is focussed to build up a part of knowledge-based geospatial database system that would be of great help to assist marine geoscientists in interpreting various geoscientific characteristics across the ocean and ocean bottom surfaces and assessing the potentiality of a specific area from onboard and laboratory test results. The structuring and development of digital geospatial database ensures the ability to store attributes of twenty-three discrete mappable geometric feature elements of seabed maps. The Entity-Relationship approach to data modelling reflects the heterogeneity of the data sources and the E-R model was achieved through data normalization, finding out identity of data records, their hierarchisation, spatial data structure and temporal considerations. Each event driven thematic table/entity records a set of information representing a data group, the characteristic properties or parameters of a specific survey area. The relation between entities as defined by composite keyed fields allow data to be aggregated over logically related master-detail entities. Integration of GIS solutions is viable both at the desktop and the middle-tier level of classical three-tier setup used for browser based deployment.

INTRODUCTION

During the last two and a half decades, Marine Wing, Geological Survey of India (GSI) is engaged in acquisition of a vast wealth of information on marine geosciences. Both spatial and non-spatial data archival in digital form, implementing an efficient data management system, speedy retrieval of data, prompt exchange of information amongst intra- and inter-department users warrant building up of a uniform data structure. This would involve organized storage of various forms of data and creation of digital database with fundamental information that will represent part of reality and fulfil the objectives of GSI or the concerned projects. The digital database creates facilities for sharing of data in the form of a multi-user database which most of the sophisticated database management software (in this case ORACLE RDBMS) allow many people to share data simultaneously without interfering the actions of other users. Besides, digital format of data also allows input into wide range of visualization and modelling software, and the digital system can impose easily complex self-checking procedures and security measures in a properly managed environment (cf. Giles, J.R.A., 1995). Rasmussen, K., 1995, however, states that the database should represent the true reflection of the structure of information contained within it and capable of handling wide range of processing need of data processing software with equal efficiency. Elmasri & Navathe (1989) opined that the related data that makes complete information in the database should be logically coherent and must have some inherent meaning (cf. Giles, J.R.A., 1995). The present study brings out a detailed account on database analysis, design and development, and scheme of implementation of 'an enterprise digital database' related to marine geosciences studies in GSI with a humble attempt to integrate GIS solutions for management of spatial data and dynamic composition of various thematic maps at desktop level and visualization and consultation of those thematic maps using a browser based basic map viewer for viewing spatial data directly fetched from the digital database system.

DATABASE ANALYSIS & DESIGN

Building up of the database was started with the task of database analysis or data modelling in which a conceptual data model of part of reality (that is to be represented in the database) has been developed. It provides the framework for storage of geological data. In order to construct the conceptual data model, a clear understanding and description of the reality was achieved through the process of iterations and in depth interactions with the experts in marine geosciences and consultation of published as well as unpublished literatures of national and international standards. Following are the events and sub-events identified for translating those into E-R model.

- Description attributes of a sample in general
- Description attributes of a core sample (subsurface material)
- Characteristics pertaining to various size fractions of a sample
- Petrologic and mineralogical attributes of a sample
- Palaeontological attributes of a sample
- Chemical determinants of a sample(solid/liquid/gas)
- Geotechnical attributes of a sample
- Bathymetric survey attributes
- Offshore mineral resource attributes
- Magnetic survey attributes
- Shallow seismic survey attributes
- Side scan sonar survey attributes
- Physical oceanographic attributes
- Weather attributes

The data modelling was guided by a set of events and sub-events which represent a set of information types e.g. types of sediment that are contained in surface and subsurface seabed

sediment deposits or placer mineral concentrates in beach sediments. In the database a set of constraints helps to ensure that the data stored in the database is logically coherent, meaningful and consistent, and maintains integrity of the database thereby reflects a possible reality. The constraints alone are not able to guarantee the correctness of data in the database; rather help to prevent non-coherent and inconsistent data being entered. For example, a typical constraint might state that a sample can be taken from only one geographical location.

The most widely used Entity-Relationship (E-R) approach was adopted for drawing up the conceptual data model. The help of a CASE (Computer Aided Software Engineering) tool of Oracle designer was used for this purpose. The information types in the E-R model are constrained with the following three concepts (Rasmussen, K. 1995)

- Entities about which the information is expressed.
- Relationships which are meaningful association between entities.
- Attributes or properties of entities which express information about entities.

Each of these concepts can be grouped into classes or types. The information types or entity types, relationship types and attribute types have been considered key components to describe the information requirements of a cruise area under study. The effect of imposing database constraints is both explicit and implicit in nature and their effect on relationship types and attributes of entities is obvious. 'Degree' and 'Optionality' are the two explicit constraints which affect the relationship types. For example, a gravity core sample can only be taken at one position is a constraint of degree whilst putting up of a gravity core sampler must be at some geographical position is a constraint of optionality. The implicit constraints affecting attributes may be of several types. For examples, all attributes must be single valued e.g. an occurrence of seabed sediment must be compositionally single type and all entities must be uniquely identifiable within the model in the same way and for all time.

The E-R diagram (Fig. 1) mainly records about system entities and information about the relationship between entities. The system entities are objects for which information is recorded and are represented by rectangular boxes. They are mainly of two types: master entities and detail entities. The line drawn between two entities along with some additional symbols attached represents the relationship between entities. Three conventional symbols such as a circle, a bar and a crows-foot have been used which indicate the cardinality of the relationship between entities. The pair of bars, nearest to each entity, represents a maximum cardinality. A single bar indicates a minimum cardinality of one and a circle indicates a minimum cardinality of zero and a crows-foot indicates a maximum cardinality of many. In fig.1, for a given location point there may be described by zero or many sampling observations whilst a sampling event must describe one and only one location.

