The development of the data system and growth in data sharing

Lead author: Sylvie Pouliquen, Ifremer, Plouzané, France

Contributing authors
Steve Hankins NOAA/PMEL, Seattle, USA
Robert Keeley ISDM, OTTAWA,Canada
Jon Blower, UREADES, London, UK
C Donlon ESA,Netherlands ESA, Netherlands)
Alex Kozyr , CDIAC, USA
Robert Guralnick University of Colorado, USA

I. Context and setting the scene

A great wealth of data exists, for a wide range of disciplines, derived from in-situ and remote sensing observing platforms, in real-time, near-real-time and delayed mode. These data are acquired as part of routine monitoring activities and as part of scientific surveys by a few thousands institutes and agencies all around the world. Not only the means to acquire these data have changed in the past ten years but also the way to use them.

In past 10 years IT technology has progressed a lot. It presently allows data exchanges of a few giga-bytes via Internet in developed countries. In the late nineties it was still considered as high technology to provide data on CDROM and no more on tapes and only small datasets were distributed via Internet. Nowadays CDROM are considered as a backup delivery system especially for countries with poor internet connections. The explosion of the Internet has provided new tools and a new way of using computers and communication means

The nature of the requirements from government agencies have changed: they want to know (or may be guess) what the future of the earth will look like and what will be the impact on their piece of land/atmosphere (climate change issues, ocean health monitoring, fisheries assessment…) but they can't pay the full bill for the data acquisition. Therefore they are pushing and nowadays more and more imposing a change in data policy and move towards "data acquired with public funds should be available to the community".

Moreover the nature of science has changed. Investigators and research funding agencies are looking for context, impacts, and synthesis, rather than just focusing on individual, well-defined processes. Most scientists need the other data too, not just their own. Most scientists cannot do their work using only data they have collected themselves.

Finally the growth of operational oceanography activities aiming at providing services which are promising in term of downscaling activities , are really important. These are in demand by users especially for real/near real time data just as operational meteorology has been doing for a long time already.

In 1998 EuroGOOS issued "The science base of EuroGOOS" publication where some limitations related to data exchanges were highlighted. The situation has improved but the following statements are still relevant and should be seen as targeted actions not only in Europe but all around the world.

1. *Lack of international infrastructure for operational oceanographic data gathering, transmission, and products, (e.g. as adopted in World Weather Watch), and consequently lack of common standards*. It is still true in general. While the situation has improved a lot in the past 10 years in the physical domain. Experiences within GODAE, Argo and GHRSST programs have shown that it was possible to reach consensus on common standards (formats, real-time and delayed mode quality control, data distribution…). In some domains like biogeochemistry there is still a long way to go.
2. *Lack of clear right or duty to collect and transmit real-time data.* Once again in the past ten years we have seen the concept of "portals" emerging with the duty to serve in real-time the users: Salto/DUACS for altimetry, Medspiration/GHRSST for SST, Argo and Gosud Global Data centers, JcommOps are examples that exist nowadays. It has proven that sharing data rapidly was not scary for the scientists but

on the contrary beneficial as problems were detected more rapidly by comparison with nearby measurements and collaboration to set up appropriate observing system facilitated.

3. *Lack of proper design of a services infrastructure, using, for example, multiple data inputs such as wind, waves, and currents, to generate predictions of oil spill movements.* With the GMES initiative in Europe we have seen demonstration of the capability to build end-to-end services for users. Some projects like Mersea/MyOcean, Marcoast or PolarView are consolidating the systems that will need to be sustained in the future.
4. *Imbalance between monitoring (measurement) technology and capacity for post-processing data and subsequent real time use of numerical models*. Once again money and man power have been put for easing data access both at national and international level to ease access to homogeneous datasets. This is illustrated by the Coriolis project at the French level, or SeaDataNet IP that is funded by the European Union within GMES, DMAC in USA,…. This effort should be sustained in the future

In past decade progress has been made driven by applications such as Operational Oceanography but also by a fundamental change in scientist cultural behavior. This important change first started in satellite community where it was possible to use together data from missions managed by different countries (Altimetry from NASA and ESA, SST from most space agencies) but also in the in-situ world where Argo has been a pilot experience for the other JCOMM networks. . This paper will discuss the progress achieved in the past ten years., lesson,s learned and future needs. Some aspects will be detailed in the other OceanObs09 plenary and community white papers related to data infrastructure issues.

II. A relationship between providers and users

But what does sharing data really mean? In fact it's simply providing a way to potential users to access data that have been acquired and processed by a provider. In fact users don't want access to raw data they want access to information or products, elaborated upon the raw data, with a level of processing that depends on the category of users and the nature of their applications. The goal is then for the data or product providers to set up data services in order to serve these users. The needs are derived from main drivers that combined together allow to classify the way a user gains access to data and products. While operational users want secure access to well defined products, the general public needs easy to understand information and the research community is asking for access to as much data as possible to allow in depth studies. This close relationship between users and product providers is very important and allows tailoring the service to the user needs once the basis of information management systems have been set up.

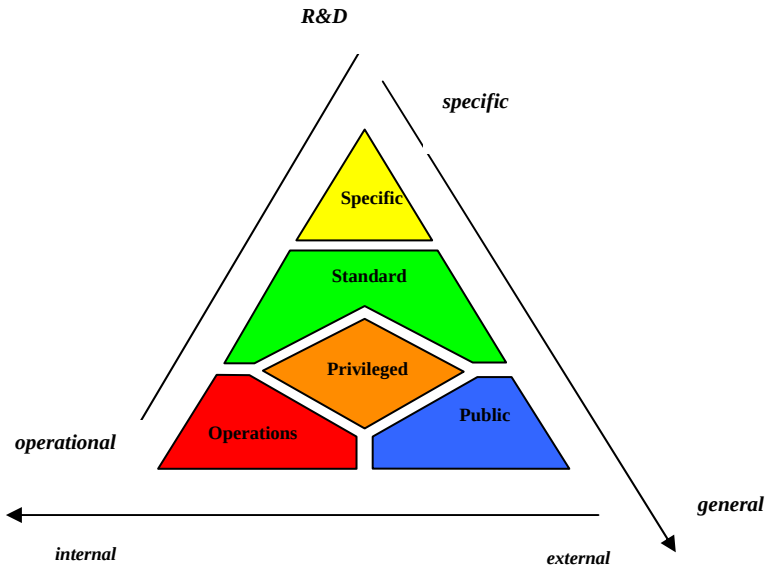

Figure 1 User classification
(courtesy to Mersea EU project)

