

Research of Data Architecture in Digital Ocean

ZHANG Feng^{1,2}, LI Sihai², SHI Suixiang²

(1.School of the Earth Sciences and Resources, China University of Geosciences, Beijing 100083 China; 2. National Marine Data and Information Service, Tianjin 300171 China)

Abstract: The characters of marine data, such as multi-source, polymorphism, diversity and large amount, determine it's different from other data. How to store and manage marine data rationally and effectively to provide powerful data support for marine management information system and "Digital Ocean" prototype system construction is an urgent problem to solve. System planning different types of data, such as marine resource, marine environment, marine economy and marine management, and establishing marine data architecture frame with uniform standard are to realize the effective management of all levels marine data, such as national marine data, the provincial (municipal) marine data, and meet the need of fundamental information-platform construction.

Key Words: Digital Ocean; Data Architecture; Data Warehouse; Data Mart; Metadata

0 Introduction

"Digital Ocean" was born along with the developing strategy of "Digital Earth". It got much attention when brought up. "Digital Ocean" means forming integrated system to recover and expect marine on the base of the data collected during marine investigation, marine monitoring, marine surveillance (including satellite, airplane, shipping, buoy, shore station), and social survey, by the technology of database, GIS, network, to construct digital integrated platform and virtual environment. "Digital Ocean" can promote rationally utilizing and efficiently developing the ocean and keep sustainable marine development through describing the modern ocean and simulating the future ocean. In the world today, major maritime powers, such as USA, Russia, UK, France, Germany, Japan and Canada, are pushing forward "Digital Ocean" information system construction. Interaction between modern ocean and digital ocean will be the basic model of developing and utilizing ocean in 21st century.

To realize the "Digital Ocean" system, to realize the rational and effective management and store different types of marine data, and to meet the application demand of marine fields, a fairly complete data architecture is needed, which must have 7 characteristics: ①Integrity: Store marine data completely and ensure all desired data can be retrieved. ②Rationality: Design rational data structures, and consider reasonably the relationships between different parts and data interchange. ③Independence: Each part of data architecture has its relative independence and performs corresponding function independently. ④Security: Guarantee the security of data when stored, accessed and modified. ⑤Efficiency: Strike the right balance between saving spaces and increasing the speed of response. ⑥Reliability: Ensure the data can be efficiently loaded even the amount of data is large. Guarantee the data which is delivered to the application system is

reliable. ⑦Consistency: The same data in different parts of system must be consistent.

This paper discusses the structure arrangement of the marine data architecture and the design of application system, and case proves the applicability of the data architecture.

1 Background

All branches, operation centers and institutes of China State Oceanic Administration saved a large amount of marine science data and related information after years of survey and collection, including marine hydro-meteorology, marine surface meteorology, marine biology, marine chemistry, marine environmental quality, marine geology, marine geophysics, marine basic geography, marine aviation and satellite RS, marine economy, and marine resources. The data covers all oceans of the globe, and the amount of data is more than hundred billion bytes. The key of management and application marine data is to keep this data accessible, maintainable and secure. The service of this system includes information exchange among institutes of China State Oceanic Administration and other department of marine affairs, data statistics, annual statistics, range statistics, social inquiry, etc.

The characters of marine data such as multi-source, polymorphism, diversity and large amount determine it's different from other data. The multi-source is derived from the different observation techniques, which bring difference of data accuracy and format. Therefore the data structure is complex. The polymorphism means marine data has different format, such as figure, image, text, etc, so the method of processing data is complex. The diversity refers that marine data includes a great variety of subjects, which lead to the complexity of data management.

On the other hand, departments of marine affairs have a large amount of marine data, but this data, which is not organized and utilized effectively yet, is scattered in different units. A lot of data stores still as original form, which makes it difficult to manage, maintain and retrieve. Getting data for specific application analysis accurately is especially difficult, and needs much time and manpower. At present, expertise and traditional statistics are still adopted to analysis, forecast and decision relating to marine, such as meteorology, sea condition, resources, economy and disaster. These non-automatic ways of analysis, forecast and decision greatly affect data utility efficiency. While the technologically backward of marine data management affects the accuracy and actual effect of data, and further influences the efficiency of analysis and decision.

