Oceanographic Data Management - A National Perspective

Pankajakshan Thadathil

Data and Information Division, National Institute of Oceanography, Dona Paula, Goa 403 004
e-mail: pankaj@nio.org

Abstract

The scope of a Distributed Database Management System (DDMS) with a Centralized Metadata Directory (CMD) is discussed considering the Indian perspectives in ocean data management. The approach presented in the proposed system is capable of dealing with oceanographic data of diverse type and optimizes the efficiency of data query, retrieval and other management aspects. An overview of the present data management that exists at national level is presented and the scope of DDMS with a CMD in managing the national data is examined. The CMD acts as a *'single window'* facility to inform the end-users about the national data warehouse. The issues addressed in the context of the oceanographic data management are common for other geophysical parameters as well. Therefore, the concept of a DDMS with a CMD is a feasible system that could be implemented for any other data management as well.

1. Introduction

A major challenge faced in today's oceanographic data management is the unprecedented growth in the geophysical data gathering. This include conventional as well as satellite observations, numerical weather and ocean prediction model output, climatological and geographic data. Such large volume of multi-disciplinary and diverse data demands an efficient database management system to maximize the data utility. In building such an efficient database system one must consider the following factors: (1) the rapid advancement in information technology, (2) the ever increasing demand on the storage space, and (3) the potential of distributed databases compared to that of conventional centralized database (Silberschatz et al., 1991; Gardiner et al.,1989).

The primary objective of oceanographic data management, like any other data management, is to ensure availability and accessibility of data to end-users and to archive it for posterity. Though the main stream scientists in oceanography are likely to be informed about the data generated by various institutes, a large sector of the end-users are unaware of the details of the oceanographic data warehouses

at national/international level. Collection of oceanographic data is often expensive and it is, therefore, important that the data is managed in a way that maximizes its utility for the benefit of the national and international users. To achieve the above goal there should be an efficient management system to handle large volume of multi-disciplinary oceanographic data (physicochemical, geological, meteorological and biological data) gathered by various organizations.

In a centralized data management system, one center/organization takes the whole responsibility of data management. However, the multidisciplinary nature of the National Oceanographic Programs (*NOPs*) and the existing diversity in the data management systems make the centralized system more complex and therefore, advocate involvements of oceanographic institutions and data centers in the management of the oceanographic data at national level. Such a data management system that shares the responsibilities among various centers through a network of databases is called Distributed Database Management System (DDMS).

In this study, the scope of a DDMS with a Centralized Metadata Directory (CMD) is examined by considering the oceanographic data management in India as a case study. Though the paper addresses the data management aspects pertaining to the Indian national oceanographic data, the issues and concepts reported in the paper are, in general, are applicable to other national oceanographic data and its management. Section 2 provides an overview of the present data management system that exists at the Indian national level. In Section 3, the article discusses scope of a DDMS that involves national funding agencies, core national oceanographic institutions and data centers. Section 4 provides a general discussion on the DDMS along with a Centralized Metadata Directory (CMD), which is followed by conclusion in section 5.

2. Management of oceanographic data – Indian scenario

Vast coastline and abundance of marine resources (living and non-living) are the two primary factors that prompted the establishment of national oceanographic institutions (Table 1) which are directly involved in oceanographic research and development activities. Together, these oceanographic institutions have thrived to improve the understanding of the seas around India and have built a large knowledge base for the Indian Ocean. While the national oceanographic R&D activities have progressed significantly over the last few decades, advancement in oceanographic data management has not gained the required momentum that is essential to deal with the large volumes of multi-disciplinary oceanographic data. The national ocean policy statement (http://dod.nic.in/dodpol.htm) highlights the necessity of a centralized data system for collection, collation and dissemination of information acquired both indigenously and from foreign sources.

Under National Oceanographic Programs (NOPs) large volumes of data is generated for the Indian Ocean by various R&D institutions in ocean sector. Most

of the NOPs justify the data utility, mainly within the scope of the project concerned, in other words 'primary users'. It appears that sufficient importance is not presently given to document the metadata (information about data) and data for the benefits of a wider (secondary) user community. Because of the poor documentation and management of the data and metadata, the user community at present has to make *'extra effort'* to trace the existence of the data and its availability.

Status and format of data/metadata available with national oceanographic organizations or data centers vary significantly. Integrating all these data of varying level of format and accessibility, and providing an efficient system for national oceanographic data management has remained a mammoth task that need to be addressed (Sarupria and Reddy, 2004; Sarupria and Bhargava, 1994). Recent developments in information technology assist in coping with both the diversity and volume of data flows (Tsui and Jurkevics, 1992).

