Title

DataZoo: an Oceanographic Information System Supporting Scientific Research

Permalink

https://escholarship.org/uc/item/139019q8

Authors

Baker, Karen S
Kortz, Mason
Conners, James

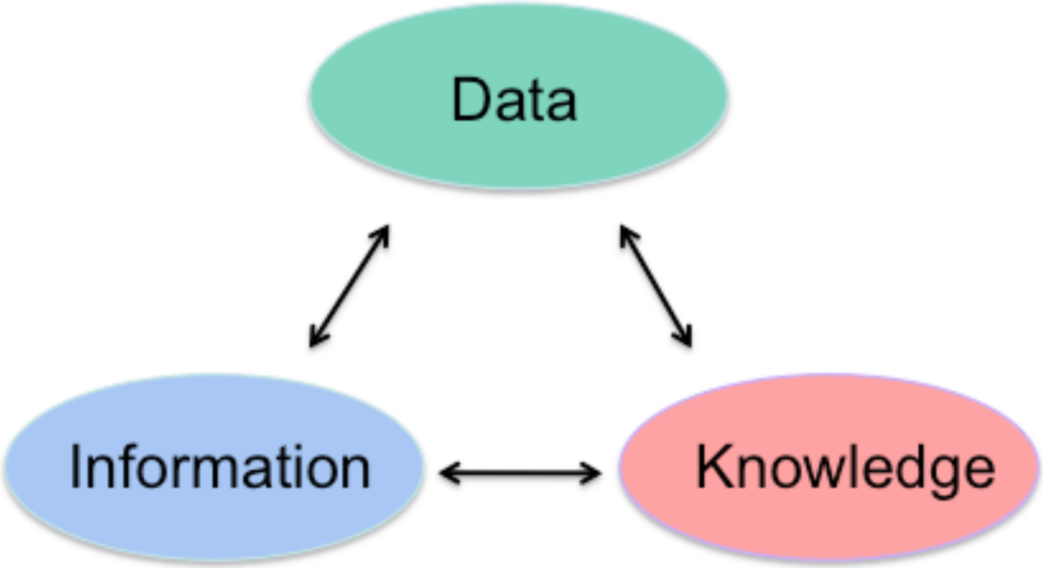
Publication Date

2011-06-01

DataZoo: an Oceanographic Information System Supporting Scientific Research

Karen S. Baker, Mason Kortz, and James Conners

Integrative Oceanography Department
Scripps Institution of Oceanography, UCSD



This report is about the development, management, and continuing design of an information system within a field-oriented, scientific research environment in collaboration with those who collect and use the data.

Scripps Institution of Oceanography Technical Report
June 2011

Table of Contents

DataZoo: an Oceanographic Information System Supporting Scientific Research

1 Introduction

This report is about the development of DataZoo, an information system. Design, development, and use of this computer-based technology that aggregates, organizes, and delivers scientific research data is brought about through the efforts of a small information management team aiming to meet the data needs of field-oriented scientific research. The system supports ship-based oceanographic projects whose participants measure and observe and thereby produce an array of biological, physical, and chemical field data. The report could be considered the story of a software system but it becomes a larger narrative, a suite of interdependent tales when the data, the multiple contexts, and the associated roles are considered. To be understood fully, the DataZoo story must be considered from a variety of perspectives: a tool view provides a technological account of the information system with its incremental growth over three generations; an infrastructure view gives a situated understanding of the configuration and the work involved in creating a multi-faceted information infrastructure that interfaces with existing practices and procedures, and an ecosystem view describes the various phases, components, scopes and sphere's of context with data arrangements. Our intent with this report is to reach out to a diverse set of audiences by including non-technical material along with technical material from various perspectives.

The scope and scale of DataZoo have changed over time but since its inception a broad view has existed of it conceptually as an information system design opportunity and its role in close proximity to the data origin as part of a scientific program's information infrastructure and of the broader information ecosystem in which it is enmeshed. Together these three concepts – an information system, an information infrastructure and an information ecosystem - refer to the mix of software and hardware, of data types and procedures, of people and practices, and of funding and organizational structures that work collectively in the management of data.

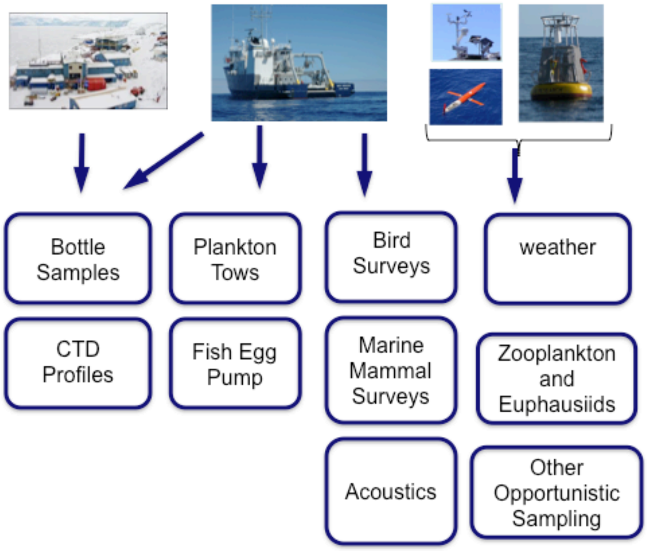


Figure 1.1. From land stations, ships, buoys, and autonomous systems, a variety of data types are collected in oceanographic studies.

DataZoo developed originally to serve one team of researchers, growing over the period of ten years and three generations of system development to serve four long-term projects in addition to a number of related individual efforts. Starting with suites of datasets collected from individual oceanographic cruises and made available digitally to the research team, DataZoo includes after 20 years approximately 200 time-series datasets that are made publically accessible and queriable. There are a variety of data origins, sources originating from land stations and ship cruises as well as moorings and satellites (Figure 1.1). The data types are diverse – from small manual datasets to large automated streams – and the processing often is analysis intensive. Data collection types have grown from highly structured tabular data with metadata to include highly complex data and collections involving a very large number of files. The latter are handled by new applications that together with the core DataZoo system now form a multi-component architecture.

Frequently a number of research groups, each with differing data management arrangements and capabilities, are associated with a scientific project. As an information system and a data repository, DataZoo effectively creates a digital data commons both for research groups and for a variety of projects (Figure 1.2). The existence of the data commons is seen as adding to rather than precluding or replacing interactions between and among individual participants and project-based groups.

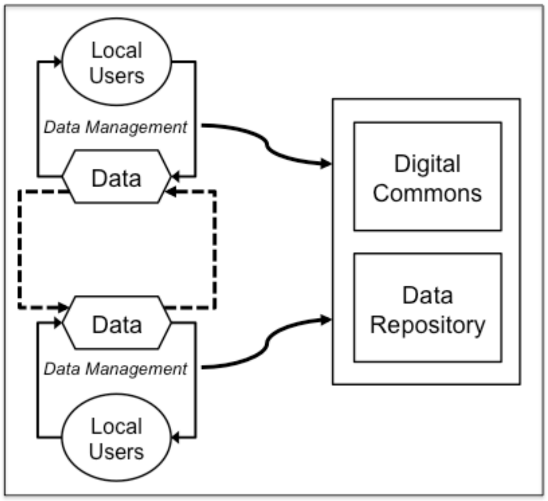


Figure 1.2. Two collections of data flowing into a digital data commons or an information management data repository like DataZoo.

The three successive stages of this information system's development are designated generations within an information system trajectory. Major developments in the DataZoo suite of applications were often marked by a vision of adaptability and sustainability gained from experience in striving to maintain some measures of **simplicity and coherence in the software system while ensuring the flexibility** to meet new requirements. As part of this effort, information management strives to maintain in practice **stability of architecture for developers and users while still accommodating change**. Change may emanate from local circumstances as well as from larger arenas, i.e. community, network, domain, national, and international. New requirements may involve adding features to existing code structures or developing applications to expand system capabilities. In addition, changes may include update by replacement of code or libraries to improve performance and enhancement amidst ongoing development of conventions and standards in multiple arenas.

The DataZoo history illustrates an adaptive approach to growth of a local or site-based information system and a 'fit-for-purpose' philosophy of information. The 'fit-for-purpose' or 'good enough' philosophy of information management recognizes decisions about activities and elements may be made that are appropriate for a particular intended use of the data. We use an iterative, step-wise approach with DataZoo, making it an organic system that grows via interdependent repeating cycles and sub cycles of design, development, testing, deployment, and use. This approach creates a generative understanding of data requirements and practices as well as of user and developer needs.

In the text that follows, Section 2 provides some background. Initial circumstances relating to the DataZoo information system are described, and then, the Ocean Informatics Initiative and the Long-Term Ecological Research (LTER) program are introduced as the contextual framework within which DataZoo develops. A summary of the DataZoo context of use and influential factors is presented in Section 3. The technical architecture is described in Section 4. Section 5 reviews major findings and is followed by final thoughts in Section 6. The appendices provide a variety of materials relating to DataZoo.

2 Background

The type of extended environment within which design and development occur also plays a significant role in shaping the development of an information system. The DataZoo concept and its realization were first made possible by the alignment organizationally of a Long-term Ecological Research site with a site-based, multi-project infrastructure initiative called Ocean Informatics.

2.1 Initial Circumstances

DataZoo is an information system that incorporates a data repository and provides accessibility to scientific data, effectively aggregating, describing, organizing, managing, and serving data in various forms originating with field measurements and observations. Environmental information systems structure the data seen and influence inquiry (Fortun, 2004). DataZoo's design, shaped by the oceanographic research projects it serves, began in 2005. To data consumers, the web interface is the most visible and thereby familiar aspect of DataZoo. To data managers, DataZoo is a set of tools for storing and describing data. To designers, it is an application composed of many interlocking technologies. The initial need by a single project for a shared digital data commons grew first into the idea of an information system and eventually into development of a suite of applications. Over time DataZoo has become a core component of a larger information system of systems.

The information management expertise required to support our current undertaking is organized within the conceptual framework of the Ocean Informatics Initiative (Millerand and Baker, 2011). The Ocean Informatics Initiative is located at Scripps Institution of Oceanography, a directorate of the University of California San Diego. Ocean Informatics began in 2003 as a way of addressing a diversity of data needs. The heterogeneity of data (e.g. CLASS, 2008) and the situated nature of biotic data with specialized requirements became evident early. The notion of being situated is well documented in theory and practice (Suchman, 2002). Two complementary and strategically synergistic groups were brought together purposefully for this initiative: Computational Infrastructure Services (CIS) and Information Management (IM) (Baker et al, 2005; Baker and Wanetick 2010). Initially computational infrastructure for information management was provided by the Institute for Computational Earth System Sciences (ICESS) at University of California Santa Barbara. In 2003 computational infrastructure support was migrated to a local technical support group that began in the Center for Coastal Studies at SIO and eventually expanded to provide services for the Integrative Oceanography Division. Organizational placement occurred some years later in the form of a CIS recharge facility in October 2009. This partnered approach meant the information management team at SIO was able to focus on DataZoo's architecture and work with heterogeneous data while the CIS team took the lead on computational platform administration, network arrangements, security, and change associated with development and maintenance of computational platforms and services.

DataZoo, the application that is the focus of this report, began as the information management component for the Palmer Station Long-Term Ecological Research (PAL LTER) project. Over time, three additional projects – the California Current Ecosystem LTER (CCE LTER), the California Cooperative Fisheries Investigation at Scripps Institution of Oceanography (CalCOFI-SIO) and at the Southwest Fisheries Science Center, a National Ocean and Atmospheric Administration laboratory (CalCOFI-SWFSC) – opted to use DataZoo to meet some of their information management needs. The Palmer LTER study area is located off the west coast of the Antarctic Peninsula while the other projects all study the region off the coast of Southern California.

Table 1.1 Overview Timeline

1949 CalCOFI begins
1990 PAL LTER begins
2003 Ocean Informatics Initiative begins
2004 CCE LTER begins and joins Ocean Informatics
2004 Design concept for DataZoo begins
2005 DataZoo first generation release
2006 CalCOFI joins Ocean Informatics
2007 DataZoo second-generation release
2010 DataZoo third generation release

An overview timeline of this multi-project, multi-agency information management effort (Table 1.1, see also Appendix 8.1) shows the two LTER sites funded by the National Science Foundation (NSF), PAL and CCE beginning in 1990 and 2004, respectively. The two CalCOFI sites funded today largely by National Ocean Atmospheric Administration (NOAA) began in 1949 and began work with Ocean Informatics in 2006. What began as a coalition-building data effort developed into an information infrastructure across multiple projects.

The LTER and the Ocean Informatics contexts have both influenced the development of DataZoo. Summaries of these contexts are given in the next sections followed by information on the conceptual foundations including design, infrastructure and information ecosystems.

2.2 LTER Context

The LTER began as a network in 1980, having been influenced by its predecessor, the International Biological Program (Callahan, 1984; Aronova et al, 2010). It is a site-based network funded by the National Science Foundation (NSF). DataZoo emerged from the work of the information management component associated with the first oceanographic site within a well-established network of LTER sites. This site-level effort exists synergistically alongside domain-level efforts such as that involving the application Metacat (Berkley, 2001), network-level efforts with the LTER Network Information System (Brunt, 1998; Baker et al, 2000; Servilla 2006, 2008), and community-specific efforts by LTER Information Management Committee participants with co-design and co-development in co-ordination with the LTER network office staff (Baker et al, 2006a, b; Kortz 2009, 2011).

2.2.1 Scientific culture

LTER sites as well as the Network itself is constantly evolving, reinventing itself to address new challenges. It has changed organizationally over time from an informal, consensus-governed network of six sites in 1980 to twenty-six sites in 2006 with a more structured governance via by-laws. LTER participants recognized and actively incorporated early on three understudied concepts: long-term research, collaboration, and data management. Each site is expected to plan for the technical and logistical elements required for coordinated field work and synthetic research as well as to address the social and organizational arrangements necessary for carrying out collaborative science over the long term at two levels - site and network.

Sites have in common a holistic understanding inherent to ecological studies (Likens, 1983; Franklin, 1989). For ecological researchers and information managers alike, the LTER network over three decades has been proactive in expanding community awareness of and experience with time-series datasets, regional-level analyses, and cross-site synthesis (Hobbie et al, 2003; Baker and Millerand, 2010). Consequently, traditional interdisciplinary approaches to site-based biome studies develop alongside the collective aim to understand complex systems over time and at multiple levels.

A focus on time-series data foregrounds the concept of “long term” that is recognized and emphasized by inclusion in the name of the LTER program itself (Franklin, 1989; Magnuson, 1990). The LTER provides some examples of strategies for addressing short and long-term needs including one described as ‘continuing design’ (Karasti et al, 2010). Explicit strategies are needed in development of information systems for meeting immediate needs concurrently with long-term needs. Further, the network configuration presents distinct levels of data work or systems to consider: data collection for biome studies by many distinct research groups, data use by the site team within its local organizational setting, and the LTER network pool of data in conjunction with LTER partnerships. Each site studies the local biome. The array of interdependent populations and systems is called the natural ecosystem.

2.2.2 Information management

The LTER network recognized and established the role of data management by requiring each site to identify their data manager explicitly in plans, proposals, and reports. It has become a well-established part of the LTER culture that each site have an information manager who participates in an all-site information management committee (Baker et al, 2000; Baker and Karasti, 2004). This role has developed over time to include mediation work within one arena or intermediation across arenas (Baker and Millerand, 2010). Information management involves balancing of time spent with a multitude of social, organizational, and technical elements involved in making, maintaining, and modifying an information system as part of an information infrastructure.

Early site data practices followed the disciplinary tradition of having data analyzed and stored in laboratories of individual scientists. LTER is unusual in that from the start data management was given priority and instantiated in a concrete manner (Risser and Treworgy, 1986; Gurtz, 1986; Michener, 1986; Gorentz, 1992; Olson et al, 1994). LTER activity at a site often has represented the first formalized data management efforts outside the arrangements within the laboratory of a single investigator. For example, for the first thirty years of the Hubbard Brook research site, it is reported that “from 1955 to 1985 there was little effort to consolidate and archive the masses of data produced.” (Veen et al, 1994). The focus was on the science, on the generation of scientific knowledge at the individual investigator level.

Historically, data management is part of the scientific research process, occurring as part of project planning, data collection, data analysis, and report or paper generation. Data management traditionally focuses on a science-driven, hypothesis-specific activity to be pursued in the short-term. In contrast, the LTER established both a traditional conceptual framework tied to scientific activity but also an administrative framework that identified and funded a role dedicated to data management (Stafford et al., 1994). The melding of these frameworks represents a new approach, in essence an infrastructure plan for data over the long-term. It set the stage for designing and tending to time-series ecological datasets. Work on understanding the concept and ramification of the long-term view of the natural ecosystem suggested considerations of the long-term of information systems. With this in mind, a view developed of DataZoo not as a one-time data management solution but rather as an information system subject to continuing design (Karasti et al, 2010). Given the continuity of effort over time, it has been possible for the Ocean Informatics Initiative to develop effective collaborations within the Information Management Committee as well as the Science Studies communities. There have been various reports, surveys, and working groups providing feedback on DataZoo theory and practice. From an LTER perspective, DataZoo represents an LTER-stimulated response addressing one of the LTER foundational claims that ecological science takes responsibility for managing and making accessible datasets longer than the traditional 1-3 year NSF funding cycle.

At least one information manager from each LTER site is required to participate as a member of the LTER Information Management Committee (IMC). This committee has grown into what has been described as a Community Of Practice (Karasti and Baker, 2004). The Information Management Committee with its annual meeting provides a forum conducive to data sharing and communication about data strategies. This committee represents a professional arena focusing not on technology or science alone but placing emphasis on the integration of science and data needs and their technological support. When Palmer joined the LTER network in 1990, metadata forms and templates were a prominent activity in the information management community. Subsequent LTER IMC endorsement and adoption of the Ecological Metadata Language in 2005 kept metadata at the forefront of community activities. Interaction, both formal and informal, with the IMC has had a significant influence on design and development of DataZoo.

The data management work developed into information management activities and responsibilities. The work of information management has been described as tri-modal, supporting science, stewarding data, and mediating technology (Karasti and Baker, 2004). These arrangements contribute to understanding and growth of a multi-dimensional infrastructure that supports the expansion from sharing of data to data curation, from service support to designing elements of an information ecosystem, and best practices development to participation in both sense making and standard making activities.

The identification of a separate budget at the site-level that is overseen by the information manager creates flexibility in both devising and responding to opportunities. The independence provided by a delineated budget effectively motivates a responsibility that leads to proactivity and stimulates creativity. For PAL and CCE, having funds available from two sites prompted development of an umbrella effort, the Ocean Informatics Initiative, which added a telling impetus toward generality of design despite being a local-level effort. It created what can be described as a built-in, cross-site energetics. Further, the blending of efforts for two sites creates a critical mass of personnel that has enabled an Ocean Informatics team approach with dynamic microsystems. First there is a lead information manager working with developers to co-design administratively and conceptually as well as to implement, test, and assess applications collectively. In addition, there are two developers co-designing at a code level. This buddy system has significant ramifications in terms of code quality and hence robustness of development and sustainability in that the programming is not only done in the head and by the hands of an individual but occurs between the two individuals as well. Shared coding also has social ramifications that mitigate the isolation of human-machine interactions inherent to programming.

2.2.3 Site-based network model

The LTER emphasis on shared categories of scientific study (i.e. biome populations, nutrient movements, site disturbances), concepts (long-term, network, information management, metadata), and activities (working groups, Network Information System, ClimDB, SiteDB, EcoTrends, Unit Registry) has stimulated and informed the data discussions and activities of Ocean Informatics. The work by local investigators with network-level working groups provides opportunities for joint learning as cross-site scientific activities are carried out in conjunction with data sharing and exchange activities. Economists speak of the value of competition between different methods to achieve a common goal. Our experience within the LTER network of sites and its twenty-six approaches to building information systems casts this competition into the form of a cooperative comparative study. The LTER plurality of approaches in practice leads to meaningful dialogue. The LTER configuration sets-up an arena where site similarities create a common ground conducive to respect for differences in design.

The LTER model has had a major influence on the design and development of DataZoo. We currently provide information management services to four long-term sites using the LTER site-based network configuration as a model where Ocean Informatics provides a conceptual umbrella and communication hub for work driven by the needs of participating project participants. The LTER requirement for mandatory funding for information management at each site catalyzed long-term planning for development of an information system. The LTER site-network configuration prompted personnel to design for local situations but simultaneously stimulated awareness of activities and future interfaces at the network, domain, national or international levels. The LTER network of 26 sites represents a unique comparative laboratory for information management and technology that provides both test cases and use cases, some of which provide glimpses into situations that we may face in our future. The LTER network configuration provides an intellectual space for juxtaposing technologies, design work, and emergent scientific data needs while contributing to development of new data practices.

A capacity for multi-scale, multi-perspective planning contributes to construction of sustainable scientific networks. The LTER network provides scientists with funds to pursue site science that in the best of cases synergizes with an individual's own local scientific interests. Yet, they must tend also to network activities, thereby expanding their horizons in unanticipated ways and representing an orientation to larger-scale, interdisciplinary research. LTER participation frequently enhances abilities to ask cross-site or regional questions as well as to imagine new approaches to global issues and new data practices. Changes in scale of scientific questions and instrumentation leads to changes in the scope of data and expectations. In particular, changes include expanding from a focus on sharing site-based scientific insights about joint themes to participating in cross-site synthetic data efforts that require new types of collaboration. And changing from a focus on data systems for aggregation and access of data for local use has expanded to include expectations regarding data curation, synthesis, and exchange.

2.3 Ocean Informatics Context

A series of technical reports summarizes various aspects of Ocean Informatics including the notion of informatics (Baker, 2005), long-term information management (Baker and Karasti, 2005), and the use of a design studio (Donovan and Baker, 2011), and the early history of the initiative (Millerand and Baker, 2011). A series of papers summarizes our circumstances and approaches (Baker and Chandler, 2008; Karasti and Baker, 2004, 2008; Karasti et al, 2004, 2006, 2010).

