Overcoming obstacles to data integration: a case history

Abstract

Hans Mose Jensen[1], Julie Gillin[2], Jacqueline Jones[3], and Patrick Roose[4]

ICES Data Centre is currently developing the Database on Oceanography and Marine Ecosystems, DOME. This database integrates over a century of physical, chemical and biological data which have previously been segregated into a variety of databases and file structures. DOME's development is supported by OSPAR, Cefas and ICES.

Integration of oceanographic data has presented several challenges including *near*-duplicated data, incompatible coding, loss of links between data, normalisation-breakers, etc. To provide guidance in dealing with these issues – and sometimes take hard decisions - OSPAR established an Intersessional Correspondence Group, ICG-DOME. ICG-DOME as well as Cefas is represented on DOME's steering group.

Typical problems of data integration and pragmatic solutions will be presented. DOME's structure will be outlined, and selected DOME features will be demonstrated.

Keywords: Integrated oceanographic database system; integrated marine ecosystem database system; DOME.

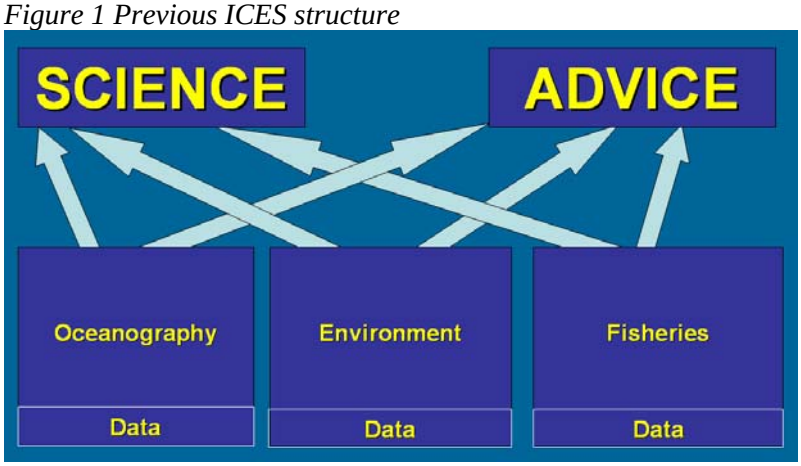
[1] ICES, H.C. Andersens Blvd. 44-46, DK-1553 Copenhagen V, Denmark [tel: +45 33 67 36; fax: +45 33 93 42 15; email: hans.jensen@ices.dk] [2] ICES, H.C. Andersens Blvd. 44-46, DK-1553 Copenhagen V, Denmark [tel: +45 33 67 12; fax: +45 33 93 42 15; email: julie@ices.dk] [3]Cefas Burnham Laboratory, Remembrance Avenue, Burnham-on-Crouch, Essex CM0 8HA, United Kingdom [tel: +44 (0)1621 787200; fax: +44 (0)1621 784989; email: jacqueline.jones@cefas.co.uk] [4]MUMM, 3de en 23ste Linieregimentsplein, B-8400 Ostend, Belgium [tel: +32 (0)59 24 20 54; fax: +32 (0)59 70 49 35; email: proose@mumm.ac.be]

Background for data integration in ICES

In order to understand the challenges faced in the data integration process within the ICES Data Centre (ICES DC), it is important to understand the history of data management in ICES. The Database on Oceanography and Marine Ecosystems (DOME) is presently the main integrated data system being developed in ICES, however the process of integrating data is going on at many levels involving many persons and skills outside the ICES Data Centre.

Before 2004, the ICES secretariat was divided in three separate scientific areas (see ***Figure 1***), Oceanography, Environment and Fisheries. Each of the scientific areas had their own scientific manager responsible for the data management and advice within their specific scientific area. Each of the sections handled their own data in their own data systems. The workflows from submission to data system were also independent including quality control procedures etc. The handling of data queries was also handled separately depending of the scientific area.