In 1991, DOD (presently 'Ministry of Earth Sciences') made a serious effort to coordinate oceanographic data generated by various national institutes/agencies involved in ocean related R&D activities by establishing the National Ocean Information System (NOIS). While establishing the NIOS, DOD set the following objectives:

(a) To develop efficient data handling and database structures use state-of-the-art strategies to exercise stringent quality control on data and its efficient dissemination for operational prediction systems and to a variety of bonafide users.

(b) To substantially augment data acquisition systems using an optimal blend of satellites, data buoys and oceanographic research vessels.

In order to realize the objectives set for the NOIS program, DOD recognized 13 data centers located in expert institutes engaged in R&D activities in ocean sector and designated them as National Marine Data Centers (NMDCs). The list of the 13 NMDCs is given in Table 1. After the establishment of NOIS in 1991, three more core marine institutes were established under the DOD and each generates voluminous oceanographic data. These three institutes are National Institute of Ocean Technology (NIOT), National Centre for Oceanographic and Antarctic Research (NCOAR) and Indian National Centre for Oceanographic Information System (INCOIS).

2.1 Limitation of the existing system.

Despite the good effort of the DOD and the existence of these NMDCs, NOIS could not fully achieve the set goals of oceanographic data management at national level for the following reasons:

2.1.1 NMDCs and other oceanographic institutions do not have sound data management policies and strict adherence to such policies

Broadly, all NMDCS and other institutions should adopt the Inter-governmental Oceanographic Commission (IOC) data exchange policy that promotes timely, free and unrestricted access to oceanographic data and associated metadata, and aim to maximize the amount of data exchanged without infringing the rights of data originators. However, considering such constrains as cost of data generation, security classifications, etc. NMDCs and other institutions shall develop their own data policies. In the case of some institutes, policies are established but not adequately enforced. Therefore, there should be mechanism to evaluate the level adherence to the data policy.

2.1.2 Lack of proper co-ordination among NMDCs and other oceanographic institutions

Considering the multidisciplinary aspects of the oceanographic observations and the existing diversity in data management policies among NMDCs and other institutes, there is a need for proper co-ordination among NMDCs and other institutes to avoid duplication in all sectors of data management.

2.1.3 Lack of data management component at project formulation level

NOPs or other projects should make proper allowance for management of data not only for the project concerned but also for realizing the wider objectives of data at national and international level.

Project level involvement of NMDCs will allow them not only to identify adequate funding for data management activities but also to monitor the data and information flow within the project enabling 'Beginning-to-End Data Management'.

2.1.4 Most of the NMDCs have not taken the full advantage of IT in data management.

IT has direct impact on data management activities such as data flow monitoring, data archival, database development and updating, networking the NMDCs and other institutes for online accessibility and data mining and archaeology. IT can assist well in coping with the diversity and volumes of data flow (Tsui and Jurkevics,1992).

2.1.5 NMDCs and other core oceanographic institutions do not have a well defined operational metadata directory.

Metadata enables visibility to the data holdings of NMDCs. From the metadata directory, the nodal agencies (DOD) shall get an overview of the data holding and the level of adherence to the data policies followed by each NMDCs and other core oceanographic institutes. This will enable the DOD to identify those NMDCs or institute which are patent defaulters of the national data policies. As long as

DOD continues its funding for NMDCs, it has the right to point finger towards such defaulting NMDCs or institutions.

3. Scope of a DDMS at national level

Distributed Database is a collection of databases that are connected through networking and it could involve different database management systems. Scope of a DDMS at national level stems from two factors:

(a) The diversity of the national oceanographic data management systems.

(b) The necessity to provide an easy access to the voluminous multidisciplinary data of varying format and accessibility to the end-users.

Figure 1 shows schematic illustration of a DDMS with the three categories of data/databases along with a Centralized Metadata Directory (CMD). Based on accessibility, at any point of time, the data existing with any NMDC could be classified to three category; online, offline and non-accessible data. In figure 1, the continuous arrows indicate online access and the broken arrows indicate offline/ classified accessibility.

(a) Online data

Digital and computerized data that are available through a Local Area Network (LAN) or Wide Area Network (WAN) is referred to on-line (public domain) data.

(b) Offline data

Data available in digital format (computerized or hard copy) which is physically accessible, but not on-line is referred to off- line data. Data in analogue or sample form that are not accessible on-line, but having physical accessibility is also referred to off-line data. A large volume of the national oceanographic data falls under this category.

(c) Non-accessible data

Those category of data that do not have either on- line or physical accessibility is referred to non-accessible (classified). Though it is advisable to convert the analogue/samples form of data to digital data and computerize the available hard copies, at any given time, there will be some percentage of the data lying in such primitive forms. In some cases, the data will be computerized in reasonable time after its collection. However, in some other cases, it takes long time for analyzing the samples and converting to digital data. Cases also may exist wherein the primitive form of data maintains statuesque forever.