2.3.1 Information management philosophy

The initial Palmer LTER data management effort established a project digital commons for datasets with associated metadata, thereby addressing the need for shared data use by project participants. The Palmer LTER science components (e.g. seabirds, zooplankton, phytoplankton, bio-optics, microbial ecology, physical oceanography, sea ice, information management and modeling) each has specific data management needs but also must share data with other components. Over a decade, the digital commons prompted a shift from relying solely on an individual investigator's local laboratory data caches to use of a centralized storage arena, a data repository built atop a UNIX-style hierarchical file structure. During this time, investigators and data managers alike gained experience with data sharing while contributing to the shared repository. Such a repository would be considered a '**research repository**' in the three category classification of repository types – research, resource, and reference – reported by the NSB (2005) or 'local' in the idealized repository type categories designated local and remote (Baker and Yarmey, 2008). Repositories and repository types are complex organizational units (Cragin and Shankar, 2006). An early definition of digital repositories identified characteristics such as a variety of data contributors, management of content as well as metadata, minimum services for data handling, and sustainability. Also included in the list are a number of characteristics that are not well defined: trusted, well supported and well managed (Heery and Anderson, 2005):

As LTER information systems and data repositories matured, information management efforts began to be documented in written form by LTER sites (Veen et al, 1994; Brigg et al, 1994; Wasser 1998; Ingersoll, 1997). LTER IMC-organized forums (Benson, 1996; Baker, 1996,1998; Porter et al, 1996) and the development of the LTER Network Information System (Baker et al, 2000) furthered this effort. Eventually in 2006 a book written by one site included a chapter on information management (Benson et al, 2006). Lessons learned and articulated involved the need for simplicity and flexibility as well as for some combination of site-level and network-level activities.

Encounters with the diversity and complexity of information management issues within the LTER began to make evident the need for further expertise in data collecting and curation on topics such as classification, communication, collaboration, and community building. Support by NSF for new types of partnerships led to Science Studies informing the development of Ocean Informatics (Baker et al, 2005; Ribes and Baker, 2007). These partnerships spurred ethnographic study of the work of information managers (Karasti and Baker, 2004; Baker and Karasti, 2004; Karasti et al, 2006; Baker and Millerand, 2010) highlighting topics such as invisible work, infrastructure growth, information systems design, standards development, and data stewardship (Bowker et al, 2010; Millerand and Bowker, 2009; Bowker et al, 2010). From these studies emerged the Ocean Informatics focus on design, infrastructure and information ecosystems.

2.3.2 Enacting the information management philosophy

LTER support for information management is not exorbitant – on average enough for a part-time information manager in 1990 and a full-time information manager in 2010. Given the embedded but minimally funded approach to LTER data management efforts, development is conservative and dominated by a constant influx of requirements and review of needs, an approach that results in a pragmatic, incremental growth that frequently involves acceptance of 'limited fit-to-purpose' or 'good-enough for now' arrangements. It is our hypothesis that this limited but steady funding approach focusing on enabling scientific research is conducive to local learning about data practices, information systems, and information management. The information manager working in close partnership with research scientists provides a unique, situated view – an "eyes up" perspective - on the information management trajectory with its cycles and sub cycles of activity. Close proximity within a scientific research network focused on community-building today tends to encourage a balance, on one hand maintaining the status quo of information management as a technique supporting the immediate use of data while on the other hand developing systems that support information management as part of a long-term strategy. It is the combination of these tasks that frames job requirements for the information management professional (Karasti et al, 2010).

Data workflow begins with planning that may be described as occurring within two arenas: investigator-based and information manager-based. Quality assessment, data collection, and quality control as well as data calibration, processing, and analysis traditionally have been the responsibility of the individual investigator together with their research assistants. Data cleaning, normalization, reformatting, ingestion, curation, preservation, and delivery are responsibilities of an information manager. When located close to the data origin, data irregularities and surprises are the norm so flexibility – with equal measures of accommodation and innovation – is a dominant requirement for managing data. New practices and requirements arise and are identified at this point; data demand local accommodation of change. When there is a confluence of diverse measurements and sampling types, data arrangements and formats have been described as requiring "guidelines rather than rules because some questions can only be answered on a case-by-case basis" (Veen et al, 1994).

2.4 Drawing Upon Conceptual Foundations

There has been a trio of concepts influencing development of DataZoo over the years: design, infrastructure, and information ecosystems. Elaboration of such concepts has fostered a practice-based understanding of information management as well as an awareness of technology-driven forces prevalent in the scientific work arena.

The DataZoo design process borrows from an array of models pertinent to software design (e.g. spiral design, iterative design, rapid prototyping, adaptive design) and information system co-design at the local-level (e.g. participatory design, user-centered design, inquiry-based design, evidence-based design) to name a few. We strive to create a professional environment including arrangements for both development and production work areas and mechanisms in place for moving code between the two. We are site-based rather than a large-scale enterprise. Therein lies both our strength and our weakness, or perhaps more to the point, recognition of our place in the continuum of efforts required in data work that spans multiple levels and locations.

Information systems viewed within their larger context may be seen as elements of infrastructure. On one hand, an ecologist familiar with environmental systems might be ready to conceive of information systems as components of an information ecosystem. Hanseth (2010) has written about such systems as an ensemble of social, technical, and organizational elements cultivated with a collective aim in mind. He builds on the concept of a sociotechnical web of integrated elements (Kling and Scacchi, 1982; Markus and Robey, 1988) where technical aspects are enmeshed within a network of other elements including machines, humans, organizations, and alliances (Orlikowski and Iacono, 2001). This concept of infrastructure crosses boundaries; it has reach, able to tie together datasets, groups, and communities (Pollock and Williams, 2010; Edwards et al, 2007; Hanseth et al, 1996). Within the larger context, network and to community activities with data influence how sites regularize and standardize their efforts with sharing core data and contributing to developer- producer/user-developer interface as well as concept of site-network interface

LTER ecologists study the natural ecosystem; information managers work with a less-recognized ecosystem, the information ecosystem. The various arrangements of and perspectives on the array of interdependent technologies together with their uses and users represent a dynamic system constituting what may be recognized as an "information ecosystem" (Nardi and O'Day, 1999) with the complexities recognized as associated with ecosystems natural and otherwise (Ulanowicz, 1997; Cowan et al, 1994) and with the issues involved with communicating such knowledge outside the fields of study from which they originate (Taylor, 2005). The ecosystem concept is one of those concepts able to telescope and/or to slide their baseline so one must take care to specify the scope of any discussion, i.e. the local ecosystem, the community or domain ecosystem, and the national or international ecosystems. This scope has been dubbed the 'sphere of context' (Baker and Yarmey, 2009)

Section 3 follows and presents the DataZoo context of use. Major technologies, design philosophies, and community needs associated with DataZoo are then discussed in Sections 4.1-4.3 followed by more technical influences and impacts on DataZoo in Section 4.4.

3 Context of Use

DataZoo is, technologically speaking, a tool for the scientific community we support (see Appendix 8.2). As such, the organization and presentation of data in the DataZoo system are strongly influenced by that community's perceptions of how data are organized and used. Some of these perceptions are representative of the larger ecological research community; some are quite specific to our local environment where local data contributors are frequent data users. As a result, there is no single approach to data on which we have based the construction of DataZoo. Within both the local and larger-scale or global communities, we have found multiple valid approaches to data management. Because of this, the design of DataZoo tends towards a balance between options rather than strict dedication to a particular approach to data management.

3.1 Scientific Organization of Data

Organization of data is a central activity in information management, and the primary criteria of many information systems. In this context, organization of data refers to an identification or classification system, structured and implemented so as to enable location of data. Because organization of data is so critical to the purpose of DataZoo, we examined what researchers, specifically ecologists, perceive as useful modes for organizing data. While we were able to identify broad common methods of organizing data, we also found that there are many viable combinations and implementations of these methods.

In practice, the ecological community generally organizes data in three ways. The first is organization by **organization**. In this context, an organization could be as broad as a national research network or as narrow as a single researcher's lab. The organization from which data are generated is important for proper attribution of effort. Organization of data by organization often provides a high level topical grouping – data produced by a particular lab, institution, or research division are often related. The second common method of organizing data is by **study**, or sampling effort. Data that are sampled under similar conditions, whether spatially, temporally, or methodologically, are good candidates for integrated analysis. Organizing data by study preserves this relationship for future analysis. The third method is organization by **parameter**, which includes measurements and observations of phenomena. Depending upon the work arena, parameters may also be known as variables, values, entities, or elements. Data that represent measurements of related phenomena are often used together in analysis. Organizing data by the parameter or parameters measured makes related data more findable, and therefore more useful.

These criteria alone do not create an organizational system. A system of organizing data may implement some, all, or none of the above metrics of organization, and may implement them at varying levels of granularity. For example, organization by parameter can vary from very broad categories such as biological, chemical, and physical data, to very specific categories based on the particular taxonomy, compound, or physical phenomena observed. Some systems require all data to be in exactly one organizational category; other systems will allow data to be labeled with multiple categories at once. Any of these organizational schemes may be implemented hierarchically as well, further increasing the number of permutations of these three basic organizational approaches. The result is a number of possible, practical approaches and implementations.

3.2 Scientific Uses of Data

As with data organization, there are trends in data system usage that are common within the broad scientific community. One common usage is the identification of data system content. **Identification of data** includes browsing, searching, or filtering the superset of available data. Identification of data is a reflection of the data organization, and so identification by organization, study, and parameter are common, although other modes of identification are possible. **Identification via resources** provides a capacity to locate metadata necessary for data interpretation. Typical examples of resources include code dictionaries, methodology manuals, and glossaries. Data systems may allow resources to be identified through related data, through separate browse, search, and filter mechanisms, or both.

Once data and resources are identified, users generally interact further with the data system in two modes. **Data exploration** includes any capability to inspect or analyze data within the data system, such as viewing data, plotting data, or creating data summaries and statistics. **Data access** is the capability to transfer data out of the data system in a format that can be used with other applications, usually as a file that is downloaded to the user's local environment.

The final common usage of data systems we have identified is the **management** of contents. Management includes any capacity to alter the contents of a data system. In many communities there is an expectation that the management use case applies to a different subset of users than the other use cases (i.e. that researchers do not manage the contents of the data system directly). In other cases, the users who interact with the system in the identification, exploration, and access use cases are the same users who manage the data system contents.

3.3 Impact of Local Considerations

In addition to the generalized views on data organization and use cases, each data system is shaped by local perceptions. In the case of DataZoo, providing data system support in a primarily ship-based oceanographic data environment establishes certain considerations for data management. Sampling takes place over many distinct studies, most of which are ship-based cruises. Data are organized by parameter into a single time-series dataset that may span multiple sampling efforts with considerable gaps, so the data and metadata organization must enable retrieval of a dataset as a time-series and also must support identification of cruise-specific subsets. Also, because cruises are generally collective studies by many investigators, with ship time as a shared resource, we often have data from multiple researchers from multiple institutions organized as a single sampling effort. A single study or even a set of parameters often falls under the purview of multiple institutions. Therefore a rigid hierarchy where organization by parameter is subordinate to organization by study, and study is subordinate to organization, is not viable. Instead, we have created a data model in which multiple organization, study, and parameter categories can be applied to each dataset equally (see *Section 4.2*).

These considerations also affect the handling of the use cases in DataZoo. Because of the non-hierarchical data organization, identification of data and resources is enabled from multiple entry points. This allows browsing and searching of data and resources from both the study and the parameter organizational perspectives. For instance, identifying a set of datasets from a single cruise and identifying a single dataset over multiple cruises are both supported use cases. Data exploration and access capabilities are also influenced by local needs. For example, online visualization provides plot types that are commonly used within our community, and data downloads are offered in commonly requested file formats.

3.4 Balancing Tensions

DataZoo is designed with a balance in scope between local and larger-scale or global perceptions of data in mind. Global may refer to a network or discipline of regional, national or international scale. Designing to local perception is essential to creating a data system that meets the needs of site users. Designing to global perceptions provides extensibility as a data system expands to include unanticipated data types and functionality. The tension created between these different perceptions leads to a well-rounded data system but also one that is both immediately useful and viable in the long-term. During the design process, in addition to scope, we have identified several other tensions as have others (e.g. Hanseth and Monteiro 1996). Table 3.1 identifies six of these relating to architecture, user expertise, content management, and system functionality that, when balanced, inform us and help us to make broadly applicable design decisions as the system undergoes continuing design.

Table 3.1

Scope:	Local and Global
Architecture:	Simplicity and Flexibility
Expertise:	Expert User and New User
Content	IM-managed and Researcher-managed
Functionality	In-system Feature and External Application

Simplicity vs. Flexibility: A simple data system requires less time and effort to develop and maintain. Such systems usually consist of a small, self-contained code base that can be quickly designed, prototyped, and released. However, simplified systems are usually designed for a limited range of needs with regards to data types and application elements. Accommodating future needs requires either redesign of the original data system, which can create compatibility concerns, or creation of a separate system, which make data integration more difficult. Thus investing some additional time in the development and maintenance of a more flexible system can be desirable. Flexible systems require a greater understanding of a broad set of use cases. This can be provided via a participatory design model, wherein system users are involved in the design process from the initial stages. Such systems will also have a more complex code base, which requires both more design effort and a longer development period.

Expert Users vs. New Users: Users who are familiar with a data system or the scientific domain it represents often require specialized tools for interacting with data systems. The complexity of these tools can be a barrier to entry for users that are new to the data system and its contents. Designers may try to create a balanced set of tools that is appealing to both types of users. Alternatively, extra time may be invested to create multiple distinct tools and interfaces for different user groups.

IM-managed vs. Researcher-managed: Many data systems are managed by dedicated information managers who can leverage expertise in system design, development, and maintenance. By allowing researchers to manage data system contents directly, scientific domain expertise is added to this list, improving the quality of the data and metadata in the system. However, while many information managers are comfortable interacting with a data system though scripting, database interfaces, and other direct channels, researchers often require more user-friendly management interfaces, which in turn require more developer time and effort to create and maintain.

In-system Features vs. External Applications: A data system's utility is largely determined by the functionality it provides. Features that are integrated with the data system obviate the need for the user to download and possibly reformat data, providing a lower barrier to use. However, adding functionality requires time and effort, and many use cases – analysis, synthesis, visualization, and so on – can be handled with external applications. Designers must decide what functionality to provide within the data system, and what functionality to expect to be handled via external applications.

4 Technical Architecture

4.1 History of Development

To paraphrase an old saying: DataZoo was not built in a day (Appendix 8.1 Timeline). DataZoo has evolved over several iterations during a period of five years (Figure 4.1). Because of the cyclic nature of the iterative design process, major decisions made in the DataZoo design process can be broadly grouped into generations - periods of design and development followed by a period of testing, use, and gathering feedback to inform future design. Though the development of a particular feature may span multiple development iterations and bug fixes and high-priority changes happening in time frames that defy easy categorization, this generational model is an accurate representation of how major design questions as well as our solutions have evolved over time.

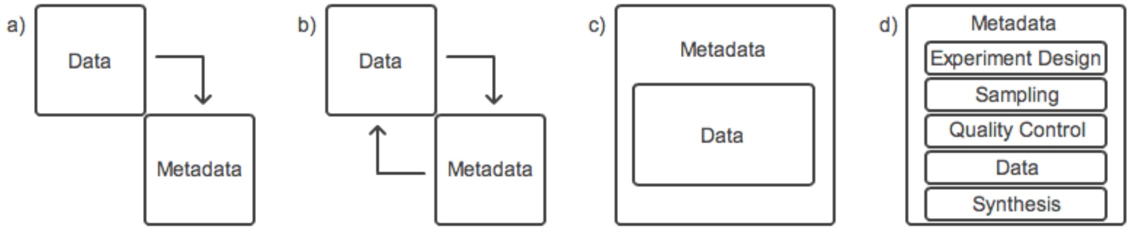


Figure 4.1 Data and metadata across several generations of DataZoo. (a) Prior to the development of DataZoo, the Palmer data system used metadata as a supplement to the data files, defined by the data and generally created after data collection. (b) With the first generation of DataZoo, metadata also provided structural information, defining how the data was stored. The focus remained on the relationship between data and metadata as distinct entities. (c) In the second generation, metadata was recast as the defining not just the data, but also the environment in which data existed. (d) The third generation of DataZoo incorporated metadata documenting the entire data lifecycle, rather than focusing solely on collection and structure.

4.1.1 Information System Prior to DataZoo (Before 2005)

DataZoo has its roots in the Palmer Station LTER data system that was developed over the first 15 years of the site's operation (Baker 1996, 1998). This data system defined the basic data and metadata structures that DataZoo would be built upon, as well as many of the use cases towards which DataZoo would be designed.

The Palmer data system initially consisted of a series of data files organized hierarchically, first by study (cruise or season) and subsequently by sampling (bird census or underway chlorophyll) or high-level data type (bird or biomass datasets). These files were stored on a locally shared disk, providing a simple way for local users to identify and access data files. From early on, influenced by existing approaches at the then sixteen other LTER sites, a practice was adopted of creating a metadata file corresponding in name and location to each data file, in a standardized tagged format. This allowed users to identify easily the metadata resource associated with each data file. Finally, to improve accessibility a web interface to the data system was created. This web interface consisted of a script that would display the results as a catalog of data from cruise and season studies, with each having two types of data representations: the set of datasets associated with a study and one dataset type from all studies. This allowed users to see the full list of available data and metadata files, organized either by study or dataset.

The limitations of this early data system and users' responses to them, helped to inform the features that would be included in DataZoo. Although data and metadata files were discoverable and (at a very simple level) sortable in the data system, they were not searchable. Data files could not be queried or subset dynamically, and there were no tools for data exploration prior to downloading. Furthermore, implementing any of these features was difficult in the original configuration because, although there were standard formats for data and metadata files, these were not enforced for there were no data and metadata models implemented.

4.1.2 First Generation (2005 - 2007)

The main goals of the first release of the DataZoo information system were to improve discoverability of data and metadata and to allow dynamic subsetting of data. We created well-defined and enforceable models for representing data and metadata. Datasets were modeled as tables, with each table containing several parameters and potentially spanning multiple studies and organizations. Metadata was modeled using an entity-relation model that hierarchically arranged organization, study, dataset, and parameter entities in many-to-many relationships. Both of these models were implemented in MySQL relational databases.

This new model allowed us to represent complex relationships between institutions, sampling efforts, and data measurements. Specifically, data could now span multiple studies, which in turn could be supported by multiple organizations. The relational model also allowed us to enforce controlled vocabulary use by linking metadata fields to lookup tables. In the first generation of DataZoo, scientific units, sampling platforms, and keywords were implemented as lookup tables.

Well-defined data and metadata models also allowed us to create new interface elements. In addition to browsing for data by organization, study, or parameter, the metadata model allowed us to implement searching on these elements. Use of dictionaries enhanced searching by providing set lists of terms to query on. The queriable data model allowed us to provide paged previews of data and basic line and scatter plots prior to downloading, thus supporting more data exploration within our system. Finally, structured metadata allowed us to begin exporting information in standardized exchange formats, the first of which was EML, the standard used by the LTER network. Having our metadata in a local relational database meant we would be able to export as needed into alternative formats as well.

The first generation of DataZoo had several limitations, some of which were recognized during the design phase, and others that were not evident until the system had been used for several months. The simplicity of the data model made it difficult to represent very large datasets, due to the lack of optimization available in the database structure. The plotting package we were using, JPGraph, also suffered from size constraints and a limited feature set. The lack of a coherent, user-friendly management system made data updates time consuming; most of our management was done directly through MySQL interfaces. The most significant limitation we found was in our ability to represent metadata at the level of an individual parameter within a dataset. Metadata of this specificity was more common than we had anticipated, and we realized we would eventually need a more detailed substructure at the parameter level to deal with it.

4.1.3 Second Generation (2007 - 2010)

The second generation of DataZoo began with a focus on redesigning the parameter level metadata. We divided parameter metadata into two entities. The attribute entity represents the general classification of the measured variable and contains unit, storage type, and broad subject information. The column entity represents the use of an attribute in a specific dataset, and contains context-specific name and subject information as well as optional qualifiers that can be used to encode specific parameter-in-use information. We eventually formed a list of approximately 30 qualifiers, including sampling and instrument details, history of derived data products, and quality control methods. This allows us to relate parameters between datasets at a broad level (using attributes) while preserving important metadata specific to the dataset (using columns and qualifiers). It represents an extension beyond most existing standards and will likely need to be modified as experience with various classification schemes and semantic approaches matures within scientific communities.

The DataZoo interface also saw many improvements in the second generation. Navigation and search interfaces were completely redesigned based on user feedback, and the JPGraph plotting system was replaced with Matplotlib,. This library allowed us to generate a greater variety of plots with larger datasets. A full suite of management tools was also created, allowing update of metadata and data by users and information managers without direct interaction with the MySQL database. An AJAX architecture was used for both the management tools and the redesigned user interface in order to provide a smooth user experience.

The dictionary tables that were created in the first generation were revised as well. The keywords dictionary was migrated to a controlled vocabulary application with its own search and management interfaces. The contents of this application were expanded beyond keywords to include species codes, quality flags, and other variable-level code lists. These code lists could then be referenced in column qualifiers, allowing data providers to include standardized variable codes and data users to refer back to those codes when interpreting data.

The second generation of DataZoo still had some limitations in place from the first generation, most notably the difficulty storing, retrieving, and displaying very large datasets. Expanding the scope of DataZoo to include more institutions, projects, and laboratories also revealed limitations in our organization and study level metadata, which was originally designed to cover the fairly narrow scope of LTER oceanographic sites and cruises. Our efforts in standardizing variable code dictionaries also revealed several areas in which standardization efforts would best be focused in the future, including unit and attribute-level metadata.

4.1.4 Third Generation (2010 - ongoing)

The second generation of DataZoo was very stable; it was in use by several local research groups and the general public for approximately two years before another major revision was made. During this time we collected feedback from data providers and users and monitored our own use of the system. Having a system that we knew would be stable in the short-term also gave us time to consider longer-term changes. As result, the development of the third generation of DataZoo consisted of more focused and technically involved changes than the second generation.

As the scope of DataZoo continued to expand, we realized that not every organization, study, and dataset could be described by the same metadata fields, and so the design of the metadata for these entities was updated to provide a more flexible metadata model. Using a model similar to column qualifiers, several metadata field templates were created that could be assigned to individual organizations, studies, and datasets dynamically. The dataset and study search interfaces were updated to take advantage of the new structure. The DataZoo management tools were updated to allow data managers to create and edit templates as needed, making it possible to update the metadata schema without changing the MySQL implementation. The use of templates added a great deal of flexibility to the development of metadata and facilitated work on templates by both programmer developers and programmer analysts.

To deal with the issue of very large datasets in DataZoo, an asynchronous data access web service was created. The data access service allowed data queries to be processed in the background, without risking timeouts in the client software. The data access service also standardized the interface and returns for data queries between DataZoo and other databases, allowing us to use previously DataZoo-specific interaction tools such as data previews, plotting, and reformatting with a broader range of data sources. The development of the data access service, originally prompted by the need to solve a client limitation, brought about a fundamental improvement in our approach to dataset access.

The standardization efforts that began with the controlled vocabularies and code lists continued. The controlled vocabulary application was redeveloped as a REST web service and expanded with additional content. The list of scientific units in DataZoo was replaced by an interface to the newly developed LTER Unit Registry (see Appendix 8.7.3; Kortz, 2010; Karasti et al, 2010), a web service enabled database of units used across the LTER network. This contributed to standardization efforts by both making local units available to the network and using units submitted by other sites in DataZoo.