To adequately take advantage of the application service of marine data, as the marine administrative department, China State Oceanic Administration provides basic data support for "Digital Ocean". Upgrading, updating and establishing database of national marine data, developing high quality marine data basic product, and realizing marine data-sharing by the greatest degree are important tasks. Therefore research on architecture of marine data is extremely important.

2 Architecture of marine data

In consideration of bulky architecture of marine data and complex relation among the parts, this study tries ensuring the system structure, application, deployment and maintenance as simple as possible during the process of

design and implementation on the premise of meeting need.

Marine data architecture is divided into 5 parts roughly: primary layer, basic layer, integration layer, product layer, subject layer. As shown in Figure 1.

2.1 Primary layer

Primary layer which contains all data source of system is composed of original data which is distributed in all servers. This original data includes text file, binary data, XML file for describing metadata (derived from tool of metadata input), and data file derived from other application system. This data can be divided structurally into structured data, semi-structured data and non-structured data. Maximum data source of marine data architecture comes from standard dataset, which is semi-structured and non-self-describing data. The standard dataset stores different subject data got from marine field survey. After data is preprocessed and pre-calculated based on file, and data file is corrected and calibrated, this data can be loaded into database. This data source is named semi-structured scientific text data source in this study.

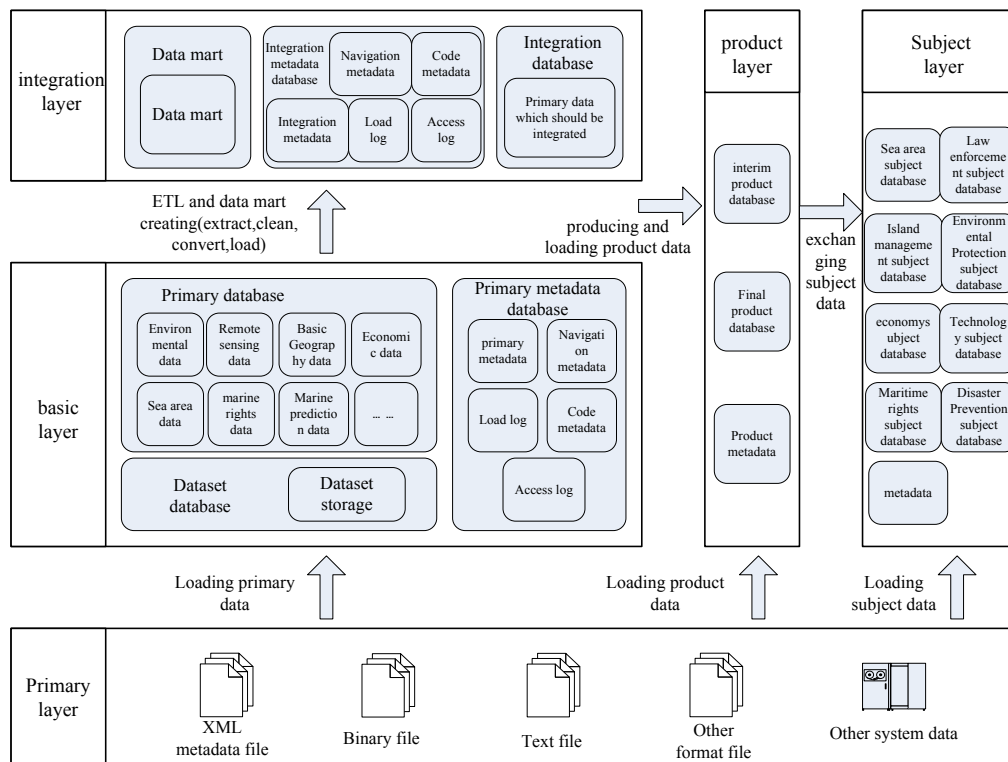


Fig.1 Marine Data Architecture

Structured relational database is an important data source of marine data architecture as well. At present, all units of China State Oceanic Administration have established lots of relational databases for supporting professional data processing and business system applying. All these databases are important data source of marine data architecture. In addition, non-structured hyper text data sources (Web page), such as the data of marine economy, marine rights and interests (including Marine Laws and Regulations) collected from internet, international public data updated at regular period in website, international exchange data, are the important data sources of marine data architecture as well.

