Biological data and metadata initiatives at CSIRO Marine Research, Australia, with implications for the design of OBIS

Tony Rees and Kim Finney
CSIRO Marine Research • Hobart Australia

Introduction

In 1997, our present agency was formed from a merger of two pre-existing research Divisions of Commonwealth Scientific and Industrial Research Organization (CSIRO) Australia, CSIRO Fisheries and CSIRO Oceanography. Following the creation of this new enlarged Division, named CSIRO Marine Research (CMR), a "critical mass" of staff concerned with processing and managing research data, in particular those collected by our research vessels, were gathered to form the nucleus of a new Divisional Data Center.

Prior to the establishment of the Data Center, data collected by individual Divisional research projects remained with project personnel, as did any general knowledge of their existence from a corporate perspective. As a focal point for Divisional data management activities, one of the key functions of the Data Center was to gain a measure of what data existed within the Division and to characterize the accessibility of these data. This information needed to be collated and communicated within the Division and distributed to external research collaborators and other interested external parties.

A second and related function was to develop mechanisms to provide a common access point for Divisional data. Divisional research data could be expected to exist in a plethora of formats and on a variety of media. But merely accessing these data in their raw and disparate forms was considered limiting, so an additional task for Data Center staff was to devise a solution that would allow for some degree of data integration, data mapping and data visualization. All of these requirements were complicated by the need to deliver Data Center services across three sites, located within different Australian States.

The strategic approach adopted to improve CMR's data management . . . has relied heavily upon the World Wide Web (WWW) as the standard framework for gathering information.

Setting up the systems necessary to achieve these goals has taken the Data Center three years and the task is still ongoing. The strategic approach adopted to improve CMR's data management has encompassed the introduction of both procedural and technological changes and has relied heavily upon the World Wide Web (WWW) as the standard framework for gathering information from Divisional staff and for disseminating information back to them. Choosing to use the WWW as the backbone of our systems has proved invaluable because it has allowed us to deliver the same service no matter what type of platform is being used by CMR research staff and no matter where these staff reside. If they have access to the web, they have access to the complete range of Data Center online services. Service delivery is influenced only by the speed of the user's computer and the type of network connection available.

The CMR Divisional data management solution has involved the development of two new databases and the substantial upgrade of a third. The result is a loose "federation" of applications that can interact and which all communicate with users and with each other over the internet. These applications operate on a common database system (ORACLE 8i), and are therefore easily able to exchange and share data "views."

The new systems that we have developed are the Marine Laboratories Information Network (MarLIN) and the Structured Query and Information Delivery (SQuID) System. MarLIN is a sophisticated, mediated, metadata directory system, while SQuID is a Divisional data warehouse with data visualization capability. The pre-existing system we have upgraded is called CAAB (Codes for Australian Aquatic Biota), which is a taxon

management system that assigns taxonomic codes to species or groups of species. These codes have the advantage of being consistent and persistent over time, and are utilized for taxon handling in both MarLIN and SQuID.

The remainder of this discussion explores in more detail the various components of the CMR Divisional data management system and focuses on design issues that we have had to address that will no doubt be encountered again by designers of the future global OBIS system.

"MarLIN" – CMR's metadata system

Metadata is a term for information about data (rather than the actual data points or values themselves). This information is captured in a structured and meaningful way and is generally made readily accessible, in much the same way that a library catalog is made available to describe a library's holdings. By storing metadata about CMR datasets in a relational database, information entry, data searches and information retrieval can be made comparatively simple for the user. Most repetitive items of information are stored only once, which not only minimizes the volume of managed data and simplifies its maintenance, but also facilitates the design of applications incorporating picklists and pull-down menus, both of which aid the entry and storage of key metadata items in a rapid and standardized manner.

Metadata is useful in a variety of ways, for example in supporting:

1. data *discovery:* e.g. to find out what data exist on certain topic or species, collected from specified geographic regions and time periods;
2. dataset *assessment:* that is, will data, identified in (1) above, be useable for the desired purpose, in terms of adequacy of detail, spatial and temporal coverage and data accuracy;
3. dataset *context:* e.g. supporting information that gives background about research voyages or projects during which the data were collected, equipment used, staff names, or publications that assist with interpretation of the dataset; and
4. dataset *access:* that is, where the data are physically stored, how can they be accessed (e.g. via the web or by contacting the data custodian) and what access conditions may apply.

