Metadata and Data Management Activities at CSIRO Marine Research, Australia

Kim Finney and Tony Rees

Divisional Data Centre, CSIRO Marine Research, GPO Box 1538, Hobart, Tasmania 7001, Australia
(E-mail: Kim.Finney@marine.csiro.au, Tony.Rees@marine.csiro.au
world wide web: http://www.marine.csiro.au/datacentre)

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

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

Abstract

CSIRO Marine Research holds marine biological, oceanographic and atmospheric data collected in the Australian region and elsewhere over a 60+ year period, and its data holdings are continuously growing as new research is carried out or data are acquired from elsewhere. Formed in 1997, the Divisional Data Centre has been developing and implementing new policies for data management which include the incorporation of "best practice" technology in large scale data management and metadata systems, and developing and adopting standards for data storage and access, data distribution and access policy, and other issues. This paper introduces our current activities in three areas: (1) development of a new Research Database to store data of a wide variety of types; (2) design and implementation of a metadata tool (our own "*MarLIN*" system) to describe Divisionally-held datasets; and (3) ongoing development of the Division's CAAB (*Codes for Australian Aquatic Biota*) taxon management system, used in our own and other agencies' databases across Australia. Example of other noteworthy activities in data and metadata management systems in Australia are also briefly discussed.

1. Introduction

The CSIRO Division of Marine Research (CMR) consists of about 300 staff, with the organisation's main headquarters in Hobart, Tasmania. Two other CMR laboratories are located in Queensland and in WA respectively. CMR is one of a number of research agencies funded by the Australian government, but we like to think of ourselves as the lead marine agency. We tackle most aspects of marine research except for marine geology.

The area of ocean in which we operate covers about 16 million km^2, which is more than twice Australia's continental area. This includes 61,000 km of coastline for mainland Australia and Tasmania alone. Australia is one of the most biologically diverse nations on earth and it is estimated that in our southern temperate oceans approximately 80% of the species found are endemic. To investigate this ocean expanse, CMR operates two ocean-going research vessels, one of which is a national facility. I think it would be fair to say that Australia needs more vessels and/or more vessel time if Australian biologists are to have any hope of adequately mapping the biota in the area under national jurisdiction.

CMR is actually a relatively new agency, having been formed from the merger of two existing CSIRO Divisions in 1997. At the time of merger the opportunity was taken to establish a Divisional Data Centre to provide a focus for data management within the new Division. The Data Centre has 12 permanent staff, qualified in various fields, mainly in the physical and biological sciences but also has staff with an IT background.

This presentation provides some background on CMR's approach to data management, briefly discusses the Data Centre's newest initiative in developing a marine data warehouse and outlines in some detail technology already developed to manage CMR's metadata and biological species coding.

2. Strategic Approach To Data Management

The Data Centre's function includes supporting the data collection and IT activities of the Division's two ships and through the Data Centre Manager's position also develops data management policy for CMR. The Data Centre promotes best practice in data management and steers Divisional data management activity through a Data Management Strategy (http://www.marine.csiro.au/datacentre/ext_docs/dms_web.htm - developed in mid 1997). This short to medium term Strategy is geared towards bringing the data management practices of CMR into the 21st century and is focused on affecting cultural change. It tackles policy, procedural and technology issues, all of which are essential, if the culture of the Division is to be changed to one that better embraces information management and one that is strategically equipped to exploit new information technologies.

The types of data management issues that require attention, within a research agency such as CMR, are often complex and need to be tackled from a number of angles. As you can imagine, resources as always, limit how much can be done at any one time, so although we have an overall vision within CMR of where we would ultimately like to be, we are taking an incremental approach to getting there.

Figure 1 demonstrates how we have approached building our data management infrastructure. The metaphor used is that of building a wall, with each brick cemented into place, making the next layer of the wall capable of supporting further, more technologically complex layers. We view the commitment of senior managers in promoting data policies and the cooperation of researchers in implementing those policies, as the fundamental foundation layer. On top of this layer we have implemented metadata database technology (i.e. the Marine Laboratories Information Network – MarLIN). This quite sophisticated tool has the capacity to catalogue, document and link to CMR's research data and products, as well as to the Division's existing financial and records management systems.

Another important layer in our wall is the development and implementation of standards (e.g. for data collection, data archiving and data clearinghouse development). In parallel with the creation of these standards the Data Centre has designed and is now developing a research agency marine data warehouse. This system will interface with the Division's metadata directory but also run as an independent application across the WWW.

Eventually we will distribute most of the Division's data through the MarLIN/Marine Warehouse interface, requiring development of appropriate technology to handle such things as data licensing agreements and e-commerce modules (to cater for the payment of any fees that may be imposed for data transfer).

