

VENUS/NEPTUNE DMAS Examination

Final Report

**Barrodale Computing Services Ltd. (BCS)
Hut R, McKenzie Ave., University of Victoria
Victoria BC V8W 3W2 Canada**

October 18, 2004

BCS acknowledges the contributions of Kent Berger-North (kbergernorth@ccesm.com) in this project, and also thanks the many respondents to our questionnaires, e-mails and phone calls.

Table of Contents

1	EXECUTIVE SUMMARY.....	1
2	INTRODUCTION.....	5
3	REPRESENTATIVE INSTRUMENTATION FOR A CABLED OBSERVATORY.....	7
3.1	LIST OF INSTRUMENTS EXAMINED	7
3.2	GENERAL CONSIDERATIONS.....	8
3.3	SEABIRD CTD MODEL SBE 16PLUS.....	8
3.4	AANDERAA OPTODE 3975 OXYGEN SENSOR.....	13
3.5	GTD-PRO GAS TENSION DEVICE	16
3.6	RDI DEEP WATER WORKHORSE ADCP.....	18
3.7	IMENCO IMDV 3018 DIGITAL VIDEO CAMERA	21
3.8	JASCO AIM-2000 ORIENTATION SENSOR.....	24
3.9	BROADBAND HYDROPHONE SYSTEM	26
3.10	MBARI-ISUS NITRATE SENSOR	28
3.11	D-A INSTRUMENTS OBS-3A OPTICAL BACKSCATTER SENSOR	31
3.12	ASL ZOOPLANKTON ACOUSTIC PROFILER.....	34
3.13	FLOWCAM FLOW CYTOMETER.....	37
3.14	GURALP CMG-1T3 SEISMOMETER	40
3.15	SUMMARY TABLES.....	43
3.16	TABLE 1. CONTACT INFORMATION	44
3.17	TABLE 2. SUMMARY OF QUANTITIES AND DIMENSIONALITY	46
3.18	TABLE 3. SUMMARY OF SAMPLING RATES AND VARIABILITY	48
3.19	TABLE 4. SUMMARY OF INTERFACES, DATA FORMATS AND SOFTWARE	49
3.20	CONSIDERATIONS FOR INCORPORATING NEW INSTRUMENTATION INTO VENUS/NEPTUNE SYSTEMS	51
4	DATA MANAGEMENT PRACTICES	54
4.1	DATA REPOSITORIES	54
4.2	OCEAN OBSERVING SYSTEMS	68
4.3	DATA ORGANIZATIONS WITH OTHER-THAN-OCEANOGRAPHIC DATA	76
4.4	LOCAL SCIENTIFIC DATA ORGANIZATIONS	80
4.5	SUMMARY OF DATA MANAGEMENT PRACTICES	86
4.6	RESPONDENT RECOMMENDATIONS	91
5	RECENT DEVELOPMENTS IN DATA MANAGEMENT	96
5.1	NETCDF	96
5.2	OPeNDAP	97
5.3	DEVELOPMENTS IN METADATA MANAGEMENT	98
5.4	UNIDATA INTERNET DATA DISTRIBUTION	99
5.5	THEMATIC REALTIME ENVIRONMENTAL DISTRIBUTED DATA SERVICES (THREDDS)	100
5.6	WEB COVERAGE SERVERS (WCS)	101
5.7	DATABASE VERSUS FILE STORAGE	101
5.8	DATABASE EXTENSIONS	102
5.9	SUMMARY OF KEY POINTS.....	103
6	ARCHITECTURAL OPTIONS FOR VENUS/NEPTUNE DMAS.....	105
6.1	BACKGROUND.....	105
6.2	GENERAL FRAMEWORK.....	108
6.3	OPTION A: TRADITIONAL APPROACH	110

6.4	OPTION B: ORDBMS APPROACH	112
7	APPENDIX A – FILES PROVIDED ON SUPPLEMENTAL CD-ROM.....	114
7.1	BROCHURES	114
7.2	MANUALS	115
7.3	DATA FILES AND FORMATS.....	116
7.4	SOFTWARE	117
8	APPENDIX B – QUESTIONNAIRES AND RESPONSES	118
8.1	QUESTIONNAIRE 1A: FOR ORGANIZATIONS CONCERNED WITH PERFORMING MEASUREMENTS	118
8.2	QUESTIONNAIRE 1B: FOR ORGANIZATIONS CONCERNED WITH COLLECTING AND DISSEMINATING INFORMATION FROM OTHER REPOSITORIES.	129
8.3	QUESTIONNAIRE 2: FOR ORGANIZATIONS DEALING WITH LARGE DATA VOLUMES	143
8.4	QUESTIONNAIRE 2 RESPONSES	150
9	APPENDIX C – INTERVIEW SYNOPSES.....	176
9.1	PHASE 2 INTERVIEW CONTACTS.....	176
9.2	BC ACTIVE CONTROL SYSTEM	176
9.3	BC LAND AND RESOURCE DATA WAREHOUSE	179
9.4	FLEET NUMERICAL METEOROLOGICAL AND OCEANOGRAPHY CENTER.....	181
9.5	HERZBERG INSTITUTE OF ASTROPHYSICS / CANADIAN ASTRONOMY DATA CENTRE.....	183
9.6	INCORPORATED RESEARCH INSTITUTIONS FOR SEISMOLOGY.....	186
9.7	INSTITUTE OF OCEAN SCIENCES: DATA COLLECTION AND TRANSMISSION PROCEDURES	189
9.8	INSTITUTE OF OCEAN SCIENCES: ARGO SYSTEM.....	191
9.9	JET PROPULSION LABORATORY / PHYSICAL OCEANOGRAPHY DISTRIBUTED ACTIVE ARCHIVE CENTER	192
9.10	NATIONAL OCEANIC & ATMOSPHERIC ADMINISTRATION / NATIONAL ENVIRONMENTAL SATELLITE, DATA AND INFORMATION SERVICE	195
9.11	NATIONAL OCEANIC & ATMOSPHERIC ADMINISTRATION / PACIFIC MARINE ENVIRONMENTAL LABORATORY	197
9.12	PACIFIC FORESTRY CENTRE / NATIONAL FOREST INFORMATION SYSTEM	199
9.13	PACIFIC GEOSCIENCE CENTRE	201
10	APPENDIX D – TABLE OF CONTENTS FOR “DATA MANAGEMENT AND COMMUNICATIONS PLAN FOR RESEARCH AND OPERATIONAL INTEGRATED OCEAN OBSERVING SYSTEMS”	205
11	APPENDIX E – LIST OF INSTITUTIONS HOSTING OPENDAP SERVERS.....	207
12	APPENDIX F – OBJECT-RELATIONAL DATABASES.....	208
12.1	RELATIONAL DATABASES	208
12.2	DRAWBACKS OF RELATIONAL DATABASES.....	209
12.3	OBJECT-RELATIONAL DATABASES.....	210
12.4	BENEFITS OF OBJECT-RELATIONAL DATABASES.....	212
13	APPENDIX G – DATA MINING.....	214
14	APPENDIX H – LIST OF “MUST-READ” MATERIAL	219
15	APPENDIX I – QUALITY MANAGEMENT ISSUES FOR THE OPERATION OF THE DMAS AND DATA QA/QC	220
16	APPENDIX J – GLOSSARY OF COMPUTER TERMINOLOGY	222

1 Executive Summary

This Final Report by Barrodale Computing Services Ltd. (BCS) for the VENUS/NEPTUNE DMAS Examination project describes the results of our investigations into data characteristics and data management practices relevant to VENUS and NEPTUNE.

In Section 3 of the report, a list of 12 sensors provided by UVic was examined in detail to characterize the data in terms of dimensionality, sampling rates, data transfer rates, metadata, operational modes, formats, and transmission practices. The sensors investigated and the corresponding vendors were as follows:

Sensor	Vendor
CTD	Seabird SBE 16plus
Oxygen Sensor	Aanderaa Oxygen Optode 3975
Gas Tension Device	ProOceanus GTD-Pro
Acoustic Doppler Current Profiler	RDI Deep Water Workhorse (300 kHz)
Digital Video Camera	Imenco IMDV 3018
Orientation Sensor	Jasco AIM-2000
Broadband Hydrophone System	IOS – Svein Vagle
Nitrate Sensor	MBARI-ISUS
Flow Cytometer	Fluid Imaging FlowCAM
Optical Backscatter Sensor	D-A Instruments OBS-3
Zooplankton Acoustic Profiler	ASL Water Column Profiler
Seismometer	Guralp CMG-1T 3

Each of the above sensors was characterized in terms of the following

- requirements and limitations;
- physical quantities and the dimensionality;
- typical applications;
- how the data are applied to answer specific questions;
- sampling rates and variability;
- typical and peak data output (transfer) rates;
- metadata;
- operational modes and dynamic configuration;
- data formats, processing software, protocols and interfaces;
- VENUS/NEPTUNE considerations.

Textual descriptions of the above sensors and data are provided in Section 3, and the information is also summarized in tabular form, to allow for convenient reference. In addition, a CD-ROM containing more than 1000 pages of supplemental information (brochures, manuals, data files and formats) was also provided separately.

In Section 4 of the report, we describe the results of our investigations to identify and survey a wide range of organizations involved in the collection, distribution and archiving of data. The aim of this examination was to provide guidance to VENUS/NEPTUNE by considering the practices of, and lessons learned by, other organizations, both oceanographic and non-oceanographic, which have a mandate to collect, process, and disseminate large amounts of scientific data.

The BCS project team identified a number of organizations that we felt would have data management experience that is relevant to the data management tasks facing VENUS/NEPTUNE. The primary requirement for selecting an organization was that they share one or more of the following VENUS/NEPTUNE characteristics:

- they serve as a repository for a large volume of scientific data and associated (complex) metadata;
- they have a real time data acquisition and storage component;
- they have some degree of post-processing and data product creation;
- they have facilities to allow access to the data soon after they are collected;
- they have a wide range of users and user requirements (e.g., data formats, client applications, etc.).

Four categories of organization were identified and are described in the report:

- data repositories;
- ocean observing systems,
- organizations with other-than-oceanographic data, and
- local scientific data organizations.

These agencies included:

- National Oceanic and Atmospheric Administration;
- Marine Environmental Data Service;
- Institute of Ocean Sciences;
- Incorporated Research Institutions for Seismology;
- World Data Center (WDC) System;
- International Oceanographic Commission (IOC) / International Oceanographic Data and Information and Exchange (IODE) Responsible National Oceanographic Data Centers (RNDOC's);
- GOOS (Global Ocean Observing System);
- Herzberg Institute of Astrophysics (Canadian Astronomy Data Centre);
- University of Victoria (ATLAS);
- BC Ministry of Sustainable Resource Management (BC Active Control System and BC Land and Resource Data Warehouse);
- Pacific Forestry Centre (National Forest Information System);
- Pacific Geoscience Centre;
- LANDSAT/RADARSAT;

- Canadian National Data Centre for Earthquake Seismology and Nuclear Explosion Monitoring;
- Environment Canada (Weather Buoy System);
- Pacific Marine Environmental Laboratory (TAO-TRITON and EPIC);
- Fleet Numerical Meteorology and Oceanography Center (USGODAE server and weather modelling);
- NASA EOS (Physical Oceanography Distributed Active Archive System);
- Incorporated Research Institutes for Seismology;
- NOAA NESDIS (satellite programs);
- Natural Environment Research Council (NERC) Datagrid (UK);
- Coriolis project (France);
- Ifremer (France).

These agencies were found through Web searches, reading of literature found on the Web, and through interviews and questionnaires that were conducted/administered during the project. Major sources of contact information were:

- “Enabling Ocean Research in the 21st Century: Implementation of a Network of Ocean Observatories”, written by the US National Research Council Committee on the Implementation of a Seafloor Observatory Network for Oceanographic Research (2003);
- The US National Oceanic and Atmospheric Administration (NOAA) Web site (<http://www.noaa.org/index.html>);
- The Canadian Department of Fisheries and Oceans (DFO) Marine Environmental Data Service (MEDS) Web site (http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Home_e.htm);
- The International Oceanographic Data and Information Exchange (IODE) Web site (<http://ioc.unesco.org/iode/>);
- The Ocean Portal Web site (<http://www.oceanportal.org/>).

Investigation of the Web sites for the individual repositories and programs produced a list of contact people, and these contact people were sent questionnaires. Two questionnaires were sent to data repositories: one targeted at sites in the field that actually collect raw data, and one targeted at sites that receive data from other sources. A separate questionnaire was designed and sent to selected recipients in the other three categories of agencies. Some organizations were sent more than one questionnaire. In addition, more than a dozen phone interviews were conducted with representatives of agencies of all categories. Generally local organizations were given interviews and non-local organizations were sent questionnaires, although some local organizations preferred to fill out the questionnaire, and some non-local organizations preferred to be interviewed. All interviewees were sent the questionnaire prior to being interviewed.

The agencies, contacts, contents, distribution lists, and results of the questionnaires and interviews are detailed in Section 4.

Very recently (May 2004) Ocean.US published the document “Data Management and Communications Plan for Research and Operational and Integrated Ocean Observing Systems”. This plan, which pertains to IOOS, the proposed US implementation of the global initiative GOOS, is outlined in Section 4.2.7 of this report. It is anticipated that VENUS and NEPTUNE will fall, to some degree, under the umbrella of IOOS. The 300-page document is very thorough and presents:

- immediate steps to be taken by associated programs (such as VENUS/NEPTUNE);
- short term and long-term implementation plans;
- background information on data procedures / standards and on existing and emerging data technologies.

A tabular summary of the data management practices revealed in this part of the investigation is given in Section 4.5. Each of the questionnaire respondents and interviewees was also given the opportunity to provide specific data management advice to the VENUS/NEPTUNE project. For convenience, these comments have been gathered together and are presented in Section 4.6.

In Section 5 of the report, we identify and describe a number of technologies that have been applied to oceanographic, meteorological, and seismographic applications. Some of these have become well established; for others, the possible use is still emerging. These technologies include:

- NetCDF;
- OPeNDAP;
- developments in metadata management;
- Unidata Internet Data Distribution (IDD);
- Thematic Realtime Environmental Distributed Data Services (THREDDS);
- the Web Coverage Services OpenGIS standard;
- new database technology, including extensions.

In Section 6 of the report, we present some architectural options for the development of a DMAS for VENUS/NEPTUNE. These are illustrated by several diagrams showing general, traditional and ORDBMS options, and a discussion of each of these approaches.

Finally, several appendices are provided, including questionnaire details and summaries of interviews, and a glossary of technical terms related to VENUS/NEPTUNE DMAS.

In summary, this report provides an extensive compendium of information, background material, experience, and resources on which to base the development of a DMAS for VENUS and NEPTUNE.

2 Introduction

In June 2004, Barrodale Computing Services Ltd. (BCS) was awarded a contract with the University of Victoria to provide the VENUS/NEPTUNE Team with information on data types, current data management practices, data products and analysis/access tools, and other workable data management systems that relate to the scientific disciplines expected to use VENUS/NEPTUNE. This report describes the work that was done in fulfilling the aims of the project.

The first major goal of this project was to examine and define the characteristics of the data for the instruments that will be used by VENUS/NEPTUNE in the near future, and to examine how those data are managed. Specific objectives of this phase were:

1. To define, catalog and characterize a set of identified sensors and instruments.
2. To characterize the corresponding data from the identified instruments.
3. To formulate a list of relevant data issues when new sensors are identified.

Section 3 of this report contains the results of BCS's investigations concerning sensors and data characteristics. The section includes detailed information on the sensors and their data, both in the form of textual descriptions and summary tables. The information in the text is supplemented by a CD-ROM containing documentation for data formats, command and control of instrumentation, sample data files, product brochures, and product manuals (see Appendix A). This information will be critical in assisting the VENUS/NEPTUNE team in its DMAS design.

The second major goal of this project was to examine the data management practices of agencies involved in managing large amounts of data, for both ocean-related and non-ocean-related applications. The aim of this examination was to provide guidance to VENUS/NEPTUNE by considering the practices of, and lessons learned by, other organizations, both oceanographic and non-oceanographic, which have a mandate to collect, process, and disseminate large amounts of scientific data. The specific objectives of this investigation were to contact and obtain information about data and data management practices, using questionnaires and interviews to:

1. Identify relevant national and international repositories of oceanographic data and determine the types and volumes of data stored, and the practices used to manage these data;
2. Identify other organizations and initiatives that handle data from ocean observing systems; examine and document their data management initiatives, status, and plans.
3. Identify five examples of large-scale scientific data management in non-ocean fields; examine and document their overall strategies, practices, successes, and failures, and highlight the similarities and differences in the various approaches.
4. Identify other large data management projects in the lower BC mainland and Vancouver Island, and determine opportunities to benefit VENUS/NEPTUNE DMAS design.

Section 4 of this report describes the process and results of these investigations. Three questionnaires were designed and sent out to more than 40 recipients in ocean-related and non-ocean-related organizations, and nearly 20 responses were received. The questionnaires and responses are provided in Appendix B. In addition, in-person and phone interviews were conducted with 13 representatives (some of whom also returned questionnaires) from a cross section of agencies involved in handling large amounts of data. Detailed notes from the interviews are provided in Appendix C. Also, during this stage, a document describing the implementation plan for the Data Management and Communications (DMAC) subsystem of the Integrated Ocean Observing System (IOOS) was identified, and its Table of Contents is presented in Appendix D.

A third aim of this project was to identify relevant developments in data management and storage that are planned or envisioned. Based on Web searches, information provided by the above sources, and our own experience, a number of recent advances in data management that are relevant to the design and operation of the VENUS/NEPTUNE DMAS were identified. These advances are described in Section 5, and a list of institutions hosting OPeNDAP servers (one of the key developments) is presented in Appendix E.

A final goal of this project was to work with VENUS/NEPTUNE staff to provide a basic framework to guide DMAS design and development. This involved identifying important issues and outlining strategic options for VENUS/NEPTUNE DMAS development. These topics are discussed in Section 6. Further information related to this topic, in the form of a discussion of object-relational databases and data mining, are presented in Appendices F and G, respectively.

Finally, some general information is presented in the last three appendices. Appendix H contains a list of “must-read” material, Appendix I provides a discussion on quality management issues, and Appendix J contains a glossary of DMAS-related technical terms.

3 Representative Instrumentation for a Cabled Observatory

3.1 List of Instruments Examined

This section summarizes the characteristics of the data that is produced by each of a representative set of the sensors to be used in the VENUS and/or NEPTUNE projects. This list, which was provided by UVic, consists of the following sensors and models (or suppliers):

- | | |
|--------------------------------------|------------------------------------|
| • CTD: | Seabird SBE 16plus |
| • Oxygen Sensor: | Aanderaa Oxygen Optode 3975 |
| • Gas Tension Device: | ProOceanus GTD-Pro |
| • Acoustic Doppler Current Profiler: | RDI Deep Water Workhorse (300 kHz) |
| • Digital Video Camera: | Imenco IMDV 3018 |
| • Orientation Sensor: | Jasco AIM-2000 |
| • Broadband Hydrophone System: | IOS – Svein Vagle |
| • Nitrate Sensor: | MBARI-ISUS |
| • Flow Cytometer: | Fluid Imaging FlowCAM |
| • Optical Backscatter Sensor: | D-A Instruments OBS-3 |
| • Zooplankton Acoustic Profiler: | ASL Water Column Profiler |
| • Seismometer: | Guralp CMG-1T 3 |

This section addresses the following issues:

- the physical quantities being observed and the dimensionality of the measurements;
- the nature of the data and how they are applied to answer scientific questions;
- the typical sampling rates, and whether the sampling is continuous or sporadic (and if so, the likely variability in sampling rates);
- the typical and peak data transfer rates to the DMAS;
- any required metadata, whether provided by the instrument or available by other means;
- operational modes of the instrument (polled, streaming, internal logging, duty cycles);
- proprietary data formats, processing software and interfaces;
- calibration and post-processing;
- current practices in transmission mechanisms and whether an (IP) interface to cable is available.

Supplemental documentation is provided separately (on a CD-ROM) where available for:

- data formats;
- command and control of instrumentation;
- sample data files;
- product brochures and manuals.

3.2 General Considerations

The following characterizations of representative instruments intended for deployment in the VENUS and/or NEPTUNE arrays focuses primarily on the physical capabilities of the measuring systems. Specific sensors can often be deployed in differing configurations in order to satisfy the scientific objectives of an investigation. It is important to consider that sampling rates and frequencies, required accuracy and resolution, and other sampling parameters will be determined by the underlying science objectives of the application. In many cases it will be possible to satisfy a wider user community by sampling at the highest frequency and resolution required of the most demanding application, and processing the data (through averaging and other means) to produce additional data products that meet the requirements of other applications.

Not all of the sensors and instruments described in this section are self-contained devices that report calibrated digital information. Some of the devices have analog outputs (for example, 0 to 5 volts) that are processed by an ADC (analog to digital converter) and transformed to a value through the application of calibration equations. The ADCs are typically resident in a related instrument or in a Data Acquisition Module. The range of reporting, accuracy, resolution, and data formats are thus dependent on the combined sensing, acquisition and reporting system.

With the exception of protocols and interfaces, the text of this section will address only issues that are relevant to data and/or software and will not detail hardware and electrical requirements (e.g., voltage supply, pressure casing options, sensor hardware configurations, etc.).

Specific metadata issues are addressed individually for each instrument. However, there are metadata requirements related to instrument clusters, node configurations, deployment, and operational programs. These requirements are not addressed in this section but are essential for overall DMAS operations.

The text of this section has been developed from product brochures and manuals, Web site information, and personal communication with the manufacturers and scientists.

3.3 Seabird CTD Model SBE 16plus

The SBE 16plus SEACAT is a temperature and conductivity recorder (pressure optional) intended for moorings and other long-duration, fixed-site deployments. The SBE 16plus SEACAT offers the improved C, T, and pressure specifications of MicroCATs, and also includes 8 Mbytes of memory and four differentially-amplified A/D input channels with 14-bit resolution. The 16plus is power-efficient: 9 alkaline D-cells will record 380,000 samples of C and T. Conditioned power (500 ma) is available for auxiliary sensors (dissolved oxygen, pH, turbidity, fluorescence, PAR, ORP, etc.); their cabling is simple and reliable with two auxiliary input connector ports. The SBE 16plus uses the same temperature and conductivity sensors proven in 5000 SEACATs and MicroCATs, and

(optionally) a superior new micro-machined silicon strain gauge pressure sensor developed by Druck, Inc. Improvements in design, materials, and signal acquisition techniques yield a low-cost instrument with superior performance that is also easy to use. Calibration coefficients, obtained in computer-controlled high accuracy calibration baths, are stored in EEPROM memory. They permit data output in ASCII engineering units (degrees C, Siemens/m, decibars, salinity [PSU], sound velocity [m/s], etc.). The SBE 16plus sample interval is soft-programmable in one-second increments ranging from 10 to 14,400 seconds. Between samples, the 16plus powers down, drawing only 30 μ amps of current. Data are recorded in non-volatile FLASH memory for 38.4 kbaud upload after recovery. Real time monitoring is practical using the SBE 16plus's 3-wire RS-232C data output. The 16plus is well-suited to networked sensor arrays, where its operation can be triggered by satellite, radio, or hardwire telemetry equipment. Optional RS-485 (2-wire) and inductive modem (1-wire loop) interfaces allow multiple SEACATs to share a simple and robust telemetry cable.

3.3.1 Requirements and Limitations

To mitigate sensor drift of the conductivity cell due to biofouling, the instrument requires an anti-fouling kit of TBT (Tri-Butyl-Tin) available from the manufacturer. Periodic cleaning and calibration of the conductivity cell is required; the interval is determined by *in situ* conditions and rate of fouling. The TBT plugs are consumables and need periodic replacement.

3.3.2 Physical Quantities and the Dimensionality

Parameter	Units	Dimensionality	Resolution	Accuracy ⁽¹⁾	Range
a) Temperature	a) deg C	a) Scalar	a) 0.001	a) 0.005	a) -5 to +35
b) Conductivity	b) S/m	b) Scalar	b) 0.00005	b) 0.003	b) 0 to 9
c) Pressure	c) dbar	c) Scalar	c) 0.002% ⁽²⁾	c) 0.1% ⁽²⁾	c) 0 to 7,000 ⁽³⁾

(1) Accuracy affected by sensor drift and biofouling (especially for conductivity)

(2) Relative to full scale

(3) Pressure rating dependent on selected sensor and casing

3.3.3 Typical Applications

The CTD is the primary instrument used by oceanographers to characterize the physical properties of the water column. CTDs are capable of producing high resolution vertical profiles, measuring conductivity, temperature and pressure. These quantities are used to compute values of salinity, depth, density, and sound velocity. Many models of CTDs are capable of accommodating third-party sensors (transmissometers, fluorometers, nephelometers, gas tension devices, dissolved oxygen sensors, etc.) for coincident data acquisition.

There is a very wide range of applications for CTD data, including (but not limited to):

- water mass characterization (in T/S space);
- water column structure;
- dynamic topography and geostrophy;
- tidal analysis (signals on daily/weekly/annual scales);
- circulation modeling;
- climate modeling;
- environmental assessment (habitat);
- microstructure and turbulence;
- wave phenomena.

CTDs can be deployed as profiling instruments from a ship, in moored configurations, or on autonomous platforms, buoys and floats.

3.3.4 How the Data are Applied to Answer Specific Questions

See above (Section 3.3.3).

3.3.5 Sampling Rates and Variability

The sampling rate of the SBE 16plus is dependent on the mode of deployment: profiling or moored. In a profiling application the data acquisition rate is 4 Hz (the maximum sampling rate of the instrument). In moored deployments the scans are averaged and output at a user-defined interval ranging from 10 minutes to four hours.

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
4 Hz	736	0.25 sec to minutes	Continuous output mode available. Polled output mode available. Specified interval output available.

High resolution profiling CTDs can acquire data at frequencies in the neighborhood of 30 Hz.

3.3.6 Typical and Peak Data Output (Transfer) Rates

The maximum data output rate for the SBE 16plus is limited by its RS232 communications port, which has a default setting of 9600 baud. The data volume is determined by the number of channels reporting, the sample interval, and the output data format. The following table provides typical data volume calculations:

	Configuration			
	Profiling	Profiling with 4 aux. channels	Moored	Moored with 4 aux. channels
Sampling Rate	4 Hz	4 Hz		
Temperature	24 bits	24 bits	24 bits	24 bits
Conductivity	24 bits	24 bits	24 bits	24 bits
Pressure	40 bits	40 bits	40 bits	40 bits
Time	n/a	n/a	32 bits	32 bits
Aux. channels	0 bits	64 bits	0 bits	64 bits
TOTAL (bits/sec)	352	608	480	736

3.3.7 Metadata

A comprehensive list of instrument-specific metadata is provided in the instrument manual (available on the companion CD-ROM – see Appendix A).

3.3.8 Operational Modes and Dynamic Configuration

The SBE 16plus can be operated in profiling or moored mode. It has a comprehensive command set (detailed in the SBE 16plus manual available on the companion CD-ROM) that enables the configuration of (for example):

- acquisition interval;
- number of samples to average;
- output formats.

A shared serial interface is used for both data transfer and instrument control and configuration.

3.3.9 Data Formats, Processing Software, Protocols and Interfaces

Data Formats: The SBE 16plus can be configured to output ASCII data in hexadecimal format or in calibrated engineering units.

Processing Software: The Seasoft-Win32® data processing package provides the following functionality:

Type	Module Name	Module Description
Instrument Configuration	Configure	Define instrument configuration and calibration coefficients (equivalent to SEACON in DOS version).

Data Conversion	Data Conversion	Convert raw .hex or .dat data to engineering units, and store converted data in .cnv file (all data) and/or .ros file (water bottle data).
	Bottle Summary	Summarize data from water sampler bottle .ros file, storing results in .btl file.
	Mark Scan	Create .bsr bottle scan range file from .mrk data file.
Data Processing	Align CTD	Align data (typically for conductivity, temperature, and oxygen) relative to pressure.
	Bin Average	Average data, basing bins on pressure, depth, scan number, or time range.
	Buoyancy	Compute Brunt Väisälä buoyancy and stability frequency.
	Cell Thermal Mass	Perform conductivity thermal mass correction.
	Derive	Calculate salinity, density, sound velocity, oxygen, potential temperature, dynamic height, etc.
	Filter	Low-pass filter columns of data.
	Loop Edit	Mark a scan with <i>badflag</i> if scan fails pressure reversal or minimum velocity test.
	Wild Edit	Mark a data value with <i>badflag</i> to eliminate wild points.
	Window Filter	Filter data with triangle, cosine, boxcar, gaussian, or median window.
File Manipulation	ASCII In	Add header information to a .asc file containing ASCII data.
	ASCII Out	Output data and/or header portion from .cnv file to an ASCII file (.asc for data, .hdr for header). Useful for exporting converted data for processing by non-Sea-Bird software.
	Section	Extract rows of data from .cnv file.
	Split	Split data in .cnv file into upcast and downcast files.
	Strip	Extract columns of data from .cnv file.
	Translate	Convert data in .cnv file from ASCII to binary, or <i>vice versa</i> .
Data Display	SeaPlot	Plot data (C, T, P as well as derived variables and data from auxiliary sensors) from a single file or from multiple files on the same plot (overlay plots). Create TS plots with contours. Plots can be sent to a printer or saved to a file or the clipboard. SeaPlot can plot data at any point after Data Conversion has been run.
Miscellaneous	SeacalcW	Calculate derived variables from one user-input scan of temperature, pressure, etc.

Protocols and Interfaces: The SBE 16plus uses a single RS-232 serial interface for instrument programming and data transfer. The default communications parameters are 9600 baud, 8 bits, 1 stop bit, no parity (9600N81).

IP Interface Available	Interfaces Available	Data Formats	Comments
N	a) RS-232 b) RS-422	ASCII, Hex	a) Processing software available. b) Can output ASCII stream. c) Shared serial port for control and data.

3.3.10 VENUS/NEPTUNE Considerations

To resolve fine structure and detect very small changes in property measurements, the sensors must be calibrated periodically. In particular, the conductivity cell is susceptible to biofouling and needs both cleaning and calibration at an interval determined by the rate of algal growth on the sensing elements.

According to the manufacturer, the VENUS/NEPTUNE applications are really hybrid applications of ship-based and moored CTDs; the issues that must be balanced in this respect include:

- biofouling of moored instruments;
- dynamic response required for high resolution profilers;
- the stability of long-term instrument deployments.

3.4 *Aanderaa Optode 3975 Oxygen Sensor*

The Aanderaa Oxygen Optode is based on the ability of selected substances to act as dynamic fluorescent quenchers. The fluorescent indicator is a special platinum porphyrin complex embedded in a gas permeable foil that is exposed to the surrounding water. A black optical isolation coating protects the complex from sunlight and fluorescent particles in the water. The sensing foil is attached to a sapphire window providing optical access for the measuring system from inside a watertight titanium housing. The foil is excited by modulated blue light, and the phase of a returned red light is measured. By linearizing and temperature compensating with an incorporated temperature sensor, the absolute O₂ concentration can be determined.

3.4.1 Requirements and Limitations

None.

3.4.2 Physical Quantities and the Dimensionality

The Aanderaa Oxygen Optode measures absolute oxygen concentrations.

Parameter	Units	Dimension- ality	Resolution	Accuracy	Range
a) Oxygen conc. b) Air saturation	a) μM b) %	a) Scalar b) Scalar	a) < 1 b) 0.4	a) 8 b) 5	a) 0 to 500 b) 0 to 120

3.4.3 Typical Applications

Monitoring of dissolved oxygen in areas where the supply is limited relative to demand:

- in shallow coastal areas with significant algal blooms;
- in fjords or other areas with limited water exchange;
- aquaculture applications;
- environmental monitoring of industrial activity (mine tailings, dredging activities).

3.4.4 How the Data are Applied to Answer Specific Questions

See Section 3.4.3

3.4.5 Sampling Rates and Variability

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
1 Hz	430 (est.)	1 sec to 255 min	

3.4.6 Typical and Peak Data Output (Transfer) Rates

The maximum data rate of the Aanderaa Oxygen Optode over RS-232 serial interface is 9600 baud.

3.4.7 Metadata

Instrument configuration information and calibration coefficients are stored in the instrument's EEPROM. These include (but are not limited to):

- phase coefficients;
- temperature coefficients;
- foil batch number;
- temperature coefficients for phase to O_2 calculations;
- salinity;
- calibration in air (phase, temperature, pressure);
- calibration in standard solution (phase, temperature, pressure);
- sampling interval;
- offset and slope corrections for I2C output to analog adapter;
- output setting.

A list of instrument-specific metadata is provided in the instrument manual (available on the companion CD-ROM).

3.4.8 Operational Modes and Dynamic Configuration

The sensor can be configured for either polled operation or automatic output at a specified interval.

The instrument has an ASCII command set for interactive control. The types of operations available include:

- reading and writing coefficient information;
- loading and saving configurations;
- sampling;
- calibrating;
- testing;
- setting sample intervals;
- setting output format.

Script files of multiple commands can be sent for sequential execution.

More information regarding command sets, properties, and scripting is provided in the instrument manual (available on the companion CD-ROM).

3.4.9 Data Formats, Processing Software, Protocols, and Interfaces

The Aanderaa Oxygen Optode interfaces directly to the top end-plate of RCM-9 and RCM-11 current meters.

Its RS-232C serial communications configuration is 9600 baud, 8 data bits, 1 stop bit, no parity (9600N81) with XON/XOFF handshaking.

The data output formats are dependent on the selection of the *OUTPUT=* property stored in the instrument's EEPROM. More information regarding data formats is provided in the instrument manual (available on the companion CD-ROM).

IP Interface Available	Interfaces Available	Data Formats	Comments
N	RS-232	ASCII	

3.4.10 VENUS/NEPTUNE Considerations

None.

3.5 GTD-Pro Gas Tension Device

3.5.1 Requirements and Limitations

The GTD-Pro is supported by firmware in the Seabird SBE 16plus CTD. The GTD-Pro will operate with other instruments and data acquisition systems, but the SBE 16plus interface has been specifically developed for automatic control and integration of this instrument.

3.5.2 Physical Quantities and the Dimensionality

The GTD-Pro measures gas tension, or total dissolved air pressure. All pressure measurements can be reported in any standard engineering units. A sensor in the default factory configuration will report in mbar.

Parameter	Units	Dimension-ality	Resolution	Accuracy	Range
Total dissolved air pressure	mbar	a) Scalar b) Scalar	0.0001	0.1	0 to 2000

3.5.3 Typical Applications

The primary application of the GTD-Pro is the long-term monitoring and calculation of air-sea gas exchange rates and fluxes. See Section 3.5.4 for more detail.

3.5.4 How the Data are Applied to Answer Specific Questions

The measurement reported by the GTD-Pro is the sum of the partial pressures (more accurately, fugacity) of all dissolved gases in the water. By far the dominant gases are N₂ and O₂, and to a lesser degree H₂O, Ar and pCO₂. By applying corrections for the minor gases it is possible (along with independent measurements of dissolved O₂) to provide an accurate estimate of dissolved N₂. As dissolved N₂ is essentially abiotic, the differences between the dissolved N₂ and O₂ provide an estimate of the true biological dissolved O₂ utilization and production.

3.5.5 Sampling Rates and Variability

The sampling rate is typically every 60 seconds. This can be changed from 0.1 second to 90 seconds. If the instrument is to be sampled on a duty cycle, then it is typically configured to provide data automatically, 60 seconds after power up. The exact timing of the samples varies a little with the measurement as data returned is an integrated value. Therefore, if the sampling rate is set to 60 seconds, the instrument can provide the data burst anywhere from 58 to 62 seconds. When attached to a CTD, the GTD-Pro is usually polled for a value using a serial command over RS-232. Multiple instruments (each having a unique ID) can be daisy-chained. The data can be automatically returned after a specified interval, or when polled by another instrument.

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
10 Hz	9600	As per CTD	Usually an auxiliary sensor; data rates as per Seabird SBE 16plus CTD (for example).

3.5.6 Typical and Peak Data Output (Transfer) Rates

9600 baud is the standard communications rate over a serial interface.

3.5.7 Metadata

Instrument-specific metadata include serial numbers, calibration coefficients stored in EEPROM, etc.

3.5.8 Operational Modes and Dynamic Configuration

The GTD-Pro can be operated in two modes: polled or automatic reporting after a specified interval.

3.5.9 Data Formats, Processing Software, Protocols and Interfaces

The ASCII data strings are short and typically look like “*01001013.12345”. The initial 5 characters of the data string are the instrument identifier, and the remaining characters represent the pressure measurement.

Proprietary real time data processing routines (Matlab code) are available from the manufacturer to convert the GTD-Pro data along with co-located dissolved oxygen, temperature and salinity measurements to provide dissolved N₂ concentration and saturation levels.

IP Interface Available	Interfaces Available	Data Formats	Comments
N	RS-232 RS-485	ASCII	Control/data interface exists in Seabird SBE 16plus CTD.

3.5.10 VENUS/NEPTUNE Considerations

Some post-processing of the calculated N₂ saturation levels are required to account for calibration adjustments with dissolved oxygen sensor. The GTD-Pro does not generally need post-deployment calibrations as it is stable over long periods.

The Seabird SBE 16plus has on-board firmware compatible with the GTD-Pro.

3.6 RDI Deep Water Workhorse ADCP

3.6.1 Requirements and Limitations

The RDI Workhorse series ADCP can be ordered from the manufacturer with 75 kHz, 150 kHz, 300 kHz, 600 kHz, or 1200 kHz transducers. The frequency of the transducer determines the effective range of the profile and the bin size (or vertical resolution) for a constant signal to noise ratio; lower frequencies have longer range but larger bins. The 75 kHz model has a nominal range of 500-600 m but a large bin size (8-32 m) and therefore a coarse resolution; the 300 kHz model has a nominal range of 126 m (typically 80-90 m) with a bin size of 4-8 m. Larger bins have a better signal to noise ratio but reduced resolution. The 150 kHz model is a special order for the VENUS and NEPTUNE projects and provides a balance between range and bin size. The range (approximately 300 m) is appropriate for the water depths of the planned VENUS deployment.

A summary table of the Workhorse nominal ranges is provided below for transducer frequencies of 150 kHz, 300 kHz, 600 kHz, and 1200 kHz.

			Long Range Mode	
Frequency	Range	Cell Size	Range	Cell Size
1200 kHz	14m	1m	19m	2m
600 kHz	47m	2m	67m	4m
300 kHz	126m	8m	165m	8m
150 kHz	350m	16m	450m	16m

Note: Ranges for 150 kHz are estimated.

3.6.2 Physical Quantities and the Dimensionality

The parameters listed in the table below are the most commonly requested measurements from an ADCP, referenced to Earth coordinates (*versus* beam or instrument coordinates). However, for VENUS and NEPTUNE it is possible that the instrument will be required to operate in a specific high resolution mode, reporting velocities and backscatter in beam coordinates. These parameters are characterized in the companion CD-ROM.

Parameter	Units	Dimensionality	Resolution	Accuracy	Range
a) Velocity	a) cm/s	a,b) Vector	a) 0.1	a) 0.5%	a) 0 to 500
b) Direction	b) deg	array	b) 0.01	b) 2	b) 0 to 360
c) Tilt	c) deg	c) Scalar	c) 0.01	c) 0.5	c) 0 to 15
d) Temperature	d) deg C	d) Scalar	d) 0.01	d) 0.4	d) -5 to 45

3.6.3 Typical Applications

Like the CTD, the ADCP is a widely used instrument in oceanography and has many applications. Typical applications include (but are not limited to):

- tidal analysis (signals on daily/weekly/annual scales);
- mean currents (synoptic and seasonal scales);
- turbulence (high frequency component; scales of seconds to days);
- wave phenomena, especially internal waves (scales of minutes to days);
- relations and interactions among the listed applications;
- seasonal and long-term variation in zooplankton abundance and distribution.

Data products include:

- contour maps of velocity;
- beam velocities and beam variances (especially for turbulence);
- real time currents (that could possibly be animated with .AVI files).

3.6.4 How the Data are Applied to Answer Specific Questions

See Section 3.6.3.

3.6.5 Sampling Rates and Variability

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
2 Hz	50720	0.5 sec	Using PD0 output format with 150 cells at 3170 bytes per scan

3.6.6 Typical and Peak Data Output (Transfer) Rates

The ADCP is capable of a wide range of data output rates, depending on the application. Turbulence studies require the highest resolution data in time and space. In PD0 mode the instrument will generate 6340 bytes per second (assuming 150 cells of 2 m size).

The maximum baud rate on the RS-232 interface is 115200; 57600 should suffice and the instrument default is 9600.

3.6.7 Metadata

The RDI family of ADCPs has an extensive command set for instrument configuration and operation. For metadata, operational modes, dynamic configuration and data formats, please refer to the document “Workhorse Commands and Output Data Format” in the companion CD-ROM.

3.6.8 Operational Modes and Dynamic Configuration

The RDI family of ADCPs has an extensive command set for instrument configuration and operation. For metadata, operational modes, dynamic configuration and data formats, please refer to the document mentioned in Section 3.6.7.

3.6.9 Data Formats, Processing Software, Protocols and Interfaces

IP Interface Available	Interfaces Available	Data Formats	Comments
N	RS-232	ASCII, Binary	a) Processing software available. b) Can output ASCII engineering units. c) Data formats available.

The RDI family of ADCPs has an extensive command set for instrument configuration and operation. For metadata, operational modes, dynamic configuration and data formats, please refer to the document mentioned in Section 3.6.7.

RDI provides a wide range of software for deployment planning, instrument configuration, real time display and data collection. Software available includes:

- PLANADCP software for deployment parameters – WINDOWS;
- WinSC – Plan, Deploy, Recover Data (Self Contained) – WINDOWS;
- WinADCP™ – Data Playback, WINDOWS;
- WinRiver River Discharge/ Shallow Water Acquire and Display Data – WINDOWS;
- RDI Tools – WINDOWS;
- Utilities for Workhorse – DOS;
- River Discharge Plan, Deploy, Acquire Data – DOS;
- Plan, Deploy, Acquire Data with Bottom Track – DOS;
- Navigator Plan, Deploy – DOS.

3.6.10 VENUS/NEPTUNE Considerations

The sampling mode and data output frequency required of the high resolution turbulence experiments requires that the instrument be configured in a special mode (MODE 12) for maximum data output (2-4 Hz). The application also requires that the instrument report values referenced to beam coordinates. These output formats are not compatible with real time UV current velocities in earth coordinates (as used and reported by NOAA, for example). However, both applications can be satisfied by feeding the MODE 12 output to a process running RDI software, which in turn can compute the more commonly used data products. It is not possible to obtain the required high-resolution data if the instrument were to be set up in a mode that reports ensemble-averaged data (for example), as information is lost in the transformation. Careful selection of acquisition modes and parameters should be examined to ensure that information essential to more demanding applications is not discarded.

Calibrations are not typically required for the transducers, although the compass may require swinging. The sensing elements are not particularly susceptible to biofouling.

3.7 Imenco IMDV 3018 Digital Video Camera

The Imenco IMDV 3018 Digital Video Camera was identified as a possible sensor to be used in the VENUS and NEPTUNE projects. Due to the limited amount of information available at the time of this report, this section will address the applications and general issues associated with data from digital still cameras, digital video cameras, and HDTV cameras deployed on IP networks.

Other still digital cameras investigated include:

- Imenco SDS3040;
- C-Map Systems Cyclops;
- Kongsberg OE14-208.

3.7.1 Requirements and Limitations

HDTV (high definition television) video signals generate very large data volumes in real time, demanding more bandwidth than that of all of the other identified sensors combined (on a per node basis).

HDTV over IP is an emerging technology that is primarily in the research domain. The University of Washington has been a leader in developing the technology and has demonstrated successes since 1999. Commercially, HDTV signals are typically broadcast via satellite and the Internet is not a popular signal medium due to the bandwidth requirements. Lower resolution digital video is more commonly used in Webcasting.

HDTV signals will require a camera, HDTV encoders, network interfaces, HDTV decoders, streaming video services, and large data storage devices.

3.7.2 Physical Quantities and the Dimensionality

Parameter	Units	Dimension- ality	Resolution	Accuracy	Range
Imagery	n/a	n/a	N/a	n/a	n/a

3.7.3 Typical Applications

Applications for HDTV may include:

- monitoring of benthic communities;
- monitoring of hot vents;

- monitoring of species interaction;
- species identification;
- outreach services.

Applications for digital still cameras include:

- monitoring of benthic communities;
- monitoring of node and instrumentation.

3.7.4 How the Data are Applied to Answer Specific Questions

See Section 3.7.3.

3.7.5 Sampling Rates and Variability

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
a) streaming b) 1 Hz (est.) c) 1 Hz (est.) d) 1.7 Hz	a) 1244 Mbps b) c) 1.5 Mbps d) 100 Mbps	a) streaming b) c) 15 sec d) 0.5 sec (est)	a) HDTV b) Imenco SDS3040 c) Kongsberg OE14-208 d) C-Map Cyclops

3.7.6 Typical and Peak Data Output (Transfer) Rates

Data rates are dependent on the resolution of the image, the frame rate, and whether the signal is interlaced or progressive scan. The following table summarizes data rates for varying resolutions, frame rates, and interlacing:

Resolution	Data Rate
640x480x30i	184Mbps
720x480x30p	207Mbps
1280x720x30p	553Mbps
1280x720x60p	1106Mbps
1920x1080x30i	1244Mbps
1920x1080x30p	1244Mbps

Demonstration projects include:

- a demonstration at Portland SC99 involved 5 channels of 200+Mbps HDTV (1.1Gbps) transmitted over 2GbE NICs and generated 54.7 terabytes in 90 hours;
- 5x270 Mbps HDTV/POS transmission over 300 km;
- 1.2 Gbps TCPIP between desktops over 300 km;
- 1.5 Gbps HDTV/ATM transmission over 500 km.

Compression can significantly reduce the bandwidth required. For example,

- a 720p (progressive scan), 60 fps, 1280x720, 20 bits/sample transmission, or a 1080i (interlaced) 30 fps, 1920x1080, 20 bits/sample media stream is 1.485 Gbps uncompressed;
- the same transmission, compressed, requires 19.4 Mbps.

In October 2001, NTT Japan announced the world's first successful transmission of 1.5 gigabit/second uncompressed High-Definition TV signals over an Internet link between two sites 20 kilometers apart. Soon after, American researchers working independently used home-grown technology to demonstrate the same capability over longer distances, first between Seattle and Denver, and then from Seattle to Washington, D.C.

The first public demonstration of the U.S. project was at the "SuperComputing 2001" (SC01) conference held in Denver. In January 2002, the same system was used to stream real time uncompressed HDTV across the country to a meeting of researchers in a Washington D.C. hotel via the Internet2 national backbone. HDTV content was originated at the University of Washington's "ResearchChannel" facilities in Seattle. HDTV comes in many quality levels. Consumer-grade HDTV is compressed to less than 20 Mbps, whereas "studio quality" HDTV feeds require over 200 Mbps. Fully uncompressed HDTV signals of the highest quality require 1.5 Gbps. One motivation for trying to send uncompressed HDTV signals is to avoid the latency (delay) that occurs whenever real time data streams are compressed. Reducing latency is critical for interactive applications, such as video conferencing.

3.7.7 Metadata

No information.

3.7.8 Operational Modes and Dynamic Configuration

No information.

3.7.9 Data Formats, Processing Software, Protocols and Interfaces

The protocol for transmitting the video information over IP at the SC01 conference was defined by engineers from ISI, Tektronix, and UW, using the IETF standard Real Time Protocol (RTP) specification as a foundation.

The current 292M standards are provided in the supplemental information on CD-ROM:

"RTP Payload Format for Society of Motion Picture and Television Engineers (SMPTE) 292M Video"

IP Interface Available	Interfaces Available	Data Formats	Comments
N	a) IEEE-1394 b) Composite c) Y/C		a) Requires IEEE-1394 to IP converter.
Y	a) Ethernet (data); RS232 (control)	JPG, RAW	a) Imenco SDS3040
Y	b) USB (data); RS232/485 (ctrl)	JPG, RAW	b) Kongsberg OE14-208
Y	c) 100T (data); RS485 (control)	JPG, RAW	c) C-Map Cyclops

3.7.10 VENUS/NEPTUNE Considerations

- The camera fields of view will require artificial lighting, which may have impacts on the observed ecosystems, particularly when lights are used for extended periods of time.
- At the time of this report, the engineering contractor (OceanWorks) was investigating the possibility of analog video capability as an alternative to digital video. A definitive answer had not yet been provided.
- SMPTE-292M is the widely used interconnect standard for HDTV equipment.
- An IP network can emulate the SMPTE-292M circuit.

3.8 *Jasco AIM-2000 Orientation Sensor*

The AIM series of underwater instrumentation packages provides a system for measuring the tri-axial orientation of any submerged platform or device, transmitting the data via cable as a digital data stream.

3.8.1 Requirements and Limitations

None.

3.8.2 Physical Quantities and the Dimensionality

Parameter	Units	Dimensionality	Resolution	Accuracy	Range
a) Magnetic Dir.	a) deg	a) Scalar	a) 0.1	a) 1 (level)	a) 0 to 360
b) Inclination	b) deg	b) Scalar	b) 0.1	b) 0.5 (to 30°)	b) -60 to +60
c) Pressure	c) mbar	c) Scalar	c) 0.1	c) 0.1% f.s.	c) variable
d) Temperature	d) deg C	d) Scalar	d) 0.1	d) 0.2	d) -40 to +85

3.8.3 Typical Applications

The intended application of the AIM-2000 for the VENUS project is to provide a measurement of platform heading, tilt (or change in platform tilt) and supplemental

pressure and temperature data. Other applications include determining the real time attitude of towed bodies and aiming sonar transducers.

3.8.4 How the Data are Applied to Answer Specific Questions

The data are not intended for scientific experimentation but for indicating physical orientation of node hardware. This information may also provide additional positional metadata for other sensors.

3.8.5 Sampling Rates and Variability

Maximum reporting rate without stale data is 8 Hz. Typical reporting rate is once per minute, and once per hour is possible.

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
8 Hz	6400	1 sec to minutes	

3.8.6 Typical and Peak Data Output (Transfer) Rates

The maximum output frequency of the system is 8 Hz, with NMEA sentences on the order of 100 bytes. Total data volumes are on the order of 800 bytes per second (with all auxiliary sensors). The instrument is capable of higher baud rates than the specified NMEA standard of 4800.

3.8.7 Metadata

Configuration data and calibration coefficients are available through the use of instrument software available from JASCO.

3.8.8 Operational Modes and Dynamic Configuration

The instrument can be operated in an automatic mode reporting at a specified interval. A software interface is available for interactive configuration and control of the instrument.

3.8.9 Data Formats, Processing Software, Protocols and Interfaces

Data are output in NMEA-0183 ASCII strings (or NMEA “sentences”). Among the sentences output by the device are:

- the standard NMEA HDG sequence;
- IIXDR (unspecified transducer) sentences for pitch etc.; data formats are provided in the companion CD-ROM;
- depth;

- temperature;
- a proprietary status register PJAS that contains a status code (in hexadecimal format) with state, out of range information, error codes, etc.

A full software package is available for use by VENUS and NEPTUNE.

IP Interface Available	Interfaces Available	Data Formats	Comments
N	a) RS-232 b) RS-485	NMEA-0183	a) Format available for proprietary data sentence. b) Shared serial port for control and data.

3.8.10 VENUS/NEPTUNE Considerations

With respect to a virtual IP connection, the bi-directional communications capability is not perfect as tested with third-party 232/IP converter. The tested converters do not implement with 100% compatibility the RS-232 functionality. (Note: this is not unique to the AIM-2000 but is an artefact of the converter).

3.9 **Broadband Hydrophone System**

The broadband hydrophone is a prototype custom development system designed by Svein Vagle at IOS. The 200 kHz instrument is controlled *in situ* with a Windows-based computer and can be operated remotely via IP interface.

3.9.1 Requirements and Limitations

None.

3.9.2 Physical Quantities and the Dimensionality

The measurement property is pressure amplitude. The signal frequency is 200 kHz, and is digitized with a 16-bit A/D converter. The maximum acoustic bandwidth is 100 kHz (taking Nyquist frequency into account).

Parameter	Units	Dimensionality	Resolution	Accuracy	Range
Pressure amplitude	Decibels	Scalar array	*see text	*see text	0 to 65535 (16 bit)

3.9.3 Typical Applications

Typical applications of the broadband hydrophone include:

- marine mammal tracking;
- algorithm development for acoustic array steering;
- ambient sound monitoring (surface wave generation phenomena);
- monitoring the effect of shipping noise on marine mammals.

3.9.4 How the Data are Applied to Answer Specific Questions

See Section 3.9.3.

3.9.5 Sampling Rates and Variability

The 200 kHz transducers provide a signal with 100 kHz acoustic bandwidth. Marine mammal tracking typically requires 10-12 kHz (maximum 15-20 kHz) acoustic bandwidth. The exception to this is the Dahl's porpoise, which can vocalize in the 100 kHz range.

The bandwidth can be reduced at source for a hydrophone, or can be shared among hydrophones connected to the same instrument. For example, a 3-hydrophone system would have to share its bandwidth, with each hydrophone allocated 67 kHz.

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
200 kHz	3,200,000 (3.2 Mbit/s)	1 Hz	200 kHz data rate with 16 bit A/D sampling per hydrophone

3.9.6 Typical and Peak Data Output (Transfer) Rates

Peak data rates are 3.2 Mbit/sec per hydrophone (200 kHz with 16 bit resolution)

3.9.7 Metadata

A synchronized time signal is required for the acoustic signals. Software gain settings, bandwidth, number of hydrophones, calibrations, and array geometry are required metadata fields.

3.9.8 Operational Modes and Dynamic Configuration

The broadband hydrophone operates continuously in a streaming mode. The instrument is controlled via batch file configuration. Interactive configuration and control of the instrument is possible by sending batch command files.

Additionally, neural network routines are used for detecting from data stream whales and other marine mammals. This software would enable event triggering (possibly changing the sampling or recording regime) and is available for use by VENUS and NEPTUNE.

3.9.9 Data Formats, Processing Software, Protocols and Interfaces

There are several options for handling and processing the high volume data produced by the hydrophone. These include:

- logging the full data stream to off-line storage;
- buffering a specified time period of data in on-line storage;
- enabling data streaming to end-users;
- automatically reducing the signal to 3-10 kHz as a data product for Web streaming applications;
- calculating spectra and storing them on-line as data products. A 5-second raw data stream requires 2MB storage; a 4096 point FFT will result in 2048 complex numbers requiring 16 kB (uncompressed) resulting in a factor of 128. Five second windows may be appropriate for mammal tracking; shorter windows (0.5 to 1.0 sec) may be required for surface processes.