To cover these areas, each MarLIN record has some 50 "metadata elements," including dataset title, abstract, collection start and end dates, keywords, formats, access constraints, bibliographic references, supporting information regarding research projects and research voyages. These elements represent a "superset" of the core metadata elements recommended by the Australia and New Zealand Land Information Council (ANZLIC) for spatial dataset descriptions in Australia and New Zealand (ANZLIC, 1996), and as such, allow interoperability of MarLIN with other systems in our region that subscribe to the same set of agreed core metadata elements.

In the context of designing systems for locating (and then accessing) spatially referenced data on marine organisms, MarLIN has several features of interest. The most obvious is that universal access to the system is made available via the WWW. Individual MarLIN records are tagged as to whether or not they are to be globally viewable, visible to specific internet domains only (i.e., accessible via the CMR intranet), or password accessible only. These security measures enable dataset descriptions of differing levels of confidentiality to be stored on the same system, harnessing a single user interface.

MarLIN metadata records are assigned bounding limits in space and time, which enables queries to be constrained either by geographic region, time period, or both. In the present version of MarLIN, geographic regions are approximated by the construction of a "bounding rectangle" within which all designated data points are deemed to occur, and spatial searches are also executed using "bounding rectangles." Optimally, searches should operate on irregular polygons and this feature will be included in the next MarLIN release. Time information is held in the present version of MarLIN as start and end dates for an observation series. These features, sufficient for a "first pass" identification of possible data of interest, will be refined in the next version of MarLIN, to take account of data whose distribution may be discontinuous in either space or time.

MarLIN records are tagged with a variety of keywords, covering the types of research equipment used, parameters measured and species groups represented, as well as subject categories, all of which can be invoked at the "search" stage. The keywords list for "species groups" follows the list of controlled terms propagated by the Australian "Blue Pages" marine metadata system (AODC, 1996), and includes a hierarchy of terms ranging from "vertebrates" and "invertebrates" down to more specific categories such as crabs, whales, dolphins, or selected groups of fishes. In addition, MarLIN records can be tagged with up to 15 individual taxon identifiers, corresponding to particular species or designated species groups as defined in CAAB. All internal (database level) taxon-based searching and retrieval within MarLIN is performed using these numeric identifiers, while CAAB is used as a look-up table from which to derive the currently used taxon name for presentation to the user, in both the MarLIN search interface and in any results returned. The significance of this is explored further below, in the section on CAAB.

The result of a successful MarLIN search, which can be undertaken from anywhere in the world, is a link to one or more metadata records describing individual datasets held at our agency. Attached to a given record may be one or several hyperlinks providing access to further information via the web, and also to the actual data if they are available for distribution online. At present, this last type of link operates on pre-prepared "packets" of data that have been made ready for direct

web download. These data are gradually being ported into SQuID. Future extensions to SQuID will result in these links being reconfigured to dynamically extract or display relevant data directly from our data warehouse, subject to any release constraints which may be applicable to a particular dataset. Most recently, MarLIN has been slightly modified to function as the key component of the Division's new Archival Information System (AIS). Information packages that have been forwarded by research staff to the Data Center for archiving are referenced in MarLIN, and a pointer to the off-line data storage location is then recorded in the relevant metadata record. In future, it is anticipated that purchase of a mass storage system (e.g. DVD Jukebox) would enable us to use MarLIN to directly access archived raw datasets, as well as access qualified data stored on-line in SQuID.

The concepts and database structure for MarLIN version 1 (as developed over the period 1996-98) have been described in Rees & Ryba (1999), however the system is still subject to ongoing development as external requirements evolve and new methods of interrogating the database are implemented. To see the most recent version of MarLIN, the application should be accessed "live" on the web at the URL listed at the end of this article.

"SQuID" – CMR's Data Warehouse

SQuID is the most recent addition to the CMR "federated" data management system. Like MarLIN, it uses ORACLE 8i, but with the difference that it also invokes ORACLE's Spatial Data Option for the storage and manipulation of spatial coordinates. The client software is written in Java and uses Java's Remote Method Invocation (RMI) and Java Database Connectivity (JDBC) to connect to the underlying ORACLE data store (Figure 1).