4.2 Application Architecture

The architecture of DataZoo consists of both a conceptual model of how information is represented and a concrete implementation of that model as an application. While the conceptual model provides the basis for the implementation of DataZoo, the technical constraints of implementation also feedback in to the more abstract model. This section covers both of these elements, first outlining the conceptual framework and then providing implementation details.

4.2.1 Conceptual Framework

This section covers details of the conceptual framework used for developing DataZoo in support of data management for a scientific research community. Models covered include those required for metadata and data storage as well as those used for reference in application development and usability assessment.

4.2.1.1 Metadata

Below are the primary entities that make up the DataZoo metadata model (Section 4.1; Figure 4.1). Together, these entities make up the basic set of elements that support a community understanding of how metadata is recorded for both data management and data contribution. Also described are the conceptual relationships between these entities; together with the information inherent to the entities themselves, this creates a complete metadata model.

Organizations

Organizations are the top-level entities providing program or institutional affiliation information for data in the data system, mapping to the concept of organization of data by institution. An organization entity consists of a unique name with optional location, personnel, and policy information. In general, Organizations represent a person or group of people within the data lifecycle.

Each Organization may be associated with multiple Studies. These relationships describe what sampling efforts are sponsored by a given Organization. Organizations may also be associated with each other - Organization entity may be assigned a parent Organization, thus providing a method of structuring institutions hierarchically to reflect associations with the data and other metadata entities at various levels such as funding (e.g. LTER), research program (e.g. CalCOFI) or researcher (e.g. individual laboratory).

Studies

Studies contain data sampling-level information within the metadata model. A study entity contains information about how, when, and where data was gathered, as well as associated personnel. A Study represents an activity or set of related activities in the data lifecycle.

Each Study can be related to multiple Organizations and Datasets. The relationship to Organization indicates the institutional support behind a sampling effort. The relationship between Studies and Datasets establishes sampling origin. A single study frequently comprises multiple Datasets. Like Organizations, Studies also allow for intra-entity parent relations to create hierarchical structures among studies. This allows for multiple data sampling attributions representing the various levels of participation in data collection to be organized under an umbrella study representing a broad collection effort.

Dataset

The Dataset is the primary entity of the data system, providing the basic definition for a table of data records. The Dataset entity provides organization of data at the measurement level. Its structure is composed of a set of fields for providing information such as title, description, general methods and primary organizational ownership. The specific field sets depending on the type of Dataset, but in general Datasets represent information about the source, processing, and management of a variable or group of related variables.

A Dataset is related to Organizations, Studies, and Columns. Each Dataset is associated with exactly one Organization, giving a primary institutional context to each Dataset. A

Dataset also relates to one or more Studies, indicating the sampling activities during which the measurements represented in the Dataset were made. The actual data table structure for a Dataset is defined through its relation to a set of one or more Columns.

Column

A Column is the basic abstraction of each actual column within the data table containing the data records. The Columns in a dataset provide the definitions for the actual physical storage and organization of the data, and thus contain storage type, precision, and ordering information that can be mapped directly to a database, spreadsheet, or other data storage application.

Columns are assigned an Attribute and zero or more sets of Qualifiers. Both of these relationships extend the definition of the column to provide better organization of data by measurement.

Attribute

Each Column must be assigned an Attribute, which provides information that defines the subject and unit of measurement reflected in the data values. This level of abstraction provides a basic level of comparability across datasets.

Qualifiers

As mentioned above, multiple Qualifiers may be assigned to a Column. Qualifiers essentially define re-usable metadata fields that provide a way to document with some granularity methods of data collection and analysis as well as to provide a more detailed definition of the measurement. Qualifiers are conceptually grouped into related metadata fields, allowing groups of Qualifiers to exist as meaningful sets of measurement documentation fields that can be applied to a Column.

4.2.1.2 Data

DataZoo's data model is defined by the Dataset entity, which represents a tabular set of data records over one or more periods of sampling. A single Dataset essentially represents a type of data generally sampled in a consistent manner by the research community, e.g. dissolved iron, containing a related set of measured variables using relatively consistent sampling methods. For the local community served, the majority of data is collected during cruises aboard a research vessel but can also be sampled in various other ways, such as stationary or autonomous instrument assemblages. Sets of records are assigned to their source Study and are appended to the Dataset's tabular representation.

4.2.1.3 Application

DataZoo as an application was designed primarily to meet the needs of data managers, data contributors and data users. Each role within the community requires a set of application features that supports the work of each individual. To meet the needs of data managers, DataZoo provides an interface that supports the creation and management of metadata entities and Dataset documentation as well as uploading data. The data contribution role is supported by the sectioning of data catalogs by research project or organization as well as a flexible underlying attribution model that incorporates data privacy through the ability to enforce visibility rights for registered groups of users. The data user role is broad and involves the identification, exploration, and access of data, which are also needs of the previous two groups. DataZoo was designed with many features to support data discovery and download such as a searchable browsing interface, query and download capabilities as well as exploratory visualization.

At a software level, the DataZoo application architecture is conceptualized as a series of abstracted layers: storage, model, business logic, and interface. The storage layer organizes and persists information, while the model exposes this information in a useable way. The business logic takes the information presented by the model layer and performs operations on it, the results of which are given to the interface layer for display. User input to the interface layer progress back down the layers in a similar manner. These layers interact through well-defined APIs or services, allowing development to progress on a layer-by-layer basis in a modular fashion.

4.2.2 Implementation

This section covers the technical implementation of the conceptual topics presented in the previous section.

4.2.2.1 Metadata

The metadata model for DataZoo is implemented using a relational database (MySQL). Figure 4.2 gives a high level abstraction of the schema for the basic set of entities described earlier.

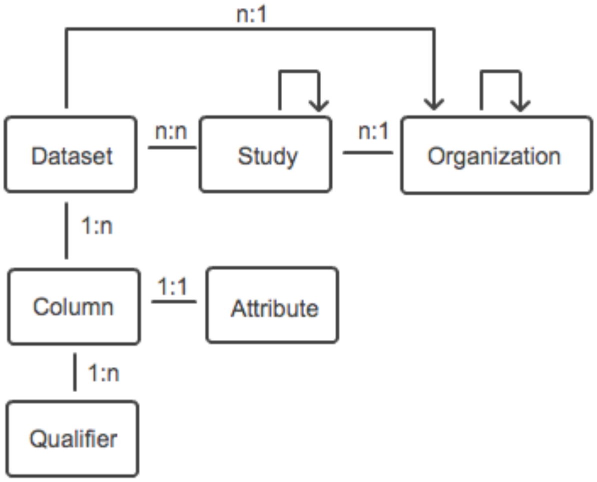


Figure 4.2. Primary metadata entity relations

The relations portrayed support the conceptual relationships for the primary model entities. The two self-referential keys shown for Organizations and Studies provide for parent-child relationships so allow for hierarchical organization within those entities, supporting the complex participatory and supportive roles that groups and institutions play in research field sampling. Another aspect that is that Organizations and Datasets are related in two ways. There is a one-to-one assignment of primary ownership of a Dataset to an Organization. Organizations are also associated with Datasets through their relationships to Studies. Several Organizations may contribute to a sampling expedition in various ways and will all be secondarily related to the sampled or analyzed values in the resulting Dataset. In this way, the intricacies of institution associated with data are supported well.

Metadata templates

Metadata templates are an element of DataZoo's relational model implementation that provides a flexible and extensible framework for defining entity metadata. Templates define sets of fields for entities with each set describing a particular variation of that entity. An example of where the need for this framework appears is in the management of data collection studies. Two studies, for example, an instrument mooring and a research cruise, are essentially the same type of entity (as defined in DataZoo's conceptual model) but require very different sets of metadata fields providing their descriptions. Templates are used in the database implementations of the Organization, Study, Dataset, and Qualifier entities.

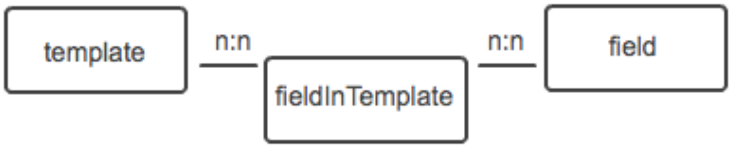


Figure 4.3. Metadata templates storage

All templates are stored in the same sub-schema of DataZoo's relational model as shown above (Figure 4.32). The template table stores entity-specific metadata template records and the fieldInTemplate relational table establishes the set of fields defining that template. Fields in the field table are defined by name and value type and can be reused across templates. For each entity implementing this framework includes a relation to the template table, indicating which template is to be used for each instance of that entity. Each entity using the template framework also has an entity-to-field relational table for template-specific metadata storage.

One of the key benefits of this structure is the flexibility it provides for making changes to entity models without having to update the actual database schema. As the understanding or requirements of a particular entity change, in terms of defining metadata, updates are made by adding or removing fields to defined templates or creating new ones.

Code layer (PHP)

In the model layer, PHP class definitions abstract the primary entities of DataZoo's metadata model and their interactions as well as auxiliary elements and functionality. For all entities that require persistent storage, a design pattern named the Store Model was implemented to abstract database storage and retrieval. When object retrieval and initialization is required through either procedural calls or related object methods, it is done through the desired entity's store object. The same approach is used for storage. The decoupling of object definitions from considerations of persistence provides for more readable code as well as maintainability when storage technologies change over time.

4.2.2.2 Data

Storage

The current data storage implementation for DataZoo is a single database, distinct from that for metadata, containing one table per Dataset. For each Dataset, the structure for the storage table is generated from the information stored in the metadata model. Within the Management interface of DataZoo the option to generate the SQL for building the data table for a Dataset is available, created from the unique IDs and storage type definitions for each Column.

In previous generations of DataZoo, application code abstracted data retrieval and query operations were performed within web server execution time limits. The latest implementation utilizes an asynchronous data access web service, decoupling database interaction from the application code.

Access

Accessing data from DataZoo is done through a web service that handles queries across multiple data storage back-ends. This web service accepts the registration of queries from the interface layer using a standardized query-string syntax and communicates with code modules that abstract individual data back-ends – which may be databases, files, or any other storage mechanism – to produce data results. Client-server communication is facilitated through a set of web service resources that provide for the following:

1. Information about data origin
2. Registration of a query for a particular data source
3. Status update for a registered query
4. Data result access for a finished query
5. Metadata access for a finished query

In whole this provides an interface with which a client can execute data queries and retrieve data and metadata results asynchronously, offloading processing and maintaining user feedback. Below is a sequence diagram (Figure 4.4) illustrating a typical client-server interaction:

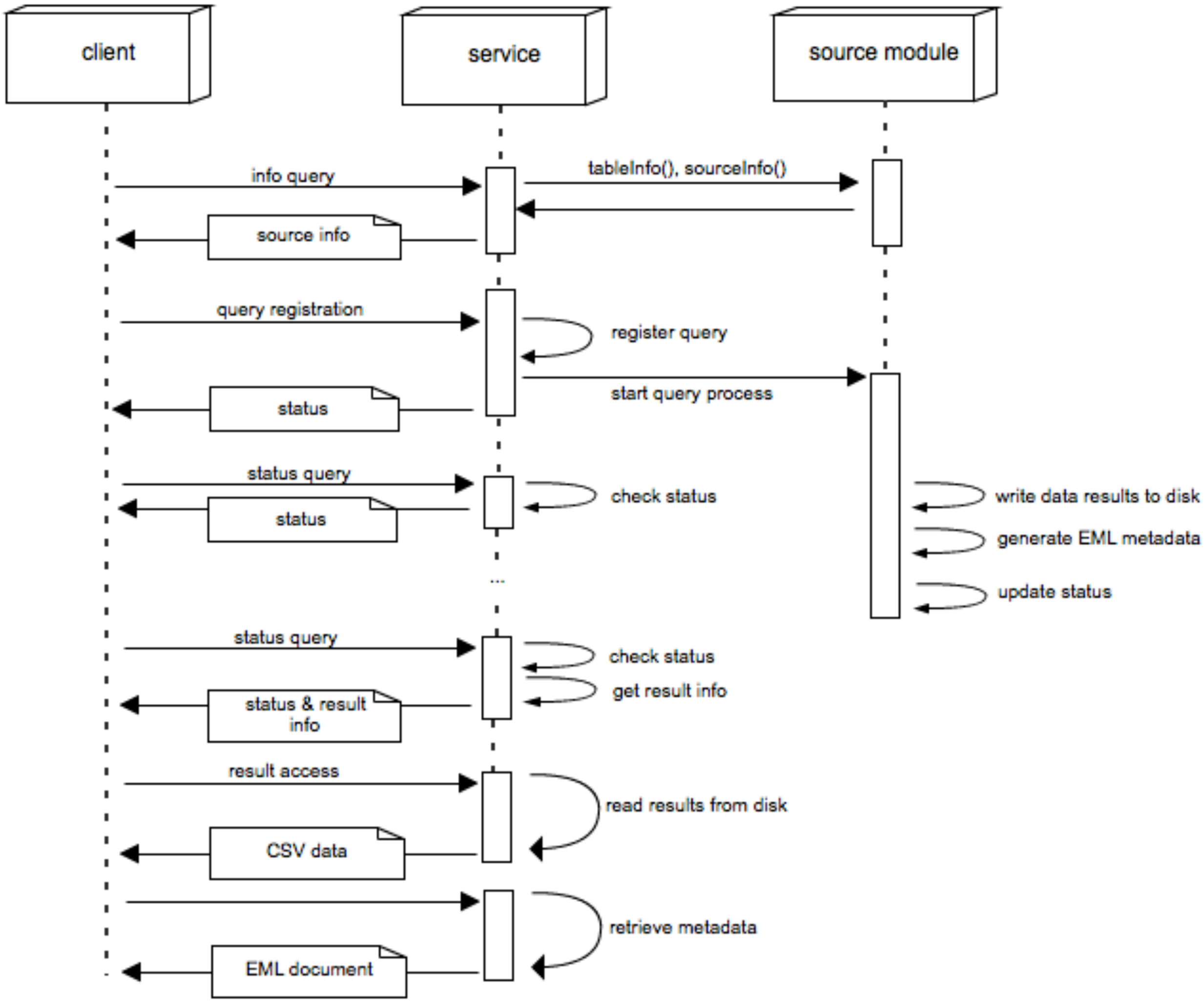


Figure 4.4. Data access web service sequence diagram

As shown, the data access process begins when the client sends an info query. Using the information in the source info response, the client constructs a data query and sends it to the service for registration. The service calls a source module to begin the query process, and at the same time sends a status message, with the status 'processing', back to the client. When the query is complete, the source module writes the results to disk and updates the query status from 'processing' to 'complete'. The client periodically requests status updates, until the service responds with a 'complete' message. When a query has finished the results are accessible as CSV-formatted data. In DataZoo all data exploration interfaces such as plotting, previewing and re-formatting interact with the data result of a finished query.

In addition to data query processing, the data access service also provides access to query-specific metadata, generated by data source modules and formatted as an Ecological Metadata Language (EML) document. As a data system serving two LTER site communities, EML generation for data sets fulfills an LTER network-wide requirement for metadata accessibility for information management components. It also provides a standard for implementation across data sources integrated into the data access web service. EML produced via the data access service is processed using administrative command-line utilities to add system-specific content for LTER contribution. In addition, an archive is maintained, utilizing the Dataset accession numbers (unique identifiers), for storing Dataset metadata and data revisions.

For DataZoo data set access through this web service, the data source module includes the application library described earlier in this section, providing object-oriented access to entity metadata. A query-specific EML document is then constructed procedurally using the PHP DOM extension, designed for creating and interacting with XML documents. This is cached by the web service alongside the CSV-formatted query results and accessed similarly. The combination of CSV data and EML metadata is used to provide several additional features, including data plots, data previews, and alternate download formats such as Excel and NetCDF.

4.2.2.3 Application

DataZoo's application implementation can best be described in terms of its abstracted software layers. At the interface layer, DataZoo was designed as a web site, developed using PHP, HTML and JavaScript. Interface requests are routed by a set of patterns for re-direction, using Apache's *mod_rewrite* module, to PHP scripts that perform business logic operations and create useful displays and interfaces. These operations consist primarily of interactions with DataZoo's model layer, a library of PHP class definitions that abstract data system entities, which persists and retrieves information via the storage layer, a set of PHP class definitions for abstracting persistent data storage and retrieval. Additional operations include authentication logic, which interfaces with the system's LDAP server to authorize users for aspects of the application that require this, such as metadata management and the viewing of non-public data.

4.2.2.4 Integrated Elements

A number of application elements have been integrated into the DataZoo architecture to provide a variety of supportive roles. These elements vary in model, accessibility and implementation as well as their level of integration into DataZoo's architecture. The two primary methods of application integration are via either web service or server-side inclusion of APIs. For those accessed using web services, a caching mechanism was implemented. Caching is done on a per-request basis and stores both the request URL (as an MD5 hash) and the response body, enabling service-agnostic caching across multiple REST compliant web services. This allows DataZoo to function normally even when particular web services are unavailable.

Plotting

Data visualization is done via a web service that provides an interface for creating various 2-dimensional data plots. This application was written in Python in order to utilize a well-supported open source library, matplotlib, for generating these plots. Requests specifying a plot type, various plot attributes and data source are made to the service and a base64-encoded image is returned along with metadata. The data source can be a reference to a result for data retrieval from the data access service described earlier or a CSV-formatted data table included in the request body.

Units

Originally part of DataZoo's relational metadata schema, the unit model used for measurement qualification now exists as a unit registry, an LTER network-wide resource available through its web service interface (Ref). The dictionary within the registry defines units of measurements based on the International System of Units (SI). It also supports a wide range of administrative functions for inter-site unit comparison, documentation and development.

Controlled vocabularies

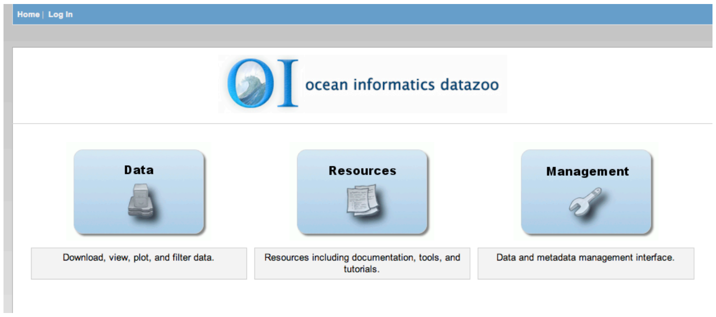
Controlled vocabularies are utilized in DataZoo for storing and accessing metadata that exist as structured lists. Some examples are dictionaries that provide definitions for code values used in data sets and controlled vocabularies used for application interface development. These lists are maintained through a local stand-alone application that provides vocabulary exploration and administration. Controlled vocabularies are integrated into DataZoo through a web service interface.

Study participants

Personnel participation in data collection and analysis is an important component of data set metadata. For data collection Studies, personnel are documented using a stand-alone personnel management application called PeopleZoo. This application provides support for managing personnel and their relations to defined resources. Data collection Studies are defined as resources within this system and personnel records are associated with these. In DataZoo, personnel records are obtained using PeopleZoo's PHP API and are integrated into the application interface.

4.3 Data system interface

DataZoo's interface is divided into three main categories of interest to the user and presented via using three prominent buttons on the front page: *Data*, *Resources* and *Management (Figure 4.5)*. Under each button is a short description of the tools in that section of DataZoo. This division of DataZoo's interface was



Download, view, plot, and filter data.

Resources including documentation, tools, and tutorials.

Data and metadata management interface.

Figure 4.5. The entry into DataZoo showing three interface choices: data, resources, and management.

implemented in the second-generation system in order to provide the user with a clear task orientation based on our primary use cases described earlier. The following narrative describes the features of the data system interface from a typical user's perspective, beginning the "Data" section.

4.3.1 Data

After clicking on the *Data* button a list of data catalogs, one for each Organization supported by the data system, is presented. Within the data system, Datasets are associated with one or more research projects, represented in the metadata model as Organizations. Each Dataset has one primary project and any number of secondary projects. These associations result in the set of project-specific data catalogs shown on this page, with each containing all Datasets associated with the corresponding Organization.

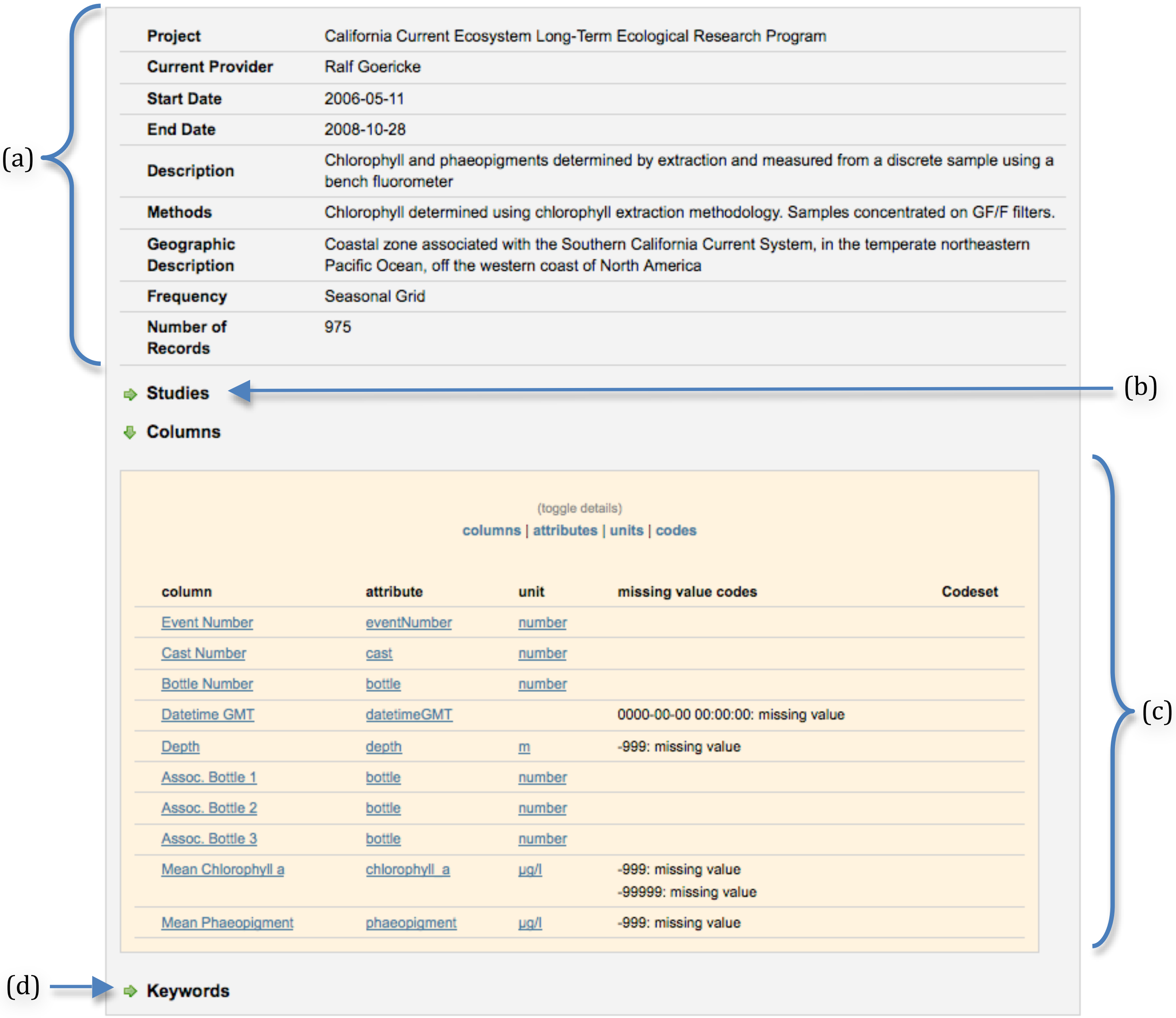
Each project entry has a link to that project's primary web site, a logo or set of logos, a short description and the number of datasets associated with that project. Clicking one of the links brings the user to that project's data catalog. Within a data catalog there are three main sections, as shown in the left-hand navigation menu: "Datasets", "Studies" and "Prepared Views". The first section allows the user to search across dataset titles and primary metadata including measurement fields. The "Studies" section provides for the searching and identification of data via their source(s). The third section, "Prepared Views", will be discussed later. Additional elements of the left-hand navigation menu include a selection dropdown, allowing quick switching between catalogs while maintaining current view. There are also three context-specific elements dependant on the current catalog. These are a link to the primary website and the data use and acknowledgement policies for the current catalog's project.

