ICES CM 2010:A14

ICES CM 2009/A

Operational Data distribution at Institute of Marine Reasearch

Sjur Ringheim Lid, Helge Sagen, Trond Westgård

The Norwegian marine datacenter (NMD) started serving operational research data from vessels operated by the Institute of Marine Research in the early 2000. Through the EU FP7 project MyOcean NMD has become the thematic assembly center for Arctic in-situ data. As a thematic assembly center NMD deliver in-situ data to the global assembly center and to the ocean forecasting centers where it's used for assimilation or validation of model output. The data service has been expanded to include data from other data sources like the Coriolis data service. Real time quality control procedures have been defined in the MyOcean project and all data goes through these procedures and are flagged according to the SeaDataNet Quality flag scale.

The Institute of Marine Research has also started work to operationalise other data types gathered on research vessels. These data types include different kinds of biological samples and chemical data that will become available to scientists in near real time.

Keywords: Operational data distribution, MyOcean, SeaDataNet, automatic quality control

Contact author: Sjur Ringheim Lid, Institute of Marine Research, Norwegian Marine Datacentre P.O. Box 1870 Nordnes 5817 Bergen, Norway . Sjur.ringheim.lid@imr.no,

Introduction

Operational data gathering and distribution has been something the Institute of Marine Research has worked on since the early 21st century. As the ships operated by the institute got Internet connections we saw the possibility to make the data available to all scientists with a much shorter delay than earlier. This operational data distribution gave us some problems with regards to data quality and operational quality procedures had to be developed to give the scientists a minimum level of assurance in the data. The data made operationally available at the institute has been mostly focused on oceanographic data such as CTDs and Thermosalinographs, but we are now moving towards a much broader coverage of data types and are working towards the goal that in the future most data types will be available to all scientists shortly after collection.

Methodology

Figure 1 shows how operational data acquisition and distribution is done at IMR. The data is collected from multiple data sources, it is then put into various databases at IMR before real time quality control is performed