All NMDCs and other core oceanographic institutes hold all the three categories of data. As depicted in figure 1, in the case of online data/database,

the NMDCs shall provide accessibility to database as well as to meta-database. However, in the case of the other two categories (offline and classified), though databases could not have online accessibility, metadata could be made available online. Therefore, by networking the meta-databases of online, offline and classified databases of NMDCs to the CMD, the DDMS provides a 'Single-Window Facility' to the end-users to know about the national oceanographic data warehouse. This 'Single Window' facility of CMD enables the end-users to access the metadata online without going through the archives of individual NMDCs and institutes, thereby avoiding the 'extra effort'. As seen in figure 1, end-users have online accessibility to the CMD. Even if the data is non-accessible, once the end-user is aware of the data through the CMD, the data could be made available through the user-custodian contact. In some of the NMDCs/institutes, it is likely that data is in hard copy or in analogue form or in samples, or in further primitive form of data. In all such cases, if the NMDCs make the metadata available to CMD, it could provide visibility to those data which are in primitive formats.

Though there are not any prescribed norms to limit the fields in a metadata table, from the user point-of-view, the first requirement is to know what type of data is available, with whom, and in what format. For this purpose, such information should be made available concerning all data/information that exists with the NMDCs and other national marine institutes. On an optimal level, the CMD should have such fields in the metadata records as name of the data, summary, attributes (parameters), geographic location, data currency (dates), quality, custodian, how to access, purpose for which data is collected and whom to be contacted. Once the metadata is known, the user is informed about the accessibility of the data. Applications of a metadata directory are the following:

(a) Provides visibility to details of the data holdings

(b) Enable accurate search and resource discovery

(c) Enable effective management of resources

(d) Accompany a dataset when it is transferred to another computer so that the data set can be fully understood.

(e) Metadata directory adds to the safety of the data as the custodian is held responsible.

4. Discussions

As mentioned in the section 2 of this article, there are many inherent problems among the NMDCs that inhibit data flow and hence good data management. However, a careful analysis of the problems would reveal that the underlying core issue of national oceanographic data management is non-availability of information about each NMDC's data holdings, in other words, NMDC's meta-databases to showcase their data warehouse. As good management depends on good assessment of the resources, it is imperative for NMDCs to have metadata

directories for efficient data management. It is this rationale of the data management that signifies the importance of the CMD in the DDMS

In effect, NOIS concept is similar to the DDMS, wherein the responsibilities of national data management have been shared among the NMDCs and DOD acts as a nodal agency for dissemination of information on the data holdings of NMDCs. The centers also were networked through Microwave Earth Station (MES). Though NMDCs were networked no data was made for public domain by NMDCs. Neither the NMDCs made any serious effort to classify data and make it available for online access nor DOD enforced so. Very few NMDCs have imported data to operational databases that could be made public access data. Though DOD desired to host the NMDCs data holdings information through its site, a well defined metadata (whose format is accepted by all NMDCs) and a mechanism to routinely update the CMD and make it online was not designed within the NOIS system.

Since DOD acts as a nodal agency not only in funding the NOPs but also in coordinating the data management activities of the NMDCs and other data generating agencies, DOD or a DOD designated institute could be the ideal agency to host the Central Metadata Directory. If implemented properly, the DDMS presented in this paper should alleviate most of the existing problems of NMDCs. While CMD enables the required centralization for a nodal agency like DOD for data visibility, the former also acts as a common link that canalize the data flow (online, offline, and classified) in the DDMS.

The data policy of one NMDC may differ from that of the other. However, if all NMDCs could follow a unified national policy for oceanographic data management, it helps to simplify the data flow within the NMDCs. Therefore, it is important for the national data managers to make sincere effort to formulate a national policy for oceanographic data management that is acceptable for all NMDCs. Also, there should be some centralized system/mechanism to closely monitor the adherence of each NMDCs to the national policy.

5. Conclusion

The scope of a DDMS with a CMD is discussed considering the Indian perspectives in ocean data management. By networking the meta-databases of online, offline and classified databases, the DDMS provides a *'Single-Window Facility'* to the end-users to know about the national oceanographic data warehouse. This *'Single Window' facility* of CMD enables the end-users to access the metadata online without going through the archives of individual data centers and institutes, thereby avoiding the *'extra effort'*. The approach presented in the proposed system is capable of showcasing diverse nature of oceanographic/ meteorological data to end users and optimizes the metadata/data retrieval efficiency.

