UC San Diego
Scripps Institution of Oceanography Technical Report

## Title
CalCOFI Data Management, White Paper

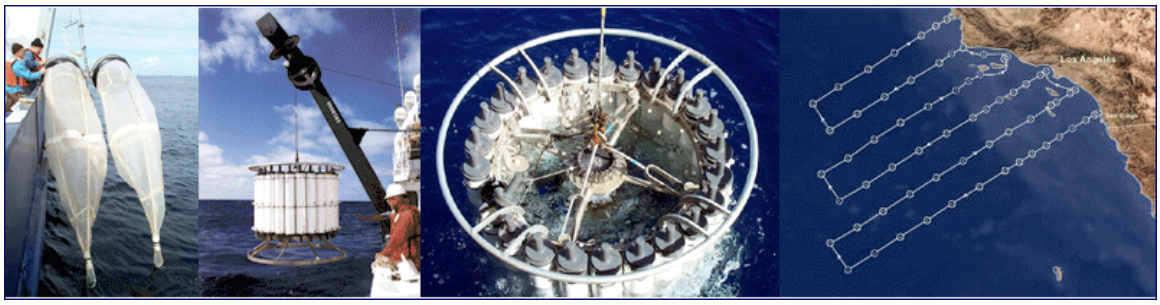
**Permalink**
https://escholarship.org/uc/item/9gr058dq

**Authors**
Stocks, Karen I
Baker, Karen S

**Publication Date**
2005-09-01

eScholarship.org Powered by the California Digital Library
University of California

# CalCOFI Data Management
## White Paper



Karen I. Stocks[1] and Karen S. Baker[2]

[1]San Diego Supercomputer Center, UCSD
[2] Scripps Institution of Oceanography, UCSD

Scripps Institution of Oceanography Technical Report
September 2005

# Table of Contents


Purpose 1
Acknowledgements 1
1. Introduction 3
2. Current Management of CalCOFI Data 6
    2.1 Hydrographic "bottle" dataset 6
    2.2 Ichthyoplankton 8
    2.3 Zooplankton 9
3. Evaluation of CalCOFI Data Management, Recommendations, and Recent Activities 10
    3.1 Database software and organization 10
    3.2 Backup and Archive 10
    3.3 Metadata and Data Documentation 10
    3.4 Data acquisition and processing 13
4. Developing an Integrated Access System 15
    4.1 Organizational Considerations 15
        Coordination of Development and Collaborative Design 16
        Community Communication 16
    4.2 Technical Considerations 18
        Common Database Elements for Integration 18
        Integrated Ocean Observing System (IOOS) Compatibility 19
        Centralized Versus Distributed Systems 20
        Data Contribution, Use, and Acknowledgement Policies 20
        Standards 21
        Assessing Available Tools 21
        Hardware and Software 23
    4.3 A Prototype Online Access System 24
5. Expanding Beyond Focus Data 25
    5.1 Data Types 25
        Discrete Data Acquired on CalCOFI Cruises on CalCOFI Stations 25
        Continuous Data Acquired on CalCOFI Cruises on CalCOFI Stations 26
        Continuous Data Acquired on CalCOFI Cruises between CalCOFI Stations 26
        Raster/Grid Data in the CalCOFI Region 27
    5.2 Data Inventory 28
6. Summary of Recent Activities 29
7. Scaling up: Considerations for PaCOOS 31
8. Conclusions 32
Acknowledgements 32

# Table of Contents, continued
# Tables, Figures and Appendices


Table 1: Summary of Recommendations 2
Figure 1: CalCOFI Partnerships 4
Figure 2: CalCOFI Data Diagram 7

Appendix 1: Personnel list 33
Appendix 2: Meeting Summaries 35
Appendix 3: Data Inventory 38
Appendix 4: Data Inventory Template 44
Appendix 5: 2005/2006 Proposal 45
Appendix 6: Glossary of Acronyms 47

## Purpose

The California Cooperative Ocean Fisheries Investigations (CalCOFI) program is one of the longest-running, multidisciplinary ocean monitoring and observing programs in existence. For many years, the emphasis of data management within CalCOFI was to quality control and curate the individual datasets collected on CalCOFI cruises, and make them available to researchers and fisheries managers through printed reports and requests for data to the data curators. Today, a new goal is emerging of having CalCOFI datasets available online and, eventually, interoperable with other CalCOFI-related datasets and the larger, developing federation of the Ocean Observing System data.

In this document we review the current state of data management within three of the primary CalCOFI datasets (hydrographic, ichthyoplankton, and zooplankton) and then make recommendations for moving towards the integrated online system that is envisioned, including expanding to other data types. Concrete recommendations for moving forward are summarized in Table 1 and explained in more detail throughout the text.

## Acknowledgements

We give thanks to Rich Charter, Ralf Goericke, John Hunter, Mati Kahru, Mark Ohman, Jesse Powell, Elizabeth Venrick, Jerry Wanetick, and Jim Wilkinson for their contributions and to the CalCOFI community for supporting this work. Scripps Institution of Oceanography CalCOFI provided the support for this planning document.

## **Table 1: Summary of Recommendations** – detail provided throughout the text

### Organization

- Establish decision-making structures, committee members, and responsibilities. In particular, we recommend that the CalCOFI Committee create an Information Management subcommittee composed of the data managers of each main CalCOFI dataset to minimize divergent development. The decision-making structure should then develop:
  - short-term development priorities (i.e. central vs. distributed system, key functionality, data priorities, who will host the online access system, etc.),
  - long-term strategies (open source vs. proprietary software, processes for expansion, personnel training needs, etc.).
- Develop community communication mechanisms through meetings, prototype assessments, etc. within the informatics community and between informatics and oceanographic researchers. User outreach mechanisms will be particularly important.
- Continue designing and maintaining the personnel directory, and make the appropriate content available online.
- Formalize a data submission policy for main CalCOFI data types, a data use policy, and a data acknowledgement statement.
- Continue the ongoing task of the data inventory.

### Local Elements and Work Flow

- Continue developing the integrated online access system being created by the SWFSC.
  - Conduct an evaluation, including user input, of the online access prototype as early in development as possible.
- Review and update the data input and processing procedures.
- Devote modest additional funds for systems administration at SIO.
- Document data system components and data management best practices.

### Cross Community Coordination

- Participate in IOOS data management working groups to expand IOOS technologies, such as OPeNDAP, to better handle CalCOFI data types.
- Archive data with the National Ocean Data Center until IOOS archive centers develop.
- Review existing external tools, standards, protocols that can be useful for CalCOFI, including transport specifications (OPeNDAP, DiGIR, OGC), metadata standards (FGDC, EML, GM3/GETADE), species observation standards (OBIS, BioCASE), taxonomy standards (NCBI), data formats (NetCDF, HDF, XML, RDB, ascii), controlled vocabularies (MRIB, MMI), catalogues (Metacat, Mcat, Thredds), analysis/visualization tools (LAS, ACON, ODV, IDV), and the International Council for the Exploration of the Sea (ICES) Oceanographic standards and Data Management Software.
- Launch local design teams on OPeNDAP, metadata standards, and other interoperability strategies such as dictionaries, thesauri, and ontologies. Select pilot and prototype activities strategically.

### Data System Design and Development

- Devote resources to redesigning the SIO collections zooplankton database, ideally before new survey lines begin operation.
- Create FGDC metadata for the primary CalCOFI datasets, as is required for all federally funded data, and make this available in conjunction with all data being served online.
- Expand metadata to hold information on methods of collection, taxonomic resolution, precision information, and quality control procedures.
- Register standard metadata with the Global Change Master Directory and/or the NOAA Coastal Data Development Center.
- Focus in the short term on readying datasets locally for flexible integration and data delivery in a variety of formats.

# 1. Introduction

The purpose of this document is to overview the current state of data management within the California Cooperative Ocean Fisheries Investigations (CalCOFI) program and outline steps and recommendations for building towards an integrated, online information system for CalCOFI. It will also include consideration of how this effort could be scaled up to contribute to the emerging Pacific Coast Ocean Observing System (PaCOOS).

CalCOFI is a long-term sampling program aimed at understanding the marine environment off southern California and improving living resource management in the region. It represents a consortium of the NOAA/NMFS Fisheries Resource Division of the Southwest Fisheries Science Center (SWFSC), the California Department of Fish and Game, and the Integrative Oceanographic Division (IOD) of Scripps Institution of Oceanography (SIO). A number of journal articles provide an introduction to the CalCOFI program (e.g. Ohman and Venrick, 2003; Bograd et al, 2003; Hewitt, 1988; CalCOFI, 1989, 1999) and demonstrate integrative uses of the data to address particular scientific questions (e.g. Deep Sea Research, 2003; McGowan et al, 1998; Roemmich and McGowan, 1995).

The core of CalCOFI is the regular sampling: 4 times per year research vessels visit a standardized grid of 66 stations to measure physical, chemical, biological, and meteorological properties of the California Current ecosystem including censuses of organisms from phytoplankton to birds. Though the spatial and temporal scope has varied through time, this program has persisted since its inception in 1949.

A program of this scope creates an enormous volume of heterogeneous data that can be valuable to a wide variety of oceanographic research and resource management applications. A recent interest in making the full range of data readily available online has led to an effort to understand holistically the current state of data management of CalCOFI data and lay a path towards building an integrated online access system.

A second driver for recent CalCOFI data management interest comes from the emerging array of efforts with which to partner. Figure 1 shows some of the developing local, national, and international ocean data projects relevant to CalCOFI. In the United States, one of the larger efforts, the Integrated Ocean Observing System (IOOS),[1] is in the early stages of development. The goal of IOOS is to create a unified and cost-effective approach for providing data needed to meet the suite of ocean-related challenges facing society: improving the safety and efficiency of marine operations, mitigating the effects of natural hazards, predicting the effects of climate change, protecting and restoring natural ecosystems, sustainably using marine resources, etc.

[1] http://ocean.us

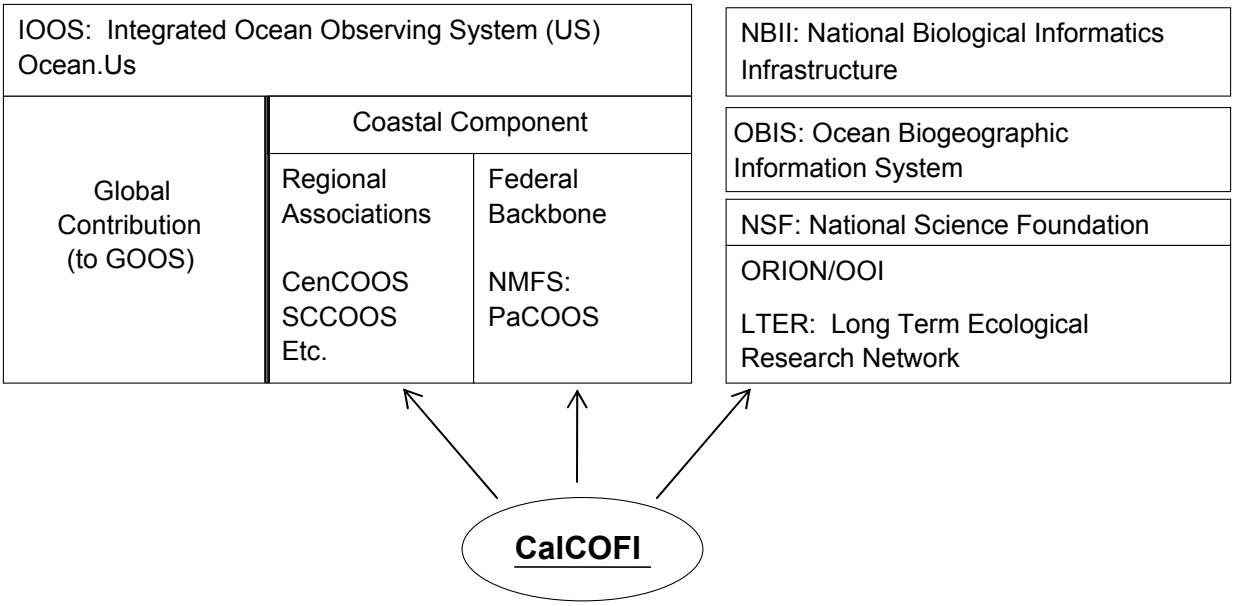



*Figure 1: Relationships of CalCOFI to other ocean data initiatives. CalCOFI has a direct relationship with the Pacific Coast Ocean Observing System (PaCOOS), the Southern California Coastal Ocean Observing System (SCCOOS), and the California Coastal Ecosystem site of the Long Term Ecological Research (LTER) program. The OBIS and NBII boxes are representative of the multiple national and international biological data initiatives that have overlapping goals or could share technologies with CalCOFI.*

Part of the work of IOOS is to develop an integrated data management system across the many IOOS components.