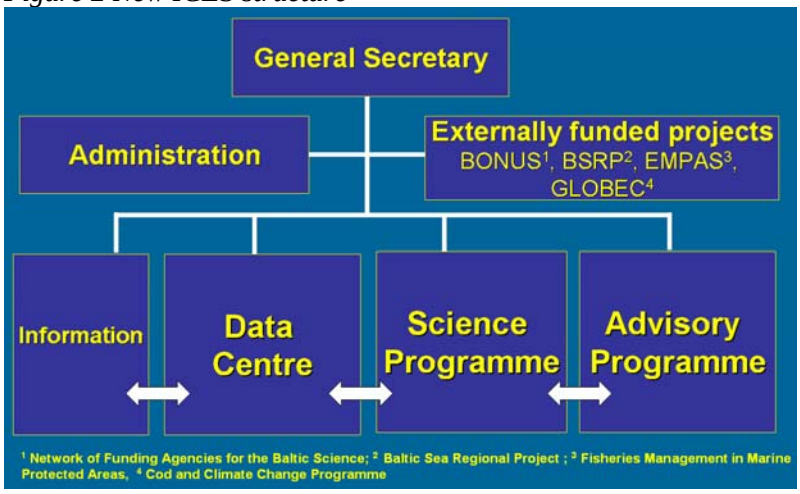
Figure 1 Previous ICES structure



These three separate scientific sections served the requirements of the ICES science programme and ICES advisory process independently.

In 2004, the ICES secretariat was reorganised (see ***Figure 2***). One of the main reasons for the reorganisation was to be better prepared for giving advice according to the ecosystem approach. In the new ICES structure, a central Data Centre is responsible for managing the data of all the scientific data disciplines covered by ICES. The Data Centre serves both the Science and Advice programmes according to their specific requirements.

Figure 2 New ICES structure



One of the important tasks for the new ICES Data Centre (ICES DC) is the integration of data, data systems, workflows etc. that was previously taken care of in three separate sections – and in three very different ways. The integration process has proven to be challenging at many different levels because the previously separated data handling procedures are deeply buried within a cascade of data systems, conversion programmes, quality control procedures, exchange formats, coding systems etc.

DOME background

The DOME (former SKY) project was conceived in 1999 in conjunction with the development of a new environmental reporting format (version 3.2). During 2002 and 2003, the ICES secretariat analysed possibilities for integrating the environmental and oceanographic data systems and concluded that is was necessary to have a database with a

single entry point for environmental and oceanographic data, but that is was not practical to make a fully integrated database.

These conclusions resulted in the SKY database being developed as a number of discipline – specific tables but with common elements placed in a generic shared area. SKY was abandoned for a number of reasons – mostly of an organisational nature.

In 2004/2005, SKY's basic concept was revived via DOME (supplemented with support by OSPAR and Cefas). The DOME database system is being developed in three sequential phases:

Phase 1 (2005-2006):

- Implementation of the overall structure
- Implementation of essential functions (e.g., data selection and export, automatic update of external codes such as BODC PD, ITIS, ERMS)
- Implementation of the internal interface
- Integration of the following data:
  - Contaminants and Biological Effects of contaminants including diseases in Biota (CDEF)
  - Contaminants and Biological Effects of contaminants in Sediment (CES)
  - Contaminants and Biological Effects of contaminants in Seawater (CEW)
  - Water Bottle data types

Phase 2 (2006-2007):

- Enhancement of essential functions (e.g., additional export formats)
- Implementation of optional functions (e.g., production of generic data products)
- Implementation of unit conversion routines for products (tentative)
- Integration of the following data:
  - Biological Community data
  - CTD and underway data

Phase 3 (tentative - 2007):

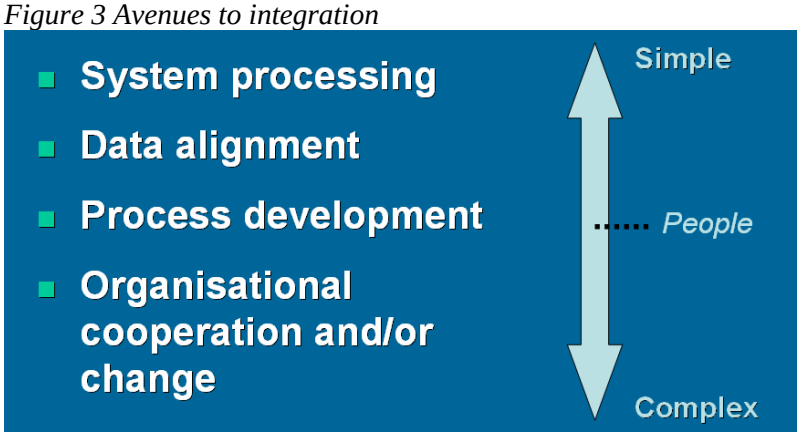
- Implementation of data editing functions
- Implementation of web interface (online data downloading, etc.)
- Integration of the following data (if feasible)
  - fisheries data