File formats include VOC and WAV. WAV files are appropriate for Internet distribution as it is a well-supported standard. VOC file formats (or similar) are required for more demanding applications, as VOC supports block types (e.g., GPS, time, headers, etc. embedded in the file).

IP Interface Available	Interfaces Available	Data Formats	Comments
Y	IP (Windows PC at source)	Binary, .WAV	

3.9.10 VENUS/NEPTUNE Considerations

A synchronized time signal needs to be embedded in the data stream for many applications. Breaking wave processes in particular require these signals for computing time delays between hydrophones in array configurations. While the VENUS/NEPTUNE design documents indicate that a time signal will be available, how this signal is embedded in a format that is decipherable by the hydrophone processing system has yet to be determined.

3.10 MBARI-ISUS Nitrate Sensor

The MBARI-ISUS (commonly referred to as ISUS) is a spectrophotometer that uses ultraviolet absorption spectroscopy techniques to measure *in situ* dissolved chemical species. The sensor is a chemical-free, solid state instrument that offers easy, accurate, real time and continuous nitrate concentration measurements.

Features of the ISUS include:

- real time nitrate measurements;
- analog output port for easy integration with CTD and other sensors;
- high sample frequency;
- high data precision and accuracy;
- spectral analysis software.

3.10.1 Requirements and Limitations

The ISUS nitrate sensor can be equipped with a copper antifouling guard which has proven to be effective for periods up to and exceeding 6 months. The instrument requires servicing to clean the biofouling guard and replace the small nitex sleeve that is mounted inside the guard.

The ISUS sensor has a UV deuterium lamp with a 1000 hour lamp life before a lamp replacement and calibration is required.

3.10.2 Physical Quantities and the Dimensionality

Parameter	Units	Dimension-ality	Resolution	Accuracy	Range
Nutrient conc. *also salinity and spectral intensity	μM	Scalar	0.05	2	0.5 to 200

3.10.3 Typical Applications

Nitrate is one of the main nutrients required for growth of phytoplankton. Understanding the distribution of nitrate in the oceans is essential to understanding biological processes.

3.10.4 How the Data are Applied to Answer Specific Questions

See Section 3.10.3.

3.10.5 Sampling Rates and Variability

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
0.5 Hz	13688	2 seconds	

3.10.6 Typical and Peak Data Output (Transfer) Rates

The data string is a fixed length 1711 byte ASCII frame containing the full UV absorption spectrum and calibrated nitrate readings (light frames) and the same length fixed frame with spectral dark data (dark frame). There are 10 light frames of data provided from the instrument for every dark frame.

The instrument baud rate is 38400 bps, and the maximum data output rate is 0.5 Hz or 1 frame every 2 seconds.

3.10.7 Metadata

Metadata associated with the ISUS includes (but is not limited to):

- instrument serial number;
- deployment logging mode;
- scheduled logging parameters (number of minutes between readings, when the first logging event will occur, and whether or not to log real data);
- spectrophotometer parameters (integration period, per scans, spectra scans, dark current scans, collection rate of DC data, coefficients);
- fiberlite lamp parameters (total on time, power on warmup period, shutter during warmup, UV cleaning cycle, reference detector);
- curve fitting options (fitting range, concentrations, baseline model, coefficients).

3.10.8 Operational Modes and Dynamic Configuration

The ISUS is designed to be operated in three modes: continuous, scheduled, and triggered (polled). In continuous mode operation, the instrument can be started and stopped with ASCII commands. Scheduled mode is typically used in moored configurations; the user specifies the number of samples to take and the interval between sampling periods (in minutes). Triggered mode has not yet been implemented on the ISUS.

3.10.9 Data Formats, Processing Software, Protocols and Interfaces

Interfaces available include RS-232 and RS-422. The RS-232 interface is bi-directional; the RS-422 interface operates in a transmit-only mode. The default baud rate of the instrument is 38,400 bps.

The ISUS has been interfaced to and extensively tested with the Seabird series of CTD profilers.

Data reported by the instrument is in the Satlantic Data Format Standard. For each sample, the instrument will compose and transmit one frame of telemetry. Two telemetry modes are available in the ISUS: Full and Concentration. Full telemetry provides all information to the user. Telemetry fields are output in ASCII format in comma delimited fields. Data formats are fully described in the MBARI-ISUS Operation Manual.

SatView and SatCon are Windows-compatible software packages compliant with the proprietary Satlantic Data Format Standard. SatView is a data acquisition and real time display application. SatCon is a post processing and conversion application for telemetry logged with SatView.

IP Interface Available	Interfaces Available	Data Formats	Comments
N	a) RS-232 b) RS-485	ASCII	

3.10.10 VENUS/NEPTUNE Considerations

The ISUS sensor has a UV deuterium lamp with a 1000 hour lamp life before a lamp replacement and calibration is required. This means that discrete event sampling via power up/down should be incorporated to save lamp life but still collect good temporal nitrate data.

Example:

Collect Nitrate data for 0.5 minutes, every 5 minutes, 24 hours per day. For a 6 month deployment the lamp would use approx. 438 hours. In this case the Nitrate sensor is operational for a year before lamp servicing is required.

If power cycling is not possible, the sensor can be programmed to log and send data at any interval required.

3.11 D-A Instruments OBS-3A Optical Backscatter Sensor

The OBS-3A monitor is an optical backscatter sensor for measuring turbidity and suspended solids concentrations by detecting infrared (IR) radiation scattered from suspended matter. The IR scattered between 140 and 160 degrees is detected after passing through a filter that absorbs visible light. The OBS signal is proportional to the turbidity or sediment concentration and a polynomial function is used to provide the specified accuracy.

The OBS-3A includes a temperature sensor and may be equipped with pressure and conductivity sensors.

3.11.1 Requirements and Limitations

Light attenuation is caused by the combined absorption and scattering properties of everything in the water column, including the water itself. The attenuation coefficient is the sum of the absorption coefficient and the scattering coefficient ($a+b=c$). The OBS-3A provides an estimate of the scattering coefficient (b), but a value for the absorption coefficient (a) is required to completely characterize the optical attenuation of a sample. A transmissometer or related instrument can provide the absorption value.

The OBS-3A is calibrated with a Formazin standard solution to produce a voltage output that is proportional to NTU (Nephelometric Turbidity Units). It is important to recognize that two different models of optical backscatter instruments will produce different NTU values for the same sample due to the differences in sensing elements and principles of

operation. Cautious interpretation of NTU values should consider whether the same model of instrument and mode of data collection was used between data sets.

3.11.2 Physical Quantities and the Dimensionality

The physical quantity measured by the OBS-3A is optical backscatter. This measurement is quantified by value into NTUs.

Parameter	Units	Dimensionality	Resolution	Accuracy	Range
a) Turbidity	a) NTU	a) Scalar		a) 0.1*	a) 0 to 4,000
b) Sediment conc.	b) mg/l	b) Scalar		b) 0.1	b) 0 to 5,000*

3.11.3 Typical Applications

Turbidity describes the relative clarity of water, ranging from perfectly clear and transparent to cloudy, hazy or opaque. Turbidity in water is caused by suspended matter such as clay, silt, finely divided organic and inorganic matter, colored organic compounds, algae and other microscopic organisms. Optical backscatter instruments provide measurements that can be used to compute turbidity.

3.11.4 How the Data are Applied to Answer Specific Questions

See Section 3.11.3.

3.11.5 Sampling Rates and Variability

Scan rates and numbers of scans to average per sample depend on the application and the environmental conditions. Typical sampling rates are:

Environment	Scan Rate (Hz)	Scans per sample	Sample interval (sec)
River	2	60-200	900
Beach	4-10	256-1024	300-900
Estuary	4-8	60-1024	900-3600

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
2 Hz	1600 bits/sec (approx)	0.5 sec	

3.11.6 Typical and Peak Data Output (Transfer) Rates

Data volumes are dependent on the fields selected for output and the reporting interval. Peak output rates would be on the order of 100 bytes per record at 2 Hz (1600 bits/sec). Typical output rates would be on the order of 100 bytes per record every 5 minutes, or possibly once an hour.

3.11.7 Metadata

Instrument metadata for the OBS-3A include:

- sensor serial numbers (OBS, pressure, temperature, conductivity, auxiliary OBS);
- pressure units;
- OBS coefficients (NTU).

Many of these values are accessible through the supplied software.

3.11.8 Operational Modes and Dynamic Configuration

The OBS-3A can be configured for automatic sampling for specified durations at specified intervals.

3.11.9 Data Formats, Processing Software, Protocols and Interfaces

The default RS-232 communications parameters for the OBS-3A are 19200 baud, 8 data bits, no parity, 1 stop bit (19200N81).

The data are output in ASCII strings, with either tab or space delimiters. Fields include:

- date and time (to 0.1 seconds);
- depth (optional);
- wave height and period;
- turbidity in NTUs;
- standard deviation of measurements;
- sediment concentration (mg/L);
- statistics;
- temperature;
- conductivity;
- salinity;
- battery voltage.

A sample data file is provided in the supplemental CD-ROM.

Configuration and display software is available from the manufacturer.

IP Interface Available	Interfaces Available	Data Formats	Comments
N	RS-232	ASCII	

3.11.10 VENUS/NEPTUNE Considerations

The infrared light-emitting diode (IRLED) deteriorates with use and will become less bright over time (on the order of a few percent per year). The optical face may change with biofouling on the order of weeks to months. Optically clear TBT compounds are available to mitigate these effects. The instrument needs periodic cleaning and calibration. Calibration procedures are documented and can be performed by the user.

To characterize beam attenuation for a given sample, coincident transmissometer data should be collected.

The sensor results (in NTU units) should be compared only to results from similar sensors.

3.12 *ASL Zooplankton Acoustic Profiler*

The ASL WCP (Water Column Profiler) is a 200 kHz monostatic hydrophone system with a 6 degree beam width. It measures biomass in the overlying water column (in the case of a bottom-mounted, upward-looking sensor) with a target spatial resolution on the order of centimeters in length scale (nominally 16 mm for a 200 kHz transducer).

3.12.1 Requirements and Limitations

The WCP has a maximum effective range of approximately 200 m when supplied with a 200 kHz transducer.

3.12.2 Physical Quantities and the Dimensionality

The WCP measures acoustic backscatter and reports a volume scattering return (a power measurement with units of decibels referenced to a standard). Bin sizes and locations are computed by gated time return. Nominally the instrument is configured for 1 m bins.

An instrument configured for 200 bins and 1 m bin resolution will return a record containing header information and 200 power values (8 bits each) representing the water column.

Parameter	Units	Dimensionality	Resolution	Accuracy	Range
Acoustic backscatter (volume)	Decibels	Scalar array	*see text	*see text	0 to 255

3.12.3 Typical Applications

The ASL WCP is designed to monitor zooplankton distribution, activity and abundance. The abundance value provided is a rough estimate based on unit area. Typical deployments are both internally-recording and real time streaming systems over cable.

3.12.4 How the Data are Applied to Answer Specific Questions

The application of data is best illustrated by example. Since September 1999 a 200 kHz Water Column Profiler has been operating on a moored buoy in Saanich Inlet. Time series of acoustic backscatter have been collected, showing the evolution of the behavior and abundance of the dominant species of zooplankton (Euphausiids and Amphipods) over timescales from diurnal to seasonal.

3.12.5 Sampling Rates and Variability

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
1 Hz	13,264	1 sec	

3.12.6 Typical and Peak Data Output (Transfer) Rates

The WCP maximum output rate is 1 Hz of continuous data. The instrument provides a complete output record, as per the attached format description, for each ping sent out. Each data record consists of a header (58 bytes including auxiliary sensors) and 1 byte of data for each bin. The instrument provides a spatial or bin size resolution for the measured volume backscatter return at 0.125 m (1/8 of meter). For the maximum range setting of 200 m, there are $8 \times 200 = 1600$ measurements of volume backscatter return in each water column profile.

The total information provided for each one second ping amounts to 1658 bytes (13,264 bits) per second. This will require a serial port with at least a 19200 baud rate capability.

3.12.7 Metadata

Most of the instrument-specific metadata are provided in the header of each WCP record. These include:

- deployment name;
- phase number;
- start date/hour of this phase;
- ping interval;
- ping length;
- lockout;
- maximum range;

- pressure interval;
- burst interval;
- burst length;
- data map;
- samples per bin;
- synchronization;
- pressure (if present);
- temperature (if present);
- tilt.

3.12.8 Operational Modes and Dynamic Configuration

The WCP automatically outputs an ASCII data stream at a specified interval, nominally 1 Hz. When connected via RS-422 (or virtual RS-422) interface, the instrument can be configured using supplied software. Parameters that can be configured include:

- the duration of the outgoing acoustic ping in microseconds (usually 300);
- the receiver gain setting; choices are 1, 2, 3 and 4; the gain for setting 4 is about 20 dB greater than for setting 1;
- the maximum range for measurement periods (usually set to the maximum allowable value of 200 m).

3.12.9 Data Formats, Processing Software, Protocols and Interfaces

The WCP outputs data in an ASCII hex format. Data formats are described in the companion CD-ROM. The instrument can be supplied with desktop software (Windows-based) for real time display of 10-minute windows and logged to ASCII text files and raw data files. Interface protocols are listed in the following table.

IP Interface Available	Interfaces Available	Data Formats	Comments
N	RS-232 RS-485	ASCII hex	Requires 19,200 baud for max data rate

3.12.10 VENUS/NEPTUNE Considerations

- as with any upward-looking acoustic device, the WCP signal can be affected by structures in the proximal water column, and in particular by mooring hardware;
- an ocean bottom upward-looking configuration is more stable than a downward-looking buoy-mounted configuration;
- for the VENUS deployments, the standard aluminum pressure casing has been replaced with a synthetic material to reduce the effects of corrosion for long-term deployments;
- the acoustic transducer has been modified to permit variable sensor geometry without having to rotate the entire instrument; the sensor head can easily be fixed to a mounting structure in any orientation;

- as a single-frequency instrument, the WCP cannot discriminate specific types of biomass (e.g., perform species identification, size distribution, etc); multiple-frequency instruments are capable of providing more information to assist in classifying targets.

3.13 FLOWCAM Flow Cytometer

FlowCAM® is a state-of-the-art instrument for rapid monitoring of particles in fluid. It combines the capabilities of flow cytometry and microscopy by automatically counting, imaging, and analyzing the cells in a discrete sample or a continuous flow.

The processing system captures a digital image of each cell and presents the data in an easy-to-read spreadsheet or through their patented, interactive scattergram.

FlowCAM® lends itself to numerous oceanographic applications, offering both discrete sampling and continuous, *in situ* analysis. It enables the collection of important information from aquatic systems, providing up-to-the-minute, time stamped data on particles and cells from 1 µm to 3 mm. It allows for the easy integration of other instrumentation, such as bulk fluorometer or temperature and salinity monitors.

The submersible version of the FlowCAM has the following features:

- user-friendly operation and flexible analysis with adjustable settings to maximize acquisition of target data;
- three visualization packages, with four levels of magnifications, 20X, 10X, 4X, and 1X, allow users to target specific cell types or particles;
- magnification is enlarged 100X by the CCD camera;
- each counted cell or particle is imaged and kept in focus by FlowCAM's patented imaging system;
- specific particle measurements include: length, width, size (ESD), area, light scatter, and fluorescence (chlorophyll and phycoerythrin are factory standard);
- adjustable flow rates from 1 ml/minute to 12 ml/minute;
- flow cell and calibration settings are designed to increase run time and allow for rapid and cost effective maintenance;
- the interactive scattergram enables rapid and immediate review of cell data and images;
- FlowCAM's user interface provides real time viewing of analysis and particle images in both capture and live video modes;
- acquired data and images are stored in common data formats and can be used in any archiving and image processing programs;
- optional 488 nm or the 532 nm excitation light.

3.13.1 Requirements and Limitations

The FlowCAM requires an IP network connection; no other interfaces are available. Interactive operation requires the proprietary IMS software package.

3.13.2 Physical Quantities and the Dimensionality

The FlowCAM measures 10 properties of particles in the water or of the water itself. These properties are particle time of passage, particle chlorophyll fluorescence, particle light scatter, particle phycoerythrin, particle length, particle width, particle ESD, water bulk fluorescence, water temperature, and water salinity. Images of every particle are also taken.

Parameter	Units	Dimension- ality	Resolution	Accuracy	Range
Imagery, spatial and spectral metrics (length, width, size, area, scatter, fluorescence)	µm	Scalar	1		1 – 3,000,000

3.13.3 Typical Applications

The FlowCAM can be configured for monitoring:

- coastal waters (and, in particular, red tide phenomena);
- fluid borne particulate contamination;
- phytoplankton;
- ballast water;
- aquaculture, etc.

3.13.4 How the Data are Applied to Answer Specific Questions

The primary function of the flow cytometer is to facilitate the identification and classification of phytoplankton and other single celled organisms.

FlowCAM generates qualitative and quantitative information that can be used to develop normal baselines which can be referenced to compare algae levels and algal blooms. By spot-checking in the field, FlowCAMs can help pinpoint when and where disturbances occur and visually display the specific microorganisms. This real time information helps managers assess water quality on-line, as well as determine whether isolated events are accidental or introduced.

Post filtration, FlowCAM can be used for continuous process control to monitor ozone or ultra violet treatment and to maintain quality control in water treatment plants.

The data are stored in spreadsheet-compatible form and is also stored in a database which allows the user to use an interactive scattergram; from this, the user can select a range of properties and display particles that fall within that range. These data can be used to show cell types present and when and where they passed by the instrument.

3.13.5 Sampling Rates and Variability

The Flow Cytometer can be configured for continuous sampling, provided sufficient power is available.

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
N/A	250,000 (assuming 100 MB/hr)	N/A	Data are copied over TCP/IP network connection using filesystem operations

3.13.6 Typical and Peak Data Output (Transfer) Rates

Data and images combined may result in data rates up to 100 Mbytes per hour unless a sample strategy is used to reduce this.

3.13.7 Metadata

There are no specific metadata fields apart from the CSV descriptors of cells identified in the image files.

3.13.8 Operational Modes and Dynamic Configuration

The instrument logs data and images internally. Files and data can be accessed (uploaded and/or deleted) via network connection using filesystem operations.

3.13.9 Data Formats, Processing Software, Protocols and Interfaces

Image files are stored in .TIFF format in a file-based structure on the instrument. Computed metrics are stored in a CSV (comma separated value) file that can be imported into processing and archiving packages.

IMS software (available from the manufacturer) is recommended for offsite processing. The software can also be used to control the instrument and to review data.

Post processing can be performed for particle type recognition using libraries of different cell types. This allows the automated determination of cell types sampled by the instrument at different times.

IP Interface Available	Interfaces Available	Data Formats	Comments
Y	TCP/IP	.TIFF (images) CSV (data)	

3.13.10 VENUS/NEPTUNE Considerations

The instrument can be calibrated for particle size measurements. It also self-calibrates every half hour to correct for image chamber fouling. The image chamber will require periodic cleaning.

3.14 Guralp CMG-1T3 Seismometer

Technical information is available on the supplemental CD-ROM.

3.14.1 Requirements and Limitations

None.

3.14.2 Physical Quantities and the Dimensionality

The types of data collected are digitized velocity and acceleration values (20, 40, or 100 samples per second per sensor). Data samples are 27-28 bits uncompressed. Not much acceleration (strong motion) data are collected at this time in the Canadian National Seismic Network. Normally acceleration is collected for broadband sensors in 3 axes.

Parameter	Units	Dimension -ality	Resolution	Accuracy	Range
a) Velocity b) Acceleration	a) m/s b) m/s ²	a) Vector b) Vector	*see text	*see text	*see text

Frequency Response

Sensor	Standard Bandwidth	Optional Bandwidth
CMG-1T	0.008 Hz (120 s) to 50 Hz	0.0027 Hz (360 s) to 50 Hz
CMG-3T	0.01 Hz (100 s) to 50 Hz	0.018 Hz (60 s) to 50 Hz

3.14.3 Typical Applications

Canada has 120 land stations in the National Seismic Network. About 70% of the data in Canada comes into PGC in real time. The national network is divided into eastern and western regions. The Ottawa branch has 85% of the data in real time (they get all the POLARIS data; PGC gets only western Canada POLARIS data). Automated hourly batch transfers via FTP to Ottawa address any missing data.

Most post-processing addresses the determination of location and how the earth moved. The analyses are not automated. Location computation is automatic; an analyst then checks and verifies the trigger and location data (for example, an event could be a mine blast that is not easily distinguishable for classification by machine).

No data products are generated *per se*, but information is used for national building codes, and hazard and risk analysis.

Data disseminated to other agencies includes stations, locations, magnitudes, and fault mechanism.

3.14.4 How the Data are Applied to Answer Specific Questions

See Section 3.14.3.

3.14.5 Sampling Rates and Variability

Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
100 Hz	3200	1 - 6 sec	

3.14.6 Typical and Peak Data Output (Transfer) Rates

Average throughput is 1000-2000 bits/sec; 2400 baud links are sufficient. When there is a lot of activity the bandwidth exceeds 2400 baud. The data are buffered in the digitizer and will get out eventually. For one-way data transmission (no error checking possible) the data may be sent twice, several minutes apart. The data acquisition system discards any duplicate packets.

3.14.7 Metadata

Location, station identifier, calibration coefficients, software gain settings, and instrument configuration information are unique to each station. Settings are not typically changed over the course of a deployment, though time signal drift may occur due to GPS lockup.

3.14.8 Operational Modes and Dynamic Configuration

Calibration information can be sent remotely to the seismometer. Gain settings or firmware changes are possible over a 2-way link. Although possible, in the Canadian Seismic Network there is no interactive instrument management or configuration.

3.14.9 Data Formats, Processing Software, Protocols and Interfaces

Guralp products store their data in Guralp Compressed Format (GCF). Comprehensive software available from the manufacturer and in the public domain enables export and

conversion to a range of widely used formats (GCF, SUDS, SEG-y, SAC, and miniSEED).

The Canadian Seismic Network uses .CA data file formats.

Matlab code is available for import of data in GCF format (see supplemental CD-ROM).

Scream! is a Windows 9x/NT/2000/XP application for Seismometer Configuration, Real time Acquisition and Monitoring. This comprehensive package is capable of auto-sensing connected instruments and initiating data acquisition and storage. Multiple serial ports can simultaneously receive stream data from any GSL instrument, display it in any scaling factor, and record to disk. Network facilities enable stream data broadcasting over a TCP/IP supporting network (e.g., LANs, dial-up lines, etc). Files recorded to disk can be replayed.

IP Interface Available	Interfaces Available	Data Formats	Comments
Y*	RS-232 RS-422 TCP/IP		

3.14.10 VENUS/NEPTUNE Considerations

QA and post-processing is not performed in real time. Communications between sites use CRC on data packets. Typical failure modes include the occasional GPS timing problems (lockup or clock drift); for many applications, a 3-4 second error is OK.

Calibration information is entered into the data acquisition system and consists of compensation for component tolerances. This type of calibration is not needed periodically and is only performed when the instrument is serviced.

3.15 Summary Tables

For ease of reference, the information in the above sections (i.e., Sections 3.3– 3.14) is summarized in the following four tables:

Table 1. Contact information. For each sensor, this table contains the Web sites for the sensor manufacturers, along with email addresses and phone numbers for those contacts that have been identified. It also indicates what type of documentation is currently available.

Table 2. Summary of quantities and dimensionality. For each sensor, this table lists the parameter measured, the units of measurement, the dimensionality of the data, and the resolution, accuracy, and range of the measurements.

Table 3. Summary of sampling rates and variability. For each sensor, this table contains the maximum data acquisition frequency, the maximum data reporting rate, the reporting interval, and comments related to this information.

Table 4. Summary of interfaces, data formats and software. For each sensor, this table outlines whether an IP interface is available, and lists the interfaces that are available, the data formats and relevant comments.

Note that in the tables below, the symbol (*) denotes a tentative or qualified entry.

3.16 Table 1. Contact Information

Instrument	Model	Brochure Available / Obtained	Response from Supplier	Product Manual Obtained	Contact Information
CTD	Seabird 16plus	Y/Y	Y	Y	www.seabird.com Doug Bennet: dbennet@seabird.com Phone: (425) 643-9866 Fax: (425) 643-9954
Oxygen Sensor	Optode 3975	Y/Y	N	Y	www.aanderaa.com Richard.butler@aanderaa.no , info@aanderaa.no erik.riber-mohn@aanderaa.no elin.legreid@aanderaa.no Phone: (+47) 55 109900 Fax: (+47) 55 109910
Gas Tension Device	GTD Pro	N/N	Y	N	www.pro-oceanus.com Craig McNeil: mcneil@pro-oceanus.com Phone: (902) 852-4433 Fax: (902) 852-4433
Acoustic Doppler Current Profiler	RDI Deep Water Workhorse (300 kHz)	Y/Y	Y	Y	www.rdinstruments.com Paul Devine: Pdevine@rdinstruments.com Phone: (858) 693-1178 x 3011 Cell: (858) 254-7204
Digital Video Camera	Imenco IMDV 3018	Y/Y	N	N	www.imenco.no jon-inge.pederson@imenco.no Phone: (+47) 52 864100 Fax: (+47) 52 864101
Orientation Sensor	Jasco AIM-2000	Y/Y* (interim)	Y	N	www.jasco.com Chris Sundstrom: chris@jasco.com Phone: (250) 483-3300 x 1003 Fax: (250) 483-3301
Broadband Hydrophone System	IOS – Svein Vagle	N	Y	N	Svein Vagle: vagles@pac.dfo-mpo.gc.ca Phone: (250) 363-6339 Fax: (250) 363-6798

Nitrate Sensor	MBARI-ISUS	Y/Y	Y	Y	www.satlantic.com Cyril Dempsey: cyril@satlantic.com Phone: (902) 492 4780 Fax: (902) 492 4781
Flow Cytometer	FlowCAM	Y/Y	Y	N	www.fluidimaging.com Chris Sieracki: sieracki@gwi.net Phone: (207) 882-1100 Fax: (207) 882-4800
Optical Backscatter Sensor	OBS-3	Y/Y	Y	Y* (different model)	www.d-a-instruments.com John Downing: John@D-A-Instruments.com Phone: (800) 437-8352 or (360) 385-0272 Fax: (360) 385-0460
Zooplankton Acoustic Profiler	ASL Water Column Profiler	N/N	Y	N	www.aslenv.com Dave Fissel: dfissel@aslenv.com Phone: (250) 656-0177 Fax: (250) 656-2162
Seismometer	Guralp CMG-1T 3	Y/Y	Y	N	www.guralp.net Bruce Pauly: dta_pauly@compuserve.com (Agent: Digital Technology Associates) Support: support@guralp.com Sales: sales@guralp.com Phone: (+44 (0)118 981 9056 Fax: (+44) (0)118 981 9943

3.17 Table 2. Summary of Quantities and Dimensionality

Instrument	Model	Parameter	Units	Dimension- ality	Resolution	Accuracy	Range
CTD	Seabird 16plus	a) Temperature b) Conductivity c) Pressure	a) deg C b) S/m c) dbar	a) Scalar b) Scalar c) Scalar	a) 0.001 b) 0.00005 c) 0.002%	a) 0.005 b) 0.003 c) 0.1%	a) -5 to +35 b) 0 to 9 c) 0 to 7,000
Oxygen Sensor	Optode 3975	a) Oxygen conc. b) Air saturation	a) μ M b) %	a) Scalar b) Scalar	a) < 1 b) 0.4	a) 8 b) 5	a) 0 to 500 b) 0 to 120
Gas Tension Device	GTD Pro	Total dissolved air pressure	mbar	a) Scalar <u>Scalar</u>	0.0001	0.1	0 to 2000
Acoustic Doppler Current Profiler	RDI Deep Water Workhorse (300 kHz)	a) Velocity b) Direction c) Tilt d) Temperature	a) cm/s b) deg c) deg d) deg C	a,b) Vector array c) Scalar d) Scalar	a) 0.1 b) 0.01 c) 0.01 d) 0.01	a) 0.5% b) 2 c) 0.5 d) 0.4	a) 0 to 500 b) 0 to 360 c) 0 to 15 d) -5 to 45
Digital Video Camera	Imenco IMDV 3018	Imagery	n/a	n/a	N/a	n/a	n/a
Orientation Sensor	Jasco AIM-2000	a) Magnetic Dir. b) Inclination c) Pressure d) Temperature	a) deg b) deg c) mbar d) deg C	a) Scalar b) Scalar c) Scalar d) Scalar	a) 0.1 b) 0.1 c) 0.1 d) 0.1	a) 1 (level) b) 0.5 (to 30°) c) 0.1% f.s. d) 0.2	a) 0 to 360 b) -60 to +60 c) variable d) -40 to +85
Broadband Hydrophone System	IOS – Svein Vagle	Pressure amplitude	decibels	Scalar array	*see text	*see text	0 to 65535 (16 bit)
Nitrate Sensor	MBARI-ISUS	Nutrient conc. *also salinity and spectral intensity	μ M	Scalar	0.05	2	0.5 to 200
Flow Cytometer	FlowCAM	Imagery, spatial and spectral metrics (length, width, size, area, scatter, fluorescence)	μ m	Scalar	1		1 – 3,000,000
Optical Backscatter Sensor	OBS-3	a) Turbidity b) Sediment conc.	a) NTU b) mg/l	a) Scalar b) Scalar		a) 0.1* b) 0.1	a) 0 to 4,000 b) 0 to 5,000*
Zooplankton Acoustic Profiler	ASL Water Column Profiler	Acoustic backscatter (volume)	decibels	Scalar array	*see text	*see text	0 to 255

Seismometer	Guralp CMG-1T 3	a) Velocity b) Acceleration	a) m/s b) m/s ²	a) Vector b) Vector	*see text	*see text	*see text
-------------	-----------------	--------------------------------	-------------------------------	------------------------	-----------	-----------	-----------

3.18 Table 3. Summary of Sampling Rates and Variability

Instrument	Model	Max Data Acquisition Frequency	Max Data Report Rate (bits/sec)	Reporting Interval	Comments
CTD	Seabird 16plus	4 Hz	736	0.25 sec to minutes	Continuous output mode available. Polled output mode available. Specified interval output available.
Oxygen Sensor	Optode 3975	1 Hz	430(est.)	1 sec to 255 min	
Gas Tension Device	GTD Pro	10 Hz	9600	As per CTD	Usually an auxiliary sensor; data rates as per Seabird SBE 16plus CTD (for example).
Acoustic Doppler Current Profiler	RDI Deep Water Workhorse (300 kHz)	2 Hz	50720	0.5 sec to minutes	Using PD0 output format with 150 cells at 3170 bytes per scan
Digital Video Camera	Imenco IMDV 3018	a) streaming b) 1 Hz (est.) c) 1 Hz (est.) d) 1.7 Hz	a) 1244 Mbps b) c) 1.5 Mbps d) 100 Mbps	a) streaming b) c) 15 sec d) 0.5 sec (est.)	a) HDTV b) Imenco SDS3040 c) Kongsberg OE14-208 d) C-Map Cyclops
Orientation Sensor	Jasco AIM-2000	8 Hz	6400	1 sec to 1 minutes	
Broadband Hydrophone System	IOS – Svein Vagle	200 kHz	3,200,000 (3.2 Mbit/s)	1 Hz	200 kHz data rate with 16 bit A/D sampling per hydrophone
Nitrate Sensor	MBARI-ISUS	0.5 Hz	13688	2 seconds	
Flow Cytometer	FlowCAM	N/A	250,000 (assuming 100MB/hr)	N/A	
Optical Backscatter Sensor	OBS-3	2 Hz	1600 (approx)	0.5 sec	Dependent on intermediary acquisition and processing system.
Zooplankton Acoustic Profiler	ASL Water Column Profiler	1 Hz	13,264	1 sec	
Seismometer	Guralp CMG-1T 3	100 Hz	3200	1 to 6 sec	

3.19 Table 4. Summary of Interfaces, Data Formats and Software

Instrument	Model	IP Interface Available	Interfaces Available	Data Formats	Comments
CTD	Seabird 16plus	N	a) RS-232 b) RS-422	ASCII, Hex	a) Processing software available. b) Can output ASCII stream. c) Shared serial port for control and data.
Oxygen Sensor	Optode 3975	N	a) RS-232 b) Analog 0-5V	a) ASCII b) N/A	In analog mode the sensor output and format is dependent on the intermediary collection system
Gas Tension Device	GTD Pro	N	RS-232 RS-485	ASCII	Control/data interface exists in Seabird SBE-16 plus CTD.
Acoustic Doppler Current Profiler	RDI Deep Water Workhorse (300 kHz)	N	RS-232	ASCII, binary	
Digital Video Camera	Imenco IMDV 3018	N	a) IEEE-1394 b) Composite c) Y/C		a) Requires IEEE-1394 to IP converter. Requires streaming video server
Digital Stills Camera		Y Y Y	a) Ethernet (data); RS232 (control) b) USB (data); RS232/485 (ctrl) c) 100T (data); RS485 (control)	JPG, RAW JPG, RAW JPG, RAW	a) Imenco SDS3040 b) Kongsberg OE14-208 c) C-Map Cyclops
Orientation Sensor	Jasco AIM-2000	N	a) RS-232 b) RS-485	NMEA-0183	a) Format available for proprietary data sentence. b) Shared serial port for control and data.
Broadband Hydrophone System	IOS – Svein Vagle	Y	IP (Windows PC at source)	Binary, .WAV	
Nitrate Sensor	MBARI-ISUS	N	a) RS-232 b) RS-485	ASCII	
Flow Cytometer	FlowCAM	Y	10baseT	TIFF, CSV	a) Requires proprietary software interface for acquisition, control, processing. b) Summary files in CSV, images in TIFF.
Optical Backscatter Sensor	OBS-3	N	RS-232	ASCII	Must be integrated as an auxiliary channel on another instrument.

Zooplankton Acoustic Profiler	ASL Water Column Profiler	N	RS-232 RS-485	ASCII hex	
Seismometer	Guralp CMG-1T 3	Y*	RS-232 RS-422 TCP/IP		

3.20 Considerations for Incorporating New Instrumentation into VENUS/NEPTUNE Systems

1. Requirements and Limitations

- 1.1. What are the power requirements and characteristics?
 - 1.1.1. Voltage
 - 1.1.2. Current
 - 1.1.3. AC/DC
 - 1.1.4. Power spikes, surges or drains when operating or powering the instrument up/down
- 1.2. Are the pressure casing ratings sufficient for the deployment depth?
- 1.3. To what degree are the sensors subject to biofouling?
- 1.4. Are there any components that are consumable or subject to wear and degradation?
 - 1.4.1. Reagents (including anti-biofouling measures)
 - 1.4.2. Light sources that degrade with use
 - 1.4.3. Moving mechanical assemblies
- 1.5. What are the typical failure modes and their MTBF (Mean Time Between Failure) for the instrument?
- 1.6. Will the presence or operation of the sensor affect other sensors in the node?
 - 1.6.1. Vibration affecting motion sensors
 - 1.6.2. Light sources affecting optical sensors
 - 1.6.3. Heat dissipation affecting temperature measurements
 - 1.6.4. Noise affecting acoustic sensors
 - 1.6.5. Reagents used or generated (e.g., chlorine/bromine affecting conductivity measurements)
- 1.7. If the sensors are subject to calibration drift, what are the characteristics of this drift (linear, nonlinear)? What is the long-term stability of the instrument?
- 1.8. How frequently do the sensors need servicing or calibration?

2. Data and Metadata

- 2.1. What are the physical quantities and their dimensionality for the sensors?
- 2.2. What are the resolution, precision and accuracy of the sensors?
- 2.3. What are the sampling rates and the variability of sampling rates for the sensors?
- 2.4. What are the typical and peak data output rates for the instrument?
- 2.5. What metadata describes the instrument? Which of these metadata are dynamic?
 - 2.5.1. Configuration
 - 2.5.2. Calibration
 - 2.5.3. Sampling regime
 - 2.5.4. Operational modes
- 2.6. How are the data presented by the instrument?
 - 2.6.1. Stream-based
 - 2.6.2. File-based

3. Operational Modes and Dynamic Configuration

- 3.1. What are the operational modes of the instrument?
 - 3.1.1. Streaming
 - 3.1.2. Polled
 - 3.1.3. Automatic reporting at specified intervals
 - 3.1.4. Other
- 3.2. What parameters or configurations can be modified dynamically or interactively?
- 3.3. Does the instrument operation need to be interrupted for dynamic configuration?
- 3.4. Can the instrument change operational modes autonomously (e.g., on an event trigger)? Can this change in operational mode be detected and logged as metadata?
- 3.5. Are there high frequency data streams that could not be recovered if an instrument were operated in a particular mode (e.g., 1Hz measurements cannot be recovered from 10 minuted averaged data)? What are the fundamental operational modes for maximum data collection?

4. Data Protocols and Interfaces

- 4.1. What are the available protocols and physical interfaces to the instrument?
 - 4.1.1. RS-232/422/485
 - 4.1.2. Internet Protocol, TCP/IP
 - 4.1.3. Proprietary
 - 4.1.4. Analog/Other
- 4.2. Are interfaces used in a proprietary or non-standard way (e.g., use of specific pins for instrument control)?
- 4.3. Does the instrument make use of control lines (e.g., RS-232 CTRL-BREAK) that cannot be replicated over TCP/IP?
- 4.4. Are the data and instrument control interfaces separate or shared?

5. Data Formats and Processing Software

- 5.1. What data products are produced
 - 5.1.1. In real-time?
 - 5.1.2. Through post-processing?
- 5.2. What are the data formats used by the instruments? Are they proprietary or published?
- 5.3. What are the instrument operation command sequences (if applicable)?
- 5.4. Does the instrument require specific or proprietary software for:
 - 5.4.1. Setup and configuration
 - 5.4.2. Operation
 - 5.4.3. Data processing
- 5.5. Can multiple instruments be operated with a single instance of a software application? Can multiple instances of the software controlling separate instruments run on the same computer?

- 5.6. For data processing and QA/QC (Quality Assurance/Quality Control):
 - 5.6.1. What data processing can be performed automatically (in real-time)?
 - 5.6.2. What QA/QC can be performed automatically (in real-time)?
 - 5.6.3. What data processing or QA/QC must be performed offline or by a third party?
 - 5.6.4. What is the expected turn-around time for processed data products?
 - 5.6.5. What are the data formats for processed data products?
 - 5.6.6. Do standard or widely accepted utilities exist for post-processing and QA/QC?

4 Data Management Practices

A major aim of the VENUS/NEPTUNE DMAS Examination project was to identify and survey a wide range of organizations involved in the collection, distribution and archiving of data. The aim of this examination was to provide guidance to VENUS/NEPTUNE by considering the practices of, and lessons learned by, other organizations, both oceanographic and non-oceanographic, which have a mandate to collect, process, and disseminate large amounts of scientific data.

BCS first identified a number of data repositories (as described in Section 4.1) and sent two questionnaires out: one for collectors of data, the other targeted to data archivers. Then BCS identified agencies corresponding to three broad categories. These were:

- ocean observing systems,
- organizations with other-than-oceanographic data, and
- local scientific data organizations.

The actual organizations identified are listed and described in Sections 4.2 - 4.4. BCS designed and sent another questionnaire to selected recipients in the agencies identified in each of these categories. We also conducted a number of phone interviews with representatives of agencies of all categories.

The contents, distribution lists, and results of the questionnaires are contained in Appendix B, and the minutes of the phone interviews are contained in Appendix C.

A summary of data management practices is given in Section 4.5, and a statement of the most important and salient points made by respondents during these investigations are contained in Section 4.6.

4.1 Data Repositories

This section identifies national and international repositories of oceanographic and marine seismic data, and provides contact information for these repositories. It also discusses associated programs and projects. The repositories, programs, and data discussed in this section are those that are relevant to the geographical area covered by VENUS/NEPTUNE and those that share some of the data management characteristics of the VENUS/NEPTUNE project. We have concentrated on sub-surface data, largely ignoring satellite-derived data such as sea surface height and temperature. We have also not considered point-in-time compendiums such as World Ocean Atlas 2001.

Phase 2 of this project looks beyond VENUS/NEPTUNE to try to learn from the experiences of other large initiatives; that phase examines other data management initiatives that are not directly relevant to VENUS/NEPTUNE and that have not been considered in Phase 1.

4.1.1.1 Background Information

This section discusses in general terms the programs and repositories that have data that are relevant to VENUS/NEPTUNE; specific details on data types, volumes and data management are covered in the next section. The organizations, repositories and programs discussed in this section are:

- i) the National Oceanic and Atmospheric Administration (NOAA),
- ii) the Marine Environmental Data Service (MEDS),
- iii) the Institute of Ocean Sciences (IOS),
- iv) the Incorporated Research Institutions for Seismology (IRIS),
- v) the World Data Center program (WDC),
- vi) the Responsible National Oceanographic Data Centers program (RNODC), and
- vii) the Global Ocean Observing System (GOOS).

4.1.1.1.1 National Oceanic and Atmospheric Administration

The mission of the National Oceanic and Atmospheric Administration (NOAA), a bureau of the US Department of Commerce, is two-fold:

1. ***Environmental assessment and prediction*** – to observe and assess the state of the US environment, while protecting public safety and the US's economic and environmental security through accurate forecasting; and
2. ***Environmental stewardship*** – to protect ocean, coastal and living marine resources while assisting their economic development.¹

To this end, NOAA has a number of sub-organizations, the following of which are relevant to VENUS/NEPTUNE:

1. The Office of Oceanic and Atmospheric Research (<http://www.oar.noaa.gov/oceans/>) operates a number of laboratories and programs including:
 - 1.1. National Undersea Research Program (supporting underwater research via scuba, submersibles (manned and ROV/AUV) and seafloor observatories) – see <http://www.nurp.noaa.gov/>;
 - 1.2. Pacific Marine Environmental Laboratory (specializing in long-term monitoring of the ocean environment through open ocean observatories) – see <http://www.pmel.noaa.gov/>.
2. The National Environmental Satellite Data and Information Service (NESDIS) (<http://www.nesdis.noaa.gov/index.html>) is mandated with “providing timely access to global environmental data from satellites and other sources.” NESDIS operates four data centers:

¹ See <http://www.commerce.gov/organization.html>

- 2.1. The National Climatic Data Center (NCDC), an archive of climate data, including some marine data;
- 2.2. The National Geophysical Data Center (NGDC), a repository of the following types of ocean-related data: bathymetry, seismic reflection, trackline data, sediment thickness, and marine geology;
- 2.3. The National Ocean Data Center (NODC), which “manages the acquisition, ingest processing, quality control, and long-term preservation of oceanographic data.” The role of the NODC has been described as follows:

“NODC is one of the national environmental data centers operated by the National Oceanic and Atmospheric Administration (NOAA) of the U.S. Department of Commerce. The main NODC facility is located in Silver Spring, Maryland. The NODC also has field offices collocated with major government or academic oceanographic laboratories in Miami, FL; La Jolla, CA; Seattle, WA, and Honolulu, Hawaii. The NODC holds physical, chemical, and biological oceanographic data collected by U.S. Federal agencies, including the Department of Defense (primarily the U.S. Navy); State, and local government agencies; universities and research institutions; and private industry. NODC does not conduct any data collection programs of its own; it serves solely as a repository and dissemination facility for data collected by others. A large percentage of the oceanographic data held by NODC is of foreign origin. NODC acquires foreign data through direct bilateral exchanges with other countries and through the facilities of World Data Center (WDC) for Oceanography, which is operated by NODC under the auspices of the U.S. National Academy of Sciences.”²

- 2.4. The National Coastal Data Development Center (NCDDC) maintains a searchable metadata catalog of coastal data and provides mechanisms for delivering data (metadata and ocean data (from other sites)) via the Web to customers.
3. The National Weather Service (NWS), which provides weather and hydrologic forecasts and makes weather-related data (including marine weather data) available to the public.
4. The National Marine Fisheries Service, which provides stewardship of marine life and habitat.
5. The Geospatial Data and Climate Services (GDCS), which provides service to all of NOAA regarding data, information, and other services needed to support data management activities. In particular, GDCS administers the following programs:
 - 5.1. The Environmental Services Data and Information Management (ESDIM) program (systems development, modernization, and inter-agency data management coordination);
 - 5.2. The NOAA central metadata repository (NOAAServer);
 - 5.3. NOAA participation in the Global Climate Observing System (GCOS) program, which provides a coordinating role in the world-wide gathering and dissemination of weather data (including marine weather data).

² See http://ioc.unesco.org/oceanportal/browse.php?pg_which=5&cat=5

4.1.1.2 Marine Environmental Data Service

The Marine Environmental Data Service (MEDS) is a branch of Canada's federal Department of Fisheries and Oceans (DFO). The mandate of MEDS is:

- 1) to manage and archive ocean data collected by DFO, or acquired through national and international programs conducted in ocean areas adjacent to Canada, and
- 2) to disseminate data, data products, and services to the marine community in accordance with the policies of the Department.

MEDS supports the following international initiatives:

- 1) Argo Project;
- 2) GOOS (Global Ocean Observing System);
- 3) GTSP (Global Temperature and Salinity Profile Program), and
- 4) Acts as Responsible National Oceanographic Data Center (RNODC) for drifting buoy data.

See http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Home_e.htm

4.1.1.3 Institute of Ocean Sciences

The Institute of Ocean Sciences / Ocean Science and Productivity (IOS/OSAP) data archive contains the holdings of oceanographic data generated by the Institute of Ocean Sciences and other agencies and laboratories, including the Institute of Oceanography at the University of British Columbia and the Pacific Biological Station. See http://www-sci.pac.dfo-mpo.gc.ca/osap/default_e.htm. A synopsis of data management procedures at IOS can be found in Appendix C.

4.1.1.4 Incorporated Research Institutions for Seismology

The Incorporated Research Institutions for Seismology (IRIS) maintains an archive of data that includes passive and active source waveform data, and channel response data. See <http://www.iris.washington.edu/data/data.htm>.

4.1.1.5 World Data Center (WDC) System

The World Data Center System is an organization of WDC's, each of which accepts, maintains, and disseminates various forms of scientific data. A list of current members can be found at <http://www.ngdc.noaa.gov/wdc/list.shtml>. This list includes:

- 1) the WDC for Marine Environmental Sciences
(<http://www.ngdc.noaa.gov/wdc/europe/mes.html>) (Berlin, Germany);
- 2) the WDC for Marine Geology and Geophysics
(<http://www.ngdc.noaa.gov/wdc/usa/mgg.html>) (Boulder, CO, USA);

- 3) the WDC for Oceanography (<http://www.ngdc.noaa.gov/wdc/usa/ocean.html>) (Silver Spring, MD, USA);
- 4) the WDC for Seismology (<http://www.ngdc.noaa.gov/wdc/usa/seismol.html>) (Denver, CO, USA).

4.1.1.6 IOC/IODE RNODC's

The IOC/IODE (International Oceanographic Commission/International Oceanographic Data and Information and Exchange) Responsible National Oceanographic Data Centers (RNODC) are data centers that hold special responsibilities for either specific data types or specific regions. The current list of RNODC's relevant to this study includes:

- 1) RNODC for Drifting Buoys Data: Canada (MEDS);
- 2) RNODC for IGOS (BATHY and TESAC): Japan, USA and Russian Federation;
- 3) RNODC-ADCP: Japan.

4.1.1.7 GOOS (Global Ocean Observing System)

GOOS, supported by the Intergovernmental Oceanographic Commission (IOC) of UNESCO and by the World Meteorological Organization (WMO), is a global system for gathering and processing ocean data. It consists of previously established systems and pilot projects; those of particular relevance to VENUS/NEPTUNE are described in the following sections.

4.1.1.7.1 TAO-TRITON

TAO-TRITON is an array of moored buoys arranged in a roughly 7x12 grid extending from 8°S, 95°W in the southeast to 9°N, 137°E in the northwest. Instruments at these buoys measure wind, air temperature, humidity, precipitation, short wave radiation, water temperature, and salinity and current profile (via ADCP). The water temperature and salinity are measured down to a few hundred metres in depth. Daily averages are transmitted via the Global Telecommunications System (GTS) (except profiles); higher resolution measurements and profiles are physically recovered from buoys annually.

4.1.1.7.2 Ship-of-Opportunity Program (SOOP)

The Ship-of-Opportunity program collects oceanographic data from participating merchant ships. There are approximately 100 dedicated ships of opportunity, which travel predefined routes and report upper ocean temperature (via XBT and XCTD), conductivity (via CTD and XCTD), current profiles (via ADCP), and sea surface characteristics such as temperature, wave height, pCO₂, etc. Measurements are also taken from other ships (navy, research, etc.) that are not formally participating in the program.

4.1.1.7.3 Drifting Buoys

Data from moored and drifting buoys form a component of GOOS and are available from the Drifting Buoy Data Assembly Center (DAC) at the NOAA Atlantic Oceanographic and Meteorological Laboratory (AOML) in Florida (<http://www.aoml.noaa.gov/phod/dac/dacdata.html>), and from the RNODC – Drifting Buoys in Canada (http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Databases/DRIBU/drifting_buoys_e.htm).

Contacts for these organizations are:

The DAC at AOML: mayra.pazos@noaa.gov;

The RNODC at MEDS: keeley@meds-sdmm.dfo-mpo.gc.ca.

4.1.1.7.4 Global Telecommunication System (GTS)

The Global Telecommunication System is an integrated network connecting meteorological telecommunication centers via terrestrial and satellite links. GTS is used by several of the programs discussed here (e.g., Argo, TAO, SOOP).

4.1.1.7.5 Global Temperature and Salinity Profile Program (GTSP)

The following description of the GTSP was copied from http://www.gos.udel.edu/goos/GTSP_data_access.htm

“Temperature and salinity Global Temperature-Salinity Profile Project (GTSP) data are freely available on-line by request or by subscription at time scales from hours after the observation to historical time scales. The first versions of the data are mostly low vertical resolution representations of the profiles. These data are replaced with higher resolution versions of the data that have had more stringent quality checking applied in days or weeks after the observation. Thus the quality of the data in the databases improves with time after the observation. Data inventory are available on-line in the form of monthly maps of observation points.

“The Canadian Marine Environmental Data Service (MEDS) operates the real time database that provides data in operational time frames. The US National Oceanographic Data Center (NODC) operates the ‘continuously managed’ data base where replacement of the low resolution data with higher resolution and better quality controlled data is done.”

The GTSP program acts as a conduit of information provided by other programs, in particular SOOP. A data flow diagram for GTSP is shown in Figure 1. This diagram illustrates how data flow from their point of collection (performed in conjunction with any of several different programs at any of several different sites) to their final archival site(s).

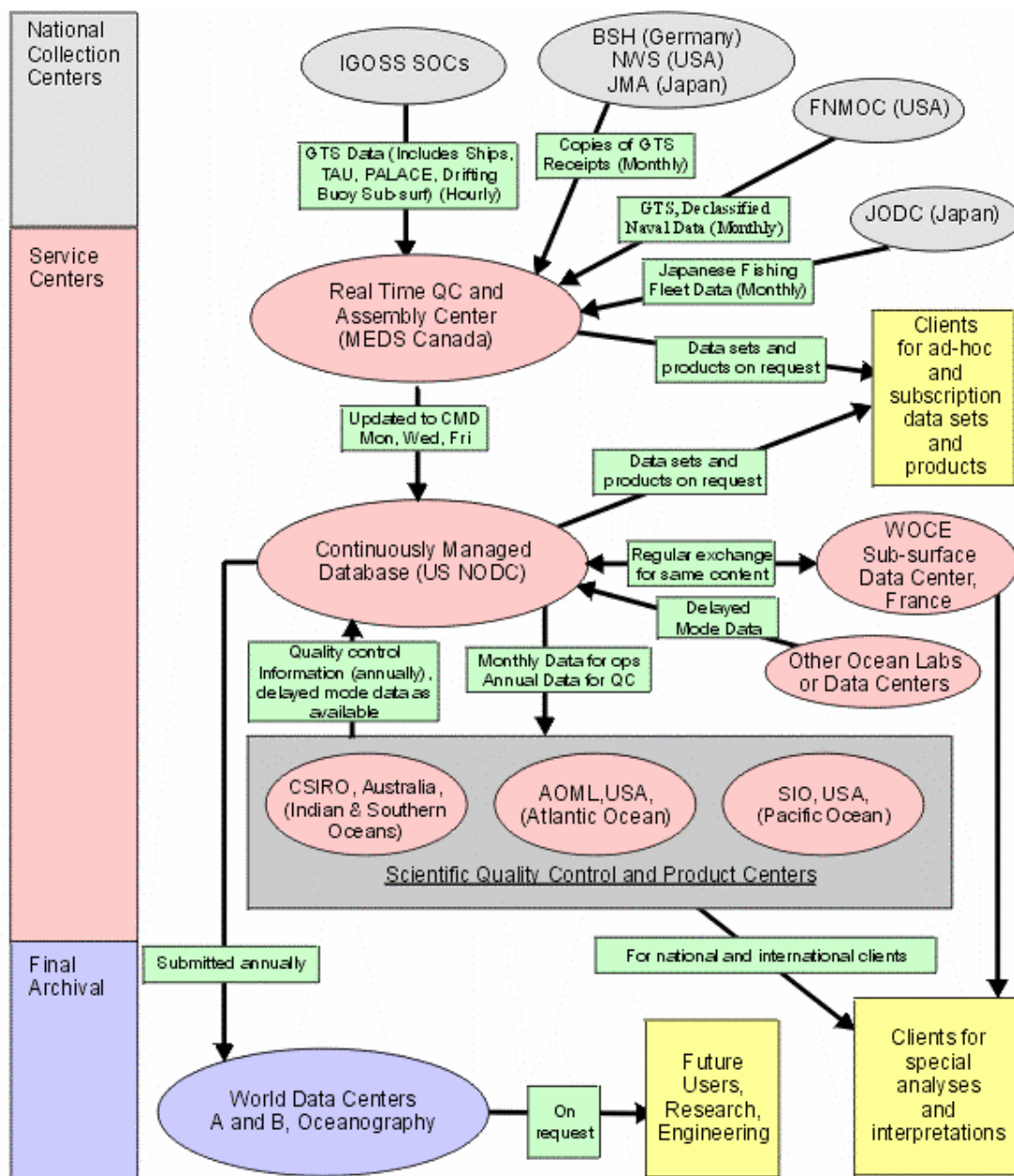


Figure 1. GTSP Data Flow Diagram (from http://www.gos.udel.edu/goos/GTSP_data_flow.htm#Diagram)

4.1.1.7.6 Global Ocean Data Assimilation Experiment (GODAE) / Argo

GODAE is based on the vision of

“A global system of observations, communications, modeling and assimilation, that will deliver regular, comprehensive information on

the state of the oceans, in a way that will promote and engender wide utility and availability of this resource for maximum benefit to the community.”³

To this end, GODAE operates a number of observational networks, one of which is “Argo” (<http://www.argo.ucsd.edu/index.html>). Argo, which is being developed through a collaboration with the Canadian Marine Environmental Data Service (MEDS), is a system of floats designed to measure temperature and salinity in the upper 2000 m of the oceans. It will eventually consist of an array of 3000 profiling floats, providing 100,000 temperature/salinity/pressure profiles and velocity measurements per year distributed over the global oceans at an average 3-degree spacing. There are currently approximately 100 active Argo floats in the Northeast Pacific region and about 1300 floats distributed worldwide. The primary repository for Argo data is on the USGODAE (US Global Ocean Data Assimilation Experiment) Fleet Numerical data server at Monterey, California (<http://www.usgodae.org/usgodae.html>). A thorough discussion of the Argo program can be found in Appendix C.