Data source of marine data architecture includes real-time data (for example satellite data), vector graphics data

(for example remote sensing data), multimedia data, and other complex format data sources. As shown in Figure 2. This study considers that data source of marine data architecture is a typical data source with various types, which can be classified as web data source (non-structured), text data source (semi-structured) and relational data source (structured).

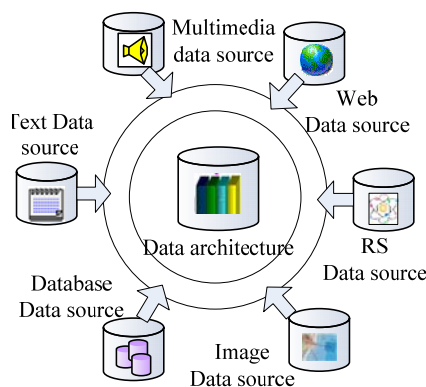


Fig.2 Data Sources of Marine Data Architecture

2.2 Basic Layer

Basic layer includes dataset database, basic database and basic metadata database. Dataset database archives data as file unit. Basic database is a relational form of storage of dataset file. Metadata database stores XML metadata, code data, navigation data, loading log, access log, etc.

2.2.1 Dataset Database

What dataset database stores is quality tested, standard dataset. Dataset of marine science is generally limited by spatial extension, element type, investigation and gathering batch, which is composed by one or several data file. Standard dataset file realizes the centralized archiving management in the form of dataset database, stores in database in the form of object or file pointer, and logically forms a whole with data in basic database, and makes backup and recover of basic database convenient. At the same time, dataset database can realize quickly retrieving dataset file based on metadata.

2.2.2 Basic Database

Basic database is composed of marine data, such as Environmental data, Remote sensing data, Basic Geography data, Economic data, Rights data, Sea Areas data, Marine Environment Forecast data, and other relational metadata. Taking Environment data for example, Marine Environment data is composed of Marine Hydrology data, Marine Meteorology data, Marine Physics data, Marine Chemistry data, Marine Biology data, Marine Geology data, Sea-bed's Topography data, and Marine Geophysics data. The scheme of basic database almost corresponds to format of dataset database, and is the relational storage of dataset file.

2.2.3 Basic Metadata Database

Basic metadata database includes navigation data, value code metadata, loading log of code metadata, access log and basic metadata. Marine basic metadata is the describing data about the data, which is used to describe basic marine

data content, structure, access mode, etc. The use of metadata eliminates the semantic independence and heterogeneity among data resources to a certain degree, and helps realize integration and exchange of data resources. Marine basic metadata is divided into two parts mainly: core metadata and whole metadata.

a) Core metadata: To provide least metadata element set what marine data source needs. Core metadata includes the metadata elements of the following question: what, where, when and who.

b) Whole metadata: To provide obligatory and optional metadata element set what complete marine data source (individual dataset, dataset series, all kinds of marine elements) needs. It can completely define all metadata for identifying, evaluating, excerpting, using and managing marine data.

Data processing professional retrieves data of data layer by core metadata. On the other hand, data processing professional needs check the whole metadata according to data. Some metadata which is relevant to basic layer, such as loading time, information of operator, is common searching criteria and retrieving target employed as well. Not only relational metadata but also XML metadata can be searched by SQL language (the latter must be embed XPath expression in SQL).

Basic metadata includes other metadata, for example, logic organizational structure of basic database is stored in navigation data to fast queries; Value code metadata keeps record of relationship between all field value code of basic database and what it represents for transforming when figure showing; Loading log records the information of data loading in basic layer; Access log ensures tracking who has modified data, what, why, when and how for maintaining basic layer data and guaranteeing database security in some degree.