In the past ten years we have therefore seen the development of value added products, mainly gridded products, climate change indicators, reanalysis of big data holdings, to fulfill the needs of users who were confident enough in the accuracy of the observations and were focusing on the oceanographic applications they could study. This has lead to the setting up of thematic centers or service providers that were able to build their products for the users.

On the other end the providers come from various domains from meteorological and environment agencies, satellite agencies, research communities, private sector with different culture and customs in term of data

processing and sharing. They usually organize their information system and data distribution channels to serve their users. Being able to serve other users need additional efforts and harmonization with the rest of the world…

The following table sketches the figurative distance between an observing system and an end user and the role of each actor. Depending on where a user gets the data from, the amount of processing to be done in house can be different as well as the information provided with the data. To achieve the goal "acquire once and use many times" it's important to secure a lot of metadata with the data as there is no direct contact with the scientist who operates the platform. It also secures the data for next generation!

When	From	To	Who	What
Since Ocean Observation started	Platform	Media	The Provider: can be a scientist, an agency, a private company,…	• Collect the data , transform sensor measurements into oceanographic information. • Record/Describe the way the data has been collected
Started around 1960 and began to be organized within IODE with the setting up of National Data centers	Media	Repository	A Data Center	• Archive, quality control, distribute the data to the external users • Ensure that required metadata, have been filled in and trace the history of the processing of the data (version tracking)
Late 90's after the success of the WOCE experiment	Repository	Service provider	Thematic Assembly Centers	• Integrate various data into a coherent product, check the coherency of the data coming from various platforms, provide feedback to Data centers when anomalies are detected. • Derive value added products designed for a kind of application • Ensure distribution to international community
Nowadays	Service provider	End users	Service providers	• Customized product from specific applications combining a wide variety of observation, models outputs and expertise.

Table 1: Evolution of data distribution in past decades.

III. What do we need to share?

Data are acquired at coastal, regional, pan-continent, global levels and all together they make the observing system needed by operational and research applications. Access to these data often looked like a plate of spaghetti where you never knew how long the way will be to reach the data you need. What users are asking for are portals, that build for them the connection to all the relevant datasets and provide access to these data as if they were all in a single place.

Sharing data? Fine, but what and with whom do we need to share data? One can easily understand that a coast guard having to perform ship routing in the frozen Gulf of St Lawrence is more interested in water temperature in the nearby seas than in the Mediterranean sea unless he is preparing for his holidays.. While for the Gulf of St Lawrence he will ask for temperature/salinity/wind/ice time series at a specific point with a few minutes/hours delay, for his holidays climatological maps on the WWW of sea surface temperature and wind will be enough.

This simple example shows that in fact the international data management infrastructure as to be built as a system of systems of systems offffffff…… of systems. The more you go to global scale applications the fewer parameters you need.(Figure 2). The global data management systems can therefore be seen as an organized network of integrated systems, integrated level by level from national to regional to continental to international. One level only needs to care about its interfaces with the level above and below as represented in figure 2. This way was paved in the late nineties within WOCE relying on thematic centers integrating each part of the WOCE observations. This is what lead to thematic assembly centers either within IODE (SeaDataNet in Europe) or GODAE (Aviso, Coriolis, GHRSST), ICES, OBIS(Ocean Biogeographic Information System) , OTN(Ocean Tracking Network),… It permits building value added products at different levels targeted from specific regional or global applications in term of time and space resolution, parameters required, …

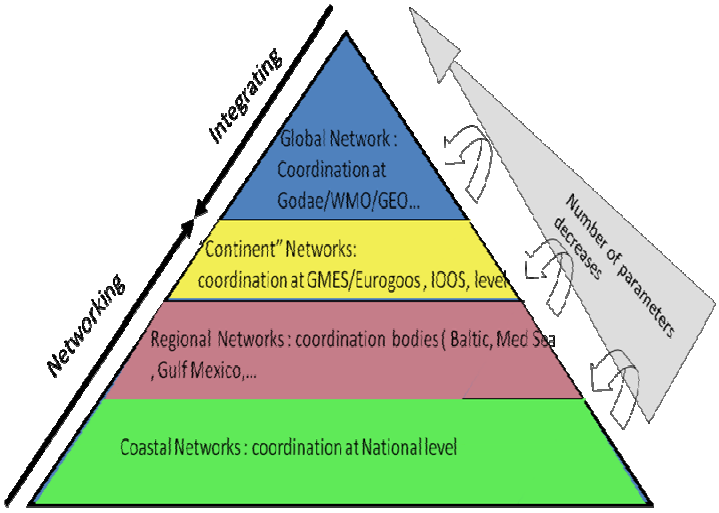

Figure 2 From coastal to global observing system

IV. Different types of Data Management architectures

A data management system is designed according to the type of data handled (images/ profiles/ timeseries/ kilobytes versus gigabytes, etc), the user access needs (individual measurements, geographical assess, integrated datasets, etc), the level of integration needed, etc.

In the past decade, with the improvement of the computer technology, the internet revolution, the increase of network speed and capacity, data management systems have been progressively moving from centralized to distributed systems. Two main architectures are nowadays commonly used:

1. Distributed processing and centralized distribution: data are processed in different places and are than copied to a single place for distribution to users.
2. Distributed processing and distribution: data are processed in different places and stay where they are. To ease user access a virtual WWW portal is implemented that uses networking techniques to find the data that fit the user needs.