4.1.1.7.7 JCOMM Oceanographic Data Exchange

JCOMM is the Joint WMO-IOC Technical Commission for Oceanography and Meteorology and is responsible for the implementation and international co-ordination of operational oceanography. In this role they have produced a “general guide to the operational procedures for the collection, encoding, quality control and exchange of oceanic surface and sub-surface temperature, salinity and current (BATHY, TESAC and TRACKOB) data.” This document can be found at <http://ioc.unesco.org/iocweb/iocpub/iocpdf/m03.pdf>.

4.1.1.7.8 IOC/IODE Ocean Teacher Data Management Resource Kit

This Web site (<http://ioc.unesco.org/oceanteacher/resourcekit/index.htm>) provides a basic foundation for understanding what is required to establish an oceanographic data center in the context of the Intergovernmental Oceanographic Commission (IOC) / International Oceanographic Data and Information Exchange (IODE) organizations. While much of the information is very basic, the Web site does provide a good primer for people lacking in either oceanographic or systems backgrounds.

³ See <http://www.bom.gov.au/bmrc/ocean/GODAE/>

4.1.2 Repository Details and Contacts

This section provides repository data details, where available, and contact people for the repositories and programs discussed in the previous section.

4.1.2.1 West Coast and Polar Regions Undersea Research Center

The Undersea Research Center in Fairbanks Alaska manages the NOAA Undersea Research Program for the West Coast and Polar Region. The Center leases and makes available to marine scientists an array of advanced undersea technology systems, including low-cost and advanced remotely operated vehicles, geophysical instruments such as side-scan sonars and high-resolution seismic reflection systems, and shallow and deep-diving submersibles. The contact information for this center is:

PO Box 757220

213 O'Neill

Fairbanks, AK 99775-7220

907.474.5870 phone

907.474.5804 fax

westnurc@guru.uaf.edu

Dr. Ray Highsmith, Director; email: highsmith@guru.uaf.edu

4.1.2.2 Pacific Marine Environmental Laboratory (PMEL)

PMEL is involved in several programs relevant to VENUS/NEPTUNE, as follows.

4.1.2.2.1 NeMO (New Millennium Observatory)

NeMO is located at Axial Seamount, an active volcano located along the Juan de Fuca Ridge, 250 miles off the coast of Washington/Oregon. NeMO performs measurements using RAS samplers (measuring temperature every 10 minutes, and pH and H₂S twice per week), and bottom pressure recorders, taking measurements four times per hour. (See

<http://www.pmel.noaa.gov/vents/nemo/realtime/index.html>)

Contact: Dr. David A. Butterfield; email: butterfield@pmel.noaa.gov

4.1.2.2.2 TAO-TRITON

Web site: <http://www.pmel.noaa.gov/tao/index.shtml>)

Contact: atlasrt@noaa.gov

Contact: Dr. Michael McPhaden, TAO Project Director,

Michael.J.Mcphaden@noaa.gov

Contact information:

TAO Project Office

NOAA - Pacific Marine Environmental Laboratory

7600 Sand Point Way NE

Seattle, WA 98115

4.1.2.2.3 Argo Profiling Floats

Web site: <http://floats.pmel.noaa.gov/>

Contact: Dr. Gregory Johnson; email gregory.c.johnson@noaa.gov

4.1.2.2.4 Vents Program Acoustic Monitoring (Seismic Activity Localization)

Web site: <http://www.pmel.noaa.gov/vents/acoustics.html>

Contact : Dr. Robert P. Dziak, Principal Investigator; email Robert.P.Dziak@noaa.gov

4.1.2.2.5 Measurement of Partial Pressure of CO₂, via Ship Cruises

Since 1985 PMEL has been measuring surface sea water pCO₂ during sea cruises of 5 NOAA ships. Several of these cruises have covered the North East Pacific area.

Web site: http://www.pmel.noaa.gov/uwpc2/north_pac.html.

Contact: dana.greeley@noaa.gov, richard.a.feely@noaa.gov, chris.sabine@noaa.gov

4.1.2.2.6 Global Surface Seawater Dimethylsulfide (DMS) Dataserver

PMEL provides a dataserver for delivering surface seawater DMS concentrations.

Web site: <http://saga.pmel.noaa.gov/dms>

Contact: james.e.johnson@noaa.gov

4.1.2.2.7 EPIC

EPIC provides public and PMEL internal access to PMEL data (e.g., CTD, XBT, bottle data, time series from moored instruments, shipboard ADCP data, and drifting buoy data). Currently, EPIC consists of 90000 data sets.

Contacts: Donald W. Denbo - donald.w.denbo@noaa.gov - 206-526-4487 (Lead)

Mick Spillane - mick.spillane@noaa.gov - 206-526-6780

Willa H. Zhu - willi.zhu@noaa.gov - 206-526-6208

Nancy N. Soreide - nancy.n.soreide@noaa.gov - 206-526-6728 (Advisor)

Email: epic@noaa.gov

4.1.2.3 National Data Buoy Center (NDBC)

One of the programs undertaken at NDBC is DART (Deep-ocean Assessment and Reporting of Tsunamis). DART measures bottom pressure every 15 minutes; when an event is detected, measurements are taken every 15 seconds to a minute for four hours. DART consists of a system of 6 buoys/units.

Generally, the system of buoys operated by NDBC takes hourly measurements of air temperature, sea surface temperature, dewpoint temperature, sea level pressure, wind speed and direction, gust speed, wave height, average and dominant wave period, mean

wave direction, and water height. Some buoys also record current profiles (ADCP) and salinity.

Web site: <http://www.nodc.noaa.gov/General/NODC-Archive/f291.html>

Contact:

Email: webmaster.ndbc@noaa.gov

Mailing address and telephone number:

Data Products Team, Operations Branch

National Data Buoy Center

1100 Balch Blvd.

Stennis Space Center, MS 39529

228-688-2805

4.1.2.4 National Climatic Data Center (NCDC)

The NCDC operates the following programs relevant to VENUS/NEPTUNE:

4.1.2.4.1 Global Buoy (CLIMVIS)

CLIMVIS measures sea surface temperature, pressure, and precipitation.

Web site: <http://www.ncdc.noaa.gov/cgi-bin/res40.pl>

Contact: ncdc.info@noaa.gov

4.1.2.4.2 Voluntary Observing Ship CLIMate Project (VOSCLIM)

This project gathers sea surface information. VOSCLIM is more meteorologically-based than the Ship-of-Opportunity program (SOOP), discussed in this document.

Web site: <http://www.ncdc.noaa.gov/oa/climate/vosclim/vosclim.html>

Contact: ncdc.info@noaa.gov

4.1.2.4.3 National Geophysical Data Center (NGDC)

The NGDC repositories hold gridded and trackline bathymetry data, sediment thickness data, and seismic data.

Bathymetry contact:

David.Divins@noaa.gov

GEODAS (GEOphysical DATA System) technical contacts:

Dan.R.Metzger@noaa.gov, 303-497-6542

John.G.Campagnoli@noaa.gov, 303-497-3158

Geology contact:

Carla.J.Moore@noaa.gov

4.1.2.5 National Ocean Data Center (NODC)

Data Acquisition Specialist: Francis.Mitchell@noaa.gov

Contact: Donald.Collins@noaa.gov

4.1.2.5.1 Buoy data

See Section 4.1.2.3.

4.1.2.5.2 Joint Archive for Shipboard ADCP

This archive currently consists of data from 780 cruises; values reported are hourly averages and 10 m depth increments.

Web site: <http://ilikai.soest.hawaii.edu/sadcp>

Contact: Mr. Patrick Caldwell, caldwell@hawaii.edu

4.1.2.5.3 Coastal Ocean Time Series Database

This database contains time series measurements of horizontal velocity, vertical velocity, water pressure, water temperature, and salinity at a specific location (specific depth), with the time series including data at ten minute intervals for periods on the order of months.

Web site: <http://www.nodc.noaa.gov/dsdt/tsdb>

Contact: Dr. Wayne L. Wilmot, Chief – Coastal Ocean Laboratory
301-713-3272 x119
Wayne.L.Wilmot@noaa.gov

4.1.2.5.4 Global Temperature-Salinity Profile Program (GTSP)

Web site: <http://www.nodc.noaa.gov/GTSP/gtspp-home.html>

Contact: Dr. Wayne L. Wilmot, Chief – Coastal Ocean Laboratory
301-713-3272 x119
Wayne.L.Wilmot@noaa.gov

4.1.2.5.5 Marine Environmental Data Service (MEDS)

Contact: Dr. Savithri (Savi) Narayanan, Director
Email: narayanans@dfo-mpo.gc.ca

MEDS is involved in the following programs:

4.1.2.5.6 Argo

Web site: http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Prog_Int/argo/ArgoHome_e.html

Contacts: Dr. Howard Freeland (IOS): freelandhj@pac.dfo-mpo.gc.ca

4.1.2.5.7 Global Temperature-Salinity Profile Program (GTSP)

Web site: http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Prog_Int/GTSP/GTSP_e.htm

Contact: services@meds-sdmm.dfo-mpo.gc.ca

4.1.2.5.8 RNODC for Drifting Buoy Data

MEDS is the world data centre for drifting buoys (Responsible National Oceanographic Data Centre - RNODC). As part of its role, MEDS acquires, processes, quality controls and archives real time drifting buoy messages reporting over the Global Telecommunications System (GTS) as well as delayed mode data acquired from other sources. Over 200,000 new records are captured monthly from the GTS.

Web site: http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Prog_Int/RNODC/RNODC_e.htm

Contact: service@meds-sdmm.dfo-mpo.gc.ca

4.1.2.6 Institute of Ocean Sciences/OSAP

4.1.2.6.1 Station-P / Line-P program

The Institute of Ocean Sciences (IOS) has CTD (pressure, temperature, salinity, and transmissometer) profiles and chemistry profile data (oxygen and various nutrient levels) for each of (currently) 26 fixed points along a line (Line-P) extending from the mouth of the Juan de Fuca Strait to the former weather station position 50°N 145°W (Station-P). Data were collected daily at Station P between 1949 and 1981. Line-P measurements have been taken between 1959 and the present, with measurements being taken 3-6 times per year via IOS cruises.

Web site: http://www-sci.pac.dfo-mpo.gc.ca/osap/projects/linepdata/default_e.htm

Contact: Mr. Frank Whitney, program coordinator
whitneyf@dfo-mpo.gc.ca

4.1.2.6.2 Marine Ecosystem ObServations (MEOS)

Fisheries and Oceans Canada and Environment Canada are currently planning to deploy new instruments on their network of weather buoys. These instruments will measure irradiance, air temperature, water temperature, salinity, wind speed and direction, fluorescence (phytoplankton chlorophyll), and zooplankton concentration. Measurements are currently being taken on two buoys (Saanich Inlet and Halibut Bank, between Nanaimo and Gibsons).

Web site: http://www-sci.pac.dfo-mpo.gc.ca/ecobuoys/default_e.htm

Contact: Jim Gower, IOS, GowerJ@pac.dfo-mpo.gc.ca

4.1.2.6.3 North Pacific Marine Science Organization (PICES)

This is an intergovernmental scientific organization, “established in 1992 to promote and coordinate marine research in the northern North Pacific and adjacent seas.” PICES is currently involved in the following projects:

- North Pacific Ecosystem Metadatabase
- The Continuous Plankton Recorder Survey of the North Pacific
- The PICES Carbon Dioxide Related Data Integration for the North Pacific (PICNIC)

Web site: <http://www.pices.int/>

Contact: Mr. Robin Brown, brownro@dfo-mpo.gc.ca

4.1.2.7 World Data Centers and RNODC's

Several RNODC's (Responsible National Oceanographic Data Center) have been established to perform a specialist role in assisting the World Data Centers for Oceanography. This section lists the contact information for those World Data Centers and RNODC's that may be relevant to VENUS/NEPTUNE, as follows:

4.1.2.7.1 WDC for Marine Environmental Sciences

Web site: <http://www.ngdc.noaa.gov/wdc/europe/mes.html> (Berlin, Germany)

Prof. Dr. Gerold Wefer, Director (info@wdc-mare.org)

4.1.2.7.2 WDC for Marine Geology and Geophysics

Web site: <http://www.ngdc.noaa.gov/wdc/usa/mgg.html> (Boulder, CO, USA)

Dr. George F Sharman, Director (George.F.Sharman@noaa.gov)

4.1.2.7.3 WDC for Oceanography

Web site: <http://www.ngdc.noaa.gov/wdc/usa/ocean.html> (Silver Spring, MD, USA)

Mr Sydney Levitus, Director (Sydney.Levitus@noaa.gov)

The WDC for Oceanography was originally established during the International Geophysical Year of 1957-58 and is operated by the U.S. National Oceanographic Data Center (NODC). The following URL describes its function in detail:

<http://www.nodc.noaa.gov/General/NODC-dataexch/NODC-wdca.html>

The WDC for Oceanography intakes data collected in conjunction with the Oceanographic Commission (IOC) Global Oceanographic Data Archaeology and Rescue (GODAR) and World Ocean Database (WOD) projects and quality-controlled and post-processing by the NOAA Ocean Climate Laboratory (<http://www.nodc.noaa.gov/OC5/>). In turn it produces the World Ocean Database. This database contains a collection of scientifically quality controlled ocean profile and plankton data that includes measurements of temperature, salinity, oxygen, phosphate, nitrate, silicate, chlorophyll, alkalinity, pH, pCO₂, and tCO₂. These and all associated products and data are described at <http://www.nodc.noaa.gov/OC5/indprod.html>.

New versions of the WOD are normally released every two or three years as time and

funding permit. During this time the WDC for Oceanography continues to add new data and replace old data that have been further post-processed. These data are not normally released until the official new version is ready, but some updates can be found at http://www.nodc.noaa.gov/OC5/WOD01/pr_wod01.html.

4.1.2.7.4 WDC for Seismology

Web site: <http://www.ngdc.noaa.gov/wdc/usa/seismol.html> (Denver, CO, USA)
Dr Stuart Sipkin, Director (sipkin@usgs.gov)

4.1.2.8 RNODC for IGOSS (BATHY and TESAC): Japan, USA and Russian Federation

Operated by the NODC of USA (NOAA, Silver Spring, MD) (see above).

4.1.2.8.1 RNODC-ADCP: Japan

Operated by the NODC of Japan (<http://www.jodc.go.jp/>).

4.1.2.8.2 RNODC for Drifting Buoy Data: Canada (MEDS)

Web site: http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Prog_Int/RNODC/RNODC_e.htm
Contact: service@meds-sdmm.dfo-mpo.gc.ca

4.1.3 Soliciting Information from Data Repositories

A goal of Phase 1 of this project was to understand how repositories currently manage their data, and, from this, glean relevant lessons for the VENUS/NEPTUNE DMAS. Two questionnaires were designed and sent out: one targeted at organizations that are collecting measurements and forwarding on related information (data and metadata), and one targeted at organizations that are receiving this information (and further forwarding it or archiving it). These questionnaires, and the responses obtained, are given in Appendix B.

4.2 ***Ocean Observing Systems***

Ocean observing systems considered in this phase were:

- 1) TAO – TRITON (PMEL)
- 2) EPIC (PMEL)
- 3) Coriolis Project (France)
- 4) NERC Datagrid (UK)

- 5) USGODAE (Monterey)
- 6) NASA EOS – EOSDIS (Physical Oceanography DAAC)

4.2.1 TAO – TRITON

TAO-TRITON is an array of approximately 70 moored buoys arranged in a grid-like fashion in the Equatorial Pacific Ocean. Instruments at these buoys measure wind speed and direction, air temperature, relative humidity, short wave radiation, precipitation, water temperature, salinity and current profile (via ADCP). The water temperature and salinity are measured down to 500 metres in depth. The buoys transmit some data continuously, via satellite telemetry (Service Argos), during several windows each day. The Pacific Marine Environment Laboratory (PMEL) receives these data from Service Argos in the form of single daily deliveries. Other data, including the profiles, is held on the buoys and recovered by PMEL when the instrumentation on the buoys is serviced or replaced in the field.

As part of the data quality monitoring, PMEL identifies and informs Service Argos of data that should not be submitted to the Global Telecommunications System (GTS) for distribution to other organizations. Typically, about 80% to 90% of data that is received at PMEL make it onto the GTS. Due to a variety of factors there is a finite time lag between observations and availability on the GTS. Over half of the TAO data which are placed on the GTS are available within 3 hours or less and over 90% are available within 7 hours. See http://www.pmel.noaa.gov/tao/proj_over/gts.html for a description of GTS data delivery, and http://www.pmel.noaa.gov/tao/proj_over/qc.html for a description of TAO data quality control.

Figure 2 below (from http://www.pmel.noaa.gov/tao/proj_over/diagrams/gif/tao_sys.jpg) illustrates these processes.

PMEL stores TAO data in flat files, with a Web-application-accessed metadata database stored using MySQL. The following Web sites can be used to retrieve TAO data:

- a) data delivery: http://www.pmel.noaa.gov/tao/data_deliv/deliv.html
- b) data display: <http://www.pmel.noaa.gov/tao/jsdisplay/>
- c) combined: <http://www.pmel.noaa.gov/tao/disdel/>

An “availability” button on the “combined delivery and display panel” can be used to determine in advance whether data are available for a particular buoy and date.

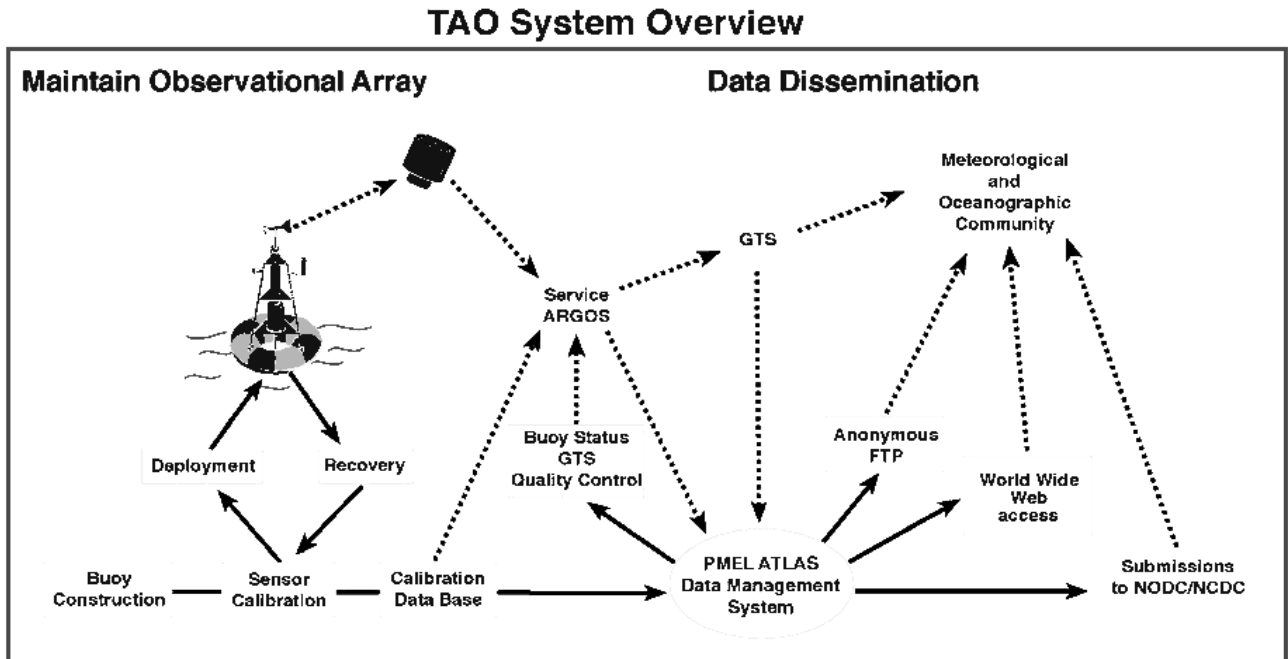


Figure 2. Overview of TAO system.

4.2.2 EPIC (PMEL)

EPIC was developed at NOAA's Pacific Marine Environmental Laboratory (PMEL) to manage the large numbers of hydrographic and time series oceanographic *in situ* data sets collected as part of NOAA climate study programs, such as EPOCS and TOGA in earlier years, WOCE and CLIVAR and others more recently. Data types include CTD, XBT, bottle data, time series from moored instruments, shipboard ADCP data, and drifting buoy data. The EPIC system provides data archival, retrieval, display and analysis procedures for oceanographic time series and hydrographic data. Users select data by specifying data type, latitude, longitude and time range or other identifying characteristics. There is a complete suite of routines for graphical display and data analysis. Multiple data file formats (including netCDF) are supported for access from C and Fortran, using the EPIC system library EPSLIB (http://www.pmel.noaa.gov/epic/eps-manual/epslib_toc.html). Matlab compatibility is provided through MexEPS for Matlab (http://www.pmel.noaa.gov/epic/eps-manual/epslib_toc.html). In addition to data analysis and display functionality, EPIC also includes MySQL-based database software that can be used to manage EPIC data once it has been downloaded to the desktop (see http://www.pmel.noaa.gov/epic/software/ep_dbprograms.htm).

At present, PMEL maintains approximately 100,000 individual data sets in the EPIC. These data are available on-line to researchers inside PMEL on the desktop and on the Web via PMEL's internal Intranet. Portions of the data are publicly available outside PMEL via the World Wide Web (<http://www.epic.noaa.gov/epic/>).

All EPIC software is freely available via FTP
(<http://www.pmel.noaa.gov/epic/download/index.html>).

4.2.3 Coriolis Project (France)

The seven institutes involved in operational oceanography in France (CNES, CNRS, IFREMER, IPEV, IRD, Météo-France, SHOM) have created the Coriolis project (2002-2005) in order to

- organize data collection in real time and delayed mode of *in situ* measurements necessary for operational oceanography,
- set up an operational *in situ* data center,
- develop and improve the technology necessary for operational oceanography.

Data for Coriolis come from measurements from ships, from moored buoys, and from drifting buoys. Ship measurements currently include XBT and thermosalinograph, with support for ADCP coming later in 2004.

The Coriolis project will contribute to the Argo program by deploying up to 250 PROVOR floats, by the end of 2005, in the Atlantic and Indian Oceans.

The Coriolis data center provides QC'd data in both real time and delayed mode. These data are available over the Web at <http://www.coriolis.eu.org>. Documentation on QC applied by Coriolis is available at <http://www.coriolis.eu.org/cdc/documents/cordo-rap-04-047.pdf>.

4.2.4 NERC Datagrid (UK)

The primary aim of the NERC Datagrid (NDG) project is to provide access to data holdings that lie in a (possibly very) loosely coupled federation of sites sharing at a minimum a set of common discovery, authentication, authorization, and access protocols. Even though much of the NDG data are publicly available, the NDG aims to provide secure access to protected data where trust relationships exist between data providers. The NDG design doesn't assume one central repository for all metadata, nor does it assume common policies. However, following the "search engine" philosophy, search engines will be free to harvest publicly available discovery metadata using standard digital library protocols to produce high performance discovery engines. Figure 3 (taken from http://ndg.badc.rl.ac.uk/public_docs/AHM2004Woolf.ppt) illustrates the enterprise level design of the NDG.

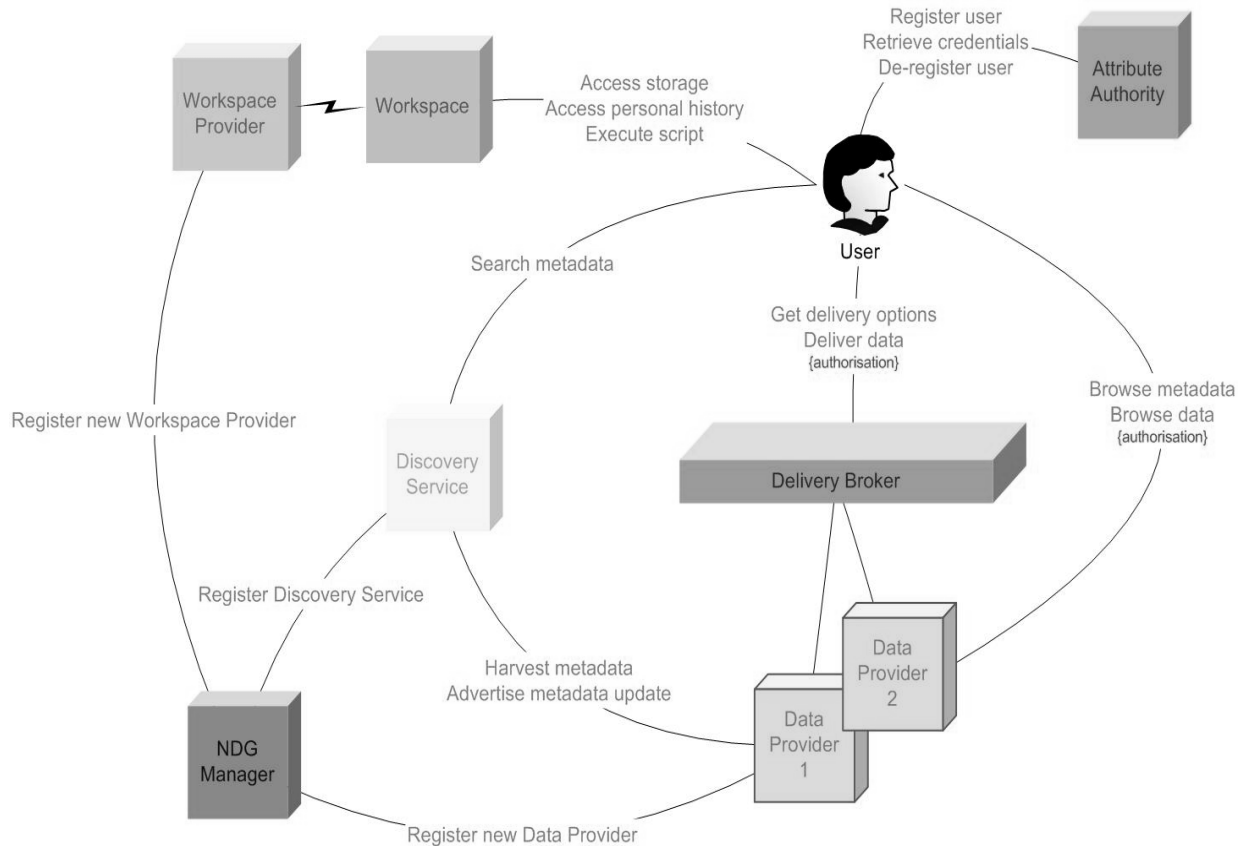


Figure 3. Enterprise level design of the NERC Datagrid (NDG).

While the “datagrid” approach may not be appropriate for VENUS/NEPTUNE at this stage, NDC appears to have taken ground-breaking steps in the areas of data discovery, user authentication, and user authorization. These aspects of NDC may serve as a model for VENUS/NEPTUNE.

4.2.5 USGODAE (Monterey)

The USGODAE (US Global Ocean Data Assimilation Experiment) Monterey Data Server (<http://www.usgodae.org>) resides at the Fleet Numerical Meteorology and Oceanography Center (FNMOC), and was developed through a partnership of FNMOC and the NOAA Pacific Marine Environmental Lab (PMEL). This server hosts near real time *in situ* oceanographic data available from the GTS and other FTP sites, atmospheric forcing fields suitable for driving ocean models, and unique GODAE data sets, including demonstration ocean model products. It supports GODAE participants, as well as the broader oceanographic research community.

Data from the various sources arrive in a scheduled fashion, hourly, four or six times per day, or daily, depending on the source.

All of the USGODAE data sets are available in their original format via HTTP and FTP. In addition, USGODAE data are served by Local Data Manager (LDM), OpenDAP (formerly DODS) and the GODAE Live Access Server (LASGODAE) from PMEL.

LASGODAE facilitates online visualization of gridded and *in situ* GODAE data sets. The application enables Web users to visualize data with on-the-fly graphics, request custom subsets of variables in a choice of file formats, access background reference material about the data (metadata), and compare (difference) variables from distributed locations. Data may be selected by provider, measurement, or measuring device. A thumbnail generator can be used to create preview plots on the server.

To provide global surface forcing for ocean model research, USGODAE serves grids from the Navy Operational Global Atmospheric Prediction System (NOGAPS) and Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). These grids are produced four times per day.

Observational data from the FNMOC Ocean QC process are posted in near real time, to USGODAE. These data files contain both the measured observations, and all of the climate and background (forecast) values used in the QC process. The FNMOC QC includes temperature and salinity profiles, sea wave height and tides, *in situ* SST and ship track, high resolution satellite SST/flux/sea ice products. Also, the Naval Oceanographic Office (NAVO) provides daily satellite level 2 SST and SSH retrievals to USGODAE.

The USGODAE server acts as the North American mirror for the METEO (France) Satellite Application Facility SAF FTP site, serving high-resolution, satellite-derived SST fields, structures, and radiative fluxes.

The USGODAE server is also one of two Argo Global Data Assembly Centers, serving data from the Argo global array of temperature/salinity profiling floats (the other center being IFREMER/Coriolis in France). The complete set of global real time and delayed-mode (QC'd) float data is available on USGODAE.

The approximate volume of data stored on the USGODAE server is 1.5 terabytes.

4.2.6 NASA EOS – EOSDIS (Physical Oceanography DAAC)

The Physical Oceanography Distributed Active Archive Center (PO.DAAC) is responsible for archiving and distributing data relevant to the physical state of the ocean. They currently deal with three satellite data sets: sea surface temperature, sea surface height, and sea surface winds. The PO.DAAC distributes products from NASA's Earth Science Enterprise (ESE) Earth Observing System (EOS) project, as well as through partnerships and cooperative agreements with other organizations and institutions. These

groups include the Jet Propulsion Laboratory (JPL) and other NASA centers, NOAA, the U.S. Naval Oceanographic Office (NAVOCEANO), the French Space Agency, the Centre National d'Etudes Spatiales (CNES), the National Space Development Agency (NASDA) of Japan, as well as the University of Miami and Brigham Young University. PO.DAAC's data development, management, archival, and distribution activities are guided in large part by a group of colleagues in the form of an EOSDIS advisory User Working Group (PO.DAAC UWG).

Currently about 70-80% of the community are researchers, but usage statistics show that this is changing. As data become available more quickly (and become more near real time) they can become useful for operational purposes.

Web interfaces have been developed that allow users to extract data (e.g., specify area rather than global data sets, which reduces file transfer requirements). The main interface is called POET and can be cursor or form driven. Data are assembled and put on the FTP site; the user receives an email when it is ready. Most requests take no more than a few minutes. A section of the Web site is dedicated to "known problems" where feedback is documented. This serves as useful information for the ongoing development of the system. The PO.DAAC is trying to develop better FAQs as data holdings and streams are growing while the support resources remain constant (need to maintain efficiency).

4.2.7 Ocean.US (IOOS) DMAC Subsystem Implementation Plan

On May 10, 2004, the US National Office for Integrated and Sustained Ocean Observations (<http://www.ocean.us>) published an implementation plan for the Data Management and Communications (DMAC) subsystem of the Integrated Ocean Observing System (IOOS) (see http://dmac.ocean.us/dacsc/imp_plan.jsp). The table of contents of this report has been reproduced in Appendix D. The document defines the following core functions of the DMAC:

1. **Data transport.** The DMAC shall provide capability for the collection/transmission of data from sensor subsystems at entry points where the data become available using DMAC standards and protocols either on the Internet or a supplied IOOS backbone, to assembly centers, users, and archive centers in real time and delayed mode, for operational, research, and product generation applications.
2. **Quality control.** The DMAC shall provide a mechanism for assuring that data are of known, documented quality. QC operations are a partnership among data observation/collection components, processors, analysts, other users, and the DMAC.
3. **Data assembly.** The DMAC shall provide mechanisms for aggregation and buffering of data streams over useful spans of time and space. Data assembly allows users to more easily exploit real time data, especially data from distributed sensor arrays.
4. **Product generation.** Products include data products such as assimilation-friendly, real time measurements, model nowcasts and forecasts, GIS layers and climatological reference fields; graphical information products such as scientific plots and maps; and text information products such as written forecasts and numerical tables. The DMAC will provide a minimal level product-generation capability, only—the guarantee of a uniform, interactive, geo- and time-referenced browse capability suitable for quick evaluation of data by IOOS scientists. Most product generation is the responsibility of the IOOS Modeling, Data Assimilation Subsystem and the value-added information product producers that will address the needs of specialized end-user groups.

5. **Metadata management.** The DMAC shall provide simple, clear guidelines and extensible standards for metadata; ensure that the linkages between data and metadata are maintained with great reliability; provide for communication of metadata among components of the system; provide training and tools to increase end users' and data providers' capacity in metadata generation and management.
6. **Data archeology.** The DMAC shall directly or indirectly facilitate activities to rescue, digitize, and provide access to legacy/historical data sets; retrieve data in danger of loss due to deteriorating media, out-of-date software, not in digital format, etc.
7. **Data archival.** The DMAC shall provide for the long-term archive and stewardship of IOOS data sets; conform to national archive standards, as well as IOOS standards and user requirements.
8. **Data discovery.** The DMAC shall provide a means for determining what data are available within the IOOS based upon queries that may be issued by users or by other machines. Data Discovery shall be seamlessly integrated with data and metadata access functions provided by the Data Transport and Metadata Management components, respectively.
9. **Administrative functions.** The DMAC shall provide oversight mechanisms to ensure the proper functioning and smooth evolution of IOOS. These include fault detection and correction, security, monitoring and evaluation of system performance, providing for system extensibility, establishing and publicizing policies for data availability, soliciting and responding to user feedback, and establishing and maintaining international linkages.

In the section "Immediate Priorities for Implementation – Concrete Guidance for Data Providers" the report makes the following recommendations:

1. That all data providers:
 - i) create metadata that are compliant with the Federal Geographic Data Committee (FGDC) standards for both current and legacy data holdings and inventories.
 - ii) submit metadata to the NASA Global Change Master Directory (GCMD) and/or the NOAA Coastal Data Development Center (NCDDC) so that data sets may be easily found through an open data discovery process.
 - iii) participate in the DMAC Metadata Working Group to ensure that the special characteristics of their data will be thoroughly considered during the formulation of DMAC metadata standards.
2. That data providers of:
 - i) *gridded data* install servers providing access to their data through OPeNDAP data access protocol. (OPeNDAP is an *operational* component of IOOS for access to gridded data. OPeNDAP servers are available for download without licensing costs.)
 - ii) *complex data collections in a relational data base (SQL)* make data accessible to DMAC by installing an OPeNDAP relational data base server. (OPeNDAP is a *pilot* component of IOOS for access to unstructured data collections.) Full operational support for relational data bases will be developed early in the evolution of DMAC.
 - iii) *Geographic Information System (GIS) data collections (for groups that are already participating in enterprise GIS networks)* continue pursuing these efforts. Gateways that provide translation from the GIS network protocols to the OPeNDAP protocol will be developed by IOOS.

- iv) *large collections of files that comprise a single (logical) data set* install an OPeNDAP “aggregation server” or participate in a DMAC data aggregation pilot project.
3. That all data providers :
 - i) participate in the DMAC Transport (Semantic Data Model) Working Group to ensure that the special characteristics of their data (if any) will be thoroughly considered during the formulation of DMAC data transport standards.
4. That all data providers either:
 - i) install a Live Action Server and notify DMAC (at Ocean.US) of its existence, so that it can be integrated into the community-wide data browsing environment, or
 - ii) notify the National Virtual Ocean Data System (NVODS) that their data set has become available through OPeNDAP and request that it be added to the community-wide data browsing capability.
5. That all data providers:
 - i) review their current data holdings to ensure that irreplaceable data are archived at a responsible entity;
 - ii) contact the archive entity that is responsible for their classes of data and make arrangements for archiving the data.

In addition to making the suggestions on immediate actions listed above, the report states that the following developments may have an effect on future standards and practices:

- Advanced data discovery techniques, such as the “Semantic Web” (www.w3.org/2001/sw/), will be monitored and considered for incorporation into the system as they mature.
 - Strategies for storage media migration will be included.
 - “Data Mining”, the ability to search within the actual data, will be incorporated as the technologies for doing so mature.
 - The following data systems will be considered for integration with the DMAC:
 - Storage Resource Broker (SRB), a generic data management infrastructure;
 - Ocean Biogeographic Information System (OBIS), “a globally distributed network of systematic, ecological, and environmental information systems”;
 - OpenGIS Web Mapping Services and Geospatial Fusion Services;
- Data Exchange Infrastructure (DEI) and the Field Spatial Data Model (FSDM).

4.3 Data Organizations With Other-Than-Oceanographic Data

Large-scale *other-than-oceanographic* scientific data management organizations considered in this phase were:

- 1) Incorporated Research Institutes for Seismology(IRIS)
- 2) Fleet Numerical Meteorological and Oceanography Center (FNMOC)
- 3) NOAA NESDIS (satellite)

4) Canadian National Data Centre for Earthquake Seismology and Nuclear Explosion Monitoring

4.3.1 Incorporated Research Institutes for Seismology (IRIS)

The Incorporated Research Institutions for Seismology (IRIS) is a university research consortium dedicated to exploring the Earth's interior through the collection and distribution of seismographic data. IRIS programs contribute to scholarly research, education, earthquake hazard mitigation, and the verification of the Comprehensive Test Ban Treaty. Support for IRIS comes from the National Science Foundation, other US federal agencies, universities, and private foundations.

The IRIS Global Seismographic Network (GSN) is one of the four major components of the IRIS Consortium. The goal of the GSN is to deploy over 128 permanent seismic recording stations uniformly over the Earth's surface. The GSN provides funding to two network operation centers: IRIS/ASL in Albuquerque, NM (operated by USGS), and IRIS/IDA in La Jolla, CA (operated by Scripps Institute of Oceanography). As of 2003 the IRIS GSN was made up of over 128 stations with affiliations to USGS, UCSD/IDA, GEOFON, Pacific21, NCDSN, GEOSCOPE, MedNet, BGR, BFO, USNSN, BDSN, TriNet, AFTAC and several other national and international networks. The IRIS GSN stations continuously record seismic data from very broad band seismometers at 20 samples per second (sps), and provide for high-frequency (40 sps) and strong-motion (1 and 100 sps) sensors where scientifically warranted. It is also the goal of the GSN to provide for real time access to its data via Internet or satellite. Over 75% of the IRIS GSN stations meet this goal.

There are basically two types of waveform data stored at the IRIS Data Management Center (DNC): continuous and event-related.

For access to continuous data there are two options. If the user wants near real time, but not QC'd, data, an online Buffer of Uniform Data (BUD), updated continuously and organized by channel day, can be accessed. Alternatively, the user can specify parameters to extract delayed, QC'd data.

SPYDER data are event-oriented data products created shortly after an earthquake occurs. This un-QC'd data come from a variety of sources, primarily from the BUD system described above, but also from autoDRMs around the world and from direct station dial-up.

The DMC routinely pre-assembles data from earthquakes that exceed magnitude 5.7 at any depth, and for events down to magnitude 5.5 if the depth is greater than 100 km. These pre-assembled data sets consist of data collected from stations all over the world and from many different networks. See <http://www.iris.edu/manuals/DATutorial.htm> for more information.

4.3.2 Fleet Numerical Meteorology and Oceanography Center (FNMOC)

FNMOC is one of the US Navy's two MetOc (METeorological and OCeanographic) productions centers (the other being NAVO – Naval Oceanographic Office), with FNMOC dealing primarily (but not strictly) with atmospheric data, and NAVO dealing primarily with oceanographic data.

FNMOC serves as the US Department of Defense's primary numerical prediction center for operational meteorological and oceanographic analysis and forecast products worldwide. The center uses several supercomputers to produce over 450,000 data products each day. These products are distributed to users (military and non-military) around the world through a variety of communication links, dial-up networks and the Internet.

FNMOC has developed *Metcast*, a request-reply and subscription system for distributing, disseminating, publishing and broadcasting of real time weather information. It is a data-oriented Web service, acting as an intermediary between clients communicating over HTTP and several meteorological databases. These databases contain weather observation reports (from stations, ships at sea, and buoys), forecasts, advisories, gridded data produced by weather models, as well as satellite imagery and plain text messages and discussions. A sub-system of Metcast includes a comprehensive set of message decoders that convert complex (and often ill-formed) World Meteorological Organization (WMO) bulletins into well-tagged XML documents.

4.3.3 NOAA NESDIS (Satellite)

The role of the National Environmental Satellite, Data, and Information Service (NESDIS) is to "provide timely access to global environmental data from satellites and other sources." To this end, NESDIS currently operates 16 meteorological satellites in geostationary (GOES) or Polar-orbiting (POES) constellations.

One of the satellite programs NESDIS is involved in is the Defence Meteorological Satellite Program (DMSP). Data from four polar-orbiting satellites are downlinked to Thule AFB and transmitted to the Air Force Weather Agency (AFWA) via communications satellite (AFWA and the US Navy Fleet Numerical Meteorological and Oceanography Center (FNMOC) are the two primary users of these data). The data are then decrypted and sent to NOAA/NESDIS National Geophysical Data Center (NGDC) via T1, where they are inventoried and copied to tape. The data are then processed by software developed by the NGDC staff and software contractors. Currently, NGDC receives and processes approximately 8.5 gigabytes of data per day.

The processing of the data involves complete renavigation of the satellite, time re-ordering the data, splitting the data apart by sensor, fixing data problems due to bit flips during transmission, organizing the data into orbits and writing the data to a robotic tape library system.

4.3.4 Canadian National Data Centre for Earthquake Seismology and Nuclear Explosion Monitoring

The Canadian National Data Centre (CNDC) supports the mission-critical Seismic Monitoring needs of Earthquakes Canada and fulfills the role of Canada's Nuclear Explosion Monitoring (NEM) National Data Centre (in support of Comprehensive Test Ban Treaty obligations).

Over 4.3 gigabytes of continuous, compressed seismic ground motion data from the stations of the Canadian National Seismograph Network (CNSN), the POLARIS network, and a few cooperating US stations, are acquired, quality controlled, processed, and archived within the CNDC on a daily basis. Event detection, automatic location, and issuing email alerts are accomplished within minutes of local and regional events. Automatic processing of Yellowknife Array (YKA) data allows global events to be detected, locations refined, and email alerts issued to national and international subscribers within three minutes of p-wave arrival. The National Earthquake Database (NEDB) is the national repository/index for all raw data, phase measurements, and derived parameters.

The National WaveForm Archive (NWFA) contains event-related digital waveforms acquired by regional networks operating in Eastern and Western Canada since 1975, and continuous data from the Canadian National Seismograph Network (CNSN) since 1992. The Automatic Data Request Manager (AutoDRM) program provides convenient access to continuous waveform data plus selected data from the NEDB. The NWFA Web server has been active since April 1995. Access to event-related data has recently been added.

In its role as the Nuclear Explosion Monitoring body, the CNDC supports the development of the CTBT International Monitoring System (IMS) in part by forwarding continuous data in near real time from the Yellowknife Array (YKA) and 4 other stations to the International Data Centre (IDC) in Vienna. The IDC regularly requests additional data from Canada's 6 auxiliary seismic stations via AutoDRM. Continuous data from 10 broad band sites designated "Federation" (FDSN) stations are supplied to the IRIS DMC for archive. Subsets of the data are also sent to Blacknest, UK.

A thorough description of the activities of the CNDC can be found at http://www.seismo.nrcan.gc.ca/cndc/cndc_detail_e.php. Figure 4 below illustrates those activities.

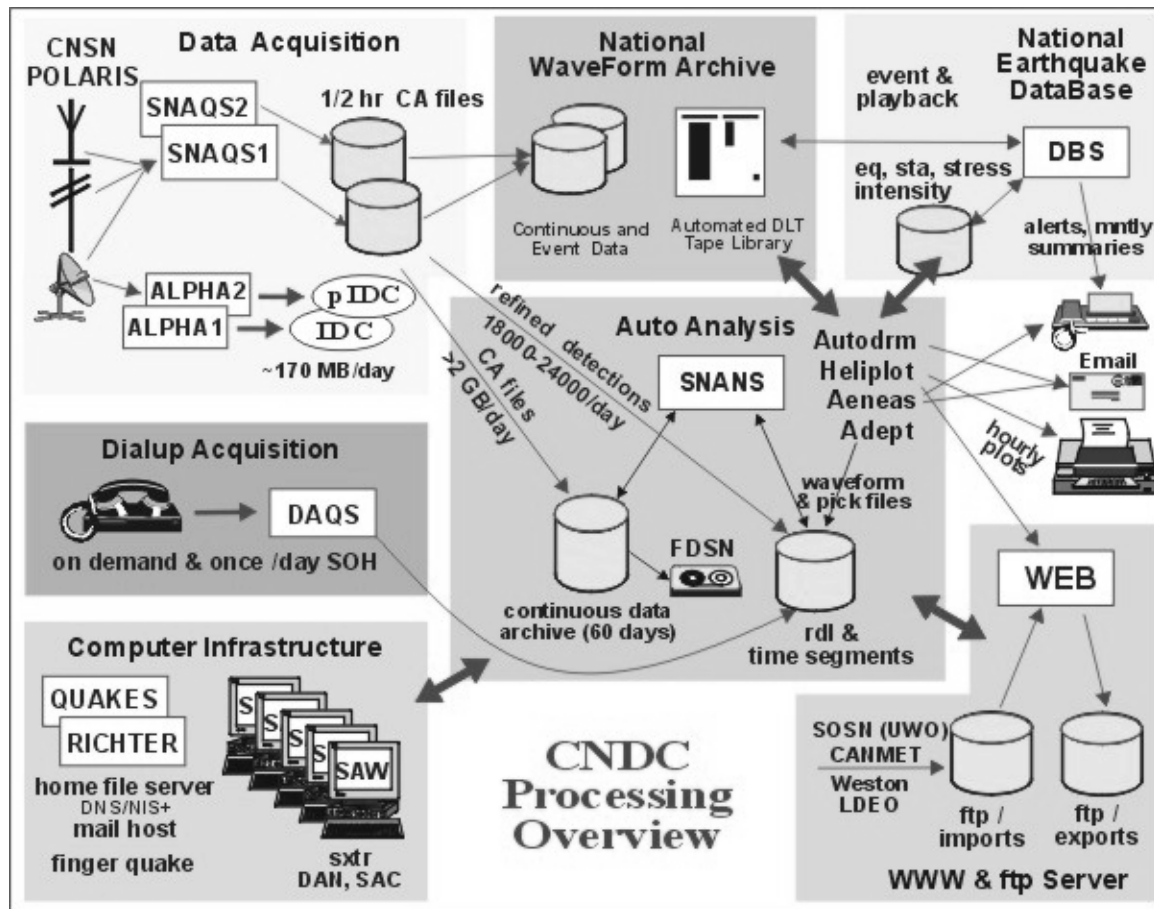


Figure 4. Activities of the CNDC (from http://www.seismo.nrcan.gc.ca/cndc/images/blkdiag0211_600.jpg).

4.4 Local Scientific Data Organizations

Large-scale local scientific data management organizations considered in this phase were:

- 1) Herzberg Institute of Astrophysics – Canadian Astronomy Data Centre
- 2) ATLAS (UVic Department of Physics)
- 3) BC Ministry of Sustainable Resource Management – BC Active Control System
- 4) Pacific Forestry Centre – National Forest Information System
- 5) BC Ministry of Sustainable Resource Management – BC Land and Resource Data Warehouse
- 6) Pacific Geoscience Center
- 7) LANDSAT / RADARSAT

The Canadian Astronomy Data Centre (CADC) was established in 1986 by the National Research Council of Canada (NRC), through a grant provided by the Canadian Space Agency (CSA), as one of three world-wide distribution centres for astronomical data obtained with the Hubble Space Telescope (HST). In subsequent years the CADC's mandate has expanded significantly to include the development and support of archives for other astronomical facilities whose operational costs are shared by the NRC. These include the Canada-France-Hawaii Telescope, the James Clerk Maxwell Telescope, and the twin 8-m telescopes of the Gemini Observatory.

The CADC is also currently developing the Canadian Virtual Observatory (CVO). The CVO will seamlessly combine the content of all of the CADC's archives as well as other publicly available astronomical data sets, and provide archival researchers ready access to astronomical source catalogues derived from these holdings, calibrated images and spectra, and powerful tools to query the content of the CVO in an intuitive manner.

The CADC handles 12 terabytes per year (in and out), and has the capacity to handle 20 terabytes per year. Almost all of these data come from intermediary repositories. A fundamental and important difference between CADC and VENUS/NEPTUNE is that CADC is not directly involved in the operation of instruments (telescopes). Individual contributing observatories serve as a buffer and perform their own QA and data packaging (for example, the HST Space Science Telescope Institute in Baltimore). CADC does, however, perform additional metadata generation and packaging, produces raw and enhanced metadata catalogues, and produces additional data products.

Currently approximately 40 terabytes of unique data (12 million distinct files) are stored on the system (totalling 80 terabytes for 2 copies). The raw metadata database is approximately 100 gigabytes in size, and the enhanced metadata database (data mine) is two terabytes in size and is growing at a rate of one terabyte per year.

ATLAS is an international collaboration of scientists currently building a particle detector that will probe nature at the TeV energy scale, or, equivalently, at the 0.0000000000000000001 m scale. ATLAS will study proton-proton collisions at a center of mass energy of 14 TeV that will be provided by the Large Hadron Collider, LHC, currently under construction at the European Laboratory for Particle Physics, CERN. The particle physics group at the University of Victoria has been an active member of the ATLAS Collaboration since its inception in 1992 (<http://particle.phys.uvic.ca/~web-atlas/atlas/>).

The ATLAS project at UVic uses the Research Computing Facility (<http://rcf.uvic.ca/>), a facility consisting of:

- a Minerva High Performance Computer (with 128 375MHz RS/6000 processors and 64 gigabytes of RAM), designed for parallel applications;

- the Mercury (Linux) Cluster, a Xeon processor-based cluster of 84 Blade servers; and
- Storage, which consists of 36 terabytes of fast disk storage and robotic tape library managing 200 terabytes of online tape storage.

4.4.3 BC Ministry of Sustainable Resource Management – BC Active Control System

In recent years, global positioning system (GPS) technology has advanced to the point where GPS can now be used to augment and replace conventional positioning and surveying techniques.

The BC Active Control System (ACS) is a system that provides GPS corrections in post-mission and real time. Real time communication is supported via a UHF radio link. Real time accuracy levels range from approximately 1 m (using RTCM code corrections and a mapping grade GPS receiver) to centimetre level (using dual frequency survey grade receivers). It provides users with the ability to survey and lay out points, accurate to a few centimetres, instantaneously. The system substantially improves the efficiency of engineers and surveyors as they no longer need to transfer precise coordinate information over long distances from control monuments to perform their duties; the BC ACS gives them the accuracy they need to complete the required tasks in real time.

BC ACS also provides the capability to correct data in a post-mission fashion. Time-stamped survey data are collected in the field and then subsequently corrected by a download of GPS data from BC ACS covering the same time period. BC ACS data are available on a subscription basis through the Ministry or through Land Data BC for per-download access. Currently approximately 60 percent of the users are commercial, 20-30 percent are government agencies, and 10 percent are institutional/educational.

4.4.4 Pacific Forestry Centre – National Forest Information System

The purpose of the National Forest Information System (<http://www.nfis.org/>) is to provide Canadians, and the international community, “with authoritative information about the state of Canada’s forests and how they are being sustainably managed.”

Canada’s forests cover almost 500 million hectares and stretch across the country. Responsibility for these forests falls under a great many jurisdictions and management agencies including governments, industry and other organizations. Organizations that contribute data include the various provincial governments, the federal Department of Fisheries and Oceans (DFO), CTI (Center for Topographic Information, Sherbrooke), Canadian Geoseismology Network, Environment Canada, Industry (Cubewerx, NASA, GEODEN), and UVic.

Data collections are OGC (Open GIS Consortium) compliant and are accessed through the Geoconnections Web portal (<http://cgdi.gc.ca> or <http://www.geoconnections.org>).

The data are organized via a distributed database system with 17 nodes across Canada, with each node representing a jurisdiction. Each node has a large data warehouse with collections corresponding to the node's jurisdiction. Metadata are polled from each of the servers and centralized in an Oracle metadata database at PFC.

Data access is via the Distributed Access Control System (DACS), a system designed and implemented by Distributed Systems Software (DSS) of Richmond, BC (<http://dss.bc.ca/>). DACS is a general-purpose, distributed system that combines Single Sign-On capability and role-based access control for Web services.

Data products include 640x480 PNG files, individual features, shape files, and GML (Geography Markup Language) streams.

4.4.5 BC Ministry of Sustainable Resource Management – BC Land and Resource Data Warehouse

The Land and Resource Data Warehouse (LRDW) is the (BC Government) corporate repository for integrated land, resource and geographic data that supports a variety of business requirements for the natural resource sector, other government agencies, industry and the public.

Data in the LRDW originate from various BC Government Ministries and organizations, and arrive in a variety of formats through a variety of mechanisms.