The integration and normalisation of data in the DOME project has revealed several challenges. To guide ICES DC in dealing with these issues and find solutions, OSPAR established an intersessional Correspondence Group, ICG-DOME. Cefas is also represented in the DOME steering group.

Avenues to integration and challenges

The integration of data proceeds on along different paths (see ***Figure 3***), that in many cases are interrelated with each other. The figure ranks processes from 'simple' to 'complex'. As more people get involved in changes, processes become more and more complex. Simple should however not be understood as 'easy', even simple system processing tasks can obviously be cumbersome and time consuming.

Figure 3 Avenues to integration



System processing

System processing comprises technical issues such as formats and data system structures. In ICES DC, a variety of exchange formats are being used or have been used in the past. The exchange formatted files are used for screening and QC of submitted data and in some cases the exchange formatted files is directly related to the data system in use. All data processing in ICES DC is in one way or another dependant on the usage of exchange formats. Data submission may not always be submitted to the Data Centre in an exchange format in which case the conversion to exchange format is part of the data processing task.

In the past a collection of exchange formats has been in use for environmental data. With the new Environmental Reporting Format version 3.2 (ERF32) they have all been combined into one format, however a lot of data is still held in the previous exchange formats. For oceanographic data, the ICES Oceanographic Format (IOF) has been in use since 1979. It was originally used on punch cards. In DOME Phase 1, only ERF32 and IOF data can be imported. Therefore a large amount of environmental data will have to be converted into the ERF32 format as a preparation for their import into DOME.

Data alignment

The different data systems and exchange formats in use in the ICES DC each have their own collection of codes used in the data (e.g. country codes, parameter codes, species codes). Country codes seem trivial, but issues such as ‘is Scotland a country’ (see below), and history (e.g., the split then reunification of Germany) has resulted in different coding systems for Country being used for different data systems within ICES DC. Converting from one format into another therefore requires mapping of country codes. There is no separate country code for Scotland in the ISO 3166 country code list. Because of this code lists based on ISO has been modified (by ICES) to include e.g. Scotland.

For environmental data, a parameter coding system (PARAM) based on CAS (Chemical Abstract Service – register of chemical substances) has been developed in cooperation with ICES working groups. Matrix and method information are stored as separate entities linked to a parameter. For the IOF formatted oceanographic data, a number of fixed parameters (e.g. salinity, depth/pressure, temperature) are used, that can be mapped with the ICES-PARAM coding list. For other parameters in the IOF formatted data, the BODC data dictionary (BODC DD) has been used. Because the BODC DD parameter codes also include method information and matrix, the conversion to ICES-PARAM codes is not trivial. Also mapping only the parameter names between the two coding systems cannot be done explicitly. One parameter code in the ICES-PARAM coding list can map to more than one parameter code in

the BODC DD and again these codes may each be mapped back into many codes in the ICES-PARAM list. It has therefore been decided to use the ICES-PARAM coding list for ERF32 formatted data and the BODC DD parameter codes for IOF formatted data in DOME. This has implications for extractions of data from DOME, where some parameters will have to be extracted using two or more parameter codes from the BODC DD and the ICES-PARAM coding list.

Another data alignment consideration is unit handling. Traditionally the exchange format used standard units for different parameters, but with the introduction of ERF32, free units were introduced. In order to make integrated calculations and output products from DOME it is necessary to make conversions between units. Converting to a set of standard units is a function of both the matrix and the parameter analysed. In DOME, parameter units will be imported as they have been submitted without being converted, the unit conversions will be done on the output side of the system. Some conversions cannot be done explicitly. The classic example is the conversion between parameter measured per mass and per volume. A parameter is submitted in µmol/kg and another in µmol/l. A correct conversion between these two units would require a complete profile data series in order to integrate the equation of state - which hardly ever is available. Another example is conversion between parameters measured on dry weight and wet weight basis requiring knowledge of the water content in the matrix.