The task of database design, however, involves mapping of the data model onto a computer data structure that helps to manage, process and display the model effectively. It also follows subsequent modification and transformation of the data model in order to make the processing efficient depending on the availability of typical database management features of ORACLE RDBMS software. Another important aspect of database design is normalization of the logical database model. It involves organizing the data into more than one table and

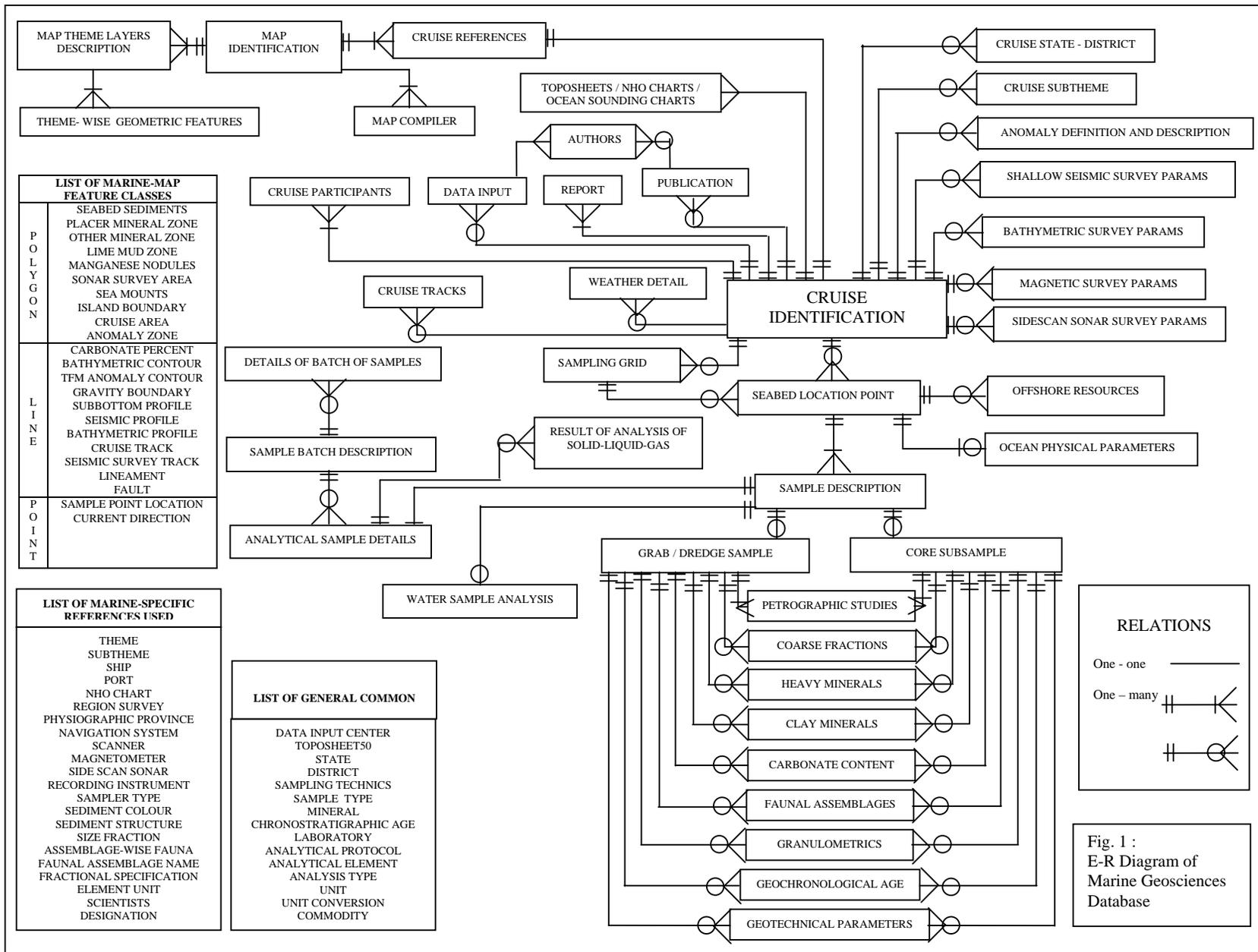
introduces faster sorting and index creation, few indexes per table, few nulls and an increase in the compactness of the database. It also helps to simplify the structure of tables. During designing process, normalization results in the formation of table satisfying certain specified constraints, represents upto 3rd Normal Form or 3NF. The normal forms have been attained to ensure that various types of anomalies and inconsistencies are not introduced in the database.

DATABASE STRUCTURE AND GUI

Once the designing of database using E-R diagram is complete, system entities and their relationship have been mapped onto a logically related tables. A table is defined with a set of rows and columns where the attributes of entity are represented as column heading and the data about the entities as rows. The Entity-Relationship (E-R) diagram for the logical model (Fig 1) can be divided into several logical areas such as locations with location descriptions, sampling grid description, samples, sample descriptions, batch information and result of analytical determinants, description of bathymetry, magnetic, shallow seismic, side scan sonar surveys, offshore resources, offshore weather, oceanographic physical attributes, anomaly identification, anomaly descriptions and above all the related cruise description. The E-R diagram also shows a separate logical area for thematic marine geological maps along with its related detail entities representing twenty-three geometric feature elements with a link to all other logical areas through the cruise references. The model is reasonably compact and has been implemented on object-oriented relational database management system i.e. Oracle RDBMS.

In order to enter the observed geological attributes and laboratory results as rapidly and accurately as possible, the oracle forms have been developed with the help of oracle developer tools which have very powerful Graphic User Interfaces (GUIs). These application forms serve both data entry and data retrieval. Each application form is specially suited to give access to a particular form of data e.g. grab/dredge sample, core sample, etc. There are number of such specific application forms for the input/output of different data types and sets of related attributes, as well as particular aspects of data manipulations. While developing the applications, various rules and validations have been imposed through writing hierarchical triggers at form level, block or table level and item level respectively. Two sets of Oracle forms have been developed for deployment of application forms in the client/server architecture and in the browser based three-tier architecture.