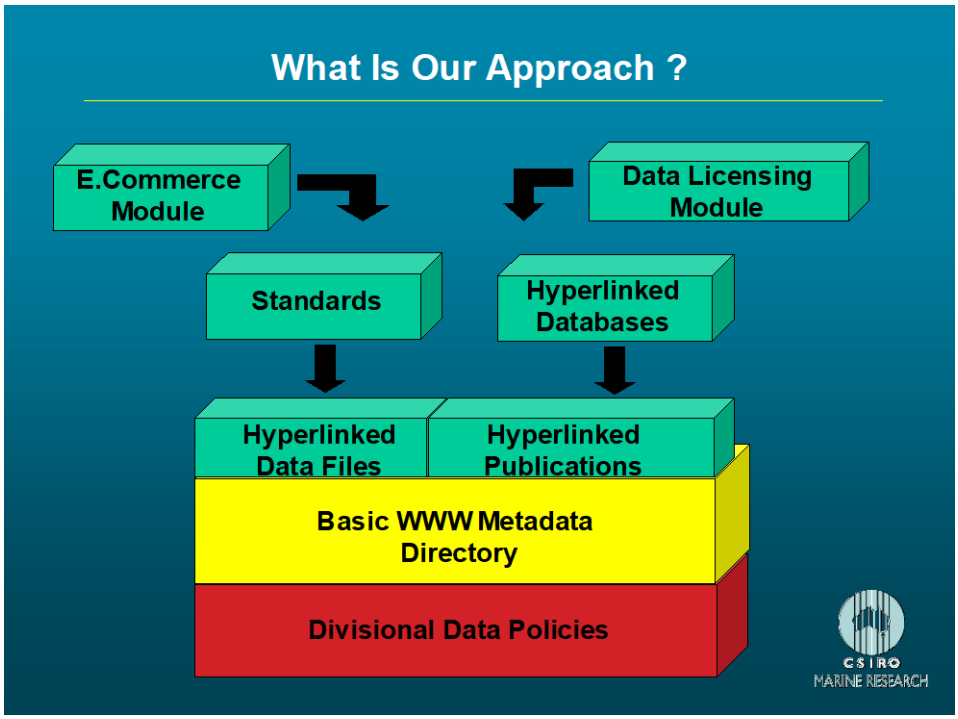

Figure 1 – Building CMR's Data Management Foundations

3. New Marine Data Warehouse For CMR

The new CMR marine data warehouse has been under development for a few months and has been named SQuID for Structured Query and Information Delivery system. Figure 2 shows the main components of the system. Using client-server technology the application interface is a JAVA applet connecting to an ORACLE 8i backend, through JAVA's remote method invocation (RMI) and WWW http protocol. The ORACLE database has been spatially enabled using ORACLE's spatial option software.

Considerable effort has already gone into developing an appropriate data model for the warehouse. This is conceptually shown in Figure 3. Most data managed within CMR can be considered to have been sourced/generated from some type of physical platform (e.g. ship, satellite, mooring - generally observational type data) or other source, such as software (e.g. marine model or GIS application). All of these sources are always related to a research project. Data sources are deployed (either physically or conceptually) to yield data streams. These data streams are in turn related to instruments or data collection devices (not shown in the diagram) and can be indexed into a summary table to facilitate high-level data retrieval. The data warehouse model is complex because of the wide variety of data streams that it must accommodate and because of the flexibility that should be provided to the user in querying the data.

The JAVA applet will provide user access to the data by a combination of spatial, aspatial and temporal queries without the need for the user to understand the underlying data model. Having retrieved the data of interest users will be able to visualize spatial distributions of subsetted data and produce simple data plots. If the data is of interest to the user and he/she wishes to acquire it for further analysis the user will have the option of down-loading the data (in a variety of supported formats).

Development of the CMR marine data warehouse is being partially sponsored by the Australian Surveying & Land Information Group (AUSLIG) under the Australian Spatial Data Infrastructure (ASDI) Partnership Program.