In DOME, units will be stored as they are submitted. Conversion to standard units will be performed on the output side of the system. This ensures that potential conversions errors and conversion factors (based on average condition assumptions) are isolated and do not affect the actual data in the database.

Process development

It has been necessary to initiate processes in order to facilitate the integration process. Species coding is an example where a process has been initiated in order to align standards.

In 2002, ICES decided to use ITIS (Integrated Taxonomic Information System - see www.itis.usda.gov) as our one common species coding system. The decision was made because the previously used species coding systems, NODC taxonomic number and RUBIN (RoUtine for Biological Information) ceased to be updated and thus needed replacement.

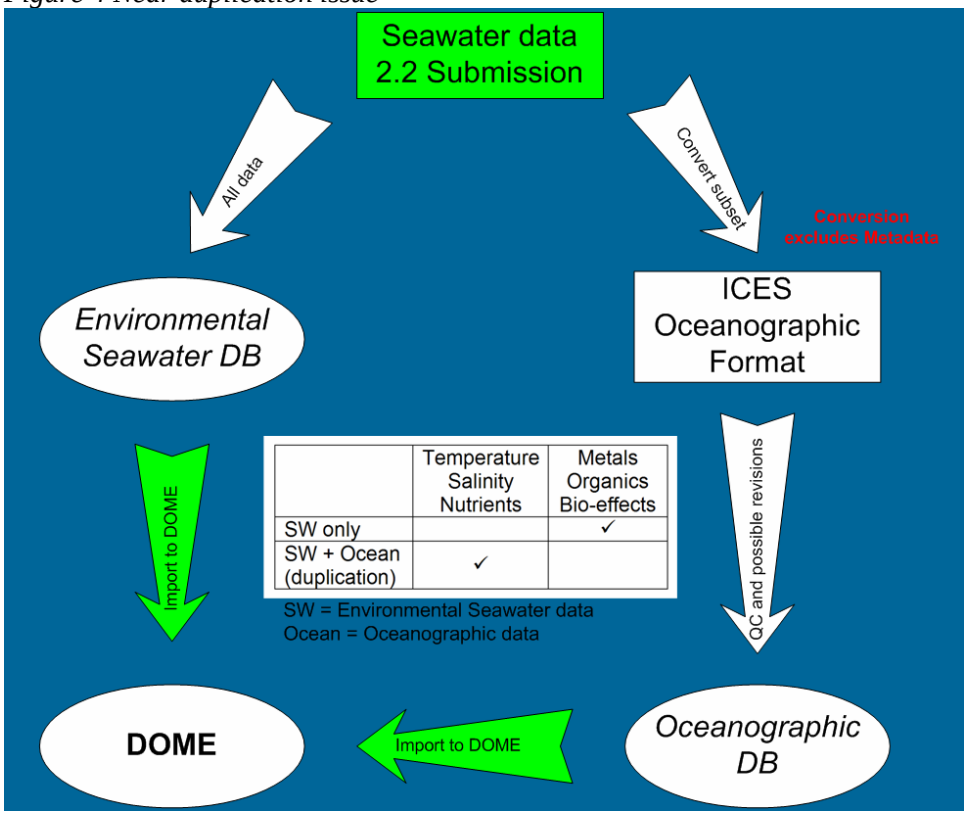
ITIS accommodated the need for fish species names, but was less adequate for biological community data. Many plankton and sediment-living species are not included in ITIS. Also, ITIS focuses on North America, so many European species are missing.

To overcome these deficits, ICES has initiated a cooperation process between ITIS and ERMS (European Register of Marine Species - see www.marbef.org/data/erms.php). A meeting with ITIS, OBIS, EurBIS, ERMS, BODC, GBIF and ICES was held in 2005. This resulted in a project outline on how the cooperation can be implemented. The intension is that ERMS will cooperate with ITIS to act as an external reviewer of European marine taxa to ensure a faster update for new species into ITIS.