An interface at <http://srmwww.gov.bc.ca/g/lrdw.html> provides access to the following services:

- i) The Discovery Service allows users to access detailed information about data that resides in the Land and Resource Data Warehouse such as: custodian, frequency of update, quality of the data, format, limitations, contacts and more. It provides several ways to search the data: by category; by selecting an area on a map; and by a standard text search. Once the required information has been found, there is a link to an ordering process which allows the data to be downloaded to the user's computer. The Discovery Service utilizes a Corporate Metadata Repository that supports not only Land Information BC but also metadata holdings contributed by other natural resource sector clients and business partners. The Discovery Service repository is based on the international ISO TC211 19115 geospatial metadata standard (see <http://www.isotc211.org>).
- ii) The Distribution Service allows users to order information for download from the warehouse to their own computers.
- iii) The Access Service allows users to view warehouse spatial and attribute data directly through a Web browser using mapping software.

In addition to this Web interface, data are also available via a terminal server and through FTP. Data in the LRDW are stored in an Oracle database, using an ESRI ArcSDE front end.

4.4.6 Pacific Geoscience Centre

The Pacific Geoscience Centre (PGC) in Sidney, British Columbia, Canada, houses staff of the Geological Survey of Canada (GSC) Pacific Division. The GSC is a part of Natural Resources Canada, the Canadian government department that specializes in energy, minerals and metals, forests and Earth sciences. GSC Pacific Division is one of several Divisions in the Minerals and Regional Geoscience Branch of the GSC, which falls under the Earth Sciences Sector of Natural Resources Canada.

Staff at PGC conduct research into areas of geology and geophysics within the region of Western Canada known as the “Canadian Cordillera”, as well as along the continental margin that is Canada’s West Coast. PGC researchers are involved in national programs on earthquake seismology and geodynamics, including the evaluation of earthquake hazards. They also undertake studies in reflection seismology, geomagnetism, paleomagnetism, geothermics, the Earth’s gravitational field and marine sedimentology.

The principal research programs at the Pacific Geoscience Centre are:

- ***Earthquake Seismology*** - the western component of Canada’s National Earthquake Hazards Program, which is aimed at understanding the causes of, and hazards associated with, earthquakes in Canada.
- ***Geodynamics*** - monitoring and investigating movement of the Earth’s crust in support of research into earthquake hazards and global change.
- ***Cordilleran & Continental Margin Tectonics*** - studies into the geological architecture of the Canadian Cordillera, its history, and recent activity along its tectonic plate boundaries in order to increase understanding of the various processes that have formed, and are forming, this region.
- ***Marine Geoscience*** - research into the marine environment of Canada’s West Coast both from a global and regional geoscience perspective, including seafloor mapping, sediment distribution, active faulting, and geological processes along the continental margin.

Data coming into PGC are transmitted by satellite, microwave, telephone, or radio. Canada has 120 land seismology stations. About 70 percent of the data from these stations come in real time into PGC, and 85 percent come into the Ottawa branch in real time. Automated hourly batch transfers via FTP to and from Ottawa bring the two systems into synch. Some data come from UW (University of Washington) and UAF (University of Alaska at Fairbanks). Data also flow to them and the US national network. These data are exchanged in real time. Incoming data rates are approximately 1.3 gigabytes per day (0.5 terabytes per year) into PGC and about 2 gigabytes per day (0.75 terabytes per year) into Ottawa. At PGC, data are maintained on a 12 terabytes RAID

system. Current data holdings consist of 2.1 terabytes on CD and 8 gigabytes of metadata.

4.4.7 LANDSAT / RADARSAT

RADARSAT International (RSI) was formed in 1989 as a joint venture between the Government of Canada, MacDonald Dettwiler and SPAR Aerospace. RADARSAT-1 was a government-funded program that successfully launched a Synthetic Aperture Radar (SAR) satellite in 1995. RSI was originally operated by MacDonald Dettwiler and SPAR, with the Government of Canada contributing the world-wide distribution rights for RADARSAT-1 data. In the late 1990s, SPAR sold its RSI shares to MacDonald Dettwiler, so that RSI is now operated exclusively by MacDonald Dettwiler, with RSI headquarters in Richmond, BC. RSI supplies data from most commercially-available Earth Observation (EO) satellites, including RADARSAT-1, LANDSAT 4/5, LANDSAT 7, IKONOS, IRS, ERS, QuickBird, EROS and ENVISAT. By combining EO satellite information with other data sources such as climate/field data, baseline geographic information, and Internet-based information, RSI has successfully responded to the changing business needs of the spatial data community. This integrated, multi-sensor approach keeps its clients at the cutting edge of advanced information solutions, and offers support to resource managers, disaster agencies, decision-makers, and other users of satellite-based information. The company delivers satellite image and data products and services for many applications, such as crop monitoring, ice reconnaissance, forest management, land use management, offshore oil and gas exploration, oil seep monitoring, mapping, ship detection and monitoring, and disaster management such as floods and oil spills.

RSI also works with global partners and customers to identify new applications for satellite imagery. The partners include an international network of more than 80 distributors, ten of whom operate their own ground receiving stations, and nine resource centres, all of whom offer near real time services, data processing, and/or training to local users.

The RSI operation in Richmond, BC receives, archives, processes and distributes in near real time 4.8 terabytes per year of LANDSAT satellite imagery, and other sensors from the ENVISAT, QuickBird and IKONOS satellites, with capacity for 18.6 terabytes/year. The operations in Prince Albert, SK and Gatineau, PQ receive, archive, process and distribute in near real time 6.9 terabytes/year of RADARSAT-1 satellite imagery, with capacity for 26.8 terabytes/year. As well as basic data ordering and distribution, RSI also provides scientific consultancy services and monitoring services, such as monitoring RADARSAT imagery for oil slicks at a Strait of Juan de Fuca choke point for the Government of Canada. RADARSAT-2 is the second high-resolution imaging SAR satellite and associated ground segment to be built in Canada. For RADARSAT-2 a partnership was formed between MacDonald Dettwiler and the Canadian Space Agency (CSA) to build, launch and operate a next generation imaging platform on a commercial basis. MacDonald Dettwiler is responsible for the overall system engineering and design,

and subcontract management for the Bus, Payload, Extensible Support Structure, Ground Segment, Operations Segment and Launch Services. While ensuring the continuation of RADARSAT-1 C-Band SAR image products, RADARSAT-2 represents a significant leap forward in spacecraft technology and available SAR image products. When launched, it will likely be the most advanced commercial SAR satellite in operation and will be used to monitor the environment and manage the Earth's natural resources.

4.5 Summary of Data Management Practices

While the organizations and programs researched in this phase are very diverse, each has some characteristics shared by VENUS/NEPTUNE. Each of these organizations can be compared along each of a set of dimensions:

- 1) What type of data does the organization or program deal with?
- 2) How complex is the metadata? Are the metadata enhanced through data analysis?
- 3) Are additional data products generated?
- 4) What technical architecture is used? How are data managed?
- 5) How important is processing and delivering data in a timely fashion?
- 6) Are the data stored in a distributed fashion?
- 7) How sophisticated are the user interfaces used to access the data?
- 8) Does the organization act primarily as a collection point, or is it required to deliver data for operational uses?

The table starting on the next page offers this comparison. In the table, the following abbreviations are used:

UVic	University of Victoria (ATLAS)
ACS	BC Active Control System
LRDW	BC Land and Resource Data Warehouse
NFIS	National Forest Information System (Pacific Forestry Center)
PGC	Pacific Geoscience Centre
LSAT	LANDSAT/RADARSAT
CNDC	Canadian National Data Centre for Earthquake Seismology and Nuclear Explosion Monitoring
ECWBS	Environment Canada Weather Buoy System
TAO	PMEL TAO-TRITON
EPIC	PMEL (EPIC)
USGODAE	Fleet Numerical Meteorology and Oceanography Center (USGODAE server)
FNMOC	Fleet Numerical Meteorology and Oceanography Center (weather modelling)
PO.DAAC	NASA EOS (Physical Oceanography Distributed Active Archive System)
IRIS	Incorporated Research Institutes for Seismology
NESDIS	NOAA NESDIS (satellite programs)

NDC Natural Environment Research Council (NERC) Datagrid
CORIOLIS Coriolis project, Ifremer
CADC Canadian Astronomy Data Centre

Dimension	Organization	Comment
What type of data does the organization or program deal with?	UVic	Particle physics; “petabytes of storage” (see http://particle.phys.uvic.ca/~web-atlas/atlas/computing/)
	ACS	GPS position information
	LRDW	Geospatial (geometry plus attributes)
	NFIS	Geospatial
	PGC	Seismic
	LSAT	Satellite
	CNDC	Seismic
	ECWBS	Oceanographic
	TAO	Oceanographic
	EPIC	Oceanographic; processing and access system
	USGODAE	Various; processing and access system
	FNMOG	Weather models
	PO.DAAC	Sea surface height, temperature, and waves, deduced from satellite
	IRIS	Seismic
	NESDIS	Satellite
	NDC	Various
	CORIOLIS	Oceanographic
	CADC	Astronomy

How complex is the metadata? Are the metadata enhanced through data analysis?	UVic	Unknown – awaiting reply
	ACS	Very simple – just static base station information
	LRDW	Contributor-supplied prior to publishing; conforms to metadata standard ISO-19115; no supplementation
	NFIS	Varies according to jurisdiction; gathered together into an Oracle metadata database at PFC
	PGC	Basic instrument – level metadata
	LSAT	Created at ground stations, but supplemented to describe post-processing
	CNDC	Under review; using CSS-3.0 standard
	ECWBS	Weak at present – under further development
	TAO	No reply
	EPIC	No reply
	USGODAE	Various
	FNMOC	Various
	PO.DAAC	Metadata provided at source; project underway to standardize metadata from satellite sensors
	IRIS	Data are continually repackaged (to reflect events) and metadata are redefined accordingly
	NESDIS	Created at ground stations, but supplemented to describe post-processing
	NDC	Metadata key to operation
	COROLIS	No reply
	CADC	Produces enhanced metadata catalogs
Are additional data products generated?	UVic	No
	ACS	No
	LRDW	No
	NFIS	No
	PGC	No
	LSAT	Yes (satellite image production)
	CNDC	Yes (event-oriented repackaging)
	ECWBS	No
	TAO	No
	EPIC	N/A
	USGODAE	Yes
	FNMOC	Yes (weather models)
	PO.DAAC	No
	IRIS	Yes (event-oriented repackaging)
	NESDIS	Yes (satellite image production)
	NDC	No
	COROLIS	No reply
	CADC	Yes

What technical architecture is used? How are data managed?	UVic	Supercomputers, mass storage, high network bandwidth
	ACS	File-based
	LRDW	Oracle database with SDE
	NFIS	File-based with Oracle metadata database and OGC Web-based access
	PGC	File-based
	LSAT	File-based, but Orbit information and ground control points are stored in an Empress embedded database
	CNDC	File-based, plus CA OpenINGRES database
	ECWBS	No reply
	TAO	File-based, plus mySQL database for metadata
	EPIC	No reply
	USGODAE	File-based, plus Informix database
	FNMOCC	Informix database, plus supercomputers for weather modelling; observation and gridded data stored in database.
	PO.DAAC	File-based
	IRIS	File-based, plus Oracle database for metadata
	NESDIS	File-based, plus mySQL and PostgreSQL for metadata
	NDC	No reply
	CORIOLIS	No reply
	CADC	File-based, plus Sybase database for raw metadata and DB2 for datamine.
How important is processing and delivering data in a timely fashion?	UVic	N/A
	ACS	Low
	LRDW	Low
	NFIS	Low
	PGC	High
	LSAT	Medium
	CNDC	High
	ECWBS	Medium-High
	TAO	Medium-High
	EPIC	N/A
	USGODAE	High
	FNMOCC	High
	PO.DAAC	Medium-High (becoming more important)
	IRIS	High
	NESDIS	Medium
	NDC	Low
	CORIOLIS	No reply
	CADC	Low

Are the data stored in a distributed fashion?	UVic	No
	ACS	No
	LRDW	No
	NFIS	Yes
	PGC	No, but replicated between Victoria and Ottawa
	LSAT	No
	CNDC	No
	ECWBS	No
	TAO	No
	EPIC	N/A
	USGODAE	No
	FNMOc	No
	PO.DAAC	No
	IRIS	No
	NESDIS	No
	NDC	Yes
	CORIOlIS	No reply
	CADC	No
How sophisticated are the user interfaces used to access the data?	UVic	No reply
	ACS	Low
	LRDW	Medium – High: see data discovery interface at http://srmapps.gov.bc.ca/metastar and interfaces at http://srnwww.gov.bc.ca/g/lrdw.html
	NFIS	Medium: google-like interface at http://www.geoconnections.org
	PGC	N/A – see CNDC
	LSAT	Medium: google-like interface at http://geodiscover.cgdi.ca/gdp/search
	CNDC	Medium: see http://www.seismo.nrcan.gc.ca/cnsn/
	ECWBS	No Web access yet
	TAO	Medium: see http://www.pmel.noaa.gov/tao/disdel
	EPIC	N/A
	USGODAE	Medium: see http://www.usgodae.org/cgi-bin/datalist.pl?generate=summary
	FNMOc	Medium – High: Metcast subscription service and API; see http://www.metnet.navy.mil/Metcast/
	PO.DAAC	Medium: see http://podaac.jpl.nasa.gov/catalog/
	IRIS	Medium – High: see http://www.iris.edu/data/data.htm
	NESDIS	N/A
	NDC	High – at least in design: see http://ndg.badc.rl.ac.uk/public_docs/AHM2004-02.ppt
	CORIOlIS	No reply
	CADC	High – see http://cadewww.dao.nrc.ca/

Does the organization act primarily as a collection point, or is it required to deliver data for operational uses?	UVic	No reply
	ACS	Collection point
	LRDW	Collection point
	NFIS	Collection point
	PGC	Operational
	LSAT	Operational
	CNDC	Operational
	ECWBS	Operational
	TAO	Operational
	EPIC	N/A
	USGODAE	Operational
	FNMOE	Operational
	PO.DAAC	Collection point
	IRIS	Operational
	NESDIS	Collection point
	NDC	Collection point
	COROLIS	Operational
	CADC	Collection point

Given the limited scope of this phase of the project, our treatment of the organizations and programs has been necessarily superficial. We have, however, investigated these organizations to the extent necessary to identify which aspects of these organizations / programs are relevant and which are not, and which aspects deserve further attention and investigation.

4.6 Respondent Recommendations

Each of the questionnaire respondents and interviewees were given the opportunity to provide specific data management advice to the VENUS/NEPTUNE project. These comments can all be found in Appendix B, but for convenience, these “words of wisdom and caution” have been gathered together and are presented in this section.

DFO/MEDS (Jean Gagnon)

A good general reference guide regarding marine data management of multi-disciplinary data sets has been documented in

<http://ioc.unesco.org/oceanteacher/resourcekit/index.htm> with some good representative sample projects <http://ioc.unesco.org/oceanteacher/Data/data.htm> used for illustration.

It is generally recognized that the World Ocean Circulation Experiment (WOCE) was a success story from a perspective of coordinating data management and attaining scientific objectives (<http://woce.nodc.noaa.gov/wdiu>).

Perhaps a more recent and local example of integration of data management within the context of a regional monitoring program would be the Atlantic Zone Monitoring Program (AZMP) http://www.meds-sdmm.dfo-mpo.gc.ca/zmp/main_zmp_e.html where the data management plan

http://www.meds-sdmm.dfo-mpo.gc.ca/zmp/Documents/e_DataMgmtPlan.htm forms a key integral and ongoing activity within the program activities.

DOC/NOAA/NESDIS National Oceanographic Data Center (Donald Collins)

Complete, accurate metadata are essential. A format description is only the beginning of metadata. No detail of metadata is too small to ignore. Details like codes, acronyms, calibration regimen, non-recording periods, instrument serial numbers, etc., are all important to someone and should be kept associated with the relevant data. Who will be available to interpret these data in 10 years? 50 years?

Avoid inventing 'meaningful codes' or new controlled vocabularies for describing data, instruments, etc. By 'meaningful code', we mean something like the NODC Taxonomic Code. Compare it to the ITIS Taxonomic Serial Number, which is a sequential integer, rather than each digit or pair of digits being imbued with a meaning. There are several pre-existing controlled vocabularies for describing instruments, organizations, platforms, data types, discovery keywords, etc. Pick one that best suits your needs and work with the maintenance organization to keep it up to date.

Identify your primary user community and as many secondary user communities that you can. How do these communities differ in the way that they look for information? Will generic discovery keywords be sufficient (i.e., NASA GCMD Parameter Valid) or do you need more specific terms (e.g., ASFA Thesaurus).

There is a growing number of similar efforts in the US, Canada, and the global community. The Joint Committee on Oceanography and Marine Meteorology has a couple of Expert Teams that are trying to define similar systems and standards. The Intergovernmental Oceanographic Commission is involved in similar efforts, as well as the various IOOS activities. Be aware of other development projects and participate in crosscutting activities as much as possible.

IRIS (Tim Ahern)

This is a very hard question to answer. Let me just say that IRIS already does, and has been doing, a very similar thing for nearly a decade. We adopted format standards early on and that has proven to be a great benefit for managing these data. We attempt to homogenize the data flow and formats before they enter (or as they enter) the DMC and therefore most data appears similar to us in terms of our automated system.

This is an incomplete answer but your question is difficult to fully answer.

MDA/LANDSAT/RADARSAT (Andrew Westwell-Roper)

Our relevant experience cannot be compressed into a questionnaire response. One important first step is to define the DMAS external interfaces - we have prepared a draft discussion paper on DMAS interfaces, which we will forward to Barrodale when comments on the draft have been received from UVic.

Another important approach is to define DMAS functional requirements (in more detail than that of the 21/01/02 DRD) at the same time as the Wet Plant requirements. Otherwise, DMAS will become a convenient place to hide all the difficult system-level problems, such as resource management, that will affect the Wet Plant and hence the ability of the system to do what is expected.

NOAA/NODC (Patrick Caldwell)

Any well described format is fine.

NOAA/PMEL, Argo Program (Greg Johnson)

VENUS and NEPTUNE should plan on significant spending for data management of a large and heterogeneous data set.

NOAA/PMEL, Vents Program (Robert Dziak)

Do you plan to provide any type of first-order analysis of your wide range of data types? Or do you simply plan to archive data then let investigators sort through the data online? Your varied data streams will require some type of intelligent algorithm (with an analyst providing quality assurance) to sort and classify the data. For example, I would think some kind of earthquake pick file will have to be produced from the hydrophone and seismometer data streams, otherwise you will quickly fall behind. Perhaps the solution is to request that outside organizations that already have data analysis infrastructure in place should coordinate VENUS/NEPTUNE records with their existing records to provide a preliminary online database.

NOAA/Undersea Research Center (Rob Cermak)

There are so many data formats. Pulling a bunch of scientists together from different institutions means you will have to deal with several different commercial software packages that all want different formats – some with closed proprietary formats. Some may have import features, but once the data are in that package and results are found, getting the results out so that others may utilize it is very difficult in some cases.

Any DMAS that wants to be successful has to provide data to PIs in as many formats as possible. This is technically infeasible due to the closed, proprietary formats used in some software and sampling applications.

Due to funding and technical restrictions, the data are made available in ASCII format. We make it available in ODV since the format for data import is somewhat documented and ODV is used by a large number of faculty. Matlab is another well used application and can readily import ASCII tables.

No matter what format you decide to store data and information in, make sure there is a reasonable way to export these data into modern systems. Usage of a single proprietary format will automatically close the use of archived data to other researchers.

NRCAN (Jim Lyons)

Have to balance advantages of handling/managing/storing disparate data types in one common waveform format vs. ease of use for clients. We have found format conversion on-the-fly to be practical using only modern SUN workstations as servers.

PFC (Robin Quenet)

Ensure that the data structure is searchable with Google and Yahoo tools. Also note that DAC (Distributed Access Control) is very useful for determining who has access to resources over the Web. It provides a means of identifying where you are from and provides authentication by jurisdiction using browser “cookies”. It permits granular control of users, groups and URLs. ACLs for resources at the file or parameter level can be checked automatically.

TAO/PMEL (Paul Freitag)

A significant part of the data processing and quality control procedures concerns use of instrument calibration data. All transmitted and stored data from TAO instruments must be combined with calibration data to produce user data. This process occurs at the time of real time distribution on GTS by Service Argos, for (near) real time processing and quality control on site, and for processing and quality control of delayed data. Management of the calibration systems and data requires about the same level of resources as the acquisition of instrument data from deployed platforms. This may be a metadata issue in your scheme, but not obviously so from our perspective.

Our experience suggests that it will be important for the VENUS/NEPTUNE DMAS to allow user-defined subsets of the data to be extracted and downloaded. Even for users who want all of the data, there are often bandwidth limitations, particularly in the third world, which make the option to download a subset of the data an essential capability.

The design of our data delivery system includes data availability information in the user interface, which quickly shows the user what data are available, prior to the user-request. For example see

http://www.pmel.noaa.gov/tao/data_deliv/

Many delivery systems require the user to make their request first, only to be informed that there are no data fitting their criteria, which can be frustrating, inefficient, and confusing.

Another suggestion is to make as much metadata as possible available from the Web or served with the data, as this will undoubtedly limit the number of questions you receive from users and make some responses as simple as providing a URL. We provide much of our metadata at

http://www.pmel.noaa.gov/tao/proj_over/proj_over.html.

5 Recent Developments in Data Management

Several recent advances in data management are relevant to the design and operation of the VENUS/NEPTUNE DMAS. These advances are described in the following sections.

5.1 NetCDF

NetCDF (<http://my.unidata.ucar.edu/content/software/netcdf/index.html>) is a binary format for storing array-oriented data, and has become a *de facto* standard for storing gridded oceanographic data. It is an open standard, and many commercial and freely available software packages for accessing and displaying netCDF files have been written. (See <http://my.unidata.ucar.edu/content/software/netcdf/software.html?unidata=cb123758405440c7b007e5b49977c5a9> for a list of these).

Some benefits of using netCDF are:

- the metadata can easily be wrapped together with the actual data, making the data set “self-describing” and removing the risk that the metadata will get lost;
- the format is application- and machine-independent;
- some capability for subsetting of data is provided by the netCDF access libraries;
- new components can be added to existing netCDF files without breaking applications that don’t know about these new components.

Some limitations are:

- the data are not stored in ASCII form and thus cannot be viewed directly with a text editor / viewer;
- there are limits to the “self-describing” feature of netCDF files – in practice, conventions must be defined and adhered to;
- there are limits to the sorts of arrays that can be conveniently represented - for example, “jagged arrays” (files consisting of several variables each with a different number of elements) are not efficiently supported.

5.2 OPeNDAP

OPeNDAP⁴ (<http://opendap.org/>), the Open source Project for a Network Data Access Protocol, is a data transport protocol in wide use in over 40 US oceanographic and meteorological organizations. OPeNDAP provides a means for data analysis and visualization packages such as Matlab, IDL and Ferret to access files over a network (local or Internet) without needing to know how the data are stored at the server site. Even though the data might be stored in netCDF, HDF, Matlab binary format, a JDBC-enabled Relational Database Management System (RDBMS), or some other supported format (see <http://www.unidata.ucar.edu/packages/dods/home/faq/whatServers.html>), the Matlab user can open the file as if it were a Matlab binary format file sitting on his or her own workstation.

Even though OPeNDAP currently supports a limited number of server and client file formats, it is possible to create custom clients and servers using the OPeNDAP development frameworks (C++ and Java Toolkits).

A list of organizations that are hosting OPeNDAP servers is provided in Appendix E.

Two OPeNDAP servers deserve special mention:

- 1) the GrADS Data Server (GDS);
- 2) the OPeNDAP Aggregation Server.

The GDS can serve a variety of file formats (including netCDF, GRIB, Binary, HDF, and BUFR) and it has the added capabilities of returning just a subset of the served file (based on either a temporal or spatial restriction) and of performing analysis operations (e.g., basic math functions, averages, smoothing, differencing, correlation, and regression) on the data. Documentation on GDS can be found at <http://grads.iges.org/grads/gds/>.

The OPeNDAP Aggregation Server can be used to present a view of data built from (aggregated from) several files. These files can either be local netCDF files or files that are in turn presented by some other OPeNDAP server. Three types of aggregation are supported:

1. JoinNew: N-dimensional arrays are joined together to form an (N+1)-dimensional data set; for example, two XYZ grids for different times can be joined together to form an XYZT grid.
2. JoinExisting: A file containing an Nx by Ny by Nt array can be joined with a file containing an Mx by Ny by Nt array to form an (N+M)x by Ny by Nt array; for

⁴ formerly known as DODS – *Distributed Oceanographic Data System*.

example, two adjacent XYZ grids can be butted together to form another (larger) XYZ grid.

3. Union: Two files, each with a different array variable (e.g., one with a temperature array and one with a salinity array), can be combined to form a single file with two variables (e.g., temperature and salinity).

Documentation on the OPeNDAP Aggregation Server can be found at <http://www.opendap.org/server/agg-html/agg.html>.

5.3 Developments in Metadata Management

Metadata can be considered to be “the information necessary for someone who is not previously acquainted with a data set to make full and accurate use of that data set.”⁵ (<http://books.nap.edu/books/NI000157/html/92.html#p200033adppp92>). Metadata are of two types: *syntactic* (information on how to access the data: data types, structures, etc.), and *semantic* (information on how to interpret the data: accuracy, provenance, units, etc.).

5.3.1 FGDC Metadata Standard

By executive order, all US Government agencies must use the Federal Geographic Data Committee (FGDC) Metadata Standard (MS) to document data sets. NOAA has developed templates that can be used in describing most types of NOAA data. See <http://www.eis.noaa.gov/fgdc/fgdc1.html>.

5.3.2 MarineXML

One of the inherent difficulties in managing metadata (especially semantic metadata) is ensuring that the metadata are updated appropriately when the referenced objects are operated on. For example, if A and B are two data objects with metadata Ma and Mb, respectively, then care must be taken to ensure that the metadata for some derived object F(A,B) are constructed appropriately. MarineXML is one of the technologies intended to address this issue. The “Marine XML Portal” (<http://marinexml.net>) contains the following description and list of benefits of Marine XML:

The development of a marine XML will support the tracking of data from collection through to the generation of integrated global and regional data sets. XML can support the metadata describing the data collection, quality control and subsequent processing. The generation of data tagged with XML at the instrument level would provide the ability to automate such processes as generation of metadata descriptions.

⁵ Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government: Working Papers (1995), Report of the Ocean Sciences Data Panel.

There are a number of reasons for using a marine XML:

- **Exchange of data.** A major strength and source of potential of XML is that it facilitates the exchange of data between different applications and operating systems. One of XML's strongest points is its ability to do data interchange. Because different organisations (or even different parts of the same organisation) rarely standardise on a single set of tools, it takes a significant amount of work for two groups to communicate. XML makes it easy to send structured data across the Web so that nothing gets lost in translation. XML is potentially the answer for oceanographic data exchange, as long as all sides agree on the markup to use.
- **Extensibility.** Extensible means that it is not a fixed format like HTML. While HTML tags must follow pre-set standards, new XML tags can be created by anyone at any time. XML will allow groups of people or organisations to create their own customized markup languages for exchanging information in their domain. Examples of existing industry-specific XML include music, chemistry, electronics, linguistics, engineering and mathematics.
- **Plain Text.** Since XML is not a binary format, files can be created and edited with a standard text making it useful for storing small amounts of data. At the other end of the spectrum, an XML front end to a database makes it possible to efficiently store large amounts of XML data. XML provides scalability for anything from small configuration files to an industry-wide data repository.
- **Data Identification.** The XML standard specifies how to identify data, not how to display it. HTML, on the other hand, describes how things should be displayed without identifying the content. Because the different parts of the information have been identified, they can be used in different ways by different applications.
- **Stylability.** When display is important, the style sheet standard, XSL, can dictate how to portray the data. Since XML is inherently style-free, different style sheets can be used to produce output in postscript, PDF, or any other format.
- **Hierarchical.** XML documents are hierarchical in structure. Hierarchical document structures are, in general, faster to access because you can drill down to the part you need, like stepping through a table of contents.

The use of XML for metadata has received a great deal of consideration by Canadian officials (in particular Robert Keeley of MEDS). See, for example, "Developing an eXtensible Markup Language (XML) Application for DFO Marine Data Exchange via the Web" (<http://ioc3.unesco.org/marinexml/files.php?action=dlfile&fid=28>) and "The XML Bricks Concept" http://ioc3.unesco.org/marinexml/files.php?action=viewfile&fid=4&fcid_id=3.

5.4 Unidata Internet Data Distribution

Unidata is a "community" of roughly 150 universities and research institutions, established with the goal of sharing data and data access tools. Among the members of Unidata are:

- the Woods Hole Oceanographic Institution;
- Environment Canada;
- Fleet Numerical Meteorological and Oceanography Center (FNMOC);

- the University of Alaska – Fairbanks (Institute of Marine Science and the Atmospheric Sciences Group);
- the University of British Columbia (Geography Dept.);
- the University of Washington (Atmospheric Sciences);
- the University of Oregon.

A complete list of members can be found at

<http://my.unidata.ucar.edu/content/community/participatinguniversities.html>.

The Unidata Internet Data Distribution (IDD) is a system that allows participating institutions to “subscribe” to data directly from the observing system (bypassing the “data center”). Currently, IDD distributes 50 gigabytes of data per day to 130 institutions; much of the data take the form of meteorological observations or weather model predictions.

5.5 Thematic Realtime Environmental Distributed Data Services (THREDDS)

THREDDS (see <http://my.unidata.ucar.edu/content/projects/THREDDS/index.html>) is actually an infrastructure built upon many of the technologies already discussed in this section. Collaborators to the THREDDS project include NOAA/NCDC, NOAA/NGDC, NOAA/PMEL, FNMOC, and IRIS (see <http://my.unidata.ucar.edu/content/projects/THREDDS/Overview/Collaborations.html> for a complete list).

The THREDDS infrastructure includes:

- simple THREDDS servers (containing catalogs and metadata);
- THREDDS/IDD servers (serving IDD data);
- OPeNDAP Aggregation server;
- Catalog Generator and Validator for automatically creating / checking catalogs;
- ADDE Cataloger for constructing catalogs for ADDE servers (ADDE is the Abstract Data Distribution Environment, an alternative to OPeNDAP⁶);
- the Live Access Server (LAS), a Web server built on top of OPeNDAP and Ferret (a data visualization and analysis tool) – see http://ferret.pmel.noaa.gov/Ferret/LAS/ferret_LAS.html and <http://ferret.pmel.noaa.gov/Ferret/>.

⁶ ADDE was developed as part of the McIDAS (Man Computer Interactive Data Access System) project at the Space Science and Engineering Center at the University of Wisconsin – Madison. McIDAS continues to be distributed as an SSEC proprietary application, widely used within the atmospheric science community.

5.6 Web Coverage Servers (WCS)

The Open GIS Consortium has published a specification (<http://www.opengis.org/docs/03-065r6.pdf>) for delivering regularly spaced 3- or 4-dimensional gridded data to GIS clients over the Web. In the future this specification will be extended to handle unequally spaced and non-gridded data.

Current WCS-compliant servers include the NASA HDF-EOS (NWGISS) server for serving satellite imagery. See [http://esto.nasa.gov/conferences/estc-2002/Papers/PS2P1\(YangW\).pdf](http://esto.nasa.gov/conferences/estc-2002/Papers/PS2P1(YangW).pdf) for a description of NWGISS.

THREDDS and the University of Florence are currently in the process of developing a WCS for netCDF files. See http://www.unidata.ucar.edu/projects/THREDDS/Nativi/WCS_Service/THREDDS_WC_S_Service.pdf?unidata=30b81713f3a0893b3842fa7daf1541d9.

5.7 Database versus File Storage

Traditionally, in oceanographic systems, data have been stored in files, with metadata being stored in a relational database. Each row in the database stores the metadata for a data file, including a reference to where the file is actually stored. This arrangement supports the practice where a scientist first decides what files are relevant to his/her needs (by querying the database), and then retrieves either the entire files or possibly (if the appropriate OPeNDAP servers are used) the relevant portions of the files.

Until recently there has been little reason to actually store the actual oceanographic data in the database, as traditional databases have offered very limited support for complex data objects. That being said, there are some disadvantages to storing oceanographic data outside the database, even in traditional database environments:

1. Whenever metadata are separated from data there is a risk that the metadata will be lost and the actual data will become meaningless and hence useless.
2. Data “mining” opportunities are lost. The only terms that can be used in a query are ones that reference metadata objects. Looking for data on the basis of new relationships (ones that weren’t considered when the metadata were derived) is impossible.
3. Storing data in database tables with a well-constructed schema will force data standards to be checked before the data are stored. With filesystem storage these standards are too often “checked” only when an attempt to access the data is made (and it is found that the data lack integrity).

In contrast to traditional RDBMS’s, modern day Object-Relational Database Management System (ORDBMS) and Object Oriented Database Management System

(OODBMS) products allow binary large objects to have exposed methods. In an oceanographic database, for example, there might be special *GRID*, *TIMESERIES*, and *PROFILE* data types with methods such as *REPROJECT*(GRID), *FUSE*(GRID1,GRID2), *MAXTEMP*(PROFILE), and *FFT*(TIMESERIES). Each row of the database would represent a single data object (grid, profile, timeseries), with columns storing the metadata fields and a single column storing the actual data. Revisiting the three problems outlined above, we see:

1. Metadata are not separated from data; hence the chance that the metadata will be lost is reduced.
2. Data “mining” opportunities are possible. Queries can be based on both metadata columns and data methods. As new relationships and properties are considered, new methods can be added.
3. A well-constructed schema, including data import/export methods will force data standards to be checked before the data are stored.

Some may argue that a transition from file-based storage to database-based storage might be very impractical given the number of applications in use today that have been written expecting Matlab data or netCDF files. This is exactly where the benefits of OPeNDAP become really apparent. Given a database with certain special data types and methods, custom OPeNDAP servers can be written to serve the data, allowing existing client applications to continue with the illusion that they are looking at Matlab or netCDF data.

5.8 Database Extensions

In the past decade, many of the major RDBMS products have become extensible. Each of these products offers (to varying degrees) the ability to define new data types and operations (methods) on those data types. Oracle, IBM/DB2, IBM/Informix, and PostgreSQL have spatial extensions (offering point, line, and polygon data types), and IBM/DB2, IBM/Informix, and PostgreSQL offer the facility to allow third parties to develop extensions (referred to as *Extenders*, *DataBlades*, and *Extensions*, respectively). The extensions relevant to VENUS/NEPTUNE are described in the following sections.

5.8.1 CopperEye DataBlade

Relational databases typically use “B-tree” indexes, a form of index that offers very stable, predictable, and scalable query performance. The cost of this query performance, however, is a higher update expense. CopperEye (<http://www.coppereye.com>) has produced a DataBlade for IBM/Informix that offers an alternative to the B-tree. CopperEye claims that transactional performance can be increased by a factor of ten in environments where the tables are rapidly changing (e.g., in a real time data collection environment).

5.8.2 Time Series DataBlade

IBM/Informix offers a real time data loader and DataBlade capable of ingesting over 40,000 readings per second. Each element of a time series object is tagged with a timestamp (with a precision of 10 microseconds), and the object can consist of arbitrary collection of data types. The DataBlade provides a set of SQL functions (e.g., for aggregating and subsetting time series) and Java and C APIs for writing programs to access datum values directly (and for writing custom methods callable from the SQL interface).

See <ftp://ftp.software.ibm.com/software/data/informix/blades/timeseries/GC27-1495-00.pdf>, <http://publib.boulder.ibm.com/epubs/pdf/6577a.pdf>, and <http://publib.boulder.ibm.com/epubs/pdf/8921.pdf>.

5.8.3 Grid DataBlade / Extension

Barrodale Computing Services Ltd. (BCS) has developed a Grid DataBlade for IBM/Informix and a Grid Extension for PostgreSQL. These products support 4D equally- or unequally-spaced grids and provide operations for spatially transforming (reprojecting) grids, subsetting grids, slicing grids (orthogonally or obliquely to the axes), and aggregating grids. See http://www.barrodale.com/grid_Demo/index.html for more information.

5.9 **Summary of Key Points**

1. NetCDF (<http://my.unidata.ucar.edu/content/software/netcdf/index.html>) is a binary format for storing array-oriented data, and has become a *de facto* standard for storing gridded oceanographic data. It is an open standard, and many commercial and freely available software packages for accessing and displaying netCDF files have been written.
2. OPeNDAP⁷ (<http://opendap.org/>), the Open source Project for a Network Data Access Protocol, is a data transport protocol in wide use in over 40 US oceanographic and meteorological organizations. OPeNDAP provides a means for data analysis and visualization packages such as Matlab, IDL and Ferret to access files over a network (local or internet) without needing to know how the data are stored at the server site.
3. By executive order, all US Government agencies must use the Federal Geographic Data Committee (FGDC) Metadata Standard (MS) to document data sets. NOAA has developed templates that can be used in describing most types of NOAA data. See <http://www.eis.noaa.gov/fgdc/fgdc1.html>.

⁷ formerly known as DODS – *Distributed Oceanographic Data System*.

4. The Unidata Internet Data Distribution (IDD) is a system that allows participating institutions to “subscribe” to data directly from the observing system (bypassing the “data center”). Currently, IDD distributes 50 gigabytes of data per day to 130 institutions; much of the data take the form of meteorological observations or weather model predictions.
5. THREDDS (<http://my.unidata.ucar.edu/content/projects/THREDDS/index.html>) is actually an infrastructure built upon many of the technologies already discussed in this report. Collaborators to the THREDDS project include NOAA/NCDC, NOAA/NGDC, NOAA/PMEL, FNMOC, and IRIS. The THREDDS infrastructure includes:
 - a. simple THREDDS servers (containing catalogs and metadata);
 - b. THREDDS/IDD servers (serving IDD data);
 - c. OPeNDAP Aggregation server;
 - d. Catalog Generator and Validator for automatically creating / checking catalogs;
 - e. ADDE Cataloger for constructing catalogs for ADDE servers (ADDE is the Abstract Data Distribution Environment, an alternative to OPeNDAP⁸);
 - f. the Live Access Server (LAS), a Web server built on top of OPeNDAP and Ferret (a data visualization and analysis tool) – see http://ferret.pmel.noaa.gov/Ferret/LAS/ferret_LAS.html and <http://ferret.pmel.noaa.gov/Ferret/>.

⁸ ADDE was developed as part of the McIDAS (Man Computer Interactive Data Access System) project at the Space Science and Engineering Center at the University of Wisconsin – Madison. McIDAS continues to be distributed as an SSEC proprietary application, widely used within the atmospheric science community.

6 Architectural Options for VENUS/NEPTUNE DMAS

6.1 Background

The following DMAS context activity diagram (Figure 5) appears in the [NEPTUNE Data Management And Archiving System Design Requirements \(Draft\)](#) document, produced in January 2002.

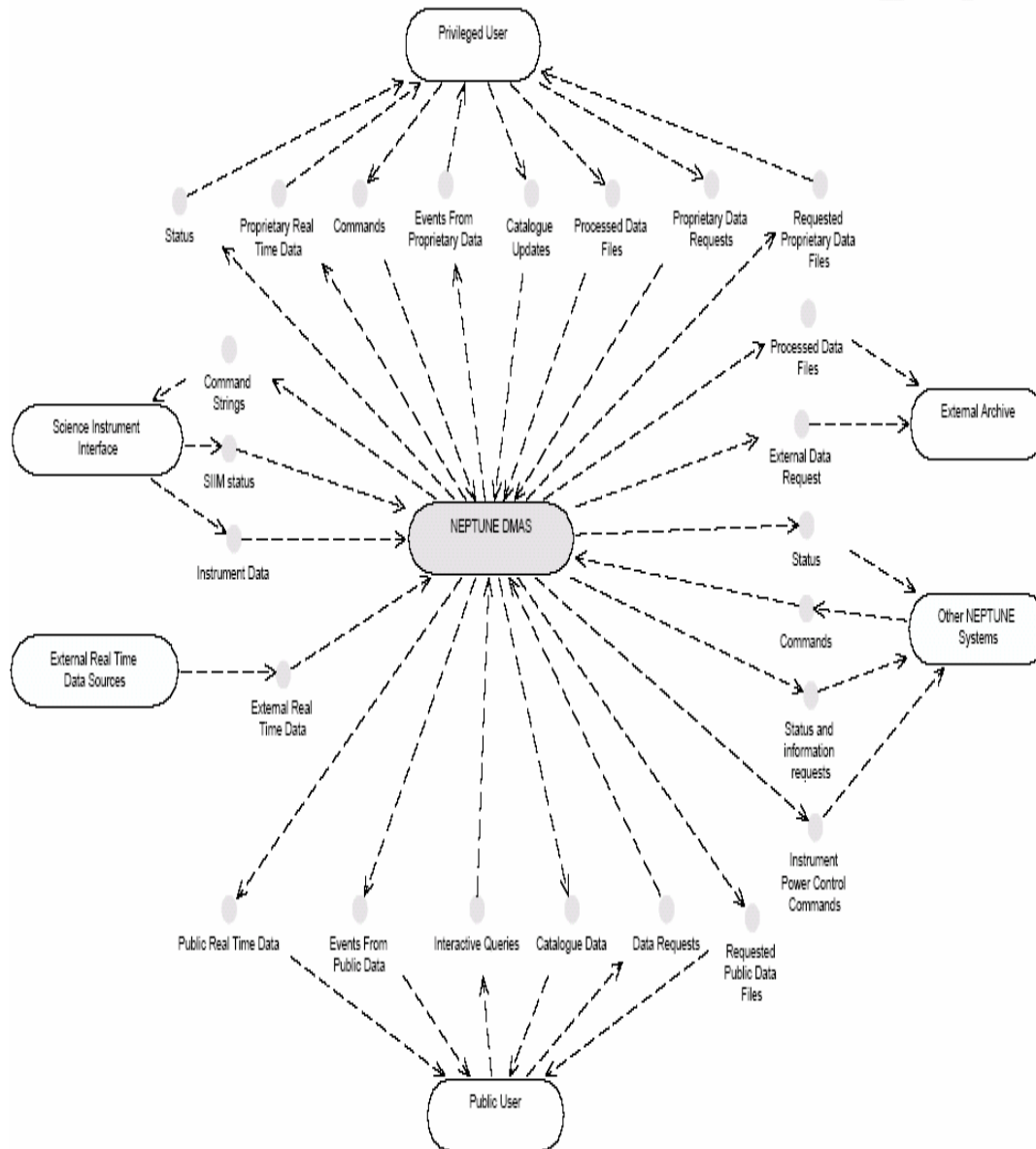


Figure 5. DMAS context activity diagram.

Figure 5 illustrates how a DMAS fits into the larger VENUS/NEPTUNE cabled-observatory framework and into the world outside (e.g., the public users). In the same report, the functionality and requirements of the DMAS are illustrated as in Figure 6.



Figure 6. Functionality requirements of DMAS.

The Shore Station Buffering activity provides an interface between the strictly real time behavior of data production, and the rest of the DMAS, which, from time to time, may not be able to keep up with the pace of the incoming data stream. The buffering activity

will present the rest of the DMAS with a more manageable, “near real time”, data stream. Further analysis of this activity is beyond the scope of this project.

The Instrument Services activity is concerned with the operation, monitoring, and control of the VENUS/NEPTUNE instruments, and with the handling of, and access to, data received from these instruments. Issues concerning instrument operation, monitoring, and control are beyond the scope of this report, but the aspect of data handling, and access to real time data, will be considered below.

Finally, the Archiving activity is concerned with the long-term storage of, and access to, VENUS/NEPTUNE data.

These activity diagrams and other sources described earlier in this document make clear the following general requirements:

- 1) Users will require access to both near real time data and to archived data.
- 2) Users will require access to catalogs of near real time data and to archived data.
- 3) Data in various forms and in each of several phases of processing will be stored (raw data, preliminary QC'd data, final QC'd data, windowed data, aggregated data, derived data, etc.).
- 4) Metadata will need to evolve accordingly and be carried along with these data.

The diagrams in the following section illustrate a framework within which alternative architectures and implementations can be presented and described. The diagrams present the preceding requirements pictorially, and illustrate where the technologies described earlier in this report could fit in.

Additional information relevant to the VENUS/NEPTUNE DMAS options described in this section is given in Appendix F (Object-Relational Databases) and Appendix G (Data Mining).

6.2 General Framework

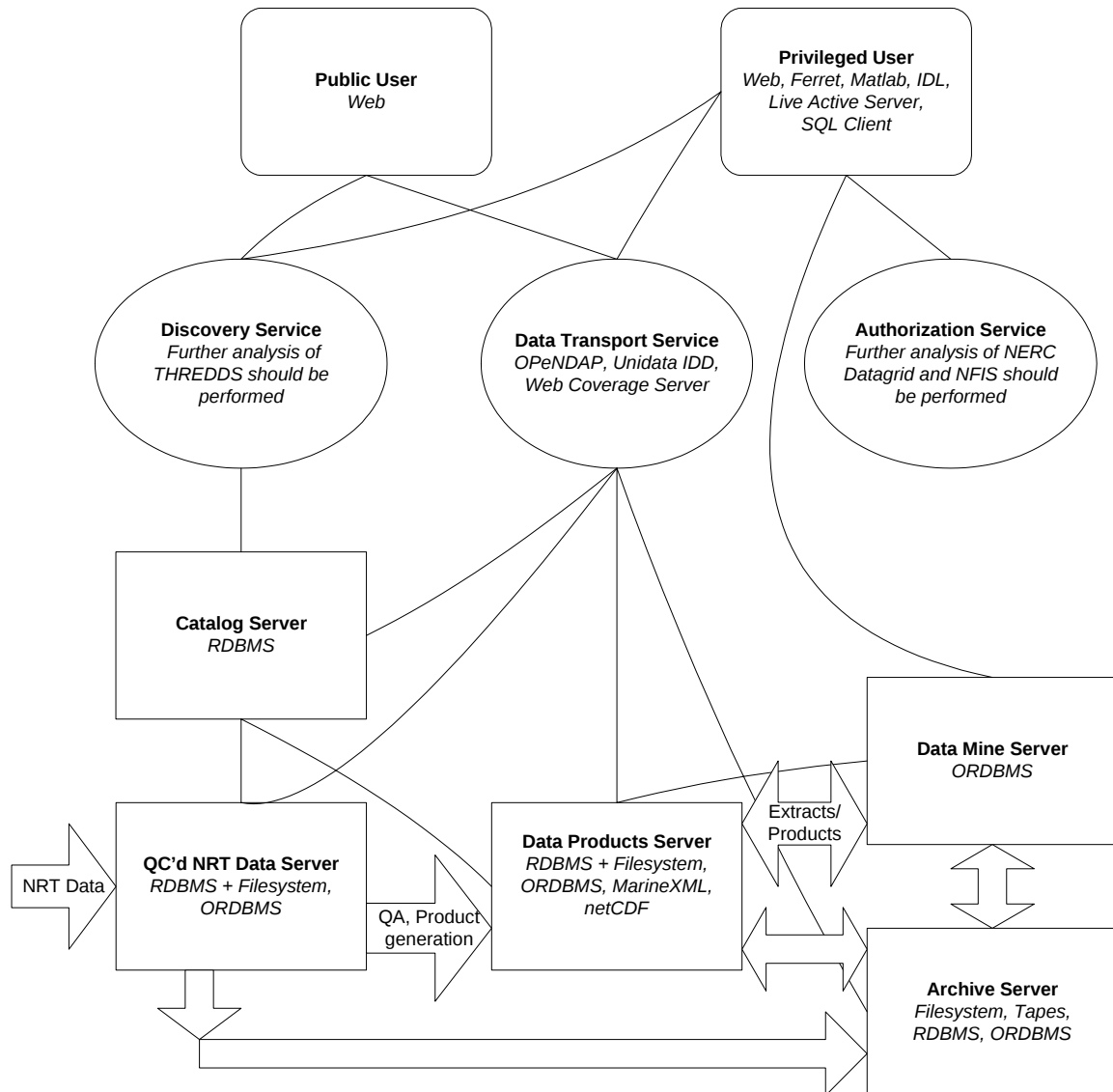


Figure 7. General framework for VENUS/NEPTUNE DMAS architecture.

Figure 7 shows a general framework for the DMAS architecture, with three levels of objects:

- 1) “Actors” are shown in boxes with rounded corners. In this simple depiction there are just two types of actors:
 - i) privileged users; and
 - ii) general users.
- 2) “Repositories” are shown in boxes with square corners. There are five repositories:

- i) the Near Real Time Data (as it exists after a rudimentary level of QC 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 Mine Server, hosting a database of data extracts within which data mining can be performed; and
 - v) the long-term Data Archive.
- 3) “Services” are shown in oval shaped boxes. Services are provided to the Actors, based on Data in the Repositories. There are three services shown:
- 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; and
 - iii) an Authorization Service, which is responsible for authentication of privileged users and determination of access levels.

Lines in the diagram simply link together actors, services, and repositories that have some relationship. No specific directionality, 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 mine), meaning that the actor can use that repository directly.

Listed within each object are the various technologies, described earlier in this report, that are relevant to that particular object. For example, public users will use the Web to access VENUS/NEPTUNE data, whereas privileged users may also use tools such as Ferret, Matlab, IDL, etc.

6.3 Option A: Traditional Approach

Figure 8 is a version of Figure 7 that presents a “traditional” approach to defining an architecture for the VENUS/NEPTUNE DMAS.

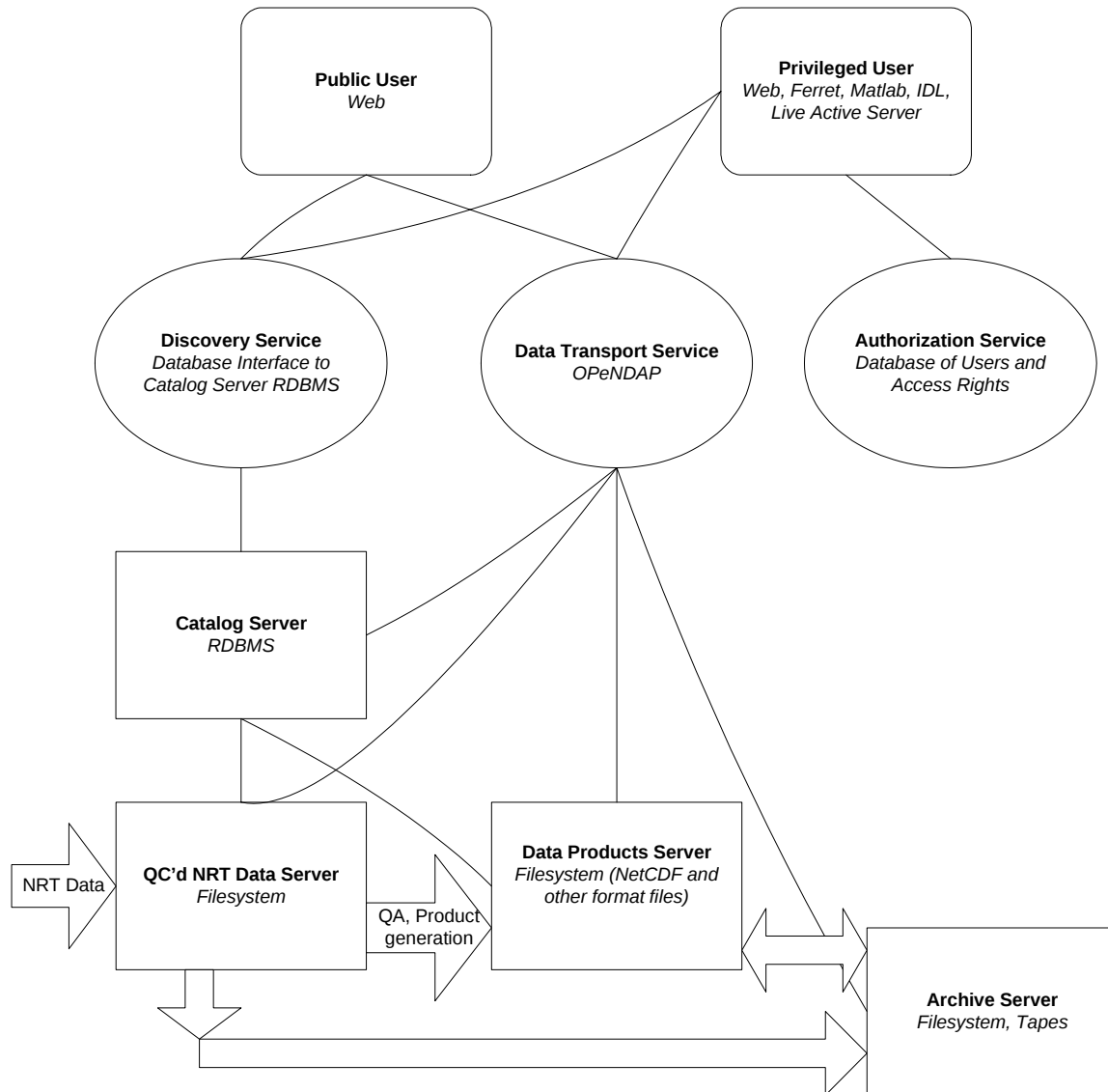


Figure 8. A traditional approach for implementing the DMAS architecture.

This architecture follows the traditional “metadata in a database, raw data in files” approach. The database can be a traditional (non-extended) relational database, since metadata elements are just simple numbers, dates, and text strings.

Limitations to this approach have been described earlier (see Section 5.7) and are essentially the following:

- 1) **Metadata / data separation.** Since the metadata are stored separately from the corresponding data there is a higher risk that they will become out-of-synch with respect to each other.
- 2) **No opportunities for data mining.** Discovery queries can access only the properties that have been elevated to “metadata” status. There may be many other properties that could be derived from the data, and useful as selection criteria, but these properties remain hidden until the data are extracted to the user’s machine and examined.
- 3) **Lost opportunities for enforcing data integrity.** When metadata are stored separately from the data they describe, there is no way to automatically ensure that referential integrity between the two is maintained.
- 4) **Limited ability to subset or aggregate data.** While OPeNDAP does provide the ability to subset and aggregate some types of files, this functionality is very limited.