The SQuID database schema accommodates marine biological, chemical and physical oceanographic parameters. At present, many of the data loaded are sourced from voyages on the Division's two ocean-going research vessels. But the data model is sufficiently flexible that it will readily accommodate the addition of parameters sampled from other platforms (e.g. satellites, floats, moorings and remotely operated vessels), as well as data from shore-based surveys. The schema has been designed so that sampled parameters are primarily referenced via spatial coordinates and a time stamp. This allows SQuID users to examine integrated datasets according to spatial and temporal constraints. A biologist interested in species distribution in a particular geographic area can also acquire any available habitat data (e.g. water column parameters and seafloor sediment composition) for that region and time period. This might be important, for example, if you are looking for any correlations between habitat type and species distribution. This is also of particular interest if models are to be employed, in order to interpolate between known data points to produce a species distribution map, or to plot the potential distribution of a species based on its known habitat or other environmental surrogate.

The SQuID data model is comprehensive in that all of the parameters recorded in the database have large amounts of associated metadata, some of which is drawn dynamically from MarLIN. This detailed metadata provides the user with enough information to determine data quality. For each recorded parameter there is information on any instrumentation used for sampling, instrument calibration data, equipment and device data including references to equipment deployment settings, and references to any analytical methods used. Each parameter recorded in the database also has a quality flag that is assigned by Data Center staff during data loading. These quality flags were developed in-house after reference to pre-existing standards developed by other agencies. None of the standards examined was considered comprehensive enough to cover the wide variety of data types found within the SQuID system.

The SQuID Java client (user interface) is invoked by the user from the CMR Data Center web page. To run the client, a browser Java Plug-in is required. If this is not already installed, it is freely available for download from the Sun Microsystems web site. The opening SQuID window is spawned by the user's browser and currently provides the option to graphically select data using spatial and temporal constraints and by selecting a specific "data stream" type (Figure 2).

"Data Streams" are units of data aggregation, chosen during the SQuID design phase. For example, a user can choose to select and view "Catch" data. The database will then be searched and all data that have

Figure 1. SQuID architecture

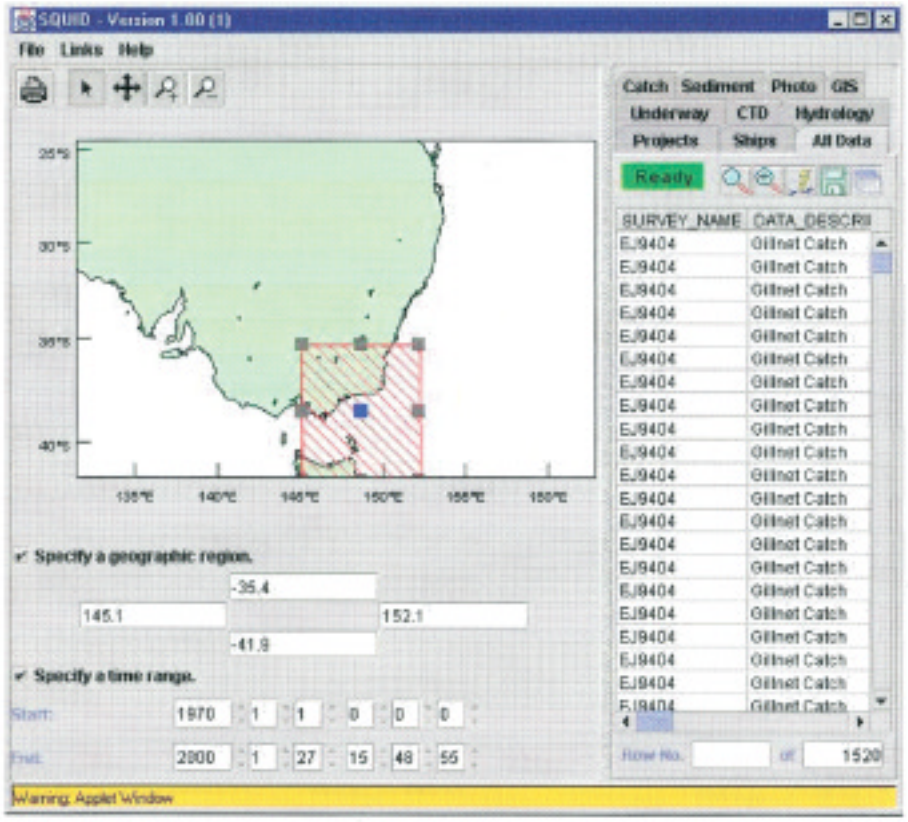

Figure 2. SQuID applet opening window.