4.3.1.1 Browsing for Data

Datasets

This section of the catalog allows an individual to browse or search for data by Dataset. It is the default placement when a user has selected a specific catalog and initially shows a listing of all datasets associated with that project. Datasets are assigned to a primary project but will show up in any catalog for which they have secondary associations as well.

For each Dataset in the list, a set of fields is displayed: dataset identifier, title, description, primary Organization, current contact and the number of records. The user can browse this complete list using the displayed fields or can utilize a search form located at the top of the page. The search form consists of a single text input, which is parsed and used to subset the list based on successful matching against a defined set of dataset metadata fields (title, description, owner, and dataset type). Selecting any dataset by clicking on its title brings the user to a summary page.



Project	California Current Ecosystem Long-Term Ecological Research Program
Current Provider	Ralf Goericke
Start Date	2006-05-11
End Date	2008-10-28
Description	Chlorophyll and phaeopigments determined by extraction and measured from a discrete sample using a bench fluorometer
Methods	Chlorophyll determined using chlorophyll extraction methodology. Samples concentrated on GF/F filters.
Geographic Description	Coastal zone associated with the Southern California Current System, in the temperate northeastern Pacific Ocean, off the western coast of North America
Frequency	Seasonal Grid
Number of Records	975

column	attribute	unit	missing value codes	Codeset
Event Number	eventNumber	number		
Cast Number	cast	number		
Bottle Number	bottle	number		
Datetime GMT	datetimeGMT		0000-00-00 00:00:00: missing value	
Depth	depth	m	-999: missing value	
Assoc. Bottle 1	bottle	number		
Assoc. Bottle 2	bottle	number		
Assoc. Bottle 3	bottle	number		
Mean Chlorophyll a	chlorophyll_a	µg/l	-999: missing value -99999: missing value	
Mean Phaeopigment	phaeopigment	µg/l	-999: missing value	

Figure 4.6. Dataset metadata interface

The summary page (Figure 4.6) displays the complete set of metadata fields for that Dataset, defined by its template (**a**). Below this information is a navigable display of additional metadata, initially hidden (except for section title) to the user and made visible by clicking the green arrow interface elements to open new areas. The first section is a list of Studies defined as the data source(s) for this Dataset (**b**). Each title in the list is clickable and navigates to the summary page for that study, which can also be arrived at by browsing the "Studies" section. Next is the list of Columns (**c**), or data fields, comprising the dataset. Opening this section (as it is shown in Figure 4.6) displays the list of Columns along with Attribute, Unit and missing value or code information for each. The list can be explored further by clicking on any one of the following elements:

Column, Attribute, Unit or code set. Doing so for any of the above displays a more complete set of metadata from the dictionaries or registries for each. In addition to the Study and Columns lists, a Dataset may also have a set of associated keywords (**d**) or additional files for download. These are displayed, if available, in similarly formatted lists.

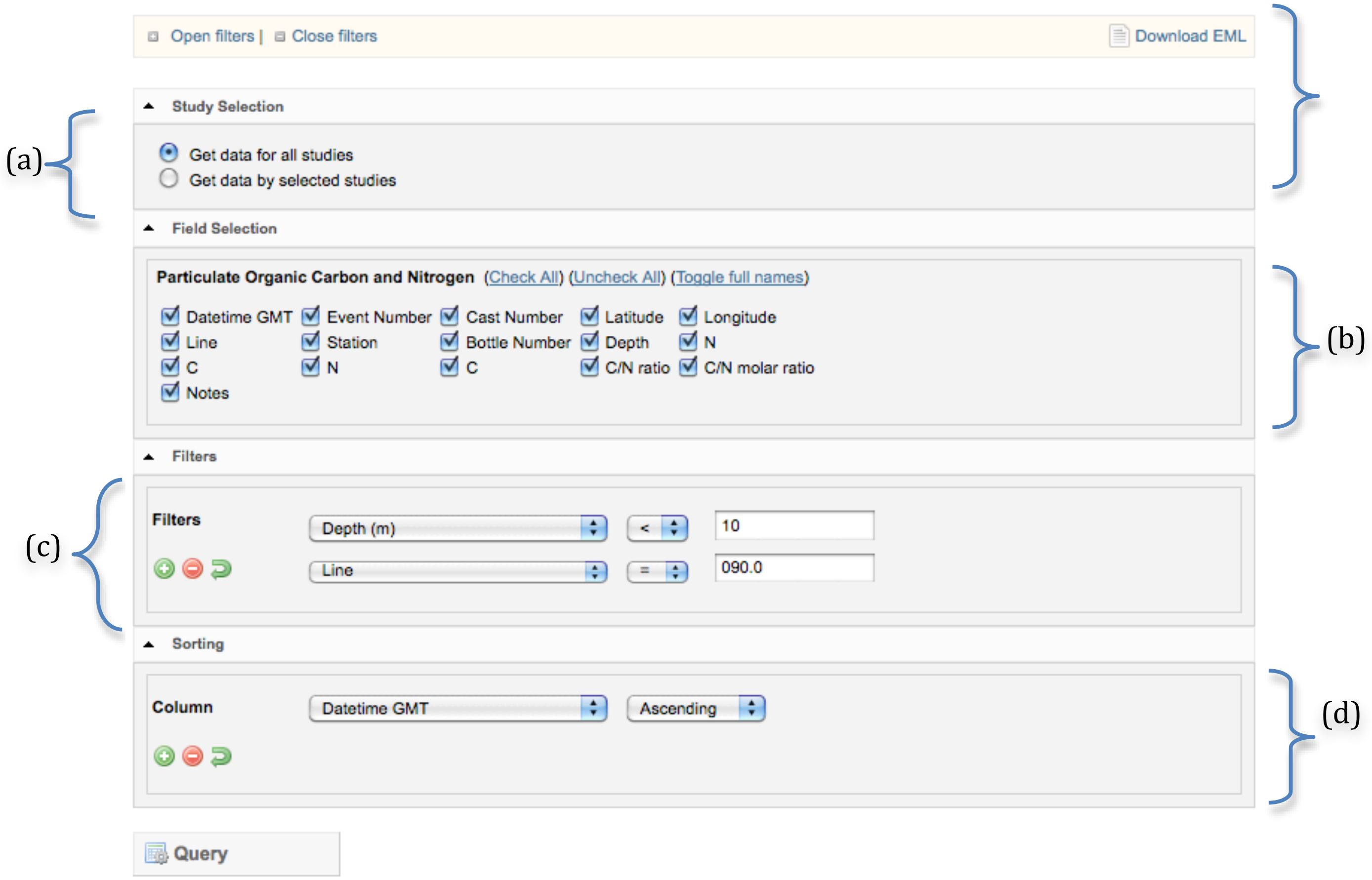
Studies
As mentioned earlier, the "Studies" section of a data catalog allows the user to search for and arrive at data from a sampling perspective; Studies define the sources of data. Navigating to the "Studies" section of a data catalog first presents the user with a list of all top-level Studies, with each entry containing a name and description. Studies can be related hierarchically in order to group data sampling events based on meaningful designation such as a specific research objective. Any Study defined this way will include an arrow that can be clicked on in order to view this hierarchy. Clicking on any individual study's name will bring the user to its summary page.

The summary page for a Study has a similar format to that of a Dataset. All primary metadata defined by a Study's template is displayed first. Next is a display (hierarchical if sub-Studies exist) of related Datasets, meaning that for each Dataset listed this study (or a sub-Study) has been identified as a sampling source. Below this is a list of participating personnel for this Study. Clicking any of the Dataset names above will bring the user to the summary page for that Dataset.

4.3.1.2 Data Access

Data exploration begins at the summary page for a Dataset. Clicking on the *Explore / Download* button above the metadata display brings the user to the interface for querying a Dataset. The user will be prompted to log in as either a **public** or **local** user before proceeding if he or she has not already done so. Once logged in, a user's identity is stored for the rest of the browser session. The public login requires that a name and email be provided for Dataset use tracking. The local user login provides authenticated access to datasets that are not yet ready for public access, as well as to system management tools.

Querying Data
At the top of the data query interface there is a link to download the entire metadata for this dataset as an EML (Ecological Metadata Language) document, the primary metadata exchange format produced by DataZoo. Below this are four query configuration sections (Figure 4.7).



Study Selection

● Get data for all studies
○ Get data by selected studies

Field Selection

Particulate Organic Carbon and Nitrogen (Check All) (Uncheck All) (Toggle full names)

☑ Datetime GMT ☑ Event Number ☑ Cast Number ☑ Latitude ☑ Longitude
☑ Line ☑ Station ☑ Bottle Number ☑ Depth ☑ N
☑ C ☑ N ☑ C ☑ C/N ratio ☑ C/N molar ratio
☑ Notes

Filters

Filters			
	Depth (m)	<	10
	Line	=	090.0

Sorting

Column		
	Datetime GMT	Ascending

Figure 4.7. Data query interface

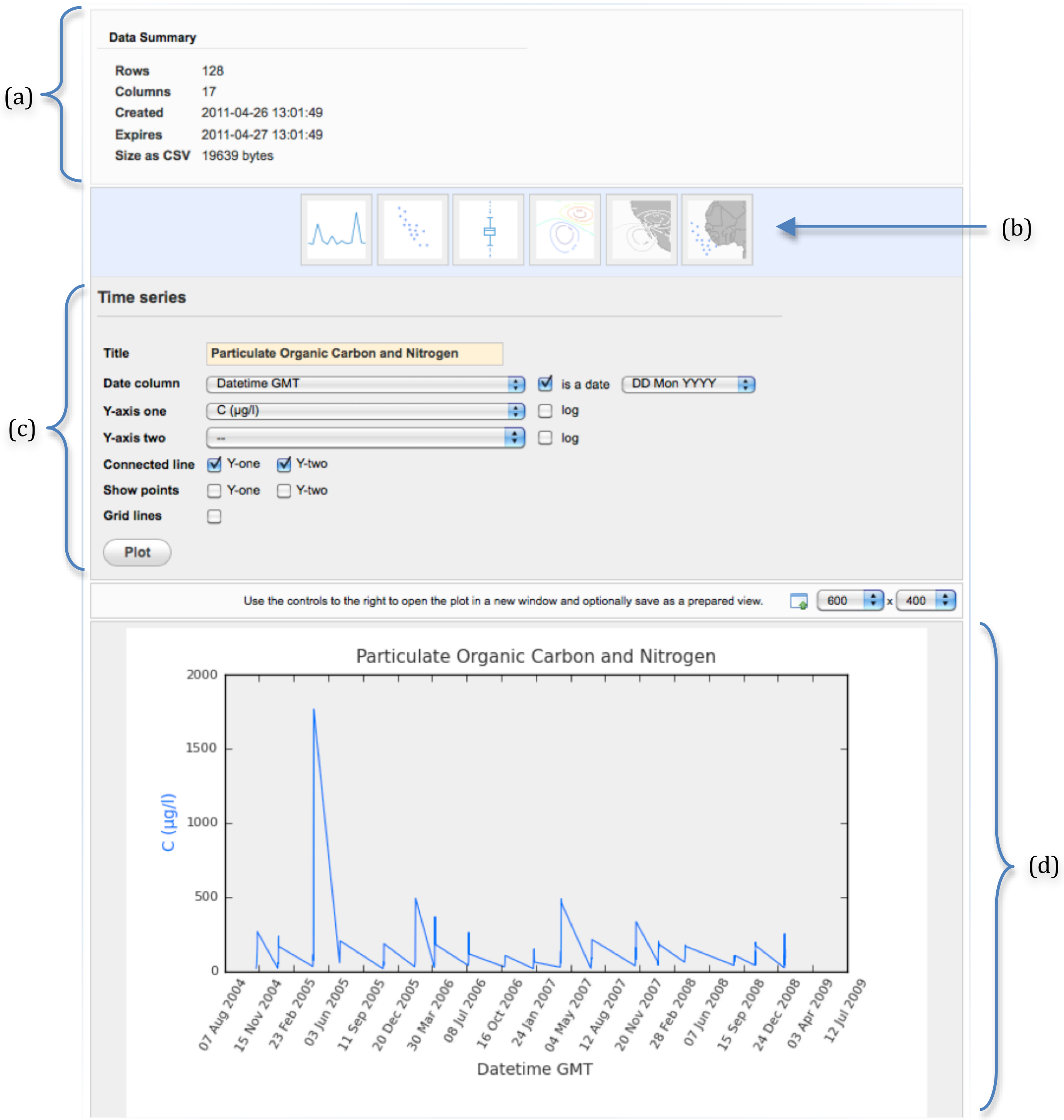
The first is “Study Selection” (**a**). This allows a user to limit the query by sampling sources and provides two options for doing so: users may either get data for all Studies or get data for a selected group of Studies. Clicking the second option then allows the user to select a range of Studies by time period or by individual selection using a multi-select form input. The next section of the query interface allows the user to select which fields (**b**) are actually queried from the Dataset, with the complete set checked by default, defining the header of the produced tabular result. An individual can then proceed to apply any field constraints and ordering parameters using the next two elements of the form. The filtering section (**c**) lets the user set any number of value constraints on a per field basis, with the option of setting multiple constraints for any one field. Each constraint is set by selecting a particular field and a Boolean operator from the set {>, <, =}. By entering a specific operand value in a free text input, a value constraint is set.

There are + and – buttons for adding or removing filters. The entire set of filters is applied as a logical conjunction. In Figure 4.7, the data is filtered for records with a **Depth (m)** of less than 10 meters and with a **Line** value of "090.0". The last option a user can set is a sequence of ordering clauses (**d**) The user can set any number of clauses by selecting a field from a dropdown and a direction (Ascending or Descending), and using the + or – buttons for adding or removing any. These ordering clauses are applied to query sequentially as set by the interface. In Figure 4.7, the data are selected to be sorted by the **Datetime** field in ascending order. A user does have the option skip setting any query options, which would result in the retrieval of the entire data set. After this is done, clicking the *Query* button initiates a request to an asynchronous data access web service for preparation and execution of the query.

When the data result is ready, the interface formats the response from the service into a display showing information about the result, including number of rows, columns and the size of the data as a CSV-formatted text file. Along with this information is a set of options for interacting with the data result. The user can reset parameters and re-run the query any number of times after this point.

Data Interaction

All data interaction is performed using the tabular result of a query. This design feature eases the burden for any subsequent interfaces obviating the need for repeated, process-intensive data queries. When a result has successfully completed, a set of options is displayed for the user to choose from. Each option opens a new browser tab/window with the appropriate interface for the selected option. Thus, the user can at any time return to the results and select a different mode of interaction. A typical first action is to preview the data. Clicking the *Preview results* link navigates to a formatted tabular display of the data. The field headers are interactive and when hovered over using the mouse display a detailed set of additional metadata, similar to what is shown for each column on the summary page for the Dataset. The user can scroll down through the data while the header remains fixed and can page through the data using a navigation menu available at the top and bottom of the page. Returning to the query page, the user can the select to download the data in one of the currently available formats: CSV (comma-separated values), Excel or NetCDF. All formats other than the CSV, which is the natively produced by the data access service, are generated using another asynchronous web service that reformats the CSV data and EML metadata results. The user selects one of these options and when the processing is finished is provided with a link to download the data directly.



Data Summary

Rows	128
Columns	17
Created	2011-04-26 13:01:49
Expires	2011-04-27 13:01:49
Size as CSV	19639 bytes

Time series

Title	Particulate Organic Carbon and Nitrogen		
Date column	Datetime GMT	☑ is a date	DD Mon YYYY
Y-axis one	C (µg/l)	□ log	
Y-axis two	--	□ log	
Connected line	☑ Y-one	☑ Y-two	
Show points	□ Y-one	□ Y-two	
Grid lines	□		

Use the controls to the right to open the plot in a new window and optionally save as a prepared view. 600 x 400

Figure 4.8. Plotting interface

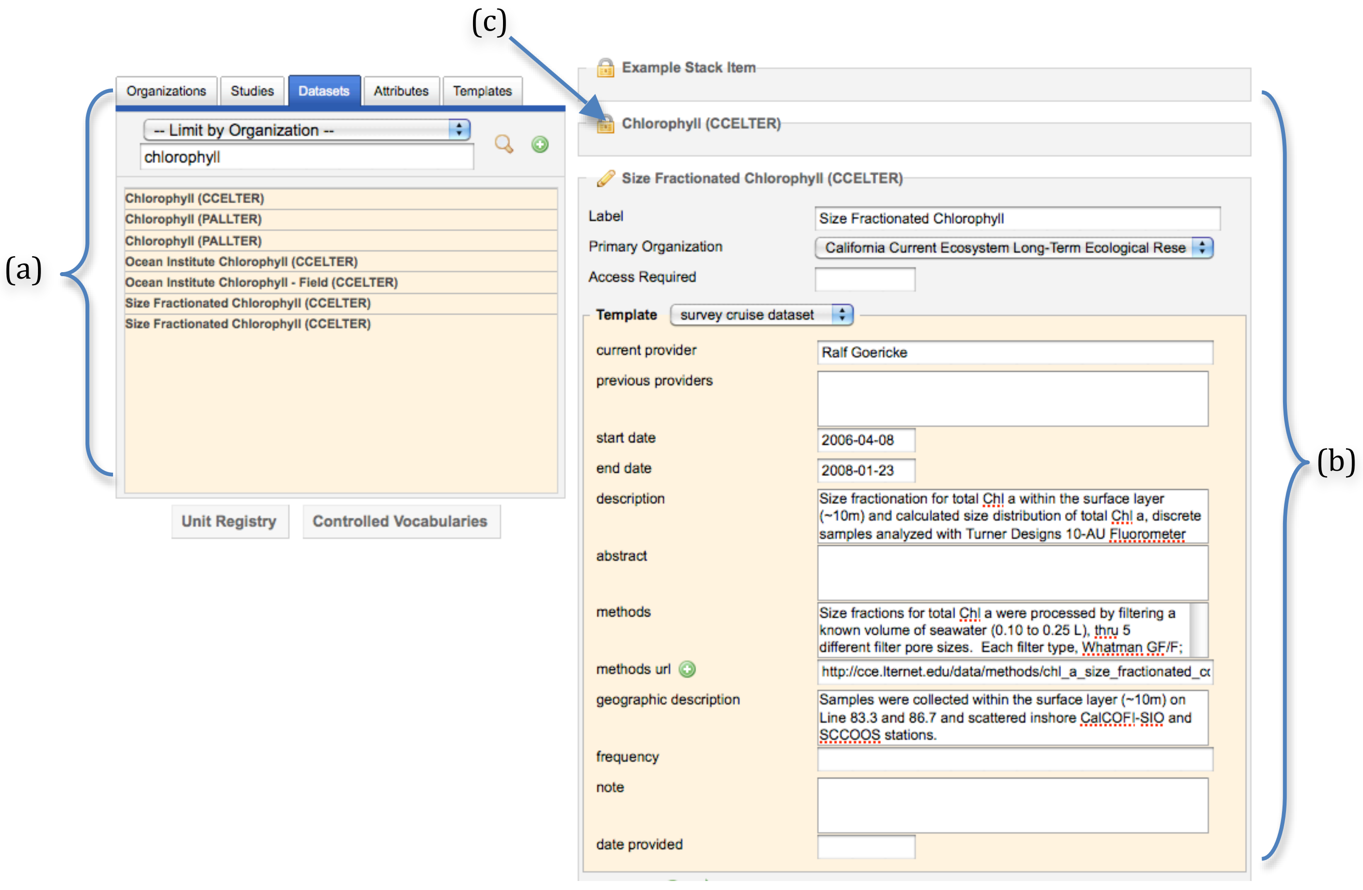
The last option for interaction is plotting, as shown in Figure 4.8. When a user selects the option to plot the data result, a new browser window/tab is opened with the plot interface. A box above the set of plot options **(a)** displays a summary for the data result being explored. Below that is a set of buttons **(b)** for selecting the type of plot to be produced: time series, scatter, box-plot, contour, contour over a map and scatter over a map. Each plot's availability is dependant on the number of data points in the result since certain visualization processes are more computationally intensive than others. Selecting a plot type brings up a form **(c)**, specific to that plot, allowing the user to select which fields are plotting on what axes as well as any additional options. When this is done, the user then clicks the *Plot* button, sending a set of parameters to a plotting web service. This web service applies the set of parameters to the data result and returns an image to the interface, which is displayed **(d)** just below the form used for setting plot parameters. The user can then optionally set new dimensions for the plot and have it opened in a new window for better viewing or saving. In the new window, above the plot image, is a small form that can be used for saving the plot to the "Prepared Views" section of the catalog by entering a title and description and clicking *Save Plot*.

