THE DEVELOPMENT OF THE DATA SYSTEM AND GROWTH IN DATA SHARING

Lead author: Sylvie Pouliquen, Ifremer, Plouzané, France
Contributing authors: Steve Hankin[1], Robert Keeley[2], Jon Blower[3], C Donlon[4], Alex Kozyr[5], Robert Guralnick[6]

[1]NOAA/PMEL, Seattle, USA - [2]ISDM, Ottawa, Canada
[3] Environmental Systems Science Centre, University of Reading, RG6 6AL, United Kingdom, Email: j.d.blower@reading.ac.uk [4]ESA, Netherlands
[5]CDIAC, Tennessee, USA - [6]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder CO 80309, USA

1. CONTEXT AND SETTING THE SCENE

A great wealth of ocean 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 thousand institutes and agencies all around the world. Both the means to acquire these data and the way in which they are used have changed greatly in the past ten years.

Over the last decade, information technology has progressed a great deal. It presently allows the exchange of gigabytes of data and more via the Internet in developed countries. In the late nineties it was considered high technology to provide data on CDROM rather than on magnetic tapes and only small datasets were distributed via the Internet. Nowadays CDROMs are considered as a backup delivery system especially for countries with poor Internet connections. The explosion in use of the Internet has provided new communications capabilities, new tools, and a new way of using computers.

The nature of requirements from government agencies have changed: they want to know or estimate what the future of the earth will look like and what will be the impact on their territories due to climate change issues, ocean health monitoring and fisheries assessment, but they can't pay the full bill for the data acquisition. Therefore they are pushing, and nowadays more often imposing, a change in data policy and a move towards increased data sharing, in which data acquired with public funds should be freely available to the community.

Moreover, the nature of science itself 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 data collected by others as well as their own. They cannot do their work using only data they have collected themselves.

Finally the growth of operational oceanographic services, based on downscaling of global model results, is 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.

In 1998, EuroGOOS issued "The science base of EuroGOOS" publication [14] where some limitations related to data exchange were highlighted. The situation has improved but the following statements are still relevant and should be seen as priority actions not only in Europe but all around the world.

- *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.* This is still true in general, although the situation has improved a lot in the past 10 years in the physical oceanographic domain. Experience within the GODAE (Global Ocean Data Assimilation Experiment), Argo and GHRSST (Global High Resolution Sea Surface Temperature) programs have shown that it was possible to reach consensus on common standards (on issues such as data formats, real-time and delayed mode quality control, data distribution). In some domains, such as bio-geochemistry, there is still a long way to go.
- *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 data to users in real-time: Salto/DUACS for altimetry, Medspiration/GHRSST for SST, Argo and GOSUD (Global Ocean Surface Underway Data) Global Data centers and JCOMMOPS are examples that exist nowadays. It has proven that sharing data rapidly is not a burden for scientists. On the contrary, it is beneficial as problems are detected more rapidly by comparison with nearby measurements and collaboration to set up appropriate observing system facilitated.
- *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 (Global Monitoring for Environment and Security) 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 be needed to be sustained in the future.

- *Imbalance between monitoring (measurement) technology and capacity for post-processing data and subsequent real time use of numerical models.* Money and man power have been allocated to homogeneous datasets both at national and international level. This is illustrated by the French Coriolis project, the EU project SeaDataNet and DMAC (Data Management and Communications) in the USA. This effort should be sustained in the future.

In the past decade progress has been driven not only by applications such as Operational Oceanography but also by a fundamental change in cultural behaviour among scientists. This important change first started in the satellite community where it was possible to bring 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 [1]. This paper will discuss the progress achieved in the past ten years, lessons learned and future needs. Some aspects will be detailed in the other OceanObs'09 plenary and community white papers related to data infrastructure issues. [4] [9] [11]. In particular this paper provides the overview of data management. Details of systems are discussed by Keeley et al [11] and by Blower et al [4]. Finally Hankin et al [9] provides a vision of future systems.

2. A RELATIONSHIP BETWEEN PROVIDERS AND USERS

But what does sharing data really mean? "Data sharing" simply means providing a way for potential users to access data that have been acquired and processed by a provider. Most users generally do not want access to raw data, but rather high-level data; information or products, which have been derived from the raw data, with a level of processing that depends on the category of users and the nature of their applications (Fig. 1). The goal for the data or product providers is to set up data services for these users. The needs are derived from main drivers combined together to define 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 the tailoring of services to the user needs, once the base information management systems have been set up.