2.3 Integration Layer

Processed and cleaned basic data is saved in integration layer for providing data service for data analysis and decision (such as OLAP and data mining) of the upper layer. This layer includes integration database, data mart and relevant metadata.

a) Integration Database

The data which organized by investigation method in basic database (such as station data, BT data) is reorganized by elements to form integration database. Basic layer data is reload into integration database after data extracted, data cleaned and data converted by ETL. Although integration database is not organized by subjects but elements, this research still follows the way of data warehouse modeling to design integration database, and defines public dimension table, subject dimension table, and private dimension table to satisfy the multi-dimension of marine data.

b) Data Mart

Data mart is the data view which is defined by data analysis professional, incremental maintained automatically, and created against the specific field, and is the subset of integrated large-scale marine data by elements. The amount of data of data mart is much less than integration database, and data mart has much more pertinence. Data analysis, such as data mining, PLAP, can be done faster and better on the base of data mart. Data mart is created by data mart creation tools, and is automatic updated and maintained by system.

c) Integration metadata database

Basic metadata in basic layer is arranged, and added some background data related to integration database and data mart.

2.4 Product Layer

Read-only and customized product data is saved in product layer, which includes interim product database, final product database and product metadata. Data product which is produced by the above layer, such as the result of basic data processing, graphic result of OLAP processing, is saved in product layer.

Marine product data may be classified in many ways. It can be classified into interim product and final product by the degree of processing, and also can be classified into normal product and service product from the view of application. These two classification methods play an important guiding role in the design of product layer.

Interim product is the data product that has been processed to a certain degree, but there is still room to process and further processing is needed to be final product. For example standard layer data which is derived from the observed temperature and salinity data is interim product, and the geometrically calibrated data obtained from the remote sensing raw data is interim product as well.

Final product is the data product which can be delivered to user directly. Final data product probably is special marine data view or calculation and statistic result. Such as acquisition time, longitude, latitude, the height and direction of wave of all data acquisition bit in certain range of time, longitude and latitude, the maximum, minimum, average of wave height per month and square area. There are many sources of final product, which may be basic database data or integration database data, results of analysis or mining, and external data. Final product database also includes the result of statistic analysis, OLAP and data mining by external system such as data table, graphic and report.

Normal product is the data product formed by daily processing, which has stable pattern, time of production and quantity. This type of product data will be deeply and widely used. Service product means the single-shot, structure stochastic data product, which is to meet the requirement of marine data user. Generally, normal product is managed by relational database and service product is managed by dataset file.

Overall, product layer includes statistical values data product, graphic and image product, and text product. All these types of product have unified XML metadata. For marine data service department, their duties are to increase product category, ensure quality of product and coverage, and other upper layer application. For example, data in “management information system” and “Prototype System” of “digital ocean”, and subject database comes from product data. So product layer is the crucial layer in “digital ocean” system.

2.5 Subject Layer

Marine subject information databases which are based on the basic databases and product databases are several practical-application-requirement-oriented subject databases which are established through comprehensive analysis and fusion. Marine subject databases are real subject-oriented database system. These subjects include Sea Area management, Island Management, Environmental Protection, Warning and Disaster Prevention, Marine Economy and Plan, Marine Supervision of Law Enforcement, Maritime Rights, Marine Technology Management, etc. Marine subject databases are very important in “digital ocean” system, which are the main basis of “management information system”.

Subject database can be classified into two categories by the organization of data: one is basic data mainly based

on basic data, the other is application-oriented business data. The first one includes product data and summarizing data which is extracted from basic database by XML data integration system, and is modeled and organized through data warehouse. The second one is derived from the process of system, and is managed by relational database.

3 Planning of Application System

Enormous data system needs a good information management system. This study plans the application of data system while designing the data system. The application system has bulky structure and complex relationship. The calling relationship among different parts is shown in figure.3. Act up to the principle of light-weight design and realization, except data architecture layer, the upper layers application with the theory of component can be divided into 3 component layers: basic component layer, application component layer and application software layer.