The National Marine Fisheries Service (NMFS) will participate in IOOS; the west-coast contribution of NMFS to IOOS will be the Pacific Coast Ocean Observing System (PaCOOS)[2]. PaCOOS will focus on improving and maintaining the observing elements needed for guiding the sustained use of marine resources and for protecting marine species and their ecosystems. It is currently in an early stage of development. An evaluation and forward-planning exercise for CalCOFI has substantial relevance to planning PaCOOS data management. With its 55+ year history, *CalCOFI is the longest-running interdisciplinary ocean observing system in existence*, developed long before the term "ocean observing system" was coined. CalCOFI is a microcosm of the variety of data types and user needs that PaCOOS will be considering; if information infrastructure for CalCOFI is developed to be scaleable and extensible, then CalCOFI can inform the PaCOOS data management development.

[2] www.pacoos.org

This report will outline:

- the current status of CalCOFI data management: what the existing systems and practices are, what improvements and additions have been made in the last year, and what the strengths and needs are (sections 2 and 3);
- planning towards an integrated online system, including both organizational and technical considerations, and a discussion of coordination with IOOS (section 4);
- considerations for expanding to new data types (section 5);
- summary of recent activities (sections 6); and
- scaling up to PaCOOS (section 7).

# 2. Current Management of CalCOFI Data

CalCOFI is a large and multi-faceted program; over time a large variety of data have been collected, both on CalCOFI cruises and otherwise, that are relevant to the CalCOFI goal of understanding the hydrography and ecology of the California Current System. CalCOFI data management systems and issues have been infrequently summarized (Charter, 1988; CODATA, 1995). Today, CalCOFI data may be found online at two websites, one maintained by Scripps Institution of Oceanography (SIO)[3] and the other by the Southwest Fisheries Science Center (SWFSC)[4].

In Figure 2, some of the many datasets relevant to CalCOFI goals are listed and categorized. Below, we focus in detail on three important CalCOFI datasets, which have been the focus to date of initial integration activities and have served as guides for future development – we will refer to these three as "focus" datasets. For each of these datasets, we discuss the current status of data management, data accessibility, recent activities to develop integrated access, and challenges for future system development. In section 5 we expand this discussion to consider the challenges presented by a wider variety of datasets. Throughout the paper we also refer to "routine" data – these are data that are now regularly collected on all CalCOFI cruises.

### 2.1 Hydrographic "bottle" dataset

Hydrographic bottle data are measurements taken from water collected at discrete depths by a bottle rosette lowered at each station. Samples from these bottles are analyzed for temperature, salinity, oxygen concentration, nutrients, chlorophyll, pigments, phytoplankton and species counts. A CTD mounted with the rosette gives continuous profiles of conductivity (salinity), temperature, and depth. Note that the bottle dataset (discrete CTD profile) and the continuous CTD profile data are both hydrographic data, but are stored separately and considered as separate datasets.

SIO curates the hydrographic datasets. SIO data files were stored originally as text files on a university VAX mainframe and manipulated using the Fortran programming language. Different parameters were held in individual datasets by date and time; later, a shift in organizational perspective resulted in pooling data into a "cruise collection" of related datasets. An early NODC-related format (SD2) evolved in 1984 into a local standard known as IEH (Is Everybody Happy), a data format that remains in use today. In 1993 the data were ported to a PC DOS system and in 1998 migrated to Windows with some Visual Basic programs added to the software toolkit.

---

[3] http://calcofi.org

[4] http://swfsc.nmfs.noaa.gov/frd/CalCOFI/CC1.htm

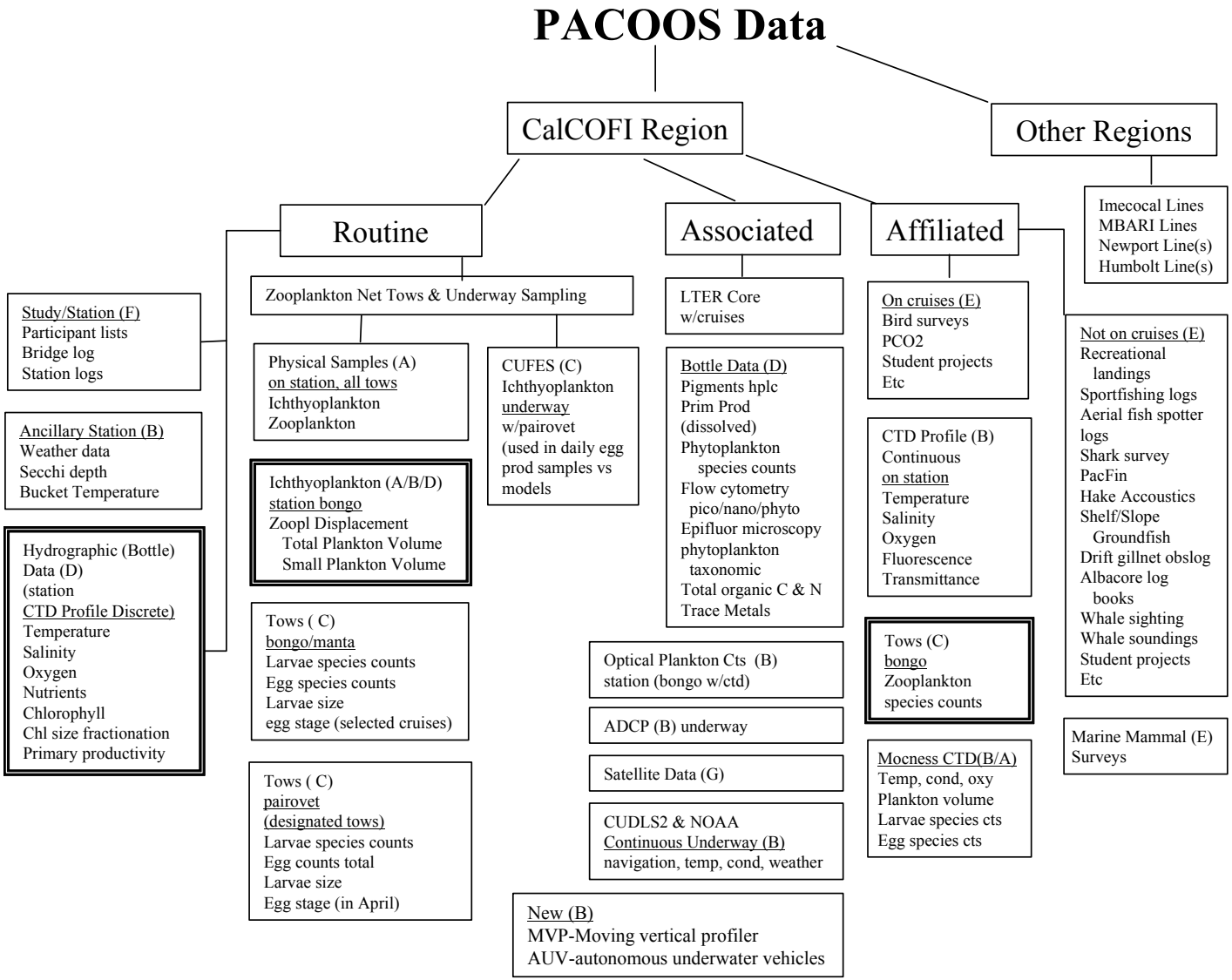



*Figure 2: Datasets and measurements relevant to CalCOFI - see Appendix 3 for details of particular datasets. The double bordered boxes indicate the three focus datasets described in this paper. The Measurement Types Key is as follows: A) Curated Physical Collections; B) Sensor in-Field; C) Microscope Analysis; D) Other Sample Analysis; E) Human Observer Counts; F) Collection Descriptions; G) Raster Sensor.*

The hydrographic datasets have since been migrated to Microsoft Access and are now managed by Jim Wilkinson at SIO. The existing database includes the cruise description and location information, hydrographic bottle sample information, and quality control data, as well as the derived sample information included in the cruise reports. The data system uses three unique identifier strings: the cruise station id# (key to linking activities on station) which builds upon cruise and station, the cast id# which adds the time information of Julian day and universal time (UTC), and the depth id# which adds depth information for the discrete bottle samples.

These data are available online[5] as a bulk download for 1949-2004 in Microsoft Access or as IEH-formatted text files. A CalCOFI IEH Formatting Tool (CIFT) for query and download is being retired as the data management group at SIO moves toward dynamic query and plotting. Currently the application SURFER is used locally for producing plots. Data from 1990 to the present can also be viewed in the Data Reports (available on the same website), either as PDF files or as interactive html pages allowing users to click on cruises and stations of interest. This site also serves data on the volume of each zooplankton sample collected in standardized bongo net tows.

### 2.2 Ichthyoplankton

Plankton nets are towed at each CalCOFI station. The collected material divided into ichthyoplankton (fish eggs and larvae) and non-ichthyoplankton zooplankton. Though collected as a single sample originally, we consider the two datasets separately because each portion of the sample is sent to a different group for taxonomic identification and both the physical specimens and the data are stored independently.

The ichthyoplankton are sorted to taxonomic group (generally species) by staff at the Southwest Fisheries Science Center. The physical samples are preserved as a specimen collection at the SWFSC. The data resulting from the taxonomic identification are entered into a database overseen by Rich Charter at SWFSC.

These data were recently migrated from Oracle to a Microsoft SQLserver system using Microsoft Access as a front-end. The database includes cruise description and location information (keying off desired station and local time) including an activity log, hydrographic and chemical measures, net information (net types, tow types, net angle, flow meter readings, and volume filtered), and species information (species, size and stage for eggs; species and size for larvae). These data are available locally at SWFSC; the complete datasets are not publicly available online, but derived maps of egg distribution are included in downloadable electronic reports. Plotting applications used locally include GMTPlot and ArcGIS.

[5] http:// www.calcofi.org

## 2.3 Zooplankton

The non-ichthyoplankton portion of the net plankton samples are curated by Mark Ohman at Scripps Institution of Oceanography. The physical samples are archived in the SIO Pelagic Invertebrate Collection, which supports the research of a variety of visiting and remote zooplankton researchers.

The data from samples that are identified to taxonomic group are currently held in a Microsoft Access relational database. However, the database is currently part-way through a re-design process (for which funds no longer exist to complete) and is not fully functional. Information about the samples, such as where and when they were collected, reside in a Foxpro database and are available online through the SIO Pelagic Invertebrates website[6]; information on the taxonomic content (i.e. "sorted" samples) are not available online. One additional complexity of these data is that, in some cases, subsamples have been taken from multiple samples and pooled before identification to taxonomic group. Therefore, these sample data are related to multiple stations rather than to a single station/cruise number as is the case for the rest of the routine data.

[6] http://collections.ucsd.edu/pi/index.cfm

# 3. Evaluation of CalCOFI Data Management, Recommendations, and Recent Activities

### 3.1 Database software and organization

All three of the focus datasets are currently in ODBC-compliant relational database structures appropriate for supporting internet-based queries. For easier security or to allow more advanced queries in the future, flowing data through a more powerful, secure platform before serving may be desirable, but there is currently no need to change the native data formats. For the hydrographic bottle and ichthyoplankton data, the databases hold the information the curators (who are the best judges, as they currently fill data requests from users) think are important, and only minor modification and extension through time are needed. For the zooplankton data, the curator indicated that the database design is not sufficient and requires a re-design after an assessment of needs. This task would ideally occur before new monitoring lines (such as the expected Humbolt line) begin as it could serve as the basis for designing a core template for new lines as they are added.

Additions and adjustments are being made to the datasets. For example, the hydrographic dataset at SIO has been expanded recently to hold data on the total zooplankton biomass capture per sample.

### 3.2 Backup and Archive

The hydrographic and ichthyoplankton datasets are backed up with nightly incremental back-ups and monthly full back-ups. Several copies of full back-ups are stored off-site. Data was originally sent to the National Oceanographic Data Center for archiving; currently NODC harvests hydrographic data from the SIO CalCOFI web site and imports it into their database. All of the main CalCOFI datasets should be regularly contributed to a national archive center such as the National Ocean Data Center.

### 3.3 Metadata and Data Documentation

Metadata is information that describes datasets, assisting the user to understand the content of the dataset. The focus datasets all have a good base level of metadata, for example describing the units in which parameters are given. Two areas, however, are not addressed currently. The first is that FGDC-compliant metadata has not been created. This is a recommendation of the Integrated and Sustained Ocean Observing System initiative. It is also a requirement for all federal data, so CalCOFI has a formal obligation to provide FGDC metadata. Metadata files for

each dataset should be created and registered with the Global Change Master Directory[7] and/or the NOAA Coastal Data Development Center[8], as recommended by the Integrated Ocean Observing System.