A few key example entities like location, sampling grid, sampling, batch of sample analysis and chemical determinants have been elaborated in order to throw some light on the mechanism and method of identification as well as normalisation of proposed related tables or entities. While describing the entities, the snap shots of oracle form developed for respective entity has been shown for better understanding of the data input, structure of the entity and retrieval processes.



Location

Location is one of the key entities amongst all the logical areas of the E-R diagram. An attempt has been made to combine all the subtypes of locations identified in the logical events into one entity but information recorded in respect of sampling, various surveys, climate details, etc. are very different. Linked to the sampling, bathymetric, magnetic, shallow seismic and side scan sonar data recording are important tables describing the surrounding area for geological description, geological features, characteristic attributes of bathymetry, magnetic behaviour, seismicity, and seafloor morphology respectively. A snap shot of location form is shown in fig. 2a.

Sampling Grid

A sampling grid is one of most important entities defined within the enterprise database. The entity attributes like grid number, grid origin x and origin y co-ordinates, and the number of x-y mesh together form a record set. The location entity has a referential link to grid entity in which a location point is uniquely identified within one grid only. The grid entity inherits the keyed attributes input center code and cruise number to which it is related. These keyed attributes when combines with a system generated unique grid identification number to form the primary key of the grid entity. The snap shot of GUI (Fig. 2b) represents the input/output user interface

Sample

A single location may be considered for various kind of sampling and it requires enforcing a relation to location entity. Three subtypes of sample have been identified such as grab/dredge sample, core sample/sub-samples and water sample. Each sub-type has its distinctive characteristic attributes. A sample inherits the keyed fields such as input centre code, cruise number and location number from the location table. These keyed fields combine with a sample number, a unique number generated from a system sequence defined in the database, to form a composite primary key of the sample table. The sample table has an alternate key which is defined by a unique field code assigned by the officers while collecting samples. Each sample sub-type has a set of master-detail tables and each table contains a set of information forming a record group. For example description of petrology and mineralogy of a sample forms a record group. Similarly, fraction wise heavy mineral concentrates, clay mineral contents of a sample, parametric geotechnical determinations, etc. reveal different sets of record groups. The snap shot of GUI (Fig. 2c) shows the input/output user interface for description of samples.

Batch Description for Analytical Sample : The information set recorded on a batch of samples selected for chemical analyses is critical to the database design as it provides a management and quality assurance tool for all kinds of sample analyses. It also reveals a metadata index to all analytical data produced by chemical laboratories of marine wing, GSI. Samples are grouped into a number of batches on the basis of nature of analysis and group of analytical determinants selected for analyses. The information on each group of samples is represented in the batch master table/entity which has the same batch reference code, laboratory reference and a unique batch

priority number. Whereas the batch detail table/entity records information on the set of analytical determinants which are grouped for analyses, analytical protocol identification number (as shown in Fig. 2d) that has a reference link with an analytical protocol reference entity which holds information on method of analysis of the batch, and the element or species determined by the method, quality control data such as lower and upper detection limits in order to maintain precision. An entry in the batch master and detail tables may have reference link with samples from different cruises/projects. Sample analyses without batch information is an incomplete information which may suffer from several critical issues like missing analytical methods, variable quality data such as precisions or detection limits, unknown laboratory reference, and even post analyses normalization of analytical data.

Sample Information and Analytical Determinants: The information on a group of analytical sample is recorded in the analytical sample description table/entity which is considered as one of the most important entities of sample chemical analyses. It inherits the keyed references from the batch entity and sample entity of marine geosciences domain. The referential integrity constraints like laboratory code, batch priority number, batch reference code, area code (cruise_id), and an unique sample id together define the primary key of analytical sample entity.

The analytical result table is one of the largest tables/entities in the database. The table/entity has the capability of holding in tunes of millions of analytical determinants when fully populated. Each row of the table contains area reference, sample identification number, laboratory code, batch reference, analytical determinants and abundance. Three tables corresponding to three media (solid/liquid/gas) have been created in the database for easy management and fast retrieval of data. The snapshot (fig. 2e) shows the user interface of analytical sample and the analytical determinants tables.

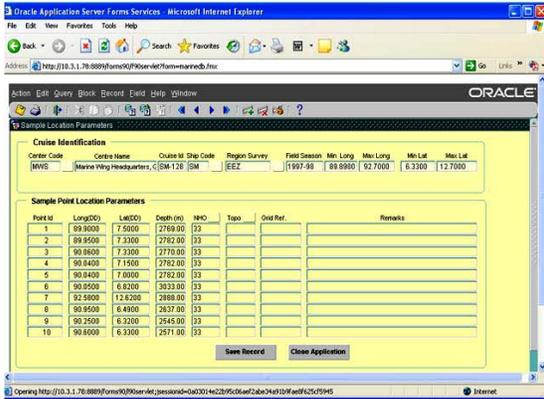
Sensor Survey Data

Bathymetry, magnetics, shallow seismics, side scan sonar and sub-bottom profiling are included in the sensor data acquisition systems. All these events have separate master and corresponding detail entities as shown in the Fig.1.

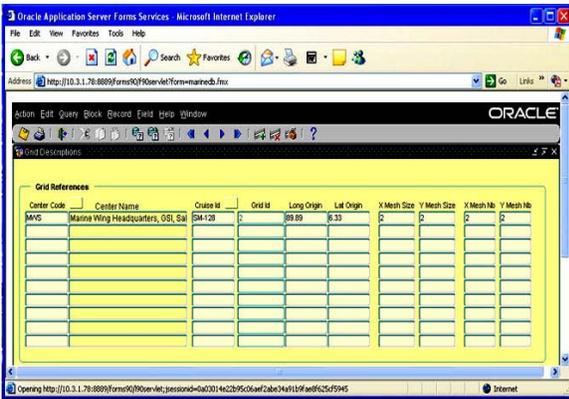
The master entity records information on the general description of the events, instruments reference for having precision information. The detail entities, on the other hand, record the actual data recorded by the instrument for a particular area of study.