Basic component layer belongs to bottom component layer. All the application of accessing data system must be through basic component. It is composed of three key components which are shared with upper layers: □ Unified authority and access control component can realize the authority and access control of system; □ Access control component monitors all the database access of data system; □ corporate access control component provides a unified way to access all parts of data system and realizes the connection management and optimization of database.

Application component layer is composed of 16 application-oriented components. These components provide component support to all subsystems of application software layer. Application component is the abstract of general or independent software modules of all subsystems. These components are shared by one or more subsystem.

Application software layer is composed of bulky software systems. According to the practical objective and requirement of marine data system management, this study designs the data directory system, data management system, data service system, data application system, professional application system and user authority management system, and several subsystems. Data system is essentially a distributed system, so how do user and application find the data needed from mass data is a challenge. Data directory system realizes the content management and directory access of all parts of data system, and makes data exchange and information query easy. Data management system realizes day-to-day management function of all database layers, such as metadata loading, metadata management, data backing up, data cleaning, ETL and establishment of subject-oriented data warehouse, etc. Data application system fulfills the tasks of retrieving and inquiring of all database layers, displays and outputs the inquire result. Data service system realizes product generating, value calculating, OLAP and data mining, and provides marine information service. Professional application system realizes some high professional application of Marine Science, which depends on manual intervention, calculation and analysis of marine professional staff, or other tools, to complete. User authority management system realizes the user authority and access control of the above software system.

All the data application systems realize the support of basic components and its application component for application layer. Marine professional can apply system function on the base of these components. These basic components encapsulate data access, user accessing control, authority control and access control. These technical details are open to marine professional. Application of basic components provides platform and component support for realizing professional application in the future, and enhances the usability and expansibility of application system to a

great degree.