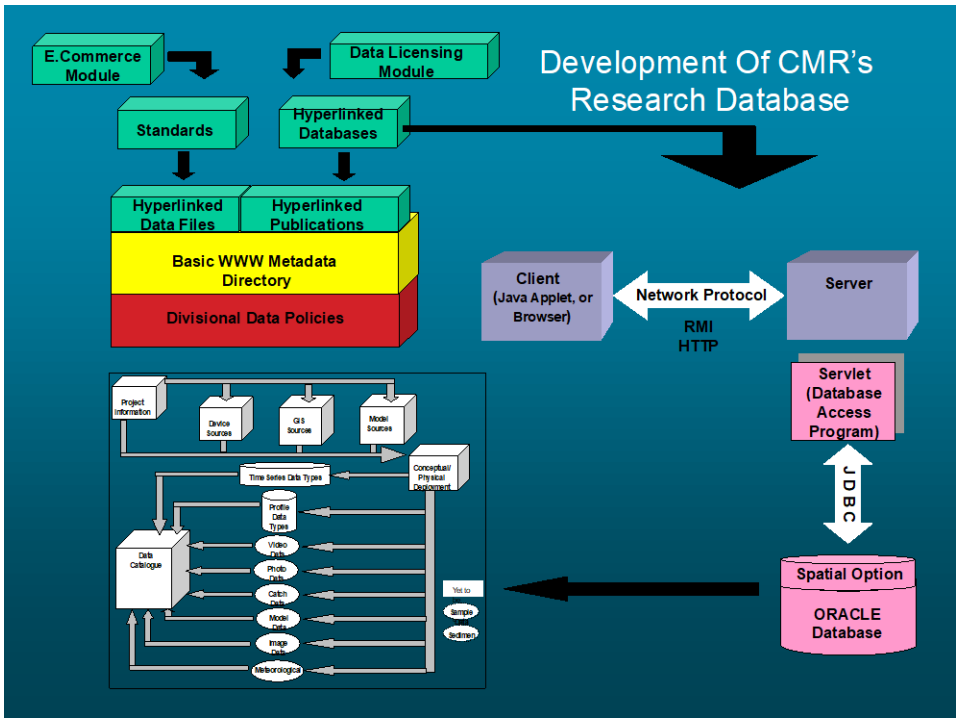

Figure 2. Components of The Marine Data Warehouse

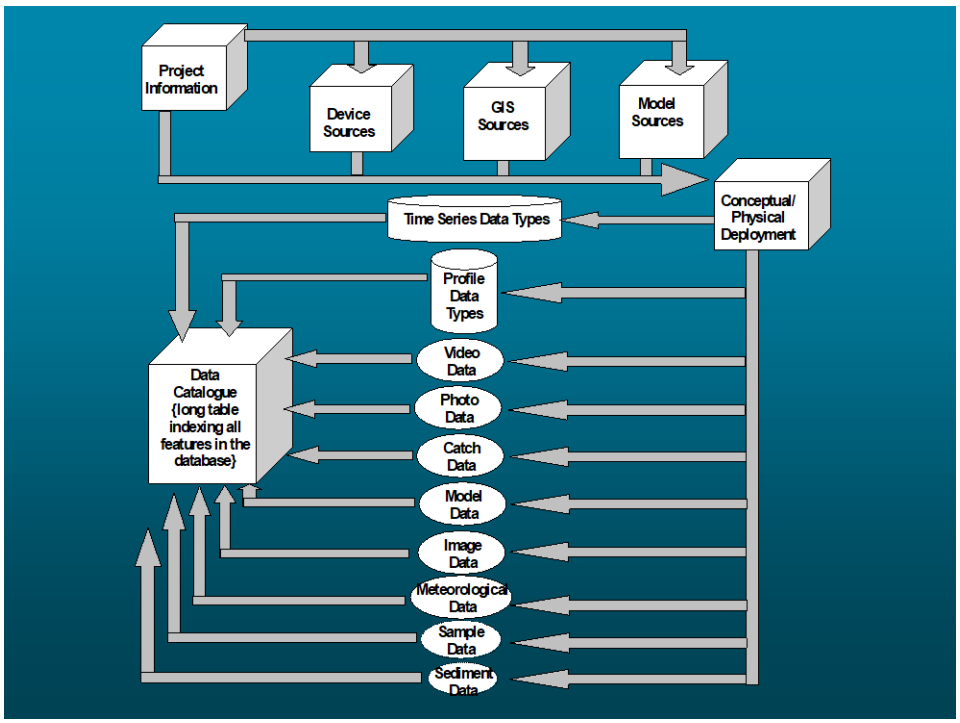

Figure 3. Conceptual Diagram Of SQuID Schema

4. *MarLIN* - CSIRO Marine Laboratories Information Network

4.1 MarLIN concept The concept for CSIRO Marine Research's MarLIN system was outlined in the Data Management Strategy developed for the Division by the Data Centre in 1997. The need was identified to design and build a metadata management system or metadatabase, to hold structured information (metadata) about the Division's data holdings, independent of the format or location of the actual data. As part of the design phase of MarLIN, key supporting information such as details of Divisional contact persons, projects, voyages of research vessels, etc. was added in relational database tables, which could be automatically incorporated into relevant metadata records as required. Examples of the types of data which can be described in MarLIN, and some of the sources of MarLIN supporting information, are shown in Figure 4. Since MarLIN has been designed from the start to be a web-based application, hyperlinks to additional documentation, graphics, and actual data (where accessible over the web) can be incorporated into the individual MarLIN records as appropriate.

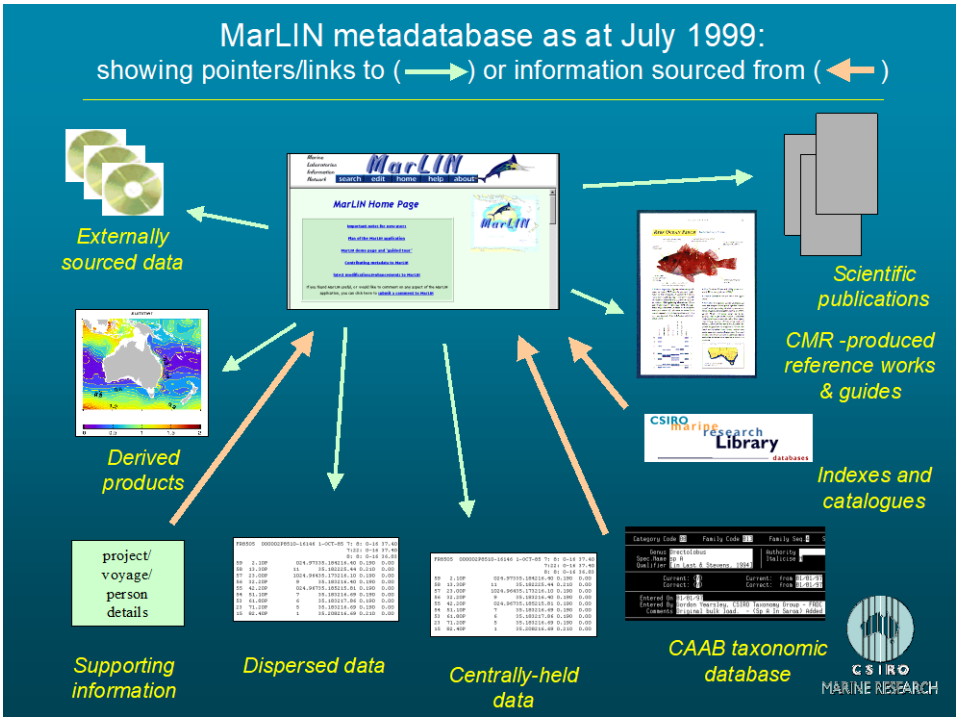

Figure 4. Conceptual position of the MarLIN metadatabase in relation to other Divisional resources

4.2 MarLIN design phase CMR's MarLIN system has been based on a pre-existing application developed for Environment Australia in 1996-7 and used by them for their in-house metadata directory ("Green Pages" or Environmental Data Directory: Environment Australia, 1996-). This software uses a combination of fixed HTML pages plus JAVA, Oracle PL/SQL and Oracle Web Agent (OWA) to allow the user to interact with an underlying Oracle Relational Database where the metadata are stored. The prototype MarLIN system was developed from the "Green Pages" application after consideration of CSIRO Marine Research's additional needs for information storage, and also the desire to be compliant with 2 external metadata standards within Australia: (1) the ANZLIC standard, primarily developed for terrestrial dataset description (ANZLIC, 1996-), and (2) the "Blue Pages" Marine and Coastal Data Directory (AODC, 1996). In addition, much of the user interface, functionality, and underlying data model was altered wherever it was apparent that improvements or extended functionality were desirable to suit the Division's needs. The MarLIN metadata format can be considered to be an "enhanced ANZLIC" scheme and currently incorporates some 50 metadata elements, some of which are themselves comprised of multiple parts or can be repeating. An overview of these, plus their correspondence to the basic

ANZLIC standard, Blue Pages elements, or additional Divisional needs, is shown in Figure 5. For a fuller description of the MarLIN system and its rationale, see Rees & Ryba (1999).