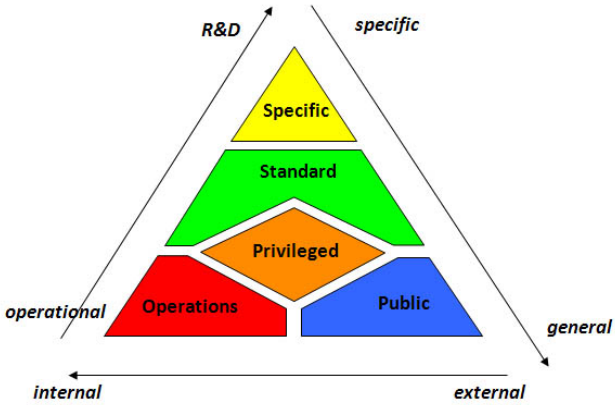

Figure 1: User classification. It shows that the more you go from operational to R&D activities the more tailored services are requested by users. It also shows you serve better the general public when you rely on standard service means easily accessible.
(courtesy to Mersea EU project)

.

Who	What	When	From	To
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 have been collected	Since ocean observation started	Platform	Media
A Data Center	• Archive, quality control, and distribute data to external users • Ensure that required metadata, have been filled in and trace the history of the processing of the data (version tracking)	Started around 1960 and began to be organized by IODE with the setting up of National Data Centers	Media	Repository
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	Late 90's after the success of the WOCE experiment	Repository	Service provider
Service providers	• Customized product from specific applications combining a wide variety of observation, models outputs and expertise.	Now	Service provider	End users

Table 1: Evolution of data distribution in past decades.

In the past ten years we have seen the development of value added products, mainly gridded products, climate change indicators, re-analysis of big data holdings, to fulfil the needs of users who were confident in the accuracy of the observations and wished to focus on the oceanographic applications they could study. This has lead to the establishment of thematic centres (centres that assemble data from various sources to serve some users on a specific theme: ie ocean physic modellers, satellite sea surface users, …) or service providers to build products for users.

The data providers come from various domains from meteorological and environment agencies, satellite agencies, research communities, and the private sector with a different culture and customs 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 needs additional effort and harmonization with the rest of the world [2] [3]

Table 1, here above, sketches the figurative distance between an observing system and an end user and the role of each actor. Depending on where a user gets data, 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 essential 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 the next generation.

3. WHAT DO WE NEED TO SHARE?

Data are acquired at coastal, regional, pan-continental and, global levels and all together they make the observing system needed by operational and research applications. Access to these data was previously extremely inconsistent, with users not being able to navigate around the systems in order to find the data they need. Users are demanding *portals* that connect to all the relevant datasets and provide access to these data as if they were all in a single place.

What data do we need to share, and with whom? 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). For the Gulf of St Lawrence, our coast guard will ask for temperature/salinity/wind/ice time series at a specific point with a few minutes/hours delay. In contrast, for his holidays, climatological maps of sea surface temperature and wind will be enough, which can be displayed simply on the Web.

This simple example shows that in fact the international data management infrastructure has to be built as a cooperative system of systems: Global scale applications require fewer parameters than local ones (Fig. 2).

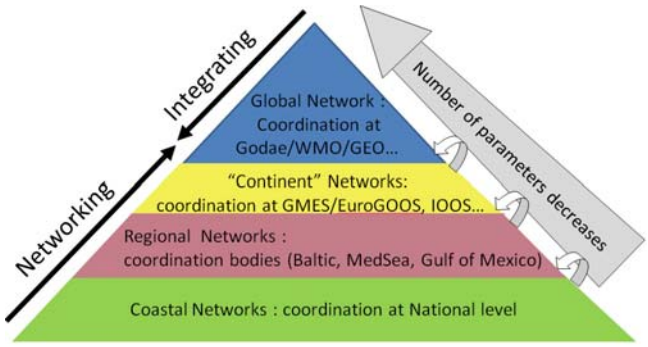

Figure 2: From coastal to global observing system

The global data management systems can therefore be seen as an organized network of systems, integrated level by level from national to regional to continental to international. Each individual 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 the WOCE (World Ocean Circulation Experiment) relying on thematic centres integrating each part of the WOCE observations. This is what lead to thematic assembly centres either within IODE (International Data and information Exchange) (such as SeaDataNet in Europe) or GODAE (Aviso, Coriolis, GHRSST), ICES (International Council for the Exploration of the Seas), OBIS (Ocean Biogeographic Information System), and the OTN (Ocean Tracking Network). It permits building value added products at different levels targeting specific, regional or global applications in terms of time and space resolution, and parameters required.

4. DIFFERENT TYPES OF DATA MANAGEMENT ARCHITECTURES

A data management system is designed according to the type of data handled (e.g. images, hydrographical profiles, time-series measurements), the typical required data volumes (kilobytes, gigabytes, terabytes, etc), the user access needs (individual measurements, geographical assessment, integrated datasets, etc), and the level of integration needed.