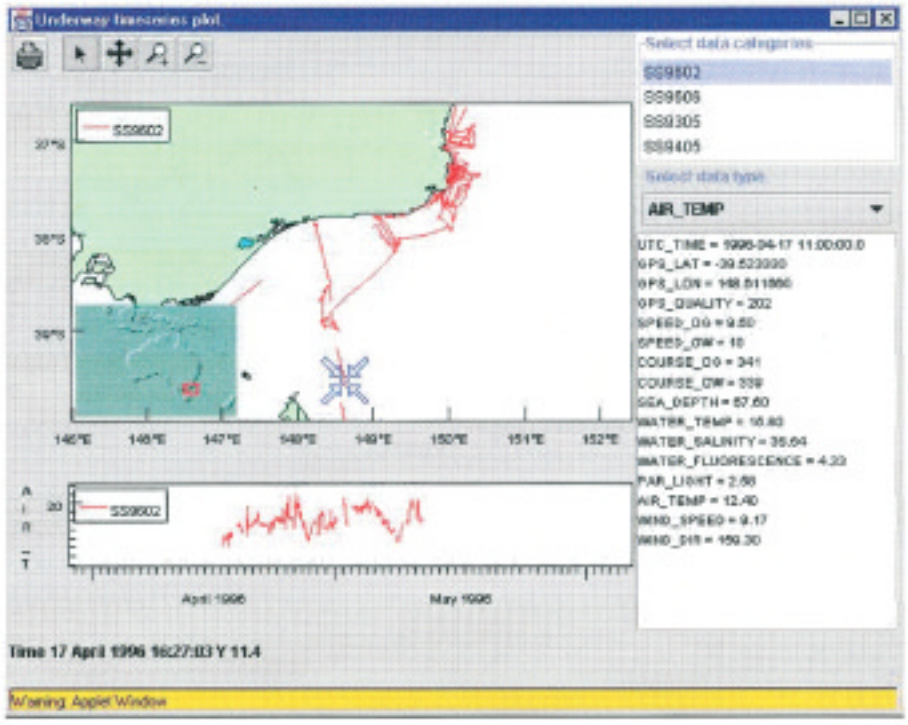

Figure 3. Example spatial display of time series data in SQuID.

been collected using trawl gear will be returned. These data are sent back to the client in tabular form and the user then has the option of selecting a variety of graphical display modes to visualise the data (e.g. data distribution plot, species length frequency plots or catch composition pie charts). If the data retrieved from the database are deemed suitable for the user's requirements, he/she can elect to have these data downloaded via email or sent as a file to the CMR FTP site, where the data can be acquired at the user's leisure.

For the same geographic location, the user can also extract other "Data Stream" types (e.g. CTD, Photo, GIS, Hydrology, Underway). The next version of SQuID will allow the user to overlay some of the graphic displays for these different "Data Streams", as well as allow for a greater degree of freedom in setting query constraints. Enhancements will include the ability to search for particular parameters or taxa within a range of catch records from a given region and/or time period.

The tools used to develop SQuID were developed in-house after reviewing a wide range of off-the-shelf products, which either did not offer the type of functionality we required or were prohibitively expensive (mostly in terms of ongoing software maintenance fees). SQuID has been developed making use of a pre-existing Java programming library, developed at CMR by staff outside of the Data Center. This library has many routines for manipulating and displaying data with spatial and time varying attributes; for example, it has enabled the Data Center to build capability into SQuID for displaying time series data, as shown in Figure 3. In the main applet window (top left) the spatial distribution of sampling for a particular parameter (Air Temperature) is shown using a ship's track. In the applet window directly below, the measured values are presented over time. Placing the mouse cursor on any data point within either of the graphical displays fills the parameter window (lower right) with a textual

summary of the data at the selected point.