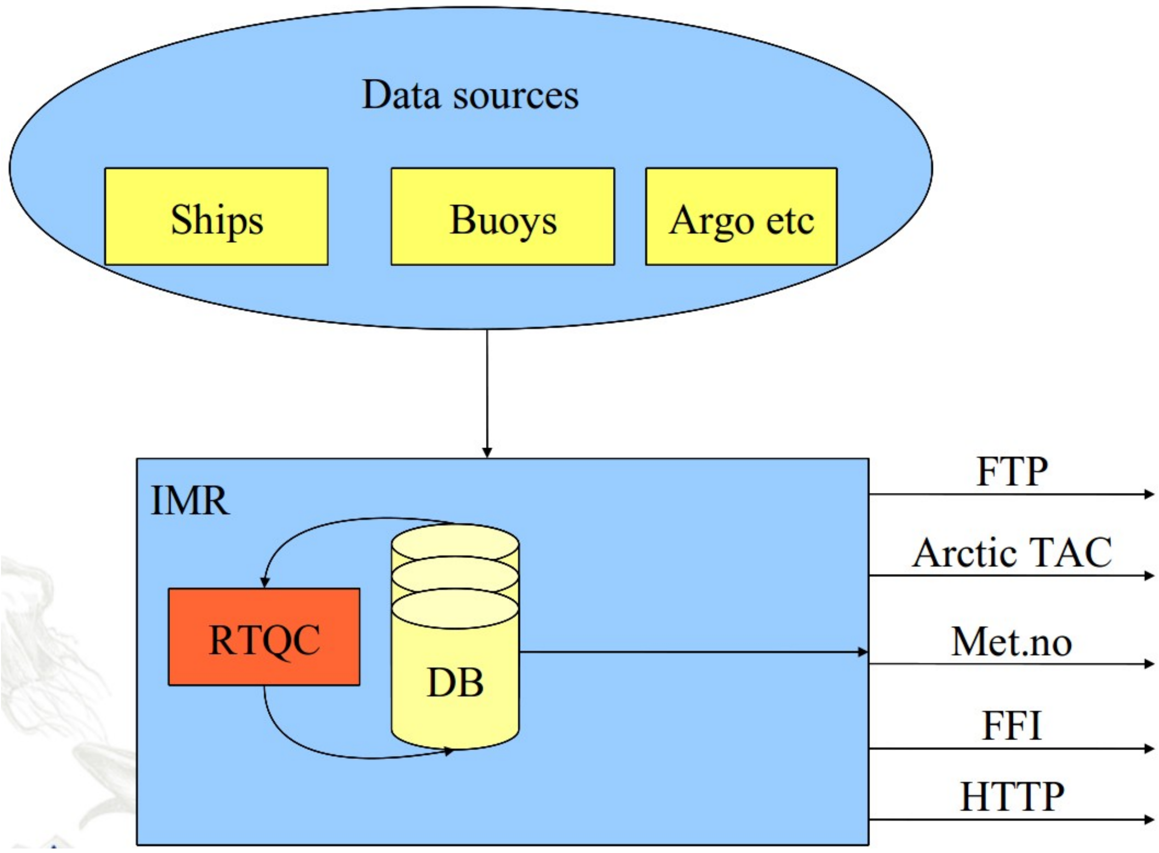

Figure 1: Operational data distribution at IMR

on the data. Data is then made available through multiple different sources. This will be explained in more detail in the rest of the document.

Acquisition

The Institute of Marine Research (IMR) with its four research ships collects large amounts of data each year that goes through our normal procedures of quality control checking and manual moving before they are made available to the users. As the computer technology has made huge leaps forward the last decades so has the collection of data, and it is now possible to make research data available in a much shorter amount of time. But it's still some way to go before all data are available to the user right after the collection has taken place. As the research ships operated by the institute got connected to the Internet the Norwegian Marine Datacentre (NMD) started working on developing software that could get the research data automatically copied into databases on land and give scientists access to these data. The software developed used the FTP protocol to collect the data from the ships and make them available in their original form on the internal network at the institute. An operational databases for oceanographic data has been created. This database were made so that it can contain data from all our oceanographic collection tools, such as CTD profilers, Thermosalinographs, drifting buoys like Argo buoys and stationary buoys. Making a database that could contain all these different types of data meant that the system easily can be extended to gather data from new sources in the future.

NMD has started data acquisition from other sources the later years and do now collect water level data and various data from the Coriolis data center like drifter data from Argo buoys.

Distribution

On top of the database an online tool for visualizing the collected data is made. The tool makes the user able to show the stations taken in a map and also make a graph of the different data profiles collected at a given station. The online tool also makes the user able to download data on the IMR developed file format called Toktfilformat.

Data is also distributed to both the Norwegian Defense Research Establishment (FFI) and The Norwegian Meteorological Institute (met.no). This distribution uses the FTP protocol to push data onto servers owned by the respective institutes.

Through the FP7 EU project MyOcean NMD has become the thematic assembly center for the Arctic. Distribution of data in this project is in the start done by setting up an FTP server at NMD where users are able to download data from. By the end of MyOcean delivery of data over the OpeNDAP protocol will also be available. The data is made available using the OceanSites NetCDF format for point and trajectory data.

Quality Control

Common quality control (QC) procedures for near real time data has been developed during the MyOcean project. These procedures has been implemented at NMD and is performed on all data made available in near real time mode. The QC procedures gives scientists a common assurance on the quality on data delivered from all TAC's in MyOcean. QC procedures is developed for multiple data types like temperature and salinity, currents, sea level and biogeochemical data. At this time only the temperature and salinity QC procedures are implemented at NMD as these are the only data types delivered from the Arctic TAC. The MyOcean QC procedures follows the SeaDatanet flagging scheme which makes the flagging compatible with data delivered through other channels.