The "Prepared Views" section of the data catalog can be navigated to from the left-hand navigation menu, and presents a complete list of saved plots for the current catalog. Each plot is listed with a title and description. Clicking on the title of any one of these opens a new window, similar to the one described above for the plot interface, with the saved plot reproduced. A prepared view is saved as the set of parameters required for reproducing the plot a user has previously created using the DataZoo interface. Therefore, when a prepared view is accessed the saved parameters are used to re-run the configured query and plot. Thus a prepared view will always reflect the current state of the dataset used to create the plot, incorporating any updates that have been made to the data since creation.

This concludes the overview of the "Data" section and covers all elements of a users interaction with the data in DataZoo. Locating data from two primary orientations, measurements and sampling, i.e. Datasets and Studies, was covered. Also, DataZoo's interfaces that provide for data exploration such as querying, browsing and plotting were described. The next section covers the "Management" section of DataZoo, which provides an interface for data management personnel to administer the metadata and data contents.

4.3.2 Management

From the home page, clicking on the "Management" button brings the user to the section of DataZoo that provides an interface for metadata administration. This section requires the user to be authenticated as a **local** user, meaning an individual with a local system account supported by our computational infrastructure group. The image below (Figure 4.9) provides a typical view of the two-panel management interface in use.



chlorophyll

Chlorophyll (CCELTER)
Chlorophyll (PALLTER)
Chlorophyll (PALLTER)
Ocean Institute Chlorophyll (CCELTER)
Ocean Institute Chlorophyll - Field (CCELTER)
Size Fractionated Chlorophyll (CCELTER)
Size Fractionated Chlorophyll (CCELTER)

Label	Size Fractionated Chlorophyll
Primary Organization	California Current Ecosystem Long-Term Ecological Rese
Access Required	
Template	survey cruise dataset
current provider	Ralf Goericke
previous providers	
start date	2006-04-08
end date	2008-01-23
description	Size fractionation for total Chl a within the surface layer (~10m) and calculated size distribution of total Chl a, discrete samples analyzed with Turner Designs 10-AU Fluorometer
abstract	
methods	Size fractions for total Chl a were processed by filtering a known volume of seawater (0.10 to 0.25 L), thru 5 different filter pore sizes. Each filter type, Whatman GF/F;
methods url	http://cce.lternet.edu/data/methods/chl_a_size_fractionated_cc
geographic description	Samples were collected within the surface layer (~10m) on Line 83.3 and 86.7 and scattered inshore CalCOFI-SIO and SCCOOS stations.
frequency	
note	
date provided	

Figure 4.9. Management interface

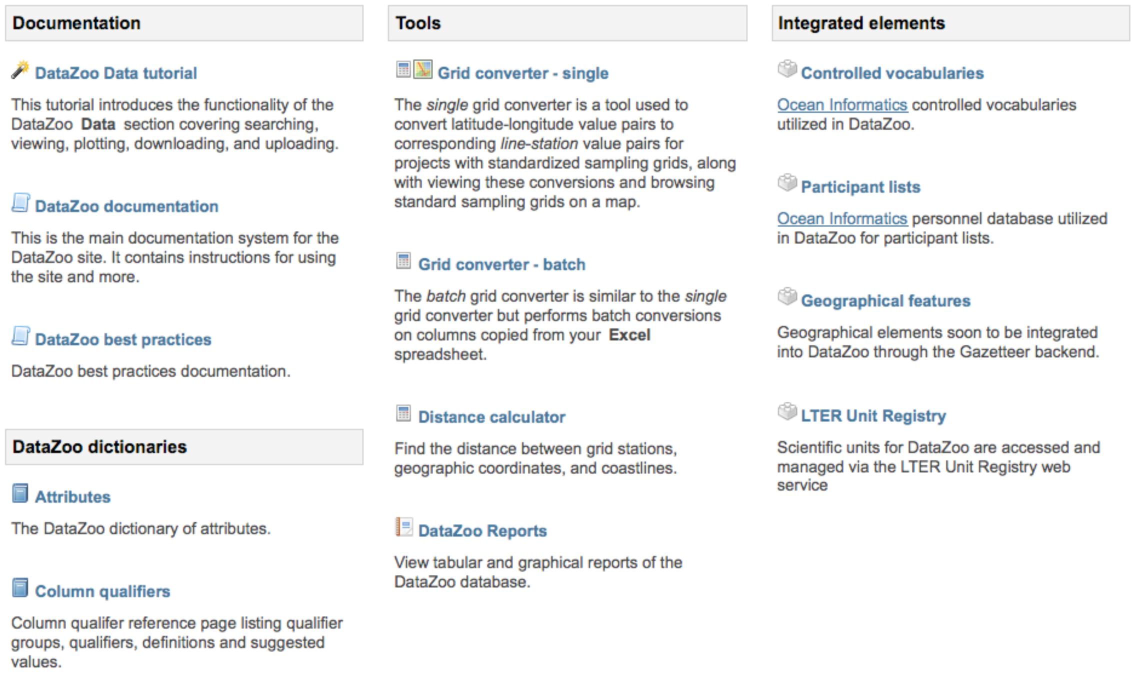
The left-hand navigation panel (**a**) is used to search the primary entities implemented by our metadata model. At the top of the panel are five tabs that can be selected to set the mode for navigation. The first four represent elements described earlier in the description of our conceptual model: Organizations, Studies, Datasets and Attributes. The fifth tab allows the user to search the defined metadata templates, which provide the primary set of fields for a given entity.

Entering a search in the input box in the left-hand navigation panel and clicking on the search icon (magnifying glass) brings up a filtered list of records for that entity; a search done without entering text results in the complete set of records. Clicking on one of the titles presented in the results window brings up an editing form in the right-hand panel (**b**). If a form is currently open and the user clicks on the link for another entity, the current form will be closed and "stacked" above the newly opened one (**c**). This allows a person to traverse data model elements to perform editing as needed while preserving opened forms. An example of where this would be useful is when a user would like to assign an Attribute to a Column in a Dataset, but finds that the desired Attribute does not currently exist in the database. A new form can be opened to create the required Attribute, stacking the current form. When finished the user can close the current form, which then re-opens the last stacked form (in this case, a Column form) and continues editing. Each form provides the ability to edit metadata, as well as entity relationships, for that specific record. Clicking the *Save* button below a form commits all edits to the database, which are then visible in all other areas of DataZoo.

Two historical points of note. First, management functions were initially distributed across the DataZoo interfaces but proved unwieldy both conceptually and in practice. Second, we built the interface expecting technical liaisons with the various research groups to use it but their upload of data once or twice a year was insufficient for them to maintain enough familiarity with upload procedures and with the several interlocking levels of metadata.

4.3.3 Resources

The "Resources" section of DataZoo provides the third and final grouping of interface elements; it is accessible from the home page by clicking on the *Resources* button. This section of the data system organizes a variety of additional resources into four groups: Documentation, DataZoo dictionaries, Tools and Integrated elements (Figure 4.10). These four categories emerged over the period of a year as we gained experience with the similarities and differences among the software artifacts.



Documentation

DataZoo Data tutorial

This tutorial introduces the functionality of the DataZoo **Data** section covering searching, viewing, plotting, downloading, and uploading.

DataZoo documentation

This is the main documentation system for the DataZoo site. It contains instructions for using the site and more.

DataZoo best practices

DataZoo best practices documentation.

DataZoo dictionaries

Attributes

The DataZoo dictionary of attributes.

Column qualifiers

Column qualifer reference page listing qualifier groups, qualifiers, definitions and suggested values.

Tools

Grid converter - single

The *single* grid converter is a tool used to convert latitude-longitude value pairs to corresponding *line-station* value pairs for projects with standardized sampling grids, along with viewing these conversions and browsing standard sampling grids on a map.

Grid converter - batch

The *batch* grid converter is similar to the *single* grid converter but performs batch conversions on columns copied from your **Excel** spreadsheet.

Distance calculator

Find the distance between grid stations, geographic coordinates, and coastlines.

DataZoo Reports

View tabular and graphical reports of the DataZoo database.

Integrated elements

Controlled vocabularies

Ocean Informatics controlled vocabularies utilized in DataZoo.

Participant lists

Ocean Informatics personnel database utilized in DataZoo for participant lists.

Geographical features

Geographical elements soon to be integrated into DataZoo through the Gazetteer backend.

LTER Unit Registry

Scientific units for DataZoo are accessed and managed via the LTER Unit Registry web service

Figure 4.10. DataZoo resources

The "Documentation" group contains reference materials providing detailed descriptions on topics related to DataZoo. The first of these is a tutorial that serves as an instructional guide for using and understanding the various interfaces present in DataZoo, covering much of what has been described within Section 4.3 of this paper. The second resource provides more general overview of DataZoo, including design practices and other useful descriptions. The third is a document that outlines data management best practices for DataZoo administrators, and also serves to make these practices viewable to the public. Each documentation resource is formatted as a set of browse-able wiki-like pages, containing formatted text, images and links to references. All are editable by authenticated local users with administrative privileges.

Below this section is a set of links to "DataZoo dictionaries". Each of these provides browse and search interfaces to various metadata dictionaries. A dictionary is defined as being a set of re-usable metadata entities that serve to provide a common semantic framework within DataZoo's conceptual model (Section 4.2). These lists provide a reference for administrators, users and supported community members.

The next group of resources, "Tools", contains a set of interfaces that provide functionality related to DataZoo or the supported local research communities. The first two links bring the user to the "Grid Converter" interface for converting between the standard latitude-longitude and project-specific coordinate systems, and vice versa. The first provides an interface for converting individual coordinate pairs as well as visualizing the points on a map, while the second does not support mapping but allows batch conversions for many coordinate pairs. The next tool is the "Distance Calculator" that supports distance calculations between points across all supported coordinate systems, as well as point-to-shoreline distances. The last tool, "DataZoo Reports", provides a set of small widget-like displays for specific views of the DataZoo metadata backend, such as a chart for visualizating dataset distribution among projects. This section serves to corral informative displays of the database that can be re-configured frequently without affecting the consistency of the rest of the data system interfaces.

The last group of resources is the "Integrated Elements" section, which groups references to non-core elements of DataZoo that have been integrated into the data system through an available API. Some of these are local applications existing in the broader information system in which DataZoo is situated, while some are outside resources. They have been integrated into the data system architecture in order to provide support for specific metadata elements like controlled vocabularies, personnel records and units of measurement.

4.3.4 Summary

The DataZoo application interface is divided into three primary sections: Data, Management and Resources. Each of these provides an orientation for user interaction based on task. The "Data" section is designed to support data and metadata browsing as well as data access and interaction. The "Management" section provides the primary interface for metadata database administration. The "Resources" section organizes data system and research community supporting interfaces as well as documentation and references.

4.4 Influences and Impacts

The DataZoo data system is the product of several years of effort, changes in technology, many iterations of design and development, and several influencing factors. The technologies upon which any data system is constructed, and the capabilities and constraints inherent to them, play a major role in shaping that system. The design philosophies of the Ocean Informatics group are another important factor in the ongoing development of DataZoo. Finally, the features and capabilities of DataZoo are molded by the constantly evolving, emergent needs of the scientific communities it supports.

4.4.1 Major Technologies

Several technical decisions have been made during the development of DataZoo. While some of these decisions have had a fairly small impact, and others were reversed or superseded during the development process, many have had a profound effect on the design and usage of the data system.

One of the most essential decisions was to build DataZoo as a web-based system. As a web-based system, DataZoo is broadly accessible – it requires no specialized software, is not platform dependent, and can be accessed from any computer capable of browsing the web. However, the limitations of a web framework can be seen in almost every aspect of the system. Interface elements are limited to what can be processed in a web browser, a limitation felt particularly in the data output and visualization interfaces. The stateless nature of HTTP transactions, as well as the relatively short timeouts imposed by some browsers, have both required special attention from developers to overcome. The wide accessibility of DataZoo has also created issues of data security and user privacy that were addressed during the design process.

The decision to store both data and metadata in a relational database has also influenced much of DataZoo's design. The general entity-relationship model we use for thinking about our metadata is derived largely from classic RDB schema models. The relational database imposes a level of consistency on the data model that in turn allows rapid querying and transformation of the elements it stores. Many of DataZoo's features, such as the search options and the ability to subset datasets, are designed to take advantage of this model. However, this same consistency makes dealing with special cases more difficult, as they must conform to the data model. The templated metadata system developed for DataZoo is a direct result of this dichotomy in that it provides flexibility for special cases, but encodes that flexibility in such a way that requires only the content, not the structure, of the database to be updated when making changes.

A recent decision that has impacted our design process is the move to a REST web service-oriented architecture. This decision has improved the integration of system elements such as user interfaces, data and metadata engines, visualization tools, and controlled vocabularies. Integration takes place within DataZoo as well as between DataZoo and other data systems, both local and global. This integration is achieved by using accessible, stable, and well-described web services as abstraction layers between system elements. At the same time, the development of future tools is constrained by the goal of remaining within a REST architecture. Also, increased reliance on external web services, while saving time and effort, can require adjustments to local development plans to maintain compatibility.

Other technologies that have influenced the design and development of DataZoo include the software stack upon which it is built. DataZoo currently uses Linux, Apache, and PHP as a web development stack, with Perl and Python scripting used for specialized services. The decision to use these tools came about due to our interest in open-source software, and in turn these tools influence our own design with the capabilities they do, or do not, provide. However, the use of open-source software is perhaps more remarkable for the lack of influence it has had on our work; almost universally, we have found that the capabilities of open-source software match those of their commercial counterparts within the scope of our development needs.

4.4.2 Design Philosophies

Our design philosophies, like DataZoo itself, have changed over time as we experiment, improvise, and learn what approaches meet our particular needs. Much like the technologies we employ, design practices are tools that enable us to work efficiently, but unlike technologies, there is rarely an industry standard to which one can look for guidance. Design philosophies influence the work we do, but our work also pushes back, shaping and improving our theoretical knowledge.

Participatory design forms one of the cornerstones of our design process (Schuler and Namioka, 1993; Blomberg et al., 1993). Participatory design involves engaging potential users of products earlier on in the design process, and keeping them engaged and informed, through community meetings, design sessions, testing activities, surveys, and written reports, throughout the entire design and development life cycle. For DataZoo, we actively recruited users, including both data providers and data consumers, to take part in this process. The most direct impact of this approach is that our design process is informed by very specific feedback at all stages, and subsequently less dependent on generalized use cases representing anonymous users. This has shaped DataZoo into a system that is strongly adapted for local users. The participatory design philosophy also affects our development schedule, as users are not always available for consultation, and participation can become a rate-limiting factor. To mitigate this constraint, we engage many user participants, not all of whom will necessarily give feedback at every stage of a project.

Other design philosophies we have embraced are those with rapid, iterative design and development with releases that introduce applications to the community in phases. This includes continuing design (Dittrich et al, 2002; Karasti et al, 2010) and Design in Use (Henderson and Kyng; 1991). This allows us to deploy applications that are usable, albeit with incomplete feature sets. This practice enables participatory design, because users can interact with an application while development is ongoing, and feedback and improvement can happen in tightly coupled cycles. However, the fragmentation of the release schedules also means developers spend more time on temporary and transient interfaces and features, and less on major system redesigns, which do not fit well into a rapid release schedule. Users may also find themselves working with incomplete applications, expecting features that are not present, or even put into the role of testers. To mitigate these problems, developers must commit even more time to responding to user requests.

One design philosophy that emerged over the course of DataZoo's development is the move away from monolithic generalized applications in favor of many applications with specialized roles. While DataZoo is quite broad, encompassing over 150 datasets, the original design was even more comprehensive. However, creating a generalized application conflicted with our other design philosophies – we had trouble incorporating very specific user requests into the feature set, and the complex code base made rapid releases difficult. Earlier on DataZoo's development, we migrated complex, specialized interfaces out to separate applications. With the incorporation of web services into our designs, we also began separating controlled vocabulary management, a specialized task that was integrated into DataZoo, into separate applications as well. While this approach requires us to balance design and development efforts among several applications, it also allows us to stagger release schedules and development-feedback cycles, enabling our other philosophies to be realized.

4.4.3 Community Needs

Both technology and design theory shape the design of DataZoo, but it is the needs of the communities DataZoo serves that drive that design forward. The DataZoo metadata structure has been consistently informed by the needs of our users, both those who contribute data and those who extract and use data. Each design decision is made with respect to a need, either present or anticipated, from the users, managers, funders, and developers of DataZoo. Because of our commitment to participatory design, we receive a great deal of direct input from users on features they would like to see in DataZoo – too much to completely enumerate here. A few needs, ones that are ubiquitous among our many users, may be identified as having had a particularly strong influence on the continuing design of DataZoo, and have elicited features that have become central to the data system.

Because DataZoo is a data repository close to the source of the data, the ability to preserve that information is of paramount importance, and drives many design decisions. Many of the unique features of our metadata model, such as the templated field structure and the column qualifiers, are direct expressions of our efforts to preserve metadata with high fidelity. Currently, the influence of metadata representation needs is also notable for how much it shifts over time. New datasets, new protocols for existing datasets, and new uses for data that elicit previously overlooked information all have contributed to the ever-changing nature of our metadata needs. Because of this, our metadata model has seen a high degree of redesign over the last six years.

In addition to handling metadata, DataZoo is ultimately an application for information exchange, that is, it is responsible for the exchange of the data themselves. The challenge of preserving information accurately is further complicated by the sheer size of some of the datasets. These concerns have resulted in two important design decisions. First, it has prompted recognition of the limits of DataZoo, and stimulated development of a multi-component system with multiple data systems for information that is simply too large, complex, or irregular to store in DataZoo (Appendix 8.3). Second, is has caused us to rethink the traditional synchronous data transfer model used in many web-based data systems. The need for a suite of query, visualization, and formatting tools that could handle millions of points of data led to the development of the data access layer that currently handles all DataZoo data transactions.

One area in which DataZoo has been influenced by a diversity, rather than a consensus, of user needs is the data interaction tools – specifically, visualization and data formatting. These tools define the boundary between DataZoo as a data access system and DataZoo as a data analysis system. Because of the many possible ways data can be analyzed, analysis tools tend to address very specific needs. When DataZoo was first released, these tools were minimal, allowing only one type of download (CSV file), two dictionaries (unit and attribute), and two types of visualization (scatter and line plots) with very little customization. User feedback has consistently called for more file formats, plot types, and user-definable options when interacting with DataZoo. DataZoo currently supports six plot types, three file formats, and dozens of subsetting and manipulation options. However, such tools are not only specialized, catering to a subset of users, they are also often very time intensive to develop. To minimize future development times, we have design generic data and metadata interaction subsystems, such as the metadata store model and the data access service, to improve the scalability and extensibility of our suite of tools.

5 Discussion

The continuing design, development and use of DataZoo is more than just an effort to create a data system – it is also a process that provides valuable insight into how these tasks will evolve in the future. The complementary philosophies of 'research-as-learning' and 'design-as-research' create an environment in which each task is subject to analysis both from a practical perspective ("How can we improve our data system?") and a theoretical perspective ("What does this tell us about design practices in general?"). This section addresses some of the major findings, both practical and theoretical, that have emerged from the process of creating DataZoo, and the analysis of that process.

5.1 Organizational Situation and Perspectives on Data

One of the most important lessons we learned from our work with DataZoo is that there are many valid philosophies and approaches to data management. Often a defining factor in the development of a working data management plan is the data managers' placement within the path of data travel, that is, the dataflow within the web of data repositories

(Figure 5.1). A data manager working directly with data producers and consumers will perceive and address challenges much differently than a data manager working with a national archive who is dealing with thousands of researchers across hundreds of institutions. We try to maintain the perspective that there is no 'best' plan for data management – just a 'best fit' for the environment and purpose at hand. Further, we recognize the value in planning for a multiplicity of locations for any single dataset or collection of datasets with the requisite data provenance.

The location of Ocean Informatics team is close to the point of data origin. As a result, we work directly with the scientists and technicians who perform experiments and generate data on and after research cruises. This placement shapes our philosophy and implementation of many data management tasks. For example metadata generation, quality control, and data system design are handled as a continuing dialogue between data producers and data managers. The capacity to have these dialogues led us to adopt and enact a philosophy of participatory design.

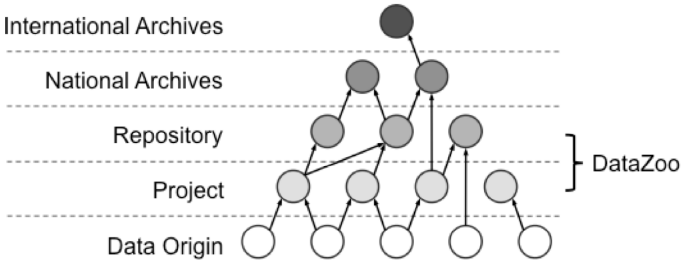


Figure 5.1. A 'web of repositories' comprised of multiple entities and their relations. Data systems that are close to the data origin are often more specialized because they deal with a smaller scope. National and international archives are usually very generalized because of the variety and volume of data they handle. DataZoo as a multi-project information system, has features of both a project-level data system and a data repository.

Because we work closely with a small number of researchers, we are very sensitive to the needs of individual labs and even individual datasets. We have found that data that may appear homogenous from a broader, more distant view is in fact very heterogeneous when viewed more closely. The concept of differing 'spheres-of-context' at each granularity or level of work has been suggested to account for the shifts in perspective (Baker and Yarmey, 2009). This in turn has led to the philosophy of developing many, fit-to-purpose applications that address case-specific needs while supporting broader-level system integration through APIs rather than attempting, from the start, a single solution for all data needs. Ocean Informatics is able to implement this philosophy in practice because of the relatively small number of researchers and labs we support. Our organizational placement as a local information management group also allows us to maintain a rapid iterative development and release schedule.

A data management group for a larger, national archive data management group may find it difficult to implement these kinds of philosophies and, due to a different placement and perspective on the data, such philosophies might not be the 'best fit'. Global-scale projects often have an expectation that the data being submitted is already processed to an archive-ready state, so they are not as involved in the quality control and metadata annotation process as a local data management group. An archive dealing with tens of thousands of datasets cannot afford to treat each one individually, and may take a homogenized view of the data, focusing more on similarities than differences. Such a philosophy is appropriate for a national archive, but would not be a good fit for the level at which Ocean Informatics operates.

5.2 Change and Flexibility

Another important finding that emerges from reflection upon the DataZoo design process is the need for flexibility. When we began work on DataZoo, we attempted to create data and metadata models that could describe all of the datasets we managed. What we found was that creating a data and metadata structure that was both universal and complete was an unrealistic task. With each cruise, new and unexpected data types were added that required more and more extensions to the existing schema. At the same time, existing datasets had changes to methods and measurements that required us to continually revise existing metadata. Over the course of the past five years, the DataZoo data and metadata models have moved towards greater flexibility, including the addition of column qualifiers and templates to the metadata schema, introduction of data access services for data queries, and the creation of alternative metadata resources such as methods manuals. Our code base has also evolved to be more flexible, relying more on abstracted APIs and services and less on closed, application-specific implementations.