Other Data

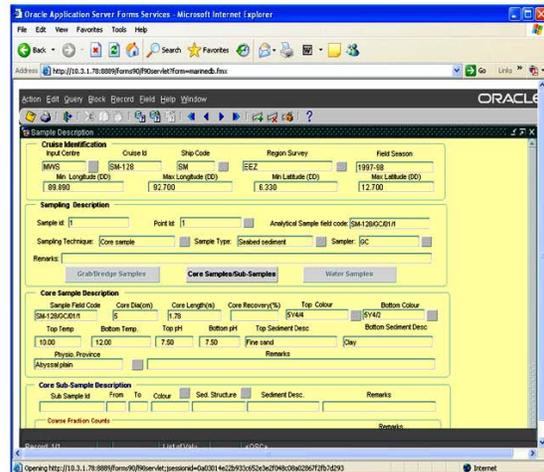
This section involves recording of information on prevailing weather, offshore mineral resources, ocean physical parametric studies, offshore resource specific anomaly zone identified in a specific study. All these events have their separate master and detail entities in the database for recording various attribute information.



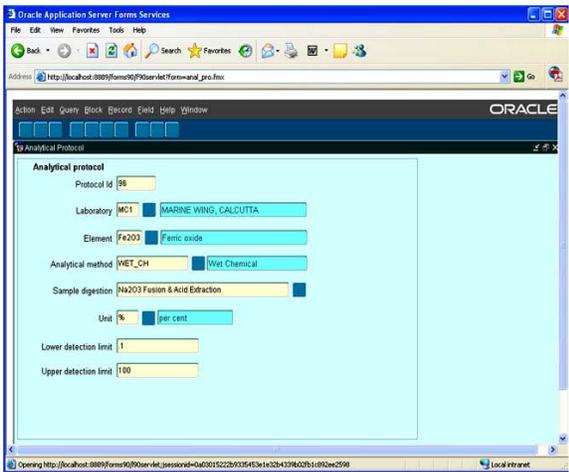
a



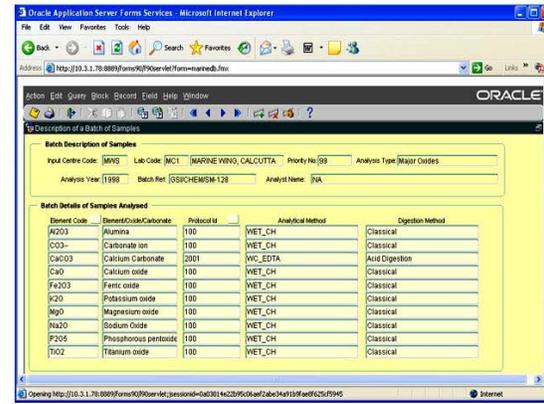
b



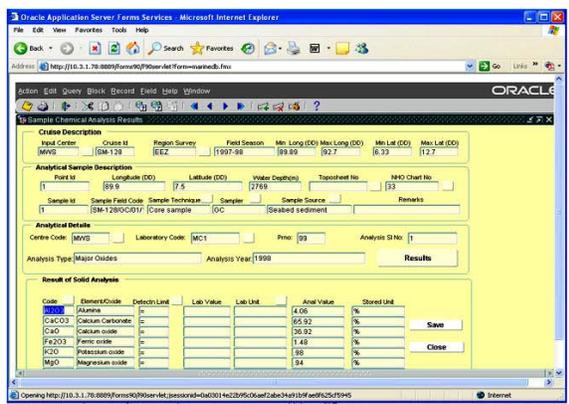
c



d



e



f

Fig. 2 : Snap shots of Graphic User Interfaces (GUI) : (a) location (b) sample grid (c) sample description (d) analytical protocol (e) batch description of analytical sample (f) analytical results of chemical determinants

Map feature elements

In all, twenty-three map feature elements of various geometric classes (Point/Line/Polygon) have been identified (as shown in Fig.1) considering all possibilities by studying in depth all the published and unpublished thematic seabed maps of Marine Wing, GSI. In the database, the map master table/entity records general information of the map that includes a unique map identification number, edition number, edition year, beginning year, ending year, restricted or unrestricted category information and so on. Corresponding to all the geometric feature elements there are separate detail entities which hold attribute information about feature elements and have a reference link with a unique geometry identification number. All the map feature elements are digitized as layers either in the desktop ARC GIS or Autocad Map which has a unique geometry identification number assigned by the user and it is used as a join-item in order to establish a seamless link with corresponding feature table/entity in the oracle database for making proper attribution to the map features during map composition and visualization. Once the a map is composed with all attribute information can be stored in oracle spatial for browser based deployment. The integration approach has been dealt with separately in the subsequent section of this paper.

DATABASE IMPLEMENTATION

In order to reduce the cost of design and maintenance, implementation of common database structure and common software standards across the distributed centers of the organisation has been planned. Similarly, widely agreed and used lexical references have been considered for data standardizations and to ease the understanding.