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

- Distributed processing and centralized distribution (Fig. 3): data are processed in different places and are then copied to a single place for distribution to users.
- Distributed processing and distribution (Fig. 4): data are processed in different places and stay where they are. To ease user access a virtual Internet 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 using examples operating at present.

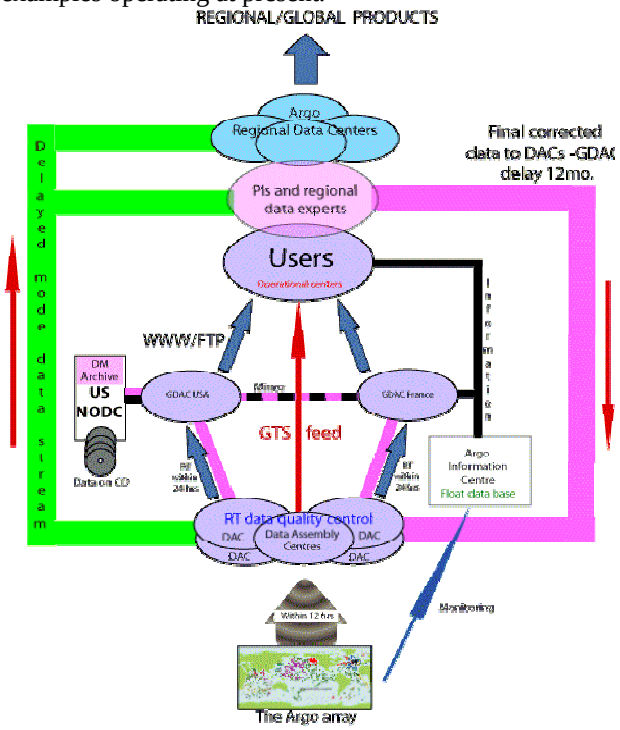

Figure 3: Real Time Data stream for Argo. Data available within 24h from acquisition

The first solution has been implemented with Argo (Pouliquen et al [15]) and adopted by other programs such as GOSUD (Smith et al [20]) and OceanSites (Send et al [19]). It is based on data processing using common procedures in DACs (Data Assembly Centres) and centralized distribution from two GDACs (Global Data Assembly Centres) that synchronize their databases daily.

The advantages of this system are:

- "One stop shopping" for the users where they can get the best available data in an single 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
- Easy to guarantee a quality of service in data delivery because GDACs have 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.
- The data format used for data exchange cannot evolve easily as it requires coordination among all actors before implementation. Since users, especially operational ones, prefer to change their formats infrequently, this is not such a big problem.
- If only one main server is set up then the system is fragile. Setting up a mirroring system can overcome this problem with additional synchronization mechanisms.

The second architecture described 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 the European SeaDatNet. 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.

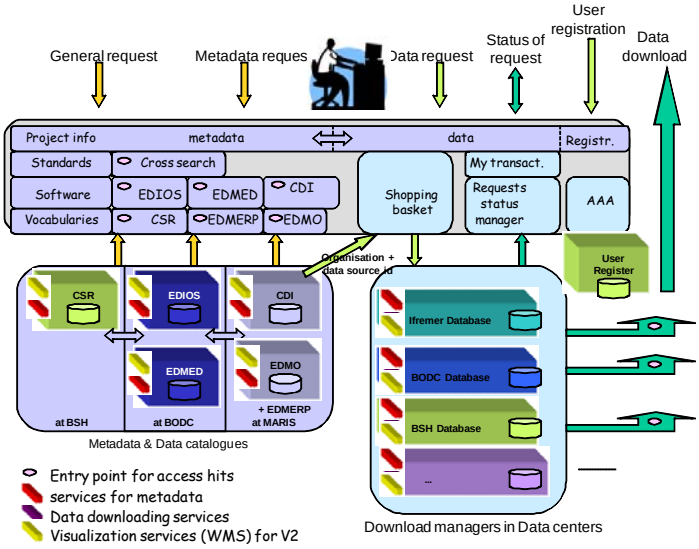

Figure 4: SeaDataNet Data Management Architecture

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 guarantee efficiency in the long term and adaptability because it will benefit from international shared developments.

The disadvantages:

- The system is not easy to set up because it needs significant international coordination, especially for metadata.
- More work for small contributors because it requires important technical expertise
- It can be unreliable if some data providers cannot guarantee data service in the long term. To be reliable such a system must rely on sustained data centres (ie centres that have secured funding on the long term to provide such services).