6.4 Option B: ORDBMS Approach

An alternative architecture, based on object-relational databases, is illustrated in Figure 9:

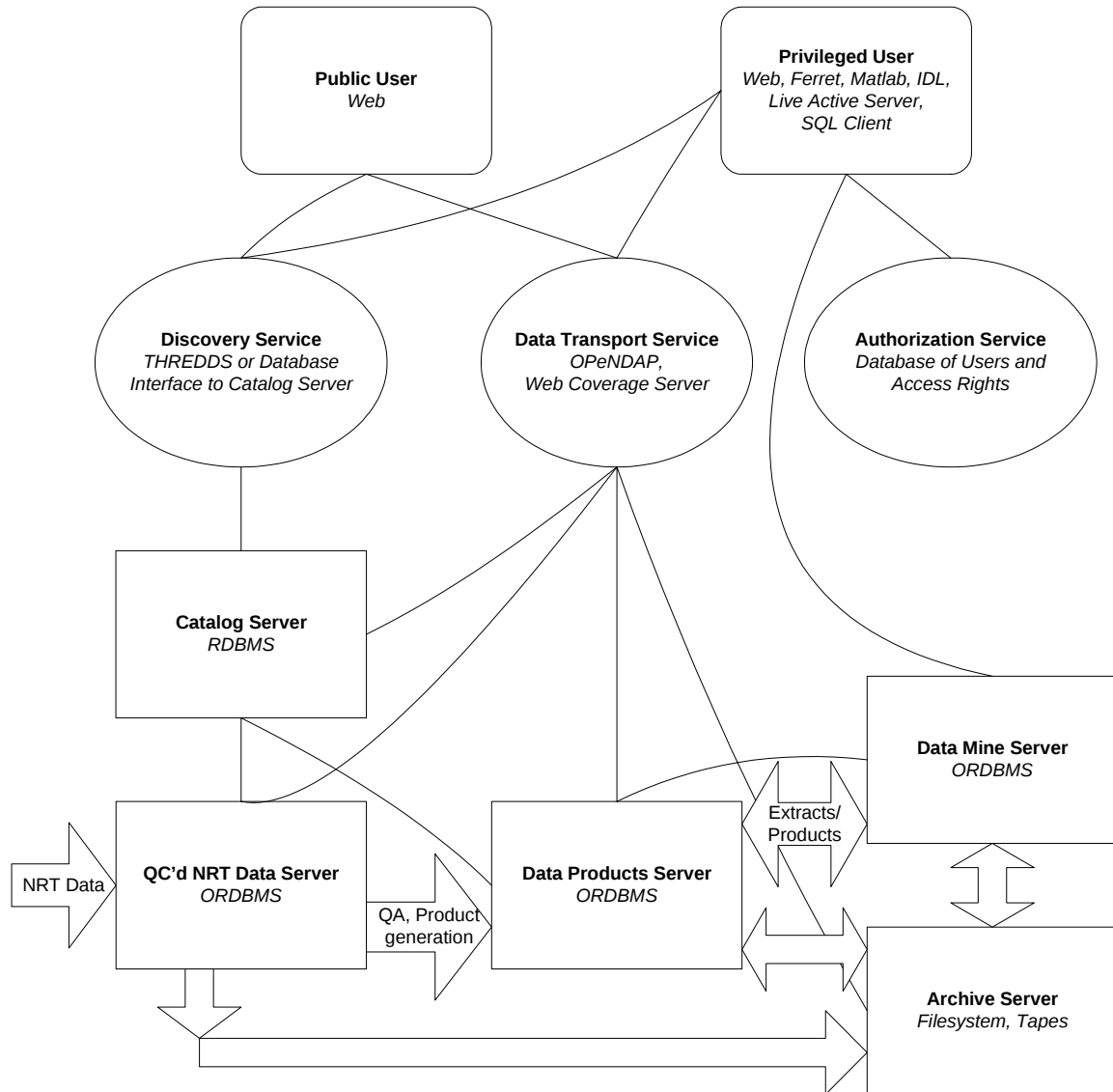


Figure 9. An ORDBMS approach for implementing the DMAS architecture.

In this architecture, data and metadata are stored together, in the same rows, in tables in object-relational databases. A short discussion of object-relational databases, and their features and benefits, is provided in Appendix F. With respect to an architecture for the VENUS/NEPTUNE DMAS, the object-relational approach offers the following benefits:

- 1) **Complex data type /operation support.** The structure of measurements need not be lost. Profiles and time series can be stored as objects in the database, and methods can be defined on these objects.
- 2) **Metadata / data treated in a uniform manner.** Since the metadata are stored together with the data they describe, the risk that they will become out of synch with respect to each other can be eliminated. A method (operation) that is written to operate on a particular type of object (e.g., a time series), can be written 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 that 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 just talked about) 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. (An introduction to data mining principles is provided in Appendix F.
- 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. Some sample queries might be:

Selection:

Find all temperature profiles in a specific region, gathered between DATE1 and DATE2, where the maximum temperature occurred between DEPTH1 and DEPTH2, and this maximum temperature was between TEMP1 and TEMP2.

Aggregation:

Return the “average” profile, for profiles satisfying the preceding selection criteria.

7 Appendix A – Files Provided on Supplemental CD-ROM

7.1 Brochures

Sensor (Directory)	Brochures
Aanderaa Oxygen Optode 3975	Oxygen_Optode3830_3930_3975_D335.pdf
ASL Water Column Profiler	wcpb0201.pdf
D-A Instruments OBS3	obs3.pdf obs3a.pdf
Digital Video, Still Cameras	Imenco DV3018 DVcamera.pdf Imenco SDS3040.pdf Kongsberg oe14208.pdf C-Map Systems Cyclops.htm
Flowcam	flowCAM_brochure.pdf
Guralp Seismometers	Guralp CMG-3T.pdf Guralp OBS_long.pdf DM24.pdf
Jasco AIM-2000	
RDI Workhorse ADCP	workhorse_monitor_ds_lr.pdf RDI WinADCP03 Display Software.pdf
Satlantic ISUS (MBARI)	
Seabird SBE 16plus CTD	SBE 16plus brochure 0704.pdf

7.2 Manuals

Sensor (Directory)	Manuals
Aanderaa Oxygen Optode 3975	Aanderaa TD218_Oxygen_Optode_3830_3930_Manual.pdf
ASL Water Column Profiler	
D-A Instruments OBS3	obs3amanual.pdf OBS3A Software.pdf
Digital Video, Still Cameras	
Flowcam	
Guralp Seismometers	
Jasco AIM-2000	
RDI Workhorse ADCP	Broadband Primer.pdf feature upgrade setup card Mar03.pdf Mariner Setup Card.pdf Mariner User Guide.pdf Monitor Setup Card.pdf Monitor User Guide.pdf Rio Grande Setup Card.pdf Rio Grande User Guide.pdf Sentinel Setup Card.pdf Sentinel User Guide.pdf WH Technical Manual.pdf WorkHorse Installation Guide.pdf WorkHorse Maintenance Guide.pdf WorkHorse Read This First.pdf WorkHorse Test Guide.pdf WorkHorse Troubleshooting Guide.pdf
Satlantic ISUS (MBARI)	Satlantic Nitrate (ISUS).pdf ISUSPro-Manual.pdf SatCon-Manual.pdf SatView-Manual.pdf
Seabird SBE 16plus CTD	SBE 16plus CTD Manual (RS232).pdf SBEDataProcessing_5.31.pdf Seasave_5.31.pdf

7.3 Data Files and Formats

Sensor (Directory)	Data Files and Formats
Aanderaa Oxygen Optode 3975	(see manual)
ASL Water Column Profiler	RawDataFmt_DF200407.doc
D-A Instruments OBS3	(see manual)
Digital Video, Still Cameras	RFC 3189 DV over IP.doc RFC 3497 HDTV over IP.doc
Flowcam	Flowcam Sample Data File.fcm Flowcam Spreadsheet Metadata.doc
Guralp Seismometers	
Jasco AIM-2000	AIM-2000 sentences.doc Sample data file for AIM-2000.txt
RDI Workhorse ADCP	(see manual) WorkHorse Commands and Output Data Format.pdf WorkHorse Command Quick Reference Card.pdf
Satlantic ISUS (MBARI)	(see manual)
Seabird SBE 16plus CTD	(see manual)

7.4 Software

Sensor (Directory)	Software
Aanderaa Oxygen Optode 3975	
ASL Water Column Profiler	
D-A Instruments OBS3	OBS3ALPATb.exe
Digital Video, Still Cameras	
Flowcam	
Guralp Seismometers	
Jasco AIM-2000	
RDI Workhorse ADCP	plan!.exe BBPRIME!.EXE WHEXPRT!.EXE XFORM!.EXE
Satlantic ISUS (MBARI)	ISUSPro-1.2-Installer.exe SatCon-1.4-install.exe SatView-2.7.1-install.exe
Seabird SBE 16plus CTD	SBEDataProcessing_Win32_V5_31b.exe Seasave_Win32_V5_31a.exe SeatermAF_Win32_V1_14.exe Seaterm_Win32_V1_48.exe

8 Appendix B – Questionnaires and Responses

8.1 Questionnaire 1a: For Organizations Concerned with Performing Measurements

8.1.1 Questionnaire 1a

Dear {specific person(s)|Sir(s)},

As {oceanographers|seismologists involved in *program|institution*} you will probably be familiar with two new ocean observatory projects that are currently being implemented: VENUS (Victoria Experimental Network Under the Sea www.VENUS.uvic.ca) and NEPTUNE Canada (North East Pacific Time Series Underwater Networked Experiments www.NEPTUNECANADA.ca and <http://www.NEPTUNE.washington.edu>). Barrodale Computing Services Ltd. (BCS) - see www.barrodale.com, has recently been awarded a contract in an open competitive bid process issued by the VENUS/NEPTUNE project offices to examine DMAS (Data Management and Archive Systems) issues associated with these planned observatories.

In order to provide guidance and information to VENUS/NEPTUNE, one of the aspects we would like to determine is how data currently make their way from the point of measurement to the final archive. What issues have already been encountered and dealt with in managing the sorts of data that will be gathered by VENUS/NEPTUNE? To this end, we are seeking specific information about the oceanographic and seismological data (sea surface, water column, sea floor and ocean crust) that are dealt with at your institution. A questionnaire follows below. **Please feel free to answer as few or as many questions as your schedule allows; whatever feedback you have time to provide would be appreciated.**

Your feedback will assist the VENUS/NEPTUNE team to address the data management requirements for the respective systems. Any additional insights and recommendations would be most helpful, as would any documents, or links to documents, that would describe current data holdings and data management procedures. As this information is being collected for a report due in a matter of weeks, your timely response would be appreciated. Please feel free to contact me directly by phone (250-472-4370) or email (<mailto:mike@barrodale.com>) for any further clarification or discussions.

1. What types of data measurements are collected by your institution (e.g., XBT, pressure sensor readings, seismic measurements, video, HDTV, etc.)?
2. What platforms are used in performing each of these types of measurements (e.g., moored buoy, drifting buoy, moored instrument, shipboard instrument, cabled system, etc.)?

3. How are the measurements retrieved from the platform (e.g., manual recovery after deployment, dedicated communications link or line, VHF/UHF radio, satellite telemetry, etc.)? Which measurements, if any, are received in real time or near real time?
4. Are the measurement data stored locally at your repository? If so, what storage mechanism is used (e.g., database, filesystem, metadata in database; data in filesystem, etc.)? If a database is used, what type is it (e.g., Oracle, DB2, OODBMS, etc.)?
5. Are the measurements made available online locally? What form does this take (e.g., Web site, FTP site, local filesystem, OPeNDAP (DODS) server, etc.)? What file formats are used in presenting the data? Are there standard data formats used to present the data?
6. How long does it typically take for collected information to be submitted to the archive or database? Is any of this information processed and stored in real time / near real time?
7. Are the measurements subsequently transmitted (e.g., via GTS) to national or international repositories? Which ones? What form(s) of post-processing and quality control are performed before the data are shipped? How frequently are measurements transmitted? Are there data format standards and metadata standards that pertain to these transmissions?
8. Does the frequency of update depend on the type of data or measurement or staff/funding limitations? Is this difference due to technological issues (e.g., limited bandwidth) or application (science, business) requirements (e.g., some data elements may be used in weather forecasting and therefore must be current)?
9. With respect to the following list of initial VENUS instruments, do you know of any format standards, emerging format standards, or internal formats used by organizations concerned with these types of measurements?

CTD (Seabird 19+)

Oxygen Sensor (Optode 3975)

Gas Tension Device (GTD Pro)

Acoustic Doppler Current Profiler (RDI Deep Water Workhorse (300 kHz)

Digital Video Camera (Imenco IMDV 3018)
Orientation Sensor (Jasco AIM-2000)
Broadband Hydrophone System
Nitrate Sensor (MBARI-ISUS)
Flow Cytometer (FlowCAM)
Optical Backscatter Sensor (OBS-3)
Zooplankton Acoustic Profiler (ASL Water Column Profiler)
Seismometer (Guralp CMG-1T 3)

10. VENUS and NEPTUNE will be collecting and disseminating a wide range of data types at differing sample rates and intervals from a common geographical region, resulting in a real time heterogeneous data set. Are there any lessons learned from your experience that could assist in the development of the VENUS/NEPTUNE DMAS, or that could assist in avoiding potential pitfalls in the design of the DMAS? Any information you feel to be relevant would be appreciated.

8.1.2 Questionnaire 1a Recipients and Responders

Questionnaire 1a was sent to the 9 recipients in the following table.

Recipient	Responder
Dr. Raymond C. Highsmith Director – West Coast and Polar Regions Undersea Research Center	Rob Cermak NOAA/Undersea Research Center
Dr. David A. Butterfield NeMO contact Pacific Marine Environmental Laboratory, NOAA	
Dr. Michael J. McPhaden TAO Project Director NOAA	Paul Freitag TAO Project, NOAA/PMEL
Dr. Gregory C. Johnson Argo contact Pacific Marine Environmental Laboratory, NOAA	Gregory C. Johnson
Dr. Robert P. Dziak Vents Program contact Pacific Marine Environmental Laboratory, NOAA	Robert Dziak
Mr. Donald W. Denbo EPIC contact Pacific Marine Environmental Laboratory, NOAA Data Products Team	

(webmaster.ndbc@noaa.gov) Operations Branch NODC, NOAA	
Mr. Frank Whitney, Program Coordinator – Station P / Line P program Institute of Ocean Sciences	Frank Whitney
Dr. Tim Ahern Program Manager – IRIS	Tim Ahern interviewed (Section 9.6), and questionnaire returned.
Mr. Robin Brown, Head – Ocean Science and Productivity Institute of Ocean Sciences	Interviewed (Section 9.7)

8.1.3 Questionnaire 1a Responses

Responses were received from 6 of the 9 recipients of Questionnaire 1a (and Robin Brown was interviewed). The written responses for each responder are given in the table below.

Question	Response(s)
1. What types of data measurements are collected by your institution (e.g., XBT, pressure sensor readings, seismic measurements, video, HDTV, etc.)?	IOS, Station P / Line P Program (Frank Whitney) CTD (T, S, transmissometry, fluorescence, oxygen), water properties (oxygen, nutrients, dissolved inorganic carbon, dissolved organic carbon, particulate carbon, primary productivity, chlorophyll a, etc.), ocean currents by drifters and moored instruments, satellite imagery, plankton, acoustic data (both ambient noise and return echo). IRIS (Tim Ahern) The IRIS DMC collects data from roughly 20 different types of sensors. All sensors deliver data streams as a time series sampled at a constant rate in time. Sensors include seismometers, hydrophones, geophones, atmospheric sensors such as wind speed, atmospheric pressure, temperature, magnetic field, electrical field etc. NOAA/PMEL, Argo Program (Greg Johnson) Within the ocean we collect data on pressure, temperature, salinity, dissolved oxygen, nutrients, CFCs, CO₂, alkalinity, fluorescence, velocity, sound, video, and many other things. NOAA/PMEL, TAO Project (Paul Freitag) [Since the TAO Project collects, distributes and archives TAO/TRITON data I filled in both of the questionnaires, although several of the answers are duplicated. You can also find more about the TAO project from our Web site at www.pmel.noaa.gov/tao.] Standard measurements: surface winds, air temperature, relative

	humidity, water temperature from 1 m to 500 m. -Optional measurements: precipitation, short-wave radiation, long-wave radiation, barometric pressure, ocean conductivity (salinity), ocean currents. NOAA/PMEL, Vents Program (Robert Dziak) We use sound channel hydrophones to record the acoustic signals of seafloor earthquakes. NOAA/Undersea Research Center (Rob Cermak) The three types that I know are CTD, zooplankton and chlorophyll measurements. There are others, but those have not been recataloged or verified. Some weather data also has been collected from the NWS and National Buoy Center. - Most data in our archive are voluntarily supplied. Projects are NOT required to use our archive. We store RAW and Processed data upon request. We also forward this information to a reporting agent upon request.
2. What platforms are used in performing each of these types of measurements (e.g., moored buoy, drifting buoy, moored instrument, shipboard instrument, cabled system, etc.)?	IOS, Station P / Line P Program (Frank Whitney) Moored instruments, drifting buoys, ship based measurements, satellites, sampling of opportunity (lighthouses, commercial vessels). IRIS (Tim Ahern) Mostly land based observatories but also some ocean bottom sensors attached to undersea cables. We also have traditional marine seismic data from airguns and multichannel streamers. NOAA/PMEL, Argo Program (Greg Johnson) We use instrumented moored buoys, drifting surface buoys, profiling floats, shipboard systems, etc. NOAA/PMEL, TAO Project (Paul Freitag) Moored buoys. NOAA/PMEL, Vents Program (Robert Dziak) We use both moored buoys and cabled seafloor instruments. NOAA/Undersea Research Center (Rob Cermak) Moorings, shipboard, buoy.
3. How are the measurements retrieved from the platform (e.g., manual recovery after deployment, dedicated communications link or line, VHF/UHF radio, satellite telemetry, etc.)? Which measurements, if any, are received in	IOS, Station P / Line P Program (Frank Whitney) CTD data - real time Water properties - after analyses (from 1 day to 2 years) Moored instruments - post mooring recovery (few months to 1-2 years) Drifting buoys - near real time Satellite images - near real time to 6 months IRIS (Tim Ahern) A huge variety here... some are connected by land based IP systems, others by satellite both full period and periodic access. We also run private Frame Relay circuits, Business DSL circuits. Some stations are so remote that physical media transfer is still utilized although that is becoming less frequent. Much of our

real time or near real time?	data are now received through automated electronic means. NOAA/PMEL, Argo Program (Greg Johnson) The measurements are retrieved in a number of ways but mostly manual recovery after deployment, and satellite telemetry (Service Argos, Iridium, etc). NOAA/PMEL, TAO Project (Paul Freitag) Daily means and a few hourly samples: NOAA POES and Service Argos. Complete high temporal resolution: manual recovery. NOAA/PMEL, Vents Program (Robert Dziak) The moored buoys are manually recovered with significant delay times. The cabled hydrophones are near real-time. NOAA/Undersea Research Center (Rob Cermak) Manual recovery after deployment. CTD data take about a year to process before it reaches the archive. Zooplankton and chlorophyll additional time beyond a year is necessary.
4. Are the measurement data stored locally at your repository? If so, what storage mechanism is used (e.g., database, filesystem, metadata in database; data in filesystem, etc.)? If a database is used, what type is it (e.g., Oracle, DB2, OODBMS, etc.)?	IOS, Station P / Line P Program (Frank Whitney) Data are stored at IOS; contact Joe Linguanti for details. Many data sets are sent to MEDS (DFO Ottawa) which is linked to international data archives. IRIS (Tim Ahern) The time series data are stored locally. We have a 1 petabyte capacity tape robotic mass storage system. We have roughly 12 terabytes of mid-range high performance RAID as a front end. The data are available through a variety of IRIS developed request tools that ultimately access the storage systems. Event segmented data are stored on a large Internet accessible RAID system where they can be accessed by FTP or through a variety of powerful data access tools. All metadata are stored in an Oracle DBMS running on a parallel SUN servers. NOAA/PMEL, Argo Program (Greg Johnson) We have our own database, EPIC, to store most of our profile (CTD, XBT, XCTD, etc.) data, but I am not the right person to talk with about this. NOAA/PMEL, TAO Project (Paul Freitag) Database (mysql) and filesystem; includes both observations and metadata. NOAA/PMEL, Vents Program (Robert Dziak) We store our data locally on DLT tape and on hard drives in file systems. NOAA/Undersea Research Center (Rob Cermak) PostgreSQL (for Linux) is now being utilized. However, we are using mysql (Linux) more heavily in past years. We have intentions to move the PostgreSQL database to mysql in the near future.

5. Are the measurements made available online locally? What form does this take (e.g., Web site, FTP site, local filesystem, OPeNDAP (DODS) server, etc.)? What file formats are used in presenting the data? Are there standard data formats used to present the data?	IOS, Station P / Line P Program (Frank Whitney) Many data sets are available on-line (Web sites, FTP); see http://www-sci.pac.dfo-mpo.gc.ca/osap/default_e.htm and http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Home_e.htm. IRIS (Tim Ahern) File formats are either FDSN SEED format (www.fdsn.org) or SEG-Y format. Data are available through Web servers, FTP servers, a rich set of data request tools, and a comprehensive set of CORBA based services that allow application to application connectivity. NOAA/PMEL, Argo Program (Greg Johnson) Much of the PMEL data are made available online locally through a variety of Web sites, FTP sites, and the like. NOAA/PMEL, TAO Project (Paul Freitag) Web site and legacy FTP site. Flat ASCII and netCDF. NOAA/PMEL, Vents Program (Robert Dziak) We make our moored hydrophone data (both processed and raw) available online on our Web site. We use a local binary format as well as CSS3.0. The raw, cabled hydrophone data are classified and cannot be publicly distributed. NOAA/Undersea Research Center (Rob Cermak) A PHP script on our Web site makes some CTD and chlorophyll data available in ODV format.
6. How long does it typically take for collected information to be submitted to the archive or database? Is any of this information processed and stored in real time / near real time?	IOS, Station P / Line P Program (Frank Whitney) Near real time - Argo profilers (drifters that collect T and S data every 10 days to 2000 m) http://w3.jcommops.org/cgi-bin/WebObjects/Argo. Up to 2 years (or more) for data from water samples. CTD data are typically processed and archived within 4 to 6 months. Some data sets remain proprietary. IRIS (Tim Ahern) A large fraction of our data is received in near real time... tens of seconds to a couple of minutes. This is for roughly 800 stations worldwide. Outages happen on a global basis and delays can happen of days to weeks to months. Some data still come by physical media although this is becoming less common but it can take 1-3 months to acquire some data. NOAA/PMEL, Argo Program (Greg Johnson) Some data (TAO Program, Argo Program, maybe others) are processed and stored in near real time. It can take as much as two years for some cruise data that require the highest possible accuracy and precision to be finalized and archived. NOAA/PMEL, TAO Project (Paul Freitag) Daily means updated daily in near real time, about 1-2 days after observation. High temporal resolution data are available within

	months of mooring recovery. NOAA/PMEL, Vents Program (Robert Dziak) It takes a few weeks to put our moored hydrophone data into our archive once data are recovered and brought back to my lab. Our cabled hydrophone data are stored in near-real-time. Earthquake locations derived from processed data take between 24 hours to 6 months to get to our online database depending on the interest of the measurement. NOAA/Undersea Research Center (Rob Cermak) No real time data are available at the present time. Typical delay is about one year.
7. Are the measurements subsequently transmitted (e.g., via GTS) to national or international repositories? Which ones? What form(s) of post-processing and quality control are performed before the data are shipped? How frequently are measurements transmitted? Are there data format standards and metadata standards that pertain to these transmissions?	IOS, Station P / Line P Program (Frank Whitney) See above. IRIS (Tim Ahern) We are the repository for all IRIS data. We are the archive for FDSN Continuous data (an international body) and we are the archive for most but not all USGS generated data in the US and worldwide. -We have extensive quality control done at several different nodes around the US. We are also just beginning to do automated QC in near real time for most of the seismic data. There are data format standards developed by the FDSN and SEG that we follow. -Communication protocols are of many different types and there are no real standards here other than all are based on IP.... NOAA/PMEL, Argo Program (Greg Johnson) I know the Argo Program and TAO program data are transmitted in near real time over the GTS, and much of our data eventually goes to NODC and thus the WODC on longer time-frames. NOAA/PMEL, TAO Project (Paul Freitag) Real time data are placed on GTS in WMO BUOY code by Service Argos within hours of transmission to satellite. Data are made available via Web and/or FTP for download by NODC, NCDC, etc. NOAA/PMEL, Vents Program (Robert Dziak) No. NOAA/Undersea Research Center (Rob Cermak) GLOBEC is the only project that has a reporting requirement known to data management. The reporting agency knows how to access the Globec information from our ASCII archive. Typical format they use are fixed columns of ASCII. Flat text files.
8. Does the frequency of update depend on the type of data or measurement or	IOS, Station P / Line P Program (Frank Whitney) Data availability depends on the structure of various programs. Argo has been established to provide near real time data to users, Line P attempts to make data available on-line within months,

staff/funding limitations? Is this difference due to technological issues (e.g., limited bandwidth) or application (science, business) requirements (e.g., some data elements may be used in weather forecasting and therefore must be current)?	other programs that are of less general interest will provide data to archives over a 2 year period. Novel or developmental data may never be archived (e.g., acoustic data on sea surface noise, wind speed). IRIS (Tim Ahern) Systems are automated. Most of the delay is for fundamental technical, political or infrastructural reasons, most systems do not rely on people to do them... We are highly automated. NOAA/PMEL, Argo Program (Greg Johnson) A mix of variables decide whether data are made available in near real time including funding, operational needs, and scientific purpose. NOAA/PMEL, TAO Project (Paul Freitag) Daily mean updates are sufficient for climate monitoring. A few hourly samples are sent for weather forecast purposes. Transmission limited by overpass schedule of POES, and telemetry costs. NOAA/PMEL, Vents Program (Robert Dziak) Data updates do depend on funding limitations, which control staffing levels. Our cabled hydrophone data have a first-order earthquake location analysis performed on a daily basis as part of proposal to provide volcanic event detection information to university colleagues. NOAA/Undersea Research Center (Rob Cermak) Huge under-staffing and under-funding are likely the reasons for low submission rates from PIs into the archive. The archive is very dysfunctional which likely enhances the disuse and low submission rate.
9. With respect to the following list of initial VENUS instruments, do you know of any format standards, emerging format standards, or internal formats used by organizations concerned with these types of measurements? • CTD (Seabird 19+) • Oxygen Sensor (Optode 3975) • Gas Tension Device (GTD Pro)	IOS, Station P / Line P Program (Frank Whitney) I am unaware of standards. However, I am most interested in CTD and nitrate sensor data, which I find easiest to use when retrievable as text files that can be imported into spreadsheets. Contact PICES http://www.pices.int/ for information on metadata archives for the North Pacific. IRIS (Tim Ahern) SEED manages data from broadband hydrophone systems and seismometers. SEED can manage just about any kind of regularly sampled time series data and several of yours are of that type. NOAA/PMEL, Argo Program (Greg Johnson) No response. NOAA/PMEL, TAO Project (Paul Freitag) The OceanSites Data management group has been working in collaboration with other international groups to define a common format for the OceanSites program time series data. Contact Sylvie POULIQUEN [Sylvie.Pouliquen@ifremer.fr] for

• Acoustic Doppler Current Profiler (RDI Deep Water Workhorse (300 kHz)) • Digital Video Camera (Imenco IMDV 3018) • Orientation Sensor (Jasco AIM-2000) • Broadband Hydrophone System • Nitrate Sensor (MBARI-ISUS) • Flow Cytometer (FlowCAM) • Optical Backscatter Sensor (OBS-3) • Zooplankton Acoustic Profiler (ASL Water Column Profiler) • Seismometer (Guralp CMG-1T 3)	details. NOAA/PMEL, Vents Program (Robert Dziak) The United Nations certified test ban treaty project has decided on CSS3.0 as their hydrophone data format. NOAA/Undersea Research Center (Rob Cermak) See Question 10.
10 .VENUS and NEPTUNE will be collecting and disseminating a wide range of data types at differing sample rates and intervals from a common geographical region, resulting in a real time heterogeneous data set. Are there any lessons learned from your experience that could assist in the development of the VENUS/NEPTUNE DMAS, or that could assist in avoiding potential pitfalls in the design of the	IOS, Station P / Line P Program (Frank Whitney) Data archiving expertise lies within MEDS. Contact them regarding pitfalls and data standards (Dr Narayanan, director). See http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Contact_US/Staff_e.htm. IRIS (Tim Ahern) This is a very hard question to answer. Let me just say that IRIS already does, and has been doing, a very similar thing for nearly a decade. We adopted format standards early on and that has proven to be a great benefit for managing these data. We attempt to homogenize the data flow and formats before they enter (or as they enter) the DMC and therefore most data appear similar to us in terms of our automated system. -This is an incomplete answer but your question is difficult to fully answer. NOAA/PMEL, Argo Program (Greg Johnson) VENUS and NEPTUNE should plan on significant spending for data management of a large and heterogeneous data set. NOAA/PMEL, TAO Project (Paul Freitag) Our experience suggests that it will be important for the VENUS/NEPTUNE DMAS to allow user-defined subsets of the

DMAS? Any information you feel to be relevant would be appreciated.	data to be extracted and downloaded. Even for users who want all of the data, there are often bandwidth limitations, particularly in the third world, which make the option to download a subset of the data an essential capability. -The design of our data delivery system includes data availability information in the user interface, which quickly shows the user what data are available, prior to the user-request. For example see http://www.pmel.noaa.gov/tao/data_deliv/ -Many delivery systems require the user to make their request first, only to be informed that there are no data fitting their criteria, which can be frustrating, inefficient, and confusing. -Another suggestion is to make as much metadata as possible available from the Web or served with the data, as this will undoubtedly limit the number of questions you receive from users and make some responses as simple as providing a URL. We provide much of our metadata at http://www.pmel.noaa.gov/tao/proj_over/proj_over.html NOAA/PMEL, Vents Program (Robert Dziak) Do you plan to provide any type of first-order analysis of your wide range of data types? Or do you simply plan to archive data then let investigators sort through the data online? Your varied data streams will require some type of intelligent algorithm (with an analyst providing quality assurance) to sort and classify the data. For example, I would think some kind of earthquake pick file will have to be produced from the hydrophone and seismometer data streams, otherwise you will quickly fall behind. Perhaps the solution is to request that outside organizations that already have data analysis infrastructure in place should coordinate VENUS/NEPTUNE records with their existing records to provide a preliminary online database. NOAA/Undersea Research Center (Rob Cermak) There are so many data formats. Pulling a bunch of scientists together from different institutions means you will have to deal with several different commercial software packages that all want different formats – some with closed proprietary formats. Some may have import features, but once the data are in that package and results are found, getting the results out so that others may utilize it is very difficult in some cases. -Any DMAS that wants to be successful has to provide data to PIs in as many formats as possible. This is technically infeasible due to the closed, proprietary formats used in some software and sampling applications. -Due to funding and technical restrictions, the data are made available in ASCII format. We make it available in ODV since the format for data import is somewhat documented and ODV is
--	---

	used by a large number of faculty. Matlab is another well used application and can readily import ASCII tables. -To this day, we still have a couple of professors running databases on Windows 98 machines because: -The data only reside in a proprietary system / data format, and the company has then ceased to exist or software is not available for modern operating systems. -Have insufficient funds to purchase a new version of the database or application to make data available on modern operating systems. -No matter what format you decide to store data and information in, make sure there is a reasonable way to export these data into modern systems. Usage of a single proprietary format will automatically close the use of archived data to other researchers. -The only successful DMAS team is JOSS. Although highly expensive for data storage, they are getting sufficient funds; it seems to make their data archive work. http://www.joss.ucar.edu/. -The EVOS/GEM project has not been successful getting their data management section on-line. They have gone through at least 3 different data managers. http://www.evostc.state.ak.us/gem/ -UAF Alaska has started a central data archive, but has yet to embrace the usage of other data sets much beyond those from satellite systems. See GINA: http://www.gina.alaska.edu/ -I hope you are ready for a huge undertaking...
--	--

8.2 Questionnaire 1b: For Organizations Concerned with Collecting and Disseminating Information from Other Repositories.

8.2.1 Questionnaire 1b

Dear {specific person(s)}|Sir(s)},

As {oceanographers|seismologists involved in *program*|*institution*} you will probably be familiar with two new ocean observatory projects that are currently being implemented: VENUS (Victoria Experimental Network Under the Sea www.VENUS.uvic.ca) and NEPTUNE Canada (North East Pacific Time Series Underwater Networked Experiments www.NEPTUNECANADA.ca and <http://www.NEPTUNE.washington.edu>). Barrodale Computing Services Ltd. (BCS) - see www.barrodale.com, has recently been awarded a contract in an open competitive bid process issued by the VENUS/NEPTUNE project offices to examine DMAS (Data Management and Archive Systems) issues associated with these planned observatories.

In order to provide guidance and information to VENUS/NEPTUNE, one of the aspects we would like to determine is how data currently make their way from the point of measurement to the final archive. What issues have already been encountered and dealt with in managing the sorts of data that will be gathered by VENUS/NEPTUNE? To this end, we are seeking specific information about the oceanographic and seismological data (sea surface, water column, sea floor and ocean crust) that are collected from other sites by your institution (and possibly further passed on). A questionnaire follows below.

Please feel free to answer as few or as many questions as your schedule allows; whatever feedback you have time to provide would be appreciated.

Your feedback will assist the VENUS/NEPTUNE team to address the data management requirements for the respective systems. Any additional insights and recommendations would be most helpful, as would any documents, or links to documents, that would describe current data holdings and data management procedures. As this information is being collected for a report due in a matter of weeks, your timely response would be appreciated. Please feel free to contact me directly by phone (250-472-4370) or email (<mailto:mike@barrodale.com>) for any further clarification or discussions.

1. From what sites, systems, or data sources are data collected?
2. What types of data are received from each?
3. What are the update frequencies for each data type / site?
4. Does your site receive any near real time or real time data from other sites? What types of data are these? From what sites do you receive them?
5. Do you receive and/or archive video or HDTV data?
6. Are the data that are received from other sites subsequently post-processed at your site? What is the nature of this post-processing?
7. From your site are the data further transmitted to other sites? Which types of data are transmitted, and to which sites? Are the data transmitted on a regularly scheduled basis? Are there data format standards and metadata standards that pertain to these transmissions? Are there any other inter-repository data interoperability issues?

8. What storage mechanism is used for the data at your site (e.g., database, filesystem, metadata in database, data in filesystem, etc.)? If a database is used, what type is it? (e.g., Oracle, DB2, other DBMS/OODBMS/ORDBMS, etc.)
9. What mechanisms are used to provide users access to these data (e.g., Web site, FTP site, OPeNDAP server, etc.)? Are there standard data formats used to present each of these types of data?
10. What is the retention period for each of these types of data? How much data are routinely discarded?
11. With respect to the following list of initial VENUS instruments, do you know of any format standards, emerging format standards, or internal formats used by organizations concerned with these types of measurements?

CTD (Seabird 19+)
Oxygen Sensor (Optode 3975)
Gas Tension Device (GTD Pro)
Acoustic Doppler Current Profiler (RDI Deep Water Workhorse (300 kHz)
Digital Video Camera (Imenco IMDV 3018)
Orientation Sensor (Jasco AIM-2000)
Broadband Hydrophone System
Nitrate Sensor (MBARI-ISUS)
Flow Cytometer (FlowCAM)
Optical Backscatter Sensor (OBS-3)
Zooplankton Acoustic Profiler (ASL Water Column Profiler)
Seismometer (Guralp CMG-1T 3)

12. VENUS and NEPTUNE will be collecting and disseminating a wide range of data types at differing sample rates and intervals from a common geographical region, resulting in a real time heterogeneous data set. Are there any lessons learned from your experience that could assist in the development of the VENUS/NEPTUNE DMAS, or that could assist in avoiding potential pitfalls in the design of the DMAS? Any information you feel to be relevant would be appreciated.

8.2.2 Questionnaire 1b Recipients and Responders

Questionnaire 1b was sent to the 18 recipients in the following table.

Recipient	Responder
Dr. Michael J. McPhaden TAO Project Director NOAA	Paul Freitag TAO Project, NOAA/PMEL
Mr. Donald W. Denbo EPIC contact Pacific Marine Environmental Laboratory, NOAA	
Data Products Team (webmaster.ndbc@noaa.gov) Operations Branch NODC, NOAA	
CLIMVIS / VOSCLIM program contact (ncdc.info@noaa.gov) NCDC, NOAA	
Mr. David Divins Bathymetry contact NGDC, NOAA	Interviewed (comments incorporated into body of report)
Mr. Dan R. Metzger and Mr. John G. Campagnoli GEODAS contact NGDC, NOAA	
Ms. Carla J. Moore Geology contact NGDC, NOAA	
Mr. Francis Mitchell NODC Data Acquisition Specialist NODC	
Mr. Donald Collins IODE contact NODC, NOAA	
Mr. Patrick Caldwell Joint Archive for Shipboard ADCP contact NODC, NOAA	Patrick Caldwell
Dr. Wayne L. Wilmot Chief – Coastal Ocean Laboratory (Coastal Ocean Time Series Database and Global Temperature-Salinity Profile Program (GTSPP) contact) NODC, NOAA	
Dr. Savi Narayanan Director – Marine Environmental Data Service (MEDS)	Jean Gagnon Chief – Data Management and Client Services

Mr. Robin Brown, Head – Ocean Science and Productivity Institute of Ocean Sciences	Interviewed (Section 9.7)
Dr. Gerold Wefer, Director – WDC for Marine Environmental Sciences	
Dr. George F. Sharman Director – WDC for Marine Geology and Geophysics	
Dr. Sydney Levitus Director – WDC for Oceanography	Robert Gelfeld
Dr. Stuart Sipkin Director – WDC for Seismology	Stuart Sipkin Harold Bolton Manager – Data Collection Centre
Dr. Tim Ahern Program Manager – IRIS	Tim Ahern (also interviewed – Section 9.6)

8.2.3 Questionnaire 1b Responses

Responses were received from 6 of the 18 recipients of Questionnaire 1b (and Robin Brown, David Divins, and Tim Ahern were interviewed). The written responses for each responder are given in the table below.

Question	Response(s)
General	WDC for Seismology – Denver (Stuart Sipkin) The WDC for Seismology-Denver is the repository for parametric seismological data. If I understand your query correctly, you are primarily interested in the acquisition, archiving, and distribution of waveform data. The U.S. Geological Survey and IRIS have an arrangement whereby the IRIS Data Management Center in Seattle is the primary repository for waveform data, including that collected by the USGS. Dr. Tim Ahern, the director of the IRIS DMC is in a much better position to give a full and complete answer to your questions. WDC for Seismology – Denver (Harold Bolton) I am the manager of the Data Collection Center that collects, processes, distributes and archives the waveform data from the broad band Global Seismic Data (GSN). We (USGS) have a collaborative agreement with IRIS that relegates various aspects of the responsibility for these seismic data. -Our data center applies various QC methods to the data. We then ship data to the IRIS DMC. We do this on a daily basis. All the data we ingest into our system for processing and

	distribution are converted to SEED (rather mini-SEED). This is the format that is mandated by our agreements with IRIS. Total daily volume of data is on the order of 1.5 gigabytes. We then later apply QC again after the latent data have been received (from hard disk, tapes, etc.) and reship to the IRIS DMC the appended data sets. This again averages about 1.5 gigabytes/day. DFO/MEDS (Jean Gagnon) The Marine Environmental Data Service (MEDS) is a branch of Canada's federal Department of Fisheries and Oceans (DFO) Science Sector. MEDS' mandate is to manage and archive ocean data collected by DFO, or acquired through national and international programs conducted in ocean areas adjacent to Canada, and to disseminate data, data products, and services to the marine community in accordance with the policies of the Department. -Marine data management within DFO Science is a distributed function coordinated nationally through the Working Group on Marine Data Management (WGMDM) chaired by MEDS. DFO/MEDS, as a supporter of the VENUS/NEPTUNE project, offers the following advice and information to the Barrodale Computing Services Ltd. (BCS) questionnaire. NOAA/WDCO (Robert Gelfeld) Syd Levitus, Director of the World Data Center for Oceanography, Silver Spring, has asked me to respond to your request about how we collect oceanographic information. It is quite a lengthy and complicated process as you can well imagine. I would like to refer you to our Web site: www.nodc.noaa.gov. This would be a good starting point for you to get an overview of all activities as both a national oceanographic data center and collectively a World Data Center. I have spoken to my counterpart (Mr. Bob Keeley) at the Canadian National Oceanographic Data Center (MEDS, Fisheries and Oceans Canada in Ottawa) and he has told me that they too were contacted by you. We normally work through them and other international centers to coordinate our international data acquisition and exchange. You might like to visit the intergovernmental Oceanographic Commission's Web site (www.iode.org) for further information. -World Data Center (WDC) for Oceanography is one component of a global network of discipline subcenters that facilitate international exchange of scientific data. Originally established during the International Geophysical Year of 1957-58, the World Data Center System functions under the guidance of the International Council of Scientific Unions (ICSU). WDC for Oceanography, Silver Spring, is collocated with, and operated
--	---

	by, the U.S. National Oceanographic Data Center (NODC). The following URL describes it function in detail: http://www.nodc.noaa.gov/General/NODC-dataexch/NODC-wdca.html -To specifically answer your questions, WDCA does not collect cabled observatories, similar to VENUS/NEPTUNE. QC and post-processing are done by our Ocean Climate Laboratory (http://www.nodc.noaa.gov/OC5/). We produce the World Ocean Database. These are collected in conjunction with the Oceanographic Commission (IOC) Global Oceanographic Data Archaeology and Rescue (GODAR) and World Ocean Database (WOD) projects. This database contains a collection of scientifically quality controlled ocean profile and plankton data that includes measurements of temperature, salinity, oxygen, phosphate, nitrate, silicate, chlorophyll, alkalinity, pH, pCO₂, and tCO₂. These and all associated products and data are described at: http://www.nodc.noaa.gov/OC5/indprod.html. There is a lot of information here to review. Please contact me if you have follow-up questions.
1. From what sites, systems, or data sources are data collected?	DFO/MEDS (Jean Gagnon) MEDS no longer conducts in situ field activities <i>per se</i>; however, as the national data center it does acquire, process, quality control, archive, and generate / disseminate data and data products collected from a variety of national programs conducted by DFO regional institutes, other government Departments, Universities, Private industry, etc., in the Canadian Area of Interest (35N-90N, 40W-180W) as well as from <u>global</u> international programs (ARGO, Drifting Buoy, GTSP, etc.) as described on its Web site http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Home_e.htm. Sites, systems, and data sources are data type specific and generally described in the above site. More detailed system flow information on any particular program is available upon request. DFO Science Pacific Region conduct multidisciplinary regional field activities as described on http://www-sci.pac.dfo-mpo.gc.ca/sci/divisions/default_e.htm which may also be relevant to the VENUS/NEPTUNE projects. IRIS (Tim Ahern) We collect data from roughly 1500 observatory locations worldwide. NOAA/NODC (Patrick Caldwell) Can get off Web: http://ilikai.soest.hawaii.edu/sadcp/ http://ilikai.soest.hawaii.edu/UHSLC/jasl.html http://ilikai.soest.hawaii.edu/HILO/ Additional notes provided by BCS: Joint Archive for Shipboard ADCP

	(http://ilikai.soest.hawaii.edu/sadcp/) include this gif: ftp://ilikai.soest.hawaii.edu/caldwell_pub/adcp/ESM/march_95/gdb.gif Joint Archive for Sea Level (http://ilikai.soest.hawaii.edu/UHSLC/jasl.html) NESDIS/NODC/NCDDC Hawaii/Pacific Islands Liaison (http://ilikai.soest.hawaii.edu/HILO/) spreadsheet of data sets at http://ilikai.soest.hawaii.edu/HILO/data/sort_PCID.html NOAA/PMEL, TAO Project (Paul Freitag) TAO/TRITON and PIRATA.
2. What types of data are received from each?	DFO/MEDS (Jean Gagnon) At MEDS, the present “active” data types processed at MEDS include: ocean profile data (physical properties such as T, S; nutrients, optical properties; chemistry; currents, etc.); plankton data; tides and water level time series data; surface wave time series data; drifting buoy data; profiling float data; thermosalinograph data; contaminants data (water column, animal sediment, in fresh and salt water); environmental data from the offshore oil and gas industry. -For more detailed information on any of the above see: http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Databases/Data_e.htm http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Databases/Data_e.htm http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Prog_Int/Prog_Int_e.html IRIS (Tim Ahern) Again this is very extensive and varies by site. Seismic sensors are the dominant source for us but we have the other data types I mentioned previously as well. Roughly 20 sensor types flowing in mostly real time. NOAA/NODC (Patrick Caldwell) Can get off Web: http://ilikai.soest.hawaii.edu/sadcp/ http://ilikai.soest.hawaii.edu/UHSLC/jasl.html http://ilikai.soest.hawaii.edu/HILO/ NOAA/PMEL, TAO Project (Paul Freitag) Standard measurements: Surface winds, air temperature, relative humidity, water temperature from 1 m to 500 m.

	Optional measurements: Precipitation, short-wave radiation, long-wave radiation, barometric pressure, ocean conductivity (salinity), ocean currents.
3. What are the update frequencies for each data type / site?	DFO/MEDS (Jean Gagnon) Daily: ocean profile data; tides and water level time series data; surface wave time series data; drifting buoy data; profiling float data; thermosalinograph data. Ad-hoc: contaminants data; environmental data from the offshore oil and gas industry. IRIS (Tim Ahern) Real time in general. Tape based transfer systems send data weekly or monthly. NOAA/NODC (Patrick Caldwell) Varies. NOAA/PMEL, TAO Project (Paul Freitag) Daily means updated daily in near real time, about 1-2 days after observation. High temporal resolution data are available within months of mooring recovery.
4. Does your site receive any near real time or real time data from other sites? What types of data are these? From what sites do you receive them?	DFO/MEDS (Jean Gagnon) The Canadian Atlantic Zone Monitoring Program provides on-line access to ocean monitoring data collected in the North West Atlantic. Hydrographic data (temperature, salinity, dissolved oxygen, fluorescence, chlorophyll a and nutrients); climate indices; plankton; sea levels; remote sensing; and meteorological data. (http://www.meds-sdmm.dfo-mpo.gc.ca/zmp/main_zmp_e.html) -MEDS is the <u>world</u> data centre for drifting buoys (Responsible National Oceanographic Data Centre - RNODC). As part of its role, MEDS acquires, processes, quality controls and archives real time drifting buoy messages reporting over the Global Telecommunications System (GTS). (http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Databases/DRIBU/drifting_buoys_e.htm) -The Canadian ARGO profiling float data and information is available on-line with links to that international global program. Data are received every six hours, processed automatically and subjected to duplicate checks and data quality control. They are then transmitted to the Global Telecommunication System (GTS) and Argo Servers within 24 hours. (http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Prog_Int/Argo/ArgoHome_e.html) -MEDS receives, processes and archives global surface

	thermosalinograph temperature and salinity observations transmitted on the Global Telecommunications System in the international TRACKOB code form. Monthly maps show the availability of these data. (http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Databases/TrackOB/Trackob_e.htm) -MEDS acquires, processes, quality controls, archives, and distributes physical, chemical and biological profiles of oceanographic data from several National and <u>global</u> International programs. Lower resolution data are available in real time. (http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Databases/OCEAN/Realtime_e.htm) -MEDS acquires, processes, quality controls, archives and distributes tide and water level (TWL) data reported on a daily to monthly basis from the DFO Canadian Hydrographic Service (CHS) water level gauging network. (http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Databases/TWL/TWL_e.htm) -MEDS acquires wave buoy data from buoys operated by the Meteorological Service of Canada as transmitted through the GOES satellite network on a daily basis. (http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Databases/WAVE/WAVE_e.htm) IRIS (Tim Ahern) Yes we do. We receive data from roughly two dozen different centers all over the globe. So some real time data come to us directly from the stations while other comes through as through relay nodes such as data centers around the world. NOAA/NODC (Patrick Caldwell) Focus is on delayed-mode for posterity archive. NOAA/PMEL, TAO Project (Paul Freitag) TRITON data are obtained from JAMSTEC and merged with TAO. Data types are nearly identical.
5. Do you receive and/or archive video or HDTV data?	DFO/MEDS (Jean Gagnon) Not at present, however we have had some experience with remote sensing data in the past, as described in http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Databases/Satellite/omw/OMW_e.htm. IRIS (Tim Ahern) No we don't. NOAA/NODC (Patrick Caldwell) No. NOAA/PMEL, TAO Project (Paul Freitag) No.
6. Are the data that are received from other sites subsequently post-processed at your	DFO/MEDS (Jean Gagnon) Yes. The nature of the post-processing functions are: Assuring that standards for metadata required for proper post-processing and archival of the source data are maintained; Ensuring consistent processing for the different instrumentation

site? What is the nature of this post-processing?	and communications media into homogeneous national databases is maintained; Standardized quality control methodologies for the national databases are maintained; and Assuring long-term off-site archive back-ups of marine science data are maintained. These functions ensure that data and data products derived from these data are reliable and generally accessible. IRIS (Tim Ahern) They are not post processed. Quality is assessed on all time series entering the DMC through our real time automated systems. NOAA/NODC (Patrick Caldwell) No. NOAA/PMEL, TAO Project (Paul Freitag) After TRITON data are received they are joined with TAO data and gridded uniformly for display purposes only.
7. From your site are the data further transmitted to other sites? Which types of data are transmitted, and to which sites? Are the data transmitted on a regularly scheduled basis? Are there data format standards and metadata standards that pertain to these transmissions? Are there any other inter-repository data interoperability issues?	DFO/MEDS (Jean Gagnon) Yes. Most data and data products are generally made available through MEDS Web site and other DFO data portals. Other special arrangements are based on client requirements ranging from specialized daily real time data files of drifting buoy data as input to regional Search and Rescue models, to yearly exchanges of delayed-mode higher resolution profile data to meet international exchange agreements. Data formats can be client-specific or as described on our Web site by data type. Metadata standards generally adhere to international guidelines as described in www.ices.dk/committe/occ/mdm/guidelines . MEDS also conforms to the open DFO Science Management Policy for Scientific Data as described in http://www.dfo-mpo.gc.ca/science/data-donnees/datapolicy_e.htm IRIS (Tim Ahern) We do supply significant amounts of data to other data centers. Again most of this is in the SEED format through a variety of transmission protocols (SEEDlink, LISS, autoDRM, CORBA/DHI). Any of the data that we receive electronically can be sent electronically in near real time. The CORBA based systems can send any data from our 50 terabyte archive of time series to end users. NOAA/NODC (Patrick Caldwell) No, sea level and ADCP, in-house format, will migrate to FGDC in coming years. NOAA/PMEL, TAO Project (Paul Freitag) Real time data are placed on GTS in WMO BUOY code by Service Argos within hours of transmission to satellite. Data are made available via

	Web and/or FTP for download by NODC, NCDC, etc.
8. What storage mechanism is used for the data at your site (e.g., database, filesystem, metadata in database, data in filesystem, etc.)? If a database is used, what type is it? (e.g., Oracle, DB2, other DBMS/OODBMS/ORDBMS, etc.)	DFO/MEDS (Jean Gagnon) All of the above, dependent upon operational requirements for access to the data. Storage media include paper, analog microfilm/microfiche, and a variety of magnetic DLT, digital CD-ROM or DVD storage media. On-line databases range from VAX indexed sequential binary files to Oracle relational databases. IRIS (Tim Ahern) Same as above question. 1 .2 petabyte Storage Tek robot with 12 terabyte from end RAID disk. On-line RAID system containing event windowed data is available over the Internet using FTP, Web-based or a variety of request tools. NOAA/NODC (Patrick Caldwell) Flat files, in-house software. NOAA/PMEL, TAO Project (Paul Freitag) Database (MySQL) and filesystem; includes both observations and metadata.
9. What mechanisms are used to provide users access to these data (e.g., Web site, FTP site, OPeNDAP server, etc.)? Are there standard data formats used to present each of these types of data?	DFO/MEDS (Jean Gagnon) See Question 7 response. Our present approach is to use existing, Web-based technologies as the window to the data. We propose to use the Geoportal initiative as the way to show what data are available and their locations and times. Achieving integration of the data will require linking the data “discovery” of Geoportal to the source archives and making those data available for download. We also provide special and secure services based on client requirements to update and retrieve data remotely via the Web from certain databases. MEDS intends to install an OPeNDAP server. IRIS (Tim Ahern) FTP, Web, DHI/CORBA, user request tools are all options. Standard formats are SEED or SEG-Y. We still do send out physical media from tape to DVD as requested by users. NOAA/NODC (Patrick Caldwell) Yes, we do all those. NOAA/PMEL, TAO Project (Paul Freitag) Web site and legacy FTP site. Flat ASCII and netCDF.
10. What is the retention period for each of these types of data? How much data are routinely discarded?	DFO/MEDS (Jean Gagnon) Marine scientific data at DFO are viewed as a capital asset and identified for perpetual back-up retention with National Archives of Canada. Although the retention media may change with technology and some data may be flagged as bad or erroneous, none is discarded unless explicitly instructed to do so by the data originator. IRIS (Tim Ahern) All data are retained forever....

	NOAA/NODC (Patrick Caldwell) None. NOAA/PMEL, TAO Project (Paul Freitag) The retention period is indefinitely. Data are never discarded.
11. With respect to the following list of initial VENUS instruments, do you know of any format standards, emerging format standards, or internal formats used by organizations concerned with these types of measurements? • CTD (Seabird 19+) • Oxygen Sensor (Optode 3975) • Gas Tension Device (GTD Pro) • Acoustic Doppler Current Profiler (RDI Deep Water Workhorse (300 kHz)) • Digital Video Camera (Imenco IMDV 3018) • Orientation Sensor (Jasco AIM-2000) • Broadband Hydrophone System • Nitrate Sensor (MBARI-ISUS) • Flow Cytometer (FlowCAM) • Optical Backscatter Sensor (OBS-3) • Zooplankton Acoustic Profiler (ASL Water Column Profiler) • Seismometer	DFO/MEDS (Jean Gagnon) MEDS does not have this instrument level specific detail or experience, other than the international guidelines as described in www.ices.dk/committe/occ/mdm/guidelines and possibly through http://www.oceanportal.org/. IRIS (Tim Ahern) SEED can manage just about any kind of regularly sampled time series data and several of yours are of that type. NOAA/NODC (Patrick Caldwell) No. NOAA/PMEL, TAO Project (Paul Freitag) The OceanSites Data management group has been working in collaboration with other international groups to define a common format for the OceanSites program time series data. Contact Sylvie Pouliquen [Sylvie.Pouliquen@ifremer.fr] for details.