Each system has its advantages and drawbacks, depending on the type of datasets to distribute and the contributors to the network. These different architectures will now be quickly described through examples operating at present.

The first solution has been implemented with Argo (e.g Pouliquen & al CWP 2009) and then been endorsed by other programs such as GOSUD or OceanSites (e.g Send & al CWP 2009). It is based on data processing using common procedures in DAC (Data Assembly Centres) and centralized distribution from two GDACs (Global Data Centres) that synchronize daily their databases..

The advantages of this system are:

- One stop shopping for the users where they get the best available data in an unique format
- Data discovery and sub-setting tools are easy to implement as all the data are in the same place
- A robust system, as the probability that both GDACs will fail is very small

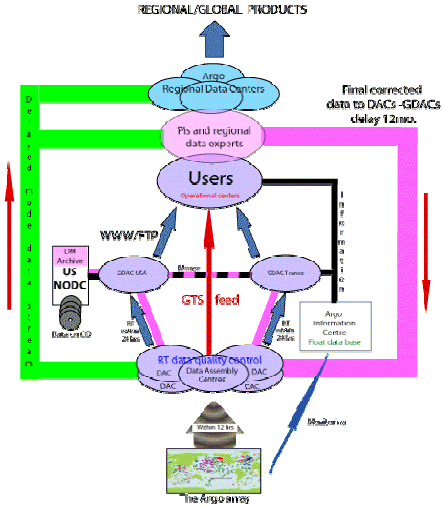

Figure 3 Real Time Data stream for Argo. Data available within 24h from acquisition

- Easy to guaranty a quality of service in data delivery because GDAC have the control of all the elements in-house

The disadvantages:

- Data are moved around the network and must rely on the "professionalism" of the DACs involved in the system to be sure that GDACs have the best profiles available.
- Additional work at DAC level to convert their data from their home format to the common format. This may be hard to do for small entities.
- Data format used for data exchange cannot evolve easily as it requires coordination among all actors before implementation. Since users, especially operational ones, do not like format changes it is not such a big problem.
- If only one main server is set up that the system is fragile. Setting up a mirroring system can overcome this problem with additional synchronization mechanisms.

The second one is used to integrate existing data systems as a cooperative integration of independent systems that will continue their mission independently while participating in an integrated data system. It's the architecture used in US-IOOS (Integrated Ocean Observing System) or European SeaDatNet infrastructure. In such a system the data processing is distributed and the data stay on physically distributed repositories, some containing huge amounts of data. The user connecting to the system website will be able to query for data without knowing where they physically reside. The key elements of such a system are the metadata management, the data discovery system and the data transport protocols.

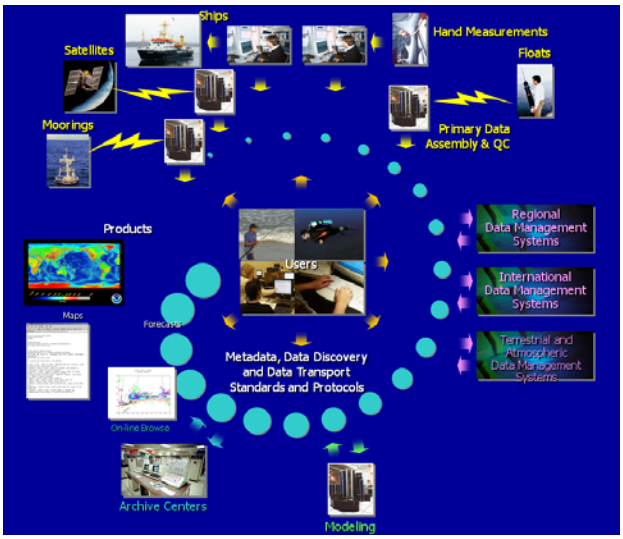

Figure 4 IOOS Data Management and Comminication system

The advantages of such system:

- Optimisation of the resources (network, CPU, Memory, etc) among the contributors,
- Data stay where they are generated preventing non compatible duplicates among the network
- Built on internationally agreed standards that guaranty its efficiency in the long term and its adaptability because it will benefit from international shared developments.

The disadvantages:

- The system is not easy to set up because it needs a lot of international coordination, especially for metadata.
- Even more work for small contributors because it requires important computer expertise
- It can be unreliable if some data providers cannot guaranty data service on the long term. To be reliable such a system must rely on sustained data centers.

V. What are the key elements for efficient data sharing

Data processing and distribution must be designed properly to be able to deliver the data in time for use. First, data have to be publicly available in real-time for forecasting activities, and within a few months for re-analysis purposes. This raises the important issue of an open data policy to be solved by the funding agencies at national and international levels. Second is the organization of the data flow among the different contributors in order to have an efficient data management network able to answer the needs. For a long time, data management aspects have been neglected in projects and too little funding was devoted to this activity both for in-situ and satellite data processing.

The past 10 years have seen the development of autonomous platforms able to acquire accurate measurements for years (Argo, gliders …) and transmitting in real-time as much data in one year than what has been acquired in the past century. (eg Roemmich & al CWP 2009 ,Testor & al CWP 2009) . RT data transmission from moorings has also increased significantly (eg Mc Phaden & al CWP 2009, Send & al CWP 2009) as well as better use of commercial and research vessels (e.g Smith CWP 2009). As a consequence it has become clear that the workload for data processing had to be spread over institutes and that harmonization of data processing

and distribution was a priority if we wanted to use these data as a network and not as individual platforms. It has also become clear that it was important to set up portals to ease access to the data of a specific network: the concept of Global Data Centers was born (e.g Pouliquen & al CWP 2009) which is an update of the WOCE Data centers but with a mission to aggregate and distribute data within 24 hours and no more after 2 years .