5. WHAT ARE THE KEY ELEMENTS FOR EFFICIENT DATA SHARING

Data processing and distribution systems must be carefully designed in order to deliver the data in time for use. First, data have to be publicly available in real-time for forecasting and monitoring activities, and within a few months for re-analysis purposes or long term analysis. This raises the important issue of an open data policy to be agreed 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 that is 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 (e.g. surface drifters, Argo and gliders) that are able to acquire accurate measurements continuously for years. These transmit in real-time as much data in one year as has been acquired in the past century (Roemmich et al [18], Testor et al [22]). Real - time data transmission from moorings has also increased significantly (Mc Phaden et al [13], Send et al [19]) as well as increased use of commercial and research vessels (Smith et al [20]). 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 in order to be able to integrate data from multiple platforms. It also became clear that it is important to set up portals to ease access to the data of a specific network, hence, the concept of Global Data Assembly Centres was born (Pouliquen et al [15]). This concept is an update of the WOCE Data Centres but with a mission to aggregate and distribute data within 24 hours and no more than 2 years after collection.

Even if, at present, there is no formally agreed consensus on data management and a 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 is possible with minimal coordination. Moreover there are numerous suites of standards, including OGC/ISO and de facto community standards (NetCDF/CF, OPeNDAP) [3,4,8] endorsed by information system managers that allow a certain level of interoperability between systems.

Providing access to data can be seen as a layering of application levels that will in the end allow applications to be built by potential users and not by the data providers themselves. Figure 5 shows 4 connected blocks of activities to provide a full service data system from archive to production. Within each block are more details of the kinds of activities. Activity of the production unit is the responsibility of data centers or data assembly centers. The information system allow for layering of systems through use of standards. Activities in the Web Portal and User blocks are the responsibility of service providers. In the past decade progress has been made mainly in the first two blocks.

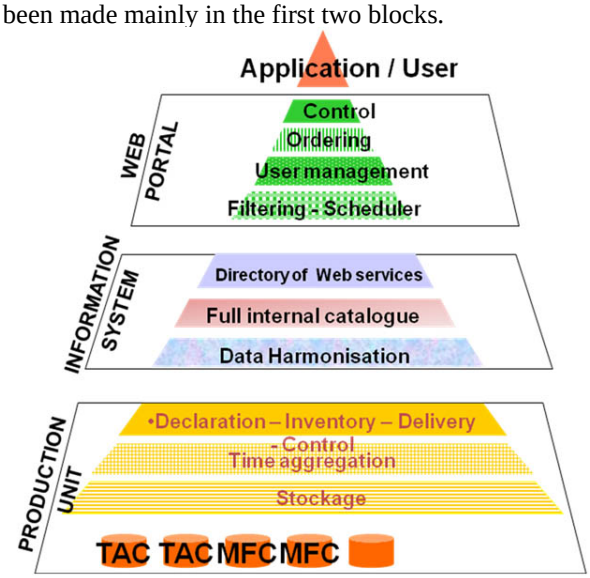

Figure 5: A layered approach (courtesy to F Blanc CLS)

5.1 The data assembly centre Layer :

Data are obtained by diverse means including 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 data volumes from a few bytes to gigabytes. The duty of the assembly centres is then

- to integrate data coming from a wide variety of platforms and providers (including scientists, national data centres, satellite data centres and operational agencies),
- to get enough information from the originators to be able to know exactly how the data have been acquired and processed (documented and commonly agreed QC procedures, history of the processing),
- and then to distribute them in an agreed standard ("speaking the same language").

In some domains where it was evident that no single scientist, agency or country would be able to acquire alone the required data, a free and open data policy was a requirement at the beginning of the project. The TAO/TRITON/PIRATA[13] array, which was 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, an array of ~3000 profiling floats, which each provide measurements 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 shift towards an open data policy has also been achieved for satellite data with the Salto/DUACS service for altimetry products and the GHRSST services for SST products. Some other programs such as GOSUD [20] for data acquired underway by research and commercial vessels or OceanSITES for reference mooring sites have adopted open data policies in the past ten years to enhance usage of the data. In all cases Global Data Assembly Centers have been set up that give access to the best copy of the data at that time requested by a user. In the same way the OBIS portal is providing a unique access to many biogeochemical datasets containing information on when and where marine species have been recorded.

The explosion in the use of the Internet has led to the creation of multiple duplicates of the same data circulating on the Internet that may be the exact copies, but may also be a modified version of the same original data; in general the user cannot easily find out whether a dataset has been modified. The Global Data Assembly Centers adopted by some programs are a possible work around but for other datasets no real solution has been found for these problems. Developments are on going to better track and tag data and identify more easily different copies of a single dataset.