FGDC metadata alone, however, will not be sufficient. The second area that needs to be addressed relates to metadata extensions. There are four aspects of the CalCOFI data that require additional documentation:

### Methods and Quality information

Over time, the methods with which data were collected have changed, for example from silk to nylon plankton nets. This information is needed by users to determine when samples are comparable across time and needs to be delivered with the data rather than located separately. Information on the methods, and also the quality control procedures, are required for the user to assess whether the data are appropriate for a given analysis or task. Some information on the current equipment and methods used are online, but how the collection methods and quality control processes have changed through time is not available except by asking the data curators. The routine databases should be expanded to include a method description table to which each sample is referenced. When the online access system is built, this can then be used to include method data, and give warnings when users try to download or plot data across time periods that are not comparable.

### Taxonomic resolution

The ability to identify ichthyoplankton and zooplankton has improved over the years so that some species that were originally identified only to genus level can now be identified to species level. If a user requests data on such a species, it can be misleading because it may look like the species was never present before a certain date, when really it was present but not identified. Information on the "resolvable" species list through time should be stored in each database, and users downloading data on a given taxa warned when they compare or download data across different species lists. How to best store and represent this information is a topic of active development in groups such as the Canadian Department of Fisheries and Oceans and the Ocean Biogeographic Information System. CalCOFI should consider whether solutions developed externally will meet its needs before inventing a new protocol.

### Precision estimates

An estimate of precision for all values given should be part of the metadata files. This includes parameters such as temperature, chlorophyll, etc., as well as location information.

---

[7] http://gcmd.gsfc.nasa.gov/

[8] http://www.ncddc.noaa.gov/

Precision estimates for currently-collected routine analyses were made on CalCOFI cruise 9003 (SIO Ref 91-4) but need to be incorporated into the metadata. Legacy precisions should be investigated subject to time and support available. If methods for measuring parameters have changed over time, then the metadata needs to reflect this changing precision. Location information is of particular importance. Methods of estimating sample location have varied from the current GPS to dead reckoning. It is important that users of the older data understand how far off the sample location estimate might be. This task will require some time to document historic methods and attach precision estimates to the data.

**Controlled vocabularies and dictionaries.**

To query a data system about data held in multiple collections depends upon nomenclature and categories held in common across the datasets. Controlled vocabularies and set dictionaries of terms, including measurement units and attributes, can greatly facilitate interoperability. For example, if two datasets used the same gear to collect a sample, but one called it "bongo net" and another called it "plankton net: bongo" the system would not know that they are equivalent. Agreeing on common terms makes integration across common elements automatable. Another example is unit dictionaries. If one dataset holds a measurement in mg per ml and another in grams per liter, a unit dictionary that defines parent units and the mathematical relationship between the units will allow data from across the two datasets to be delivered in the same units. As CalCOFI scales to take in data from more sources, the need for controlled vocabularies will grow.

As CalCOFI designs extended metadata, an effort should be made to follow existing standards, where applicable, rather than inventing new standards. Members of the ecological community have worked for a number of years on an Ecological Metadata language. EML[9] extends FGDC to provide a standard for documenting the contents of ecological datasets. Like FGDC, EML defines a way to describe a dataset as a whole in a standard way. When EML metadata are available for searching, it can help users unfamiliar with CalCOFI find datasets of interest and understand their content. It also helps a user who wants to bulk download a whole dataset as a block and work with it. There are tools such as Morpho[10] and MERMAid[11] under development to facilitate EML implementation. The National Center for Ecological Analysis and Synthesis is supporting Metacat, an EML-based metadata catalog (i.e. a place where data providers can describe their data and data users can go to find datasets of interest). The Storage Resource Broker at the San Diego Supercomputer Center provides a uniform interface for connecting to heterogeneous resources described in an EML catalog. By creating FGDC and/or

[9] http://knb.ecoinformatics.org/software/eml/
[10] http://knb.ecoinformatics.org/morphoportal.jsp
[11] http://www.ncddc.noaa.gov/Metadata/tools

EML metadata, CalCOFI data could be found and accessed through these systems, but the cost is that EML metadata can be complex and time consuming to create and maintain. Other metadata standards are evolving for specific types of data, such as satellite and ADCP data[12], and should be evaluated at the appropriate time.

Marine XML[13] is an international initiative being led by UNESCO's Intergovernmental Oceanographic Commission's (IOC) Committee on International Oceanographic Data and Information Exchange (IODE). While not yet mature, it may hold promise as a framework for CalCOFI data documentation and should be considered in the future. The Marine XML Data Exchange document developed by the Canadian Department of Fisheries and Oceans, which has similar data types and user needs as CalCOFI, offers one implementation model[14].

Despite the appearance of stability that published standards give, metadata is an extremely active area of research and development in fields such as information management, library science, and computer science, and one that is crucial to larger-scale interoperability. It is important to recognize that metadata development is an important process to undertake and support; it is much more than a matter of finding a single 'right' solution. Keeping informed and contributing to these efforts, choosing standards, implementing them, and "crosswalking" local standards with external ones, are tasks that will take devoted effort, particularly as the vision enlarges from an integrated CalCOFI data system to an integrated national ocean observing system.

### 3.4 Data acquisition and processing

One aspect of data management is the process by which data enters a system and is quality controlled, not just how it is held once it is in the system. Perspectives on data collection and processing have evolved during this ongoing era of dramatic changes in hardware availability, operating systems, facilities, programming styles, architectures, and federations. As expectations for information availability grow, it is valuable to step back and consider the work flow process and data system as a whole, from the field measurement collection moment to the integration of data with an as yet unspecified other dataset.

At present, most CalCOFI data are acquired digitally (the exceptions are the handwritten weather and station data logs, visual counts of eggs, and visual assessment of zooplankton displacement volume), but the process of quality control is time consuming and personnel intensive. Further, training new staff requires substantial time. The current data processing has served to produce quality data for more than 50 years, and quality control will likely always be

---

[12] http://www.nodc.noaa.gov/General/adcp.html

[13] http://marinexml.net/

[14] http://ioc3.unesco.org/marinexml/files.php?action=viewfile&fid=28&fcatid=1

an area that needs substantial attention and time. However, the system uses a mix of legacy routines and systems and an overall review of the process has not been conducted for many years. Streamlining is needed to create a more efficient system.

This is particularly important for facilitating the flow of data from at-sea data systems to on-land data systems at the end of a cruise. While data streams and samples are collected in the field, data checking and analysis often occur or continue on land in the laboratory. Given recent advances in technology, it would be timely for the CalCOFI community to consider its software and hardware at sea and on land jointly as part of a single data flow process – having system components in common across at-sea and on-land data management can facilitate data movement. One promising technology is web-based architectures, which are common today on land but not at sea. Such an architecture has both benefits (portability and flexibility) and costs (maintenance of web services). One of the strengths is that if the back end of this system must be changed, it does not in theory interfere with general use. With the majority of processing occurring at the central system, the client side need only support a browser. Such a system opens up the likelihood that quality control applications can be created using similar software elements. Training time is reduced when similar tools are used on land and at sea.

Changes to the data input and processing methods will be a major development task, requiring new resources, and should be undertaken carefully. Changes to interconnected protocols can have ramifications that are not obvious without a holistic review of the system. Using a collaborative design approach will allow the existing knowledge of the shipboard and data management teams to be incorporated into the redesign, facilitate the development of joint standards across datasets, and increase the awareness of how decisions in processing will affect later interoperability.

In the last year, prompted in part by discussions of this white paper and in part by the immediate need to redesign the underway data collection software, development has begun of new approaches to link at-sea and land-based data management activities for the hydrographic data. A Linux server has been deployed to prototype real-time databasing of underway data, which allows data to be accessed via web forms. This prototype permits both testing and training with new tools. Using a web-based architecture, it is built on open-source components (Apache, mySQL, and php/perl).

# 4. Developing an Integrated Access System

The above data management recommendations and activities relate to good management of the CalCOFI datasets as individual databases. Moving towards an integrated data system poses additional considerations. By integrated, in this case, we mean that the user has the ability to search and access data from across datasets. A website that allows a user to search for zooplankton data based on a station or date of interest and access that data, but then requires a second search in a second interface to get the hydrographic data is not considered integrated. One that allows the user to get both sets of data based on one search would be integrated.

The future data management goal for CalCOFI data is a system in which data 1) efficiently enter the system from shipboard sampling and shoreside analyses; 2) is fully documented for a variety of use types; 3) is archived for long-term preservation (including needed metadata); 4) can be fully accessed, queried, and downloaded; 5) can be integrated with other data types within the CalCOFI system and with external data sources (such as other Ocean Observing System initiatives); 6) can be analyzed and visualized with simple online tools; 7) will feed directly into models; and 8) are presented as synthesized data products, such as time series charts or interpolated spatial coverages, as well as raw data. In addition, the system must meet IOOS data management requirements and should be scaleable to the PaCOOS level.

This is the *long-term* goal for CalCOFI. Creating such a system will take many years and will in fact be a continuing process as new data types, new models, new software tools, changing technology, new participants, etc. are incorporated. Developing concrete, shorter-term steps to move efficiently towards the long-term vision will involve both organizational considerations (developing a decision making framework, allocating tasks, and adopting a design strategy) and technical considerations (evaluating IOOS requirements, reviewing other applicable standards, and evaluating existing data systems).

### 4.1 Organizational Considerations

We discuss organizational considerations in advance of technical considerations because, in our combined experience, information system development projects fail or succeed more often for organizational reasons than technical reasons. Developing an integrated information system is a new task, not merely an extension of current data management, and requires appropriate organizational alignments, decision-making structures, and distributions of tasks. Changes in infrastructure are needed for scaling from a small environment where individuals coordinate with respect to data management to a larger environment where communities coordinate. Today there is an emerging approach that builds from the discussion of technical environments to the consideration of sociotechnical environments, and incorporates conceptual, organizational, and social issues. Within a small group doing field work, there is coordination that is logistical, administrative and informal but when coordinating across datasets, groups, and communities, the

coordination moves to new levels involving design and collaboration at an information architecture level. This includes conceptual visioning, common vocabulary building and dataset organization. Such issues are the subject of growing research within the areas of informatics, social informatics, digital libraries, and knowledge management.

### *Coordination of Development and Collaborative Design*

Until recently, the CalCOFI data were managed as individual datasets, with more of an emphasis on data curation and access than on integration. Data expectations are now changing. Data mandates making integrated access a goal are emerging, fostering the need for increased communication and coordination between the managers of individual datasets. The past use of different station identifiers and time stamps described in section 4.2 illustrates how seemingly simple differences across datasets can increase the work needed to integrate data. We are at a time when both the SIO and the SWFSC databases are undergoing change and revision. **Without communication and collaboration in the redesign process, we risk losing an opportunity to plan now to facilitate future integration.** We recommend that the CalCOFI Committee create an Information Management subcommittee composed of data managers from each of the main CalCOFI datasets, and that this subcommittee meet regularly to discuss data integration and information systems design.

Maintaining the flexibility needed for local work while addressing the need to standardize in order to be interoperable across data from multiple communities will present a series of choices. Such issues are best addressed in partnership so that the design will be informed by the diverse requirements to be served. Note that contemporary information systems research suggests that a collaborative design process improves use, sustainability, and adaptability of such systems. Collaborative design is a user-centered iterative design approach explicitly incorporating strategies for enrollment, assessments for understanding, and heterogeneous perspectives. Further, it is through collaborative design that the process of community engagement is initiated. Although there is no perfect information system that will work for all groups, an informatics environment can be developed in which information managers participate in ongoing design and development that result in an evolving shared information architecture and a set of used and interoperable information systems. It is through collaborative design that joint understanding is developed and redesign can occur incrementally over time, adapting to ongoing scientific and technological change.

### *Community Communication*

In addition to fostering coordinated design among data managers, wider communication with a larger community that provides, interacts with, and uses CalCOFI data is desirable. We note that establishing mechanisms and devoting resources to support communication is an under-recognized and under-supported aspect of scaling up. Several activities, described below, have

recently been initiated to improve communication among the different groups involved in CalCOFI data management, and to facilitate input from interested parties outside the CalCOFI community. We recommend that these activities be continued and in some cases formalized with some form of regular reporting. For example, developing documents on data management best practices can help coordinate planning and serve as a ready introduction for new participants. And a newsletter or CalCOFI report could provide regular data news for interested participants.