Even if presently, there is no "stamped" consensus on data management and communication strategy for effectively integrating the wide variety of complex marine environmental measurements and observations across disciplines, institutions, and temporal and spatial scales, there are already some success stories that have shown that it was possible with a minimum coordination. Moreover there are either standards (ISO) or de facto common practices (NetCdf/CF, OPeNDap, ..) endorsed by information system managers that allows some interoperability between systems.

Providing access to data can be seen as a piling up of application levels that will at the end allow services to be built by potential users and not by the data providers themselves. The yellow parts of Figure 5 are the responsibility of data centers or data assembly centers, the purple part allows system of systems building, the green and red part are service providers duties. In the past decade progress has mainly been made in the first two layers.

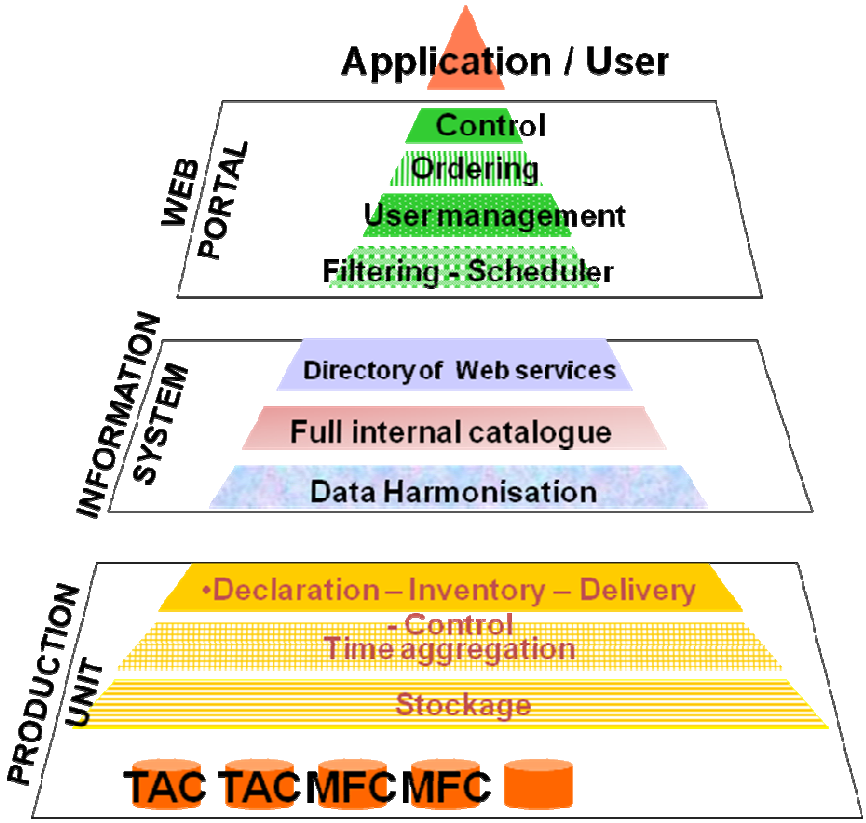

Figure 5 A layered approach
(courtesy to F Blanc CLS)

5.1 The data assembly center Layer :

Data are obtained by diverse means both in-situ (ships, drifters, floats, moorings, seafloor observatories, etc) and satellites, they come in very different forms, from a single variable measured at a single point to multivariate, four dimensional collections of data, that can represent from a few bytes a day to gigabytes. The duty of the assembly centers is then

- to integrate data coming from a wide variety of platforms and providers (scientists, national data centers, satellites data centers, operational agencies,….),

- to get enough information from the originators to be able to know exactly how they have been acquired (need for metadata) and processed (documented and commonly agreed QC procedures, history of the processing)
- and then to distribute then in an agreed standard (speaking the same language)

In some domains were it was evident that no country, no agency, no scientists would be able to acquire alone the data needed a free data policy was a requirement at the beginning of the project. TAO/TRITON/PIRATA array motivated by the 1982-1983 El Nino event, the strongest of the past century, is a great example of the benefit of a free data policy and international cooperation. Argo , a 3000 profiling float array drifting every 10 days , is another example that became a reality in the past 10 years and for which data are available within 24 hours from acquisition. The same changes to open data policy have also been achieved for satellite data with Salto/DUACS service to altimetry products and the GHRSST services for SST products. Some other programs such as GOSUD for data acquired underway by research and commercial vessels or OceanSITES for reference mooring sites have adopted open data policy in the past ten years to enhance usage of the data. In all cases Global Data Centers have been set up that give access to the best copy for the program of the data at that time . The same way for biogeochemical data the iOBIS portal is providing a unique access to many datasets containing information on when and where marine species have been recorded .

In the past 10 years IT technology has progressed a lot. . The explosion of Internet has provided new tools and a new way of using computers and communication means. It has also allowed the creation of multiple duplicates of the same data circulating on the WWW that can be the exact copy but also a modified version of the same original data. The Global Data centers adopted by some programs are a possible work around but for other datasets no real solutions has been found yet.

To be able to provide products directly usable by applications in some domains such as operational oceanography or climate change monitoring, thematic data assembly and processing centers have been set. Over the past ten years, ocean data processing centers have been considerably enhanced in order to meet the needs of ocean applications. Their role is to collect, quality control the data and check their consistency, provide an error estimate on the data, correct them in delayed mode if possible and distribute them. With the development of ecosystem models, there is a need that biologists and chemists provide data more quickly so that it can benefit from the complete data set and not just the data from individual projects. Here after are some examples of thematic assembly centers developed in past 10 years in satellite and in-situ field, in physics, chemistry and biology fields.

