



OTC 17962

Storage and Distribution of Digital Data from the Gulf of Mexico Seafloor Observatory

Paul E. Murray¹, Cedric Zala², Michael Dunham-Wilkie²

¹*Bureau of Economic Geology, Jackson School of Geosciences, University of Texas at Austin*

²*Barrodale Computing Services Ltd., PO Box 3075, Victoria, BC, Canada V8W 3W2*

Copyright 2006, Offshore Technology Conference

This paper was prepared for presentation at the 2006 Offshore Technology Conference held in Houston, Texas, U.S.A., 1–4 May 2006.

This paper was selected for presentation by an OTC Program Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the Offshore Technology Conference and are subject to correction by the author(s). The material, as presented, does not necessarily reflect any position of the Offshore Technology Conference, its officers, or members. Papers presented at OTC are subject to publication review by Sponsor Society Committees of the Offshore Technology Conference. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of the Offshore Technology Conference is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of where and by whom the paper was presented. Write Librarian, OTC, P.O. Box 833836, Richardson, TX 75083-3836, U.S.A., fax 01-972-952-9435.

Abstract

After the initial deployment of instruments at the Seafloor Observatory in Mississippi Canyon Block 118, the Gulf of Mexico Hydrates Research Consortium plans to install a fiber-optic cable for continuous transmission of data onshore for several of the instruments installed there. Using the planned configuration of instruments, data rates could exceed 180 gigabytes of data per day. The data streams must be relayed to a buffering station onshore and then retransmitted to a second facility for processing, archiving, and reformatting for public access. Both locations require sufficient bandwidth to handle the data, redundancy for backups, and computational power adequate for the prescribed tasks at each location. A software system is also necessary to handle these automated tasks, allow researchers to access subsets of the data from archive, and perform routine analyses for quality control and ongoing research.

A necessary component is the software for the data management and archive system (DMAS). This is being designed under contract by Barrodale Computing Services, and will employ an object relational database management system (ORDBMS) to store and catalog the data, allow users to select subsets of data for analysis, and perform routine processing and analysis of the data of primary interest to the research community.

Two research units of The University of Texas at Austin (Bureau of Economic Geology and Texas Advanced Computing Center (TACC)) have volunteered to provide the facility for storage and public redistribution of the data. The computer resources at TACC are sufficient for both long- and

short-term storage with redundancy, as well as high-speed connectivity through the Internet to allow users to interact with the data and instrumentation. At this time, an institution to act as the data buffering facility has not been identified.

Here we present a summary of the planned DMAS software development and the data management strategy developed for the data collected from the Seafloor Observatory.

Introduction

The Gulf of Mexico Hydrates Research Consortium and the Center for Marine Resources and Environmental Technology are currently developing a multi-sensor Seafloor Observatory to be installed on the continental slope of the northern Gulf of Mexico (OCS Block Mississippi Canyon 118). The aim of this station is to monitor and investigate the hydrocarbon system within the hydrate stability zone of the northern Gulf of Mexico, and to remotely observe changes in the physical and chemical parameters of gas hydrates. The station will be equipped with a variety of sensors that will measure the physical, chemical, microbial and thermal conditions in its local environment and detect temporal changes of those conditions. Major components of the Seafloor Observatory are geochemical instruments, temperature sensors, accelerometers, and an array of hydrophones to collect acoustic data (details can be found in [1]).

After deployment and testing of the instruments in the observatory, the data collection plan is to be divided into two phases. In the initial phase, data will be acquired using data loggers with limited storage capacity. These will be located at the observatory, periodically retrieved and redeployed to download the data. In the second phase, additional instruments will be deployed and the system is to be linked with a fiber-optic cable to an onshore location, enabling continuous monitoring and acquisition of data over extended periods. Table 1 shows the anticipated daily data volumes from each instrument for each phase. In both cases, a comprehensive plan must be in place to allow researchers to access the data.

All these data will need to be archived in an appropriately structured DMAS. We describe a general architecture for the

DMAS development in terms of actors, data repositories and services. Given the advantages of an object-relational database management system (ORDBMS) as outlined in the next section, it is planned that the DMAS will be implemented using an ORDBMS.