To be able to provide products that are directly usable by applications in some domains such as operational oceanography or climate change monitoring, thematic data assembly and processing centres have been set up. Over the past ten years, ocean data processing centres have been considerably enhanced in order to meet the needs of ocean applications (Le Traon et al [10]). Their role is to collect, quality control the data, 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 for biologists and chemists to provide data more quickly so that the community can benefit from the complete data set and not just the data from individual projects. The following sections contain some examples of thematic assembly centres developed in the past 10 years for satellite and in-situ data, in physics, chemistry and biology fields.

5.2 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.

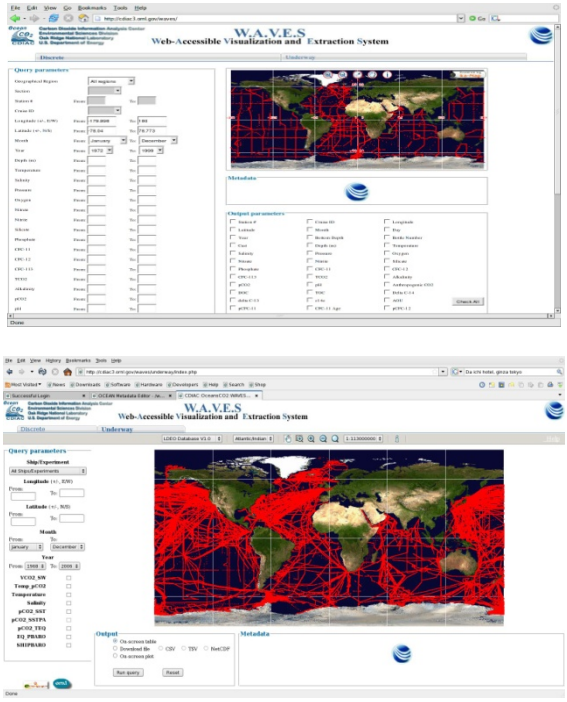

Figure6 : WAVES for the discrete measurements GLODAP data base and surface pCO2measurements LDEO database.

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 (http://mercury.ornl.gov/ocean/), then all data made available through the Web-Accessible Visualization and Extraction System (WAVES) for all discrete and surface measurements (http://cdiac3.ornl.gov/waves/).

(Fig. 6). CDIAC provides the data management support and archives the data analysis products such as Global Ocean Data Analysis Project (GLODAP) database, Carbon In the Atlantic Ocean (CARINA) database, Pacific Ocean Interior Carbon (PACIFICA) database, LDEO underway pCO_2 database, and the future Surface Ocean Carbon Atlas (SOCAT) surface pCO_2 database.

The data analysis procedure is called the second 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. (Borges & al [5]).

For Operational Oceanography Coriolis/France (http://www.coriolis.eu.org) [16] (Fig. 7) integrates into a single dataset the data from international networks (Argo[15], GOSUD[20], OceanSITES[19], DBCP [12], GTSPP – Global Temperature and Salinity Profile Project) and European regional data (EuroGOOS Regional Operational Oceanographic systems)

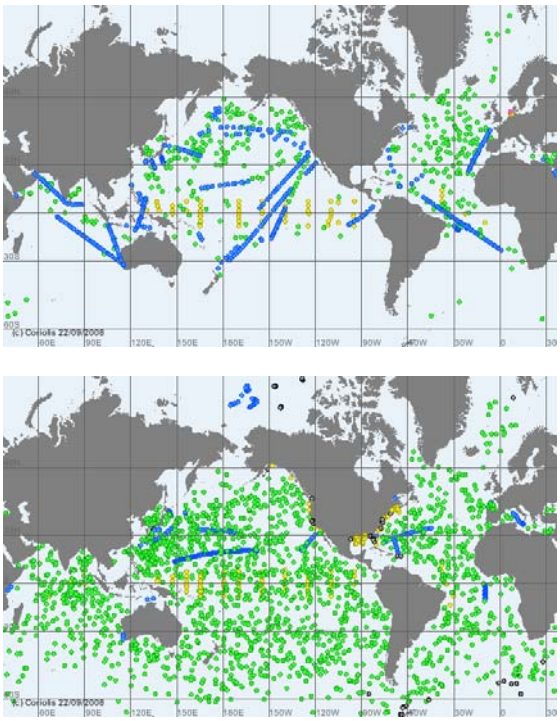

Figure 7: 10 days of profile data from Coriolis data base in Sept. 2002 (up) (about 4700 profiles)) and Sept. 2008 (below) (about 27000 profiles) : XBT(blue), Argo (green), Moorings (yellow).

Over 10 years, the amount of data processed by Coriolis has multiplied by a factor of 6 (Fig. 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 are not detected by automatic tests, or profiles/time series that are not consistent with their neighbours.

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.