5.1.1 Some success stories

Based on the development started within WOCE, CDIAC/USA (http://cdiac.ornl.gov/) CDIAC's ocean carbon data collection includes discrete and underway measurements from a variety of platforms (e.g., research ships, commercial ships, buoys). The measurements come from deep and shallow waters from all oceans. Technological advances make it possible to deliver ocean carbon data in real-time but questions about instrument reliability and data quality limit this practice at this moment. All ocean carbon data CDIAC receives come from individual investigators and groups following initial data review. CDIAC has first standardized and made an inventory of its ocean data holdings using the Mercury Metadata system, then it implemented a first version of viewing and downloading service via a Live Access Server (PMEL) and moved to web services later in 2007. CDIAC is also developing analysis products. The first task of the data synthesis project is to assemble a merged data set for each basin. As the data sets are assembled, consistency should be checked by comparing property-property and property vs. depth plots for stations that are near (within 50 to 100 km) the intersection of cruise lines (the so-called crossover analysis). This procedure is the first level of quality control and indicates, but does not eliminate, the possibility of systematic differences between cruises or oceans. The next step is to recommend adjustments to the inorganic carbon data based on a comprehensive check of analytical and data reduction procedures, analysis of crossover, and regional analysis of cruise data. This is necessary to produce a gridded data set that is both precise and accurate on a global scale. (cf Borges & al CWP 2009)

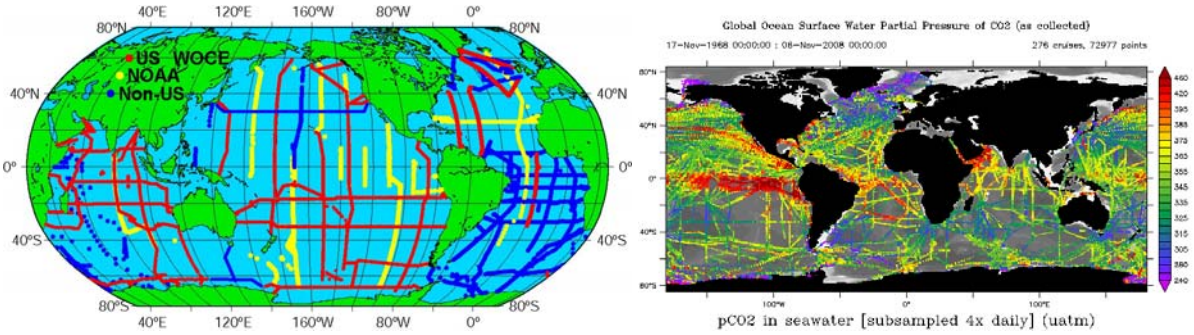

Figure 6 : GLobal Ocean Data Analysis data base and CO2 Time-series and Moorings map and derived pCO2 product

For Operational oceanography Coriolis/France (http://www.coriolis.eu.org) integrates into a single dataset data from international networks (Argo, GOSUD, OceanSITES, DBCP, GTSPP) and European regional data (EuroGOOS Regional Operational Oceanographic systems). Over 10 years, the amount of data processed by Coriolis multiplied by 6 (Figure 7) for temperature and salinity parameters in real time and delayed mode. To be able to provide such products, Coriolis developed and implemented additional quality control procedures that look at the data as a whole and are able to detect suspicious measurements that were not detected by automatic tests, or profiles/time series that are not consistent with their neighbors. Since 2005, Coriolis has also been producing global ocean weekly temperature and salinity fields from the Coriolis database using objective analysis. Statistical methods also permit detection of outliers in a data set by exploiting mapping error residuals (Gaillard et al., 2009). An alert system has been set up that detects the profiles for which the error is larger than a threshold. An operator scrutinizes outliers, discerning the difference between an erroneous profile and an oceanographic feature such as an eddy or a front. Coriolis is also setting up complementary validation activities for Argo data

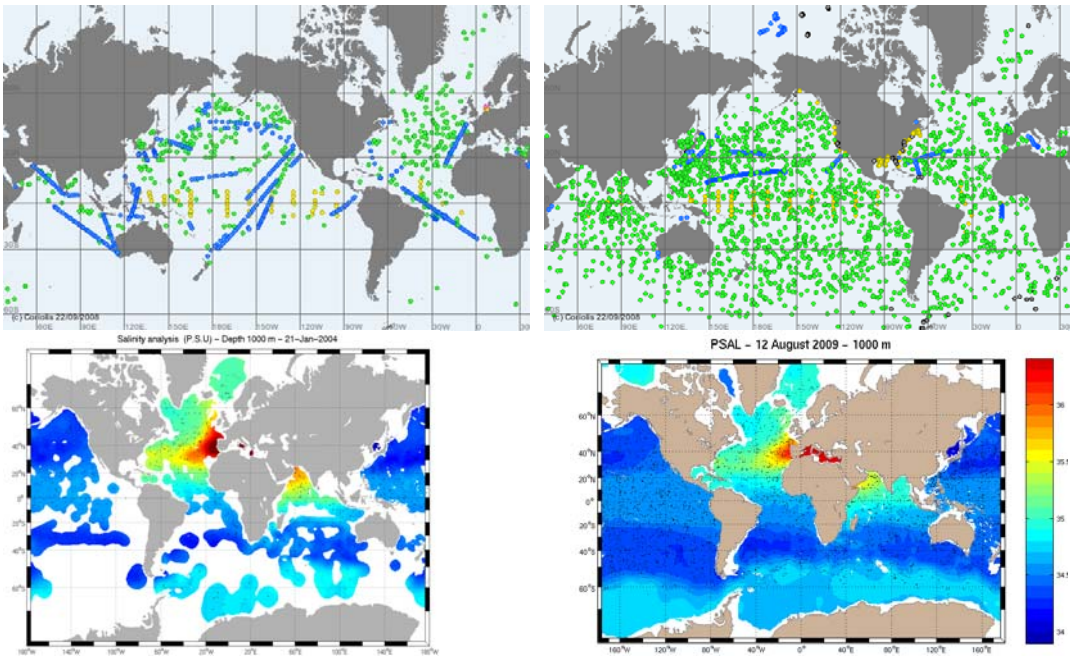

Figure 7: 10 days of profile data from Coriolis data base in Sept. 2002 (left) (about 4700 profiles)) and Sept. 2008 (right) (about 27000 profiles) : XBT*(blue), Argo (green), Moorings (yellow).and Derived salinity field at 1000m calculated from this datasets*