Future development of SQuID is planned in order to provide access to multiple databases around CMR (and possibly elsewhere) via a common Java applet, i.e. SQuID will broaden its base to form a distributed database system. Data matching the user's query will be returned from wherever they reside in the system, and will be integrated into a common package for display and/or download as desired.

"CAAB" – Codes for Australian Aquatic Biota

CAAB, together with its immediate predecessor "FISHLIST", has existed since its development in the late 1970s by CMR scientists undertaking faunal surveys off the coast of Australia. The principal aim of CAAB is to provide a numeric code for every taxon of interest (whether named or un-named), so that the code can be entered against stored data instead of a name. When data are required for use, the code is retrieved and translated back to a name using the latest information in the CAAB database.

By this means, maintenance of taxonomic names stored in databases within CMR is simplified, since if the name of the taxon changes (which can legitimately happen at any time, due to improved taxonomic knowledge), the code stored with the data stays the same. In addition, where more than one opinion as to the "correct" name for a taxon exists (a real issue for some taxa), individual researchers can use any name they wish so long as they adhere to the CAAB coding system for the stored data. CAAB codes are based on a system of eight digits, two indicating the "category", for example echinoderm, fish, mammal, etc., three digits indicating the family within that category, and the final three for taxon number within that family. For example, the CAAB code for southern bluefin tuna *(Thunnus maccoyii)* is "37 441004," where "37" indicates the category "fishes," "441" indicates the family Scombridae (tunas and allies), and "004" is a unique taxon number within that family. This means that quite a lot of useful sorting or filtering can be done based on the codes alone, which is an advantage of CAAB over some other coding systems in use elsewhere.

CAAB is used in three ways in data management activities at CMR. First, a user can simply obtain a CAAB code on demand for any taxon currently in the database, for use in his or her own application as required. Second, a custom CAAB "taxon report" can be called from other applications using a standardized query to the CAAB database passed via the web, so long as the relevant CAAB taxon number is known. The third option is that the live CAAB data tables can be addressed directly by suitable applications, and incorporated into the application's functionality. For example these tables can be utilized to build picklists of available taxa for use in data entry or data search screens. This functionality is already incorporated into MarLIN, and will be available within SQuID in the next release.

The concepts and database structure for CAAB version 1 (as developed over the period 1990-95) have been described in some detail in Yearsley et al. (1997). However the structure of the underlying database has been changed somewhat since that report was produced, in particular to facilitate web access to the database and to provide increased space for additional codes to cover marine invertebrates, reptiles, birds, mammals and selected aquatic vegetation. The most recent version of CAAB (CAAB version 2) can be accessed "live" on the web at the URL listed at the end of this article.

Working together – a "federation" of databases at CMR

In developing the systems outlined above it has been important to share information between the different pieces of technology. This has required the development of dynamic system links for real-time communication. As each type of system serves a function in its own right, it has been much simpler to design and develop smaller well-bounded applications that work together, rather than attempt development of a monolithic application that tries to serve all purposes. Approaching the problem using a "federated" system has also meant that we have been able to bring pieces of infrastructure on-line and into service as soon as each is ready, and users have had the opportunity to provide us with feedback that has been useful in shaping subsequent developments.

In developing "federated" systems, however, it is critical to pay attention to communication protocols and to use standard codes or terms wherever possible, across the various application databases in the federation. For example, in our case, taxa are handled by their CAAB taxon code for all internal MarLIN and SQuID database operations. The only place where a taxon name needs to be changed, if this is required for any reason, is in the CAAB database, and such changes are automatically promulgated throughout the system without further action being required. Similarly, voyages of research vessels are given a unique identifier which is then utilized in both MarLIN and SQuID. This identifier can be used to initiate a search of the SQuID database whilst within MarLIN or could be used to call up a relevant MarLIN metadata record whilst using the SQuID interface.

Significance of CMR experiences to date in the context of a global OBIS

There are several issues with which CMR has had to grapple that will arise again when considering the design of a global OBIS. Firstly, it seems almost mundane to mention that the web would appear the only sensible framework on which to base OBIS, holding as it does the capability to unite proprietary databases regardless of their schemas and operating platforms.

