

DMAS Report  
Nov 22

This report was largely based on notes taken by Pierre Quesnel with additional comments made by Bruce Spencer and Bill St Arnaud

**DMAS Advisory Committee Meeting Members**

**Present  
(Advisory Board)**

Bill St. Arnaud, Chief Research Officer, Canarie  
Robin Brown- Manager, Ocean Science Division, Institute of Ocean Sciences  
Bruce Spencer - Research Officer, Internet Logic, NRC Institute for Information Technology  
Pierre Quesnel - Head, Systems and Networks Management, NRC Canada Institute for Scientific and Technical Information

**Those present from UVic  
(UVic)**

Martin Taylor - President & CEO, Ocean Networks Canada  
Chris Barnes – Neptune Project Director  
Benoît Pirenne – DMAS Associate Director, Information Technology  
Mairi Best - Associate Director, Science  
Verena Tunnicliffe - VENUS Project Director  
Eric Guillemot - Manager, Software Applications  
Murray Leslie - Data Quality Control Specialist  
Nic Scott – NEPTUNE Systems Support Team  
Adrian Round – VENUS Project Manager

Overall Comments:

The Committee was tasked with a series of questions on the system availability, file management system, DMAS commercialization and communal work environment for science. Overall the committee felt that DMAS has taken appropriate steps in terms of the system availability and reliability with the only additional suggestion that backup systems should be located somewhere else in the country other than Vancouver Island.

The committee felt that direct commercialization of the DMAS raised serious concerns of practicality and the fact that attempts at commercialization may distract the DMAS team from its primary mission. However developing ocean cable research communal science tools that could be used on clouds services like Amazon might produce an indirect revenue stream when such tools were used by third parties where NEPTUNE/ONC could capture a percentage of those revenue. Using cloud services like Amazon might also address questions of scalability, file management and communal science platforms. The committee recommended that DMAS team explore platforms such as MyExperiment.org and Google Social amongst others.

The 4 specific questions and comments are discussed in more detail as follows:

**Discussion Theme One: System High Availability**

- Which High Availability environment/technology can be recommended?
- What secondary storage agreement should we enter into and with whom?

- Any advice to make those network backhaul lines more reliable? They are currently single points of failure.

The committee was asked to report and give suggestions for future directions of the HA systems. Pierre notes that CISTI has issues for HA as well and suggested it is useful to have a snapshot capability to bring back data quickly.

The committee suggested that the Data Repository not be located on the West Coast, and was assured that plans were being made to warehouse outside of the possible earthquake zone.

The Sun Systems architecture is sound and the committee suggested to stay with VMWare/Sun. As the volume grows, it is important to beef up the Web Server. Web crawlers may crawl the site and if the site is large this activity might impact performance of the servers, so this should be balanced. One option is to limit the number of connections for web crawlers at the OS level. This was already noticed for the Google crawler. It was noted that the such crawlers are not to be prevented from their task.

The 64 bit tomcat implementation across multiple machines could be considered for the Web Server implementation.

The big irons and the small commodity machines are two options for providing HA. Virtualization of servers provides the ability to be agnostic about the physical resources. The Sun machines provide the basic power for the DMAS. Most DMAS applications do not need large processing power, although in future video analysis may be done.

The data that is available now is that collected for VENUS. Access is event-driven set up through triggers, and the DBMS uses caching. The feedback that the users want is to know how their job is doing, how large the result will be, how long will it take.

The general consensus is to spend money at the shore station. This station must be very reliable. The backhaul is second priority.

### **Discussion Theme Two: A new file management system for DMAS?**

Which file system will serve best the growing needs of the ONC for storage, processing and search capabilities?

The Globus file system and Hadoop prototype were discussed. Hadoop is an open source clone of the Google file system.

Amazon offers a hosting service, EC2, which can be leased on an hourly basis and offers options for large storage.

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. [ <http://labs.google.com/papers/mapreduce.html> ]

MapReduce can be run on EC2 using Hadoop.

Unfortunately the WestGrid does not share processors cycles and cannot access other networks, so it does not seem suitable for DMAS.

As much as possible the committee recommends use of external file management systems support by Amazon and other services, especially for access by researchers and users outside of Canada.

### **Discussion Theme Three: DMAS Commercialization**

- What compelling features would make a commercial/paid support model successful?
- Do we need to limit ourselves to data management systems for ocean science?
- How do we develop documentation, online help and toll-free support lines?
- How do we exploit the time-to-market advantage we have over the competition (e.g., OOI)?
- Should we pursue all alternatives in parallel?