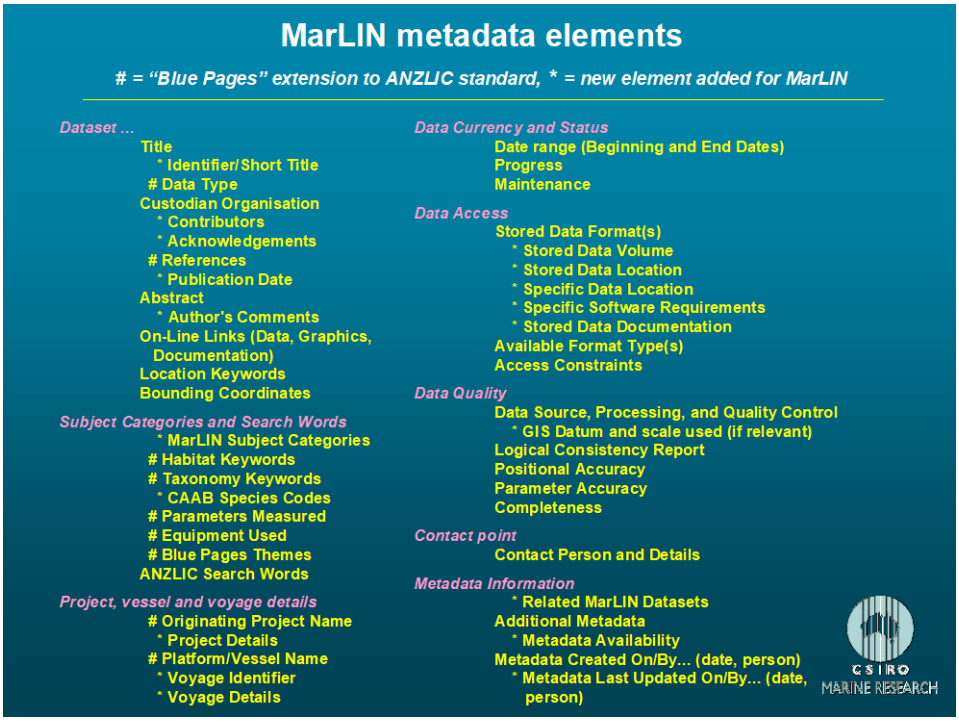

Figure 5. Overview of the main metadata elements held in MarLIN as at July 1999, as stored for any one dataset description (MarLIN metadata record)

A key "enhancement" over the subject-related indexing features in both ANZLIC (ANZLIC keywords) and Blue Pages ("Blue Pages Themes") has been the incorporation for MarLIN of a comprehensive, structured list of subject categories based, with permission, on the FAO "ASFA" (Aquatic Sciences and Fisheries Abstracts) scheme (FAO, 1983; Cambridge Scientific Abstracts, 1998). This system was considered to be the most appropriate for our agency's particular sphere of interest, and also gives the potential for future interoperability with other systems employing this scheme, for example the online version of ASFA available (to subscribing institutions) via the Cambridge Scientific Abstracts website.

4.3 MarLIN user interface MarLIN was made available on the web in a beta version in April 1998 and in a (more or less) stable "release 1" in August 1998, accessible from the CMR Data Centre website (http://www.marine.csiro.au/datacentre/). After entering via a home page, various search options are presented to the user, including "user defined search", "list datasets", "quick title search", "free text search" and "java search". As part of "user-defined search" and "java search", the user can define a region and/or time period to search by, and pick from keyword lists of various types, including project/custodian organisation names, vessel or voyage names, subject categories, habitat and taxonomic keywords, equipment or parameter listings, or individual species names, to set up their search. Some examples of these options and their implementation in the user interface as at July 1999 are shown in Figure 6.

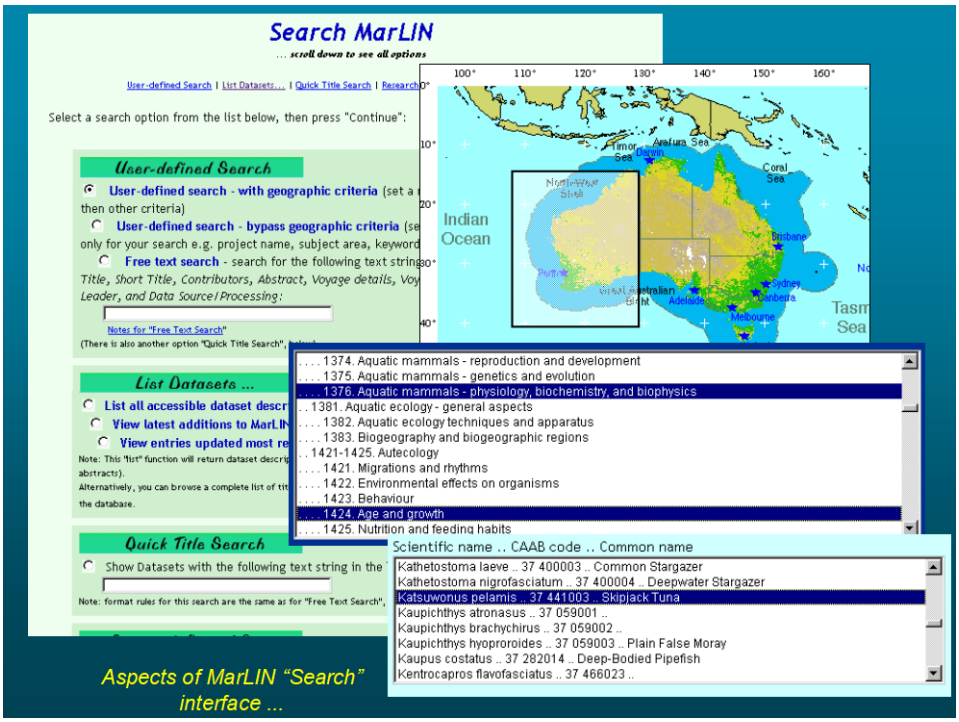

Figure 6. Sample features of the MarLIN search interface as at July 1999

After constructing and submitting a search to the MarLIN system, search results are returned to the user in one of several formats, depending on the route originally taken into the application and the type of information most relevant. For example, the "list" function returns a list of dataset titles, each hyperlinked to the full metadata record, while entering through a separate "research vessel data" interface results in a more graphical display to the user of summary information about each dataset including a thumbnail map of the dataset extent, links to cruise tracks stored as .gif images elsewhere on the Data Centre site, etc., as shown in Figure 7.