Overcoming a *near* duplication issue is another example of process development. Due to the former organisational structure, physical and chemical oceanographic data (referred to as ‘oceanographic data’) have been stored separately from hazardous chemical oceanographic data (referred to as ‘environmental seawater data’) despite an overlap in data scope.
For many years the hydrochemistry subset of the environmental seawater data has been copied into the oceanographic dataset (converted into the IOF format) excluding the metadata (see ***Figure 4***). The oceanographic data were quality controlled according to procedures used

in the former oceanographic department. The corrections were not synchronised with the mother-dataset in the environmental department and consequently the two datasets drifted apart. Because position and time may have been changed separately for the two near duplicated datasets, it is no longer possible in an easy/automated way to identify the subset of duplicate data within the oceanographic data.

Figure 4 Near duplication issue



In order to avoid importing duplicate data into DOME when importing the legacy data, a process has been approved in cooperation with OSPAR and HELCOM to outline a possible solution. It has been decided that:

- Duplicated parameters will be imported from the oceanographic dataset because these data are considered to have a higher quality control level than the duplicated subset in the environmental seawater data. The consequence of this is that meta-data for these parameters will not be imported into DOME. The metadata are however still archived in the original submitted data files.
- Non-duplicated parameters in the environmental seawater data will be imported into DOME together with the metadata information.
- Some parameters (SUSP, DOC and POC) will be imported both from the environmental seawater data and from the oceanographic data. This is because these parameters are essential for the interpretation of other chemical data and are considered cofactors. Duplication of these parameters in outputs from DOME will be avoided by flagging them as cofactors and excluding them from direct downloads.

Organisational cooperation and/or change

Organisational changes might have to be initiated in order to facilitate the integration process. The organisational changes in the ICES secretariat facilitated the move of ICES towards an ecosystem approach and non-fragmentation of the data systems. Organisational changes

cannot do the job alone, but they are sometimes needed in order to change course. This is obviously a hard and slow ongoing process.

One of the important experiences in the ICES Data Centre is the importance of a common language even on the most fundamental concepts. There has been extensive discussion on concepts like: “What is a station?”, “What is a cruise?”, “What is a data submission?” E.g. a “cruise” can be conceived as being someone tugging a bucket on a beach, but is the bucket then a “ship”? It is important to come up with a common understanding of these concepts in order to talk the same language.

The attempts to define station names as derived information to be normalised in the ICES DC data systems is a good example of the challenges to be faced when exploring the possibilities for normalisation and the importance of a common understanding.

Traditionally only “areas” (e.g. ICES rectangles) were reported with the data. Station names were not included in the oceanographic and environmental data systems. Around 5 years ago, it became apparent that areas were not sufficient and OSPAR needed station names to group trend analysis data. A common definition of “station” as a geographical window (a specific geographical position and a range common for all stations) was needed. This was, however, non-practicable because the range varied according to different environments and the purpose of monitoring. Some areas show large environmental fluctuation over small distances whereas others are very stable over large distances. Another problem is how point source stations should be identified. The point source should be measured e.g. at an outlet and the acceptable distance from the outlet is very small.

As a consequence of this and for traceability reasons, it was decided to include station name in the ERF32 format directly related to the data. However some issues remained:

- Legacy data did not contain station names
- How should submitted station names be validated
- It is possible to consider station name as derived information – can it be defined as a function of latitude, longitude, date, time, purpose of monitoring, type of station, etc? If this is the case, there would be no reason for storing the station name in DOME, rather it could be looked up

To accommodate these issues and considerations, a station dictionary was designed so that legacy data can be allocated station names as a function of a set of parameters. Data submitted with station name can be validated by looking up station name in the dictionary and comparing it to the submitted station name.