Architectural Framework for the DMAS

Figure 1 shows a general framework for the DMAS architecture, with three levels of objects:

- 1) **Actors** are shown as stylized human figures. In this simple depiction there are just two types of actors:
 - i) privileged users;
 - ii) general users (“public users”).
 Initially all users of the Seafloor Observatory DMAS will be privileged, but it is expected that in the future there may be some data that are made available to the general public.
- 2) **Repositories** are shown in boxes with rounded corners. In this model there are six repositories:
 - i) the Near Real Time (NRT) Data Server (holding the data as they exist after a rudimentary level of quality control is applied);
 - ii) the Catalog Server, hosting a database of metadata;
 - iii) the Data Products Server, hosting a database of quality controlled data products;
 - iv) the Data Exploration Server, hosting a database of data extracts within which data- and computationally-intensive processing can be performed;
 - v) the Archive Server, hosting a long-term data archive;
 - vi) the Buffering Station, which holds the data before it is ingested into the DMAS.
- 3) **Services** are shown in oval shaped boxes. Services are provided to the actors, based on data in the repositories. There are three services:
 - i) a Discovery Service, which can be used by both public and privileged users to determine what data are available;
 - ii) a Data Transport Service, which is responsible for delivering requested data to the users;
 - iii) an Authentication Service, which is responsible for authentication of privileged users and determination of access levels.

The single lines in Figure 1 simply link together actors, services, and repositories that have some relationship. No specific directionality to data flow or control flow is implied. A line joins an actor with a service if the actor uses the service; a line joins a service with a repository if the service accesses that repository. In one case there is a line directly joining an actor (privileged user) and a repository (Data Exploration Server), meaning the actor can use that repository directly.

The double-lined arrows represent data flows. Data from the instruments are collected in a buffering station, where some initial quality control is performed and the data are segmented (essentially the continuous time series are chopped

into discrete time series). Following this segmentation, the data are stored in the NRT Data Server, with metadata being copied to the Catalog Server. Once data have been stored on the NRT Data Server, they can be converted to data products with standardized file formats and stored on the Data Products Server, with metadata for these products being copied to the Catalog Server. Once the NRT Data and Data Products are no longer needed “on-line” they are moved to the Archive Server. A final server, the Data Exploration Server, is to be used for *ad hoc* computational- or data-intensive processing without risk of adversely effecting production throughput (i.e., the processing and serving of regular data products).

In this architecture, data and metadata are stored together, in the same rows, in tables in object-relational databases. The object-relational approach offers the following benefits:

- 1) **Complex data type / operation support.** The structure of measurements need not be lost. New data types can be defined to represent complex data objects such as depth versus temperature profiles and time series; instances of these data types can be stored as objects in the database, and methods can be defined to operate on these objects.
- 2) **Metadata / data treated in a uniform manner.** Since the metadata are stored together with the data they describe, the risk they will become out of synch with each other can be eliminated. A method (operation) is written to operate on a particular type of object (E.G., a time series) in a way that creates an appropriate version of the metadata.
- 3) **Data mining opportunities.** This architecture offers an isolated data mine. Privileged users can request large volumes of data-of-interest be copied to a private database on the data mine server for subsequent analysis (using the complex object methods mentioned above) and, perhaps, new data product creation. The data mine can be hosted on a server separate from the Data Products Server in order to avoid hampering the throughput and responsiveness of the Data Products Server.
- 4) **Automatically enforced data integrity.** Since metadata and the data they describe are stored together in the same rows, already-existing database mechanisms for enforcing referential integrity can be exploited.
- 5) **Unlimited ability to subset or aggregate data.** Storing data in an object-relational database rather than in files allows one to postpone the decision on what constitutes a “package”. (Storing data in files requires that one determine how much data to put in each file.) Data are essentially “packaged,” from one or more database rows, on the fly, according to the specific user’s requirements.

Handling the Data Stream

As shown in Table 1, the total daily data stream may exceed 180 gigabytes of data per day after the second phase of installation (with more instrumentation and direct connection between the DMAS and the Seafloor Observatory). Several issues arise when handling such large volumes of data. For example, there must be sufficient bandwidth across any network connections to allow the streaming of this volume of

data. We must have computer resources in place to not only capture the data stream, but have enough storage capacity to make redundant copies. We must have temporary storage available to buffer the data as they come in from the observatory, computing resources available to perform pre-processing and segmentation, and enough spare resources at critical points in the chain to compensate for equipment failures and down time to prevent data loss.

The Texas Advanced Computing Center has proposed a hardware solution with a combination of disk and tape storage and a multi-CPU cluster to act as servers for the data and the DMAS on behalf of the Bureau of Economic Geology, a member of the GOM Hydrates Research Consortium. The University of Texas at Austin (UT) is a partner in the Internet2 Consortium, a high-capacity network capable of handling the data stream. At the date of publication, the GOM Hydrates Research Consortium has not identified an organization which can act as the buffering station, but we recommend this entity be connected to the I2 network to ensure a reliable streaming of the data without loss.