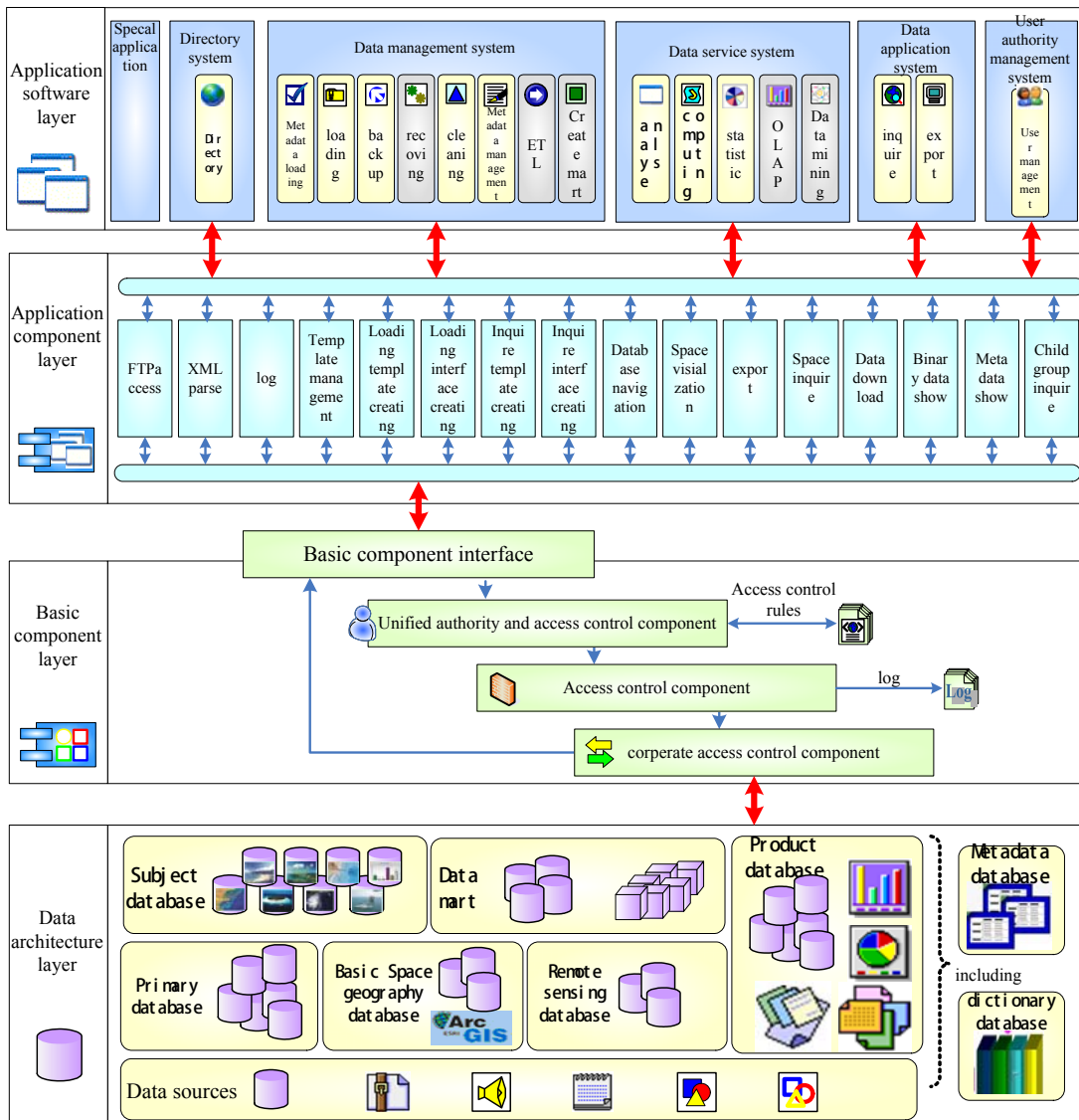


Fig. 3 Relationship of the Parts of National Marine Data Architecture

4 Case Analysis

Data is the lifeblood of the system. If data do not flow in the system, the system is just a castle in the air. The way of marine data flowing in system is very complex, and bulky data exchanges among not only different layers but also inner layer. This study takes the process of gathering marine environment investigation data of provincial node to national node as example to show how data flows in the system.

Step 1: Gathering data is tested by quality control to satisfy the data quality request of data system.

Step 2: Dataset file is converted to relational data schema, and is loaded into basic layer database. Dataset file

itself is filed to basic layer database as well. Metadata extracted from dataset by loading system is automatically loaded to basic metadata database.

Step 3: Data processing staff supplements basic metadata manually.

Step 4: Data processing staff does high data cleaning and duplicate data detecting to further ensure data quality.

Step 5: Data processing staff generates data product on the base of basic database, ensures that this data is loaded into product database of corresponding layer, and fills it with related products metadata at the same time.

Step 6: Basic database data is loaded into integration layer of integration database by ETL software after classified by elements, and is complemented and perfected with metadata.

Step 7: Customize data mart required by customer on the base of integration database, and gather data. Data mart should be stored in product layer as product database if it is data product or in integration layer otherwise. Fill data mart metadata on the base of background such as gathering arithmetic.

Step 8: Provide data service for data mining and OLAP analyzing on the base of data mart.

Step 9: Result of data mining and OLAP analyzing is saved into product layer as product, so is metadata.

Step 10: Subject layer is a type of subject-oriented data. Most marine investigation data in subject database is exchanged to subject layer from product layer by data exchange system.

Step 11: External product, such as dataset file format data and other professional application product data, is loading into product layer directly.

Step 12: New basic data or amended data generated from subject layer should be saved in product database by data exchange system.