The database has been implemented at Marine Wing headquarters on the basis of the most evolved multitiered architecture offered by Oracle RDBMS software. The architecture is a classic three-tier setup. In the three-tier architecture (Fig. 3), the lowermost Database tier contains the core Oracle database running the marine geosciences database instance, the middle-tier combines the Web tier (incorporates the Oracle HTTP Server and Web Cache) and Application server tier (includes the core Application Server 10g and its ancillary products such as Oracle Application Server Portal 10g, Oracle Developer, Oracle Reports, Oracle Form Server, and deployed all browser based application forms) and the client tier contains either a java client application or a web browser for end users those who run always the latest version of applications using the URL (Universal Resource Locator) provided the client has the proper connection to the application server. Since the platform for deploying the database applications was browser based, this has allowed easy deployment throughout the department without the overhead of installing specific software on end user machines. Browsers have proven capable of displaying all the spatial data types without any serious limitations. However, the client tier can also contain an application running on the client's desktop (thick client) and connecting directly or through HTTP to the database or application server and get access to the data through proper authorization and authentication.

Replication of the same architecture to all the operational units would help for organized storage, efficient management, easy retrieval and speedy and prompt exchange of data between

headquarters and respective operational units either through export-import or direct online input/output processes.

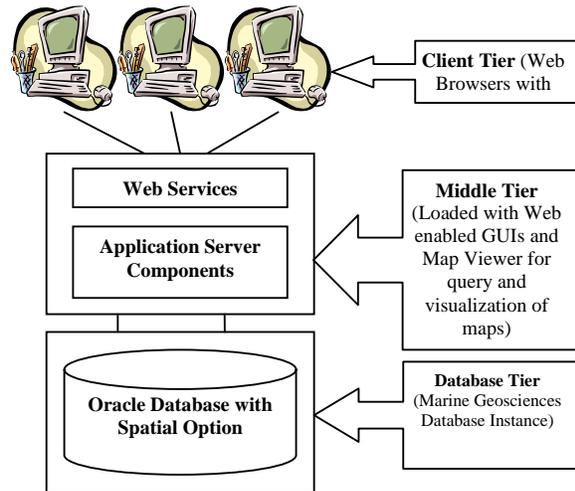
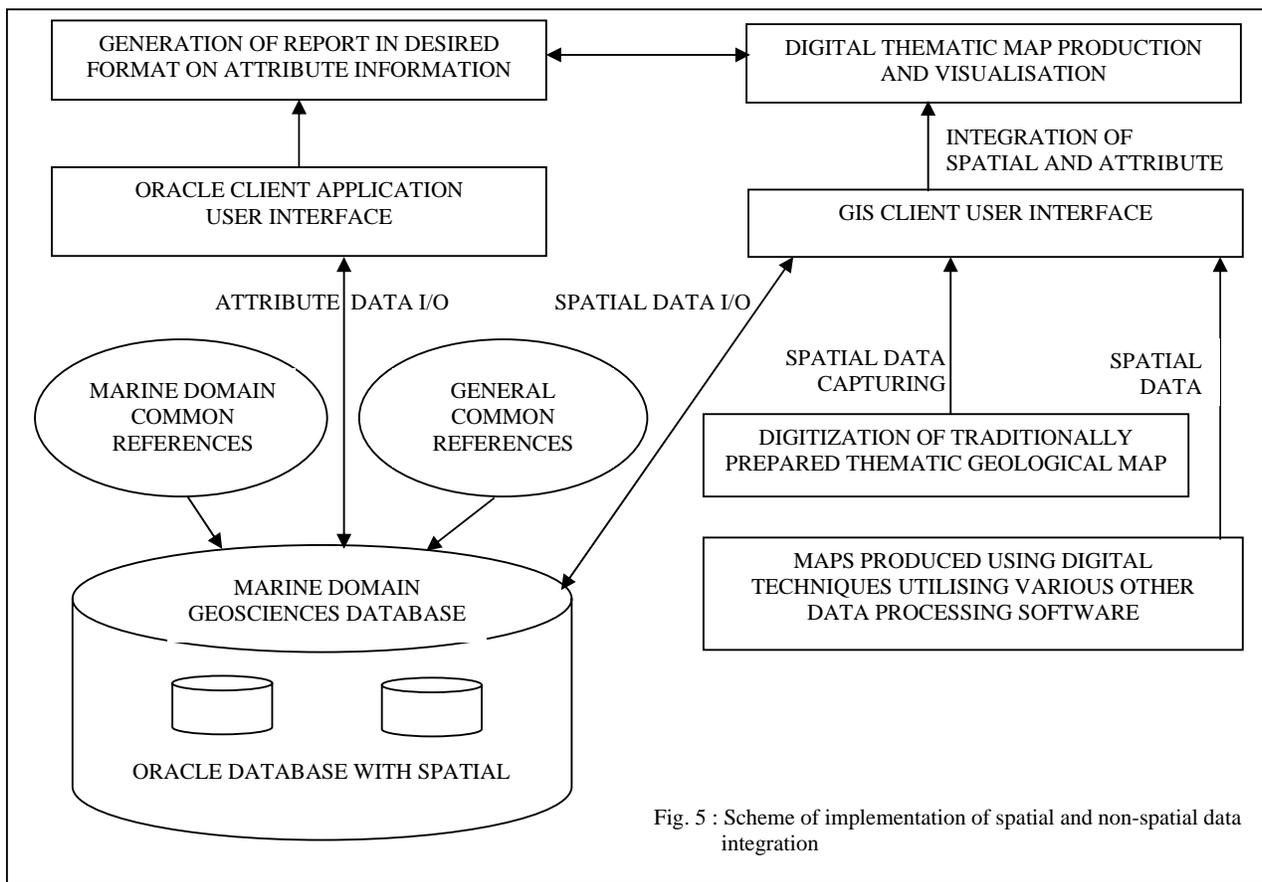
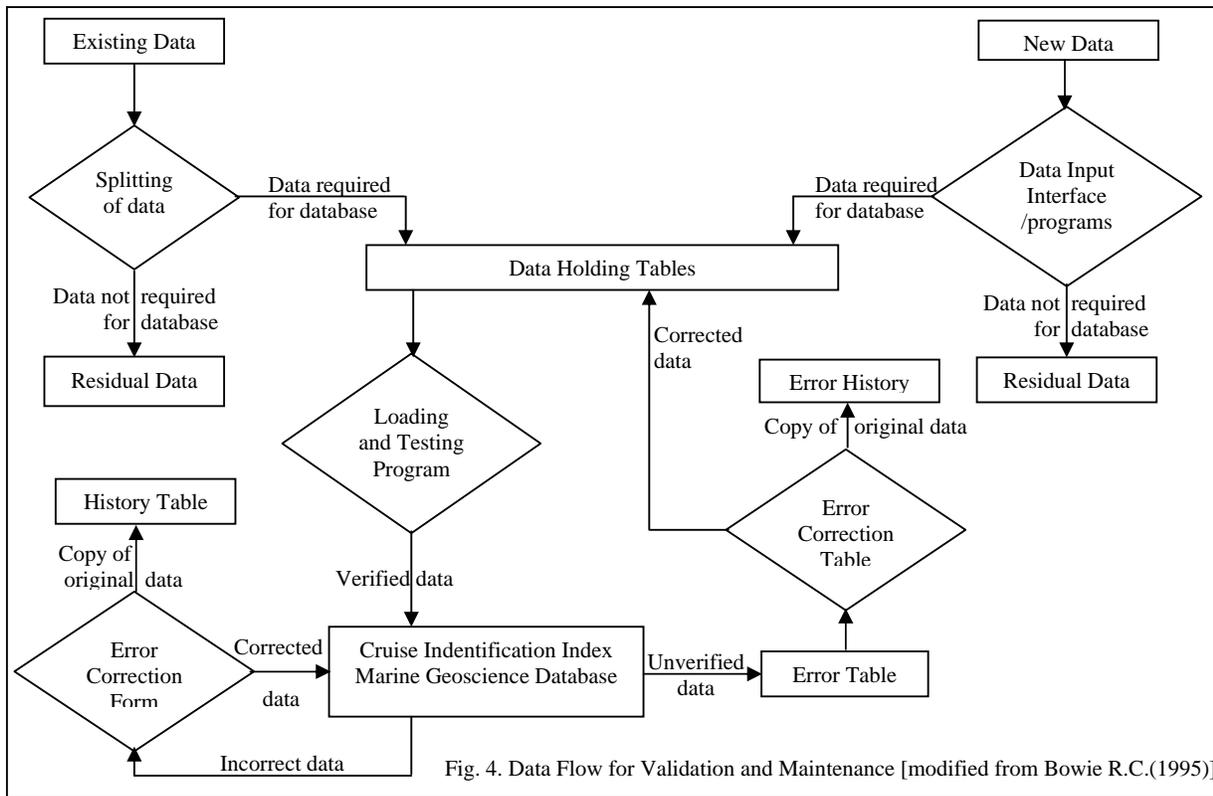