A third example is in the Satellite domain with the enhanced access to SST products within the GHRSST program (http://www.ghrsst-pp.org/). During the past ten years, a concerted effort to understand satellite and in situ SST observations has taken place leading to a revolution in the way we approach the provision of SST data to the user community. GHRSST has implemented a wide and open access in near real time to many satellite SST data products in an operational-like manner using existing data user-driven distribution protocols, tools and services. They set up two GDACS (USA, Europe) to provide unique access points to data processed by Regional assembly centers all around the world. They set up an international agreement on the definition of different SST parameters in the upper layer of the ocean and have registered them in the Climate Forecast (CF) standard name table for wide application. Diverse satellite SST data product formats and product content have been homogenized according to international consensus and user requirements to include measurement uncertainty estimates for each derived SST value and supporting auxiliary data sets to facilitate their use by data assimilation systems (eg Donlon & al CWP2009).

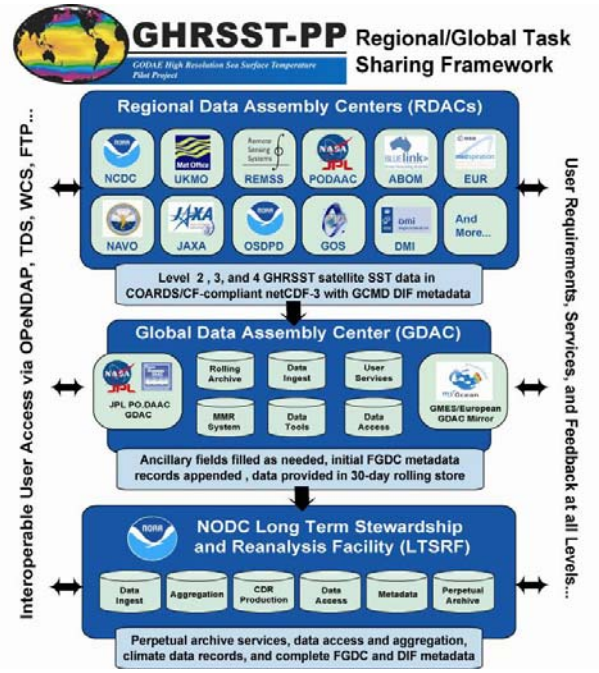

Figure 8: GHRSST Data System

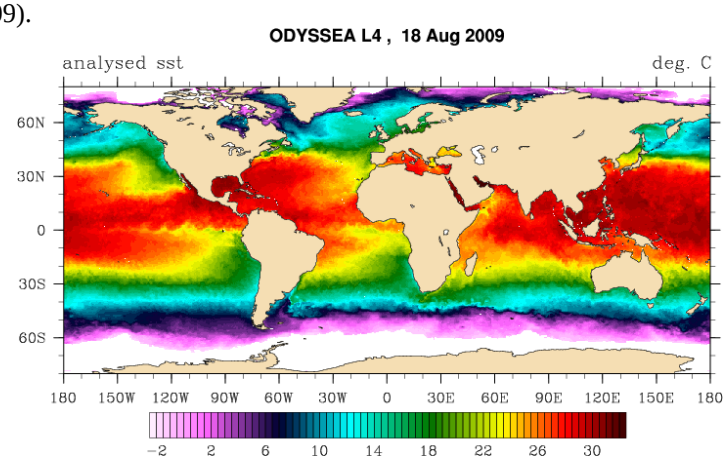

Figure 9 SST daily gridded data calculated from GHRSST data

The last example focuses on marine biodiversity. A major impediment to advancing our marine biodiversity knowledge is the paucity of digital species-occurrence data available online. Although more species-occurrence records are steadily being acquired, it is still difficult to find past and current marine biodiversity data for anything but well-studied, economically important taxa that occur in well-studied areas. It has been even harder to aggregate data from multiple sources in order to ask new questions not envisioned by those performing the initial surveys. We are often unable to answer very simple, fundamental biodiversity questions for most oceanic regions in the world, such as “what biodiversity has been found in region X?” and “has previous sampling been sufficient to support confidence in biodiversity estimates?” A partial solution to the problem of data availability is a global mechanism that facilitates sharing of biodiversity data that is housed in various repositories world-wide. It is especially important to note that a substantial percentage of the total are records collected prior to major alterations of the environment by humans. Thus, these specimen data are the best possible resource with which to construct baselines to measure changes in biodiversity over time (Suarez & Tsutsui 2004, Graham et al. 2004). Multiple agencies, but especially the Global Biodiversity Information Facility (GBIF) and International Ocean Biogeographic Information System (iOBIS) and the associated regional nodes (eg. OBIS-USA) and focused taxonomic nodes (eg. OBIS-Seamap) has developed a worldwide information infrastructure through which natural history collections (as well as other institutions and organizations) can publish their databases, and thus become part of a distributed global network of shared biodiversity data (Edwards 2004; Lane 2006, Guralnick et al. 2007). Any user with Internet connectivity can access a vast queryable global marine biodiversity data service. For example, iOBIS currently makes available 19.1 million records of 106000 species

from 643 databases. Such repositories continue to grow as new data contributors agree to share their data and metadata with the broader community.

The development of shared data standards and transmission protocols has been an essential catalyst for interoperability among biodiversity data. Because all data adhere to a common set of standards for data and metadata (Graham et al. 2004) and use the same methods for sending data over the Internet (Stein & Wieczorek 2004), search results from portals such as iOBIS, OBIS-USA and GBIF are returned to the user in a common format. As well, portals can share data with each other and with other applications via application programming interfaces (APIs). . The essential data standard is DarwinCore which specifies the minimum information content necessary for a species occurrence record (scientific name, when and where the specimen was collected and by whom). DarwinCore has been extended to meet the needs of various communities, including the ocean biogeographic community (see: http://www.iobis.org/tech/provider/schemadef1.html). As important has been development of shared transmission protocols across the publishing network. These protocols allow OBIS networks to communicate with its distributed data contributors, defining how data is exchanged. Using such systems, OBIS networks can more effectively distribute queries and can more easily synchronize datasets across the network.