### *Personnel List*

A community is made up of people. Although some idea of CalCOFI participation can be deduced from reports, ship logs, and conference attendance, and some old-time participants may be aware of the diverse participants, for newcomers and non-CalCOFI regulars the lack of a personnel list is a barrier to communication and to understanding the program. A preliminary CalCOFI data management-related personnel list was initiated at a recent community data meeting and has been used as a mechanism for prompting discussion of organizational structures and individual roles and responsibilities. The list is included in Appendix 1. We recommend that a personnel directory continue to be designed and maintained so that an appropriate personnel list can be made available on the CalCOFI website(s).

### *Community Interviews and Meetings*

In the past year, several communications activities were carried out. Communication with CalCOFI participants took the form of a series of individual interviews, small meetings, and larger workshops in addition to a conference presentation. Follow-up meetings continue as we write this report in order to further document aspects of CalCOFI data management, laying the groundwork for future dialogue and information system development. We used iterative dialogues to verify what we were understanding, to further develop this understanding, and to provide feed-back on what we learned to community participants with whom we were interacting.

Short summaries of some of the meetings are gathered together into Appendix 2. At one CalCOFI Data Management meeting, in October of 2004, sixteen participants from SIO and SWFSC discussed current efforts and generic data system elements. A coauthored report was written and circulated in order to continue the participatory nature of the meeting. At a second meeting, a dozen participants from multiple SIO departments and from WHOI attended an Ocean Informatics Exchange Workshop (November 5, 2004) in order to consider common information system issues. In addition, a CalCOFI data management presentation titled ‘CalCOFI Data Management: Today and Beyond’ was made at the Annual CalCOFI Conference (November 17, 2004) and was followed by a Data Management Work Session

bringing together CalCOFI technicians, specialists, scientists, and information managers from throughout the region.

## 4.2 Technical Considerations

### *Common Database Elements for Integration*

One of the challenges for integrating CalCOFI data is illustrated by the three focus databases, which have been developed individually and have lacked coordinated common elements. Even though hydrographic, ichthyoplankton and zooplankton samples are each taken at the same stations on each cruise, which would seem to provide a way to integrate the data, each dataset developed a unique way to name stations, record the lat/lon location, the time (local versus GMT time), the date (Julian day versus month and day), etc. As a consequence, the seemingly simple task of matching records can become a tedious exercise that must be executed differently for each analysis.

One example is the variation in how stations are identified. A series of activities coordinated at a location and time is traditionally called a 'station' and a standard for numbering CalCOFI grid lines and stations has developed. The 'station' construct is often referred to as a single point in time and space but actually involves both a time interval as well as a small area as a ship conducts tows or stops at a location. Consequently, there are multiple meanings to 'station': there is a 'desired' or 'planned' standard station, there is an 'actual' or 'held' nearest standard station location (which differs from the planned one due to changed plans or drift), and there is the start and end latitude and longitude of a sample because the ship can drift during sampling. Because the different datasets vary in their conventions as to whether they hold the intended latitude and longitude or the actual one, and of how stations are named, connecting different samples taken on a station is a human-intensive process. The experience of Steve Diggs and Christian Reiss at the SWFSC exemplifies the importance of common identifiers across databases. In their work to create a prototype online data system for CalCOFI data they have had to devote substantial time to laboriously creating a "Rosetta Stone" to connect samples and stations across the different datasets. When common identifiers are adopted, this process can be automated.

Several steps have recently been taken to facilitate future data integration. First, prompted by discussions of semantic differences, unique identifiers, and common elements, Karen Baker and Jim Wilkinson are currently developing an event number convention as a unifying element. Organizing an entire cruise around a common log using an event number and a defined activity list provides organizational elements that bring participant data together before measurements even begin. In conjunction with this, a new cruise station log notation was initiated in the 2004 winter cruise. In addition, a series of strategic design team meetings were held with participants from multiple communities across SIO. These provided forums to consider the language used to

describe data types (e.g. cruise, mooring, shore station) as well as the organizational structure that manifests in differing approaches to grouping datasets. Finally, during the 2005 Spring cruise, an electronic tablet PC tied into the ship's network was prototyped for automating log keeping.

### *Integrated Ocean Observing System (IOOS) Compatibility*

IOOS is developing recommendations to ensure interoperability within the growing IOOS network of data providers. A lengthy Data Management and Communications (DMAC) Implementation Plan[15] that outlines the many data management issues has been produced, as well as a short summary of guidance to IOOS data providers[16]. The important components of this guidance are summarized below.

**At present, the DMAC plan provides very little guidance for focus CalCOFI data types, though DMAC has plans to develop in these areas in the future.** We recommend that CalCOFI participate actively in the DMAC working groups currently being formed in order to ensure that future recommendations meet CalCOFI data needs.

#### *Metadata*

To be IOOS compliant, CalCOFI must create FGDC metadata for its datasets and register them with the NASA Global Change Master Directory and/or the NOAA Coastal Data Development Center. This does not restrict CalCOFI from following additional standards as described in section 3.3.

#### *Data Transport*

IOOS recommends OPeNDAP for the transport of raster/gridded data. The recommendation is not pertinent to the focus CalCOFI datasets since they are relational rather than raster. For relational data, IOOS suggests that data providers participate in pilot projects testing the OPeNDAP relational database server or enterprise GIS protocols, but recognizes that these technologies are still at the pilot implementation stage and thus not yet "DMAC-recommended."

#### *Online Browse*

DMAC recommends that all providers install a Live Access Server for online data browsing. LAS is a technology that is easy to implement with data served through OPeNDAP but more difficult to implement without an OPeNDAP layer. For serving data, a LAS/OPeNDAP application packaged with a THREDDS[17] catalog system is planned for 2005 release, though it is not clear if relational data will be robustly supported in this version. In the meantime,

---

[15] http://dmac.ocean.us/dacsc/imp_plan.jsp
[16] http://dmac.ocean.us/dacsc/guidance02.jsp
[17] http://my.unidata.ucar.edu/content/projects/THREDDS/index.html

the decision-making body for CalCOFI data management should weigh carefully the time required to develop a simple local interface tailored to CalCOFI data versus the importance of implementing a LAS package and contributing data to IOOS immediately.

#### *Archive*

DMAC recommends that all providers ensure that important data are archived. It is valuable to recognize the distinction between data repositories and archives. A repository is a centralized working storage for datasets held locally. An archive is a more rigorously-curated, long-term storage facility containing more finalized datasets. Long-term archiving that can ensure preservation across decades of technology changes is a service best provided by professional institutions. CalCOFI should plan to archive data with an IOOS archive center as they are developed, and in the meantime contribute data to the National Ocean Data Center for archiving (SIO hydrographic data is already contributed).

### *Centralized Versus Distributed Systems*

At present, a centralized system could serve as the basis for an initial integrated repository and an online access system for focus CalCOFI datasets. The three focus datasets are all held in La Jolla and could be duplicated to a central server with periodic updates as new cruise data become available. Data from the MBARI line could similarly be sent periodically and updated centrally. This would be the simplest system, the fastest to start up, and is similar to the strategy for the SWFSC prototype being developed by Steve Diggs and Christian Reiss. Such a solution has the advantage of focusing on and prompting needed local integration.

A centralized system does not scale in the long run. In a distributed system, data are stored on multiple, physically-separated servers that communicate with a central portal. The advantages of a distributed system are that data are generally maintained best by the group that creates or curates the data, and distributed systems minimize the creation of multiple versions of data, which can get out-of-date. So when CalCOFI wants to integrate data outside of the focus CalCOFI datasets held in La Jolla, such as when additional survey lines are created in PaCOOS or when groups such as the Pacific Fisheries Environmental Laboratory serve data relevant to the CalCOFI region of interest, a distributed system will need to be designed. A distributed system will take more time and work but is scalable.

### *Data Contribution, Sharing, and Acknowledgement Policies*

Making data public is a delicate and complex issue. Even for organizations that have a strong will to share data, issues such as quality assurance and error checking can create hesitations. We recommend developing over time a clear, written “data contribution policy” for data providers of each of the CalCOFI datatypes to be made public. This should include timelines for data submission to the repository, formats, responsibilities for quality control, and

contact information for questions. To facilitate data flow and to avoid confrontational situations, a policy discussion can serve as an opportunity for program managers, information managers, data providers, and data curators to share their perspectives on data sharing and how procedures can be developed to minimize the chance that data are misrepresented or misused.

A separate "data use policy" for data users can provide information on data use, licensing, and acknowledgement. A brief "data acknowledgement statement" might contain the repository name, sources of support, and other credits in order to provide data users with the specifics for how to acknowledge data use. These policies should be posted online.

### *Standards*

There are a variety of national and international groups developing data standards of various types – these include metadata standards such as ISO, FGDC, and EML, content standards like the Darwin Core, format standards such as netCDF, and transport protocols like OPeNDAP. While standards are valuable for creating compatibility between data systems, it is important to recognize that standards development is a rapidly-evolving field: existing national standards and collections of tools are not currently mature or fully compatible. Bridges between standards are in various stages of development. A valuable CalCOFI goal would be to become familiar with these standards efforts and remain flexible with respect to the differing and sometimes incompatible paths through the layers of data access (discovery, query, transport, publishing), analysis, and visualization that these standards and tool suites provide. A critical question to ask when considering standards is whether the standard is suited to the data and the desired functionality, and at what granularity datasets should be described. For example, a metadata standard designed to describe the contents of a dataset as a whole will not be helpful if the dataset has extremely heterogeneous contents, and a transport system aimed at moving whole datasets will not be appropriate if CalCOFI wants users to be able to find and download subsets of datasets (such as data from just one station on one date).

### *Assessing Available Tools*

CalCOFI is not the first group to build a system for serving marine data, and it should leverage where possible from existing efforts. Adopting or tailoring existing technologies generally saves development time and also facilitates future interoperability with external systems. **Several items of potential interest are listed below, but the intent is to demonstrate the variety of applicable technologies, not produce a complete list or identify those that should be adopted. Evaluating existing standards, protocols, and tools to determine if they suit CalCOFI needs is a task that will require a devoted effort and support and should be an early part of the CalCOFI system planning**. This section is aimed at developing an online system for public access to CalCOFI data – many more software tools will continue to be used in-house for acquiring, processing, reformatting, and viewing CalCOFI data.

*Species Distributions: The OBIS/GBIF Architecture*

The Ocean Biogeographic Information System[18] is an international federation of institutions providing marine species distribution records (i.e. collections or observations of a particular species at a particular location and time). They are also the primary marine provider to the Global Biodiversity Information Facility[19], which is serving terrestrial and marine species location data from almost 100 institutions in over 40 countries. OBIS and GBIF use an open source set of portal/provider software called DiGIR[20] that is based on a community standard schema. The OBIS schema[21] describes the meaning and format of ~70 fields, of which only a handful are mandatory. (The OBIS schema is a small extension to GBIF's Darwin Core schema, with a few additional fields defined, but OBIS clients can also serve as GBIF clients.) Providers are free to store their data in any database software and in any format they wish, and then "map" their data onto the schema during DiGIR installation, a task which generally takes less than a day. Once this is done, a central portal can set up a data access interface that allows users to query and access data across all the distributed resources as if it were a single, centralized database.

The OBIS system is only applicable to distributed systems. If CalCOFI decides to use a centralized architecture initially, then OBIS will not be relevant. If/when it decides to expand to a distributed system, with several servers providing data that are accessible from a central portal, OBIS offers a potential architecture. Another potential benefit of using the OBIS communication protocol is that contributing to OBIS can bring new users to the CalCOFI datasets.

Like OPeNDAP and Live Access Servers, the OBIS architecture's benefit is that it leverages substantial programming efforts that have already gone into the system, and provides a much faster way to deliver integrated data from across servers than if an individual system were developed from scratch. And, just as with OPeNDAP, it is limited in that it only works for a certain type of data, not the full spectrum of datatypes in which CalCOFI is interested. The OBIS schema is only applicable to georeferenced species data, such as the CalCOFI ichthyoplankton and zooplankton data, but not the hydrographic data. Further, fisheries management often requires more data on the quantitative aspects of sampling (densities, gear efficiencies, individual length data, etc.) than are currently provided by OBIS. However, this need has been identified and the OBIS Technical Committee is currently working on a draft extension to the schema to fit fisheries-related data. The

[18] http://www.iobis.org/
[19] http://www.gbif.org/
[20] http://digir.sourceforge.net/
[21] http://iobis.org/tech/

situation should be re-evaluated when CalCOFI is ready to move towards a distributed system.