Fig. 3 : Three-Tier Database Architecture Model adopted for Marine Geosciences database

DATABASE INTEGRATION & GIS SOLUTION

Data input, validation and production of desired output is another key area in the creation and maintenance of a digital multi-user database. The data entry into the enterprise database has been provisioned through a) key board input using common GUIs or screen forms which are mapped with different entities defined within the database and b) uploading of existing digital data in ASCII format through a specially designed programme specific application interface. However, screen specific database query with maximum possible combination of data items, update of records and even deletion of records in some selected common GUIs have also been provided. Data validation has been imposed at different levels. In order to bring out a standardization and imposed validation in data input and output and also to increase the objectivity of the attribute data, two categories of lexicon or reference entities were used. The common reference entities of general category are meant for making use across all the domains of GSI whilst the other set of common reference entities of marine geosciences domain were used in order to meet the requirement of marine purposes. Apart from primary key constraints enforced on various entities, a set of database triggers on input center code attribute has been imposed on each entity for validation of data input for respective data input centers. Similarly various rules and validation of data items have also been enforced at the GUI level in order to protect the wrong and inconsistent data entry. In case of data uploading through programs, the validation in the form of standard naming convention of input data files, checking data format error (i.e. use of proper delimiters between any two attribute values, one-to-one mapping of value with item attribute of particular entity concerned, input values whether matched with the domain characteristic of all item attributes of the entity concerned and checking of data content errors as far as possible) and duplicate records have been imposed whilst checking and validation of the errors in data contents is made possible with onscreen consultation of data through launching query for desired set of records using