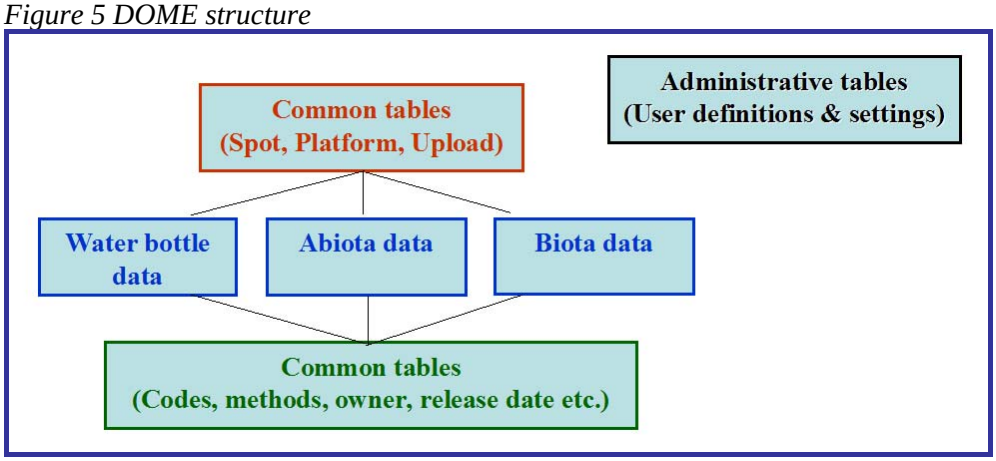
Recognition and consensus on the parameters to be used in defining a station were discussed and exposed widely differing ideas on station definitions because stations are used in many different ways. It also became apparent that looking up a station name in the station dictionary could not guarantee exactly one matching station per lookup. Consequently station name will be stored together with the ERF32 formatted data in DOME. The station dictionary can then be used to look up station names for legacy data and also to validate station names submitted as ERF32 formatted data files.

So far station name is only mandatory in data submitted for OSPAR.

DOME structure

DOME is structured (see ***Figure 5***) with a common top structure for all data types. The common top tables defines the data collection event in time and space (spot) together with other related sampling information, Cruise, Ship, Station Name, Purpose of monitoring, Monitoring Programme, Reporting organisation, Country, Monitoring year etc.

Figure 5 DOME structure

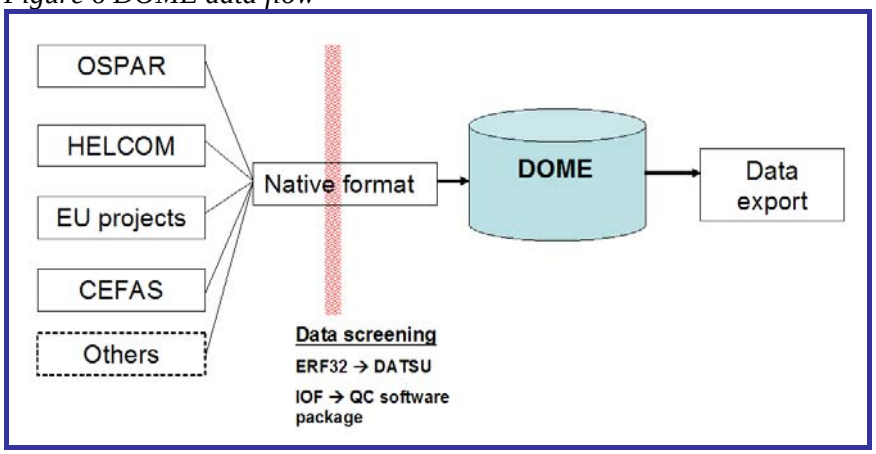


Below the common top structure are the data type specific structures: Water bottle (Oceanographic hydrochemistry data – IOF formatted), Abiota (CES and CEW – ERF32 formatted) and Biota (CDEF – ERF32 formatted). The DOME structure has not yet been entirely integrated, because of fundamental differences between the specific data types. E.g. there are differences in the hierarchical structure of Biota and Abiota data that cannot efficiently be integrated in a normalised relational database structure.

The bottom structure of DOME consists of method information, ownership and release dates together with code tables. These tables are common for all data types. DOME will contain data that references several coding systems, the ICES reference coding system (RECO), BODC DD, ITIS, ERMS and possibly others. Instead of connecting directly to each related coding system, all codes used in DOME tables will be copied to a table structure inside DOME. The DOME coding tables will be updated automatically with new codes (from external coding systems) whenever data using a new code are being imported into DOME. This will ensure a faster processing, because only codes of relevance for DOME will be stored in the DOME internal coding tables. Triggers will ensure that changes in external code tables are immediately and automatically propagated to DOME.