(Guralp CMG-1T 3)	
12. VENUS and NEPTUNE will be collecting and disseminating a wide range of data types at differing sample rates and intervals from a common geographical region, resulting in a real time heterogeneous data set. Are there any lessons learned from your experience that could assist in the development of the VENUS/NEPTUNE DMAS, or that could assist in avoiding potential pitfalls in the design of the DMAS? Any information you feel to be relevant would be appreciated.	DFO/MEDS (Jean Gagnon) A good general reference guide regarding marine data management of multi-disciplinary data sets has been documented in http://ioc.unesco.org/oceanteacher/resourcekit/index.htm with some good representative sample projects http://ioc.unesco.org/oceanteacher/Data/data.htm used for illustration. -It is generally recognized that the World Ocean Circulation Experiment (WOCE) was a success story from a perspective of coordinating data management and attaining scientific objectives. http://woce.nodc.noaa.gov/wdiu -Perhaps a more recent and local example of integration of data management within the context of a regional monitoring program would be the Atlantic Zone Monitoring Program (AZMP) http://www.meds-sdmm.dfo-mpo.gc.ca/zmp/main_zmp_e.html where the data management plan http://www.meds-sdmm.dfo-mpo.gc.ca/zmp/Documents/e_DataMgmtPlan.htm forms a key integral and ongoing activity within the program activities. IRIS (Tim Ahern) This is a very hard question to answer. Let me just say that IRIS already does, and has been doing, a very similar thing for nearly a decade. We adopted format standards early on and that has proven to be a great benefit for managing these data. We attempt to homogenize the data flow and formats before they enter (or as they enter) the DMC and therefore most data appears similar to us in terms of our automated system. -This is an incomplete answer but your question is difficult to fully answer. NOAA/NODC (Patrick Caldwell) Any well described format is fine. NOAA/PMEL, TAO Project (Paul Freitag) Our experience suggests that it will be important for the VENUS/NEPTUNE DMAS to allow user-defined subsets of the data to be extracted and downloaded. Even for users who want all of the data, there are often bandwidth limitations, particularly in the third world, which make the option to download a subset of the data an essential capability. -The design of our data delivery system includes data availability information in the user interface, which quickly shows the user what data are available, prior to the user-request. For example see http://www.pmel.noaa.gov/tao/data_deliv/

	-Many delivery systems require the user to make their request first, only to be informed that there are no data fitting their criteria, which can be frustrating, inefficient, and confusing. -Another suggestion is to make as much metadata as possible available from the Web or served with the data, as this will undoubtedly limit the number of questions you receive from users and make some responses as simple as providing a URL. We provide much of our metadata at http://www.pmel.noaa.gov/tao/proj_over/proj_over.html.
--	---

8.3 Questionnaire 2: for Organizations Dealing with Large Data Volumes

The main goal of Phase 2 of the VENUS/NEPTUNE DMAS Examination project was to look beyond V/N direct issues to provide information from other initiatives. Building on the experiences of other organizations that deal with large volumes of data will assist in the formulation of strategies that are known to be successful and efficient, and in the avoidance of those that have not proven to be effective.

To realize this objective, BCS identified agencies that deal with large data volumes, both ocean-related and non-ocean-related. We then designed and distributed Questionnaire 2 and sent it to selected contacts in these agencies, and also performed a number of phone interviews with other contacts. The content and results of this questionnaire are detailed below.

8.3.1 Questionnaire 2

Dear {*specific person(s)*|Sir(s)},

Barrodale Computing Services Ltd. (BCS) (www.barrodale.com), has been commissioned by the VENUS/NEPTUNE ocean observatory project offices to examine DMAS (Data Management and Archive Systems) issues associated with these planned observatories. (For references to these observatories, please see <http://www.VENUS.uvic.ca>, <http://www.NEPTUNECanada.ca>, and <http://www.NEPTUNE.washington.edu>).

In order to provide guidance and information to VENUS/NEPTUNE during the early stages of DMAS design, we are seeking information from other organizations (both ocean-related and non-ocean related) that have experience with systems sharing **any** or all of the following characteristics:

- serve as a repository for a large volume of scientific data and associated (complex) metadata,
- have a real time data acquisition and storage component,
- have some degree of post-processing and data product creation,
- have facilities to allow access to the data soon after they are collected,

- have a wide range of users and user requirements (e.g., data formats, client applications, etc.)

Our goal is to help devise strategies that have proven to be successful and efficient, to learn from the lessons of other organizations, and to learn about future challenges and plans.

Please answer as many questions as your schedule allows; even obtaining a few answers from you would be appreciated.

Your feedback will assist the VENUS/NEPTUNE team to address the data management requirements for the respective systems. Any additional insights and recommendations would be most helpful, as would any documents, or links to documents, that would describe current data holdings and data management procedures. As this information is being collected for a report due in a matter of weeks, your timely response would be appreciated. Please feel free to contact me directly by phone (250-472-4370) or email (<mailto:mike@barrodale.com>) for any further clarification or discussions.

TYPES OF DATA

Please describe what forms of data are collected by your archive (e.g., satellite images, sound speed profiles, time series of measurement values from a particular type of instrument, etc.)

DATA SOURCES

Please comment on the mechanisms by which data enter your site.

- Do the data come from an intermediate repository or do they come “from the field” (some measurement device or instrument (e.g., seismometer)), or is it some combination of the two?
- Do the data arrive in a streaming, continuous fashion or are they batched up (e.g., via scheduled transmissions or expeditions)?
- If the data come from another site is it “pulled” from the site (i.e., initiated by you) or pushed by the site (i.e., initiated by the other site)?
- How do the data arrive (via the Internet, an intranet, physical media delivery, etc.)?
- What media formats (for physical delivery) or protocols (e.g., FTP, HTTP, etc.) are used for delivery to your site?

- In what format(s) do the data arrive (e.g., native instrument format, netCDF, structured text, etc.)?
- If the transmissions are scheduled, how often do they occur (e.g., hourly, daily, etc.)?
- Are any (human) administrative procedures involved in these transfers?
- What is the level of effort involved in managing each of these data sources?
- What is the approximate volume of data, per unit time, for each form of data (as defined at the top of this section) entering your site?
- Do you have any other comments about how data arrive at your site?

DATA DESTINATIONS

Please comment on what happens to the data once they arrive at your site.

- Are the data stored in your repository?
- If they are stored in your repository, please describe briefly the physical nature of your repository (e.g., is it based on tapes, disk arrays, optical disks, CD/DVD's, relational databases, data warehouse systems, etc.?).
- Are the data transmitted elsewhere through real time distribution (e.g., GTS)? Please describe.
- Are the data sent to other repositories? Please list them.
- Is a long-term (permanent) archive of your data kept offsite (e.g., at some central repository, National Archives, etc.)?
- What is the level of effort involved in managing the storage of each form of data (as defined earlier)?
- Do you have any other comments about the destination of data?

QUALITY CONTROL AND POST-PROCESSING

- What forms of post-processing are performed (manual, automatic, both)?

- What automatic quality control is applied?
- Do quality assurance standards or tests exist (e.g., is there a list of tests performed on the data in real time?) If there are, do you record the outcome of the tests (e.g., flag the data with a quality stamp that accompanies the metadata)?
- Is any data assembly (e.g., aggregation and buffering) performed? Please give examples.
- Is any product generation (e.g., simulation model creation, weather forecasts, etc.) performed? Please give examples.
- How is metadata created (e.g., is it an automatic process or a manual one)?
- What are the time lags between arrival of data, post-processing, distribution, and archive?
- Are copies of the data kept for each phase of the post-processing (e.g., data as received, data as post-processed, etc.)?
- What is the level of effort involved in performing each type of quality control / post-processing?

METADATA MANAGEMENT

- How are metadata managed (kept up to date, kept synchronized with the data, etc.) throughout the evolution of the data (from initial receipt through quality control, data assembly, product generation, etc.)?
- Are the data's metadata stored with the actual data?
- through encapsulated files (e.g., netCDF, XML)? Please specify which type.
- in an object-relational database? Please specify which brand.
- in an object-oriented database? Please specify which brand.
- What metadata standards and profiles are used?

DATA STORAGE

- Are the data stored in a
- relational database system? Please specify the brand (e.g., Oracle, DB2,

Informix, Access)

- an object-relational database system? Please specify the brand (e.g., DB2, Informix, Postgres)
- an object oriented database system? Please specify the brand (e.g., ozone, GemStone)
- a dedicated data warehousing system (e.g., Sybase IQ, IBM Red Brick)
- Are the data stored in an electronic filesystem? What formats are used (e.g., netCDF, structured text, etc.)?
- What are the reasons for using this form of storage (e.g., nature of the data, nature of data access, institutional standards, other technological reasons, etc.)?
- Is any use made of hierarchical storage management techniques or a storage resource broker?
- If multiple levels of storage devices are used (e.g., disk for newer or actively used data, tape for older or less frequently used data), what factors are used to assign the data to the various levels of storage device?
- Are multiple copies of the same data kept (as insurance against data loss, media deterioration, etc.)?
- What is the approximate volume of data and metadata stored at your site?
- Data
- Metadata
- Do you have an estimate of the storage cost per gigabyte?

DATA AVAILABILITY

- How does someone, either internal or external to your organization, perform “data discovery?”
- Is the metadata catalog available for searching via the Web?
- Once someone has decided what data they want, how do they proceed to extract it (e.g., through manual request, OPeNDAP server request, FTP, etc.)?
- What file formats are used in presenting/uploading the data?

- Do you offer any facilities for “data mining?”
- Is it possible for a scientist to issue a query based not only on the metadata but on some property of the actual data (not stored in the metadata)?
- Can you provide an example of such a query that might be performed on your data?

REAL TIME ASPECTS

- Are the data collected in real time?
- Are the data disseminated in (near) real time?
- What is the time lag between data collection, dissemination, availability for access, and archive?

OTHER COMMENTS

- VENUS and NEPTUNE will be collecting and disseminating a wide range of data types at differing sample rates and intervals from a common geographical region, resulting in a real time heterogeneous data set. Are there any lessons learned from your experience that could assist in the development of the VENUS/NEPTUNE DMAS, or that could assist in avoiding potential pitfalls in the design of the DMAS? Any information you feel to be relevant would be appreciated.

8.3.2 Questionnaire 2 Recipients and Responders

Recipient	Responder
Mr. Jim Lyons Canadian National Data Centre for Earthquake Seismology and Nuclear Explosion Monitoring	Jim Lyons

Dr Catherine Maillard Coriolis Project (France) (www.coriolis.eu.org) ESONET, IFREMER/SISMER ODC	
Mr. Ron McLaren Environment Canada Weather Buoy System (ECWBS)	Ron McLaren
Mr. Donald Denbo EPIC (PMEL)	
Mr. Ryan Hofschneider FNMOC	Ryan Hofschneider*
Mr. Steve Hankin IOOS	Steve Hankin*
Mr. Colin Leavett-Brown UVic Atlas Project	
Dr. Andrew Westwell-Roper MacDonald Dettwiler and Associates (MDA)	Andrew Westwell-Roper
Dr. Jorge Vazquez NASA EOS – EOSDIS Physical Oceanography DAAC (PO.DAAC) Ms. Bonnie Seaton Data and Operations System (EDOS)	Jorge Vazquez*
Dr. Bryan Lawrence (NERC DataGrid Principal Investigator; Head NCAS/British Atmospheric Data Centre) NERC Datagrid (BODC)	
Dr. Paul Freitag TAO (PMEL)	Paul Freitag
Mr. Phil Sharfstein USGODAE Server Project – Data Manager USGODAE	Phil Sharfstein

*Also provided phone interview; see Appendix C.

8.4 Questionnaire 2 Responses

Responses were received from 7 of the 14 individual recipients of Questionnaire 2. The answers provided by the responders are given in the table below.

Question	Response(s)
TYPES OF DATA	
Please describe what forms of data are collected by your archive (e.g., satellite images, sound speed profiles, time series of measurement values from a particular type of instrument, etc.)	ECWBS (Ron McLaren) Data measurements from moored and drifting buoys (wind speed and direction, SST, air temperature, atmospheric pressure, wave spectrum, wave period, max wave height and significant wave height). FNMOOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Optical remote sensing satellite raw imagery. MDA/RADARSAT (Andrew Westwell-Roper) Synthetic Aperture Radar (SAR) remote sensing satellite raw imagery. NRCAN (Jim Lyons) The primary (raw) data are continuous time series received in real time representing seismic ground motion (velocity and acceleration), though we also capture barometric pressure, low-sample rate weather data, and recently magneto-telluric (MT) data as well. PO.DAAC (Jorge Vazquez) We deal basically with three satellite data sets (although this is expanding). They are sea surface temperature, sea surface height, and sea surface winds. The data are in different formats ranging from binary flat formats to HDF and ASCII. TAO/PMEL (Paul Freitag) TAO collects time series measurements of: wind speed and direction, air temperature, relative humidity, short wave radiation (limited), rain (limited), sea surface temperature (SST), discrete subsurface temperatures to 500 m, salinity (limited), and currents (limited). USGODAE (Phil Sharfstein) Level 2 satellite SST and SSH, atmospheric and ocean model analysis and forecast fields, <i>in situ</i> meteorological data, <i>in situ</i> sea surface and ocean profile data, high-resolution satellite sst/flux/sea ice products, <i>in situ</i> ship track, sea wave height, tides.

DATA SOURCES	
Please comment on the mechanisms by which data enter your site.	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) FTP or tape from CCRS Ground Stations in Prince Albert, SK and Gatineau, PQ. DVD from USGS. MDA/RADARSAT (Andrew Westwell-Roper) Prince Albert: Antenna, then WAN to Gatineau. Gatineau: Antenna (through LAN). Also tape from network stations worldwide. NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No reply. USGODAE (Phil Sharfstein) No reply.
Do the data come from an intermediate repository or do they come “from the field” (some measurement device or instrument (e.g., seismometer)), or is it some combination of the two?	ECWBS (Ron McLaren) Field. FNMOC (Ryan Hofschneider) A combination of the two; previously calculated forecasts (local) are used to seed the forecast model as well as observations from the field. MDA/LANDSAT (Andrew Westwell-Roper) Intermediate repository (from CCRS or USGS Raw Archive). MDA/RADARSAT (Andrew Westwell-Roper) Intermediate repository (from CCRS Raw Archive). NRCAN (Jim Lyons) Our CNSN data come directly from field sensors in 6-second serial data packets using an in-house packet format and reception protocol. The data samples are compressed using a lossless second-difference compression scheme that has become an international standard. Multiplexing/demultiplexing of data streams require in-house devices. PO.DAAC (Jorge Vazquez) Data usually come directly from centers which have applied algorithms to initially process the satellite data from the satellite so we can distribute it in gridded formats. We receive most of our data electronically via FTP and also operationally. TAO/PMEL (Paul Freitag) Our site receives data from the field in (near) real time via

	satellite telemetry, with Service Argos as an intermediary (hereafter referred to as “relayed”, “telemetered” or “(near) real time”). High resolution data enter our site when instrument platforms are serviced or replaced in the field and the onboard recording systems are dumped (hereafter referred to as “delayed”, “high-resolution”, or “recovered”). USGODAE (Phil Sharfstein) Intermediate repository.
Do the data arrive in a streaming, continuous fashion or are they batched up (e.g., via scheduled transmissions or expeditions)?	ECWBS (Ron McLaren) Transmitted hourly. FNMOC (Ryan Hofschneider) They arrive asynchronously. MDA/LANDSAT (Andrew Westwell-Roper) Batched. MDA/RADARSAT (Andrew Westwell-Roper) Batched. NRCAN (Jim Lyons) Continuous stream of packets. PO.DAAC (Jorge Vazquez) We have both. We receive data operationally and in near real time from satellite acquisition. TAO/PMEL (Paul Freitag) Data are batched. Platforms transmit data continuously during several windows each day, but relay is contingent on satellite overpass. All relayed data are delivered once each day by the intermediary, Service Argos. USGODAE (Phil Sharfstein) Not streaming; scheduled transmissions and data flow transmissions.
If the data come from another site is it “pulled” from the site (i.e., initiated by you) or pushed by the site (i.e., initiated by the other site)?	ECWBS (Ron McLaren) Pushed by the site. FNMOC (Ryan Hofschneider) Different data sources arrive via different mechanisms. Some are pushed to us, while others we go and fetch from their remote sites. MDA/LANDSAT (Andrew Westwell-Roper) No reply. MDA/RADARSAT (Andrew Westwell-Roper) No reply. NRCAN (Jim Lyons) The normal mode is a push from each sensor site once a packet is ready to send. However, our protocol allows for retransmission requests of single or blocks of missed packets. - Data that we acquire from the POLARIS seismic/MT network are “pulled” in their native sensor format and converted to our packet format “on-the-fly”.

	PO.DAAC (Jorge Vazquez) We both push and pull. TAO/PMEL (Paul Freitag) Both modes of delivery are available to us. At present, the default is to have the data “pushed” by the intermediary. USGODAE (Phil Sharfstein) Both.
How do the data arrive (via the Internet, an intranet, physical media delivery, etc.)?	ECWBS (Ron McLaren) Via meteorological circuits (combinations of landlines and satellite communication links). FNMOC (Ryan Hofschneider) The Internet, for unclassified data. MDA/LANDSAT (Andrew Westwell-Roper) No reply. MDA/RADARSAT (Andrew Westwell-Roper) No reply. NRCAN (Jim Lyons) The backbone national network uses VSAT technology where the field data are aggregated at Telesat’s Toronto Teleport and backhauled to our Sidney, BC and Ottawa data centres via private T1 links (Anikom TS to Sidney, Bell landline to Ottawa). Under cooperative agreements, some stations make use of private telecommunications facilities of the large power utilities. A few use the Internet via serial-to-IP converters at the site. - Recently we have started moving towards IP with multicasting to allow simultaneous reception at primary and backup acquisition systems at both national offices. PO.DAAC (Jorge Vazquez) Via the Internet. TAO/PMEL (Paul Freitag) Internet. USGODAE (Phil Sharfstein) Internet and intranet.
What media formats (for physical delivery) or protocols (e.g., FTP, HTTP, etc.) are used for delivery to your site?	ECWBS (Ron McLaren) Dedicated circuits as above. FNMOC (Ryan Hofschneider) Both FTP and HTTP protocols are used; FTP is being phased out. MDA/LANDSAT (Andrew Westwell-Roper) No reply. MDA/RADARSAT (Andrew Westwell-Roper) No reply. NRCAN (Jim Lyons) We use DLT and Exabyte tape for output, plus FTP and HTTP for both input and output. PO.DAAC (Jorge Vazquez)

	FTP. TAO/PMEL (Paul Freitag) SMTP (primary), Telnet (backup). USGODAE (Phil Sharfstein) FTP, HTTP, OPeNDAP, Web Service.
In what format(s) do the data arrive (e.g., native instrument format, netCDF, structured text, etc.)?	ECWBS (Ron McLaren) Formatted messages containing sensor data. FNMOC (Ryan Hofschneider) Our observation data typically arrive as text-based messages, according to some World Meteorological Organization (WMO) or DoD blessed standard. Gridded forecasts from other centers might arrive in GRIB or just raw arrays dumped to files. MDA/LANDSAT (Andrew Westwell-Roper) No reply. MDA/RADARSAT (Andrew Westwell-Roper) No reply. NRCAN (Jim Lyons) Native instrument format, which we cast into files in CA (Canadian Archive) format. PO.DAAC (Jorge Vazquez) ASCII, binary, HDF, netCDF. TAO/PMEL (Paul Freitag) Structured text for (near) real time telemetered data, native instrument format from onboard recording systems (high resolution data). USGODAE (Phil Sharfstein) IEEE Binary, C3GRID, netCDF, text, Fortran binary, HDF-EOS, FGGE, GRIB.
If the transmissions are scheduled, how often do they occur (e.g., hourly, daily, etc.)?	ECWBS (Ron McLaren) Hourly. FNMOC (Ryan Hofschneider) Many stations transmit observations hourly, or at the very least multiple times a day. MDA/LANDSAT (Andrew Westwell-Roper) As required (order driven); 10-20 transmissions per day. MDA/RADARSAT (Andrew Westwell-Roper) As required (order driven); 15-25 transmissions per day. NRCAN (Jim Lyons) A few sites lacking continuous communications employ dial-up access for “events” (time segments of interest). PO.DAAC (Jorge Vazquez) We are now in near real time; thus, we do have data sets that are transmitted hourly. TAO/PMEL (Paul Freitag) Daily. USGODAE (Phil Sharfstein)

	Depending on the source, hourly, 6x daily, 4x daily, 1x daily.
Are any (human) administrative procedures involved in these transfers?	ECWBS (Ron McLaren) No. Data are sent and processed automatically. Human intervention is required when faults are detected. FNMOC (Ryan Hofschneider) All automatic once an initial process is set up. MDA/LANDSAT (Andrew Westwell-Roper) Yes. MDA/RADARSAT (Andrew Westwell-Roper) Yes. NRCAN (Jim Lyons) Dial-up is triggered automatically by our data access s/w. PO.DAAC (Jorge Vazquez) Most of our stuff is automated but we do have human interactions to examine error protocols. TAO/PMEL (Paul Freitag) Only in cases of system problems (telemetered data). Extensively for recovered instrument data (high resolution). USGODAE (Phil Sharfstein) No, all automated, unless something fails.
What is the level of effort involved in managing each of these data sources?	ECWBS (Ron McLaren) Sensor platform maintenance is the prime activity with a high level of effort. Maintaining the data flow once systems are operating correctly is fairly low effort. FNMOC (Ryan Hofschneider) Minimal, depending on the mechanism being used to transfer; probably half an hour's worth of work initially. MDA/LANDSAT (Andrew Westwell-Roper) No reply. MDA/RADARSAT (Andrew Westwell-Roper) No reply. NRCAN (Jim Lyons) Most continuous access runs pretty seamlessly, but the dial-ups are innately more prone to problems, modem hangs, etc. PO.DAAC (Jorge Vazquez) Hard to say. TAO/PMEL (Paul Freitag) Low for telemetered data, moderate to high for delayed-mode. USGODAE (Phil Sharfstein) 1/2 to 3/4 of a full time position.
What is the approximate volume of data, per unit time, for each form of data (as defined at the top of this section)	ECWBS (Ron McLaren) 25 kilobits per hour. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) FTP speed 10 Mbps.

entering your site?	MDA/RADARSAT (Andrew Westwell-Roper) LAN speed 10 Mbps. NRCAN (Jim Lyons) We currently receive and process ~4.3 gigabytes of continuous data and only a few kilobytes of dial-up data per day. PO.DAAC (Jorge Vazquez) I would really need more time to be able to answer this question but we are dealing with 10s of gigabytes/day. TAO/PMEL (Paul Freitag) Telemetered data: 90-120 kilobytes per day. Delayed data: 2-5 megabytes per day. USGODAE (Phil Sharfstein) I actually have this information on a pretty big chart on my work computer, but it would take a few hours to recreate here at home.
Do you have any other comments about how data arrive at your site?	ECWBS (Ron McLaren) We use the GOES satellite (moored buoys) and the ARGOS system (drifting buoys) for gathering ocean data. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Data often arrive zipped. MDA/RADARSAT (Andrew Westwell-Roper) No reply. NRCAN (Jim Lyons) Our packet protocol requires the real time acquisition systems to be able to cope with missing data, data arriving late, out of sequence, and multiple times (possibly not asked for). This has stood us in good stead wrt maximal data recovery. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No reply. USGODAE (Phil Sharfstein) No reply.
DATA DESTINATIONS	
Please comment on what happens to the data once they arrive at your site.	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Unzipped, FTP'd to processing system, ingested. (Processing system is MacDonald Dettwiler's Product Generation System, "PGS"). MDA/RADARSAT (Andrew Westwell-Roper)

	Sent to processing system and ingested. (Processing system is MacDonald Dettwiler's Product Generation System, "PGS".) NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) Data are ingested into graphical interfaces for distribution, placed on our FTP site and also archived. TAO/PMEL (Paul Freitag) No reply. USGODAE (Phil Sharfstein) No reply.
Are the data stored in your repository?	ECWBS (Ron McLaren) Temporarily, then archived in National Archive. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Yes. MDA/RADARSAT (Andrew Westwell-Roper) No. NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) Yes. TAO/PMEL (Paul Freitag) Yes. USGODAE (Phil Sharfstein) Yes.
If they are stored in your repository, please describe briefly the physical nature of your repository (e.g., is it based on tapes, disk arrays, optical disks, CD/DVD's, relational databases, data warehouse systems, etc.?).	ECWBS (Ron McLaren) Disk. FNMOC (Ryan Hofschneider) Disk arrays and databases. MDA/LANDSAT (Andrew Westwell-Roper) No reply. MDA/RADARSAT (Andrew Westwell-Roper) No reply. NRCAN (Jim Lyons) For fault tolerance, the data always reside in more than one place at a time. The acquisition systems each maintain a week's worth of data in local storage. A few minutes after real time, a merged copy of the data are copied to our mass store, currently a robotic DLT tape library. After ~ 6.5 hours, when all retx activity would be done and the data files maximally complete, a final copy is transferred to the mass store. PO.DAAC (Jorge Vazquez) All. TAO/PMEL (Paul Freitag)

	Primary storage is disk, data reside in filesystem files and in relational databases. USGODAE (Phil Sharfstein) Disk arrays, RDB.
Are the data transmitted elsewhere through real time distribution (e.g., GTS)? Please describe.	ECWBS (Ron McLaren) Yes. Raw sensor data are QC'd then encoded in WMO message formats and distributed on the GTS. FNMOC (Ryan Hofschneider) Yes, via a Web service. MDA/LANDSAT (Andrew Westwell-Roper) No. (Ordered data are couriered to customer on CD.) MDA/RADARSAT (Andrew Westwell-Roper) 50% of data are delivered to customer by FTP in near real time. NRCAN (Jim Lyons) We have installed one of our acquisition systems at Carleton U to provide them a real time feed to selected Ontario stations over the Internet. We also ship about 165 gigabytes/day of primary seismic, infrasound, and hydroacoustic data to the CTBT International Data Centre in Vienna over a private T1 link after conversion to the CTBT CD protocol/format. PO.DAAC (Jorge Vazquez) Yes, we do have operational customers. TAO/PMEL (Paul Freitag) Data are distributed real time on GTS via Service Argos. USGODAE (Phil Sharfstein) Yes, transmission through Unidata's Local Data Manager (LDM).
Are the data sent to other repositories? Please list them.	ECWBS (Ron McLaren) National Meteorological Service of Canada archives in Downsview, Ontario and MEDS archive, Ottawa. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) No. MDA/RADARSAT (Andrew Westwell-Roper) No. NRCAN (Jim Lyons) Subsets of our data are sent to the IRIS DMC in Washington State (currently via FTP but soon in real time) and to Blacknest in the UK. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) NODC. Weekly digests are sent to National Weather Service/National Centers for Environmental Prediction. USGODAE (Phil Sharfstein)

	Yes, for several of the data sets: NODC, NDBC, Navy Archive in Mississippi (I can't remember the name).
Is a long-term (permanent) archive of your data kept offsite (e.g., at some central repository, National Archives, etc.)?	ECWBS (Ron McLaren) As above. FNMOC (Ryan Hofschneider) Yes. MDA/LANDSAT (Andrew Westwell-Roper) No (stored onsite). MDA/RADARSAT (Andrew Westwell-Roper) No. NRCAN (Jim Lyons) We maintain offsite backups of the complete archive, and are working towards mirrored online archives at the east and west data centres. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Not at present. USGODAE (Phil Sharfstein) Yes, for some of the data sets.
What is the level of effort involved in managing the storage of each form of data (as defined earlier)?	ECWBS (Ron McLaren) Moderate and ongoing. The current MSC marine data archive is undergoing updating and system revisions. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) No reply. MDA/RADARSAT (Andrew Westwell-Roper) No reply. NRCAN (Jim Lyons) We have experienced a number of problems with the DLT tape library/Veritas HSM s/w that have led us to move towards an online disk archive using in-house access s/w. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Probably less than two months full time involved in storage management and data backup. USGODAE (Phil Sharfstein) Up to 1/4 time.
Do you have any other comments about the destination of data?	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Order driven.

	MDA/RADARSAT (Andrew Westwell-Roper) Order driven. NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No reply. USGODAE (Phil Sharfstein) No reply.
QUALITY CONTROL AND POST-PROCESSING	
What forms of post-processing are performed (manual, automatic, both)?	ECWBS (Ron McLaren) Both. Manual input from operational weather forecasters if data look suspect, as well as automated checks. FNMOC (Ryan Hofschneider) Manual. MDA/LANDSAT (Andrew Westwell-Roper) Radiometric and geometric image correction. Generally automatic, with manual intervention for a) ground control point marking and b) quality control. MDA/RADARSAT (Andrew Westwell-Roper) Radiometric and geometric image correction. Generally automatic, with manual intervention for a) ground control point marking and b) quality control. NRCAN (Jim Lyons) The input data streams are run through a seismic event detector, which feeds an automatic earthquake locator, which in turn feeds an automatic earthquake alert system that issues emails, faxes, pages, etc. to, e.g., tell CN Rail to stop or slow trains in a certain region within minutes of the event. Further post-processing is done to scan for and refine potential events in preparation for human analyst review. PO.DAAC (Jorge Vazquez) A wide variety. This is very data specific and ranges from manually looking at images to scripts which check for correct file sizes, etc. TAO/PMEL (Paul Freitag) Both. USGODAE (Phil Sharfstein) Automatic.
What automatic quality control is applied?	ECWBS (Ron McLaren) Primarily range checks to determine if data are within reasonable limits. FNMOC (Ryan Hofschneider)

	No reply. MDA/LANDSAT (Andrew Westwell-Roper) Automatic check of and report on radiometric calibration. MDA/RADARSAT (Andrew Westwell-Roper) Automatic check of and report on radiometric calibration. NRCAN (Jim Lyons) Power spectral density plots of each station are generated based on 8 15-minute samples a day and used for human review of potential station problems such as undue noise sources. Also, station timing is checked by generating plots of sample arrival time vs. data time and issuing an alert if a statistical drift is present. PO.DAAC (Jorge Vazquez) A variety of scripts that check for data transfers, sizes, etc. TAO/PMEL (Paul Freitag) Flagging of data that fall outside broad error specifications. USGODAE (Phil Sharfstein) File format check only.
Do quality assurance standards or tests exist (e.g., is there a list of tests performed on the data in real time?) If there are, do you record the outcome of the tests (e.g., flag the data with a quality stamp that accompanies the metadata)?	ECWBS (Ron McLaren) Yes. Suspect data are flagged. FNMOCC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Radiometric calibration and geometric accuracy (best fit against ground control points). MDA/RADARSAT (Andrew Westwell-Roper) Radiometric calibration and geometric accuracy (best fit against e.g., coastline). NRCAN (Jim Lyons) This is an area of current research within the seismological community. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Yes, data remaining after the automatic error check are checked against a narrower range of error specifications, and those that fall outside this range generate an error alert message. Questionable data are not automatically removed at this point. Rather, for each error alert, the suspect data are checked for validity by experienced data analysts. USGODAE (Phil Sharfstein) No.
Is any data assembly (e.g., aggregation and buffering) performed? Please give examples.	ECWBS (Ron McLaren) Data are formatted daily into an archive which is sent to National Archive in Downsview. FNMOCC (Ryan Hofschneider)

	No reply. MDA/LANDSAT (Andrew Westwell-Roper) Order driven value added work, e.g., fusion of multispectral with pan-chromatic imagery, mosaics. MDA/RADARSAT (Andrew Westwell-Roper) Order driven value added work, e.g., multi-temporal fusion, mosaics. NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Data are assembled into files based on data type (e.g., meteorological data, subsurface data, etc.) and sampling schemes. USGODAE (Phil Sharfstein) Yes, <i>in situ</i> files aggregated by device, time, and geographic region, individual model fields aggregated by time.
Is any product generation (e.g., simulation model creation, weather forecasts, etc.) performed? Please give examples	ECWBS (Ron McLaren) Yes. Buoy data are used in numerous forecast models. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Generate output imagery, fused imagery and mosaics. MDA/RADARSAT (Andrew Westwell-Roper) Generate output imagery, fused imagery and mosaics, plus custom prediction models, e.g., for oil slicks, floods, rice crops, pests (locusts). NRCAN (Jim Lyons) Derived parameters constitute picks of the onset timing, instantaneous period, and amplitude of seismic phases appearing on the waveforms. These are fed to an earthquake location program that generates the location (lat, lon, depth), magnitude of the earthquake, and error residuals. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No reply. USGODAE (Phil Sharfstein) No.
How is metadata created (e.g., is it an automatic process or a manual one)?	ECWBS (Ron McLaren) The capture of the metadata is a weak point currently but automated systems are being developed to capture this information. FNMOC (Ryan Hofschneider) No reply.

	MDA/LANDSAT (Andrew Westwell-Roper) Created by CCRS at ground stations. We add metadata re. final processed image. MDA/RADARSAT (Andrew Westwell-Roper) Created by CCRS at ground stations. We add metadata re. final processed image. NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) This is also very data specific. TAO/PMEL (Paul Freitag) No reply. USGODAE (Phil Sharfstein) Both automatic and manual.
What are the time lags between arrival of data, post-processing, distribution, and archive?	ECWBS (Ron McLaren) Buoy data are generally received, decoded and distributed in final format on the GTS within 5-10 minutes after transmission from the buoy. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) 2 days. MDA/RADARSAT (Andrew Westwell-Roper) Near real time (processing time on order of hours). NRCAN (Jim Lyons) Data arriving more than 2 minutes behind real time are considered too late and dropped from the automatic processing pipeline. Clearly, the timeliness of earthquake alerts is paramount. However, every attempt is made to ensure the archive is maximally complete (depends on size of outstation buffer; currently ≤ 6 hr). PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) (Near) real time data are available for distribution via world wide Web within 4hours of data arrival at PMEL/TAO Project. Post processing, quality control and distribution of high-resolution data lag weeks to months after instrument recovery. USGODAE (Phil Sharfstein) Aim for less than 30 minutes lag time.
Are copies of the data kept for each phase of the post-processing (e.g., data as received, data as post-processed, etc.)?	ECWBS (Ron McLaren) Yes. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Retain copy of raw data.

	MDA/RADARSAT (Andrew Westwell-Roper) No. NRCAN (Jim Lyons) No, as they can be reprocessed if needed, which is rare for us. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Copies of data at each phase are maintained. USGODAE (Phil Sharfstein) Yes, depending on the data set.
What is the level of effort involved in performing each type of quality control / post-processing?	ECWBS (Ron McLaren) Mainly automated. Moderate level of effort. FNMOCC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) No reply. MDA/RADARSAT (Andrew Westwell-Roper) No reply. NRCAN (Jim Lyons) QC of station data are shared among the data centre ops and technical field staff, with the balance swinging more towards the former as new tools are developed. A few minutes are all that's required to scan the list of PSD plots for each station. Analyst review and interactive analysis/relocation of events requires roughly two people at each data centre. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Typically, less than one hour of human intervention per day is involved for (near) real time post-processing and quality control. Post-processing and quality control of the delayed mode high resolution data are supported by 1.5 to 3 full time employees. USGODAE (Phil Sharfstein) All are automated.
METADATA MANAGEMENT	
How are metadata managed (kept up to date, kept synchronized with the data, etc.) throughout the evolution of the data (from initial receipt through quality control, data	ECWBS (Ron McLaren) Some metadata such as buoy inspection and sensor calibration information are on paper or electronic media. Real time buoy sensor configurations are maintained within an automated computer system that processes buoy data known as WBS (Weather Buoy System). FNMOCC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper)

assembly, product generation, etc.)?	CCRS and USGS store metadata on Web sites. We download orbit information automatically daily from CCRS, and radiometric calibration files monthly from USGS. MDA/RADARSAT (Andrew Westwell-Roper) We download orbit information automatically daily from CSA. Input data are order driven, so much of the metadata (e.g., geographic location) is known in advance through order planning process. NRCAN (Jim Lyons) Date/time is used as a primary key for many parameters (e.g., earthquake event_id, phase arrival_id, etc.). We have made modifications to an international standard database schema (CSS-3.0). PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No reply. USGODAE (Phil Sharfstein) All metadata kept with the data files, or added during assembly.
Are the data's metadata stored with the actual data? (a) through encapsulated files (e.g., netCDF, XML)? Please specify which type. (b) in an object-relational database? Please specify which brand. (c) in an object-oriented database? Please specify which brand.	ECWBS (Ron McLaren) (a) Some configuration metadata are stored with the data; however, most is in other various sources. System is undergoing revision as noted above. (b) Proprietary software called Alpha Numeric Manager (AM). FNMOG (Ryan Hofschneider) In some cases metadata are tacked on as a fixed-length header to the file format; in other formats like netCDF and XML we use their respective tagging facilities. MDA/LANDSAT (Andrew Westwell-Roper) Yes. CCRS sends input data in Framed Raw Enhanced Data (FRED) format, which has some metadata interleaved. (Transcription is by CCRS GTS.) USGS sends input data in Hierarchical Data Format (HDF), which contains an MTL metadata file. Output data are in various formats: LGSOWG: includes leader and volume descriptor files containing metadata. EOSAT Fast Format: includes header file with metadata. GeoTIFF: includes interleaved geographic metadata. HDF: includes MTL metadata file. MDA/RADARSAT (Andrew Westwell-Roper) Yes. CCRS sends input data in FRED format, transcribed by MacDonald Dettwiler's Direct Archiver System (DAS). Output data are in various formats: GeoTIFF (Mr.SID compressed), which has geographic metadata interleaved. CEOS Format: includes leader file with metadata (similar to

	LGSWOG) For RADARSAT II, default format will be GeoTIFF with (separate) XML metadata file. NRCAN (Jim Lyons) No. Volume of raw waveforms is too large. (b) We use the CA-OpenINGRES RDBMS. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No reply. USGODAE (Phil Sharfstein) (a) Yes, through netCDF, GRIB.
What metadata standards and profiles are used?	ECWBS (Ron McLaren) There are International WMO suggested metadata formats which we will work to adhere to. FNMO (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) USGS and CCRS follow FGDC metadata standard. MDA/RADARSAT (Andrew Westwell-Roper) CCRS follows FGDC metadata standard. NRCAN (Jim Lyons) This is currently under investigation/development at the departmental level. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No reply. USGODAE (Phil Sharfstein) No reply.
DATA STORAGE	
Are the data stored in a (a) relational database system? Please specify the brand (e.g., Oracle, DB2, Informix, Access) (b) an object-relational database system? Please specify the brand (e.g., DB2, Informix, Postgres)	ECWBS (Ron McLaren) I do not work with this portion of the network, so do not have the answers to this section. FNMO (Ryan Hofschneider) (b) Informix and PostgreSQL MDA/LANDSAT (Andrew Westwell-Roper) (a) Orbit information and ground control points stored in Empress Database. MDA/RADARSAT (Andrew Westwell-Roper) (a) Orbit information and ground control points stored in Empress Database. NRCAN (Jim Lyons) (a) All derived parameters and locations, station site info,

(c) an object oriented database system? Please specify the brand (e.g., ozone, GemStone) (d) a dedicated data warehousing system (e.g., Sybase IQ, IBM Red Brick)	telecommunications links, and trouble-shooting are stored in a CA-OpenINGRES RDBMS using our GSC-2.0 db schema. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) (a) mySQL USGODAE (Phil Sharfstein) (a) mySQL.
Are the data stored in an electronic filesystem? What formats are used (e.g., netCDF, structured text, etc.)?	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) We actually store the data in a non-electronic filesystem (a really big cabinet next to my desk). :) For our gridded forecast data, when they are stored as files on a disk we store them as GRIB or netCDF or a raw array serialized to file. Observation data are stored as either structured text messages or as XML. MDA/LANDSAT (Andrew Westwell-Roper) Raw data are stored in FRED format files. MDA/RADARSAT (Andrew Westwell-Roper) N/A. NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) At different stages of the post processing and distribution, data are stored in structured text files and netCDF files. USGODAE (Phil Sharfstein) NetCDF.
What are the reasons for using this form of storage (e.g., nature of the data, nature of data access, institutional standards, other technological reasons, etc.)?	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) GRIB is a WMO standard and is well supported within DoD; netCDF because it is self-describing and easier to deal with than GRIB; raw arrays serialized to file for historical reasons. MDA/LANDSAT (Andrew Westwell-Roper) Compatibility with PGS processing system. MDA/RADARSAT (Andrew Westwell-Roper) Compatibility with PGS processing system. NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Text files: compatibility with legacy systems.

	NetCDF files: broad applicability in user community. USGODAE (Phil Sharfstein) Nature of data, and nature of data access, attempt to keep all files in the original format delivered.
Is any use made of hierarchical storage management techniques or a storage resource broker?	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Nothing used beyond hierarchical directory filesystems, and database as specified. MDA/RADARSAT (Andrew Westwell-Roper) Nothing used beyond hierarchical directory filesystems, and database as specified. NRCAN (Jim Lyons) As stated, the raw waveforms reside in a robotic DLT tape library that employs Veritas HSM s/w. However, we have found that for our application (fixed size in time data files with time-synchronized and easily computable filenames with mostly sequential writing in chronological order), we can and want to control the system and often end up “fighting” with the Veritas s/w. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No. USGODAE (Phil Sharfstein) No.
If multiple levels of storage devices are used (e.g., disk for newer or actively used data, tape for older or less frequently used data), what factors are used to assign the data to the various levels of storage device?	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) FTP'd data are archived on DVD. (FTP is preferred transmission method; limited only by available bandwidth.) CCRS couriered data arrive on tape. USGS couriered data arrive on DVD. Output data are always delivered on CD. MDA/RADARSAT (Andrew Westwell-Roper) Not applicable (data arrive through WAN or LAN, and are not stored). NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No reply.

	USGODAE (Phil Sharfstein) No reply.
Are multiple copies of the same data kept (as insurance against data loss, media deterioration, etc.)?	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) No. MDA/RADARSAT (Andrew Westwell-Roper) No. NRCAN (Jim Lyons) Yes. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Yes. USGODAE (Phil Sharfstein) Yes.
What is the approximate volume of data and metadata stored at your site? (a) Data (b) Metadata	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Data: 9.6 terabytes; metadata: 20 gigabytes. MDA/RADARSAT (Andrew Westwell-Roper) N/A. NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Less than 50 gigabytes aggregate. USGODAE (Phil Sharfstein) (a) 1.5 terabytes.
Do you have an estimate of the storage cost per gigabyte?	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) No reply. MDA/RADARSAT (Andrew Westwell-Roper) No reply. NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) No reply.

	TAO/PMEL (Paul Freitag) No. USGODAE (Phil Sharfstein) No reply.
DATA AVAILABILITY	
How does someone, either internal or external to your organization, perform “data discovery?” (a) Is the metadata catalog available for searching via the Web? (b) Once someone has decided what data they want, how do they proceed to extract it (e.g., through manual request, OPeNDAP server request, FTP, etc.). (c) What file formats are used in presenting/uploading the data?	ECWBS (Ron McLaren) By request. They can request specific data which will be provided; or, when our systems are fully operational, will have Web access. (a) In progress. I’m not sure of the current status. (b) See above. (c) Not sure. The National Archive manager handles this. FNMOCC (Ryan Hofschneider) (a) Via the Web service, they request a “table of contents” which list the products available from the data service. (b) They compose an XML-like request and submit it to the Web service. (c) We present the data to the client in either XML, netCDF, or GRIB. MDA/LANDSAT (Andrew Westwell-Roper) Externally: through CCRS Discovery Portal (“CEONet”), http://geodiscover.cgdi.ca/gdp. Internally, can view through own intranet. (a) Yes. (b) CEONet gives LANDSAT order desk phone number to user. (c) Browse images specified in FGDC Geoprofile and displayed by CEONet are in JPEG format. MDA/RADARSAT (Andrew Westwell-Roper) Externally: through CCRS Discovery Portal (“CEONet”). Internally, can view through own intranet and connection to CCRS archive. (a) Yes. (b) CEONet gives RADARSAT order desk phone number to user. (c) Browse images specified in FGDC Geoprofile and displayed by CEONet are in JPEG format. NRCAN (Jim Lyons) We have developed a number of custom Web interfaces, including station network maps, a “Station Book” describing each station, and established international standard data access methods. These latter include the Automated Data Request Manager (AutoDRM) and Networked Data Centres (NetDC). See, e.g., http://www.seismo.nrcan.gc.ca/cnsn/stn_book/show_station_e.p

	hp?sta=BBB (b) AutoDRM, NetDC, FTP, HTTP, manual request (for very large volumes). (c) CA, SEED, CSS, Mk2. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) They go to this Web page for data delivery: http://www.pmel.noaa.gov/tao/data_deliv/deliv.html or this one for data display: http://www.pmel.noaa.gov/tao/jsdisplay/ or this one for combined display and delivery: http://www.pmel.noaa.gov/tao/disdel/ On these pages users manipulate menus and checkboxes, and the user interface shows them what data are available. The deliv.html and disdel have data availability information built directly into the user interface, via a Java applet, which queries the Web server in response to user selections. The server then accesses large custom designed metadata files, which are optimized for speed of access and minimum bandwidth. There is also an explicit “Availability” button which shows users a schematic time series plot for each mooring site indicating the data availability. (a) Not <i>per se</i>. Metadata are designed for use with the above Web pages, and are only accessed via server access from within the applets. (b) They click “Deliver”, and a CGI request is sent to the Web server from the applet. The requested data files are then created, and they are presented within a Web browser window with links to the data files. This is done as quickly as possible, e.g., 2 seconds to 20 minutes depending on the request. (c) ASCII; NetCDF - 4-byte real data. NetCDF - 2-byte integer data; real values are linearly mapped to 2-byte integers to save disk space and bandwidth. In most cases, 2-byte integers have sufficient resolution for data precision, and where they don’t, users are advised in readme files. USGODAE (Phil Sharfstein) Through Web site, GCMD, THREDDS. (a) Yes. (b) OPeNDAP, las, FTP, HTTP, ldm (requires manual setup request), custom Web applications. (c) Same as delivery formats.
Do you offer any facilities for “data mining?”	ECWBS (Ron McLaren) No reply. ENMOC (Ryan Hofschneider)

	No reply. MDA/LANDSAT (Andrew Westwell-Roper) No (but we're looking at it). MDA/RADARSAT (Andrew Westwell-Roper) No (but we're looking at it). NRCAN (Jim Lyons) No reply. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) By manipulating the menus on the Web pages above, users get immediate feedback on what data are available. Perhaps this fits the term "data mining". Our entire public data set is available through these pages, all the way back to the earliest deployments in 1977. USGODAE (Phil Sharfstein) No.
Is it possible for a scientist to issue a query based not only on the metadata but on some property of the actual data (not stored in the metadata)?	ECWBS (Ron McLaren) Probably not. FNMOC (Ryan Hofschneider) No. We are hoping to offer some functionality like that eventually (i.e., area thresholds). MDA/LANDSAT (Andrew Westwell-Roper) No. MDA/RADARSAT (Andrew Westwell-Roper) No. NRCAN (Jim Lyons) Data access usually based on date/time of some known event of interest (earthquake, sonic boom, etc). Many researchers like to first see what data are available before actually requesting it, but in my experience this is almost as much work as just giving them what you have, and they usually ask for all that is available anyway. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No. USGODAE (Phil Sharfstein) No.
Can you provide an example of such a query that might be performed on your data?	ECWBS (Ron McLaren) No reply. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) N/A. MDA/RADARSAT (Andrew Westwell-Roper)

	N/A. NRCAN (Jim Lyons) Give me all the waveform data from stations XXX, YYY, and ZZZ starting at date/time for a duration of 2 minutes. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) No reply. USGODAE (Phil Sharfstein) No reply.
<i>REAL TIME ASPECTS</i>	
Are the data collected in real time?	ECWBS (Ron McLaren) Hourly. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Yes. MDA/RADARSAT (Andrew Westwell-Roper) Yes. NRCAN (Jim Lyons) Yes. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Yes. USGODAE (Phil Sharfstein) Yes.
Are the data disseminated in (near) real time?	ECWBS (Ron McLaren) Yes. FNMOC (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) No. MDA/RADARSAT (Andrew Westwell-Roper) Yes. NRCAN (Jim Lyons) Yes. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Yes. USGODAE (Phil Sharfstein) Yes.

What is the time lag between data collection, dissemination, availability for access, and archive?	ECWBS (Ron McLaren) 5-10 minutes. FNMO C (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) 3 days. MDA/RADARSAT (Andrew Westwell-Roper) 2-4 hours. NRCAN (Jim Lyons) A few minutes at most. PO.DAAC (Jorge Vazquez) No reply. TAO/PMEL (Paul Freitag) Approximately 8-24 hours depending on location of buoy and satellite coverage. USGODAE (Phil Sharfstein) Depends on the data set generally 2-24+ hours.
OTHER COMMENTS	
VENUS and NEPTUNE will be collecting and disseminating a wide range of data types at differing sample rates and intervals from a common geographical region, resulting in a real time heterogeneous data set. Are there any lessons learned from your experience that could assist in the development of the VENUS/NEPTUNE DMAS, or that could assist in avoiding potential pitfalls in the design of the DMAS? Any information you feel to be relevant would be appreciated.	ECWBS (Ron McLaren) No reply. FNMO C (Ryan Hofschneider) No reply. MDA/LANDSAT (Andrew Westwell-Roper) Our relevant experience cannot be compressed into a questionnaire response. One important first step is to define the DMAS external interfaces - we have prepared a draft discussion paper on DMAS interfaces, which we will forward to Barrodale when comments on the draft have been received from UVic. MDA/RADARSAT (Andrew Westwell-Roper) Another important approach is to define DMAS functional requirements (in more detail than that of the 21/01/02 DRD) at the same time as the Wet Plant requirements. Otherwise, DMAS will become a convenient place to hide all the difficult system-level problems, such as resource management, that will affect the Wet Plant and hence the ability of the system to do what is expected. NRCAN (Jim Lyons) Have to balance advantages of handling/managing/storing disparate data types in one common waveform format vs. ease of use for clients. We have found format conversion on-the-fly to be practical using only modern SUN workstations as servers. PO.DAAC (Jorge Vazquez) No reply.

	TAO/PMEL (Paul Freitag) A significant part of our data processing and quality control procedures concerns use of instrument calibration data. All transmitted and stored data from TAO instruments must be combined with calibration data to produce user data. This process occurs at the time of real time distribution on GTS by Service Argos, for (near) real time processing and quality control at our site, and for processing and quality control of delayed data. Management of the calibration systems and data requires about the same level of resources as the acquisition of instrument data from deployed platforms. This may be a metadata issue in your scheme, but not obviously so from our perspective. USGODAE (Phil Sharfstein) No reply.
--	--

9 Appendix C – Interview Synopses

9.1 Phase 2 Interview Contacts

Agency	Contact
BC Government Active Control System	Mr. Brad Hlasny
BC Government LRDW	Mr. Peter Friesen
Fleet Numerical Meteorological and Oceanography Center – non-USGODAE-related	Mr. Ryan Hofschneider
Herzberg Institute of Astrophysics – Canadian Astronomy Data Centre (http://cadc-ccda.hia-ihc.nrc-cnrc.gc.ca/)	Mr. Severin Gaudet
Incorporated Research Institutions for Seismology	Dr. Tim Ahern, Program Manager, IRIS Data Management System
Institute of Ocean Sciences	Mr. Robin Brown Dr. Howard Freeland
JPL Physical Oceanography DAAC (PO.DAAC) (http://podaac.jpl.nasa.gov/) (NASA) EOS – EOSDIS	Dr. Jorge Vazquez (PO.DAAC User Working Group) mailto:podaac@podaac.jpl.nasa.gov
NOAA / NESDIS Information Processing Division	Dr. Ted Habermann
NOAA / PMEL	Mr Steve Hankin
Pacific Geoscience Centre (Phase 1: need to understand data flow from ocean to IRIS and other destinations) (Phase 2: are they involved in any “large data management” projects?)	Dr. Garry Rogers Mr. Richard Baldwin Mr. Tim Claydon
Pacific Forestry Centre – National Forest Information System (http://nfis.org/index_e.shtml)	Dr. Robin Quenet Mr. Brian Low

9.2 BC Active Control System

Date and Time: 30-August-2004, 13:30-14:15
Location: BCS offices
Attendees: Kent Berger-North, Mike Dunham-Wilkie (BCS),
Brad Hlasny (Base Mapping and Geomatic Services Branch of the
BC Ministry of Sustainable Resource Management)
Systems discussed: BC ACS (BC Active Control System)

Background:

The British Columbia Active Control System (BC ACS) is a system that provides GPS corrections in post-mission using the standardized Provincial coordinate system. In partnership with municipalities (BCACSm), the Province also facilitates real time correction distribution via UHF radio links and/or wireless PCS providers (e.g., Bell's 1X network). Real time accuracy levels range from approximately 1 metre (using RTCM code corrections and a mapping grade GPS receiver) to the centimetre level (using dual frequency survey grade receivers). It provides users with the ability to survey and lay out points, accurate to a few centimetres, instantaneously. The system improves the efficiency of engineers and surveyors as they will not need to transfer precise coordinate information over long distances from control monuments to perform their duties; the BC ACS gives them the accuracy they need to complete the required tasks in real time.

Notes from email prior to the phone interview:

1. Approximately 18 GPS base stations are located around the Province.
2. All base stations send their post-mission data to our Provincial Master Active Control Station (PMACS) every hour on the top of the hour.
3. PMACS then performs the following:
 - a. Data are unzipped.
 - b. Dual-frequency data are run through Quality Control software (teqc).
 - c. Dual-frequency data files are then re-packaged into a number of hourly product files for our clients (e.g., dual-frequency 1-second, dual-frequency 5-second, dual-frequency 30-second, single-frequency 1-second, single-frequency 5-second, etc.).
 - d. These data are placed online (LandData BC and FTP site) for users/clients/stakeholders to download.
 - e. For non-government users or partners, this is a user-pay system; thus, public access is not allowed.
4. Data are available online for minimum of 30 days.
5. Data are archived on tape/DVD and kept for a couple of years.