The clustered computer environment is a relatively inexpensive way to provide scalable computing power to the actors utilizing the DMAS. TACC manages and maintains several clustered systems for researchers at UT, some of which are over 1000 nodes in size. They have facilities sufficient for the safe storage and operation on a system capable of serving all the functions of the DMAS. Further nodes can be added to the system to accommodate forward growth of the project, should more computational power be required for additional users or CPU-intensive data analysis.

Hard disk systems for “live” storage (meaning the data are readily user-accessible on a current file system) are also scalable once an initial investment in a controller is made. Depending on the user requirements, several terabytes of disk storage can be made available to host a recent archive of the data (for example, the previous thirty days of data from the current date, amounting to 5.4 terabytes of information).

Long-term Storage and Archiving

In addition to serving data to users, the DMAS also calls for archiving of the raw data and any designated data products. TACC implements a state-of-the-art archiving system using 9940 tape and a robotic tape silo. With such a system, files with a date/time stamp beyond a certain age are automatically archived to tape and moved off live disk storage while retaining an identity on the file system. This way, the files still appear to users of the file system. When accessed, the files are then retrieved from tape and reloaded onto the live disk storage, making the archiving process transparent to the end user (except for the increased time delay necessary for data retrieval).

Conclusions

In order for the Seafloor Observatory to provide data to the research community, a comprehensive plan to store and archive the data in a useful way is under development. This involves a software development plan, the identification of necessary computer resources and expertise, and facilities to house and maintain the systems. The software development plan described briefly here provides a flexible structure for data access and analysis as well as data archive capabilities. TACC and the Bureau of Economic Geology will cooperate to host the DMAS systems and archive the data collected from the Seafloor Observatory.

Acknowledgements

The authors would like to thank Dr. Tom McGee of the University of Mississippi, The Center for Marine Resources and Environmental Technology, and Chris Hempel, Thomas Minyard, Kelly Gaither and Tomislav Urban of TACC for their input and helpful discussions to initiate this project.

References

- [1] Sensor and Data Characterization for the Gulf of Mexico Hydrates Seafloor Observatory. Report by Barrodale Computing Services Ltd. to the Center for Marine Resources and Environmental Technology, University of Mississippi, January 31, 2005.