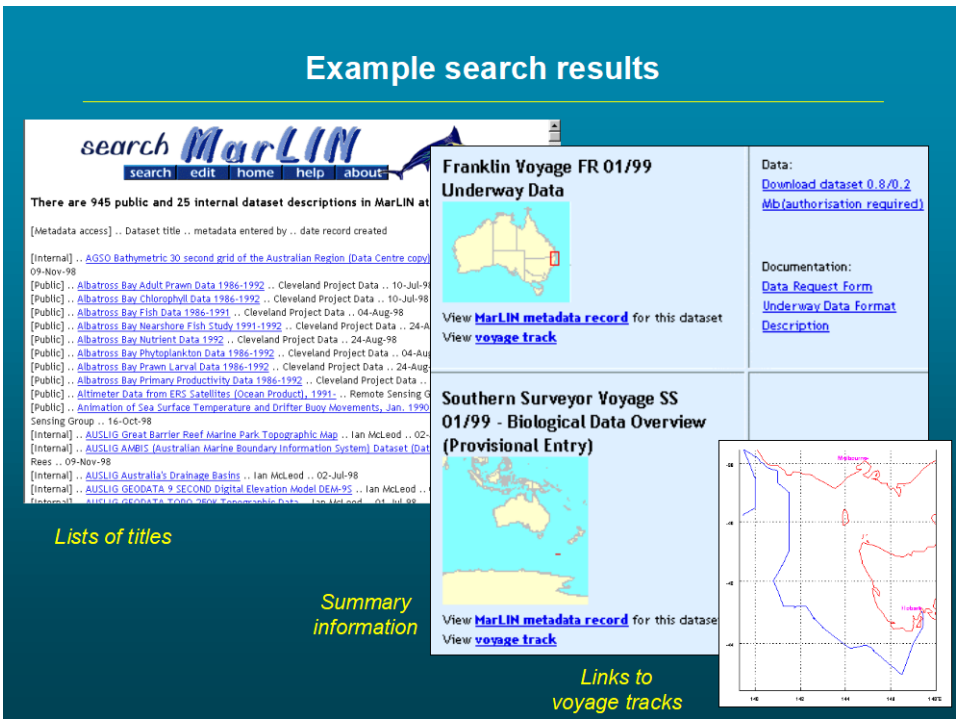

Figure 7. Examples of some formats for search results returned to the user after submitting a MarLIN search

Clicking on any hyperlink to the full metadata record results in display of detailed metadata for any dataset (e.g. as shown in Figure 8), together with on-line links to further documentation, images or data, as accessible via the web at other locations.

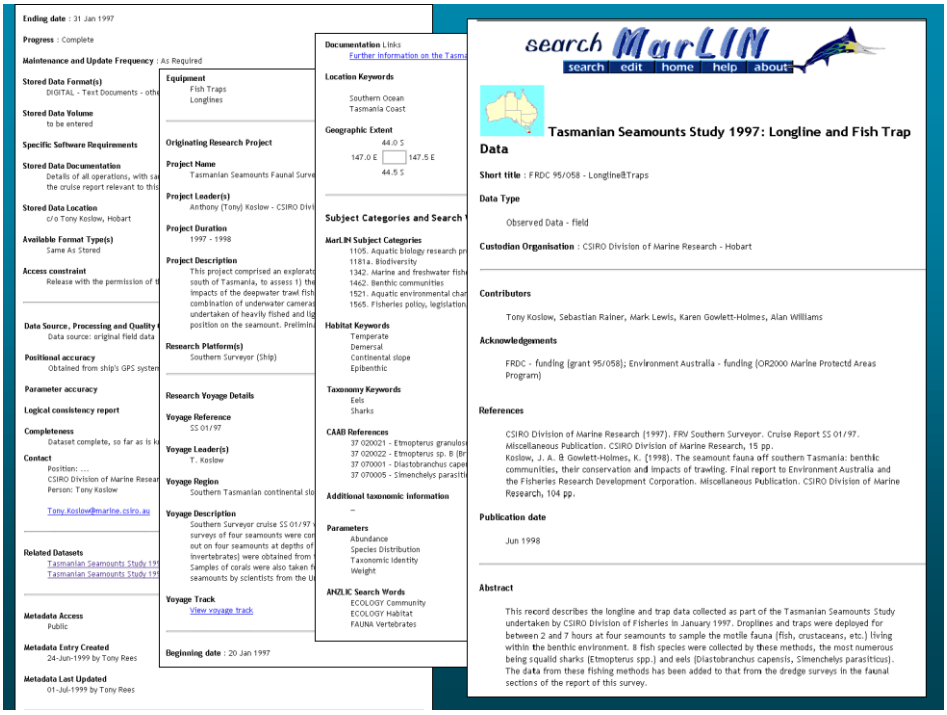

Figure 8. Sample MarLIN metadata record

4.4 MarLIN linkages While MarLIN's primary function is as an in-house metadata system for our agency, we have extended its functionality by building an automatic export system from relevant MarLIN metadata to generate the Division's contribution to the Australia-wide, distributed-system, "Blue Pages" Marine and Coastal Data Directory, of which the Division hosts a node on its server. The "Blue Pages" format requires each metadata record to be a single HTML file in a designated format, and a script has been writted to automatically create and then refresh the Division's Blue Pages entries from source data in MarLIN, on a nightly basis. This means that the metadata need only to be maintained in a single location, and the content of the Division's "Blue Pages" entries always reflects the latest data in MarLIN after the nightly update. A further advantage of this dual approach is that MarLIN records (which contain supplementary information) can be accessed directly from any Blue Pages record via a built-in hyperlink, and thus become accessible via the Blue Pages entry point and search facility. In addition, since "Blue Pages" records are discrete HTML pages, they can be accessed and indexed by proprietary internet search engines and web crawlers and their content becomes accessible to users searching the internet by whatever means, which is not normally the case for information residing in a database. These links are represented schematically in Figure 9.