*Visualization: ACON GIS Software, Ocean Data View, Integrated Data Viewer, etc.*

CalCOFI data users can benefit from visualization tools for viewing data. Many visualization options are emerging. The ACON data visualization software[22] was developed specifically for visualizing and analyzing fisheries survey data. It was created by the Canadian Department of Fisheries and Oceans, Maritimes region and is free. Ocean Data View (ODV)[23] is another package for the analysis and visualization of oceanographic and other geo-referenced profile data. ODV data and configuration files are platform-independent and can be exchanged between different systems. It was developed by the Alfred Wegener Institute for Polar and Marine Research. The Integrated Data Viewer (IDV[24]) has been developed for geosciences data and is supported by Unidata[25] in collaboration with digital library communities. ESRI offers proprietary tools for displaying geospatial data (e.g. ArcGIS) and creating tailored online mapping interfaces (ArcIMS). These are just a sampling of many available options and, as the landscape is constantly changing, the options and capabilities should be fully evaluated when CalCOFI begins developing online visualization options. One consideration in CalCOFI's choice of GIS system is that, while no recommendation from IOOS has been created, many of the IOOS Regional Associations are moving towards Open GIS Consortium specifications for Web Map Services, Web Feature Services, etc. as they are developed.

### *Hardware and Software*

At present, the focus CalCOFI datasets are all held in relational databases that provide a sufficient foundation for building an online data access system. Because of the easy interoperability that SQL and ODBC provide, there currently is no reason to enforce a uniform platform. The SWFSC has server capacity appropriate for serving the ichthyoplankton dataset. The Integrative Oceanography Division at SIO has appropriate servers for the hydrographic bottle data and zooplankton data. However, some additional modest funding for systems administration at SIO is in order. And, as the system expands beyond the focus CalCOFI datasets to data, such as satellite coverages, that require more memory, hardware and software, expansion will be required. CalCOFI will need to discuss and decide where data will be hosted and where the servers for the online access systems will be maintained.

---

[22] http://www.mar.dfo-mpo.gc.ca/science/acon/index.html
[23] http://www.awi-bremerhaven.de/GEO/ODV/
[24] http://my.unidata.ucar.edu/content/software/IDV/
[25] http://my.unidata.ucar.edu/

### 4.3 A Prototype Online Access System

Steve Diggs, overseen by Christian Reiss at the Southwest Fisheries Science Center, has been building a prototype for a centralized, online, integrated web data access for CalCOFI data. A poster describing this project was given at the 2004 CalCOFI Conference.

# 5. Expanding Beyond Focus Data

## 5.1 Data Types

A great deal of additional data is relevant to CalCOFI and should be considered in the context of a larger, integrated information system (Fig. 2). These datasets fall into several categories, each of which has its own characteristics, which are described in more detail below. But all share several common challenges for data management:

1. Each additional dataset needs to be quality controlled, and documented with appropriate metadata.
2. The system must be expanded to allow searching and retrieving of each data type – the way users will want to interact with data may vary with the type, and so more datasets will require expanded system functionality. For example, users may be happy to see a table of numbers displayed for zooplankton count data, but might want a mapping system for looking at underway cruise track data.
3. No data inventory has been done until recently: a first step for planning a system that encompasses more than the focus datasets is an inventory of the data that exist, the formats they are in, etc.
4. Creating integrated data access requires that integrating elements be present in the datasets, and that tools be developed to work with multiple data types.

An important concept is that just because a dataset is available, either in hand or through a distributed system like OPeNDAP, it does not mean that there is little effort needed to bring it into an integrated CalCOFI data system. With all these datasets, there are choices to make between raw data available quickly or delayed for quality assurance, between providing original data or standard products and averaged/composited derived products. The functionality of the web interface will need to be expanded to appropriately search and display those data and the associated metadata, integrating elements such as station identifiers will need to be aligned, etc. The Ocean Observing System Regional Associations will be working with many of these datasets, though they have a much more physical focus than the biologically-oriented. CalCOFI. IOOS tools such as OPeNDAP and LAS may work well for some of these data types. CalCOFI, focused initially on discrete datasets, is now working toward a hybrid system able to handle both discrete and continuous underway cruise datasets.

### *Discrete Data Acquired on CalCOFI Cruises on CalCOFI Stations*

Over time, additional measurements and instruments are being used that complement the core measurements on CalCOFI stations. For example, since 1998 an Optical Plankton

Counter has been integrated into the bongo net frame so that in addition to the standard integrated plankton sample, the water column profile of 'counts' can be resolved by depth. These data are binned by depth to yield discrete data points.

With some planning, the discrete data taken on CalCOFI stations is probably the easiest type of data to integrate into the CalCOFI core data management and access system If these datasets adopt the Event Code, Station Code, location, and time conventions being prototyped by the focus data curators, then their data can be returned along with focus data when users query based on station, cruise, etc. The data will need to be quality controlled and held in a queriable format with associated standard metadata as described in section 3.3.

### *Continuous Data Acquired on CalCOFI Cruises on CalCOFI Stations*

Continuous data are composed of a stream of data, instead of discrete data points. An example is a CTD sensor that, once turned on, will continue sampling conductivity, temperature, and pressure on a set time interval until stopped. As with the discrete data described above, these data will need to contain project, event, and station code identifiers to connect it back to other CalCOFI data. An additional consideration for these data is deciding how the users will want them presented. For example, will users want the full, raw set of pings, or would they want a dataset that is quality controlled and smoothed across a larger time interval? Will they need a 3-d visualization system to see the depth patterns appropriately? These datasets are more likely to be found in binary or ascii file systems than relational databases used for the focus datasets and will thus need different storage, search, and access routines written. A second consideration is in what format the users will want the data returned. The largely relational data of the focus datasets lend themselves to ascii tables. Continuous data can also be represented in ascii but other data formats, such as NetCDF, are common for individual users and OPeNDAP exchanges. As with all datasets, metadata documenting the processing methods and other aspects of the dataset will be needed.

### *Continuous Data Acquired on CalCOFI Cruises between CalCOFI Stations*

Several types of continuous data are taken during CalCOFI cruises but not on CalCOFI stations in association with the routine sampling. These include: continuous ADCP data, the continuous underway fish egg sampling (CUFES), and underway sensing of temperature, salinity, fluorescence, PAR, and other data. Beyond the general considerations for all data sets (metadata documentation, quality control, formats, etc) the challenge for these data lies in the additional ways that users will want to work with the data: will they want to subset a section of a cruise track around an area of interest? Will they want to specify a time span of interest, instead of taking the whole cruise track from a several-week cruise? If they want to plot underway data along with station data, how close in space and time must the cruise track

data be to be relevant? None of these pose insurmountable technical problems, but each additional type of data and each additional feature on the online data system requires time and planning to implement.

New types of continuous measurements are on the horizon including moorings, Moving Vessel Profilers (MVP), and Autonomous Underwater Vehicles (AUVs) such as gliders launched from ship or shore. As they are incorporated into CalCOFI, they will present the technical challenge of how to manage and display large datasets through web interfaces.

### *Raster/Grid Data in the CalCOFI Region*

Another fundamentally different type of dataset is raster or grid data. These are data that provide values for each cell (or some cells, in the cases where data are missing) on a regular grid; satellite data and model output are two common and relevant examples. Sea surface temperature (AVHRR, MODIS-TERRA, MODIS-AQUA), ocean color (CZCS, OCTS, SeaWiFS, MODIS-TERRA, MODIS-AQUA, GLI), winds (QuikScat) and altimetry (Geosat) data are among the satellite data that exist for the CalCOFI region. These data, or images, are generally spatially comprehensive although short in terms of time span. Local satellite data centers exists at SWFSC through the CoastWatch program and at SIO[26]. Satellite imagery is also available through national centers and other local archives so it would be valuable for the larger community to consider useful forms of satellite products and methods of ensuring their easy accessibility. Large differences may exist in data products derived from the same sensors when they use different processing algorithms or a different selection of images to composite. How to store and disseminate model data is an issue only beginning to be explored. Metadata for the satellite imagery is relatively routinely available; metadata for modeled data is a subject of research. Each of these types of data would benefit from wide discussion with respect to data availability, use, and reuse.

One consideration for raster data is that the data files can be very large. An alternative to serving daily raw satellite coverages is to provide composites on weekly or longer intervals. In addition, dynamic compositing may be a good option in order to cover a fixed cruise period or event timeframe. Derived products (also known as Level-4 data) may be of interest to the larger community of researchers. Net Primary Production (NPP) is routinely calculated using chlorophyll, temperature (SST) and photosynthetically available radiation (PAR) from separate sources. Similarly Export Production (Export Flux, EF of Carbon) can be derived from NPP and SST. The formats of the derived datasets are similar to primary satellite datasets. While methods exist for composite and derived products, thought will need to go into determining the most useful products for CalCOFI users.

---

[26] http://spg.ucsd.edu/Satellite_Projects/Available_Imagery/Available_imagery.htm; http://swfsc.nmfs.noaa.gov/frd/Coastwatch/Coastwatch.htm

## 5.2 Data Inventory

Stocks and Baker have initiated a data inventory for CalCOFI. For the three focus CalCOFI datasets, this has involved documenting the software platforms and database designs used, discussing how well the system is working and where areas of improvement could be identified, etc. Beyond these datasets, the goal was to identify as many disparate datasets relevant to CalCOFI as possible and establish a template for future additions. Over the years, many additional studies, such as Scripps student dissertations, have been conducted in the CalCOFI region that may have produced important data. The Ocean Observing System is gathering data from instruments in the CalCOFI region. Other fisheries surveys have come and gone over the years in the area. For each dataset, the contents (temporal, spatial, and parameters), present data management status (i.e. whether it is digital or not), appropriate contact people, etc. are being collected. Such an inventory provides a basis from which to begin discussions of prioritizing future data integration and data rescue activities. This work is ongoing; the in-progress inventory is attached in Appendix 3. In addition, a simple dataset inventory form was developed for the data management meetings and is included as a prototype in Appendix 4.

An ongoing series of publications provide summaries of CalCOFI activities and findings and give a starting place for identifying "orphan data" of potential interest. The SIO biological/chemical/physical data report series was initiated in 1950 and output as hardcopy until 2002, when it became available online in PDF format. SWFSC reports include progress reports initiated in 1950, and Ichthyoplankton data reports. The California Cooperative Oceanic Fisheries Investigations Reports is a journal published every year or two from 1950-1978 and then annually starting in 1979 with index volumes produced periodically. In addition, a series of themed atlases was initiated in 1961. In 1996, a unique, comprehensive ichthyoplankton taxonomic monograph (the “red book”) was published as part of the Atlas series (Moser, 1996). Together these publications capture data, selected data plot presentations, scientific summaries of findings, and more comprehensive manuscripts.

# 6. Summary of Recent Activities

In the past 9-12 months, during the same timeframe as this white paper has been developed, there has been substantial activity related to CalCOFI data management including program review, dataset reorganization, implementation of new hardware and software, community training, and collaborative design activities. Some of these were initiated by Stocks and Baker, but many arose and were carried out by the CalCOFI community. This progress is summarized below.

Land-based

- Developing uniform indexing elements across databases, and implementing a common "event number" and activity table across datasets to facilitate integration.
- Expanding the SIO hydrographic dataset to include zooplankton biomass data.
- Identifying CalCOFI participants by drafting a personnel list, with roles and contact information (Appendix 1).
- Writing this white paper, which has served to consolidate knowledge of CalCOFI data management and to prompt discussions of requirements and design options for scaling up the system.
- Initiating an inventory of CalCOFI-related datasets and developing an approach to categorizing CalCOFI data (Figure 2; Appendix 3 & 4).

Field-based

- Prototyping a web-services-based system for real-time databasing of underway data at sea.
- Deploying a new event log collection protocol with standardized events and activities using a PC tablet networked to a central server to replace the traditional paper log.

Local Informatics Environment

- Establishing a land-based common hardware/software infrastructure that provides shared disk storage, backups, and security.
- Starting an Ocean Informatics Reading Group at SIO to explore conceptual issues in informatics and create a mechanism for group discussion and learning.
- Participating in local working groups and design teams that span SIO departments and data types (e.g. mooring, ship, shore/pier). Several data schema design teams coming together into a data schema working group is a recent example.
- Identifying a shared interest in open source environment and tools, such as for versioning systems (Subversion) and schema generation (Omnigraffle and DBDesigner)

In Partnership

- Holding meetings and individual interviews and having a presentation and an evening workshop at the Annual CalCOFI Conference. These served to inquire and inform within the CalCOFI program as well as across CalCOFI communities and across communities that are potential CalCOFI partners (Appendix 2).
- Preparing a proposal for the Southwest Fisheries Science Center that lays out and costs a development plan for the next year (see Appendix 5). This proposal was not successful but could still inform future efforts.
- Working in concert with partners such as the Long-Term Ecological Research sites, a local Gordon and Betty Moore Foundation-funded project at the San Diego Supercomputer Center, and other national research communities such as the National Science Digital Library.