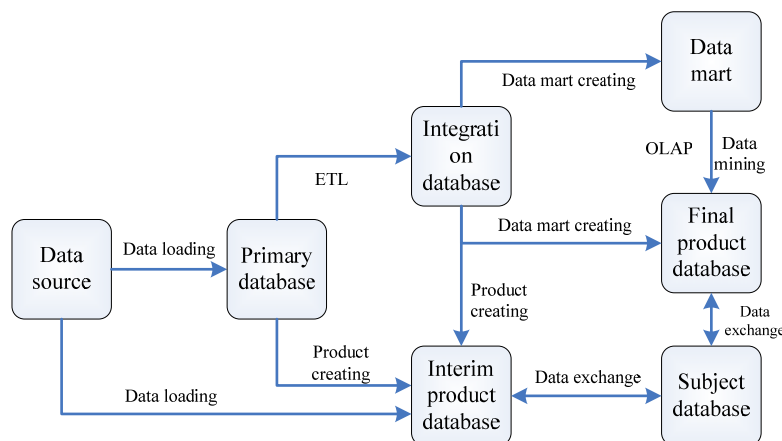


Fig. 4 Data Flow Diagram

5 Conclusion

Unified planning and layout of marine data architecture is studied from the point of database in this paper. National marine data architecture is complicate in structure, but has some regularity.

- a) Data is abstracted from detail from bottom to up, and is integrated step by step.
- b) The organization of data is gradually fine, from mean-oriented, element-oriented to subject-oriented, product-oriented and application-oriented.
- c) All of the data corresponds with its metadata. Unified metadata management is an efficient method to manage complex data.

Data architecture design is to construct unified marine data framework, is the base of effectively managing marine data of all level, and can provide powerful data support for “digital ocean” system construction. In this research, marine data architecture is designed preliminarily at present, and will be revised after deep studying. Internal logical and physical design of all parts will be further discussed.

Reference:

- [1] Core A. The Digital Earth: understanding our place in the 21st century [J]. The Australian Surveyor, 1998, 43(2): 89-91.
- [2] HOU Wenfeng. Tentative Ideas on the Development of Digital Ocean in China [J]. Marine Science Bulletin, 1999, 18(6): 1-10.
- [3] XIA Dengwen, SHI Suixiang, YU Ge. Study on the Techniques of Marine Data Warehouse and Data Mining [J]. Marine Science Bulletin, 2005, 24(3): 1-6.
- [4] XUE Hui fen, ZHOU Yanxia. A Study on the Application of Warehouse Technology to Ocean Environment Information Management [J]. Marine Science Bulletin, 2005, 24(3): 66-71.
- [5] KANG Shouling. The Sea Environment Monitoring Data and Quality Management [J]. Meteorological, Hydrological and Marine Instruments, 2003, (3): 1-6.
- [6] Shi Feng, Jie Song, Xuhui Bai, Daling Wang, Ge Yu. A Web-Based Transformation System for Massive Scientific Data [C]. Proc. of WISE Workshops 2006. 104-114.
- [7] CHEN Jixiang. Study on XML Application in Marine Data Service [J]. Marine Science Bulletin, 2004, 23(2): 46-50.
- [8] TIAN Youqiang, YU Lei, ZHANG Xiaofeng, etc. Design and Implementation of a Marine XML Data Integration System Prototype [J]. Periodical of Ocean University of China, 2005, 35(4): 691-696.
- [9] JI Peng, ZHANG Chenghui, SUN Rupeng. The System of The Real Time marine Data Transmission Network based on XML [J]. Marine Forecasts, 2006, 23(2): 45-50.
- [10] LI Zhe, QIN Qiming, WANG Hongqing, etc. Application of COM Technique in Marine Remote Sensing Multidimensional Dynamic Visualization System [J]. Marine Science Bulletin, 2006, 25(1): 70-74
- [11] HAN Litao, ZHU Qing, HOU Chengyu. Issues Related to the Construction of the Virtual Environments [J]. Marine Science Bulletin, 2006, 25(4): 1-8.
- [12] LI Sihai, ZHANG Feng, LIU Zhenmin. A Prototype System of Operational Processing and Application for Marine Environmental Data [J]. Marine Science Bulletin, 2007, 26(3): 81-86.

Author Information:

Dr. ZHANG Feng
National Marine Data and Information Service
93.Liuwei Road Hedong District Tianjin China P.C.:300171
Tel:022-24013367 Fax:022-24010926 Email:olileo@163.com