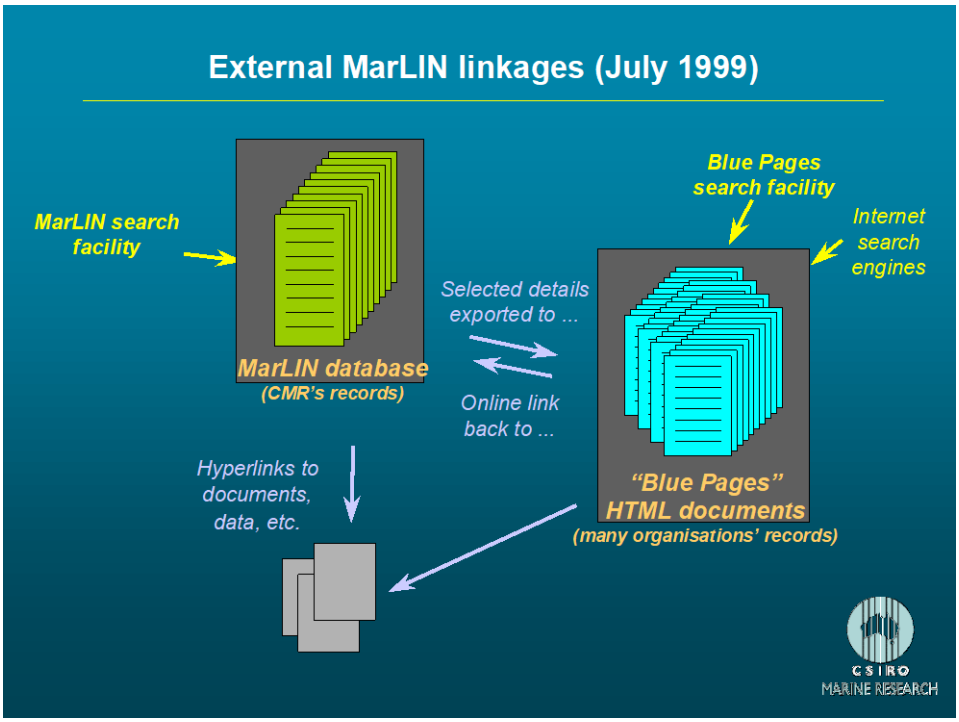

Figure 9. Schematic of linkages between MarLIN, "Blue Pages", search facilities and other www resources

4.5 MarLIN future development MarLIN is still in its initial release at this stage, and will be developed further as needs arise and opportunities are identified. Likely areas for future enhancement include incorporating "live" hyperlinks to other sources of Divisional information e.g. corporate and library databases, project data holdings, etc.; linking MarLIN to the Australian Spatial Data Directory (ASDD), a system which can interrogate remote databases simultaneously independent of their specific architecture, provided that a common core set of fields are implemented according to the ANZLIC standard; and possibly looking into cross-standard integration of metadata from a variety of institutions, using "crosswalks" to translate from one standard to another. A pilot project incorporating both the Australian MarLIN (which uses the ANZLIC standard) and NASA's Global Change master Directory (which uses the US FGDC standard) has already been undertaken by workers in another CSIRO Division, in consultation with ourselves (for a description of the background to this project, see Kuo, 1999). At the same time, since we are proceeding with the development of our Divisional research database as described above, it will be an ongoing task to establish direct links from the metadata level to the actual data, and employ common interfaces as far as possible for searching either or both systems concurrently.

5. *CAAB* - Codes for Australian Aquatic Biota

5.1 Background CAAB and its precursor, "FISHLIST", have been developed by staff of our Division and its precursors, CSIRO Division of Fisheries/Fisheries Research, over a 20+ year period. CAAB provides codes for aquatic organisms of interest to staff of our Division and other agencies around Australia, initially for fish species and commercial species of invertebrates taken in Australian waters, but also with other invertebrate taxa taken on research cruises as well as some vertebrates (marine mammals and reptiles). The CAAB system has come about in response to a variety of needs:

- Taxonomists needed a system for organising specimen collections and supporting information
- Field biologists needed a tool for rapid data entry (to include categories corresponding to "non orthodox" groups e.g. undetermined or new species, morphotypes, aggregate taxa)

- Data custodians needed a system for storing taxon-related information in a long term, stable form, independent of future name changes for taxa

- Automated sorting was desirable on the code alone, leading to the development of "intelligent" codes incorporating taxonomic hierarchy information within the code

The initial 6-digit coding system "FISHLIST" was upgraded to the first version of "CAAB", an 8-digit system, over the period 1992-1995, and the rationale for coding and a list of codes for fish species and "commercial" invertebrates has been published in the form of a Marine Laboratories Report (Yearsley et al., 1997).

CAAB codes are constructed according to a defined system, examples of which are shown in Figure 10. The first 2 digits in the CAAB code are a category code (for example 37 indicates fish, 41 indicates mammals). Of the remaining 6 digits, the first 3 are a family code within the major category, allocated according to a chosen systematic sequence so that (for example) orders, suborders etc. form contiguous blocks within the CAAB system. The final 3 digits are allocated sequentially starting with 001, as taxa are added to the system for that family, thus 37 020001 is the first taxon added for family 020 (Squalidae) of category 37 (fishes). Various other "species numbers" have pre-defined (reserved) meanings, e.g. 000 is used to indicate an undetermined member of any family, 900 and upwards are used for aggregate species, and 800-899 are reserved for local use only (temporary taxa, not added to the main system).