This list is an indication of what occurs when a community acts together. The white paper began as a way of framing questions and assessing on-going work but in doing so became a prompt for community articulation and discussion. The overwhelming task of summarizing 50 years of data management by two new CalCOFI participants (Stocks and Baker) is possible only through the interest, cooperation and participation of community members. In providing an opportunity for the community itself to consider information system design, we are taking an innovative approach training, scaling, and design by recognizing an informatics environment as a substrate for community discussion, design, and development.

# 7. Scaling up: Considerations for PaCOOS

The National Marine Fisheries Service's west coast contribution to the Integrated Ocean Observing System is PaCOOS, the Pacific Coast Ocean Observing System. PaCOOS is envisioned as an extension of the CalCOFI sampling grid along the entire US west coast. Planning for PaCOOS data management is currently underway. The vision is for data collected from sampling grids run by multiple institutions to be accessible through a single integrated online interface.

CalCOFI, in essence, represents a microcosm of the data management challenges that PaCOOS will face: all of the data types that CalCOFI encompasses must be dealt with by PaCOOS. The lessons learned to date with CalCOFI can inform the PaCOOS process.

One lesson that emerges from CalCOFI is to plan in advance for data integration. In CalCOFI, the different data sets taken on each station (e.g. bottle data, ichthyoplankton, and zooplankton) each recorded station numbers in different ways, had different methods for recording location, varied in whether they stored time as local or GMT, etc. This makes it impossible to easily integrate data from across the datasets (i.e. to allow a user to download all the data from a particular station and cruise). The more that future PaCOOS sampling programs can agree on common recording protocols, data dictionaries, etc. before sampling begins, the easier integration will be.

Another lesson is that developing an integrated data access system is a task that requires devoted attention as well as additional work for data providers. It is a new job, not an extension of the work data curators are currently conducting; staffing and budgets need to reflect the additional effort of preparing procedures for data to be presented and used in new ways as well as coordinating across organizations, work practices, and roles.

A third lesson is that the Integrated Ocean Observing system currently provides limited guidance and support for biological point stored in relational formats. PaCOOS will have to plan for development costs and may also benefit from an active role on IOOS working groups to ensure that PaCOOS data needs are considered.

# 8. Conclusions

As CalCOFI moves towards an integrated online information system, its ability to organize is a critical factor. The state of technology is such that technology will not limit the growth of the CalCOFI data management system. Rather, resources and the ability of CalCOFI to organize for sustainable development are the key issues. The need for various levels of management and for mechanisms ensuring strong and continuing user input have already been discussed. The project should plan for multiple layers of development on different timescales: the CalCOFI data management efforts will have to continue developing as PaCOOS forward planning is conducted. Modular development, with plans for flexibility through time, will better suit than one monolithic development effort with no ability to adapt as technology evolves over time and as PaCOOS and IOOS mature. System design, prototype development, community development, user enrollment, and surveys of existing tools are ongoing processes, not one-time activities.

Finally, the human element should receive equal attention as the technical. Both CalCOFI and PaCOOS will need to plan for innovative, ongoing training and support in order to achieve a long-term effort. Development of an informatics community will help create and maintain the expertise needed at local levels. The general community will need to become invested in the project's progress through meetings, outreach information, mailing lists, etc.

With appropriate planning and support, CalCOFI, which is already a model for long-term ocean observing, can become a contemporary model for interdisciplinary ocean data management.

# Appendix 1: Personnel List
# (Draft – in development)

| first_name | last name | org. | dept. | division | phone | email | CalCOFI_tasks |
|---|---|---|---|---|---|---|---|
| Dimitry | Abramenkoff | NMFS | SWFSC | FRD | 858-546-7126 | dimitry.abramenkoff@noaa.gov | field data collection |
| David | Allison | UCSD | SIO | MPL | 858-534-8947 | dallison@mpl.ucsd.edu | remote sensing, bio-optics |
| Karen | Baker | UCSD | SIO | IOD | 858-534-2350 | kbaker@ucsd.edu | LTER coordination |
| Lisa | Ballance | NMFS | SWFSC | | 858-546-7173 | lisa.ballance@noaa.gov | ecosystems studies group |
| Jay | Barlow | UCSD | SIO | IOD | 858-546-7178 | jbarlow@ucsd.edu | marine mammal data |
| Noelle | Bowlin | NMFS | SWFSC | FRD | 858-546-7155 | noelle.bowlin@noaa.gov | field data collection |
| Rick | Brodeur | NMFS | NWFSC | | 541-867-0336 | rick.brodeur@noaa.gov | Newport line |
| Richard | Charter | NMFS | SWFC | FRD | 858-546-7157 | richard.charter@noaa.gov | ichtyoplankton data coordination, analysis, db, web, requests |
| Francisco | Chaves | MBARI | | | 831-775-1709 | chfr@mbari.org | lead on MBARI line |
| Dave | Checkley | UCSD | SIO | IOD | 858-534-4228 | dcheckley@ucsd.edu | CUFES data collection |
| Teresa | Chereskin | UCSD | SIO | IOD | 858-534-6368 | tchereskin@ucsd.edu | ADCP currents |
| Ray | Conser | NMFS | SWFSC | | 858-546-5688 | ray.conser@noaa.gov | stock analysis division, CalCOFI data to serve the stock |
| Stephen | Diggs | UCSD/NMFS | SIO | STS | 858-534-1108 | sdiggs@ucsd.edu | online sysem prototyping |
| Ron | Dotson | NMFS | SWFSC | FRD | 858-546-7085 | ron.dotson@noaa.gov | field data collection |
| Ralf | Goericke | UCSD | SIO | IOD | 858-534-7970 | rgoericke@ucsd.edu | nutrient data |
| John | Greybeal | MBARI | | | 831-775-1956 | graybeal@mbari.org | metadata expert |
| Dave | Griffith | NMFS | SWFSC | FRD | 858-546-7155 | dave.griffith@noaa.gov | field data collection |
| Shaun | Haber | UCSD | SIO | IOD | 858-534-2708 | srhaber@ucsd.edu | collaborative tools, database programming |
| Amy | Hays | NMFS | SWFSC | FRD | 858-546-7130 | amy.hays@noaa.edu | field coordination, data collection, data entry |
| Roger | Hewitt | NMFS | SWFC | | 858-546-5602 | roger.hewitt@noaa.gov | director SWFSC |
| John | Hunter | NMFS | SWFSC | | 858-534-4199 | jrhunter@ucsd.edu | PaCOOS coordination |
| Steve | Joner | | | | 360-417-8946 | gofish@olypen.com | Makah tribe chief biologist |
| Matti | Kahru | UCSD | SIO | IOD | 858-534-8947 | mkahru@ucsd.edu | remote sensing, bio-optics |
| Carol | Kimbrell | NMFS | SWFSC | | 858-546-7106 | carol.kimbrell@noaa.gov | SWFSC fishery biologist |
| Nancy | Lo | NMFS | SWFSC | | 858-546-7123 | nancy.lo@noaa.gov | products for stock analysist; raw data user |
| Sue | Manion | NMFS | SWFSC | FRD | 858-546-7143 | sue.manion@noaa.gov | field data collection |

| first_name | last name | org. | dept. | division | phone | email | CalCOFI_tasks |
|---|---|---|---|---|---|---|---|
| Sherry | McCann | UCSD | SIO | IOD | 858-534-6780 | scummings@ucsd.edu | data coordination, quality control, archive |
| John | McGowan | UCSD | SIO | IOD | 858-755-2065 | jmcgowan@ucsd.edu | pelagic ecology (SIO emeritus) |
| Arthur | Miller | NMFS | SIO | CRD | 858-534-8033 | ajmiller@ucsd.edu | data user |
| Gregory | Mitchell | UCSD | SIO | IOD | 858-534-2687 | gmitchell@ucsd.edu | remote sensing, bio-optics |
| Mark | Ohman | UCSD | SIO | IOD | 858-534-2754 | mohman@ucsd.edu | zooplankton |
| Chuck | Oliver | NMFS | SWFSC | | 858-546-7172 | chuck.oliver@noaa.gov | SWFSC |
| Bill | Peterson | NMFS | NWFSC | | 541-867-0201 | bill.peterson@noaa.gov | Newport line |
| Fernando | Ramirez | UCSD | SIO | IOD | 858-534-2888 | framirez@ucsd.edu | data collection |
| Christian | Reiss | NMFS | SWFSC | FRD/AMLR | 858-546-7127 | christian.reiss@noaa.gov | online system prototype |
| Frank | Schwing | NMFS | PFEL | | 831-648-9034 | Frank.Schwing@noaa.gov | Acting director |
| Jennifer | Sheldon | UCSD | SIO | IOD | 858-822-0674 | jlsheldon@ucsd.edu | data collection and processing |
| Paul | Smith | UCSD | SIO | IOD | 858-546-7169 | paulsmith@ucsd.edu | CalCOFI data user |
| Karen | Stocks | UCSD | SDSC | | 858-534-1864 | kstocks@sdsc.edu | PaCOOS coordination |
| George | Sugihara | UCSD | SIO | PORD | 858-534-5582 | gsugihara@ucsd.edu | data user; quantitative eco |
| Elizabeth | Venrick | UCSD | SIO | IOD | 858-534-2068 | evenrick@ucsd.edu | CalCOFI/SIO director |
| Jerome | Wanetick | UCSD | SIO | IOD | 858-534-7999 | jwanetick@ucsd.edu | IOD computational data center head |
| Stephanie | Watson | MBARI | | | 831-775-1987 | swatson@mabari.org | CenCOOS coordinator (former) |
| William | Watson | NMFS | SWFSC | | 858-546-5647 | william.watson@noaa.gov | fisheries biologist |
| James | Wilkinson | UCSD | SIO | IOD | 858-822-0674 | jwilkinson@ucsd.edu | field coord., curator of hydrographic data |
| David | Wolgast | UCSD | SIO | IOD | 858-534-3857 | dwolgast@ucsd.edu | field coordination, data collection, analysis |

# Appendix 2 – Meeting Summaries

## CalCOFI Data Management Meeting Summary – 18 Oct 04

*Participants: Karen Stocks, Karen Baker, David Allison, Richard Charter, Sherri McCann, Stephen Diggs, Ralf Goericke, Mati Kahru, Gregory Mitchell, Mark Ohman, Fernando Ramirez, Christian Reiss, Jennifer Sheldon, Jerome Wanetick, Jim Wilkinson, Dave Wolgast*

This meeting brought together participants at SIO and SWFSC interacting with CalCOFI-related datasets in diverse ways. Karen Stocks and Karen Baker, tasked with forward planning in support of ongoing CalCOFI data management teams, coordinated the meeting. They opened with a presentation that covered 1) a consideration of local efforts within the context of emerging institutional, regional and national partnerships (Figure A); 2) an outline of generic data system with elements from the ingestion of multiple data types through user query and integration (Figure B); and 3) the activities Baker and Stocks are undertaking and the products they are developing.

A round-table discussion followed during which several themes emerged. First, that the integration of data is important for supporting analysis and visualization. Second, that metadata, standards, exchange protocols, and core categories (that is, joint nomenclature and language) play a critical role in data systems. Although the range of tasks is large and the data/system/organizational relationships complex, participants initiated two key processes: shared infrastructure and information flow. A series of products and processes were discussed including a recent PACOOS proposal to fisheries and a data management presentation at the upcoming CalCOFI Conference (Figure D). In addition, a personnel directory and a dataset inventory were handed out with a request for edits and updates. A follow-up technical mini-meeting will be planned for December after the November cruise, at which standards, metadata, and data integration will be considered in detail. Meanwhile, a critical issues list was begun and will be circulated prior to the next meeting.

Critical Issues

- Inventory local datasets, data types, and data sizes as well as expectations and resources
- Prioritize data tasks and expectations with respect to resources
- Identify and prioritize user community participant groups and products
- Develop mechanisms for participants to engage in identifying requirements and designing the system
- Articulate and bridge local individual metadata standards and formats
- Identify national metadata standards and exchange protocols
- Identify a data system model and mechanisms to interface with local systems
- Create a shared information infrastructure for diverse groups and diverse data types
- Consider how to build out and support partnerships
- Identify critical field collection-data system factors such as station name conventions and reporting formats
- Establish assessment mechanisms to ensure the system meets and continues to meet user needs

# Appendix 2 – Meeting Summaries, cont.

**Ocean Informatics Exchange Workshop**
November 05, 2004
CCS conference room (at the foot of the SIO pier)

The Ocean Informatics Exchange Workshop continues a dialogue initiated last year between folks managing oceanographic field data at SIO and WHOI. We are doing some forward planning with respect to the multiple dimensions of infrastructure and the design process for a contemporary information environment appropriate for communities in general as well as for organizations such as the SIO Integrative Oceanography Division (IOD).