What we learned through this process is that flexibility at the application and code levels is essential for our philosophy of emphasizing the uniqueness of each dataset. We also learned that flexibility is equally important at a more philosophical level. Design requires an element of foresight to create products that will be useful beyond the short-term, but foresight is never perfect. Having a rigid design philosophy creates stability, but can also lead to premature obsolescence in a rapidly changing field such as scientific data management. To avoid this, Ocean Informatics has adopted an organic growth approach, allowing both our data system and our practices to evolve as the needs of the community change. Maintaining this philosophical flexibility does come at a cost in the short-term, as we often find ourselves exploring several possible solutions to a challenge before settling on an approach. However, this approach has shown over time that it leads to an invaluable breadth of experience that serves as an evaluation and renewal process. This type of prototyping process leads, in turn, to more informed practices and robust systems in the long-term.

In general, what we have found is that scientific practices are rapidly changing, and data management practices must change in step to accommodate them, at least in the localized, close-to-source environment in which Ocean Informatics operates. Committing to a specific data management philosophy would require a level of predictive power that we have learned to admit we do not possess – rather, we have chosen to embrace the fact that the future holds unexpected changes, and approach each new challenge must be approached as an opportunity to revisit and revise our practices.

5.3 Specialization and Simplification

A trend that emerges from the analysis of DataZoo's evolution is the continual specialization and simplification of our data system elements. As noted in previous sections, over time the Ocean Informatics design philosophy has moved away from monolithic, manifold applications to a proliferation of simpler and more specialized applications. The move to this type of architecture was not a decision made at the beginning of the design process, but rather a discovery we made by analyzing the strengths and weaknesses of our system as it evolved. Though DataZoo appears on the surface to be a single massive application, it is in fact a suite of tools – searching, querying, visualization, management, and more – that are maintained separately but are presented together through a common interface.

The first element of this finding is the need for specialization. Because Ocean Informatics is situated so close to the point of data origin, we are able to – and expected to – be aware of very specific differences in the datasets we manage and the needs of the communities we support. Our initial approach was to create applications that were highly configurable, but we encountered two limiting factors. First, the number of different configurations would have led to an interface that was difficult to use because of its complexity. Secondly, the users did not want a high configurable interface; they wanted highly specialized, pre-configured interfaces. Effectively, users were asking for a data system with scientific expertise 'built in' – only different users wanted different built-in expertise. The end result was a compromise. DataZoo, with its configurable metadata model and data query options, handles the majority of our datasets. Datasets that require specialized interfaces, whether due to complexity, size, or specific use cases, are handled outside of DataZoo in other elements of the Ocean Informatics system (Appendix 8.3).

The specialization of our data system elements has, through necessity, led to the simplification of these same elements. As the number of application, services, and libraries we maintain increases, the time we have to dedicate to each one of them decreases. However, the burden placed on each element also decreases, allowing us to streamline and simplify the structure, and thus the maintenance, of each element. A major task in simplifying our diverse code base was to identify and normalize common elements into targeted, well-delineated APIs or web services. This allows us to support several applications for the overhead of maintaining a single system element. To support this approach, we've also had to adjust our design philosophy to be more aware of the reusability of certain system elements, and prioritizing the design and development of those elements. We are also more aware of expanding feature sets in existing applications and the point at which creating a separate, specialized tool becomes a better-fitting solution.

5.4 Developing Policies and Publishing Data

Data policy becomes especially important in plans that include data sharing. Different groups have different ideas about data policies, and all have opinions as to what needs to be included and what needs to be emphasized. Consequently, each project in DataZoo or in other applications has the option of including a locally crafted data use and data acknowledgement policy (Appendix 8.4). We experienced a breakthrough in helping groups with data sharing when the notions of 'data publishing' and 'data production' were used to extend discussions of data sharing. Such conversations seemed to open up thinking relating to data preparation and workflow as well as data ownership and accreditation.

Appendix 8.7.1 shows an interesting use of the data publishing concept where a customized interface for a specialized application was developed for a data collection with complex relations. While some domain specialists are interested in its configurable interface that presents options for making selections from a biological data collection, other researchers and managers are not. Over time, the idea of publishing a basic subset of regularly sampled data for targeted uses developed. Thus if there was a time-series sampled most consistently in Spring seasons and irregularly sampled the rest of the year, then a Spring subset was produced for ingestion into DataZoo. This results in easy access to well-described, easily-interpreted data sets made publically available through a project catalog.

Over time as participants became more at ease with the concept of data sharing, concerns and experiences with misinterpretation of data became evident. The development of derived subsets of data published into DataZoo helped with this issue. In addition, mechanisms for emphasizing data methods emerged. First, a mechanism for highlighting methods was implemented where the data could be accessed only after the methods page that covered a grayed out main dataset page was closed (see Appendix 8.4.2 and 8.4.3). A second mechanism for avoiding data misinterpretation involved presentation of data measurements made using different methods. From a data management perspective, these could be presented most efficiently as a single time-series file with a column containing a flag denoting the different methods. In order to emphasize the different methods use, however, the data contributors requested that the data be presented as four separate files that had to be downloaded separately. This meant someone who downloaded the data would have to take action to put them together and thereby be given the opportunity to think about why they were presented as separate files in the first place.

5.5 Developing our Practices

Referring with insight to the 'the mangle of practice', Pickering (1995) nudges us to find new ways – metaphorically, conceptually, and practically - of thinking about data practices given the number of emergent activities and situations. Metaphorically, in working with DataZoo we have found the use of an organic metaphor of 'growth' rather than a construction metaphor of 'building' to be an important reminder about and spur to sustainable design. Conceptually, though we have only begun to explore them, we were fortunate to begin our work with three powerful concepts that facilitate work with information systems: design, infrastructure and information ecosystems. Together these three accord complexity and interdependence their due. Practically, a healthy respect for the complexity of data associated with living systems – its organization and management, its processing and flow, its compatibility and delivery, its integration and synthesis whether the data are biological, social, or organizational – is present because our everyday work is in close association with field-based ecologists who experience with respect and wonder the complexity of the systems we strive to represent digitally so they may be better understood. This respect informs approaches, effects attitudes, and empowers design.

5.6 Considering Future Design Issues

The design of DataZoo is continually informed and influenced by changing research and data practices in the scientific communities we support. Further, as we plan for future design, our experience with various data perspectives and data publishing, change and flexibility, specialization and simplification as well as development of our own practices as described above informs and guides our work.

Data integration: One particular feature we are planning to re-visit in the near future is that of integration across distinct datasets sharing one or more sampling dimensions such as time or location. Our first implementation of this feature was embedded within the tightly coupled object and storage models of the DataZoo code base and has been put aside temporarily during re-development. Since that time, we've re-factored much of that library into modular components within the architecture such as the data access service. We've also improved specific metadata model elements that influence the design of this feature, such as dataset column documentation that includes storage type, unit of measurement and definition. When re-implementation of this feature within our data system occurs, it will benefit from both our experiences in design and an architecture that is much better staged to support a sustainable design.

Documentation: Our methods for documentation of both DataZoo and broader information management topics have evolved over time. On one hand for DataZoo, various models were implemented during different development cycles: in-line text, field-structured documentation referenced by topic name and, finally, our current documentation that relies on simple database storage and retrieval of wiki-syntax articles. The last implementation has allowed us to integrate documentation seamlessly within the varying shape of our data system, with a minimalistic interface. This model proved to be the most amenable to content updates and maintenance. Less settled is our implementation for broader-topic documentation. The lessons learned during our iterative development of application-specific documentation will be used in the future to strike a balance with ease of use and maintainability in a manner similar to the case with DataZoo. On the other hand, our documentation of information management has been pursued proactively through publication of interesting or milestone topics in the LTER newsletter Databits (http://databits.ucsd.edu), technical reports, and papers in conference proceedings and journals. Taking time for such writing fosters reflection and articulation as well as synthesis and knowledge-generation just as is needed in any scientific field if there is to be lasting knowledge-building.

Quality control: Given our focus on growth of the system itself and the priority given to ingestion of data for active research programs, we have had to make clear to date that the majority of quality control is the responsibility of the data contributors. A variety of general data checks exist from simple to complex. Initial simple checks would involve defined ranges, relations between values, and sequencing.

Metadata forms: An interesting change in our understanding of metadata occurred after the development of a set of metadata forms was completed and used. We intended that these forms would be used by a computer-savvy technician from the various research components but found that the infrequent use of the forms (i.e. once or twice a year) resulted in having to relearn how to use the forms each time. In addition, we found that any small changes in DataZoo metadata represented major barriers to use of the forms. These forms are now used by Ocean Informatics developers and analysts who update the metadata based on a variety of materials submitted by participants. We plan to readdress this issue in the near future.

Derived Datasets: Our understanding and the vocabulary associated with 'derived datasets' is nascent and expected to unfold over time. Indeed, steps in what is a continuum of work with datasets are difficult to identify and sometimes are misleading. We have today first a confusion of terms to work with – raw, field, calibrated or processed and clean, analyzed, or derived. Second, we have subcategories of derived datasets that have yet to be explicated except in large-scale projects with homogeneous datasets such as NASA satellite data and its well-defined level 1-4 derived products.

Websites: It is frequently a project web site that introduces a project and provides an integrated view of project-related materials. Materials and data may be delivered from static files or dynamically from databases with an eye to cost-effective simplicity. Ocean Informatics has designed a three-tier web template for projects that use our data systems and that require a web site. We have investigated use of content management systems (CMS) and currently are considering partnering to migrate to a Drupal CMS given potential partnering with the SIO in-house web group as well as with the LTER network community and central office. This promises to introduce enough complexity that the option has been under consideration for a number of years.

Organizational Placement: We have envisioned for some time the need for two types of digital infrastructure at the organizational level that support contemporary research efforts: a computational infrastructure service as described in Section 2 and integrative information management. The former has been established as a subscription-for-service recharge facility at SIO whereas the other is planned as an on-request recharge facility.

The nature of our work is that as we co-design and co-develop with participants, not only are new requirements identified but also new possibilities emerge. Indeed, effective planning and communication are required as design possibilities have a way of becoming perceived as pressing requirements. We have at hand a number of possibilities - funded and unfunded - for expanding DataZoo. One is to support a type of linked spatial/temporal visualization and another is to consider the relations between code dictionaries, methods manuals, and metadata at the study, dataset, attribute, and unit levels. In having used the need for multiple media galleries as opportunities for exploring different design approaches, we look forward to using the experience gained to generalize and unify gallery design. In addition, with network web services in place that establish a precedent for site-network exchange, it sets the stage for integrating personnel and bibliographic modules into DataZoo locally. And finally, there are those intriguing possibilities requiring a semantic stretch from dictionaries and standards to compatibility and integrative activities.

6 Final Thoughts

The LTER-initiated efforts at SIO have resulted in growth and enactment of local information management and in particular of **the information system DataZoo**. DataZoo is given its name in recognition of the diverse data and our aim to contain it. In containing data, it is isolated from the field environment. The name DataZoo is a reminder that care is required with how data are represented and their context documented.

DataZoo as an adaptive system is a model of ongoing enhancement and evolution. With DataZoo, we aim for an information management approach that reaches beyond the expectations that sites at the local-level, close to the data origin, meet minimum network requirements, community best practices, and review criteria. Rather as active participants we engage in activities that contribute to the notions of continuing design, working standards, and federated webs of repositories. These are elements needed to address the complexity of tasks involved in managing digital data to be used to represent natural systems. Our design-oriented approach is one way of ensuring that on-the-ground insights into the data are actively integrated into the processes developing for handling data across multiple levels and timeframes.

As DataZoo developed, **design** became recognized as critical, a sweeping category referring to design of features, systems, and communities involving ‘design as service’, ‘design as research’ and ‘design as learning’. Infrastructure became recognized as an under-appreciated category for capturing the interconnectedness of digital configurations intertwined with social and organizational arrangements, a mix maintained by adding, combining and readjusting as well as reevaluating, refactoring and redesigning. Documentation and discussion have been emphasized in Ocean Informatics efforts as a way of prompting reflection upon the conceptual and practical choices that must be made in creating information systems. Articulation enables synthesis and mindfulness to the assumptions and constraints incorporated. Attention to these elements that are soon taken for granted, may help preclude issues relating to misunderstanding or misinterpretation of data so we can continue to take delight in the order and access to data created by an information system.

Given the current state of data coordination - the lack of common practices across field arenas for naming units, attributes, geolocations and for sharing methods of collection, observation, analysis, and reporting - a great deal of time is needed for joint activities and development of mechanisms that facilitate identification, comparison, and discussion of vocabulary and procedures as well as categorization and classification as substrates to digital semantic approaches. But there are difficulties in planning support for information management given an innate impatience with the high cost of time-consuming detailed work that must be considered along with the technologically-driven promises of short-term solutions and the field-based science equipped with complex new support-intensive instrumentation.

Any philosophy or approach is affected by the state of **information management readiness** of all involved. Readiness takes different forms for participants and stakeholders. For instance, data users experienced with data sharing quickly become aware of and ready for data policy and data query while those working with data aggregated from multiple sources have an increased sensitivity to and expectations for community formats and standards. Information management readiness is influenced significantly by the extent of understanding of the concept of long-term. This concept is critical to understanding and handling ecosystems of all types

This story of DataZoo provides not only an example of meeting the data needs of a research site and the communities within which it is enmeshed but also provides an example of a trajectory of '**continuing learning'**. Rather than attending either discipline-specific classes or management-organized training, DataZoo enabled learning within a working scientific environment where change was fostered not only in systems, applications and code but also in people, practices, and the relations among them. In the course of DataZoo development programmers gained insight into design, analysts gained insight into information management, research participants gained insight into new types of scientific data practices, and all gained real-world experience with the ramifications of data sharing and data production. Over time, this approach not only bridges the technically and the scientifically focused but also raises the baseline of awareness or understanding of the local culture regarding design, infrastructure, and information management.

7 References

Aronova, E, KS Baker, and N Oreskes, 2010. From the International Geophysical Year to the International Biological Program: Big Science and Big Data in Biology, 1957-present. Historical Studies in the Natural Sciences 40(2):183-224.

Baker, KS, 2005. Informatics and the Environmental Sciences. SIO Technical Report Series. http://escholarship.org/uc/item/0179n650

Baker, KS, BJ Benson, DL Henshaw, D Blodgett, JH Porter, SG Stafford, 2000. Evolution of a Multisite Network Information System: The LTER Information Management Paradigm. Bioscience 50 (11): 963-978.

Baker, KS and CLChandler, 2008. Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics. Deep Sea Research II 55(18-19):2132-2142. Special Series Palmer Long-Term Ecological Research.
http://dx.doi.org/10.1016/j.dsr2.2008.05.009

Baker, KS and HKarasti, 2004. The Long-Term Information Management Trajectory: Working to Support Data, Science and Technology. SIO Technical Report Series.

Baker, KS, SJ Jackson, and JR Wanetick, 2005. Strategies Supporting Heterogeneous Data and Interdisciplinary Collaboration: Towards an Ocean Informatics Environment in Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS) 2005, 3-6 January, Big Island, Hawaii, pp. 1-10, IEEE, New Brunswick, NJ

Baker, K, D Pennington, and J Porter, 2006a. Multiple Approaches to Semantic Issues: Vocabularies, Dictionaries and Ontologies. LTER Databits Newsletter, Feature Article, Spring 2006.
http://databits.lternet.edu/issues/137#145

Baker, K, LYarmey, S Haber, F Millerand, and M Servilla, 2006b. Creating Information Infrastructure through Community Dictionary Processes LTER Databits Newsletter, Feature Article, Spring 2006.
http://databits.lternet.edu/spring-2010/sio-ocean-informatics-update-growing-infrastructure-support-scientific-research

Baker KS and J Wanetick, 2010. SIO Ocean Informatics Update: Growing Infrastructure in Support of Scientific Research. LTER Databits Newsletter, Feature Article, Spring 2010. http://databits.lternet.edu/issues/115#99

Baker, KS and L Yarmey, 2009. Data Stewardship: Environmental Data Curation and a Web-of-Repositories. International Journal of Digital Curation 4(2):12-27.

Baker, KS and F Millerand, 2010. Infrastructuring Ecology: Challenges in Achieving Data Sharing. In Collaboration in the New Life Sciences. J.Parker, N.Vermeulen, and B.Penders (eds), Ashgate, Surrey, England: p. 111-138.

Baker KS. 1996. Development of Palmer Long-Term Ecological Research Information Management. Pages 725-730. Proceedings of Eco-Informa Workshop, Global Networks for Environmental Information, 4-7 November 1996, Lake Buena Vista, FL. Ann Arbor, MI: Environmental Research Institute of Michigan (ERIM).

Baker, KS 1998. Palmer LTER information management. In Michener W, Porter J, Stafford S, Eds. Data and Information Management in the Ecological Sciences: A Resource Guide. Albuquerque (NM): University of New Mexico, p105-110.

Benson, BJ, PC Hanson, JW Chipman and CJ Bowser, 2006. Breaking the Data Barrier: Research Facilitation through Information Management. In Long-Term Dynamics of Lakes in the Landscape, JJ Magnuson, TK Kratz, BJ Benson (eds). Oxford University Press, Oxford, pp 259-279.

Benson, B, 1996. The North Temperate Lakes Research Information Management System. Pages 719-724. Proceedings of Eco-Informa Workshop, Global Networks for Environmental Information, 4-7 November '96, Lake Buena Vista, FL. Ann Arbor, MI: Environmental Research Institute of Michigan (ERIM).

Berkley, C, MB Jones, J Bojilova, and D Higgins. 2001. Metacat: a Schema-Independent XML Database System. 13th International Conference on Scientific and Statistical Database Management. IEEE Computer Society.

Blomberg, J, J Giacomi, A Mosher, and P Swenton-Wall. 1993. Ethnographic field methods and their relation to design. Pages 123-155 in D. Schuler and A. Namioka, eds. *Participatory design: principles and practices*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Bowker, GC, KS Baker, F Millerand, and D Ribes, 2010. Towards Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In Jeremy Hunsinger, Lisbeth Klastrup and Matthew Allen (eds), International Handbook of Internet Research (New York: Springer): p. 97-117.

Briggs JM, H Su, 1994. Development and refinement of the Konza Prairie LTER research information management program. Pages 87-100 in Michener WK, Brunt JW, Stafford SG, eds. Environmental Information Management and Analysis: Ecosystem to Global Scales.

Brunt JW. 1998. The LTER network information system: A framework for ecological information management. Pages 435–440 in Aguirre-Bravo C, Franco CR, eds. North American Science Symposium: Toward a Unified Framework for

Inventorying and Monitoring Forest Ecosystem Resources; 2–6 Nov 1998; Guadalajara, Mexico. Fort Collins (CO): US Department of Agriculture, Forest Service, Rocky Mountain Research Station. Proceedings RMRS-P-12.

Callahan JT, 1984. Long-term ecological research. BioScience 34: 363-367

CLASS, 2008. Comprehensive Large Array-data Stewardship System (CLASS); Information Heterogeneity White Paper, U.S.Department of Commerce, NOAA.

Conners, J, 2011. Notes on Design. Commentary Article, Databits Newsletter. Spring 2011. http://databits.lternet.edu/spring-2011/notes-design

Cowan, G, D Pines, and D Meltzer, 1994. Complex Ecological Systems. In Complexity: Metaphors, Models, and Reality. SFI Studies in the Sciences of Complexity, Proc. Vol XIX, Addison-Wesley, Reading (MA).

Cragin, MH and K Shankar, 2006. Scientific Data Collections and Distributed Collective Practice. Computer Supported Cooperative Work 15:185-204.

Dittrich, Y, S Eriksén, and C Hansson, 2002. PD in the Wild: Evolving Practices of Design in Use, in *Proceedings of PDC '02* (Malmö Sweden), CPSR, Palo Alto, 124-134.

Donovan, JM and KS Baker, 2011. The Shape of Information Management: Fostering Collaboration across Data, Science, and Technology in a Design Studio. SIO Technical Report.

Edwards, PN, SJ Jackson, GC Bowker, and CP Knobel, 2007. Understanding Infrastructure: Dynamics, Tensions, and Design. Ann Arbor. Retrieved November 17, 2008, from http://hdl.handle.net/2027.42/49353

Fortun, K. 2004. Environmental information systems as appropriate technology. Design Issues 20(3). Franklin, JF, 1989. Importance and justification of long-term studies in ecology. In GE Likens (Ed.) Long-Term Studies in Ecology. Approaches and Alternative. Springer-Verlag, New York, 3-19.

Franklin, M, A Halevy, D Maier, 2005. From databases to dataspaces: a new abstraction for information management. SIGMOD Record 34 (4), 27–33.

Gorentz J. 1992. Data management at biological field stations and coastal marine laboratories, Report of an invitational workshop,W.K. Kellogg Biological Station; 22–26 Apr 1990; Lansing (MI):Michigan State University.

Gurtz ME. 1986. Development of a research data management system. Pages 23–38 in Michener WK, ed. Research Data Management in the Ecological Sciences. Columbia (SC): University of South Carolina Press.

Hanseth, Ole, 2010. From Systems and Tools to Networks and Infrastructures - from Design to Cultivation: Towards a Design Theory of Information Infrastructures. In: Industrial Informatics design, Use and Innovation. IGI Global 2010, p. 122-156. See http://www.ifi.uio.no/~oleha/Publications/ib_ISR_3rd_resubm2.html

Hanseth, O. and E. Monteiro, 1996. Developing Information Infrastructure: The Tension between Standardization and Flexibility. *Science, Technology & Human Values* 21(4): 407-426.

Henderson, A and M Kyng, 1991. There is no place like home: Continuing Design in Use, in Greenbaum, J. and Kyng, M. (eds) *Design at Work: Cooperative Design of Computer Systems*. Lawrence Erlbaum, Hillsdale NJ, 219-240.

Heery, R. and S Anderson, 2005. Digital Repositories Review UKOLN and AHDS: 33. http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf

Hine, C., 2008. Systematics as Cyberscience: Computers, Change, and Continuity in Science. The MIT Press, Cambridge, MA.

Hobbie, JE, SR Carpenter, NB Grimm, JR Gosz, and TR Seastedt, 2003. The US long term ecological research program. BioScience, 53, 1, 21–32

Ingersoll RC, TR Seastedt, M Hartman, 1997. A model information management system for ecological research. BioScience 47:310-316.

Karasti, H and KS Baker, 2004. Infrastructuring for the Long-Term: Ecological Information Management. Proceedings of the 37th Hawaii International Conference on System Sciences. IEEE, Hawaii.

Karasti, H and KS Baker, 2008. Digital Data Practices and the Global Long Term Ecological Research Program. International Journal of Digital Curation 3(2):42-58.