Benoit: The palette is very wide for DMAS commercialization. It is an industry-ready sensor-based data-acquisition system. We can generalize its use to other environments: medicine, security, engineering, big science.

The Committee suggests that commercialization be explored very carefully. IBM and HP are all doing this and have the services of sales people with contacts. This effort would pull you away from your core mandate. The ability to sell even to the European and US ocean-based groups is low; they will do their own thing. There are a lot of similar products out there. The committee recommends that UVic determine what the offshore oil and gas exploration companies are using and see if there is an international market place for coastal ocean observation systems?

The committee believes there are better ways to leverage your database and exploit your intellectual property. Use Google AdSense ads on your website. Develop tools that can be used on Amazon. You generate some income and build upon Amazon's powerful toolsets.

The Partnership model is moving down a track where NEPTUNE is the new generation of wiring the oceans.

Whatever you do in commercialization, don't lose focus. You are looking at the big problems that affect the planet. Your justification is (1) public policy (2) Ecological development with spin off benefits for the marine SME's (3) public education and outreach. Can think of each of these as fulfilling a need and therefore a potential market. Perhaps the best fit for DMAS is (3).

DMAS might be better off to make tools available and let the community be supportive and develop the next layers. Use the open source model to bring the tools to the next level. Open source some of the components of DMAS. But make arrangements with Amazon and other virtualization and cloud suppliers to earn revenue when compute cycles are used in support of these open source applications

Another potential is to form partnerships with large companies, such as Alcatel/Lucent when they lay cable. Another example is partnership with instrument manufacturer.

#### **Discussion Theme Four: A communal work environments for science**

The project is relatively ambitious and we would like to hear from the Advisory Committee on the following points:

- Which technologies are available today to implement a social networking environment in which scientific processing is to take place?
- Which data processing middleware should be considered that provides transparent access to Grid processing environments?
- Which workflow tools should be leveraged that supports well the execution of procedural tasks expected from scientists?
- Which data processing environment should be made available as a default system for users to run calculations and scripts? Matlab? R? ...
- Advice on adopting/building upon community standards such as those provided in Canada by DFO?

This field is changing rapidly with the growth of popularity of social web sites like MySpace and FaceBook. Web 2.0 MashUps were considered. The suggestion was made to look at iGoogle and Google OpenSocial and MyExperiment.org:

- iGoogle supports the use of specially developed "gadgets" to display content on a user's page. The gadgets interact with the user and utilize the Google Gadgets API. The Google Gadgets API is public and allows anyone to develop a gadget for any need. [ <http://en.wikipedia.org/wiki/IGoogle>]
- OpenSocial provides a common set of APIs for social applications across multiple websites. With standard JavaScript and HTML, developers can create apps that access a social network's friends and update feeds. [ <http://code.google.com/apis/opensocial/>]

DMAS with its rich data source and interested community may be able to offer the scientific community the means to help each other to develop, share and enhance common apps and gadgets. In particular graduate and undergraduate students are interested in developing these kinds of applications. The Google Summer of Code program was used by CISTI last summer. [ <http://code.google.com/soc/2007/>]

If this can be an activity that is perceived to be led by scientists, it will have more traction. Tony Bailetti of Carleton University is leading an effort for open source development of communication-enabled applications. One important difference between that work and DMAS is that DMAS has access to content, which makes it more immediately useful.

Other topics were mentioned:

- Java Portlet: JSR 168: The specification defines a common Portlet API and infrastructure that provides facilities for personalization, presentation, and security. Portlets using this API and adhering to the specification will be product agnostic, and may be deployed to any portal product that conforms to the specification. [ [http://developers.sun.com/portalserver/reference/techart/jsr168/pb\\_whitepaper.pdf](http://developers.sun.com/portalserver/reference/techart/jsr168/pb_whitepaper.pdf)]

- My Yahoo is a customizable web page with news, stock quotes, weather, and many other features.
- The Dojo toolkit is a modular open source JavaScript toolkit (or library), designed to ease the rapid development of JavaScript- or Ajax-based applications and web sites.

A primary task is visualization. Thus a primary tool for the collaborative environment is the use of visualization tool kits, including

- MatLab.
- ManyEyes [ <http://services.alphaworks.ibm.com/manyeyes/> ] and
- LabView [ <http://en.wikipedia.org/wiki/LabVIEW> ] .

CISTI is also working on a similar sort of experimental application.