- Organized by Karen Baker, Jerry Wanetick, and Cyndy Chandler

**Agenda**

9:00 Agenda & Logistics (Baker))
Welcome (Guza)
9:15 Introductions (round-table)
9:30 Reviewing Past and Present
- Conceptual Model
- Tensions/Balances
IOD Context (Wanetick)
10:30 Break
11:00 WHOI Context (Chandler)
11:15 Semantics and Terms
Informatics: Historical Perspective
Informatics: Domain Perspective
11:45 Wrap-up
12:00 Meeting review (round-table)
12:15 Lunch

**Participants**

Karen Baker, SIO, kbaker@ucsd.edu
Cyndy Chandler, WHOI, cchandler@whoi.edu
Art Gaylord, WHOI, agaylord@whoi.edu
Shaun Haber, SIO, srhaber@ucsd.edu
Florence Millerand, LCHC, fmillera@ucsd.edu
Dawn Rawls, SIO, drawls@ucsd.edu
Uwe Send, SIO, usend@ucsd.edu
Karen Stocks, SDSC, kstocks@sdsc.edu
Wayne Suiter, SIO, wsuiter@ucsd.edu
Julie Thomas, SIO, jothomas@ucsd.edu
Jerry Wanetick, SIO, jwanetick@ucsd.edu
Lynn Yarmey, SIO, lyarmey@ucsd.edu

# Appendix 2 – Meeting Summaries, cont.

***Southwest Fisheries Science Center NOAA Fisheries***
***Scripps Institution of Oceanography***
***California Department of Fish and Game***

**CalCOFI Annual Conference 2004**
**California Cooperative Oceanic Fisheries Investigations**
**17 Nov 2004 Wednesday**

Data Management Activities:

CalCOFI Data Management: Today and Beyond
Karen I. Stocks and Karen S. Baker
1440-1500 Presentation, Sumner Auditorium

Integrating CalCOFI Datasets in a Web-Based Browser
Steve Diggs and Christian Reiss
1530-1730 Poster, IGPP Munk Conference Room, in IGPP

Data Management Workshop
K.Stocks, K.Baker, S.Diggs, C.Reiss, R.Charter
1630-1800 Helen Raitt Room, 3rd floor SIO library

# APPENDIX 3 – DATA INVENTORY

Note that the following columns could not be included due to space restrictions: format, system, date started, date ended, and notes

| Dataset | collection method | Description | Project | Contact | Inst. | Measurement Method | Availability |
|---|---|---|---|---|---|---|---|
| **CTD Profile DATA** | CTDpackage | CTD casts taken on station with standard CTD package and additional sensors measuring temperature, salinity, fluorescence, irradiance(Epar0&Eparz), light transmittance(660nm), dissolved oxygen. | CalCOFI | Wilkinson | SIO | CTD package | by request from J Wilkinson at SIO; also from R.Charter at SWFSC (for N.Lines) |
| depth | CTDpackage | | CalCOFI | Wilkinson | SIO | pressure | |
| temperature | CTDpackage | | CalCOFI | Wilkinson | SIO | temperature | |
| salinity | CTDpackage | | CalCOFI | Wilkinson | SIO | salinity | |
| fluorescence | CTDpackage | | CalCOFI | Wilkinson | SIO | fluorometer | |
| irrradiance-Epar0&Eparz | CTDpackage | | CalCOFI | Wilkinson | SIO | PAR meters | |
| light transmission @ 660nm | CTDpackage | | CalCOFI | Wilkinson | SIO | transmissometer | |
| dissolved oxygen | CTDpackage | | CalCOFI | Wilkinson | SIO | oxygenprobe | |
| distance from bottom | CTDpackage | | CalCOFI | Wilkinson | SIO | altimeter | |
| nitrate | CTDpackage | | CalCOFI | Wilkinson | SIO | nitrate sensor | |
| **Bottle DATA** | bottle | Bottle samples from CTD rosette used for temperature, salinity, oxygen, nutrients (N03,N02,PO4, SI(OH)4, ammonium), DOC,DOM,POC,PON,DOC,DON analysis. | CalCOFI | Wolgast | SIO | BottleSamples | www.calcofi.org. All data available online in ascii (IEH) format and in access db format. An old CalCOFI cgi script search program delivering IEH records on plotform gyre (CIFT CalCOFI IEH Formatting Tool) is not currently available. Steve Diggs working on integration with fisheries components. |
| temerature discrete | bottle | | CalCOFI | Wolgast | SIO | salinometer | |
| salinity discrete | bottle | | CalCOFI | Wolgast | SIO | salinometer | |
| oxygen discrete | bottle | | CalCOFI | Wolgast | SIO | autoWinkler | |
| nutrients (NO3, NO2, P04, Si(OH)4) | bottle | | CalCOFI | Wolgast | SIO | auto analyzer | |
| nutrients - ammonium | bottle | | LTER | Wolgast | SIO | isaacs/sts; odf collect | |
| discrete chlorophyll a concentration | bottle | | CalCOFI | Wolgast | SIO | Fluormetric analysis run on board ship on extracted chlorophyll samples taken from CTD bottles. | |
| primary production-particulate | bottle | | CalCOFI | Wolgast | SIO | Productivity measurements taken from CalCOFI CTD bottles. Incubation C14-uptake POC. | |

| Dataset | collection method | Description | Project | Contact | Inst. | Measurement Method | Availability |
|---|---|---|---|---|---|---|---|
| **Zooplankton Data** | calbobl | Non-ichthyoplankton portion of calcofi net tow samples. Some samples identified, some not processed. | LTER | Ohman | SIO | Physical collection exists. | Metadata and Zooplankton bulk volume is online. Taxonomically-resolved data are not available online. |
| mesozooplankton-euphausiids ~32 sentinel species | calbobl | | LTER | Ohman | SIO | microscopy | |
| mesozooplankton-zoo-pooled sentinel species | calbobl | | LTER | Ohman | SIO | microscopy | |
| mesozooplankton-size_distributions | zooscan | A flatbed scanner used in laboratory on shore to identify zooplankton. | LTER | Ohman | SIO | zooscan | |
| mesozooplankton-optical size classes-OPC | calbobl | An optical plankton counter deployed on a bongo net oblique tow within 2 nautical miles of station. Continuous sampling along a half-hour transect. | CalCOFI | Checkley | SIO | An optical plankton counter deployed on a bongo net oblique tow within 2 nautical miles of station. One tow per station carried out. Continuous sampling along a half-hour transect. Measures 270u-14mm although the majority of material is 500u to 1mm | Presently undergoing analysis. |
| meso+microzooplankton-optical size classes-LOPC | calbobl/test | | LTER | Checkley | SIO | A laser optical plankton counter deployed on a bongo net oblique tow within 2 nautical miles of station. One tow per station. Continuous sampling along a half-hour transect. Measures 100u-14u although the majority of material is 500u to 1mm. | Data handling under development. |
| ichthyopolankton-eggs-fish larvae | | Fish eggs and larvae from net tows fully documented, put in SWFSC database, and physical samples archived. | CalCOFI | Charter | SWFSC | Physical collection exists. | |
| fish larvae-neuston-tows-manta nets surface tow | ichthyoplankton | | CalCOFI | Charter | SWFSC | microscopy; fish larvae | |
| ichthyoplankton-eggs-pairovet | ichthyoplankton | | CalCOFI | Charter | SWFSC | Pairovet | |
| macrozooplankton-biomass nekton-tows-bongo net | calbobl | | CalCOFI | Charter | SWFSC | displacement volume_to_zooplankton biomass | |

| Dataset | collection method | Description | Project | Contact | Inst. | Measurement Method | Availability |
|---|---|---|---|---|---|---|---|
| ichthyoplankton-eggs-cufes | ship pipe intake | Continuous underway fish egg counter used on calcofi cruises but also on separate cruises. | CalCOFI | Charter | SWFSC | CUFES (Spring only). using a pumped sample from a ship standpipe through a concentrator so that a 2% concentrate of the water is run through a mesh filter. The filter is retrieved approximately every half hour (depending on egg density) for a ship microscopic analysis identifying eggs - sardine, anchovy, hake and macheral. Approximately 350 samples per cruise are collected. Adaptive sampling: only collect as far offshore as eggs go. Net tows taken alongside for calibration. Filters are preserved as a physical collection. | Egg maps posted online. Quantitative data available on request. |
| **Other cruise data** | | | | | | | |
| weather | on-deck | Suite of weather measurements taken on each standard CalCOFI station: wave direction, wave height, wave period, wind direction, wind speed, barometer, dryTemperature, wet temperature, weathercode, cloud type, cloud amount, visibility, and forel color. | CalCOFI | Wolgast | SIO | weather | |
| bucket temperature | bkt-temp | Measurement taken at each standard CalCOFI station. | CalCOFI | Wolgast | SIO | | |
| secchi disk | secchi | Measurement taken at each standard CalCOFI station. | CalCOFI | Wolgast | SIO | secchi | |
| primary production-dissolved DO14C | bottle | | LTER | Goericke | SIO | incubation 14-update DOC | |
| HPLC pigments | bottle | Phytoplankton taxon-specific pigments concentrations taken on calcofi cruises, starting 2003 or 2004. | LTER | Goericke | SIO | HPLC | Local desktop Msaccess database; has ids to link back to calcofi CTD/bottle data. |

| Dataset | collection method | Description | Project | Contact | Inst. | Measurement Method | Availability |
|---|---|---|---|---|---|---|---|
| phytoplankton species counts | bottle | Mixed layer bottle sample at cardinal stations pooled into regions and counted under microscope. | LTER | Venrick | SIO | Microscopy on nano and microplankton. Inverted phase contrast microscope at magnification of 100x and 250x used on formalin preserved samples. Data entered on field log and entered into digital files as ascii on desktop pc with access through basic programs. | |
| chlorophyll size fractionation | bottle | | CalCOFI | Goericke | SIO | size fractionation | |
| nano & microplankton | bottle | | LTER | Ohman | SIO | FlowCAM (trial) | |
| nanoplankton | bottle | | LTER | Landry | SIO | slides: counting-automated microscopy | |
| microplankton | bottle | | LTER | Landry | SIO | slides: counting-automated microscopy | |
| bacteria & picoautotrophs | bottle | | LTER | Landry | SIO | Flow-cytometry | |
| particulate organic matter (POM: POC, PON) | bottle | | LTER | LTERTech | SIO | dry combustion; filters preserved in liquid nitrogen stores. | |
| dissolved organic matter (DOM: DOC, DON) | bottle | | LTER | | SIO | combustion | |
| dissolved organic carbon (DOC) | pump | Selected, limited sampling | LTER | Aluwihari | SIO | combustion | |
| total organic carbon, nitrogen (TOC, TN, TON) | bottle | Total organic carbon and nitrogen | LTER | Aluwihari | SIO | combuston | By request. |
| tracemetal | TMpump | Trace metal measurement; collection using clean pump. | LTER | Barbeau | SIO | FeLume flow injection | By request. |
| tracemetal: ironconc | Tmgoflo | Trace metal measurement; collection using goflo bottle. | LTER | Barbeau | SIO | FeLume flow injection | By request. |
| tracemetal | Tmpole | Trace metal measurement; collection using two one liter bottles on 7 meter pole. | LTER | Barbeau | SIO | FeLume flow injection | By request. |
| sea surface pCO2 | SeawaterIntake | | LTER | Frederick | MBARI | IR absorbance | |
| upper ocean currents | ADCP | Accoustic instrument measuring current and biomass estimate. | LTER | Chereskin | SIO | Instrument mounted underneath ship. Sends out accoustic signal. Can determine current (from dopler timing) and estimate plankton and fish biomass (from accoustics) | |
| continuous underway system-CUDLES | underway | Underway data for temperature, salinity, dissolved oxygen, and chlorophyll | CalCOFI | Goericke | SIO | A ship underway measurement system of T, sal, DissOxy, Fluor | |