Second, if it is anticipated that OBIS will eventually unite legacy systems, in addition to (or in place of) seeking to join new databases developed specifically for OBIS, mapping between these legacy database schemas will be one of the more difficult issues to resolve. XML (Extensible Markup Language) has the potential to play an important role in making schema descriptions transparent, and object oriented approaches currently appear the most suitable methods for distributed database searching. What will be missing and what is perhaps of great importance is some form of "community" agreement on the types of data objects that will form the content of OBIS, and what types of attributes those data objects will possess.

Third, as the orbit of OBIS grows to include a multiplicity of databases around the world, it is probably optimistic to expect that the same taxon will always be identified by the same name (or version of the name) in all contributing databases. In addition, names of taxa are subject to change with time, and it will be essential to be able to search and retrieve data stored under a variety of old, current, or (potentially) new names in the future, as required. Thus in order to function, OBIS will need either a system of "standard names", with an exhaustive (and possibly automated) synonyms search facility, or a system of universal taxon codes that can be used in place of names for carrying out the searching. An obvious candidate for such a coding system is the U.S. "Integrated Taxonomic Information System" (ITIS, 1998 onwards); however at present this system has a predominantly north American focus, which would need widening before it could function properly as the taxon dictionary for a global OBIS.

Fourth, for any centralized or distributed data storage system, a "metadata layer" will be required which will allow users to discover what data exist, and to ascertain if there are limitations on the usefulness of these data because of their method of collection or other factors. Whether this layer is integrated with the data storage, or is maintained as a separate system as is currently the case with our MarLIN database, is a question for the system designers. It will also be important to recognize that new global metadata standards currently nearing release (e.g. the International Standards Organization (ISO) TC/211 metadata standard) will facilitate interchange of metadata with other systems around the world if adhered to by OBIS, leading to the potential for a greater public visibility and profile for OBIS data.

REFERENCES/FURTHER READING:

Australia New Zealand Land Information Council (ANZLIC), 1996 onwards: *Core Metadata Elements for Land and Geographic Directories in Australia and New Zealand*. ANZLIC, Canberra. Available online at http://www.anzlic.org.au/download.htm.

Australian Oceanographic Data Center (AODC), 1996: *The marine and coastal data directory of Australia—The Blue Pages, Version 1.0 Documentation*. AODC, Sydney. Available online at http://www.erin.gov.au/marine/mcdd/documentation.html. Taxonomy list available at http://www.marine.csiro.au/cgi-data/mcdd/taxonomic.options.

Finney, K, 2000: SQuID—spreading tentacles to reap from the deep. *GIS User,* 40, 40-42.

Finney, K. and T. Rees, 2000: Metadata and data management activities at CSIRO Marine Research, Australia. In: *Using marine biological information in the electronic age: proceedings of a meeting held 19-21 July 1999.* K. Hiscock, ed., Marine Biological Association of the United Kingdom, Plymouth. (CD-ROM). Available online at http://www.marlin.ac.uk/conference99/Demonstrations/CSIRO/CSIRO.htm.

ITIS partnership, 1998 onwards: *ITIS—Integrated Taxonomic Information System*. WWW publication, available online at http://www.itis.usda.gov/plant-proj/itis/index.html.

Rees, A.J.J. and M.M. Ryba, 1999: MarLIN—a metadatabase for research data holdings at CSIRO Marine Research. In: *Marine and Coastal Data Management—First Australasian Marine and Coastal Data Management Conference Proceedings November 1998*. E.L. Tanner, ed., Earth Ocean & Space, Glebe, Australia, 102-111. Available online at http://www.marine.csiro.au/datacentre/ext_docs/marlinpaper.htm.

Yearsley, G.K., P.R. Last and G.B. Morris, 1997: Codes for Australian aquatic biota (CAAB): an upgraded and expanded species coding system for Australian fisheries databases. *Report, Marine Laboratories, CSIRO Australia,* 224, 120pp. approx.

Online links discussed in the text

MarLIN is accessible at
http://www.marine.csiro.au/dmr/database/marlin/

SQuID is accessible at
http://www.marine.csiro.au/datacentre/squid/

CAAB is accessible at
http://www.marine.csiro.au/caab