Future of operational data at IMR

A large project focusing on streamlining data collection, quality control and data distribution is underway at IMR. The projects name is Sea2Data and its main goal is to make all types of data available to the users as soon as possible after collection with a quality that is as good as possible. In the project data type specific databases are created. This will make extending existing databases and addition of new databases easy in the future if that should be needed. The databases follows a set of rules for database design where a globally unique identifier (GUID) field and a original field are two important parts. These fields make synchronization of databases between different locations possible. With a GUID a new entry into any database can be made off site, so a database on board a ship can create a new record without there being a duplicate anywhere. The original field gives the possibility to make read only copies of databases and there will only be one database that can hold the original version at any time.

These databases will always have the original version at IMR. As shown in figure 2 when scientists go into the field to collect data copies of the relevant databases will be synchronized onto the computers used at the data collection site. These field versions of the databases containing new data can then be synchronized back to the original databases either after every data collection or after the data collection is finished. From the databases deployed at the research vessels data can be operationally copied to the master databases at IMR when data collection is done. This will make data available to scientists on shore in an even faster time then it is at present. Real time quality control will also be performed on all data collected using the new system that is developed.

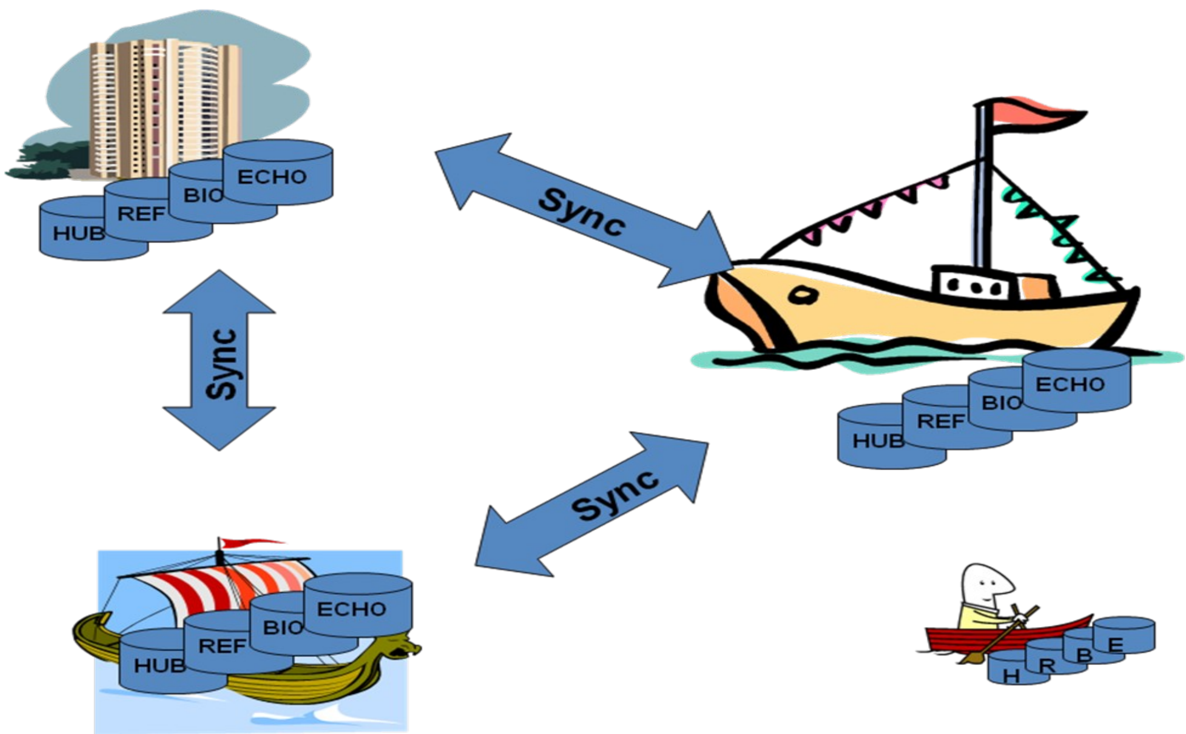

Figure 2: Synchronization of data between databases