| Dataset | collection method | Description | Project | Contact | Inst. | Measurement Method | Availability |
|---|---|---|---|---|---|---|---|
| continuous underway system-NOAA | underway | Underway data for temperature, salinity, dissolved oxygen, and chlorophyll | CalCOFI | Goericke | SIO | underway T, sal, DissOxy, Fluor | |
| optics-PRR | profile | | NASA, SCCOOS | Mitchell | SIO | | |
| optics-MER | profile | | NASA, SCCOOS | Mitchell | SIO | | |
| optics-SCCOOBOP | profile | | | Goericke | SIO | | |
| Satellite | images | | LTER, SCCOOS | Mitchell | SIO | | |
| Satellite | images | | LTER | Ramirez | SIO | MODIS-SST | |
| whale soundings | accoustic | Acoustic data tasken with whale sightings. Will start fall 2004. | MPL | Hildibrandt | SIO | whale soundings | Timeline uncertain. |
| whale sightings | observations | | MPL | Hildibrandt | SIO | whale sightings | Timeline uncertain. |
| POC, PON | bottle | | NASA | Mitchell | | | |
| Pigments HPLC | bottle | | NASA | Mitchell | | | |
| **Studies** | | | | | | | |
| bird surveys | birds | Point Reyes Bird Observatory runs bird survey transects over CalCOFI region (separate cruises from CalCOFI surveys). | Point Reyes Bird Observatory | | | underway birds | |
| Marine Mammal Surveys | program | Eastern Pacific cruises lasting 2-3 months taken every 2-3 years. Transcects every 10 miles. | NOAA | | | | |
| Albacore Logbooks | log_report | Voluntary logbooks from US albacore fishermen. Mainly albacore with geolocations. | PACOOS potential | | | | |
| Recreational Landings (LA Times Reported Catch) | log_report | Publishes records of reported catch (landings by oat) of the recreational fleet. For >40yrs. | PACOOS potential | | | | SWFSC has digitized and continues to track this dataset. |
| Sportfishing Logbooks | log_report | Same fleet as LA Times time series but based on logbooks of landings. California Fish and Game. | PACOOS potential | | | | |
| Hake Acoustics | program? | | PACOOS potential | | | | |
| Pacfin | program | Fishery-dependent data on all commercial west coast fisheries. Landings and some lat/lon catch data from both observers and logbooks. | PACOOS potential | | | | |
| Drift gillnet Observer Logbooks | program | Observer data for the commercial drift gillnet fishery (much has now been closed). Extends out to EEZ. | PACOOS potential | Rasmussen | | | |
| Shelf and slope groundfish | program | Identification of all fish in trawls. NWFSC holds. | PACOOS potential | | | | |
| SiMON | | Sanctuary Integrated Monitoring Network. | potential | | | | www.mbnms-simon.org |

| Dataset | collection method | Description | Project | Contact | Inst. | Measurement Method | Availability |
|---|---|---|---|---|---|---|---|
| **Lines** | | | | | | | |
| MBARI Line | program | Transcect line off Monterey. Done in 60's through 80's then picked up in last few years in coordination with NOAA. Contact F.Chavez. | PaCOOS | | | | Held on CD |
| Humbolt Line | program | Not started yet. Forseen in 2005 or so. No contact yet. | PaCOOS | | | | |
| Newport Line | program | No contact yet. | PaCOOS | | | | |
| Imecocal Lines | program | Independent program in Mexico that parallels CalCOFI work. | CalCOFI and PACOOS related | | | | |
| **Related Programs** | | | | | | | |
| DataZoo | Datazool | Coastal measurements made from moorings. | IOD | Wanetick | | | http://iod.ucsd.edu |
| LTER | LTER | Long-term measurement collection at Palmer Station, Antarctica (PAL) and at the California Current Ecosystem (CCE) | LTER | Baker | | | http://pal.lternet.edu |
| Ocean Biogeographic Information System | OBIS | Global marine species distribution data | | Stocks | | | http://www.iobis.org |
| NEOCO | NEOCO | Network for Environmental Observations of the Coastal Ocean. SIO pier data, plus several other stations along California coast. | OOS | Wanetick | | | Under development. |
| SCCOOS | SCCOOS | Southern California Coastal Ocean Observing System. Mexico border (and beyond with Mexico parnters) up to P. Conception. Partners include SCWRRP, universityies, JPL, PISCO data, LTER moorings. | OOS | Davis | | | |
| CenCOOS | CenCOOS | Central California Coastal Ocean Observing System. From Oregon border to San Lous Obispo. Line #67: 1988 -ongoing with gaps. 2 moorings. Includes SIMON. | OOS | Watson | | | |
| NANOOS | NANOOS | Northwest Association of Networked Ocean Observing Systems. | OOS | Barth | | | |
| PFMC | PFMC | Pacific Fisheries Management Council | PACOOS | Donal McIsaak | | | |
| **Print Documents** | | | | | | | |
| Fish Identification Guide | | Gives description for every fish species - drawing, ID keys, etc. by Jeff Moser | CalCOFI | | | | Hard Copy only |

# Appendix 4: Data Inventory Template

| **NEW** | **EXAMPLE** |
|---|---|
| Dataset:_____ | Dataset: hydrographic bottle (temp, salinity) |
| Collection Method:_____ | Collection Method: CTD-bottle |
| Common Name:_____ | Common Name: discreteTempSal |
| Data Collection Group:_____ | Datacollection Group: bottle |
| Dataset Group:_____ | Dataset Group: IEH |
| Format:_____ | Format: ascii |
| Project:_____ | Project: CalCOFI |
| Investigator:_____ | Investigator: Ralf Goericke |
| Institution:_____ | Institution: SIO |
| Measurement Method:_____ | Meas. Method: salinometer w/bottle sample |
| Measurement Category:_____ | Measurement Category: coreC |
| Begin Date:_____ | Begin Date: 19xx |
| Description:_____ | Description: samples processed on ship |

Collection Methods: Underway, CTD-profile, CTD-bottle, Profile/tow, Plankton tow, Pump, Profile, Publication, Log-report

Data CollectionGroups: ADCP, Hydro_profiles, Bottle, Zooplankton-species, Weather

Dataset Groups: IEH, Pigments, Zoopl-euphausites, Zoopl-pooled, Icthyo, Optics

Institutions: CalCOFI, LTER, MBARI

Measurement Categories: CoreC (CalCOFI), CoreL (LTER), Associated, Related

# Appendix 5: 2004/2005 Proposal

In August 2004, a data management plan and budget was provided to the Southwest Fisheries Science Center for consideration for their Federal Year 2005/2006 funding. Below, we provide the submitted text; a budget summary is given in Appendix 5. This work will ***not*** be funded this year, but it provides one model for moving towards integrated, online access to a suite of CalCOFI data types. The proposal also furthers planning for the larger PaCOOS system by 1) using CalCOFI data as a testbed for developing methods and procedures that can scale to the PaCOOS system and 2) conducting an inventory of PaCOOS data sets and resources and 3) organizing advisory committees for PaCOOS data system development.

**Proposal text as submitted:**

> **Description:** This work will create the foundation for a scaleable, IOOS-compatible online information system for CalCOFI sentinel line data. Initial system implementation will focus on three core components that are common across the planned set of PaCOOS sentinel lines: hydrographic cast data, ichthyoplankton tow data and CTD data. The project will design and implement an online interface to search and access these data types, based on DMAC-compliant OPeNDAP/LAS and Ocean Biogeographic Information System technologies, extended and supplemented as needed. Procedures will be documented to provide assistance to organizations starting new lines and thus facilitate the incorporation and integration of new data providers. In addition, the project will bring together other institutions running sentinel lines within PaCOOS to develop community consensus on standard data reporting formats and procedures to ensure that data are appropriately documented, preserved, and made accessible. Oversight, Technical and User Committees, along with other user outreach mechanisms, will provide guidance for overall development and specific functionality. An initial inventory will be carried out within the PaCOOS community to document existing datasets, technical and human resources, and system architectures. Data rescue will begin for high priority datasets identified during the inventory.
>
> **Justification**: The CalCOFI data sets offer an ideal starting point for developing a scaleable system capable of accommodating the additional PaCOOS sentinel lines expected in FY07. It is important to begin the development of the framework for this system now, so that practices can be discussed and developed before new groups begin contributing data, avoiding post-hoc efforts to integrate heterogeneous data in favor of planned data approaches. Currently, different components of CalCOFI data are housed and managed in entirely separated data sets (ichthyoplankton in one system, CTD casts in another, etc.) with limited integration and limited online access. This hampers investigations of the bio-physical interactions central to the purpose of CalCOFI and PaCOOS and limits the data from reaching their full utility. In addition, it is important for technical expertise within PaCOOS to be identified and to start coordination in order for PaCOOS management of sentinel line data to keep pace with data expansion in FY 07. Additionally, the project will contribute to IOOS-level development of biological data system requirements and recommendations, which are still early in development, to ensure that the national system is responsive to sentinel line data types and applications.

BUDGET

| | Year 1 | Year 2 | |
|---|---|---|---|
| 1. Developing an integrated information management and online access system | | | |
| Salary | | | |
| Full-time Programmer II | | | |
| 1/2 time Programmer III/IV | | | |
| 3 months K. Stocks | | | |
| 2 months K. Baker | | | |
| 2 months logistics support | | | |
| 1 months Jerry Wanetick | | | |
| 2 months summer graduate student | | | |
| **Salary Total** | 201,240 | 209,290 | |
| Other | | | |
| Hardware & Software | 15,000 | 8,000 | |
| Travel | 1,500 | 1,500 | |
| Supplies | 4,000 | 4,000 | |
| Report/publication costs | 500 | 1,000 | |
| 2. Inventory of data and systems | | | |
| Salary | 6,000 | | |
| Travel | 4,000 | 1,000 | |
| 3. Data Rescue | | | |
| Salary | 22,500 | 22,500 | |
| Supplies/Equipment | 5,000 | 5,000 | |
| 4. PaCOOS-level sentinel-line data & system coordination | | | |
| Salary | | | |
| 1 month Communications system support | | | |
| 1 month Stocks | | | |
| 1 month Baker | | | |
| 1 month logistics support | | | |
| **Salary Total** | 33,500 | 28,600 | |
| Other | | | |
| Zooplankton Meeting | 4,000 | | |
| Oversight Committee Meetings | 5,000 | 5,000 | |
| Technical Working Group Meetings | 10,000 | 10,000 | |
| User Group Meeting | 5,000 | 5,000 | |
| Meeting food (lunch, coffee) | 2,000 | 1,600 | |
| Collaboration compensation | 15,000 | 15,000 | |
| Hardware/software | 5,000 | | |
| Travel | 2,000 | 2,000 | |
| Report/publication costs | 1,000 | 1,000 | |
| Distance communication (conf. calls) | 500 | 500 | |
| Supplies | 2,000 | 2,000 | |
| | | | Total all yrs |
| Total Direct | 344,740 | 322,990 | 667,730 |
| **Total with 53.5% overhead** | **529,176** | **495,789** | **1,024,965** |
| 5. Additional Partner Participation (Optional extension) | | | |
| 1/4 time programmer (incl. grantee overhead) | 17,500 | 18,200 | |
| Subcontracts for system components | 30,000 | 30,000 | |
| UCSD Subcontract Overhead | 13,500 | 13,500 | |
| **Total with Extension** | **590,176** | **557,489** | **1,147,665** |

# Appendix 6: Glossary of Acronyms

ADCP: Acoustic Doppler Current Profiler
CalCOFI: California Cooperative Oceanic Fisheries Investigations
CCE: California Current Ecosystem LTER
CenCOOS: Central California Ocean Observing System
CTD: an instrument that measures Conductivity, Temperature, and Depth
CUFES: Continuous Underway Fish Egg Sampler
DiGIR: Distributed Generic Information Retrieval protocol
DMAC: Data Management and Communications, a component of IOOS
EML: Ecological Metadata Language
FGDC: Federal Geospatial Data Committee Content Standard for Geospatial Metadata
GBIF: Global Biodiversity Information Facility
GCMD: Global Change Master Directory
GMT: Greenwich Mean Time
GPS: Global Positioning System
IOD: Integrative Oceanography Division, a part of SIO
IOOS: Integrated and Sustained Ocean Observing Initiative
ISO: International Organization for Standards
LAS: Live Access Server
LTER: Long Term Ecological Research Network
MBARI: Monterrey Bay Aquarium Research Institute
NCEAS: National Center for Ecological Analysis and Synthesis
NMFS: National Marine Fisheries Service
NOAA: National Oceanic and Atmospheric Administration
NODC: National Ocean Data Center
NSF: National Science Foundation
OBIS: Ocean Biogeographic Information System
Ocean.US: National Office for IOOS
OPeNDAP: Open-source Project for a Network Data Access Protocol
ORION: Ocean Research Interactive Observatory Networks
PaCOOS: Pacific Coast Ocean Observing System
PAR: Photosynthetically Available Radiation
SCCOOS: Southern California Coastal Ocean Observing System
SDSC: San Diego Supercomputer Center
SIO: Scripps Institution of Oceanography
SWFSC: Southwest Fisheries Science Center, a part of NMFS
THREDDS: Thematic Realtime Environmental Distributed Data Services
UNESCO: United Nations Educational, Scientific and Cultural Organization ...
WMS, WFS, WCS: Web Map Service, Web Feature Service, Web Coverage Service