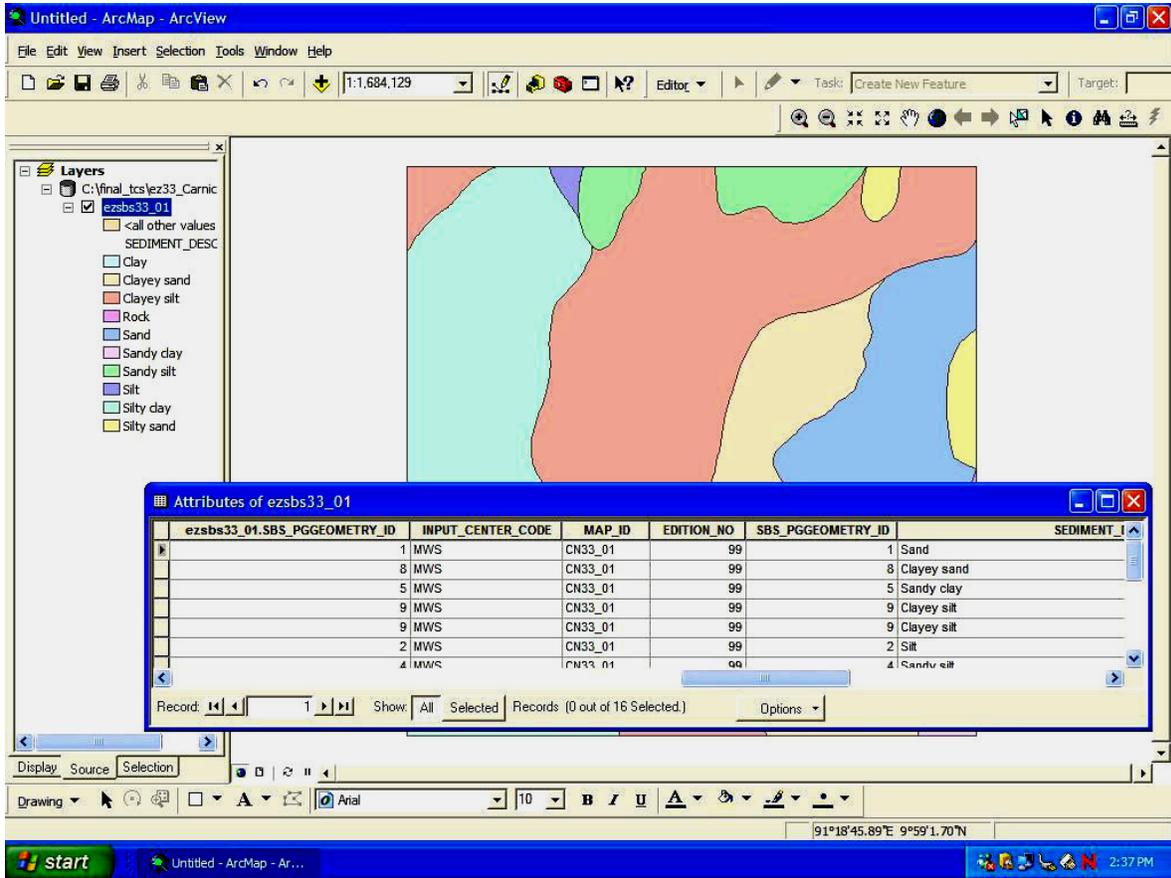


Fig. 6a Seabed sediment distribution map generated in desktop ARC GIS fetching attribute data from Oracle tables

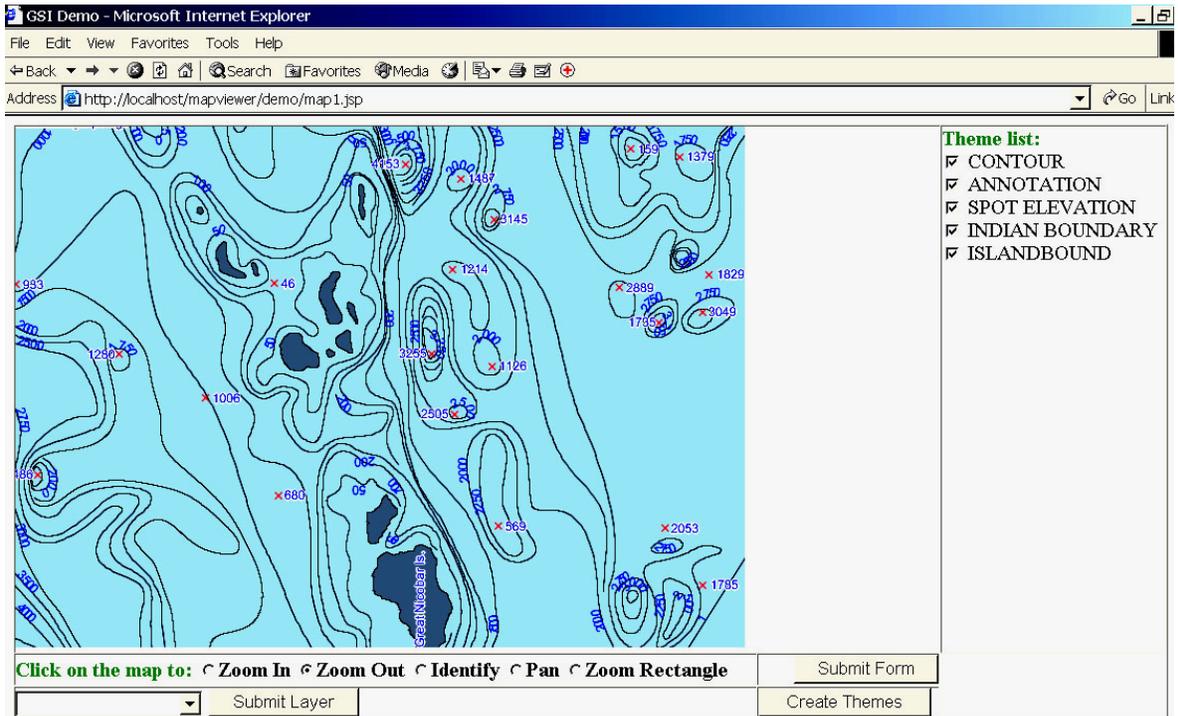


Fig. 6b. Display of bathymetric map through Map Viewer in the browser-based environment fetching data from Oracle Spatial tables

common GUIs. The fig. 4 represents the scheme of data validation being followed at Marine Wing, GSI.

In general most of the spatial and non-spatial data are stored separately into isolated data silos. The non-spatial data are mainly stored in the segregated occurrences and in various file formats such as in Oracle RDBMS and in spreadsheets, etc. This segregated data storage is detrimental to achieving the maximum benefit out of an integrated approach and introduces inefficiencies in the form of data duplication, conflicting data, complex forms of data replication, and inability to meet the scientific and technical interest of GSI. The objective of the integrated solution approach is to migrate from isolated data storage regimes to a centralised integrated enterprise database.

Integration of spatial and non-spatial data can be done either through running script pertaining to data import/export or application level integration or database level integration.