Initially built to serve the French Ocean Forecasting System Mercator-Océan, Coriolis has gradually extended its services to other European Forecasting centres as well as international research community.

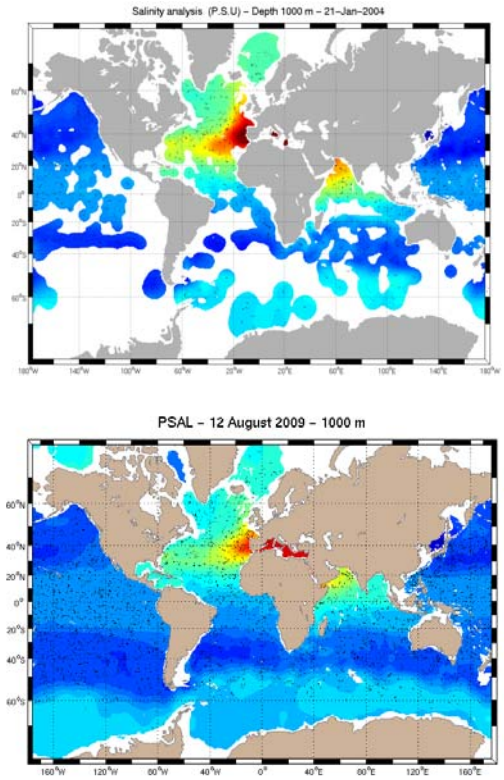

Figure 8: Derived salinity field at 1000m calculated from these datasets in 2002 and 2008

A third example is the GHRSST program (http://www.ghrsst-pp.org/), which provides enhanced access to sea surface temperature (SST) products. 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 manner using existing user-driven data distribution protocols, tools and services (Fig. 9).

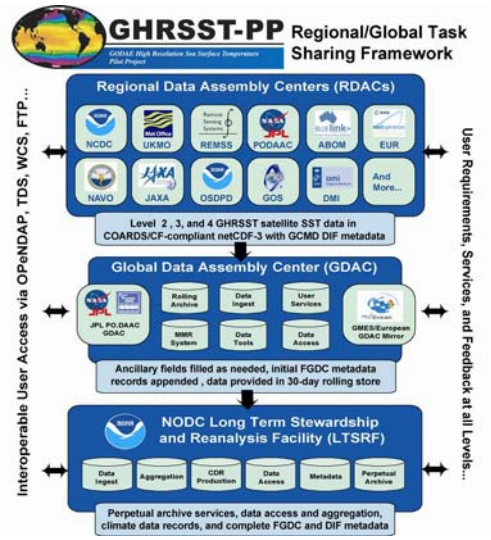

Figure 9: GHRSST Data System

They set up two GDACS (USA, Europe) to provide unique access points to data processed by regional assembly centres all around the world (Fig. 10). 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 (Donlon et al [7]).

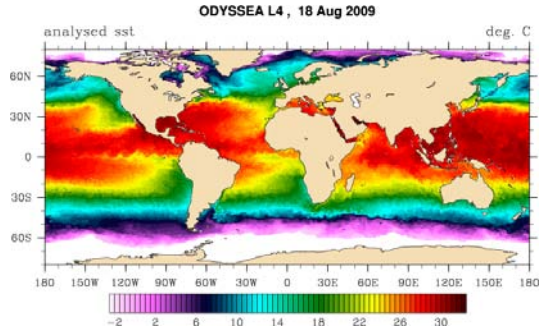

Figure 10: SST daily gridded data from GHRSST data

The last example focuses on marine biodiversity. A major impediment to advancing our marine biodiversity knowledge is the paucity of widely available digital species-occurrence data 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. Multiple agencies, but especially the Global Biodiversity Information Facility (GBIF) and the International Ocean Biogeographic Information System (OBIS) and the associated regional nodes (eg. OBIS-USA) and focused taxonomic nodes (eg. OBIS-Seamap) have 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 [24-26-27]). Any user with Internet connectivity can access a vast, global marine biodiversity data service. For example, iOBIS currently makes available 19.1 million records of 106,000 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 [28], search results from portals such as iOBIS [23], OBIS-USA and GBIF are returned to the user in a common format. 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

are exchanged. Using such systems, OBIS networks can more effectively distribute queries and can more easily synchronize datasets across the network.