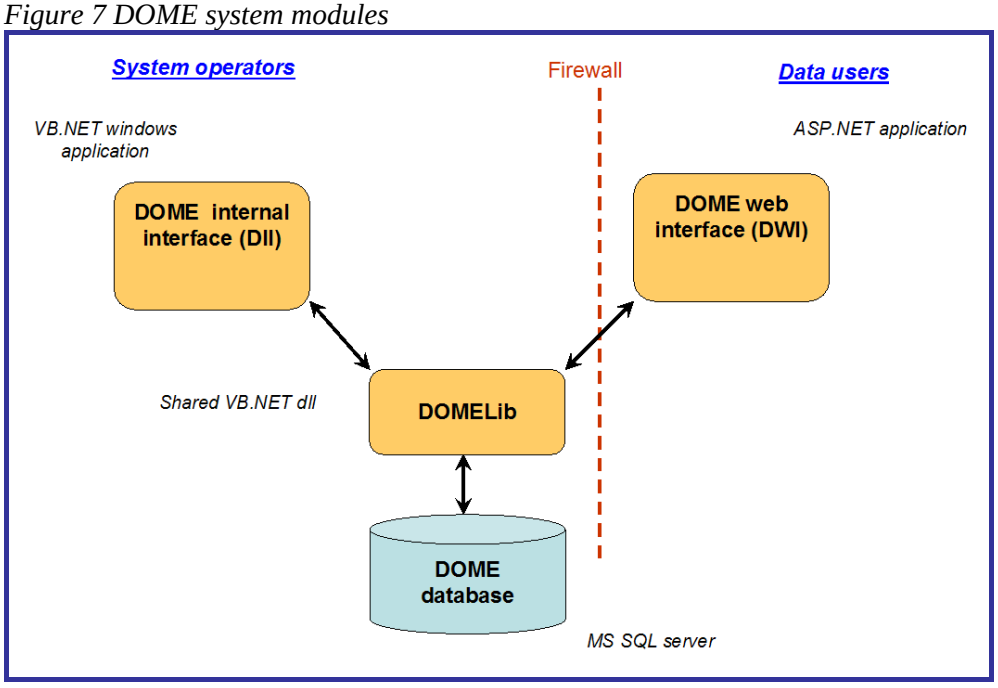
The data to be integrated in DOME arise from a variety of sources (see ***Figure 6***). They may or may not be submitted to ICES as native formatted files (native exchange formats for DOME are IOF or ERF32), but in order to be imported to DOME they must be in native format. The native formatted data files are screened using DATSU (DATa Screening Utility) for ERF32 and a QC software package for IOF formatted data files before the import into DOME. At present, data editing facilities are not planned in DOME. Data can thus only be changed by replacing previously imported data with a new dataset.

Figure 6 DOME data flow



The modular structure of DOME can be seen in ***Figure 7***. System operators will use the internal interface to operate the system, whereas external users will have access to data through a web interface (scheduled for Phase 3). Database access procedures will be shared between the internal interface and the web interface in a common dll module. This will also facilitate the usage of other possible interfaces (e.g. iSea - interactive Spatial Explorer and Administrator from Cefas).

Figure 7 DOME system modules



Concluding remarks

The integration and normalisation of data in the ICES secretariat is considered to be a long term process. DOME Phase one is one of the important steps. Organisational integration within the ICES secretariat is an ongoing process which promotes the integration of data. Ultimately, data integration has a number of costs and benefits.

Benefits

- Wider application. Having a common access point for the data increases the exposure of data to more users.

- Increased compatibility with other data types. The compatibility increases as a result of data being aligned and new international standards being applied (e.g. common code usage)
- Higher quality due to possibility for cross checks
- Rational operation due to fewer systems and structures needing maintenance
- Better exploitation of generic facilities. Handling data in data integrated systems makes the advantages of having generic facilities and procedures across data disciplines obvious, leading to better exploitation of the possibilities for building generic structures.

Costs

- Increased complexity. Developing an integrated system is complex and complex challenges will have to be faced and solved as described in this paper.
- Less specific/flexible systems. The systems might be experienced by users as being less flexible when it comes to the specific requirements of one particular data type