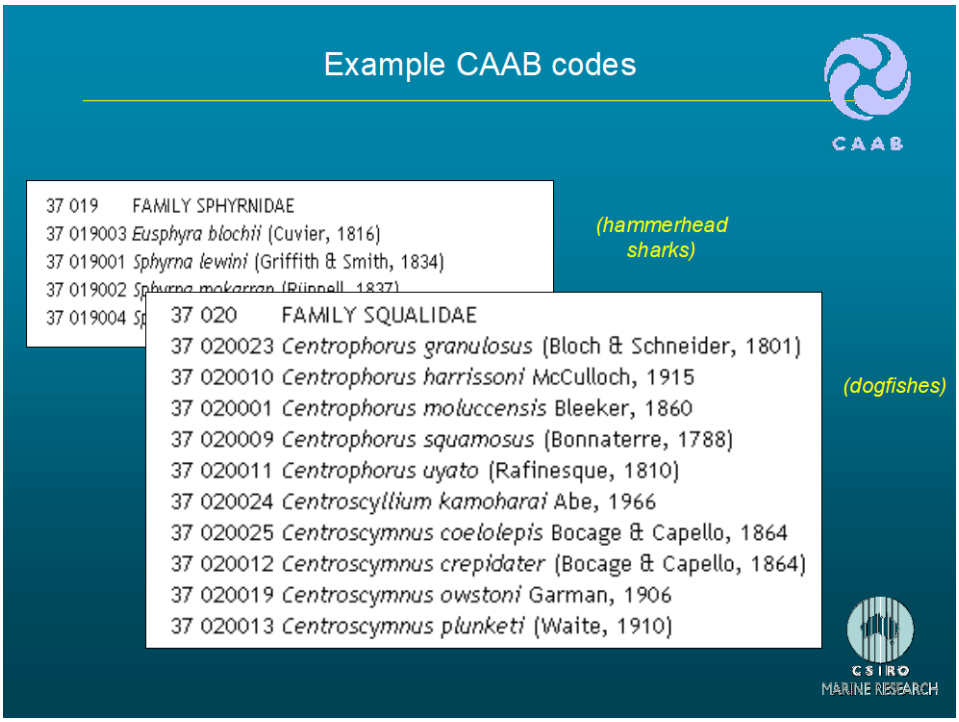

Figure 10. Example CAAB codes from two families of category 37 (Pisces).

As at July 1999, CAAB contains some 4500 codes for fishes (aiming to be a complete coverage of the Australian fauna, including some 200 taxa awaiting formal description), another c. 700 codes for invertebrates (chiefly crustaceans and molluscs), and some 100 codes for other vertebrates (marine reptiles and mammals). Further codes will be allocated on a continuing basis, including (in the near future) seabirds and marine angiosperms (seagrasses and mangroves).

5.2 Use of CAAB information CAAB codes are used in a comprehensive group of marine science agencies across Australia, including all of the key Federal Government agencies (Australian Bureau of Statistics, Australian Fisheries Management Agency, Bureau of Resource Sciences, CSIRO Marine Research, Dept. of Primary Industries and Energy) and the majority of the various State Fisheries Departments. CAAB codes are also quoted in various fishery guides and handbooks produced by the above agencies and others. At present, CAAB codes are supplied to interested parties either as a printed list (Yearsley et al., 1997), current to mid 1995, or an electronic data file for upload to users' systems. These links are illustrated in Figure 11. However, before the end of 1999 it is intended to construct a web- compatible version of CAAB so that users can query the "live" information in the database directly, without requests needing to be serviced by CSIRO staff (Figure 12). As part of this process, common name information for many taxa (not previously made available to external users) will be quality-checked and upgraded as necessary, and additional functionality will be incorporated, for example to allow generation of individual taxon-level reports, with hyperlinks to other taxa, etc. The facility to incorporate a variety of on-line links for any taxon will also be incorporated, for example to CMR-generated maps or photographs, plus additional background information held elsewhere. Using the web, it will also be possible to use CAAB as an entry point for MarLIN searches, and go from a taxon code attached to a MarLIN record to a CAAB report on that taxon, for additional information.

In addition, CAAB has been called into service as a pre-existing taxon dictionary and taxon management system for MarLIN and (when development is completed) the SquID database described earlier. This means that data and metadata stored in SquID and MarLIN, respectively, can use the taxon code as primary taxonomic identifier, and retrieve current scientific and common name information from CAAB on demand. Use of CAAB taxon codes also enables taxon-related data in MarLIN and Squid to be rapidly sorted according to CAAB major category, and systematic order within that category. In the prototype MarLIN application as at July 1999, codes, scientific names and common names of taxa were copied from CAAB during the construction phase of the MarLIN application, and reflect CAAB content as at February 1998. Later this year, MarLIN will be modified to communicate "live" with information in the CAAB database, so that changes to CAAB content will automatically be picked up when next querying MarLIN and retrieving species names or taxon codes.