Acknowledgements

The author would like to acknowledge Dr. M. Dileepkumar, scientist from the National Institute of Oceanography (NIO), Goa for going through the manuscript and providing valuable suggestions that helped to improve the quality of the manuscript. Thanks are also due to J.S. Sarupria and G.V. Reddy, scientists from the Data Centre, NIO, Goa for their valuable comments during the preparation of the manuscript.

References

Silberschatz, A., Stonebraker, M. and Ullman, J. 1991, Database systems: achievements and opportunities. Communications (New York).10:110-120.

Gardiner, B.,Bergen, W. and Mathewson, M.1989. New design concepts for the development of meteorological workstations. In Preprints,

Fifth International Conference on interactive Information and Processing Systems for Meteorology, Oceanography and hydrology, pp. 6-9.

Sarupria, J. S. and G.V. Reddy, 2004: Oceanographic Data and Information Network in the Indian Ocean, METOC-2004, 471-477.

Sarupria, J. S. and R. M. S. Bhargava, 1994: New technological trends for ocean data management. In Ocean technology perspectives, PID, New Delhi, Eds. Sushil Kumar, 893-900.

Tsui, T. and A. Jurkevics, 1992: A database management system design for meteorological and oceanographic applications, Marine Technology Society Journal, vol. 26 (2), 88-97.

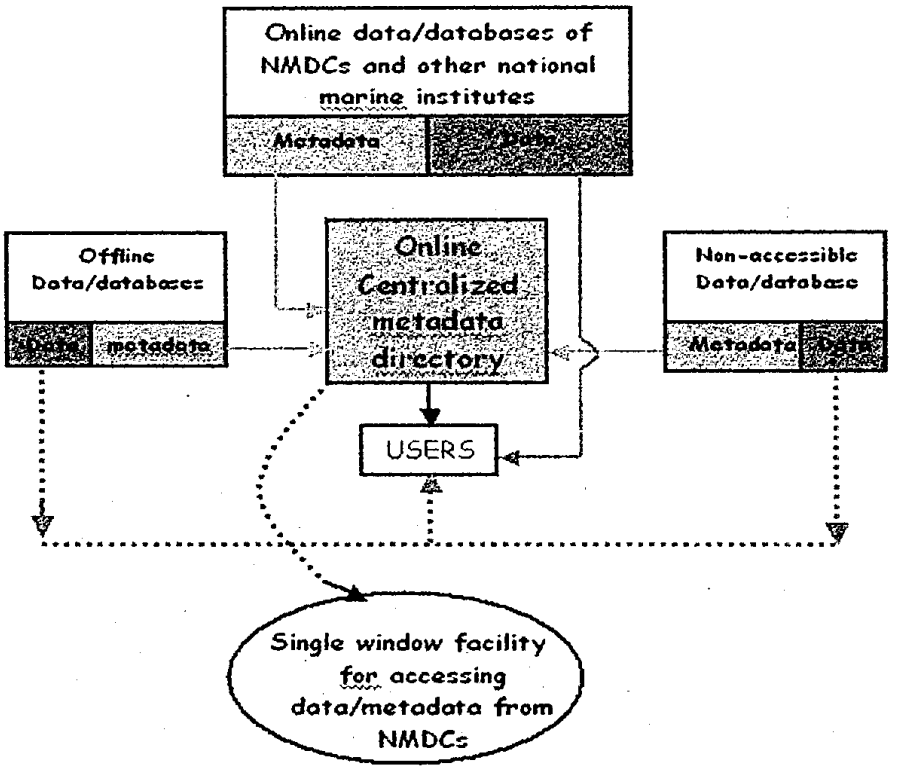

Figure 1. Schematic illustration of a Distributed Database Management System (DDMS) with a Centralized Metadata Directory (CMD). While the continuous arrows indicate online accessibility, the broken arrows indicate offline/classified accessibility.

Table 1. List of National Marine Data Centers (NMDCs)

National Institutes/Organisation	Area
National Institute of Oceanography, Goa	Oceanography
India Meteorology Department, Pune	Marine Meteorology
Survey of India, Dehradun	Sea level
Geological Survey of India, Culcutta	
Central marine Fisheries research Institute, Kochi	Marine Living Resources
Institute for Ocean Management, Madras	Coastal Zone
National Remote Sensing Agency, Hyderabad	Marine Remote Sensing
KD Mlavia Institute of Petroleum exploration, Dehradun	Marine Geophysics
Naval Hydrographic Office, Dehradun	Bathymetry
Fishery survey of India, Mumbai	Offshore fisheries
Central drug Research Institute, Lucknow	Bio-active substances
Central Salt & Marine Chemicals Research Institute, Bhavnagar	Algal Resources & Marine Chemicals
National Institute of Oceanography, Regional Centre, Mumbai	Marine pollution