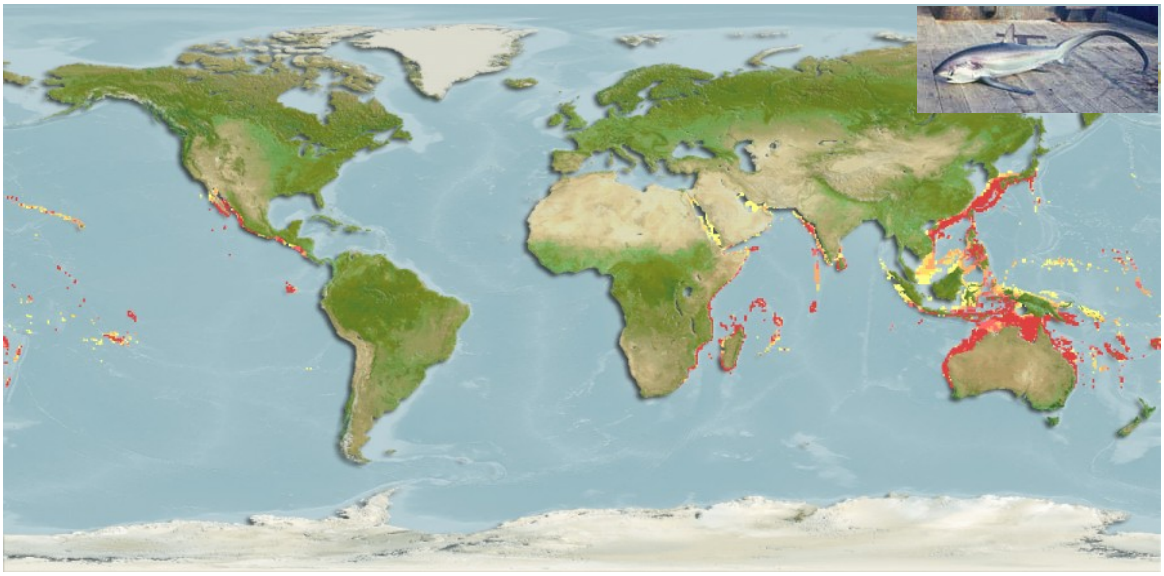

Figure 10 Shark (Alopias pelagicus) distribution in OBIS

5.1.2 Essential elements for information sharing to be implemented by providers

The tasks of these thematic assembly centers have been possible in the past ten year because of progresses made in two main areas:

- ***Data description metadata and formats***
- ***Improvement and standardization of quality control procedures***

Progress made on data format and metadata : Data must be preserved in such a manner that they will still be useful in the future when the PI who acquired the data may have moved somewhere else. They must also be distributed in a way that a user can easily merge it with other datasets relevant for his application. They must help to find the data among the network (data catalogues). That is the purpose of defining correctly distribution data formats as well as the metadata (data on the data) that need to be preserved for future processing. Data formats have always been a nightmare both for users and data managers and they are both dreaming of the "Esperanto" of data formats. Computer technology has improved a lot in the past decade and we have slowly moved from ASCII formats (easy to use by human eyes but not for software), to binary format (easy for software but not shareable among platforms (Windows, Unix, etc), to self-descriptive, multiplatform formats (Netcdf, Hdf, etc) that allow more flexibility in sharing data among a network and are read by all software that are commonly used by scientists. One important point for metadata is to identify a common vocabulary to record most of this information. This is easier to achieve for a specific community such as physical operational oceanography as the number of parameters is small, but it starts to be a bit more difficult when we want to address multidisciplinary datasets. To help in this area some metadata standards are emerging for the marine

community with CF/COARDS convention, Marine XML in Europe, and ISO19115/19139 norm. The availability of usable standards both for data description (catalogues, metadata, formats) are discussed more in detail in Keeley & al Plenary Paper and Snowden & al CWP 2009.

Progress made on standardize quality control procedures

Assessing the quality of an observation is important for a sensible and efficient use of the data. The quality control (QC) procedures have to be adapted to the allowed delay of the delivery. In Real-time most of QC is done automatically and only outliers are rejected for in-situ, or sensor drift is estimated against in-situ for satellite data. In delayed mode, more scientific expertise is applied to the data and error estimations can be provided with the data. Data quality control is a fundamental component of any ocean data distribution system because using erroneous data can cause incorrect conclusions, but rejecting extreme data can also lead to erroneous assumptions by missing important events or anomalous features. The challenge of quality control is to check the input data against a pre-established "ground truth". But who really knows this truth when we know that the ocean varies in time and space, and also that no instrument gives an exact value of any parameter but only an estimation of the "truth" within some error bars? As most of the data are processed by different actors, but used all together by users, clear documentation of the quality control procedures, good metadata to know how the data has been acquired, a homogenization of the quality flags, and reliability of different actors in applying these rules are required. In past decades a lot of progress has been made for basic physical data such as temperature , salinity ,currents,... But for other parameters there are still a lot of discussion on going before reaching agreement (eg Burnett & al QUARTOD Working group CWP 2009, Pouliquen & al Argo Data Management CWP 2009, Reed & al IODE Ocean Data Standards Forum CWP 2009) . The following step is to be able to provide uncertainty estimates on the data provided .

5.3 The Interoperability level :

Data assembly centers can do a tremendous work in building very good, documented and reliable products and if they are not easy to find by potential users, it may stay on the data center disks. With the explosion of Internet we are now used to finding in a few mouse clicks a lot of information so why not data from ocean observing systems? This is the purpose of the interoperability layer.