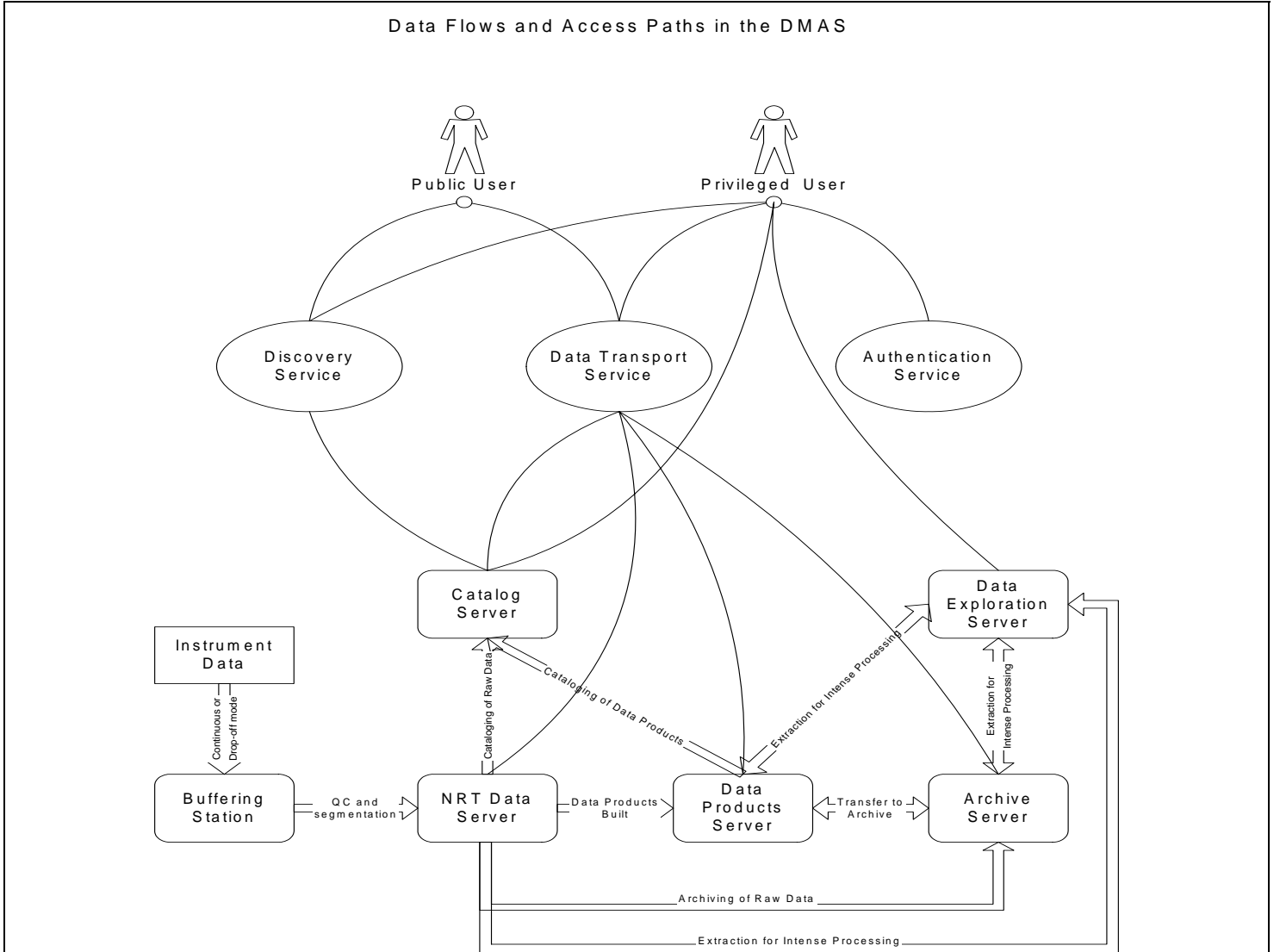


Figure 1: General framework for the Seafloor Observatory data management and archive system (DMAS).

Sensor	Number of Sensors	Average Data Rate (bits/s)	Daily Data Stream (megabytes)
Hydrophone	(a) 12 (b) 80	a) $1.83 \cdot 10^6$ b) $12.2 \cdot 10^6$	a) 19,775 b) 131,833
3D Accelerometer	(a) 6 (b) 30	a) $140 \cdot 10^3$ b) $703 \cdot 10^3$	a) 1,476 b) 7,383
Thermistor	(a) 12 (b) 18	a) $2.81 \cdot 10^3$ b) $4.22 \cdot 10^3$	a) 29.6 b) 44.5
Orientation Sensor	(a) 16 (b) 16	a) 384 b) 384	a) 4.0 b) 4.0
CTD	(a) 2 (b) 2	a) 368 b) 368	a) 3.8 b) 3.8
Current Meter (ADCP)	(a) 1 (b) 1	a) $4.9 \cdot 10^3$ b) $4.9 \cdot 10^3$	a) 51 b) 51
Fluorometer	(a) 2 (b) 2	a) $1.0 \cdot 10^3$ b) $1.0 \cdot 10^3$	a) 11 b) 11
METS Methane Sensor	(a) 1 (b) 2	a) 24 b) 48	a) 0.25 b) 0.49
IR Spectroscopic Methane Sensor	(a) 1 (b) 1	a) 40 b) 40	a) 0.41 b) 0.41
Nephelometer	(a) 2 (b) 2	a) $1.56 \cdot 10^3$ b) $1.56 \cdot 10^3$	a) 16.5 b) 16.5
Mass Spectrometer	(a) 1 (b) 2	a) 228 b) 456	a) 2.3 b) 4.7
Digital Video Camera	(a) 0 (b) 1	a) 0 b) $3.45 \cdot 10^6$	a) 0 b) 37,320
Digital Still Camera	(a) 1 (b) 4	a) $102 \cdot 10^3$ b) $410 \cdot 10^3$	a) 1,080 b) 4,320
Bubble Detector	(a) 1 (b) 1	a) 800 b) 800	a) 8.2 b) 8.2
Pore-Fluid Pressure Sensor	(a) 0 (b) 6	a) 0 b) 1440	a) 0 b) 14.8
Total		a) $2.08 \cdot 10^6$ b) $16.8 \cdot 10^6$	a) 22,458 b) 181,300

Table 1: Estimated data rates and daily data stream totals (at bottom) for the instruments in the Seafloor Observatory. Instrument configurations for the initial deployment with remote data loggers are denoted by (a). In the second phase (b), a fiber-optic link and additional instruments are deployed for continuous data streaming to the DMAS.