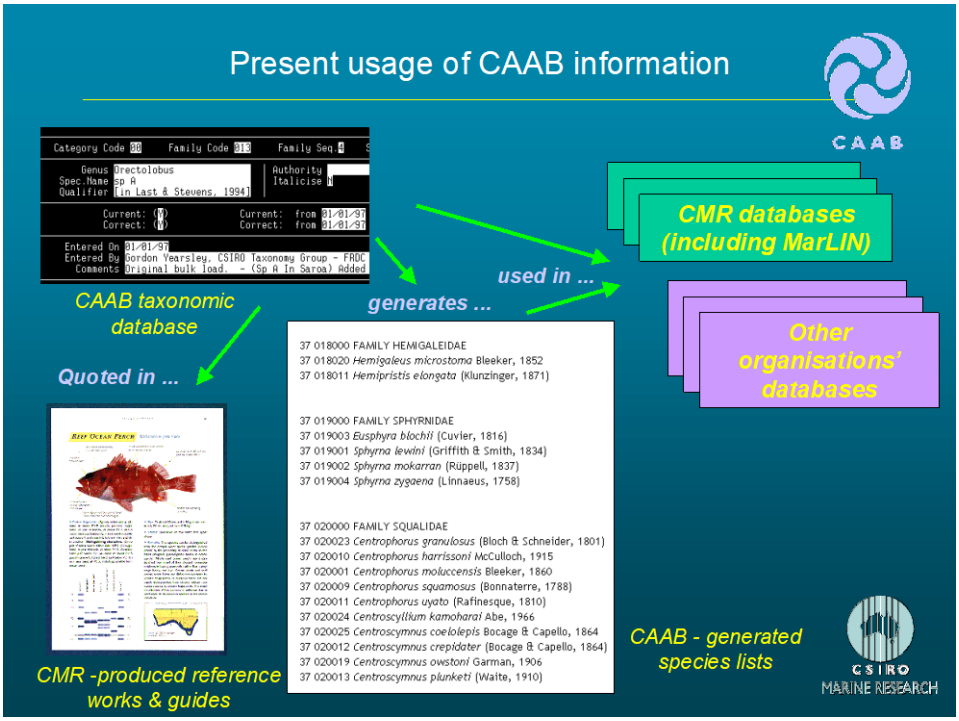

Figure 11. Method of operation of CAAB as at mid 1999

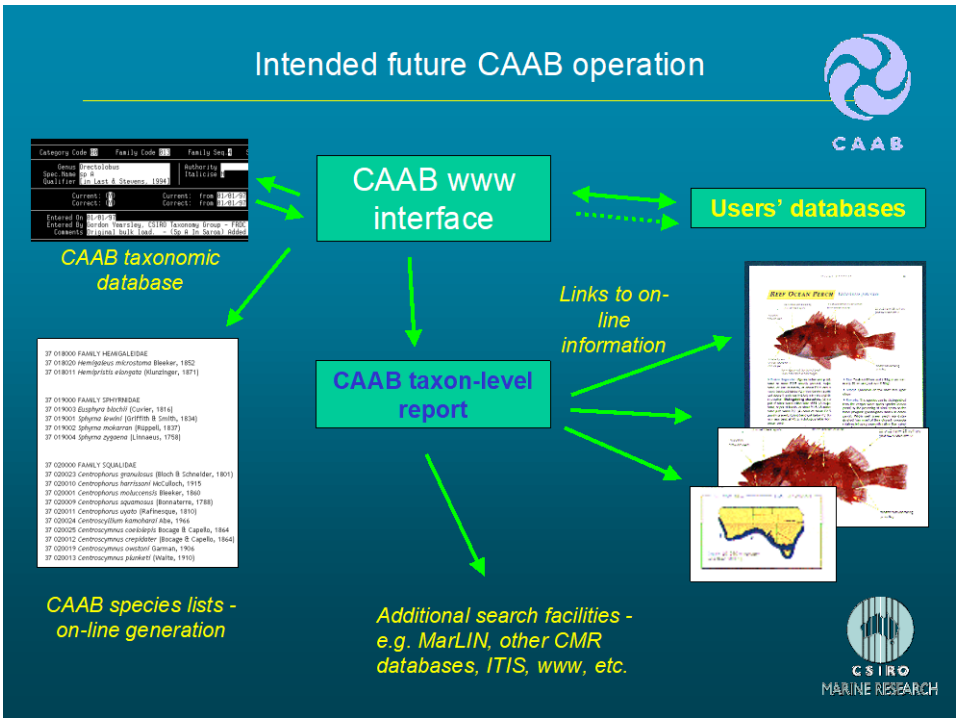

Figure 12. Intended future operation of CAAB (by end 1999)

6. Other data and metadata systems in Australia of possible interest

Before concluding this paper, it is worth mentioning that there are a number of other data and metadata activities occurring at other agencies around Australia which may be of interest to the designers of the UK MarLIN system. These include other metadata directories at other agencies (e.g. Australian Antarctic Division at http://www-aadc.antdiv.gov.au/Metadata/default.htm and the West Australian WALIS system (http://www.walis.wa.gov.au/), distributed data directories such as the "Blue Pages" (http://www.erin.gov.au/marine/mcdd/) and the Australian Spatial Data Directory or ASDD (http://www.environment.gov.au/database/metadata/asdd/), as well as prototype internet mapping applications such as the Australian Coastal Atlas (http://www.environment.gov.au:80/marine/coastal_atlas/). A more complete listing is available from the CSIRO Marine Research Data Centre site at http://www.marine.csiro.au/dmr/database/marlin/other_meta.htm.

7. Concluding remarks

The developers of the UK MarLIN system are in the fortunate position of being able to investigate a wide variety of other systems, either web enabled or not, which already emulate some of the functionality envisaged for their system. Many of the elements presented from activities at our agency have potential parallels within the proposed UK MarLIN system. While the eventual methods of implementation and underlying technologies may be different, many of the same philosophical questions will need to be addressed. In addition, the UK MarLIN faces the added complexity of (potentially) addressing data stored at remote sites and in a range of formats. Hopefully, some of the technologies and concepts already employed by ourselves and other presenters at this meeting will enable the developers of the UK system to decide more rapidly what to encorporate for their own needs, while remaining mindful of present and evolving external constraints (for example, the upcoming proposed ISO metadata standard: see http://www.anzlic.org.au/metaiso.htm for some introductory information).

8. References

Australia New Zealand Land Information Council (ANZLIC), 1996-. *Core Metadata Elements for Land and Geographic Directories in Australia and New Zealand.* Digital document, available online at http://www.anzlic.org.au/download.htm (as at November 1999).

Australian Oceanographic Data Centre (AODC), 1996. *The marine and coastal data directory of Australia – the Blue Pages, Version 1.0* [Documentation supplied to users]. Also available online at http://www.erin.gov.au/marine/mcdd/documentation.html (as at November 1999).

Cambridge Scientific Abstracts (CSA), c.1998. *Database Classification Codes: ASFA and Oceanic Abstracts*. WWW document, available online at http://www.csa.com/helpV3/classcod.html (as at November 1999).

Environment Australia, 1996-. *Green Pages - the Environmental Data Directory*. WWW application, available online at http://www.erin.gov.au/database/edd/ (as at November 1999).

Food and Agriculture Organisation (FAO) Fishery Information, Data and Statistics Service ,1983. *ASFIS-2: ASFIS subject categories and scope descriptions*. Rome: FAO, 59p.

Kuo, D., 1999. *Integrating ANZLIC and GCMD Catalogues Using Simple Fuzzy Keyword Mappings.* Paper presented at EOGEO99, February 9th - 11th, 1999 Washington Area, USA; available on the web at http://webtech.ceos.org/eogeo99/Papers/Kuo/Catalogue.html (as at November 1999).

Rees, A.J.J. & Ryba, M.M., 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* (ed. Tanner, E.L.), pp. 102-111. Glebe, Australia: Earth Ocean & Space; available on the web at http://www.marine.csiro.au/datacentre/ext_docs/marlinpaper.htm (as at November 1999).

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