It first allows discovering on the network what the data/products are available. This is based on standardized product description (what, who, when, where, how,…) providing information enabling product discovery and links to the relevant documentation and access means to the products. Such product catalogues have progressed a lot in the past ten years and even if the standards (ISO19115, ISO 19139) are still evolving a bit, they are already implemented as in SeaDataNet (European project federating European National Oceanographic Data Centers) or IOOS, recommended at European level within the INSPIRE directive and endorsed by some software providers like ArcIMS widely used in the cartography domain. Protocols like WCS (Web Catalog Service) to interrogate them through the internet are developing and should be soon certified. This discovery level allows to find not only observations data but also products built upon observations like gridded fields, climatologies,…

When data providers endorse an open free data policy then it's possible to know how to build additional services for viewing and even download the data from a central point without knowing were the data are stored. For example within GODAE experiment a lot of products (observations and model outputs) have been put freely available in NetCDF format on servers on which OPeNDAP services were implemented. That way a consistent mechanism to access the products was available to download only the needed data via Internet (see Hankin & al CWP 2009 , Blanc & al CWP 2009). Better serving ocean data are discussed in detail in Blower plenary paper 2009.

VI. Perspectives

In past decade, data exchange between partners has improved mainly fostered by the satellite and the ocean physical communities. IT technology and data management systems are no more an obstacle. It has been shown that sharing in near real time ocean data was beneficial not only for the community but also for the scientists that were deploying instruments: anomalies in platforms detected earlier by comparison with nearby measurements, cost efficiency in implementation with better knowledge of existing platforms, networking activities to improve common quality control procedures. Moreover it has eased the work of the National and Thematic data centers to

collect important quantity of metadata that are essential for future reprocessing activities and climatology improvements. Presently a lot of data have been acquired in past centuries and it is really difficult to determine whether a trend is related to the accuracy of the data or the real trend in the ocean. The present experience has allowed the development of new services and products such as operational oceanography the same way as in meteorology 20 years ago. This data and product sharing in (near)real time need to be extending to other communities such as biogeochemistry or geology. It doesn't happen by chance and needs a real willingness and involvement of the community to improve and update regularly the needed vocabularies, QC procedure, Metadata content … necessary to be efficient and sustainable in the long term.

We do need now to take necessary steps to ensure that the shared observations and products are:

- **accessible**: this goal will be achieved by free access to essential data and implies incentives from funding agencies and a change in scientists' behavior,
- **comparable**: this implies agreement on quality control procedures both in real-time and delayed mode to provide datasets independent of what platform sampled it. It also implies a good version control of the data that allows to know the process the data have been through
- **understandable**: this will be solved by common standards for data description and distribution. Computer techniques will help to solve the syntactic part of the problem but better coordination between Europe, USA ,etc is needed to solve the semantic part of the problem.
- **Recognized/cited** in scientific papers that use them : it will help people to release their data more rapidly
- **Moving from Observations to information,** use of in-situ/satellite/model outputs together to provide services (Godae at International scale, MyOcean in Europe , IOOS in USA, ..) needed by agencies link EEA (European Environment Agency) or EMSA(European Marine Safety Agency) in Europe

We also have to move from primary observation to information production. This leads to major integration of data by merging different data sources coming from various instruments both in-situ and satellite.

Finally we should not re-invent the wheel and take benefit from the expertise in other communities (meteorology, deep ocean operational oceanography…) to improve data exchange at the regional level and improve answers to monitoring and crisis management activities. At the end of the day, we will all contribute to the Global Earth Observing System of Systems (GEOSS). The way forward for data system and integration will be discussed more in detail by S Hankin Plenary paper.

VII. References

Blanc & al , "Data and product serving, an overview of capabilities developed in 10 years " , OceanObs09 CWP 2009

Blower & al, "Serving Godae Data and products to the ocean community", Oceanography Magazine Godae special issue (2009)

Blower & al, "Data discovery and delivery", OceanObs09 plenary paper 2009

Borges & al OceanObs09 CWP 2009)

Burnett & al , " Quality Assurance of Real-Time Ocean Data: Evolving Infrastructure and Increasing Data Management to Monitor the World's Environment " ,OceanObs09 CWP 2009,

Donlon & al, "Successes and Challenges for the Modern SST Observing System ", OceansObs09 CWP 2009

Hankin & al, "NetCDF-CF-OPeNDAP: Standards for Ocean Data Interoperability and Object Lessons for Community Data Standards Processes ", OceanObs09 CWP 2009 ,

Hankin & al, " Data system including data transport and product delivery / visualization", OceanObs09 Plenary 2009

Le Traon & al, Ocean, "Data Assembly and processing for operational Oceanography: Ten years of achievements", Oceanography Magazine Godae special issue (2009)

Keeley & al , "Data Assembly Infrastructure", OceanObs09 Plenary Paper 2009

Keeley & al, " The Data Management System for Surface Drifters", OceanObs09 CWP 2009

Mc Phaden & al, " The Global Tropical Moored Buoy Array", OceanObs09 CWP 2009,

N.C.Flemming, D Prandle , "The science base of EuroGOOS" , EuroGOOS publication N°6 1998

Pouliquen & al Argo Data Management OceanObs09 CWP 2009,

Reed & al IODE Ocean Data Standards Forum OceanObs09 CWP 2009)

Roemmich & al , " Argo: Observing the global ocean", OceanObs09 CWP 2009 ,

Send &t al, "OceanSITES", OceanObs09 CWP 2009)

Smith &al, " The Data Management System for the Shipboard Automated Meteorological and Oceanographic System (SAMOS) Initiative" , OceanObs09 CWP 2009

Snowden & al, " Metadata Management in Global Distributed Ocean Observation Networks", OceanObs09 CWP 2009.

Testor& al ," Gliders as a component of future observing systems " , OceanObs09 CWP 2009) .