Karasti, H, KS Baker, and F Millerand, 2010. “Infrastructure time: Long-term matters in collaborative development.” Computer Supported Cooperative Work - An International Journal 19:377-415.

Karasti, H, KS Baker, and E Halkola, 2006. Enriching the notion of data curation in escience: data managing and information infrastructuring in the Long Term Ecological Research (LTER) network. Computer Supported Cooperative Work 15, 321-358.

Kling, R and W Scacchi, 1982. The Web of Computing: Computer technology as social organization.8In M. Yovits (ed.), *Advances in Computers*, Vol. 21, 3-90.

Academic Press, New York.

Kortz, M, 2011. Review: The PersonnelDB Design and Development Workshop. LTER Databits Newsletter, Feature Spring 2011. http://databits.lternet.edu/spring-2011/review-personneldb-design-and-development-workshop

Kortz, M, 2009. LTER Unit Registry: Products and Processes. LTER Databits Newsletter, Newsbits 2009 Fall. http://databits.lternet.edu/issues/114#79

Likens, GE, 1983. A priority for ecological research. Bulletin of the Ecological Society of America, **64**, 234-243.

Magnuson JJ 1990. Long-term ecological research and the invisible present. *BioScience* 40: 495–501.

Markus, LM and D Robey, I988. Information Technology and Organizational Change: Causal Structure in Theory and Research, *Management Science*, 34: 5, pp. 583-598.

Michener W. 1986. Data management and long-term ecological research. Pages 1–8 in Michener WK, ed.Research Data Management in the Ecological Sciences. Columbia (SC): University of South Carolina Press.

Millerand, F and GC Bowker, 2009. Metadata Standards: Trajectories and Enactment in the Life of an Ontology. In S. L. Star & M. Lampland (Eds.), Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life, Cornell University Press, p. 149-176.

Millerand, F and KS Baker, 2011. Ocean Informatics Monograph: Ocean Informatics Initiative: An Ethnographic Study (2002-2006). Parts 1 & 2. SIO Technical Report.

Nardi, B, V O’Day, 1999. *Information Ecology: Using Technology with Heart*. Cambridge: MIT Press.

NSB, 2005. National Science Board: Long-Lived Data Collections Enabling Research and Education in the 21st Century. NSF NSB-05-40.

Olson, RJ, LD Voohees, J.M.Field, M.B.Gentry, D.E.Strebel, D.R.Landis, K.F.Huemmrich, and B.W.Meeson, 1994. Packaging and distributing ecological data from multisite studies. In Proceedings of the Eco-Informal ’96, Global Networks for Environmental Inforation, Environmental Research Institute of Michigan, Ann Arbor, pp. 93-102.

Orlikowski, W and S Iacono, 2001. Desperately Seeking the IT in IT Research -A Call to Theorizing the IT Artifact. Information Systems Research 12(2):121-134.

Pickering, A, 1995. The Mangle of Practice: Time, Agency, and Science. Chicago, IL: University of Chicago Press.

Pollock, N and R Williams, 2010. E-Infrastructures: How Do We Know and Understand Tehm? Strategic Ethnography and the Biography of Artefacts. Computer Supported Cooperative Work 19: 521-556.

Porter, JH, 2000. Scientific Databases. In Ecological Data: Design, Management and Processing, WK Michener and JWBrunt (eds). pp 48-69.

Porter J, B Hayden, D Richardso, 1996. Data and information management at the Virginia Coast Reserve Long-Term Ecological Research Site. Pages 731-736. Proceedings of Eco-Informa Workshop, Global Networks for Environmental Information, 4-7 November '96, Lake Buena Vista, FL.

Ribes, D and K Baker, 2007. Modes of Social Science Engagement in Community Infrastrcutucture Design. Proceedings of the International Conference on Communities and Technologies, Michigan State University , East Lansing, Michigan, June 18-30, 2007.

Risser, PG and CG Treworgy, 1986. Overview of ecological research data management. In W.K.Michener, ed. Research dadta management in the ecological sciences. Belle W. Baruch Institute in Marine Science, Number 16. University of South Carolina Press, Columbia, South Carolina, p9-22.

Schuler, D and A Namioka, 1993. Participatory Design: Principles and Practices. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Seastedt, TR, and JM Briggs. 1991. Longterm ecological questions and considerations for taking longterm measurements: Lessons from the LTER and FIFE programs on tallgrass prairie. Long-term Ecological Research: An International Perspective (SCOPE Vol. 47) 153- 172.

Servilla, M, D Costa, C Laney, I San Gil., J Brunt, 2008. The EcoTrends Web Portal: An Architecture for Data Discovery and Exploration. Proceedings of the Environmental Information Management Conference 2008.

Servilla, M, J Brunt, I San Gil, and D Costa, 2006. Pasta: A Network-level Architecture Design for Automating the Creation of Synthetic Products in the LTER Network. LTER DataBits Fall 2006: http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06fall/

Stafford SG, JW Brunt, and WK Michener, 1994. Integration of scientific information management and environmental research. Pages 3019 in Michener WK, Brunt JW, Stafford SG eds. Environmental Information Management and Analysis:

Ecosystem to Global Scales. London: Taylor & Francis.

Suchman, L, 2002). Located accountabilities in technology production. *Scandavian Journal of Information Systems 14*, 2, 91-10

Taylor, PJ, 2005. Unruly Complexity: Ecology, Interpretation, Engagement. University of Chicago Press.

Ulanowicz, RE, 1997. Ecology, the Ascendent Perspective (Complexity in Ecological Systems). Columbia University Press.

Veen, C, CA Federer, D Buso, T Siccama, 1994. Structure and Function of the Hubbard Brook Data Management System. Bulletin of the Ecological Society of America 75: 45.

Wasser, C, 1998. LTER site and network level information management: Challenges for the future. Proceedings of the ILTER Regional Workshop: Long Term Ecological Research – Examples, Methods, Perspectives for Central Europe, 16-18 September 1998. Madralin (near Warsaw), Poland.

8 Appendices

8.1 DataZoo Timeline

Summarizing projects, roles, and system development

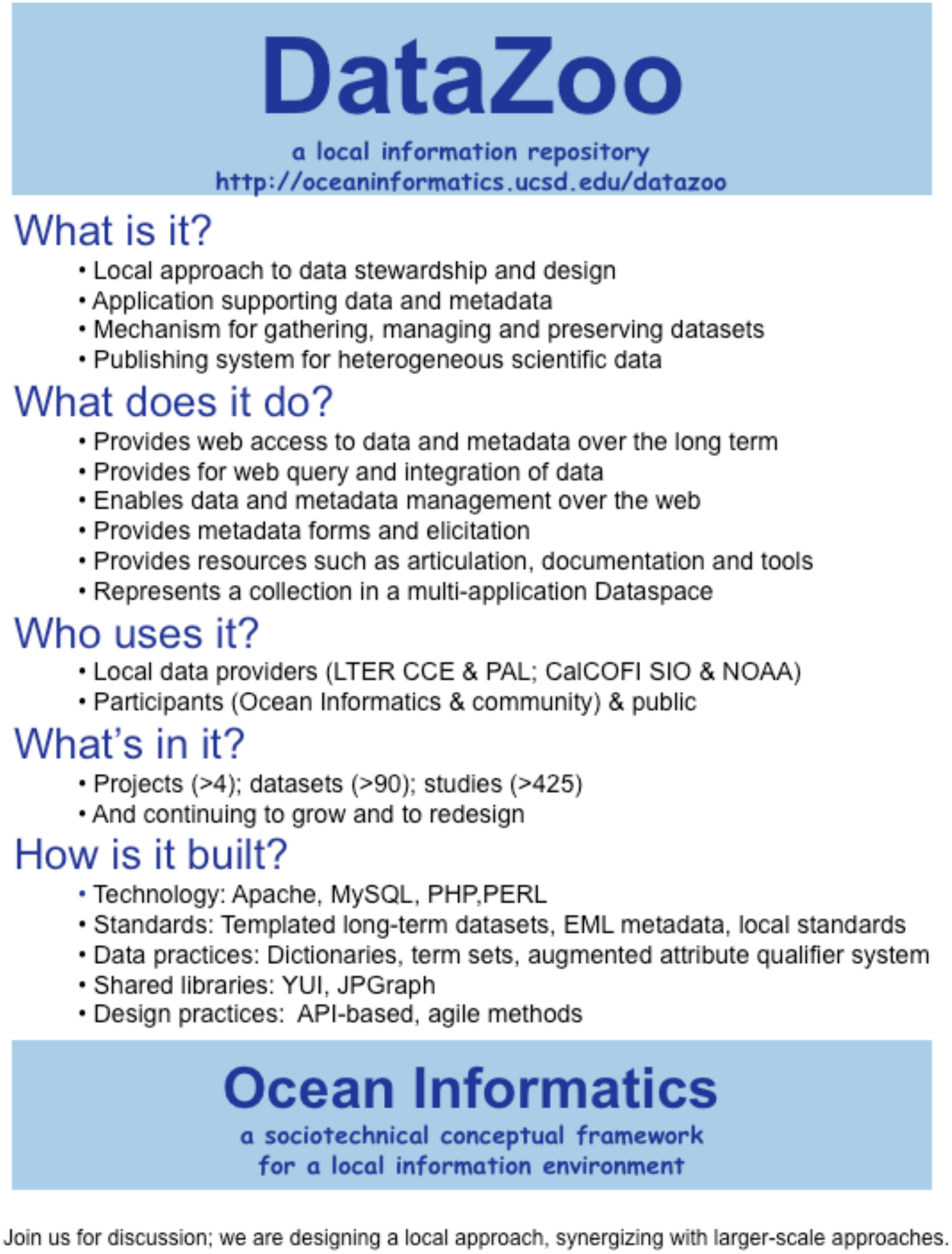
1990	Oct	PAL LTER begin with defined data management role
1991	Jan	Centralized data repository established
1993	Jun	Gopher experimental site created
1994	Jan	Metadata forms developed – desktop version for at sea use
1994	Feb	Web home page established
1994	Jun	Gopher production site for metadata launched
1996	Sep	Protocol pages on data collection methods developed
1997	Sep	Dynamic, online dataset catalog using cgi script created
1998	Aug	All-Site Description Directory prototype developed
2000	Aug	LTER NIS System with site module development
2001	Jan	UCSD/Science Studies Program participation
	Sep	Initiate longitudinal ethnographic studies
2003	Aug	Ocean Informatics Initiative begin
2004	Jan	PAL system programmer developer role begin
	Apr	Project server installed (MacOS)
	Sep	CCE LTER begin with information management role
	Oct	Interoperability reading group begin (winter/fall quarters)
	Dec	Programmer metadata analyst role begin
2005	May	Programmer development role added to make development team
	May	WebDav disk sharing deployed
	Jul	CalCOFI-SIO affiliation begin
	Aug	Eventlogger project initiated
	Sep	Unit Dictionary prototype demo
	Nov	DataCat data system design begin
2006	Apr	DataZoo information system design begin
		Local authentication begun with LDAP single sign-on
		Dictionary development with LTER begun
		JPGraph for web plotting; WordPress for blogging.
	Jul	UCSD/Science Studies formal affiliation
	Jul	Development team member replaced
	Aug	Project web server installed (MacOS)
	Sep	CalCOFI-SWFSC partnership with NOAA
2007	Feb	ZooDB and IchthyoDB planning as complex databases begin
	Jun	Design Studio move

		DataZoo help system developed
	Jul	TermZoo and management interface developed
	Aug	DataZoo launch (PAL, CCE) with code under SVN
		Single sign-on implemented
	Sep	Servers reconfigured
	Nov	Web Media Galleries developed (API)
2008	Jan	DataZoo resource-based architecture
	Mar	DataZoo management interface shift from EML to XML
	May	API capacity development
	May	Media Gallery zooplankton (API)
	Jul 01,	IOD Computer Infrastructure Services initiated (IOD-CIS)
	Jul	IchthyoDB migration begins
	Sep	Three component architecture established
2009	Jan	Service-oriented architecture initiated
		Dataset index numbers for provenance added
	Feb	Zooplankton Dataspace made public
	Jun	Zooplankton workshop by NOAA held
	Jun	Ocean Informatics summer reading group held
		Ocean Institute remote data entry form created as XML document database implementation
		LTER Unit Registry demo
		Ocean Institute application redesign
	Sep	NetCDF output created for OceanSITES CalCOFI project
	Dec	Zooplankton dataspace made public
		Zooplankton data published into DataZoo
2010	Jan	Web server purchased Del (RedHat Linux) with VMWare
	Mar	FileFinder as third component for very large file collections created
	Mar	Unit Registry LTER Post-ASM Working Group funded
		LTER Network Tiger Team Membership begun
	Jun	Ocean Informatics summer reading group held
	Jun	Unit Registry production mode launch
	Jul	Data access layer middleware production as asynchronous web service
	Jul	Templating system redesign
	Aug	Management service redesigned
	Nov	Unit Registry LTER Product Oriented Working Group funded
	Nov	Web services LTER Information Manager Buy-Out Support funded
	Nov	Data format service middleware developed (xls, netCDF outputs)
	Nov	First enterprise level communication re server migration (PAL, CCE, CalCOFI)
	Dec	Migration to new web server completed
2011		

Feb	Data format service expanded (matrix transformations)
Feb	Metadata Access Layer middleware

8.2 DataZoo Flyer

The Datazoo flyer presents a set of summary points describing elements of the data system, to be handed out during data management discussions or project scoping.



DataZoo

a local information repository
http://oceaninformatics.ucsd.edu/datazoo

What is it?

- Local approach to data stewardship and design
- Application supporting data and metadata
- Mechanism for gathering, managing and preserving datasets
- Publishing system for heterogeneous scientific data

What does it do?

- Provides web access to data and metadata over the long term
- Provides for web query and integration of data
- Enables data and metadata management over the web
- Provides metadata forms and elicitation
- Provides resources such as articulation, documentation and tools
- Represents a collection in a multi-application Dataspace

Who uses it?

- Local data providers (LTER CCE & PAL; CalCOFI SIO & NOAA)
- Participants (Ocean Informatics & community) & public

What's in it?

- Projects (>4); datasets (>90); studies (>425)
- And continuing to grow and to redesign

How is it built?

- Technology: Apache, MySQL, PHP,PERL
- Standards: Templated long-term datasets, EML metadata, local standards
- Data practices: Dictionaries, term sets, augmented attribute qualifier system
- Shared libraries: YUI, JPGraph
- Design practices: API-based, agile methods

Ocean Informatics

a sociotechnical conceptual framework
for a local information environment

Join us for discussion; we are designing a local approach, synergizing with larger-scale approaches.

8.3 *Applications for Heterogeneous Data Types*

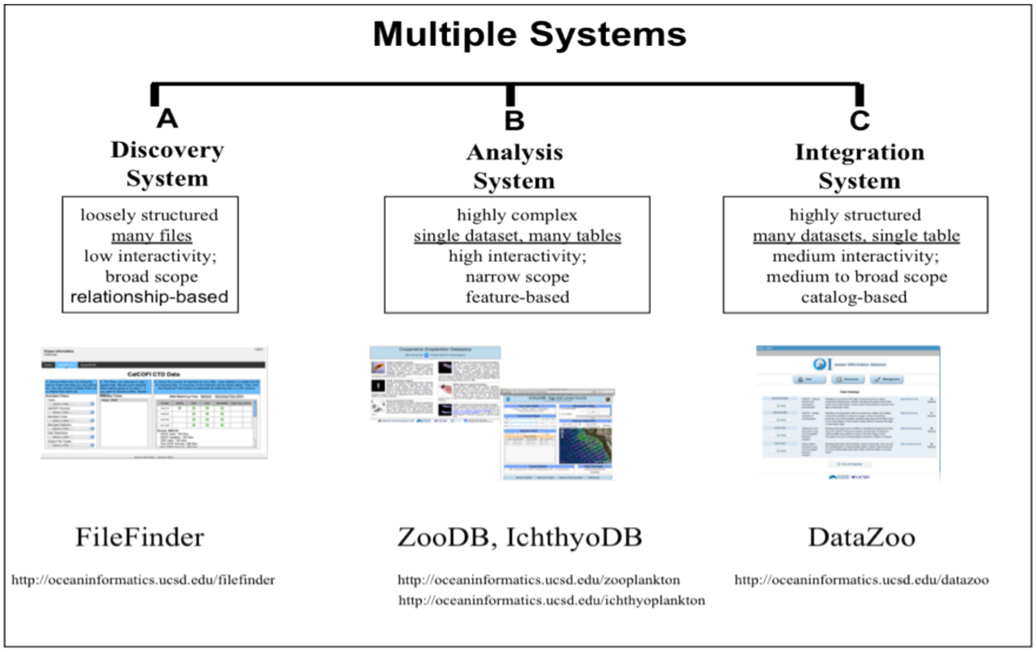
8.3.1 Multi-component information system

The DataZoo data system is part of a larger, multi-component information system. Two overviews of this environment are given below. The first portrays the assemblage of applications with targeted functionality; the second includes an 'abstraction layer' that ties together the components. The applications in this environment are characterized by the roles they play and the types of data they serve.

The first category is **discovery systems**, systems that address very large but relatively structured data products. These systems allow users to find and access data but do not provide data interaction tools such as plotting, previewing, or custom queries.

The second category of applications is **analysis systems.** These systems are feature-rich applications that provide a detailed level of access to a single dataset. Analysis systems are used when the complexity of the data or the features required prevent the use of a more generalized solution.

The third category, of which DataZoo is a member, are **integration systems**. Such systems are focused on providing consistent access to a large number of datasets. In order to provide a normalized interface to many datasets, the data must be homogeneous and highly structured. Integration systems provide a compromise between the broad scope of a discovery system and the high interactivity of an analysis system.



Multiple Systems

A
Discovery System
loosely structured
many files
low interactivity;
broad scope
relationship-based

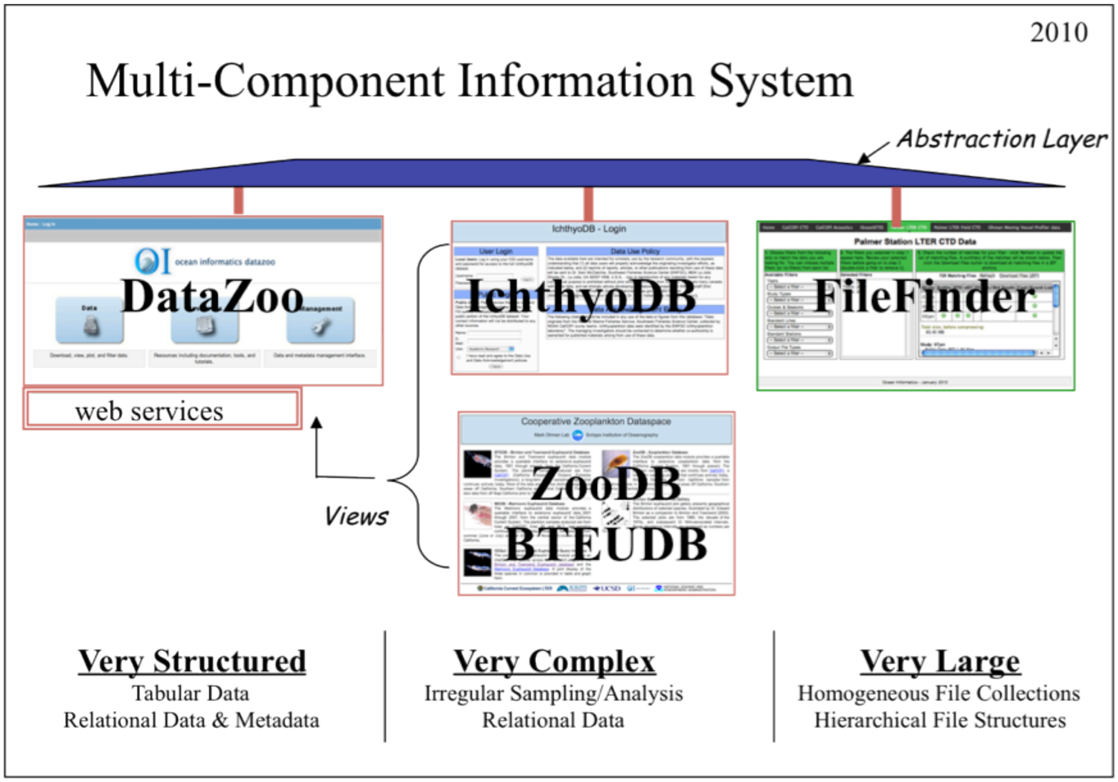
FileFinder
http://oceaninformatics.ucsd.edu/filefinder

B
Analysis System
highly complex
single dataset, many tables
high interactivity;
narrow scope
feature-based

ZooDB, IchthyoDB
http://oceaninformatics.ucsd.edu/zooplankton
http://oceaninformatics.ucsd.edu/ichthyoplankton

C
Integration System
highly structured
many datasets, single table
medium interactivity;
medium to broad scope
catalog-based

DataZoo
http://oceaninformatics.ucsd.edu/datazoo



2010

Multi-Component Information System

Very Structured
Tabular Data
Relational Data & Metadata

Very Complex
Irregular Sampling/Analysis
Relational Data

Very Large
Homogeneous File Collections
Hierarchical File Structures

Applications
http://oceaninformatics.ucsd.edu/datazoo
http://oceaninformatics.ucsd.edu/zoodb
http://oceaninformatics.ucsd.edu/ichthyoplankton
http://oceaninformatics.ucsd.edu/filefinder

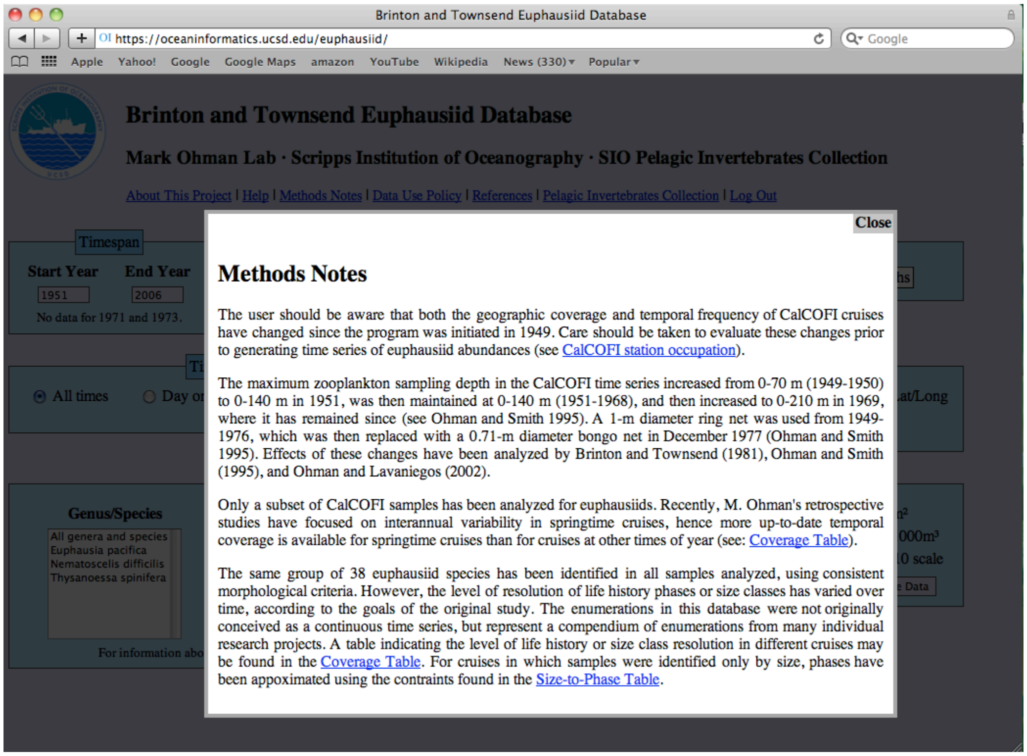
Web sites
http://pal.lternet.edu
http://cce.lternet.edu
http://oceaninformatics.ucsd.edu/
http://interoperability.ucsd.edu

8.3.2 Complex, Irregularly Sampled Measurement Data

The following application modules were developed in order to serve complex, irregularly sampled observational data. the term 'irregular' is used to describe several sampling situations common to biological sampling: a) sampling may be located spatially at regular grid stations but the entire grid is not sampled; b) sampling occurs at varying times so that months and seasons are not evenly represented over time; and c) physical water samples are pooled for counting efficiency to form a single composite sample that is representative of a larger sampling area.