Notes from the phone interview:

1. The BC ACS Service can be described as a post-mission data delivery system, wherein field trips are made to survey/inventory some data holding (e.g., cut block survey, ground water well inventory, etc.). These GPS surveys are surveyed to a specific positioning standard (NAD83(CSRS)). The "rover" GPS data collected from the field is then combined with GPS basestation data and corrected. A typical scenario involving a base station and a "rover" (in the field):
 - a. Base station file is collected continuously.
 - b. Rover file is collected simultaneously.
 - c. Upon return to the office, the rover file is cross-correlated in the time domain; this makes the position information more precise.
 - d. For most GPS manufacturers, this process is automated.

2. At the top of the hour each base station FTP's its previous hour's data to the PMAC's computer in Victoria. QA is applied automatically to the data via software and a QC file is produced. (At this stage, an alarm is triggered if a file has not been received.) The data are then automatically packaged and put into products online.
3. Automatic checks include:
 - a. File times.
 - b. 3600 blocks of information (1 hour of data at 1 second intervals).
 - c. Position.
 - d. Multipath quality.
 - e. Data outages.
 - f. Verbose format and one-line summaries are created.
4. Data transfer is via the Internet (digital cable, dialup modem).
5. Real time transmissions are not saved or archived; they are transmitted for the following clients/applications:
 - a. CDGPS (Canada-wide DGPS Service- see <http://www.cdgps.com>) corrections via MSAT satellite.
 - b. CRD (regional system with UHF and repeaters, corrections in RTCM format).
 - c. GVRD (Bell Mobility 1X system; data streams into "rover" receiver).
6. Data are processed centrally using in-house software ("teqc") and freeware QA/QC software ("UNAVCO" from the University of Colorado).
7. Previously, COTS (Commercial Off-The-Shelf) components were not available. Now they are more common and include base station control, QC, and distribution of data.
8. The system is maintained and operated by two BMGS personnel (part-time). An offline system is available as backup if the first one goes down. Remote GPS basestation maintenance is typically done through a contractor.
9. CITS (BC Government Common IT Services) looks after file transfers and hardware support; they provide user accounts and passwords on request. This is good in terms of not needing a dedicated FTP server and support within the Ministry, but reduces responsiveness to the client as CITS does not have expertise in the knowledge domain.
10. Land Data BC are used just for data distribution (per-file payment).
11. Data are maintained online for 30 days at a minimum, 60-90 days typically. Archival is achieved in a couple of ways: hard drives are swapped out, and archives of the hard drives are made to DVD. Data volume is 30 megabytes/hour, 720 megabytes/day.
12. Another system that may be of interest to VENUS/NEPTUNE is DIMS (Data Image Management System). Project management within BMGS is provided by Rostam Yazdani and IMB support is provided by Michael Ross (IMD) or Dave Skea. A lot of images are stored on the image server: ortho images, air photos, satellite images, etc.
13. With respect to data destinations for ACS, the data are not sent to anyone. There is free and paid access via FTP pull. Commercial costs are typically \$1200/year for unrestricted downloads. PGC retrieves the data for earthquake and plate

- tectonic studies. The Geodetic Survey Division of NRCAN also retrieves the data for North American network studies/adjustments.
14. Data products are international standards: RINEX (Receiver INdependent EXchange) and RMTC. A proprietary legacy format (Trimble's SSF) is also supported.
 15. Real time database and distribution was attempted at one point, but too much time was spent online and downloading.
 16. Metadata are very simple: a flat file of information on the base station.
Information includes:
 - a. Receiver name and type.
 - b. Antenna type.
 - c. Swap times.
 - d. Time up/down.
 - e. Serial numbers.
 - f. Models, etc.
 17. Data are organized in the archive by Julian date (1-365) in hourly files by product (dual frequency 1 second, dual frequency 5 second, dual frequency 30 second values). All base stations are included in each file (e.g., "BCDL" for Dease Lake). File naming convention is by manufacturer/industry standard: 8 characters (4 for site number, 3 for day of year, 1 for the hour [A,B,C...X]). IGS (International GPS Service) formats, folders and naming structures are standardized. PGC uses the same structures for its GPS data.
 18. About 60% of the users are commercial; 20%-30% are Provincial Government agencies, and 10% are institutional/educational.
 19. Other comments and suggestions: Follow the industry requirements and needs.

9.3 BC Land and Resource Data Warehouse

Date and Time: 24-August-2004, 15:30-16:30
Location: BCS offices
Attendees: Kent Berger-North, Mike Dunham-Wilkie(BCS),
Peter Friesen (BC Ministry of Sustainable Resource Management)
Systems discussed: LRDW (BC Land and Resource Data Warehouse)

Background:

The British Columbia Land and Resource Data Warehouse (BC LRDW) consists of data in ArcSDE/Oracle, architecture and services to provide comprehensive and integrated information delivery.

Metadata for the LRDW list metadata "sheets" in MS Word Document format for many of the data sets currently in the MSRM Land Resource Data Warehouse in ArcSDE format.

BCeID is an online service that makes it possible to use a Login ID and a single password to sign in – securely – to any BCeID participating Government Web site or service. With a BCeID, you can sign in to government participating sites using your BCeID and a single password so you don't have to remember a different Login ID and password at every Web site. The information you register with BCeID is stored online, securely, in the BCeID directory as your "BCeID profile". When you sign in to a BCeID participating Government site by typing your Login ID and password in the BCeID sign-in box, BCeID confirms that: a) the Login ID you typed is registered with BCeID; and b) the password you typed is correct. BCeID then notifies the site that you are who you say you are (that you have provided valid "sign-in credentials"), and you are given access to the participating site.

Notes from phone interview:

1. The LRDW contains geospatial data and all related attribute information.
2. The data sources originate in individual ministries.
3. Data arrive via various gateways:
 - a. Oracle – straight replication from other systems.
 - b. SDE sources.
 - c. Coverages to SDE (ESRI's Spatial Database Engine) layer utilities.
 - d. Others.
4. The data are not replicated within the LRDW.
5. No QA or post-processing is applied. QA is assumed applied before it arrives.
6. Rudimentary checking is applied (e.g., things that SDE would not be happy with). Some topological constraints and relational integrity are checked.
7. Derivative products (SDE layers) are produced.
8. Metadata are required from the contributor prior to publishing data. Application programs exist for collecting metadata (the user populates a form or table).
9. Metadata standard ISO-19115 is observed (metadata field required). See srmapps.gov.bc.ca/metastar (MSRM Corporate Metadata Service).
10. Data are distributed to private and public sector clients:
 - a. For government analyses.
 - b. For public and other users of spatial and attribute data.
11. There are 3 methods of system access:
 - a. Terminal server.
 - b. ARCIMS Web interface.
 - c. FTP packaging and shipping.
12. Identification and authentication components include:
 - a. BCeID (see background notes).
 - b. AIC.
 - c. IDIR for government authentication.
 - d. A proxy determines what's available.
13. With respect to time lags for the archive: as new data source is coming online, deltas are updated on a nightly basis and a full refresh is performed on a monthly basis.
14. Note: take a look at Integrated Land Registry Project.

15. The LRDW database is approximately 300 gigabytes using the old Ministry of Environment file-based system.
16. Data sources include:
 - a. Ministry of Sustainable Resource Management.
 - b. Ministry of Water, Land, and Air Protection.
 - c. Land and Water British Columbia.
 - d. Ministry of Energy and Mines.
17. The data are available instantly through the Terminal Service, but not through the packaging service (the data need to meet policy requirements).
18. Distribution is through <http://srmapps.gov.bc.ca/apps/dwds> (Data Warehouse Distribution Service) with an SDE 8.3 front end to Oracle 9.0. The Oracle server is a purely relational server, with complex objects stored in blobs and interpreted by SDE (i.e., the Oracle Spatial Cartridge is not used).
19. CITS (BC Common IT Services, a division of the BC Ministry of Management Services) manages the database backup considerations.
20. Data products include
 - a. Shape files.
 - b. .E00 files.
 - c. CSV files (for non-spatial data).
 - d. GML (Geographic Markup Language - coming in the Fall).
21. The system is being migrated to an e-commerce configuration as a replacement for LDBC distribution.
22. There are no facilities for data mining.
23. QA requirements differ between the source and target. Currently, QA is at the level a custodian feels is appropriate and accurate. LRDW performs minimal QA but does run the data model through delivery and performs a data load through tests.
24. New services are being built around the warehouse:
 - a. Large variety of gateways.
 - b. Single service for configuration.

9.4 Fleet Numerical Meteorological and Oceanography Center

Date and Time: 09-September-2004, 13:30-14:00
Location: BCS offices
Attendees: Kent Berger-North, Mike Dunham-Wilkie (BCS),
Ryan Hofschneider (FNMOC)
Systems discussed: FNMOC (Fleet Numerical Meteorological and Oceanography Center)

Background:

The Fleet Numerical Meteorological and Oceanography Center (FNMOC) assimilates real time weather data from automatic and non-automated sources, processes them with a numerical atmospheric prediction model, and then disseminates the resulting data products (weather forecasts) in near real time.

TEDS (Tactical Environmental Database Server) represents the US Navy's effort to manage the multiple types of environmental data types required by the Navy's diverse applications. TEDS fills a Fleet requirement for standardized Meteorological and Oceanographic (MetOc) databases and extraction routines. This system provides a common data framework to shore facilities and forward deployed assets to access relevant environmental data and determine impact on operations. TEDS is comprised of two components: a scaleable and modular data repository, and a middleware process that enables the transport and transformation of data to/from the data repository.

Notes from the phone interview:

(Notes from the initial email responses are not included here).

1. Input data sources are airport stations, ships at sea, fixed and mobile stations, and probably buoys.
2. Data are transmitted by batched FTP and at other times HTTP pushes. The data are broadcast as it becomes available.
3. Message formats include UAREPS (upper air reports) and METARS.
4. NAVO (Naval Oceanographic Office) forecasts come in netCDF format. AFWA (Air Force Weather Agency) are in GRIB format. FNMOC is responsible for meteorological products, and NAVO for oceanographic products (although there is some overlap).
5. When the model data are output from forecast models, they come out in flat file format. An ISIS collection of flat files goes into TEDS or a Grid DataBlade enabled database.
6. TEDS is comprised of 1D and 2D grids (e.g., a 2D grid at a particular level and time for forecasts). It is Informix version 9 based.
7. FNMOC is switching almost exclusively to Linux, getting rid of all the old Sun boxes; forecast models are still run on SGIs, etc.
8. TEDS Metcast is the primary vehicle for distributing gridded data and satellite imagery.
9. The Grid DataBlade is a back-end solution. Clients are requesting an XML interface; at the time that this is implemented, the Grid DataBlade can push data operationally.
10. DODs encodes the data in a URL (REST style Web service).
11. TEDS does not use PostgreSQL.

12. Metcast clients compose a request for products; the subscription logic is on the client side (the client must poll the server to determine what new data are available).
13. Recommendations:
 - a. Strongly recommend accepted Web services standards (usable by all platforms and languages) e.g., REST or SOAP style.
 - b. Full XML schema.
 - c. Small amounts of data in XML; large amounts of data in netCDF.
 - d. Use of OpenGIS Web coverage standards (implemented for forecast parameters; for any particular query, overlays can be applied).

9.5 Herzberg Institute of Astrophysics / Canadian Astronomy Data Centre

Date and Time: 23-August-2004, 13:30-14:15
Location: BCS offices
Attendees: Kent Berger-North, Mike Dunham-Wilkie (BCS),
Severin Gaudet (NRC)
Systems discussed: CADC (Canadian Astronomy Data Centre)

Background:

The Canadian Astronomy Data Centre (CADC) was established in 1986 by the National Research Council of Canada (NRC), through a grant provided by the Canadian Space Agency (CSA), as one of three world-wide distribution centers for astronomical data obtained with the Hubble Space Telescope (HST). In subsequent years the CADC's mandate has expanded significantly to include the development and support of archives for other astronomical facilities whose operational costs are shared by the NRC. These include the Canada-France-Hawaii Telescope (CFHT), the James Clerk Maxwell Telescope, and the twin 8-m telescopes of the Gemini Observatory.

The CADC is also currently developing the Canadian Virtual Observatory (CVO). The CVO will seamlessly combine the content of all of the CADC's archives as well as other publicly available astronomical data sets, and provide archival researchers ready access to astronomical source catalogues derived from these holdings, calibrated images and spectra, and powerful tools to query the content of the CVO in an intuitive manner.

Located at the Dominion Astrophysical Observatory (DAO) of the Herzberg Institute of Astrophysics (HIA) in Victoria, BC, the CADC staff consists of professional astronomers and software developers who, along with providing support for the above archives, have developed an abundance of other sophisticated tools to support and enhance the research efforts of Canadian (and international) astronomers. Their efforts ensure that the maximum possible scientific exploitation of every observation obtained at the world-class observatories supported in whole or in part by the NRC is achieved.

Notes from the phone interview:

1. In 1985 the Canadian Space Agency supported the establishment of a center that would serve as a data repository for Canadian astronomers needing access to Hubble Space Telescope data. As computer systems became more advanced it was determined that electronic data transfer was possible, and that there was no longer a need to be physically present at an instrument or facility. The electronic exchange of HST data was expanded to include other telescope data, including:
 - a. CFHT (Canada-France-Hawaii Telescope).
 - b. James Clerk Maxwell.
 - c. Dominion Astrophysical Observatory.
 - d. Gemini telescopes.
2. Initially the data management and distribution was handled by a scientist (astronomer). Today the data are managed by a staff of approximately 14 domain experts (astronomers) and Information Technology personnel (programmers, operators, database administrators).
3. The facility handles 12 terabytes per year (in and out), and has the capacity to handle 20 terabytes/year.
4. Most data come from intermediary repositories.
5. A fundamental and important difference between CADC and VENUS/NEPTUNE is that CADC is not directly involved in the operation of instruments (telescopes). Individual contributing observatories serve as a buffer and perform their own QA and data packaging (e.g., the HST Space Science Telescope Institute in Baltimore).
6. CADC does additional metadata generation and packaging, and produces raw and enhanced metadata catalogs.
7. Data types include spectra, images, and cubes (space and frequency).
8. File inspection is performed by automated mechanisms.
9. Data are always pulled to enable local resource management. This makes operators independent of the data management mechanisms (the data generator is independent of the archive).
10. Data arrive in unscheduled batches.
11. The CADC has a gigabit Ethernet connection; there is no bottleneck at the facility.
12. Data transfers are periodic or episodic. Upon transfer, a set of additional data products is produced. Using a CFHT example, raw data are supplemented by first order data products (calibration and instrument signal removal) and sent to Paris. Second order products are then generated (science characterization, mosaic stitching, geometric corrections, metadata enhancement). CFHT has a legacy survey to extract science from the data. All data goes to Paris for second level processing; they come back 6 times per year.
13. The data and sampling are non-homogeneous. The catalogs and data products attempt to create homogeneous descriptions of the data, which tends to be sparse (typically addressing only 20 of approximately 100 parameters). The data tend to be floating point (as opposed to integer or character).

14. The primary interest of various project participants varies. Observatories are concerned about pixels; CADC is concerned about metadata. Quality and consistency depends on the source, instrument, and people maintaining the software and operating the instruments.
15. QA standards were developed at CADC, and checking is performed at CADC. Types and nature of the checks performed (also at the source) include:
 - a. Ongoing (established) checks.
 - b. New instruments.
 - c. New modes.
 - d. New software.
16. Data aggregation is possible to enhance signal to noise ratios, etc.; for example, multiple observations of a single part of the sky will be aggregated for better SNR or exposure time. However, it is important not to perform aggregation too soon, as it may be a one-way function (with data loss). Attempt to dissociate data collection from data packaging.
17. With respect to data availability in time, the HST (for example) has two streams. The first is public and immediately available; the second is proprietary for 1 year.
18. Databases are organized by collections (e.g., CFHT). A file storage system handles data for all the collections. A link between the data set name in the collection and the file is created and stored. The real filename is not important.
19. File sizes vary from 10 kilobytes (spectra) to 400-500 megabytes (images).
20. Data formats included FITS (self documenting, contains metadata), GND, radio formats, and XML (but not for pixel data). Data formats are typically determined by the provider.
21. Data dictionaries are maintained to map header content to database content. The data dictionary identifies the units, source, and data types. Particular instruments may use the same keyword differently for different instruments. This is also a problem across data providers (even using common formats such as FITS). For example, keyword "x" may contain a floating point value and be used by something else in another part of the same file (context dependent).
22. Regarding destinations of the data (for other agencies, etc.), it is on a "they pull" basis.
23. With respect to backup, the whole collection is stored on site (2 copies) and at UVic (on tape libraries). The data are recoverable but not servable. Primary data are all on spinning magnetic disk.
24. Post processing includes producing better data products for the user, catalog augmentation, and metadata generation. Proprietary aspects of the data may have impacts on real time QA (delays varying from minutes to years).
25. Metadata are generated completely automatically and is triggered on arrival. Different processing pipelines may be used for proprietary vs. public data.
26. Time lags between arrival of data and distribution depend on the data source and corresponding data policies.
27. Raw metadata database is Sybase. The data mine is in DB2.
28. About 40 terabytes of unique bytes are stored on the system (totalling 80 terabytes for 2 copies): 12 million distinct files (in 3 copies for 36 million files), total 120

- terabytes. The raw database is approximately 100 gigabytes. The database is growing at 2 terabytes per year (linear increase).
29. With respect to data access, all Web browsing is free. Processing resources and data retrieval is by user registration (manually vetted), over FTP or HTTP proxy.

9.6 Incorporated Research Institutions for Seismology

Date and Time: 09-September-2004, 09:30-10:30
Location: BCS offices
Attendees: Kent Berger-North, Mike Dunham-Wilkie (BCS),
Tim Ahern (IRIS)
Systems discussed: IRIS (Incorporated Research Institutions for Seismology)

Background:

The Incorporated Research Institutions for Seismology (IRIS) is a university research consortium dedicated to exploring the Earth's interior through the collection and distribution of seismographic data. IRIS programs contribute to scholarly research, education, earthquake hazard mitigation, and the verification of the Comprehensive Test Ban Treaty. Support for IRIS comes from the National Science Foundation, other US federal agencies, universities, and private foundations.

The IRIS Global Seismographic Network (GSN) is one of the four major components of the IRIS Consortium. The goal of the GSN is to deploy over 128 permanent seismic recording stations uniformly over the Earth's surface. The GSN provides funding to two network operation centers: IRIS/ASL in Albuquerque, NM (operated by USGS), and IRIS/IDA in La Jolla, CA (operated by Scripps Institute of Oceanography). As of 2003 the IRIS GSN was made up of over 128 stations with affiliations to USGS, UCSD/IDA, GEOFON, Pacific21, NCDSN, GEOSCOPE, MedNet, BGR, BFO, USNSN, BDSN, TriNet, AFTAC and several other national and international networks. The IRIS GSN stations continuously record seismic data from very broad band seismometers at 20 samples per second (sps), and provide for high-frequency (40 sps) and strong-motion (1 and 100 sps) sensors where scientifically warranted. It is also the goal of the GSN to provide for real time access to its data via Internet or satellite. Over 75% of the IRIS GSN stations meet this goal.

Notes from the phone interview:

1. Data arrive from 2500 stations in total and from 1500 of these in real time. IRIS doesn't get data directly from stations but from operating centers.
2. Approximately 50 centers are routinely sending in data (e.g., Canadian data comes from Ottawa).
3. Some centers are in Kyrgystan, Kazakhstan, and the Former Soviet Union (FSU) republics. IRIS installed some sites in the 1980s to get things up and running by providing initial funding.

4. About 2/3 of the observatory data are collected in real time. Some places have no communications facilities; some data arrive via diplomatic pouch.
5. With respect to military sensitivity of the data, there is the occasional problem (e.g., in the FSU). Some data are edited or not delivered at all. Filtering or processing of the data (as sometimes performed by US military) reduces its usefulness for science applications. These issues are addressed on a station by station basis. However, once data have been released to the DMC (Data Management Center), they are available for immediate access (except for data from proprietary programs, which have up to a 2 year delay). SOSUS hydrophone data are not available.
6. Monitoring is not the mandate of IRIS, but of the USGS. IRIS has no direct involvement in emergency response (triggered by seismic events). It is primarily a scientific organization.
7. Data from Canada reach IRIS relatively quickly with a turnaround of approximately 7-9 days. This includes very long period data (hundreds to thousands of seconds in the case of free Earth oscillations).
8. Two data collection centers are supported: Albuquerque and San Diego. The NEIC (National Earthquake Information Center) gets its data directly from Albuquerque (except for data wherein IRIS is the entry point to the US). Data flow is very complex.
9. The data are both pushed and pulled. IRIS has a whole series of request tools, and the data are essentially pushed due to firewall and network address translation issues. Users can subscribe to real time streams. Other streams are available for historic data.
10. IRIS was established in 1983 as a non-profit organization. The GSN (Global Seismic Network) was established in 1988, which essentially coincides with the start of the DMC. Since then, historical data holdings have been incorporated (dating back to the early 1970s).
11. SPIDER and WILBUR are event-based systems. WILBUR is the Web-based facility to access the SPIDER and FARM repositories. Lists of tools are available on the Web site.
12. PASSCAL Portable Array for Seismic Studies of the Continental Lithosphere (portable instruments for studies with time frames ranging from a few weeks to years). Typically STS-2 and CMG-3 type of sensors (higher frequency) are used for looking at the near surface, crust, lithosphere and upper mantle on shorter time scales. GSN instruments are permanent and tend to be Teledyne borehole or STS-1 sensors.
13. GSN data flow through one of two data collection sensors where QC procedures are applied (looking at glitchy or gappy data, determining if the sensor is OK), but the harder part is to ensure that the metadata (station, orientation of the sensor, response to ground motion) is correct. First line QC metadata are produced by operations centers. All data in the DMC are passed through an automated QA framework (noise, power spectral density, RMS, glitchiness, gappiness, percent availability). IRIS spends \$1-1.5 M per year on data QA before the data goes into the archive. Data are not altered, just flagged.
14. All metadata are stored in Oracle. Access to this metadata is a work in progress.

15. FDSN (Federation of Digital Broadband Seismographic Network) file formats are used (SEED – Standard for Exchange of Earthquake Data) which includes header and waveform information. MiniSEED contains the waveform data (data records of 256 bytes to 4 kb). The paradigm here is to get the data into MiniSEED in real time.
16. Metadata and MiniSEED data are treated separately.
17. Metadata can be updated asynchronously with data-less SEED volumes.
18. For each channel, the data are stored in a day file. For each station, the data are stored for all channels in a day file. Up to 12 channels are stored (which may have different sample rates or characteristics). The primary data unit is a station day file.
19. Some metadata are computed and put into Oracle. Requests for seismograms with particular characteristics can be made. Currently this is not available interactively and online, but it is the way they would like to go.
20. The model at IRIS is to not permit accessors of data to perform data mining at their facilities as it is resource intensive (especially being a tape-based archive).
21. The amount of data at IRIS could conceivably be on disk, but is not at this point. The ATA disks have reliability issues as cheap drives historically have not been an option. It still costs approximately \$20,000 per terabyte for storage.
22. There are 125 terabytes in the archive in 4 copies (approximately 30 terabytes unique data). Two copies of two different sort orders (station and time) are maintained. A fifth copy of the time sort is stored out of the geographic region.
23. The capacity of the archive is 1.2 petabytes.
24. The robotic storage system holds 6000 tapes of 200 gigabytes each. The data retrieval is on the order of 5 seconds (1 minute worst case with removing a tape, inserting a tape).
25. Oracle 9i.x is used for metadata; the database is on the order of 60 gigabytes (approximately 1000:1 ratio of data to metadata).
26. Most of the servers are Sun servers (there are one or two 4000/4500's, the majority are midrange Sunfires of V880 class). They are rock solid machines.
27. The backup system is provided via the Oracle method of duplicate server (backup capability provided on Oracle). The switchover is not automatic and takes 1-2 minutes (cost-based decision to do this rather than have an immediate switchover mechanism).
28. IRIS does not use Veritas for backups, but for filesystem and volume manager.
29. Incremental dumps are made in Unix, synchronized with the Oracle backup. Redundant copies are shipped out twice per week.
30. The operations group (handling the data in and out) has 4-5 people plus 2 systems administrators. The total number of staff is 19. Data ingestion and delivery are automated.
31. John Delaney has been the primary contact between IRIS and NEPTUNE, and Severin Gaudet (as of 3-4 years ago) was also involved.
32. IRIS is very interested in archive and distribution of the seismic and hydrophone data from NEPTUNE.
33. Current data ingestion is 16 terabytes/year with linear growth.

34. SEED formatted data currently come in from Ottawa. Data format conversions are performed in Ottawa.
35. Real time will be provided in CD1 (Continuous Data 1) format. The CD1 group is based in Vienna, Austria (Canada contributes to that group).

9.7 *Institute of Ocean Sciences: Data Collection and Transmission Procedures*

Date and Time: 04-August-2004, 09:30-10:30
Location: IOS
Attendees: Kent Berger-North, Mike Dunham-Wilkie (BCS),
Robin Brown (IOS)
Systems discussed: Data Collection and Transmission Procedures

Background:

The Ocean Science and Productivity Division of the Institute of Ocean Sciences conducts research to provide a foundation for assessing the effect of changes in ocean conditions on marine ecosystems and understanding the role of the ocean in the global climate system. The program has three main areas:

- monitoring and understanding the functioning of marine food chains, including the environmental factors controlling distribution and abundance of fish;
- understanding the ocean's role in the global climate system;
- Arctic ecosystems and climate change.

The division has staff located at the Institute of Ocean Sciences and the Pacific Biological Station. Research activities are conducted in coastal and open ocean waters of the North Pacific and in the Arctic Ocean.

Notes from the interview:

1. Robin described the process by which DFO collects oceanographic data from a ship and submits it to MEDS. He used a typical Line P cruise to demonstrate:
 - a. Digital data are collected in real time, using instruments such as CTD's. The data files are recorded to local disks.
 - b. Bottle data are collected for chemical and biological parameters, and for correcting the CTD data.
 - c. Some of the analyses are performed on-board and some are reserved for later analyses on shore.
 - d. Bottle data are used as quality control (QC) for the CTD data. Where possible, data are traced to primary standards.
2. The process of collection to submission of data to MEDS is as follows:
 - a. Data are collected from the vessel.

- b. Quality analysis and control (QA/QC) is performed (sometimes in a semi-automated process). Flags are set where appropriate.
 - c. Data are approved (in principle) by the Chief Scientist once QA/QC and post-processing is complete.
 - d. The data are submitted to the IOS archive in the IOS Header format (human-readable ASCII data files with verbose headers; the processing history is preserved in the header).
 - e. Historically, the IOS archived data were sent to MEDS annually. Now MEDS has access to the IOS data directories and MEDS pulls the data on demand. MEDS reformats the data to their own needs and archives them. IOS has supplied software libraries for data manipulation and file format reading. Should there be any questions regarding the data, usually arising from MEDS QA/QC, MEDS will contact IOS for clarification. Typically it is Don Spear (Head, Ocean Profiles Section) who asks for clarification.
3. Locally at IOS, there are two processes for data processing and archive. The first data process deals with the data files themselves and the second with searchable on-line metadata database.
4. The data file processing is conducted as follows:
 - a. Files are processed and saved to directories on the DFO/IOS network. The directories have read-only access to many users and read-write access to only a very few.
 - b. The data are saved in IOS Header format for profile and time series data.
 - c. Working directories are transferred (on completion) to archive directories. The data are saved in a file-based system of human readable data.
 - d. Raw data, intermediate processing stages and final “working/finished data” are archived to off-line media. One copy is retained at IOS and a second is lodged with the Public Archives of Canada. Raw data and intermediate stages are removed (erased) from on-line storage to conserve disk space.
5. The metadata management and access is as follows:
 - a. Metadata are preserved in a searchable, on-line database.
 - b. Metadata are generated through an automated process, whereby a “crawler” is sent through the archives on a regular basis and metadata extracted from the archive. This process ensures that both additions to the archives (new data) and updates (and possibly corrections) are included in the searchable Metadatabase, with very little human intervention.
 - c. Metadata are accessible via the Web at the following IOS URL’s.
 - Profile/station data:
http://www-sci.pac.dfo-mpo.gc.ca/osap/data/SearchTools/SearchProfiles_e.asp
 - Moored instrument data:
http://www-sci.pac.dfo-mpo.gc.ca/osap/data/SearchTools/SearchMoorings_e.asp
 - Satellite Images:
http://www-sci.pac.dfo-mpo.gc.ca/osap/data/SearchTools/SearchSatellites_e.asp
 - d. The metadata database is updated weekly (by scanning the data archives).
 - e. Information about profiles and moored time series data is contained in the database. Metadata include critical information such as time, date, and parameters measured.

- f. The metadata database is a Microsoft Access product with a Microsoft Visual Basic front end and business logic. The database contains on the order of a few hundred thousand records.

9.8 Institute of Ocean Sciences: Argo System

Date and Time: 04-August-2004
Location: IOS
Attendees: Kent Berger-North, Mike Dunham-Wilkie (BCS),
Howard Freeland (IOS)
Systems discussed: Data Collection and Transmission Procedures

Background:

The Argo program, which is being developed through a collaboration with the Canadian Marine Environmental Data Service (MEDS), is a system of floats designed to measure temperature and salinity in the upper 2000 m of the oceans. It will eventually consist of an array of 3000 profiling floats, providing 100,000 temperature/salinity/pressure profiles and velocity measurements per year distributed over the global oceans at an average 3-degree spacing. There are currently approximately 100 active Argo floats in the Northeast Pacific region and about 1300 floats distributed worldwide. The primary repository for Argo data is on the USGODAE (US Global Ocean Data Assimilation Experiment) Fleet Numerical data server at Monterey, California (<http://www.usgodae.org/usgodae.html>).

Notes from the interview:

- 1) The Argo data collection and distribution system is as follows:
 - a) Argo floats obtain a profile (surface to 2000 m) every 10 days.
 - b) Every 10 days, the float surfaces and transmits to Service Argos, a system of polar-orbiting satellites. The baud rate is extremely low but there are sufficient "hits" each day that a profile of 80 levels can be extracted reliably in about 7-8 hours of transmission time.
 - c) Data are received by Service Argos in Maryland.
 - d) Every 6 hours, MEDS pulls data from Service Argos via FTP.
 - e) Profiles are decoded.
 - f) Real time QC is applied to the data.
 - g) The QC'd data are distributed via:
 - i) The GTS (WMO) in KKYY format;
 - ii) MEDS Web server;
 - iii) Argo Global Data Centers (located in the US at the Naval Post Graduate School (NPGS) in Monterey, California, and in France by IFREMER on the Mercator Server, one of Europe's contributions to GODAE).

- h) Delayed Mode data are delivered with higher QC (applied by the Principal Investigator for a region/country) and takes approximately 6 months to be returned to the global archives.
- 2) The KKYY data will eventually (over the next year) be phased out in favor of BUFR format. KKYY is limited to 2 decimal places for salinity. TESAC formats include KKXX and KKYY.
- 3) The Canadian Argo project has about a 90% success rate in throughput of good data. Occasionally, position information is missing. For example, on 27-July-2004, Service Argos was unable to provide positioning information for a few days. No positions were lost as position data were recomputed at a later date.
- 4) Utilities exist to produce and maintain an Argo mirror of the global data set on a PC.
- 5) The Delayed Mode system of data delivery includes more careful QC. Available data include P1,P2,T1,T2,S1,S2 fields for Pressure, Temperature and Salinity (respectively). x1 is observed data, and x2 is corrected.
- 6) Canada currently has 84 floats; the US has 680; the total globally is 1297. The target is 3000.
- 7) Over the next 2 years, the plan is to abandon Service Argos and switch to Iridium telemetry. Some of the considerations of this planned switch are:
 - a) The cost is less (factor of 4-5).
 - b) There will be a shorter latency (immediate versus daily upload).
 - c) 1000 samples in the vertical (as opposed to 80 at present) will be transmitted.
 - d) Data will be in engineering units as opposed to coded hexadecimal.
 - e) Argos PTT-days cost \$15 CAD (for Canada).
 - f) Iridium costs are USD \$0.80 per handshake (2 required) + USD \$1.86 per minute.
 - g) Iridium has no positioning capability. Dr. Steve Riser of the University of Washington has developed a single antenna for GPS and Iridium; GPS will be installed on each float.
 - h) Iridium are LEO (Low Earth Orbit) satellites.
- 8) Argo data are assembled into 1 file per day per ocean in a single netCDF file. There are 3 oceans (Indian, Pacific, Atlantic) in the data dissemination structure.
- 9) With respect to post-calibration, efforts have been made to collect coincident CTD data. However, the variability in 10 days exceeds the corrections that would be made to the CTD cast.

9.9 Jet Propulsion Laboratory / Physical Oceanography Distributed Active Archive Center

Date and Time: 09-September-2004, 11:30-12:15
Location: BCS offices
Attendees: Kent Berger-North, Mike Dunham-Wilkie (BCS),
Jorge Vazquez (JPL, PO.DAAC)
Systems discussed: Physical Oceanography Distributed Active Archive Center

Background:

The Physical Oceanography Distributed Active Archive Center (PO.DAAC) of the Jet Propulsion Laboratory (JPL) is responsible for archiving and distributing data relevant to the physical state of the ocean. Most of the products available at the PO.DAAC were obtained from satellites and are intended for use in oceanographic and interdisciplinary scientific research. Please refer to the on-line data catalog for a listing of available PO.DAAC data products, tools, and services.

The PO.DAAC, an element of the Earth Observing System Data Information System (EOSDIS), and a member of the DAAC Alliance, distributes products from NASA's Earth Science Enterprise (ESE) Earth Observing System (EOS) project, as well as through partnerships and cooperative agreements with other organizations and institutes within and outside the United States. These groups include the Jet Propulsion Laboratory (JPL) and other NASA centers, National Oceanic and Atmospheric Administration (NOAA), U.S. Naval Oceanographic Office (NAVOCEANO), French Space Agency, Centre National d'Etudes Spatiales (CNES), National Space Development Agency (NASDA) of Japan, as well as the University of Miami and Brigham Young University.

PO.DAAC's data development, management, archival, and distribution activities are guided in large part by a group of colleagues in the form of an EOSDIS advisory User Working Group (PO.DAAC UWG). This ensures that the data PO.DAAC provides do, and will, meet the ongoing needs of the oceanographic community it serves.

Notes from the phone interview:

1. The function of the PO.DAAC is to serve the physical oceanography data sets for NASA (satellite data sets): temperature, SS (sea surface) winds and SS height. Various platforms and *in situ* data (buoys) are used for validation. The primary focus is on satellite data.
2. Altimetry is a radar pulse; temperature is IR (which can't see through clouds – a big problem); winds are calculated indirectly (related to sea state and energy return of transmitter pulse).
3. The near real time operational aspect is relatively uncharted territory. The center is attempting to get data out as quickly as possible after acquisition.
4. Some users want data pushed to them (retrieved from provider and sent to researcher or agency). The system is constantly being refined. A lot of quality checks need to be performed (e.g., all of data sent). With temperature, for example, it is desirable to look at as many images as possible to mitigate cloud masking.
5. Some QC is automated. For example, if data do not come in, an event alarm is generated. Something that cannot be easily automated is image processing (looking at cloud masking, etc.). Much of the feedback for quality comes from users.

6. A section of their Web site is dedicated to “known problems” where feedback is documented. This serves as useful information for the ongoing development of the system.
7. The source of the data depends on the satellite and data set (NOAA, Navy, French Space Agency).
8. Different agencies or data providers do their own quality checks. QC is exchanged with other organizations and agencies providing the data. Examples for altimetry are close collaboration with the French Space Agency CNES and for SST with the University of Miami.
9. Responsibility for documenting data issues is also a collaborative effort. For example, some data sets can be retrieved from more than one agency, including NASA and NOAA. In some cases Web interfaces at the PO.DAAC allow for additional functionality in distributing the data. In other cases there is a clear difference in the roles of the data provider and distributor. That is, the data provider does not have the main responsibility for distribution. It falls on the PO.DAAC. An example is with SST and the relationship between the PO.DAAC and the University of Miami algorithm and science team. The clear separation between data provider and distributor allows for the PO.DAAC to be the focal point for user interactions.
10. Data are stored on an FTP site that is backed up. Other parties can provide more technical details on the archive system, which is basically robotic tape drives (terabytes of data). The center attempts to put everything on the FTP site.
11. Web interfaces have been developed that allow users to extract data (e.g., specify area rather than global data sets, which reduces file transfer requirements). The main interface is called POET and can be cursor or form driven. Data are assembled and put on the FTP site; the user receives an email when it is ready. Sometimes the amount of data is too large, so the data must be split into several requests. Most requests take a few minutes.
12. SST at 4 km resolution makes for large data files.
13. Metadata are already provided with SST (for example); usually it is collected from the providers. An international SST project is underway to standardize all the metadata for all the satellite sensors. NetCDF will be used for this to address cross-platform issues, Live Access Server, etc.
14. All user statistics are logged for all data interactions and FTP access (volumes sent out, volumes ingested). A single user sign-on mechanism (where a user is tracked through the system) is desired, as multiple logons can skew the statistics.
15. User interactions feed back into how data are served (e.g., regional versus global data).
16. The PO.DAAC is trying to develop better FAQs as data holdings and streams are growing while the support resources remain constant (need to maintain efficiency).
17. The volume of data is critical with respect to the type of hardware used, etc. Subsets of the data are useful, and it is essential to get the data out electronically. DAAC will not put a data set online without basic documentation and software for reading it. PO.DAAC is a firm believer in using FTP sites. It’s nice to build a GUI, but researchers want quick and direct access.

18. OpenDAP is a very nice search engine. POET was developed about the same time as Live Access Server and OpenDAP and has some similar functionality. POET serves ASCII, netCDF, and GIS formats. Adding formats increases the number of users of that data.
19. Usage is changing. About 70-80% of the community are researchers. Now with real time mode, things are changing as the data become available operationally. Commercial users, weather forecast models, National Hurricane Center for wind data, and coastal modeling are emerging real time applications.
20. Ground truth data from NOAA buoys are carried at PO.DAAC, but not performed there (done at the University of Miami). However, the PO.DAAC does distribute the matchup *in situ* database that is created at the University of Miami along with the satellite SST data.

9.10 National Oceanic & Atmospheric Administration / National Environmental Satellite, Data and Information Service

Date and Time: 02-September-2004, 13:30-14:15
Location: BCS offices
Attendees: Kent Berger-North, Mike Dunham-Wilkie (BCS),
Ted Habermann (NOAA NGDC)

Background:

The National Geophysical Data Center (NGDC), located in Boulder, Colorado, is a part of the US Department of Commerce (USDOC), National Oceanic & Atmospheric Administration (NOAA), National Environmental Satellite, Data and Information Service (NESDIS). They are one of three NOAA National Data Centers (NNDC).

The NGDC provides scientific stewardship, products and services for geophysical data describing the solid Earth, marine, and solar-terrestrial environment, as well as observations from space.

The NOAA NESDIS mission is to provide and ensure timely access to global environmental data from satellites and other sources to promote, protect, and enhance the Nation's economy, security, environment, and quality of life. To fulfil its responsibilities, NESDIS acquires and manages the Nation's operational environmental satellites, provides data and information services, and conducts related research.

Notes from the phone interview:

1. The most significant difference between VENUS/NEPTUNE and NGDC is that VENUS/NEPTUNE will be designed for a relatively small number of data streams. NGDC's data holdings currently contain more than 300 digital and analog databases, some of which are very large. Real time systems that would be

similar to VENUS/NEPTUNE in terms of data management include three satellite systems:

- a. The Operational Linescan System (Lights at Night, Aurora)
http://dmsp.ngdc.noaa.gov/html/sensors/doc_ols.html.
- b. GOES (Geostationary Operational Environmental Satellite). This system acquires a picture of the sun every minute (and is available in near real time at the Space Environment Center). GOES also produces a full scan of the US and northern hemisphere every 12 minutes.
- c. Space sensors (polar orbiting and geostationary), including solar wind and energetic particle measurements.
2. NGDC does not have a satellite receiving station and is not the operating agency responsible for acquiring the data. NESDIS and the Space Environment Center (next door to NGDC) receive the data and forward them.
3. The DMSP (Defense Meteorological Satellite Program) activities would be of relevance and should be of interest to VENUS/NEPTUNE.
 - a. The primary users are AFWA (Air Force Weather Agency) and FNMOC (Fleet Numerical Meteorology and Oceanography Center).
 - b. The DoD used to have a three day hold on the data; now they are available to some users in a couple of hours.
 - c. Data are transmitted in 1/2 orbit pieces, and access is available on 1/8 orbits. The data are not sent as an image, but as scanlines that need to be processed into an image. Metadata are generated automatically and some data quality assessment is performed (e.g., interpolation of noisy scanlines).
4. NGDC generates standards-level metadata. There is a wide variety of approaches to granular data (e.g., files or 1/8 orbits). Metadata are stored in MySQL or Postgres for DMSP. Crude spatial representation information is collected in real time (e.g., nearest grid number for a fixed 1-degree grid) along with a pointer to a file and a pointer to a .GIF image.
5. There are a number of customers who get the whole data stream (or only parts of it). The Web interface is appropriate for users who can browse the archives looking for a specific piece of an orbit (i.e., a particular geographical area at a particular point in time).
6. One of the challenges for NGDC is to migrate as much data as possible into a relational database or open system using established standards such as FGDC (Federal Geographic Data Committee) or ISO (International Standards Organization). (They are essentially evolving into the same thing.)
7. NGDC is interested in the Grid/Geodetic/Spatial DataBlade products to allow direct query and management of data.
8. Often cited arguments against RDBMS include:
 - a. How are you going to archive a proprietary format?
 - b. There is a need to read data without the assistance of a software program.
9. An important consideration for data migration as technology evolves is minimizing the number of data paths.
10. The CCSDS (Consultative Committee for Space Data Systems) has proposed the OAIS (Open Archival Information System) Reference Model

- (<http://www.ccsds.org/docu/dscgi/ds.py/Get/File-143/650x0b1.pdf>), available through <http://arc.nesdis.noaa.gov>. This is now an ISO standard for long-term archives, and is being proposed as a potential model for the NESDIS long-term archive. In addition to addressing documentation and quality assurance issues, it addresses submission agreements between the submitter and the archive.
11. VENUS/NEPTUNE might want to consider SensorWeb (open source development, OpenGIS formats and access). Contact Dr. Mike Botts at the University of Alabama (Huntsville). OpenGIS and a relationship to IOS are the best bets for a framework for the future.
 12. Within the next 6 months NGDC hopes to be able to take 10 representative data streams and study how they are going to be encoded for SensorWeb.

9.11 National Oceanic & Atmospheric Administration / Pacific Marine Environmental Laboratory

Date and Time: 15-September-2004, 15:30-16:00
Location: BCS offices
Attendees: Mike Dunham-Wilkie (BCS),
Steve Hankin, NOAA/PMEL
Systems discussed: THREDDS, Live Access Server (LAS), OPeNDAP

Background:

NOAA's Pacific Marine Environmental Laboratory (PMEL) carries out interdisciplinary scientific investigations in oceanography and atmospheric science. Current PMEL programs focus on open ocean observations in support of long-term monitoring and prediction of the ocean environment on time scales from hours to decades. (See <http://www.pmel.noaa.gov/>.)

Steve Hankin is a computer scientist at PMEL. He is the principal developer of the FERRET visualization and analysis program. He has served on the ANSI Computer Graphics Committee, X3H3, and has a keen interest in computer standards issues. Steve Hankin is chairperson of the IOOS Data Management and Communications Steering Committee (DMAC-SC) and serves on the DMAC-SC Data Transport expert team and the DMAC-SC Data Facilities Management outreach team. (See http://www.unidata.ucar.edu/packages/dods/archive/proposals/nopp-html/nopp_47.html.)

Notes from the phone interview:

1. Steve described how THREDDS, LAS, and OPeNDAP fit together. They are quite distinct but related. See the following points.
2. OPeNDAP is a Web services definition for requesting binary data; it forms an insulating layer between binary data sources and requesting applications, in much the

same way as ODBC (Open Database Connectivity) forms an insulating layer between applications that use data and the data sources (databases) themselves.

3. Live Access Server (LAS) is a product server (built at the “product” level of abstraction, a lower level of abstraction than the “data” level, on which OPeNDAP is built). Live Access Server delivers data products (grids, images, etc.) to client applications (e.g., FERRET, IDL, Matlab). FERRET, created at PMEL by Steve Hankin, is an interactive computer visualization and analysis environment designed to meet the needs of oceanographers and meteorologists analyzing large and complex gridded data sets.
4. The focus of LAS is on data from the physical sciences (e.g., oceanography, geology, meteorology).
5. Data for the products that LAS serves can come from an OPeNDAP source, from local files, relational databases, OpenGIS Web map or feature server.
6. THREDDS stands for “Thematic Realtime Environmental Distributed Data Services” (see <http://my.unidata.ucar.edu/content/projects/THREDDS/index.html>). It is a Unidata initiative and has an educational focus – excerpt from the Web site above:
“The mission of THREDDS is for students, educators and researchers to publish, contribute, find, and interact with data relating to the Earth system in a convenient, effective, and integrated fashion.”
7. THREDDS will provide a way for organizations with data to create XML-based hierarchical catalogs of their data holdings (identifying the data sets, some metadata, and access protocols), and a way for consumers to find and access these data.
8. Ben Domenico (ben@ucar.edu) is a contact person for THREDDS.
9. There are technologies like LAS and OPeNDAP available for use by IOOS programs, and the IOOS programs will have specific types of data and specific access requirements. Not all technologies are appropriate for all data types / access requirements. For example, OPeNDAP is highly suitable for access to gridded data (and has been made an *operational* component of IOOS for this purpose). However, OPeNDAP is relatively immature insofar as access to relational database data is concerned (and has thus been made a *pilot* component of IOOS for this purpose). Gap analyses need to be performed to determine where new tools need to be developed or existing ones enhanced.
10. VENUS/NEPTUNE may also want to look at OBIS compatibility. OBIS (Ocean Biogeographic Information System – <http://www.iobis.org/>) is an interface to marine species data from all over the world. OBIS provides access to interactive physical oceanographic data at regional and global scales and software tools for biogeographic analysis. One such tool is DiGIR (Distributed Generic Information Retrieval – see <http://www.digir.net/>), a tool for accessing data from distributed sites.
11. The DMAC plan at http://dmac.ocean.us/dacsc/imp_plan.jsp on page 49 says “... Gateways that provide translation from the GIS network protocols to the OPeNDAP protocol will be developed by IOOS.” It is unclear at this point whether this gateway will be built on top of the OpenGIS Web Coverage Server standard, as that standard may fit well with the richness of OPeNDAP.
12. IOOS, like other US organizations, is committed at present to following FGDC metadata standards, but there will likely be increasing international pressure to

migrate to the ISO 19115 standard. This is not seen as a major issue as the standards are quite similar.

13. Stephanie Watson (swatson@mbari.org) is a good person to talk to about metadata standards. She is working very closely with both IOOS and NSF/Orion efforts in moving forward agreement on metadata standards.
14. A difference between IOOS and NSF/Orion: IOOS consists of many distributed programs (Argo, etc.) and is operational in focus; the mission is to provide stable, robust, operational services. NSF/Orion, on the other hand, has more of a research focus. NEPTUNE falls under the NSF/Orion umbrella (OOI – Ocean Observing Initiative).

9.12 Pacific Forestry Centre / National Forest Information System

Date and Time: 24-August-2004, 13:30-14:15
Location: BCS offices
Attendees: Kent Berger-North, Mike Dunham-Wilkie (BCS),
Robin Quenet (Research Scientist, Integrated Resource
Management, NRCAN), Brian Low (Geospatial Scientist,
NRCAN)

Background:

The Pacific Forestry Centre (PFC) is one of five Canadian Forest Service research and development centers committed to the sustainable development and competitiveness of the Canadian forest sector. PFC provides essential forest research to secure the social, economic and environmental value of the forest for future generations.

PFC employs approximately 135 research and development staff working in four major program areas: Forest Biology; Forest Resources; Marketing and Operations; and Industry, Trade and Economics. Forest research and development activities at PFC are not limited to BC, but have a national and international scope.

PFC, as part of the Canadian Forest Service, is responsible for:

- enhancing Canada's forests and forest sector through the discovery, development, demonstration and transfer of innovations;
- conducting and publishing research in the areas of forest resources, forest protection, the forest environment, and wood utilization;
- providing technical advice and scientific information to clients;
- providing funding and technical services for forest management on Federal and Indian lands;
- generating sound economic information, statistics and advice; and
- addressing industry, trade and general international issues and opportunities in support of the forest sector.

Notes from the phone interview:

1. The primary data resource is the National Forest Information System (NFIS).
2. The data are primarily spatial data holdings with attribution. Layers are served up by individual jurisdictions and have varying amounts of information, with BC typically providing the most information.
3. Data collections are OGC (Open GIS Consortium) compliant and are accessed through the Geoconnections Web portal (<http://cgdi.gc.ca> or <http://www.geoconnections.org>).
4. Standards are important: NFIS uses OGC where possible (e.g., WFS – Web Feature Service, WMS – Web Mapping Service).
5. USD imagery for land cover and land use, and topographic information for Canada is also available.
6. Other organizations that contribute data to the collections are DFO, CTI (Center for Topographic Information, Sherbrooke), Canadian Geoseismology Network, Environment Canada, Industry (Cubewerx, NASA, GEODEN), and UVic.
7. PFC manages a distributed database system with 17 nodes across Canada. Most provinces have a node. Each node has a large data warehouse with collections corresponding to the node's jurisdictions.
8. Metadata are polled from each of the servers and centralized in an Oracle metadata database at PFC. Metadata updates are not consistent (not everybody publishes metadata to the same extent and quality).
9. The OGC compliant servers have “get” capabilities for pulling new data nightly. The project office cascades to each of the servers at 0400. The Web portal is updated on a daily basis.
10. Systems being considered include:
 - a. OGC compliant stateless catalogs (speak to Ron Lake at Galdos Systems Inc. (<http://www.galdos.ca>)).
 - b. Cubewerx Inc. (Hull, Que) (<http://www.cubewerx.com/main/>).
 - c. Open source from Scandinavia.
11. Data collections include 5 m digital orthomaps and 1 m digital orthomaps.
12. CFS initiative and National Forestry Database Program provide data entry mechanisms from other sources.
13. The NFDPA National Afforestation⁹ Inventory (NAI) provides data validation.
14. Data dictionaries are published with standards. They need specific metadata for quality sources.
15. Some provinces are using FGDC (Federal Geographic Data Committee) metadata; some think that it's not detailed enough for some of the biology work. Some research organizations are using MBII. Some search tools don't pick up MBII tags on a query.

⁹ Afforestation is the planting of trees for commercial purposes, usually on land supporting non-forest types.

16. Data access is via DACS. DACS is a general-purpose, distributed system that combines single sign-on capability and role-based access control for Web services, which are any static and computational resources provided by a Web server through HTTP. The DACS design is open and built from industry-standard protocols and APIs, making porting to new environments straightforward. Designed and implemented by DSS, development of DACS was initiated in May, 2001. DACS is a key component of Canada's National Forest Information System (NFIS) and has been deployed in conjunction with the IRMS team at the Canadian Forest Service's Pacific Forestry Centre (PFC). Current releases of DACS are Web-based and run on several varieties of Unix and Windows 2000 Server in conjunction with the Apache Web server. There is also some support for Microsoft's IIS.
17. Data products include 640x480 PNG files, individual features, shape file, GML (Geography Markup Language) stream.
18. VENUS/NEPTUNE should investigate the OGC Sensor Collection Service (used by Environment Canada – "websensor"). EC contact Tom Kralidis (tom.kralidis@ec.gc.ca). *Background:* Owsview is a Web-based thin client which supports discovery, access and visualization of supported specifications of the OpenGIS Consortium (OGC) and Canadian Geospatial Data Infrastructure (CGDI) endorsed specifications, such as Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), Styled Layer Descriptor (SLD), Web Registry Service (WRS), Web Map Context Documents, Catalog Service, Gazetteer Service, Sensor Collection Service (SCS) and the GeoConnections Discovery Portal API (searching for products and services). Owsview also exemplifies the benefits of standards based services for chaining between services for discovery, access and visualization of information holdings.
19. Metadata are maintained in an Oracle database (Cubewerx) at the project office. A potential alternative format will be XML (Galdos).
20. Comments and suggestions:
 - a. Ensure that the data structure is searchable with Google and Yahoo tools.
 - b. DAC (Distributed Access Control) is very useful for determining who has access to resources over the Web. It provides a means of identifying where you are from and provides authentication by jurisdiction using browser "cookies". It permits granular control of users, groups and URLs. ACLs for resources at the file or parameter level can be checked automatically.

9.13 Pacific Geoscience Centre

Date and Time: 23-August-2004, 16:00-16:45
Location: BCS offices
Attendees: Kent Berger-North, Mike Dunham-Wilkie (BCS),
Garry Rogers, Tim Claydon, Richard Baldwin
Systems discussed: Canadian Seismographic Network (Western Region)

Background:

The Pacific Geoscience Centre (PGC) in Sidney, British Columbia, Canada, houses staff of the Geological Survey of Canada (GSC) Pacific Division. The GSC is a part of Natural Resources Canada, the Canadian government department that specializes in energy, minerals and metals, forests and Earth sciences. GSC Pacific Division is one of several Divisions in the Minerals and Regional Geoscience Branch of the GSC, which falls under the Earth Sciences Sector of Natural Resources Canada.

Staff at PGC conduct research into areas of geology and geophysics within the region of Western Canada known as the “Canadian Cordillera”, as well as along the continental margin that is Canada’s West Coast. PGC researchers are involved in national programs on earthquake seismology and geodynamics, including the evaluation of earthquake hazards. They also undertake studies in reflection seismology, geomagnetism, paleomagnetism, geothermics, the Earth’s gravitational field and marine sedimentology.

The principal research programs at the Pacific Geoscience Centre are:

- ***Earthquake Seismology*** - the western component of Canada’s National Earthquake Hazards Program, which is aimed at understanding the causes of, and hazards associated with, earthquakes in Canada.
- ***Geodynamics*** - monitoring and investigating movement of the Earth’s crust in support of research into earthquake hazards and global change.
- ***Cordilleran & Continental Margin Tectonics*** - studies into the geological architecture of the Canadian Cordillera, its history, and recent activity along its tectonic plate boundaries in order to increase understanding of the various processes that have formed, and are forming, this region.
- ***Marine Geoscience*** - research into the marine environment of Canada’s West Coast both from a global and regional geoscience perspective, including seafloor mapping, sediment distribution, active faulting, and geological processes along the continental margin.