Since the integration at the database level has been proved to be the best level of integration, this approach has been adopted in GSI for having a number of advantages over other methods, as also observed by Betlehem, A.W., 2005. Those advantages are: a) as the integration is at the lowest database level, any system built on top of this database will have full access to the enterprise database. b) data synchronization issues are removed and, all data is used from a centralised source which is free from issues of data duplication and data inconsistency; c) all data become a part of the IT infrastructure and thus centralised backups, hardware upgrades, software upgrades and IT personnel can be implemented and can be made available on a small set of platforms, rather than widely dispersed over a variety of platforms.

The database integration process was focussed on the data and building up of an enterprise data model for all the intended data holdings. In order to accomplish this, focussing on each of the Marine Wing's area of interest a detailed review has been carried out to ensure that all data items are modelled effectively. Fig. 5 shows the scheme of implementation of spatial and non-spatial data integration

The database was created using Oracle 10g as the database platform for the storage of all data within the marine geosciences database management system. The choice was primarily based on the existing use of Oracle within the organisation and the ability of Oracle to store all the required data types, including spatial data. Oracle Spatial option has been used for the storage of all spatial features as it allows access to complex spatial queries and data using standard SQL execution methods from within application code.

Prior to integration, all spatial data were stored, managed and processed with AUTOCAD MAP and Desktop Arc GIS software. Most of the thematic maps in Marine Wing were composed in AUTOCAD MAP and were primarily in the form cartographic presentation. However, a few of which were composed fetching spatial data from personal geodatabase of desktop ARC GIS tools. Presently most of the theme-based geometric feature elements are being digitized using ARC GIS tools and the maps are being composed by making a seamless link with Oracle database for fetching non-spatial information through the process of attribution of data. For example, Fig. 6a showing a seabed sediment distribution map generated by desktop ARC GIS fetching attribute data from related table created in Oracle database. Once a map is composed successfully and declared complete in all respects, it can be

exported to oracle spatial using spatial console or export utility provided by Oracle for organized storage of map geometric feature elements along with all attributes in an integrated way.

It is evident that any integrated system that purports to make use of spatial data, must have facilities to display and edit those data. In general, this is being performed through ARC GIS tools at desktop level. In order to overcome these limitations in the present integrated database system, a humble attempt has been made to deploy a simple to use browser based map viewer along with its all basic functionalities coupled with some application specific custom functionalities (Fig. 6b). The basic map viewer has functionalities required by all map usages are zoom in, zoom out, pan, search, layer control, distance measurement, etc. Besides, attachment of customized tools to the basic map viewer is also possible and is under development that will allow it to add additional feature specific functionality to the map viewing applications. This integrated solution has already received overwhelming success in the Road Management System of Australia as described by Betlehem, A.W. (2005) in his paper dealing with an example of moving from disparate spatial and corporate data repositories into a completely integrated corporate wide database solution incorporating the browser based basic map viewer along with some customized business tools that allowed for the seamless integration of spatial and non-spatial data into a comprehensive management reporting suite. The developed road management system highlights the use of maps to present rich data to a wide variety of users without the need for specialised GIS systems.

CONCLUSION

The creation and development of a database system is a continuous process. The achievement of such an integrated database system starts with conceptualization and construction of an enterprise data model from which the database is created. The success of this database system must be dependent on its fullest use by the internal and external users. This core system will definitely stand as the centerpiece for ongoing activities pertaining to database creation and application development in marine geosciences domain in GSI. The future scope of this system must be driven by the requirements of users and the user interfaces will continue to evolve to further imitate the geological descriptive processes. The Oracle Spatial technology along with deployment of map viewer at the middle-tier, an integrated part of an open GIS technology, would be considered more than adequate for map viewing and editing capabilities which are expanding all the time.

ACKNOWLEDGEMENT

The authors are indebted to the Director General, Geological Survey of India, Sr. Deputy Director General, Marine Wing, Geological Survey of India for giving permission to present the paper in the international symposium. The authors are also thankful to the Directors, Geodata Division, Co-ordination & Technical, Technical Consultancy Services and Publication and Information Division for their constant encouragement and valuable guidance at every stages of preparation of the manuscript. Last but not the least the authors thankfully acknowledge all colleagues for their constant co-operations and reviewers of this paper for taking pains to go through critically

over the manuscript and making valuable comments for improvement of the quality of the paper.

BIBLIOGRAPHIC REFERENCES

Betlehem, A.W. (2005) Enterprise Wide Integration of Spatial Data into Core Business Practices..Proceedings of SSC 2005 Spatial Intelligence, Innovation and Praxis: The national biennial Conference of the Spatial Sciences Institute, September,2005. Melbourne: Spatial Sciences Institute. ISBN 0-9581366-2-9.

Bowie R.C., 1995. Data Management in the National Geological Record Centre, From Giles, J.R.A. (ed.) Geological Data Management, Geological Society Special Publication No. 97, pp. 117-125.

Coats J.S. & Harris J.R., 1995. Database design in geochemistry: BGS experience, From Giles, J.R.A. (ed.) Geological Data Management , Geological Society Special Publication No. 97, pp.25-32.

Elmasri & Navathe, S.B. (1989) Fundamentals of Database Systems. Benjamin Cummings, Redwood City, California.

Garmany J. (Jr.) & Burluson D.K., 2004. Oracle Application Server 10g Administration Handbook, Tata McGraw-Hill Edition. New York.

Giles J.R.A., 1995. The what, why, when, how, where and who of geological data management, From Giles, J.R.A. (ed.) Geological Data Management , Geological Society Special Publication No. 97, pp.1-4.

Giles J.R.A. & Bain K.A., 1995. The nature of data on a geological map, From Giles, J.R.A. (ed.) Geological Data Management , Geological Society Special Publication No. 97, pp.33-40.

Rasmussen K., 1995. An overview of database analysis and designed for geological systems, From Giles, J.R.A. (ed.) Geological Data Management , Geological Society Special Publication No. 97, pp.5-11