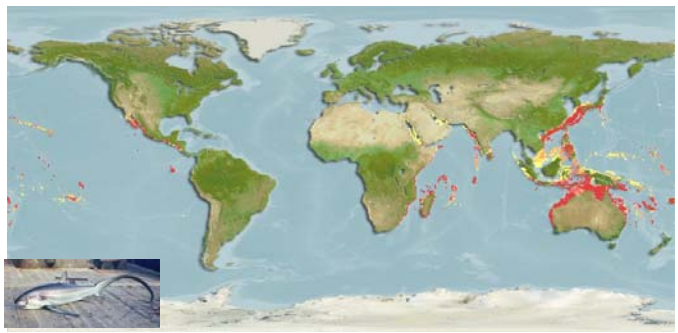

Figure 11: Shark (Alopias pelagicus) distribution in OBIS

5.3 Essential elements for information sharing to be implemented by providers

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

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

5.3.1 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 researcher who acquired the data may have moved somewhere else. They must also be distributed in a way that a user can easily merge them with other datasets relevant for his application. Metadata must help to find the data among the networks (data catalogues). That is the purpose of defining 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, who are both dreaming of the "Esperanto" of data formats, as English became the international communication language. Computer technology has improved a lot in the past decade and we have slowly moved from ASCII formats (easy to read by human eyes but not efficient for software), to binary formats (easy for software but not easily shareable among platforms (Windows, Unix, etc), to self-descriptive, multi-platform formats (NetCDF, HDF, etc) that allow more flexibility in sharing data on a network and are read by software that are commonly used by scientists [4,8]. One important point for metadata is to identify a common vocabulary to record this information. This is easier to achieve for a specific community such as physical operational oceanography as the number of parameters is small, but becomes more difficult when addressing 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 the ISO19115/19139. The availability of usable standards both for data description (catalogues, metadata, formats) are discussed more in detail in Keeley et al. [11] and Snowden et al [21].

5.3.2 Progress made on standardize quality control procedures

Assessing the quality of an observation is important for sensible and efficient use of the data. The quality control (QC) procedures have to take into consideration the allowed delay of data delivery. In real-time most 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 results 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 bar? 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 have 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 and velocities.. But for other parameters there is still a lot of discussion on going before consensus can be reached (Burnett et al. [6], Pouliquen et al. [15], Reed et al. [17]). The next step is to provide uncertainty estimates for the data.

5.4 The Interoperability layer:

Data Assembly Centres can do enormous amounts of work in building very good, documented and reliable products, but if they are not easy to find by potential users, they may reside forever only on the data centre disks. With the explosion of Internet capabilities 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.

This layer firstly allows discovery of which 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 to the products. Such product catalogues have progressed much in the past ten years and although the standards

(ISO19115, ISO 19139) are still evolving, they are already successfully implemented in SeaDataNet (European project federating European National Oceanographic Data Centres) and IOOS and are recommended at the European level within the INSPIRE directive and endorsed by some software providers like ArcIMS, which are widely used in the cartography domain. Protocols like CS-W (Catalogue Service for the Web) to interrogate products through the internet are developing and should be soon certified [4]. With this discovery level, users can find not only observational data but also derived products such as climatologies.

When data providers endorse a free and open, data policy then it's possible to know how to build additional services for viewing and downloading the data from a central point without knowing where the data are stored. For example within the GODAE experiment a lot of products (observations and model outputs) have been made freely available in NetCDF format on servers, made available through OPeNDAP services, which provide a consistent means to access precise subsets of large data holdings (Hankin et al. [8], Blanc et al [2]). Methods for improving the standardization and serving of ocean data are discussed in detail in Blower [4] and Hankin [9] plenary papers.

6. PERSPECTIVES

In the past decade, data exchange between partners has improved mainly fostered by the satellite and the physical oceanographic communities. Computer technology and data management systems are no longer an obstacle. It has been shown that sharing ocean data in near real-time is beneficial not only for the community but also for the scientists that deploy instruments because anomalies in platforms were detected earlier by comparison with nearby measurements. Cost efficiency was achieved in implementation with better knowledge of existing platforms, and benefits from networking activities to improve common quality control procedures. Moreover it has eased the work of the National and Thematic data centres to collect the important metadata that are essential for future reprocessing activities and climatology improvements.

Presently a lot of data have been acquired in past centuries and it remains difficult to determine whether a trend is related to the accuracy of the data or a real trend in the ocean. We are currently developing new services and products such as operational oceanography services the same way as in meteorology 20 years ago. This data and product sharing in (near) real-time needs to be extended to other communities such as biogeochemistry or marine geology. It doesn't happen by chance and needs a strong willingness and involvement of the community to regularly improve and update the necessary vocabularies, QC procedures and metadata content, to be efficient and sustainable in the long term.

We 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' behaviour,
- **comparable**: this implies agreement on data standards and quality control procedures both in real-time and delayed mode to provide datasets independent of which platform sampled them. It also implies good version control of the data that allows knowing the processing steps 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 encourage people to release their data more rapidly
- **Provided at an appropriate level for the user**, for example: agencies such as the European Environment Agency (EEA) and the European Marine Safety Agency (EMSA) require high-level products that integrate information from various sources including in-situ instruments, satellite platforms and numerical models.