Notes from the phone interview:

1. The types of data collected are digitized velocity and acceleration values (20, 40, or 100 samples per second per sensor). Data samples are 27-28 bits uncompressed. Not much acceleration (strong motion) data are collected at this time.
2. Data are transmitted by satellite, microwave, telephone, or radio.
3. Canada has 120 land stations. 70% of the data in Canada come into PGC. The national network is divided into eastern and western regions. The Ottawa branch has 85% of the data in real time (they get all the POLARIS data; PGC gets only western Canada POLARIS data). Automated hourly batch transfers via FTP to Ottawa addresses any missing data.

4. Some data come from UW (University of Washington), UAF (University of Alaska at Fairbanks). Data also flow to them and the US national network. Washington, Idaho and Oregon have 150 stations.
5. Data are exchanged in real time.
6. Approximately 1.3 gigabytes per day (0.5 terabytes per year) comes into PGC; about 2 gigabytes per day (0.75 terabytes per year) into Ottawa.
7. Data are maintained on a 12 terabytes RAID system (spinning magnetic disk).
8. Data are archived on site in two copies to CD; Ottawa serves as the offsite storage.
9. Real time data destinations include UW for specific stations, Montana Tech, and the Alaska Tsunami Warning System.
10. Other repositories include the NEIC (National Earthquake Information Center) in Colorado, the US national network (Golden, Co.), and the National Archive in Ottawa.
11. QA and post-processing are not performed in real time. Communications between sites uses CRC on data packets. Typical failure modes include the occasional GPS timing problem (lockup or clock drift). An error of 3-4 seconds is OK.
12. Most post-processing addresses the determination of location and how the Earth moved. The analyses are not automated. Location computation is automatic; an analyst then checks and verifies the trigger and location data (for example, an event could be a mine blast that is not easily distinguishable for classification by machine).
13. The current system is being migrated to a more automated software package (Antelope) for data processing and storage. It incorporates real time acquisition and tracking, and the creation of metadata databases.
14. No data products are generated *per se*, but information is used for national building codes, and hazard and risk analysis.
15. Data disseminated to other agencies include stations, locations, magnitudes, and fault mechanisms.
16. Data formats include CA (an in-house format), and miniSEED. In-house digitizers (27.5 bit) are used exclusively in Canada except for the Yellowknife array.
17. Data and metadata standards: IRIS PASSCAL uses Quanterra Q330.
18. Data storage is currently on CD. The system is migrating to RAID spinning magnetic disk, tape and/or DVD backup (plus offsite).
19. Data access is Web-based using AutoDRM (Ottawa office file request).
20. Current data holdings are 2.1 terabytes on CD and 8 gigabytes of metadata. Three or four formats are available via AutoDRM (CA, GSC 2.0, SEED). AutoDRM is freeware from Switzerland.
21. Not much metadata are exchanged, and this is done only on an *ad hoc* basis.
22. Other comments include:
 - a. Use the KISS principle.
 - b. Use standard Unix utilities and open standards (note: Antelope is an exception for this application).

- c. The PGC system is 100% operational; the research component has been split off into other structures.
- 23. It is hoped that there will be a NEPTUNE feed to PGC and possibly IRIS.
- 24. Alaska uses Antelope software; UW uses Earthworm.
- 25. Data collected are acceleration or velocity, normally acceleration for broadband sensors is collected in 3 axes.
- 26. Data samples are 32 bits, typically 40 samples (varies), and are compressed.
- 27. Average throughput is 1000-2000 bits/sec; 2400 baud links are sufficient. When there is a lot of activity the bandwidth exceeds 2400 baud. The data are buffered in the digitizer and will get out eventually.
- 28. Noisy stations are masked (gain is turned down). Instrument setup and configuration is through (in) the digitizer.
- 29. Data are packeted at 6 second intervals, with transmission lasting 3 of the 6 seconds. Transmission methods include microwave, RF modem, spread-spectrum radio, satellite and land line.
- 30. The digitizer will transmit and expects acknowledgement and time packages; it will send an ERR if it doesn't get them. If packets are missing, it asks for them.
- 31. For one-way data transmission (no error checking possible) the data are sent twice, 11 minutes apart. The data acquisition system discards any duplicate packets.
- 32. The national network uses the in-house format. The data end up as CA files.
- 33. Data acquisition uses in-house systems; the Antelope system may be adopted later. All stations use the CA format (and will likely stay like this for a while).
- 34. Calibration information is entered into the data acquisition system and consists of compensation for component tolerances. This type of calibration is not needed periodically and is only performed when the instrument is serviced. Calibration information can be sent remotely to the seismometer.
- 35. There is no interactive instrument management or configuration.
- 36. Gain settings or firmware changes are possible over a 2-way link.
- 37. Comments include:
 - a. Use a good-sized buffer if it is possible to drop a communications link. PGC has 6 hours available in the digitizers; several days would be nice.
 - b. If using RF, use multi-hop links (concentrators can count packets; if an intermediary loses a package, the concentrator can re-request data from the digitizer). Five serial inputs with one serial output keeps the bandwidth down.

10 Appendix D – Table of Contents for “Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems”

Executive Summary	1
Part I. Overview	7
Preface	11
Section 1. Overview	15
Introduction	15
The Vision	16
Challenges	17
Community Building	18
International Cooperation	18
Section 2. Technical Analysis	21
IOOS Data Communications	21
The Observing Subsystem and Primary Data Assembly/Quality Control	21
The DMAC Subsystem - A Data Communications Infrastructure	24
Metadata Management	24
Data Discovery	25
Data Transport	26
On-line Browse	30
Data Archive and Access	31
Modeling and Analysis Subsystem	34
Information Products and Applications	34
Interoperability with Other Data Management Systems	35
Section 3. Management, Oversight, and Coordination	36
Governance	36
IOOS Data Policy	37
IOOS/DMAC Standards Process	37
User Outreach	38
System Engineering Approach	39
DMAC Cost Model	41
Section 4. Immediate Priorities for Implementation	45
Metadata/Data Discovery/Data Location	45
Data Transport	46
Data Archive	47
System Engineering Approach	47
Concrete Guidance to Data Providers	48
Part II. Phased Implementation Plan for DMAC	51
Section 1. Functional Requirements	53
General Requirements	53
1. IOOS DMAC Vision	53
2. DMAC Overall Functional Requirements	54
3. Participating Activities	55
4. Infrastructure/Communications	56
5. Technology Infusion	56
6. Other General System Requirements	57
Data Communications Infrastructure and Archival	58
1. Metadata/Data Discovery (MD) Requirements	58
2. Data Transport (DT) Requirements	62
3. Data Archiving and Access (AA) Requirements	67

Section 2. Phased Implementation Plan	77
Metadata/Data Discovery Activities and Schedule	77
Data Transport Activities and Schedule	85
Data Archive and Access Activities and Schedule	98
Part III. Appendices	113
Appendix 1. Metadata/Data Discovery	113
Appendix 2. Data Transport	138
Appendix 3. Data Archive and Access	195
Appendix 4. User Outreach	221
Appendix 5. System Engineering Approach	275
Appendix 6. Technology Maintenance and Refreshment	281
Appendix 7. Biological Data Considerations	285

11 Appendix E – List of Institutions Hosting OPeNDAP Servers

Antarctic Cooperative Research Centre, Tasmanian Partnership for Advanced Computing (TPAC)
Carolinas Coastal Ocean Observing and Prediction System (Caro-COOPS)
Center for Ocean Land Atmosphere Studies (COLA)
Columbia University/LDEO - International Research Institute (IRI/LDEO)
Florida State University - Center for Ocean-Atmospheric Prediction Studies (COAPS)
George Mason University - Seasonal to Interannual Earth Science Information Partner (SIESIP)
Gulf of Maine Ocean Observing System (GoMOOS)
Maine - Department of Marine Resources (Maine DMR)
Monterey Bay Aquarium Research Institute (MBARI)
Naval Oceanographic Office (NAVOCEANO)
NASA/GSFC - Goddard Distributed Active Archive Center (GDAAC)
NASA/JPL - Physical Oceanography Distributed Active Archive Center (PODAAC)
NASA/JPL - Ocean Earth Science Information Partner (OceanESIP)
NASA/MSFC - Global Hydrology Resource Center (GHRC)
NCAR - Vegetation/Ecosystem Modeling and Analysis Project (VEMAP)
NOAA - Atlantic Oceanographic & Meteorological Laboratory (AOML)
NOAA - Chesapeake Bay Program Office
NOAA - Climate Diagnostic Center (CDC)
NOAA - Forecast Systems Laboratory (FSL)
NOAA - Geophysical Fluid Dynamics Laboratory (GFDL)
NOAA - National Center for Environmental Prediction (NCEP)
NOAA - National Climatic Data Center (NCDC)
NOAA - National Data Buoy Center (NDBC)
NOAA - National Oceanographic Data Center (NODC)
NOAA - Northeast Fisheries Science Center (NMFS/NEFSC)
NOAA - Pacific Marine Environmental Laboratory (PMEL)
North Carolina Ocean Observing System (NC-COOS)
Northern Illinois University (NIU)
Oregon State University - Remote Sensing Ocean Optics (ORSOO)
Science Applications International Corporation (SAIC)
Skidaway Institute of Oceanography (SKIO)
Texas A&M University (TAMU)
UNAVCO - Using Global Positioning System (GPS) technology for Earth sciences research
UCAR - Data Portal
UCAR - Unidata Internet Data Distribution (IDD)
United States Geological Survey (USGS) - Woods Hole Field Center
University of Alabama/Huntsville - Global Hydrology Resource Center (GHRC)
University of Hawaii/APDRC - Asia-Pacific Data-Research Center
University of Hawaii/SOEST - Sea Level Center
University of Kansas - Kansas Geological Survey
University of Massachusetts - School of Marine Science and Technology
University of Miami - Rosenstiel School of Marine and Atmospheric Science (RSMAS)
University of Miami - HYCOM Consortium for Data Assimilative Modeling (HYCOM)
University of Rhode Island - Graduate School of Oceanography
University of South Florida - Institute for Marine Remote Sensing (USF/IMaRS)
University of South Florida - Coastal Ocean Monitoring and Prediction System (USF/COMPS)
US GODAE - Global Ocean Data Assimilation Experiment (USGODAE)
Woods Hole Oceanographic Institution - US/GLOBEC Georges Bank Program (GLOBEC)
Woods Hole Oceanographic Institution - Upper Ocean Mooring Data Archive (UOP).

12 Appendix F – Object-Relational Databases

This appendix provides an introduction to object-relational database systems, highlights their benefits with respect to traditional relational database systems, and discusses their suitability to VENUS/NEPTUNE.

12.1 Relational Databases

A relational database is a set of tables containing data fitted into predefined categories. Each table represents either a real world object, a relationship between objects, or both. When the table represents a real world object, the predefined categories (called columns) are comprised of attributes that describe the object, and each row represents a single object. When the table represents a relationship, the predefined categories identify the objects taking part in the relationship, and each row represents a single instance of the relationship.

Consider as an example a database with three tables: STUDENT, COURSE, and ENROLLED-IN. The STUDENT table might have the following columns: StudentID (a unique identifier for the student represented by the row), Name, Address, etc. The COURSE table might have the following columns: CourseNum (unique identifier), CourseName, ClassHours, LabHours, etc. Finally, the ENROLLED-IN table, representing a relationship between STUDENTs and COURSEs, would have columns identifying the specific STUDENTs and COURSEs in a relationship, as well as possibly attributes of this relationship (e.g., Grade).

The following diagram illustrates a few rows of this database:

STUDENT		
StudentID	Name	Address
792221	Jane Doe	123 Main Street
782045	Tom Brown	12 Brown Drive
792080	John Smith	4030 Jason Ave

COURSES			
CourseNum	CourseName	ClassHours	LabHours
ABC501	Intro to ABC	3	0
XYZ599	Advanced XYZ	3	3

ENROLLED-IN		
StudentID	CourseNum	Grade
792221	ABC501	94
782045	XYZ599	73
792080	XYZ599	85

Using Structured Query Language (SQL), the user of this database can ask the following sorts of questions:

- i) What is John Smith's address?
- ii) What courses are being taken by Jane Doe?
- iii) Who is taking ABC501?
- iv) What is the average grade for Advanced XYZ?
- v) What was the highest grade for Advanced XYZ? Who got this grade?
- vi) Produce a report showing grades for Intro to ABC.

12.2 Drawbacks of Relational Databases

The relational database model is perfectly suitable for the sort of application just described, where the data types are simple (just numbers and text strings) and the operations are simple (e.g., arithmetic, comparison). Where the limitations in the relational model show is when the objects are complex. Consider for example a database that contains time series measurements taken by various instruments. The following are two possible implementations of this database:

- 1) One row per instrument:

Traditional relational databases offer just scalar data types (integer, floating point, text string) plus a "binary large object" (blob) catch-all data type for unstructured data (e.g., text pages, pictures), so if we're going to store a time series as an object in a relational database we have to store it as a binary large object.

Instrument	Time series
1	<Time series as binary large object>
2	<Time series as binary large object>
3	<Time series as binary large object>

- 2) One row per instrument per time step:

Another approach is to store the time series as a series of rows, one row per instrument per time step.

Instrument	Time	Time Series Data Value
1	1.0	43.5
1	2.0	43.2
...	...	
2	1.0	12.4
2	2.0	12.5
...	...	
3	1.0	33.1
3	2.0	34.2
...	...	

One problem with option 1 is that there is very little the user can do with a time series while it is in the database – about all the user can do is extract the entire time series for an instrument, and then perform any analysis on their own workstation. A second problem is that it is very impractical to keep the time series up to date. The time series is continually changing, but since it is stored as an atomic blob in the database, updating the time series requires completely overwriting this blob, an operation that takes time. So the effect is that what is in the database at any given time is just an out-of-date snapshot.

Option 2 suffers from neither of these two problems, but nonetheless it is a very inefficient design. Each instrument id is stored multiple times, and since there is an overhead to storing a row, this overhead is repeated multiple times for each time series.

We will come back to this example in Section 12.4 when we discuss the benefits of object-relational databases.

12.3 Object-Relational Databases

In their book **Object-Relational DBMS's – Tracking the Next Great Wave** (<http://www.amazon.com/exec/obidos/tg/detail/-/1558604529?v=glance>), Stonebraker and Brown list the following three characteristics of object-relational databases:

1. Support for base type extension in an SQL context.
2. Support for complex objects in an SQL context.
3. Support for inheritance in an SQL context.

The following three sections discuss each of these characteristics in turn.

12.3.1 Data Type Extension

As mentioned earlier, traditional relational database servers support just a few data types: integer, floating point, string, date, and sometimes simple binary large object (blob), and on each of these there is defined a limited number of operations. Object-relational technology provides the ability to define new data types and operations on both original and new data types.¹⁰

There are two varieties of extended data types:

¹⁰ Traditional relational database management systems do provide the ability to define new operations on existing data types, but it is important to note that in those cases the operations execute outside the database server, on the client side of the client-server wall. Operating on the client-side rather than the server-side results in performance degradation and means that the operations cannot be used in server-side mechanisms such as indexes.

- 1) **“Distinct” data types.** These are user-defined data types that share their internal representation with an existing type (their “source” type), but which are considered to be a separate and incompatible type for most operations. They can be viewed as restricted (in terms of their domain) and/or extended (in terms of their available operations) versions of the source data type. For example, we might define a “sound speed” data type that is restricted to floating point numbers that are in the physically realizable range of sound speed values. We might also define some operations on sound speed values that might make sense in the physics of sound speeds but wouldn’t be applicable to floating point numbers in general. Distinct data types provide a mechanism for data integrity to be enforced almost automatically – it is impossible for the database server to store a bad sound speed value in a sound speed data type column.
- 2) **“Opaque” data types.** These are data types whose precise definition is not documented (hence they are opaque), and which are intended to be manipulated only using the documented interface, which consists of a set of functions. Many Geographic Information System object relational databases have opaque data types for polygons (used to represent the geographic extents of lakes, islands, and large rivers), lines (used for highways and streams), and points (used for spot elevations, points of interest, etc.). The complement to opaque data types is a set of functions that operate on the data type. For the polygon data type the operations might include:
 - a. the geometric constructors Intersection, Union, Difference, etc. These would take two or more polygons and return a new set of geometries representing their intersection, union, difference, etc.
 - b. the boolean operations Intersects, Contains, IsContainedIn, etc. These would take two polygons and return ‘True’ if they intersected (for example), and ‘False’ if they didn’t.

In the context of VENUS/NEPTUNE we might use opaque data types for multidimensional arrays of data (e.g., grids and profiles), images, seismic recordings, etc. A time-image data type could be used to manipulate image grids (image slices stacked in time) allowing, for example, a time series of pixel values to be extracted.

12.3.2 Support for Complex Objects

The basic building blocks for creating complex types in object-relational databases are “composites” and “sets”.

Composites are built from two or more other data types. For example, in VENUS/NEPTUNE we might define a data type called “argo_recording_atdepth”, consisting of several floating point values, one for each of the properties (e.g., temperature, salinity, dissolved oxygen, pressure) measured at a specific point on the Argo float’s ascent.

A set is a grouping of objects of the same type. An “argo_profile” set data type, for example, might be defined as a set of “argo_recording_at_depth” values. A time series is another example of a set data type.

12.3.3 Inheritance

The third distinguishing characteristic of object-relational databases is inheritance, whereby one data type can inherit the characteristics (e.g., functions that operate on the data type) from one or more parent data types. Operations available for an inheritor may include the operations available on the parent, customized versions of the operations available for the parent, or completely new operations.

In the context of VENUS/NEPTUNE we might have data types called “recording_at_depth” and “profile”, with inheriting data types called “argo_recording_at_depth” and “argo_profile.” The operations defined for the argo_profile data type would include all the operations defined for the profile data type, possibly customized to deal with the specifics of Argo floats.

12.4 *Benefits of Object-Relational Databases*

Let’s revisit the time series example from Section 12.2. With an object-relational database we can construct a distinct data type called “instrument_recording” and a time series set data type called “timeseries(instrument_recording).”

Instrument	Time series
1	Timeseries(instrument_recording) for instrument 1
2	Timeseries(instrument_recording) for instrument 2
3	Timeseries(instrument_recording) for instrument 3

However, unlike the blob implementation, with this implementation we can perform operations (other than simple extract) on each Timeseries(instrument_recording) value. Operations might include:

- i) append(Timeseries(instrument_recording), instrument_recording)
 - add an instrument recording to the end of the time series
- ii) extract(Timeseries(instrument_recording), startTime, endTime, deltaTime)
 - subsample and/or window a time series, returning a new time series
- iii) fft(Timeseries(instrument_recording))
 - return a new time series that is the Fast Fourier Transformed version of the input time series.

The benefits of using an object-relational database for VENUS/NEPTUNE can be summarized as follows:

- 1) Storage efficiency is improved (relative to the one-row-per-simple-element approach).
- 2) Network use is reduced (relative to blob approach, where entire blobs, instead of just the relevant pieces, must be sent to the client machine).
- 3) Concurrency is improved, since pieces of large data type objects can be locked independently of other pieces of the same object (e.g., a time series can be appended to while earlier parts of the time series are being read). Multiple users can safely query the same data concurrently.
- 4) Composite data types allow data to be “bundled” with their metadata. This bundling enables data base implementers to build operations in such a way that metadata are updated appropriately and consistently.
- 5) Integrity is improved, since distinct data type constraints and opaque data type operations can ensure that bad data are rejected before it is stored in the database.
- 6) Database extensibility is improved: the object-oriented paradigm facilitates re-use of code. Additional data types and operations can be added easily, without breaking existing applications.
- 7) The approach fosters the uniform treatment of data items. The traditional approach is to access metadata with SQL and the actual data with file access mechanisms. By instead storing all data types in the same framework, applications can use the same SQL interface to perform complex queries that are based on any of these data items, including dynamically-derived properties of the data.
- 8) Object-relational databases can be extended to include custom data access methods. Traditional relational database management systems include b-tree indexes, which are very effective for indexing types of data that have natural linear orderings (numbers, text strings, etc.), but not effective for querying spatial data. As new data types are added to an object-relational database, appropriate access methods can be added as well.

13 Appendix G – Data Mining

The following has been excerpted and adapted, with the author's permission, from "Introduction to Data Mining" (1999) (http://www.exinfm.com/pdf/files/intro_dm.pdf) by Osmar R. Zaiane of the Department of Computer Science, University of Alberta (<http://www.cs.ualberta.ca/~zaiane/>). Two of Dr. Zaiane's research interests include Content-based Information Retrieval and Multimedia Data Mining.

Comments specific to VENUS/NEPTUNE have been added and are placed in square brackets and underlined.

What are Data Mining and Knowledge Discovery?

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Figure 10 shows data mining as a step in an iterative knowledge discovery process.

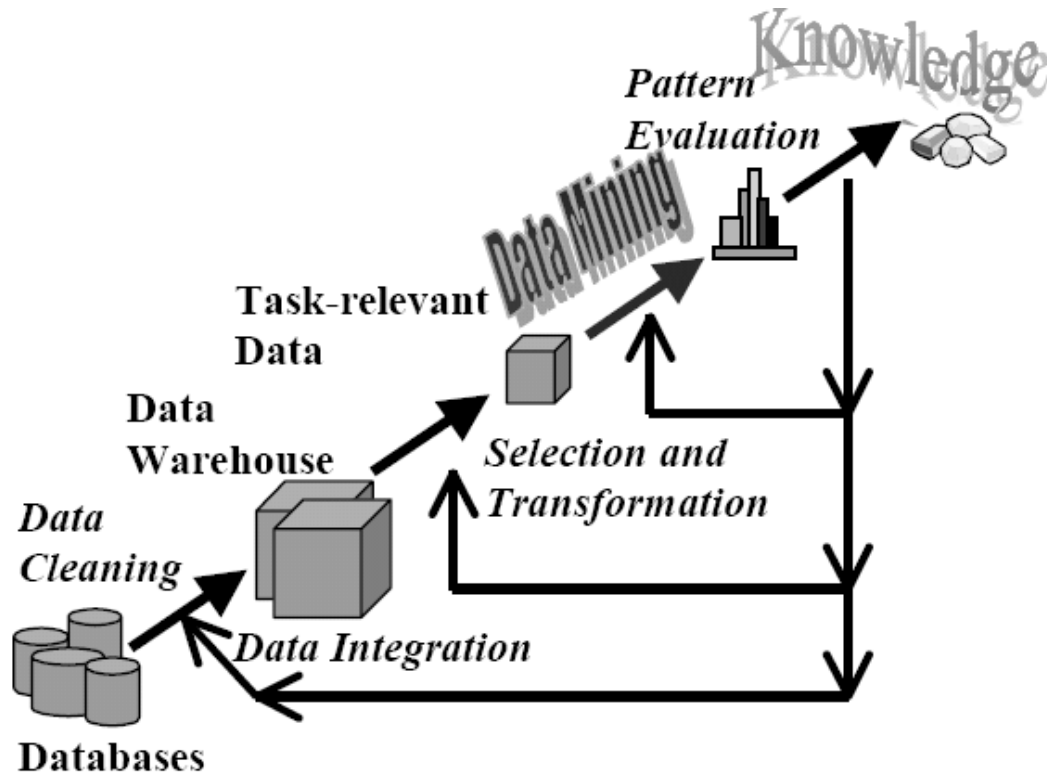


Figure 10. Data mining and its role in knowledge discovery.

[This diagram maps onto the framework illustrated in Section 6. The “Data Warehouse” boxes correspond with the Product Server, while the “Task-relevant data” box corresponds to the Data Mining server.]

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning:** also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection. [QA process in VENUS/NEPTUNE, may include just flagging the suspect data.]
- **Data integration:** at this stage, multiple data sources, often heterogeneous, may be combined in a common source. [In VENUS/NEPTUNE, the common source is the Products Server.]
- **Data selection:** at this step, the data relevant to the analysis is decided on and retrieved from the data collection. [In VENUS/NEPTUNE, the privileged user specifies which data are to be extracted to the Data Mine Server.]
- **Data transformation:** also known as data consolidation, it is a phase in which the selected data are transformed into forms appropriate for the mining procedure.

- **Data mining:** it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation:** in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation:** is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, *data cleaning* and *data integration* can be performed together as a pre-processing phase to generate a data warehouse. *Data selection* and *data transformation* can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results. [It is important to note that KDD produces new data products that can then be fed back into the VENUS/NEPTUNE Data Products Server.]

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called “gold mining” and not “rock mining”, thus by analogy, data mining should have been called “knowledge mining” instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

The following are some specific forms of data mining. [They have been described using a “video store” business example, but they can easily be seen to relate to VENUS/NEPTUNE science examples as well.]

- **Characterization:** Data characterization is a summarization of general features of objects in a target class, and produces what is called *characteristic rules*. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For example, one may want to characterize video store customers who regularly rent more than 30 movies a year.
- **Discrimination:** Data discrimination produces what are called *discriminant rules* and is basically the comparison of the general features of objects between two classes referred to as the *target class* and the *contrasting class*. For example, one may want to compare the general characteristics of the customers who rented more than 30

movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures.

- **Association analysis:** Association analysis is the discovery of what are commonly called *association rules*. It studies the frequency of items occurring together in transactional databases, and based on a threshold called *support*, identifies the frequent item sets. Another threshold, *confidence*, which is the conditional probability that an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis. For example, it could be useful for the video store manager to know what movies are often rented together or if there is a relationship between renting a certain type of movies and buying popcorn or pop. The discovered association rules are of the form: $P \rightarrow Q [s, c]$, where P and Q are conjunctions of attribute value-pairs, and s (for support) is the probability that P and Q appear together in a transaction and c (for confidence) is the conditional probability that Q appears in a transaction when P is present. For example, the hypothetical association rule: $RentType(X, "game") \wedge Age(X, "13-19") \rightarrow Buys(X, "pop") [s=2\%, c=55\%]$ would indicate that 2% of the transactions considered are of customers aged between 13 and 19 who are renting a game and buying a pop, and that there is a certainty of 55% that teenage customers who rent a game also buy pop.
- **Classification:** Classification analysis is the organization of data in given classes. Also known as *supervised classification*, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a *training set* where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the video store managers could analyze the customers' behaviors vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels "safe", "risky" and "very risky". The classification analysis would generate a model that could be used to either accept or reject credit requests in the future.
- **Prediction:** Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.
- **Clustering:** Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called *unsupervised classification*, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of

maximizing the similarity between objects in a same class (*intra-class similarity*) and minimizing the similarity between objects of different classes (*inter-class similarity*).

- **Outlier analysis:** Outliers are data elements that cannot be grouped in a given class or cluster. Also known as *exceptions* or *surprises*, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable.
- **Evolution and deviation analysis:** Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values. It is common that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity an important attribute of a data mining system.

14 Appendix H – List of “Must-Read” Material

1. On May 10, 2004, the US National Office for Integrated and Sustained Ocean Observations (<http://www.ocean.us>) published an implementation plan for the Data Management and Communications (DMAC) subsystem of the Integrated Ocean Observing System (IOOS) (see http://dmac.ocean.us/dacsc/imp_plan.jsp and Appendix D, where the table of contents of this report has been reproduced). The document defines the core functions of the DMAC.
2. The International Oceanographic Data and Information Exchange (IODE) Resource Kit is a comprehensive self-training and resource tool for newly established Oceanographic Data Centres, designed to assist managers and staff members to acquire the skills to set up and run IODE centers. It contains a range of marine data-management and information-management materials, including software, quality control and analysis strategies, training manuals, and relevant IOC documents. The Resource Kit provides a broad spectrum of background information on global data and information archiving activities, specifications for data storage in standard formats, and the software tools to perform many quality-control, subsetting, and analysis procedures. The Resource Kit is accompanied by Training Manuals and regional data sets used in training workshops. A good general reference guide regarding marine data management of multi-disciplinary data sets has been documented in <http://ioc.unesco.org/oceanteacher/resourcekit/index.htm> with some good representative sample projects <http://ioc.unesco.org/oceanteacher/Data/data.htm> used for illustration.
3. The CCSDS (Consultative Committee for Space Data Systems) has proposed the OAIS (Open Archival Information System) Reference Model (<http://www.ccsds.org/docu/dscgi/ds.py/Get/File-143/650x0b1.pdf>), available through <http://arc.nesdis.noaa.gov>. This is now an ISO standard for long-term archives, and is being proposed as a potential model for the NESDIS long-term archive. In addition to addressing documentation and quality assurance issues, it addresses submission agreements between the submitter and the archive.
4. VENUS/NEPTUNE should investigate the OGC Sensor Collection Service (used by Environment Canada – “websensor”). *Background:* Owsview is a Web-based thin client which supports discovery, access and visualization of supported specifications of the OpenGIS Consortium (OGC) and Canadian Geospatial Data Infrastructure (CGDI) endorsed specifications, such as Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), Styled Layer Descriptor (SLD), Web Registry Service (WRS), Web Map Context Documents, Catalog Service, Gazetteer Service, Sensor Collection Service (SCS) and the GeoConnections Discovery Portal API (searching for products and services). Owsview also exemplifies the benefits of standards based services for chaining between services for discovery, access and visualization of information holdings.

15 Appendix I – Quality Management Issues for the Operation of the DMAS and Data QA/QC

Quality is the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs. The term “quality” should not be used as a single term to express a degree of excellence in a comparative sense, nor should it be used in a quantitative sense for technical evaluations. To express these meanings, a qualifying adjective should be used.

The achievement of satisfactory quality involves all stages of the quality loop as a whole. The contributions to quality of these various stages are sometimes identified separately for emphasis: for example, quality due to definition of needs, quality due to product design, quality due to conformance, and quality due to product support throughout its lifetime. In some references, quality is referred to as “fitness for use” or “fitness for purpose” or “customer satisfaction” or “conformance to the requirements.” These definitions represent only certain facets of quality.

What are the differences between the concepts of quality control, quality assurance, quality management, quality planning, and total quality management?

Quality management is all activities of the overall management function that determine the quality policy, objectives, and responsibilities, and implement them by means such as quality planning, quality control, quality assurance, and quality improvement within the quality system.

Quality planning is the activities that establish the objectives and requirements for quality and for the application of quality system elements. Quality planning covers product planning, managerial and operational planning, and the preparation of quality plans.

Quality control is operational techniques and activities that are used to fulfil requirements for quality. It involves techniques that monitor a process and eliminate causes of unsatisfactory performance at all stages of the quality loop.

Quality assurance is the planned and systematic activities implemented within the quality system and demonstrated as needed to provide adequate confidence that an entity will fulfil requirements for quality.

A quality system is the organizational structure, procedures, processes, and resources needed to implement quality management.

Total quality management is the management approach of an organization, centered on quality, based on the participation of all of its members, and aiming at long-term success through customer satisfaction and benefits to all members of the organization and to society.

(Source: ANSI/ISO/ASQ A8402-1994. Quality Management and Quality Assurance—
Vocabulary)

16 Appendix J – Glossary of Computer Terminology

Access	See Microsoft Access .
ASCII	ASCII stands for “American Standard Code for Information Interchange,” pronounced “as-key.” ASCII files or “plain text format” files consist entirely of text (letters, numbers, punctuation) that do not have any special formatting. The key points about ASCII format files are that they can be easily viewed (with any text editor or word processing package), but they take up more space than other encoding schemes. The opposite of ASCII is binary .
AutoDRM	AutoDRM stands for “Automatic Data Request Manager.” AutoDRM is a software package, which when installed on a server , allows anyone with access to electronic mail to retrieve data from that server. AutoDRM is a widely used method to retrieve earthquake information (including waveform data) from seismological observatories.
Binary format	Any file format for digital data encoded as a series of bits. Usually numbers are stored in their internal representation rather than using a character-based encoding such as ASCII . The key points about binary format files are that they take up less space than an ASCII file, but are harder to read, as they require an external program to translate them into ASCII form.
Blade server	A blade server is a thin, modular electronic circuit board, containing one, two, or more microprocessors and memory, that is intended for a single, dedicated application (such as serving Web pages) and that can be easily inserted into a space-saving rack with many similar servers. Blade servers, which share a common high-speed bus, are designed to create less heat and thus save energy costs as well as space. Large data centers and Internet service providers (ISPs) that host Web sites are among companies most likely to buy blade servers. A blade server is sometimes referred to as a high-density server and is typically used in a clustering of servers that are dedicated to a single task, such as file sharing or serving Web pages. Like most clustering applications, blade servers can also be managed to include load balancing and redundancy (failover) capabilities. A blade server usually comes with an operating system and the application program to which it is dedicated

	already on the board.
BUFR	The World Meteorological FM-94 BUFR (Binary Universal Form for the Representation of Meteorological data) is a binary code designed to represent, employing a continuous binary stream, any meteorological data. It has been designed to achieve efficient exchange and storage of meteorological and oceanographic data. It is a self-defining, table driven and very flexible data representation system, especially for huge volumes of data.
C	C is a structured, procedural programming language that has been widely used both for operating systems and applications and that has had a wide following in the academic community. Many versions of Unix-based operating systems are written in C. With the increasing popularity of object-oriented programming, C is being rapidly replaced as “the” programming language by C++ , a superset of the C language that uses an entirely different set of programming concepts, and by Java , a language similar to but simpler than C++, that was designed for use in distributed networks.
C++	C++ is an object-oriented programming (OOP) language that is viewed by many as the best language for creating large-scale applications. C++ is a superset of the C language.
CA OpenINGRES	Ingres is a relational database management system (RDBMS) that evolved from a research project at the University of California at Berkeley in the 1970s. There are two different versions of Ingres: a public domain version, known as University Ingres or Berkeley Ingres; and a commercial version currently marketed by Computer Associates, and known as OpenIngres, CA-OpenIngres, or Ingres II. The commercial version of Ingres II runs on the following operating systems: Windows NT, the majority of Unix platforms, OpenVMS, and Linux. Like many of its RDBMS competitors, Ingres has begun to add object-oriented development features to address the growing paradigm shift in the RDBMS marketplace towards more object-oriented database management systems (OODBMSs).
Client	A software program that is used to contact and obtain data from a server software program on another computer, often across a great distance. Each client program is designed to work with one or more specific kinds of server programs, and each server requires a specific kind of client. A Web browser is a specific kind of client.
Client-server	A common form of distributed system in which software is split between server tasks and client tasks. A client sends requests to a server , according to some protocol, asking for

	information or action, and the server responds. There may be either one centralized server or several distributed ones. This model allows clients and servers to be placed independently on nodes in a network, possibly on different hardware and operating systems appropriate to their function, e.g. fast server/cheap client.
Computer Cluster	A computer cluster is a group of connected computers that work together as a unit. One of the more popular implementations is a cluster with nodes running Linux as the OS and Beowulf software (both free) to implement the parallelism. Two primary reasons for clustering computers are fault-tolerance (where two computers mirror one another) and higher performance (where several computers work independently on a split-up task). Clustering can provide significant performance benefits versus price. The System X supercomputer at Virginia Tech, the third most powerful supercomputer on Earth as of November 2003, is a computer cluster of 1100 Apple Power Macintosh G5s running Mac OS X. The total cost of the system is \$5.2 million, a tenth of the cost of slower mainframe supercomputers. Clusters were originally developed by DEC in the 1980s. They not only support parallel computing, but also shared file systems and peripheral devices. They are supposed to give you the advantage of parallel processing while maintaining data reliability and uniqueness. A cluster of computers is sometimes referred to as a server farm.
CORBA	CORBA stands for “Common Object Request Broker Architecture”, an Object Management Group (OMG) specification which provides the standard interface definition between OMG-compliant objects. (In 1989, the OMG consortium, which included IBM Corporation, Apple Computer Inc. and Sun Microsystems Inc., formed to create a cross-compatible distributed object standard. The goal was a common binary object with methods and data that work using all types of development environments on all types of platforms – the standard created was CORBA.)
Database extension	An extension (e.g., DataBlase, Extender) is a package defining

	new SQL data types and operations on those data types. Standard SQL, for example, includes an “integer” data type and integer operations such as “+”, “max”, “average”, etc. An (spatial) extension to SQL might include a “polygon” data type and operations such as “area”, “intersects”, and “union.”
DB2	DB2 is a family of relational database management system (RDBMS) products from IBM that serve a number of different operating system platforms. According to IBM, DB2 leads in terms of database market share and performance. Although DB2 products are offered for Unix-based systems and personal computer operating systems, DB2 trails Oracle’s database products in UNIX-based systems and Microsoft’s Access in Windows systems. In addition to its offerings for mainframe operating systems, IBM offers DB2 products for a cross-platform spectrum that includes UNIX-based Linux, HP-UX, Sun Solaris, and SCO UnixWare; and for its personal computer OS/2 operating system as well as for Microsoft’s Windows 2000 and earlier systems. DB2 databases can be accessed from any application program by using Microsoft’s Open Database Connectivity (ODBC) interface, the Java Database Connectivity (JDBC) interface, or a CORBA interface broker.
Fat Client	See Thin client – fat client
Ferret	Ferret is an interactive computer visualization and analysis environment designed to meet the needs of oceanographers and meteorologists analyzing large and complex gridded data sets. It runs on most Unix systems and on Windows XP/NT/9x. It can also be installed to run from a Web browser (“WebFerret”). Ferret was developed by the Thermal Modeling and Analysis Project (TMAP) at PMEL in Seattle to analyze the outputs of its numerical ocean models and compare them with gridded, observational data. The model data sets are generally multi-gigabyte in size with mixed 3- and 4-dimensional variables defined on staggered grids.
Flat file	A single file containing ASCII characters representing or encoding some structure, e.g. of a database, tree, or network. Flat files can be processed with general purpose tools (such as text editors) but are often less efficient than some kind of binary file if they must be parsed repeatedly by a program. Flat files are more portable between different operating systems and application programs than binary files.
Fortran	FORTRAN (FORMula TRANslation) is a third-generation (3GL) programming language that was designed for use by engineers, mathematicians, and other users and creators of scientific algorithms. It has a very succinct and spartan syntax.

FTP	File Transfer Protocol (FTP), a standard Internet protocol, is the simplest way to exchange files between computers on the Internet. Like the Hypertext Transfer Protocol (HTTP), which transfers displayable Web pages and related files, and the Simple Mail Transfer Protocol (SMTP), which transfers email, FTP is an application protocol that uses the Internet's TCP/IP protocols. FTP is commonly used to transfer Web page files from their creator to the computer that acts as their server for everyone on the Internet. It's also commonly used to download programs and other files to your computer from other servers. As a user, you can use FTP with a simple command line interface (for example, from the Windows MS-DOS Prompt window) or with a commercial program that offers a graphical user interface. Your Web browser can also make FTP requests to download programs you select from a Web page. Using FTP, you can also update (delete, rename, move, and copy) files at a server. You need to logon to an FTP server. However, publicly available files are easily accessed using anonymous FTP. Basic FTP support is usually provided as part of a suite of programs that come with TCP/IP. However, any FTP client program with a graphical user interface usually must be downloaded from the company that makes it.
GML	GML stands for "Geographic Markup Language" - a dialect of XML designed for geographic data. GML is an emerging international standard; the current version (v2.0) of the specification formally bears the status of "adopted specification" within the Open Geospatial Consortium (OGC); this is the top rung of the OGC specification ladder. Basically the designation implies that GML is mature enough to be used in other implementation activities or incorporated into software products.
GRIB	GRidded Binary, a bit-oriented format approved by the World Meteorological Organization that has compaction features for efficiently transmitting and storing large volumes of gridded data. GRIB is an alternative to NetCDF.
HDF	The National Center for Supercomputing Application's "Hierarchical Data Format." This is a flexible, standard, public-domain file format designed for sharing graphical and floating point data among different programs and machines.
HTTP	HTTP (Hypertext Transfer Protocol) is the set of rules for transferring files (text, graphic images, sound, video, and other multimedia files) on the World Wide Web. As soon as a Web

	user opens their Web browser, the user is indirectly making use of HTTP. HTTP is an application protocol that runs on top of the TCP/IP suite of protocols (the foundation protocols for the Internet). HTTP concepts include (as the Hypertext part of the name implies) the idea that files can contain references to other files whose selection will elicit additional transfer requests. Any Web server machine contains, in addition to the Web page files it can serve, an HTTP daemon, a program that is designed to wait for HTTP requests and handle them when they arrive. Your Web browser is an HTTP client, sending requests to server machines. When the browser user enters file requests by either “opening” a Web file (typing in a Uniform Resource Locator or URL) or clicking on a hypertext link, the browser builds an HTTP request and sends it to the Internet Protocol address (IP address) indicated by the URL. The HTTP daemon in the destination server machine receives the request and sends back the requested file or files associated with the request. (A Web page often consists of more than one file.)
IDD	Unidata Internet Data Distribution. The IDD allows users to “subscribe” to certain sets of data products; IDD servers then deliver the requested data to their local servers as soon as they are available from the source. With the initial national implementation in 1994, the IDD may have been the original example of Internet “push” technology. It now appears to provide the reliability, flexibility, and efficiency required by participating institutions. In the future, Unidata plans to augment the system to better serve disciplines outside the atmospheric sciences disciplines and to incorporate anticipated new networking technologies to minimize the impact of the IDD on the underlying network.
IDL	IDL (Interactive Data Language), a product of Research Systems Inc. (RSI) is a language for creating visualizations based on scientific or other data. IDL is a complete package for the interactive reduction, analysis, and visualization of scientific data and images. Optimized for the workstation environment, IDL integrates a responsive array oriented language with numerous data analysis methods and an extensive variety of two- and three-dimensional displays into a powerful tool for researchers. IDL supports an extensive data import capability, publication quality hard copy output, and user-defined graphical user interfaces. IDL users can create complex visualizations in hours instead of weeks with the aid of IDL’s high level capabilities and interactive environment. IDL is useful in physics, astronomy, image and signal

	processing, mapping, medical imaging, statistics, and other technical disciplines requiring visualization of large amounts of data.
In situ data	In situ data are produced by observations and measurements made directly at the site or the phenomenon, rather than remotely. Argo floats, for example, produce in situ data. Satellite-derived sea surface heights and temperatures are examples on non-in situ data.
Informix	The term Informix refers to a relational database management system, and for about 15 years also referred to the company which developed it. The Informix DBMS developed from the pioneering Ingres system that also led to Sybase and SQL Server. For a time in the 1990s Informix was the second most popular database system, after Oracle. In 2001 IBM purchased Informix in order to gain access to its existing market share and customer base. The Informix products have a loyal following, as they tend to perform well, are easy to manage, and are relatively easy to develop extensions for. IBM appears committed to continuing to develop and support Informix releases.
Java	Java is a programming language expressly designed for use in the distributed environment of the Internet. It was designed to have the “look and feel” of the C++ language, but it is simpler to use than C++ and enforces an object-oriented programming model. Java can be used to create complete applications that may run on a single computer or be distributed among servers and clients in a network. It can also be used to build a small application module or applet for use as part of a Web page. Applets make it possible for a Web page user to interact with the page. The major characteristics of Java are: 1. The programs you create are portable in a network. Source programs are compiled into what Java calls bytecode, which can be run anywhere in a network on a server or client that has a Java virtual machine. The Java virtual machine interprets the bytecode into code that will run on the real computer hardware. This means that individual computer platform differences such as instruction lengths can be recognized and accommodated locally just as the program is being executed. With Java, platform-specific versions of programs are no longer needed. 2. The code is robust, meaning that, unlike programs written in C++ and perhaps some other languages, the Java objects

	can contain no references to data external to themselves or other known objects. This ensures that an instruction can not contain the address of data storage in another application or in the operating system itself, either of which would cause the program and perhaps the operating system itself to terminate or “crash.” The Java virtual machine makes a number of checks on each object to ensure integrity. 3. Java is object-oriented, which means that, among other characteristics, an object can take advantage of being part of a class of objects and inherit code that is common to the class. Objects are thought of as “nouns” that a user might relate to rather than the traditional procedural “verbs.” A method can be thought of as one of the object’s capabilities or behaviors. 4. In addition to being executed at the client rather than the server, a Java applet has other characteristics designed to make it run fast. 5. Relative to C++, Java is easier to learn. Almost all major operating system developers (IBM, Microsoft, and others) have added Java compilers as part of their product offerings.
JDBC	Java Database Connectivity (JDBC) is an application program interface (API) specification for connecting programs written in Java to the data in popular databases.
Matlab	MATLAB, short for “Matrix Laboratory”, is an interactive program from The MathWorks for high-performance numeric computation and visualisation. MATLAB integrates numerical analysis, matrix computation, signal processing, and graphics in an easy-to-use environment. MATLAB is built on sophisticated matrix software for analyzing linear equations. The tools supplied can be used for applied mathematics, physics, chemistry, engineering, finance and other areas dealing with complex numerical calculations. It is used by more than 1,000,000 people in industry and academia and runs on most modern operating systems, including Windows, MacOS, Linux and Unix.
Metadata	The information necessary for someone who is not previously acquainted with a data set to make full and accurate use of that data set.
Microsoft Access	A Microsoft software product that is primarily a data management tool (database software). Access has tools to enter, edit, and index data and to retrieve it via custom forms and reports. Microsoft Access runs on Windows 9x/NT/XP desktop systems.

Microsoft SQL Server	A relational database management system (RDBMS) which is part of Microsoft's BackOffice family of servers. SQL Server was designed for client/server use and is accessed by applications using SQL. It runs on Windows NT version 3.5 or higher.
Middleware	The term "Middleware" is used to describe a software agent acting as an intermediary, or as a member of a group of intermediaries, between different components in a transactional process. The classic example of this is the separation that is attained between the client user and the database in a client/server situation. The reason for introducing middleware in such a situation is to better service client requests by reducing the number of resource-expensive connections to the database and more efficiently passing the requested data back.
MySQL	MySQL (pronounced "my ess cue el," not "my sequel") is an open source relational database management system (RDBMS) that uses Structured Query Language (SQL), the most popular language for adding, accessing, and processing data in a database. Because it is open source, anyone can download MySQL and tailor it to their needs in accordance with the general public license. MySQL is noted mainly for its speed, reliability, and flexibility. Most agree, however, that it works best when managing content and not executing transactions. The MySQL relational database system was first released in January, 1998. It is fully multi-threaded using kernel threads, provides application program interfaces (APIs) for C, C++, Java, and many other languages. MySQL currently runs on the Linux, Unix, and Windows platforms. Many Internet startups have been especially interested in MySQL as an alternative to the proprietary database systems from Oracle, IBM, and Informix.
NetCDF	"Network Common Data Format". A self describing, platform independent binary data format (created by UNIDATA).
Object-relational database	A relational database that has been extended to support additional data types and operations. See Database extension .
OGC	The Open Geospatial Consortium, Inc. (OGC, formerly known as the Open GIS Consortium) is a non-profit, international, voluntary consensus standards organization that is leading the development of standards for geospatial and location based services. Through their member-driven consensus programs, OGC works with government, private industry, and academia to create open and extensible software application

	programming interfaces for geographic information systems (GIS) and other mainstream technologies. The Web Coverage Services Standard is one such interface standard.
OPeNDAP	“Open source Project for a Network Data Access Protocol”. A data transport protocol that provides a means for data analysis and visualization packages such as Matlab , IDL and Ferret to access files over a network (local or Internet) without needing to know how the data are stored at the server site.
Oracle	Oracle is the leading (in terms of market share) relational database management system for Unix-based servers.
Plain text format	See ASCII .
PostgreSQL	PostgreSQL is a free software object-relational database server (database management system). It offers an alternative to other open-source database systems as well as to proprietary systems such as Oracle , Sybase , IBM’s DB2 and Microsoft SQL Server .
Push-pull	Push is defined as the technology that puts pre-selected content directly on your computer screen from the Internet without your need to browse for it. With this technology, introduced by PointCast and Individual, Inc. and added to 4th generation browsers, you can program your desktop, for example, to automatically receive such things as local weather, news headlines, selected stock reports, and sports scores for selected teams or events. Metcast is an example of a “Push” technology. The alternative to Push technology is “Pull.” “Pull” refers to requesting data from another program or computer. The World Wide Web is based on pull technologies, where a page isn’t delivered until a browser requests it.
RAID	RAID (redundant array of independent disks; originally redundant array of inexpensive disks) is a way of storing the same data in different places (thus, redundantly) on multiple hard disks. By placing data on multiple disks, I/O operations can overlap in a balanced way, improving performance. Since multiple disks increases the mean time between failure (MTBF), storing data redundantly also increases fault-tolerance.
Relational database	A relational database is a set of tables containing data fitted into predefined categories. Each table (which is sometimes called a relation) contains one or more data categories in columns. Each row contains a unique instance of data for the categories defined by the columns. In addition to being relatively easy to create and access, a relational database has the important advantage of being easy

	to extend. After the original database creation, new columns and tables can be added without requiring that all existing applications be modified.
Role-based access control	The earliest forms of access control systems assigned privileges to users. These early access control systems allowed the system administrator to enable defined privileges for specific users. The addition of user groups improved that situation. With groups, the system administrator could assign privileges to groups such as Sales or Accounting and add users into those groups. Role Based Access Control (RBAC) is the next evolutionary step in access control. RBAC enables privileges to be assigned to arbitrary roles. Those roles can then be assigned to real users. This provides more granular control of privileges, which enhances system security. In addition, it reduces the amount of administrative effort required to add or delete system users.
Server	A computer, or a software application that provides a specific kind of service to client software running on other computers. The term can refer to a particular piece of software, such as a Web Server, or to the computer on which the software is running. A single computer may have several different server software applications running on it, thus providing many different servers to clients on a network.
Single sign-on	An authentication process in a client/server relationship where the user, or client, can enter one name and password and have access to more than one application or access to a number of resources within an enterprise. Single sign-on takes away the need for the user to enter further authentications when switching from one application to another. Although single sign-on makes the login process more convenient for the user, it does mean that the password becomes more valuable to a hacker because of the large number of systems it can access. For this reason some consultants discourage the use of single sign-on systems, and, where there is no other realistic option, recommend that passwords are guarded safely and changed regularly.
SQL Server	See Microsoft SQL Server .
Sybase	Sybase is a computer software company that develops and sells database management systems (DBMS) and middleware . Sybase products have found extensive application, particularly in commercial, industrial, and military communications

	systems.
Thin client – fat client	In client/server applications, a “thin client” is a client designed to be especially small so that the bulk of the data processing occurs on the server. A “fat client”, in contrast, is a client where the bulk of the data processing logic resides on the client rather than the server. Ferret is an example of a fat client; the Web interface for TAO data is an example of a thin client.
THREDDS	Thematic Realtime Environmental Distributed Data Services. THREDDS primary technological focus is on building middleware that will facilitate the interoperability of data provider systems and the data analysis and display systems used by the communities working with the data sets. These systems are largely built according to standard protocol specifications and tools for implementing those protocols. These include data and metadata server packages, server side catalog tools client side access and implementation widgets, as well as XML transport mechanisms for: • Data provision and access • Metadata generation and access • Standards-based Web services.
VAX	VAX (Virtual Address eXtension) is an established line of mid-range server computers from the Digital Equipment Corporation (DEC). It followed DEC’s PDP-11 in 1978 and also introduced a new operating system, VMS. VAX included a 32-bit processor and virtual memory.
Web Coverage Services standard	The Web Coverage Services Standard (WCS) supports electronic interchange of geospatial data as “coverages” – that is, digital geospatial information representing space-varying phenomena. A WCS provides access to potentially detailed and rich sets of geospatial information, in forms that are useful for client-side rendering, multi-valued coverages, and input into scientific models and other clients. The WCS may be compared to the OGC Web Map Service (WMS) and the Web Feature Service (WFS); like them it allows clients to choose portions of a server’s information holdings based on spatial constraints and other criteria. Unlike WMS, which filters and portrays spatial data to return static maps (rendered as pictures by the server), the Web Coverage Service: • provides available data together with their detailed descriptions; • allows complex queries against these data; and • returns data with its original semantics (instead of pictures) which can be interpreted, extrapolated, etc. – and not just portrayed.

	Unlike WFS, which returns discrete geospatial features, the Web Coverage Service returns representations of space-varying phenomena that relate a Spatio-temporal domain to a (possibly multidimensional) range of properties.
Web portal	Portal is a term, generally synonymous with gateway, for a World Wide Web site that is or proposes to be a major starting site for users when they get connected to the Web or that users tend to visit as an anchor site . There are general portals and specialized or niche portals. Some major general portals include Yahoo, Excite, Netscape, Lycos, CNET, Microsoft Network, and America Online's AOL.com. Examples of niche portals include Garden.com (for gardeners), Fool.com (for investors), and SearchNetworking.com (for network administrators). A number of large access providers offer portals to the Web for their own users. Most portals have adopted the Yahoo style of content categories with a text-intensive, faster loading page that visitors will find easy to use and to return to. Companies with portal sites have attracted much stock market investor interest because portals are viewed as able to command large audiences and numbers of advertising viewers. Typical services offered by portal sites include a directory of Web sites, a facility to search for other sites, news, weather information, email, stock quotes, phone and map information, and sometimes a community forum. Excite is among the first portals to offer users the ability to create a site that is personalized for individual interests.
XML	XML (Extensible Markup Language) is a flexible way to create common information formats and share both the format and the data on the World Wide Web, intranets, and elsewhere. For example, computer makers might agree on a standard or common way to describe the information about a computer product (processor speed, memory size, and so forth) and then describe the product information format with XML. Such a standard way of describing data would enable a user to send an intelligent agent (a program) to each computer maker's Web site, gather data, and then make a valid comparison. XML can be used by any individual or group of individuals or companies that wants to share information in a consistent way. XML, a formal recommendation from the World Wide Web Consortium (W3C), is similar to the language of today's Web

	pages, the Hypertext Markup Language (HTML). Both XML and HTML contain markup symbols to describe the contents of a page or file. HTML, however, describes the content of a Web page (mainly text and graphic images) only in terms of how it is to be displayed and interacted with. For example, the letter “p” placed within markup tags starts a new paragraph. XML describes the content in terms of what data are being described. For example, the word “phonenum” placed within markup tags could indicate that the data that followed was a phone number. This means that an XML file can be processed purely as data by a program or it can be stored with similar data on another computer or, like an HTML file, that it can be displayed. For example, depending on how the application in the receiving computer wanted to handle the phone number, it could be stored, displayed, or dialed. XML is “extensible” because, unlike HTML, the markup symbols are unlimited and self-defining. XML is actually a simpler and easier-to-use subset of the Standard Generalized Markup Language (SGML), the standard for how to create a document structure. It is expected that HTML and XML will be used together in many Web applications. XML markup, for example, may appear within an HTML page.
--	---