Finally we should not re-invent the wheel but instead use the expertise in other communities (including meteorology and deep ocean operational oceanography) to improve data exchange at the regional level and improve answers to monitoring and crisis management activities. Ocean observing systems will contribute to larger integrating initiatives such as the Global Earth Observing System of Systems (GEOSS), which is discussed further in [4], therefore it is extremely important to work with other communities in moving forward. The way forward for data system and integration will be discussed more in detail by Hankin in his plenary paper [9].

7. REFERENCES

1. Belbeoch M. & Co-Authors (2010). "The JCOMM in situ Observing Platform Support Centre: A decade of progress and remaining challenges, *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2)*, Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.

2. Blanc F. & Co-Authors (2010). "Data and product serving, an overview of capabilities developed in 10 years", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
3. Blower J. & al. (2009). "Serving Godae Data and products to the ocean community", *Oceanography Magazine Special issue on "The revolution in Global Oean Forecasting-GODAE: 10 years of achievement", Vol 22, NO3,* September 2009.
4. Blower J. & al., (2009). "Ocean Data Dissemination: New Challenges for Data Integration", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 1),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
5. Borges J. & Co-Authors (2010). "A global sea surface carbon observing system: dissolved inorganic carbon dynamics in costal environments", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
6. Burnett W. & Co-Authors (2010). "Quality Assurance of Real-Time Ocean Data: Evolving Infrastructure and Increasing Data Management to Monitor the World's Environment", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
7. Donlon C. & Co-Authors (2010). "Successes and Challenges for the Modern SST Observing System", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
8. Hankin S. & Co-Authors (2010). "NetCDF-CF-OPeNDAP: Standards for Ocean Data Interoperability and Object Lessons for Community Data Standards Processes", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. &Stammer, D., Eds., ESA Publication WPP-306.
9. Hankin S. & Co-Authors (2010). "Data management for Ocean Sciences – the Next decade", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 1),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
10. Le Traon P.Y. & Co-Authors (2010). "Ocean, "Data Assembly and processing for operational Oceanography: Ten years of achievements", *Oceanography Magazine Special issue on "The revolution in Global Oean Forecasting-GODAE: 10 years of achievement", Vol 22,* NO3, September 2009.
11. Keeley R. & Co-Authors (2010). "Data Assembly Infrastructure", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 1),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
12. Keeley R. & Co-Authors (2010). "Data Management System for Surface Drifters, , *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
13. Mc Phaden M. & Co-Authors (2010). "The Global Tropical Moored Buoy Array", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
14. Prandle D. & N.C.Flemming (1998). "The science Base of EuroGOOS", *EuroGOOS, publication N°6,* 58..
15. Pouliquen S. & Co-Authors (2010). "Argo Data Management", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
16. Pouliquen S. & Co-Authors (2006). "Coriolis a French project for operational oceanography", *EuroGOOS publication, N°23.*
17. Reed G. & Co-Authors (2010). "IODE Ocean Data Standards Forum", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
18. Roemmich & Co-Authors (2010). "Argo: Observing the global ocean", *in Proceedings of the*

"OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2), Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.

19. Send U. & Co-Authors (2010). "OceanSITES", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
20. Smith S. & Co-Authors (2010). "Automated Underway Oceanic and Atmospheric Measurements from Ships", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. and Stammer, D., Eds., ESA Publication WPP-306.
21. Snowden D. & Co-Authors (2010). "Metadata Management in Global Distributed Ocean Observation Networks", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
22. Testor P. & Co-Authors (2010). "Gliders as a component of future observing systems", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. &Stammer, D., Eds., ESA Publication WPP-306.
23. Vanden Berghe E. & Co-Authors (2010). "Integrating biological data into ocean observing systems: the future role of OBIS", *in Proceedings of the "OceanObs'09: Sustained Ocean Observations and Information for Society" Conference (Vol. 2),* Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306.
24. Edwards,J.L. (2004) Research and societal benefits of the global biodiversity information facility. BioScience, 54,485–486.
25. Graham,C.H. et al. (2004).New developments in museum-based informatics and applications in biodiversity analysis. TrendsEcol.Evol. (Amst.), 19,497–503.
26. Guralnick,R.P., A. W. Hill and M. Lane. (2007) Toward a collaborative, global infrastructure for biodiversity assessment. Ecol.Lett., 10,663–672.
27. Lane, M. (2006) Information infrastructure for global biological networks. Microbiol.Aust., 27,23–25.
28. Stein B. R. and J. Wieczorek (2004) Mammals of the world: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics, 1, 14–22.