8.3.2.1 EuphausiidDB

The Brinton and Townsend euphausiid data module provides a queriable interface to the extensive euphausiid data from the California Current System. The euphausiid data are based upon nearly 10,000 zooplankton samples from approximately 200 CalCOFI (California Cooperative Oceanic Fisheries Investigations) cruises, spanning the period from 1951 to the present. In total, 39 species of euphausiids were identified and enumerated in this region. Dr. Mark Ohman initiated and directed this project as a contribution from the SIO Pelagic Invertebrates Collection.


Brinton and Townsend Euphausiid Database
https://oceaninformatics.ucsd.edu/euphausiid/
Google
Apple
Yahoo!
Google
Google Maps
amazon
YouTube
Wikipedia
News (330)
Popular
Brinton and Townsend Euphausiid Database
Mark Ohman Lab · Scripps Institution of Oceanography · SIO Pelagic Invertebrates Collection
About This Project | Help | Methods Notes | Data Use Policy | References | Pelagic Invertebrates Collection | Log Out
Timespan
Start Year
End Year
1951
2006
No data for 1971 and 1973.
All times
Day or
Genus/Species
All genera and species
Euphausia pacifica
Nematoscelis difficilis
Thysanoessa spinifera
For information abo
Lat/Long
Close

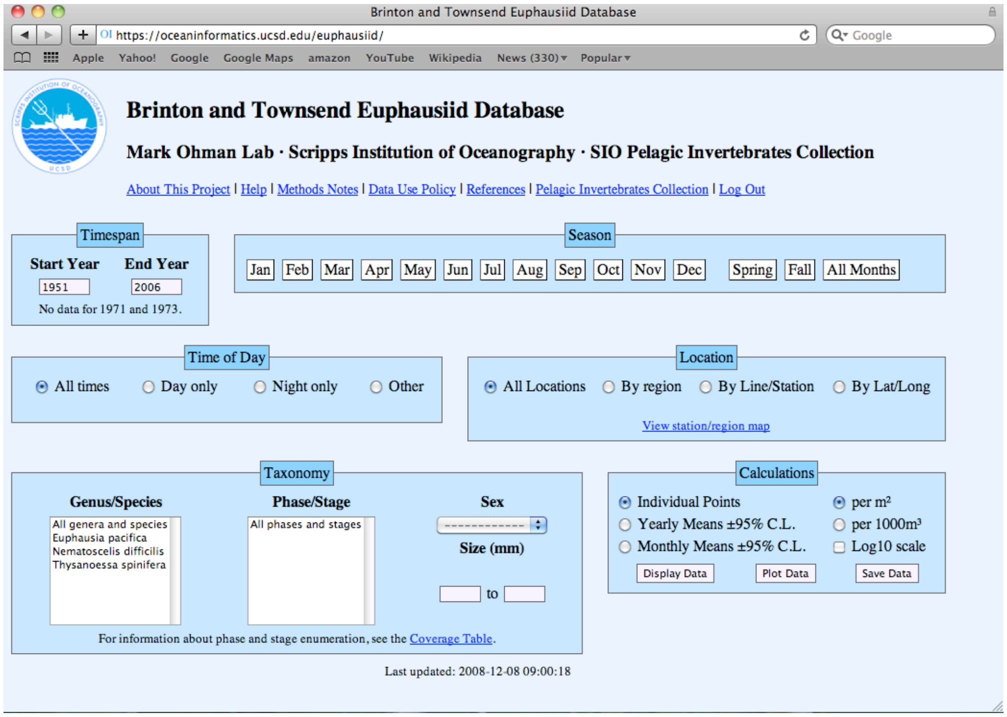
Methods Notes

The user should be aware that both the geographic coverage and temporal frequency of CalCOFI cruises have changed since the program was initiated in 1949. Care should be taken to evaluate these changes prior to generating time series of euphausiid abundances (see CalCOFI station occupation).

The maximum zooplankton sampling depth in the CalCOFI time series increased from 0-70 m (1949-1950) to 0-140 m in 1951, was then maintained at 0-140 m (1951-1968), and then increased to 0-210 m in 1969, where it has remained since (see Ohman and Smith 1995). A 1-m diameter ring net was used from 1949-1976, which was then replaced with a 0.71-m diameter bongo net in December 1977 (Ohman and Smith 1995). Effects of these changes have been analyzed by Brinton and Townsend (1981), Ohman and Smith (1995), and Ohman and Lavaniegos (2002).

Only a subset of CalCOFI samples has been analyzed for euphausiids. Recently, M. Ohman's retrospective studies have focused on interannual variability in springtime cruises, hence more up-to-date temporal coverage is available for springtime cruises than for cruises at other times of year (see: Coverage Table).

The same group of 38 euphausiid species has been identified in all samples analyzed, using consistent morphological criteria. However, the level of resolution of life history phases or size classes has varied over time, according to the goals of the original study. The enumerations in this database were not originally conceived as a continuous time series, but represent a compendium of enumerations from many individual research projects. A table indicating the level of life history or size class resolution in different cruises may be found in the Coverage Table. For cruises in which samples were identified only by size, phases have been appoximated using the contraints found in the Size-to-Phase Table.


Brinton and Townsend Euphausiid Database
https://oceaninformatics.ucsd.edu/euphausiid/
Google
Apple Yahoo! Google Google Maps amazon YouTube Wikipedia News (330) Popular

Brinton and Townsend Euphausiid Database

Mark Ohman Lab · Scripps Institution of Oceanography · SIO Pelagic Invertebrates Collection

About This Project | Help | Methods Notes | Data Use Policy | References | Pelagic Invertebrates Collection | Log Out

Timespan

Start Year	End Year
1951	2006

No data for 1971 and 1973.

Season

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Spring Fall All Months

Time of Day

● All times ○ Day only ○ Night only ○ Other

Location

● All Locations ○ By region ○ By Line/Station ○ By Lat/Long

View station/region map

Taxonomy

Genus/Species

All genera and species
Euphausia pacifica
Nematoscelis difficilis
Thysanoessa spinifera

Phase/Stage

All phases and stages

Sex

Size (mm)

to

For information about phase and stage enumeration, see the Coverage Table.

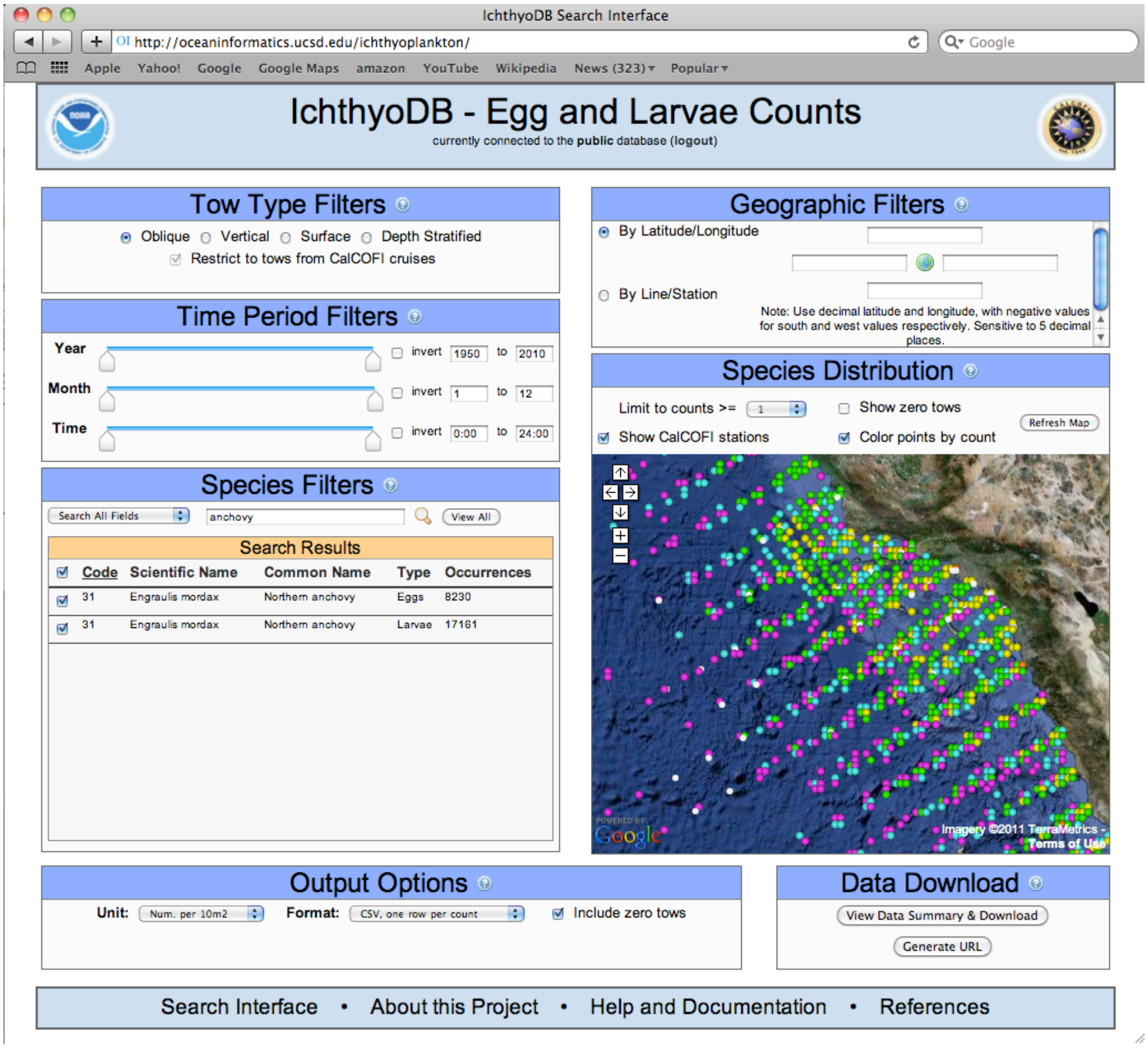
Calculations

● Individual Points ● per m^2
○ Yearly Means ±95% C.L. ○ per 1000m^3
○ Monthly Means ±95% C.L. □ Log10 scale

Last updated: 2008-12-08 09:00:18

8.3.2.2 IchthyoDB

CalCOFI participants at the NOAA Southwest Fisheries Science Center and at Scripps Institution of Oceanography have worked jointly since 1949 on a time-series of biological and physical oceanographic measurements and observations. Cruises are currently conducted quarterly. Equipment and methods are described at the CalCOFI web site and by the SWFSC where current **sampling methods** as well as **gear** are shown. The interface shows first that a selection must be made regarding net type. The use of three net types - pairoVET, oblique, and surface tow - is routine on CalCOFI cruises; the use of the Mocness net system depends upon participant interest and opportunity.



http://oceaninformatics.ucsd.edu/ichthyoplankton/

IchthyoDB - Egg and Larvae Counts

currently connected to the **public** database (**logout**)

Tow Type Filters

● Oblique ○ Vertical ○ Surface ○ Depth Stratified

☑ Restrict to tows from CalCOFI cruises

Time Period Filters

Year □ invert 1950 to 2010

Month □ invert 1 to 12

Time □ invert 0:00 to 24:00

Species Filters

Search All Fields | anchovy | View All

Search Results

☑	Code	Scientific Name	Common Name	Type	Occurrences
☑	31	Engraulis mordax	Northern anchovy	Eggs	8230
☑	31	Engraulis mordax	Northern anchovy	Larvae	17181

Geographic Filters

● By Latitude/Longitude

○ By Line/Station

Note: Use decimal latitude and longitude, with negative values for south and west values respectively. Sensitive to 5 decimal places.

Species Distribution

Limit to counts >= 1

□ Show zero tows

☑ Show CalCOFI stations

☑ Color points by count

Refresh Map

Output Options

Unit: Num. per 10m2 Format: CSV, one row per count ☑ Include zero tows

Data Download

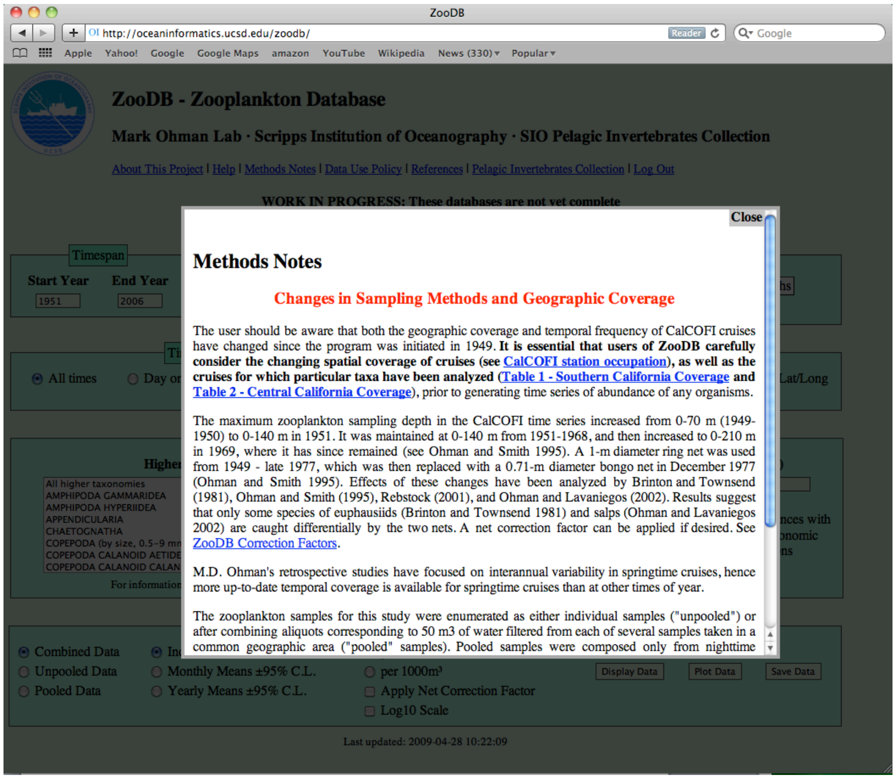
View Data Summary & Download

Generate URL

Search Interface • About this Project • Help and Documentation • References

8.3.2.3 ZooDB

The ZooDB zooplankton data module provides a queriable interface to extensive zooplankton data from the California Current System covering 1951 to the present. The plankton samples analyzed are from CalCOFI (the California Cooperative Oceanic Fisheries Investigations) beginning in 1949. Most of the data originate from springtime cruises (February, March, April or May), 1951-2008 from two distinct geographical areas off California: Southern California (CalCOFI lines 80-93) and Central California (lines 60-70). Only nighttime samples were analyzed. Dr. Mark Ohman initiated and directed this project as a contribution from the SIO Pelagic Invertebrates Collection.



ZooDB - Zooplankton Database

Mark Ohman Lab · Scripps Institution of Oceanography · SIO Pelagic Invertebrates Collection

WORK IN PROGRESS: These databases are not yet complete

Methods Notes

Changes in Sampling Methods and Geographic Coverage

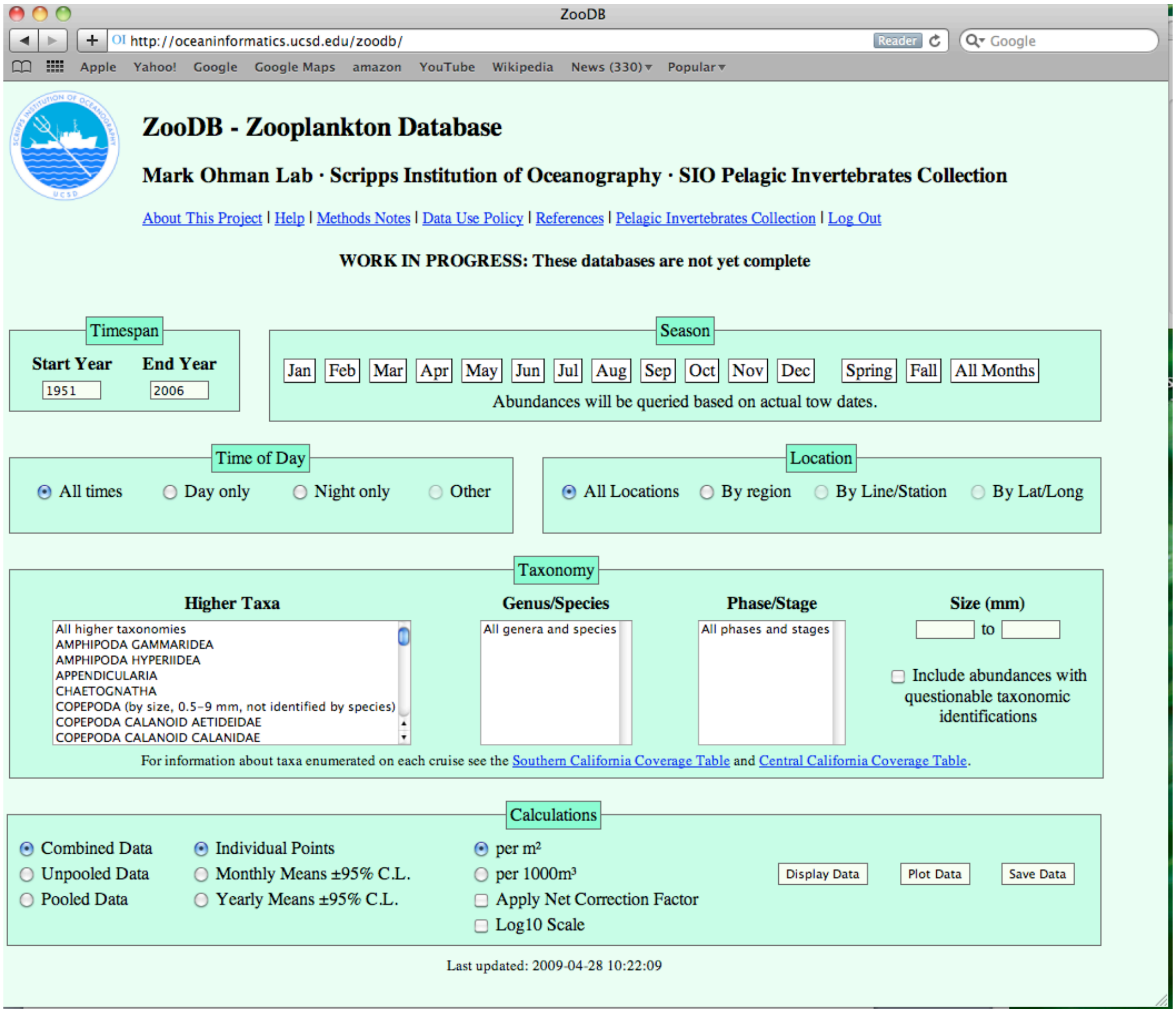
The user should be aware that both the geographic coverage and temporal frequency of CalCOFI cruises have changed since the program was initiated in 1949. **It is essential that users of ZooDB carefully consider the changing spatial coverage of cruises (see CalCOFI station occupation), as well as the cruises for which particular taxa have been analyzed (Table 1 - Southern California Coverage and Table 2 - Central California Coverage**), prior to generating time series of abundance of any organisms.

The maximum zooplankton sampling depth in the CalCOFI time series increased from 0-70 m (1949-1950) to 0-140 m in 1951. It was maintained at 0-140 m from 1951-1968, and then increased to 0-210 m in 1969, where it has since remained (see Ohman and Smith 1995). A 1-m diameter ring net was used from 1949 - late 1977, which was then replaced with a 0.71-m diameter bongo net in December 1977 (Ohman and Smith 1995). Effects of these changes have been analyzed by Brinton and Townsend (1981), Ohman and Smith (1995), Rebstock (2001), and Ohman and Lavaniegos (2002). Results suggest that only some species of euphausiids (Brinton and Townsend 1981) and salps (Ohman and Lavaniegos 2002) are caught differentially by the two nets. A net correction factor can be applied if desired. See ZooDB Correction Factors.

M.D. Ohman's retrospective studies have focused on interannual variability in springtime cruises, hence more up-to-date temporal coverage is available for springtime cruises than at other times of year.

The zooplankton samples for this study were enumerated as either individual samples ("unpooled") or after combining aliquots corresponding to 50 m3 of water filtered from each of several samples taken in a common geographic area ("pooled" samples). Pooled samples were composed only from nighttime

Last updated: 2009-04-28 10:22:09


ZooDB
http://oceaninformatics.ucsd.edu/zoodb/
Reader
Google
Apple Yahoo! Google Google Maps amazon YouTube Wikipedia News (330) Popular

ZooDB - Zooplankton Database

Mark Ohman Lab · Scripps Institution of Oceanography · SIO Pelagic Invertebrates Collection

About This Project | Help | Methods Notes | Data Use Policy | References | Pelagic Invertebrates Collection | Log Out

WORK IN PROGRESS: These databases are not yet complete

Timespan

Start Year 1951 **End Year** 2006

Season

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Spring Fall All Months

Abundances will be queried based on actual tow dates.

Time of Day

● All times ○ Day only ○ Night only ○ Other

Location

● All Locations ○ By region ○ By Line/Station ○ By Lat/Long

Taxonomy

Higher Taxa

All higher taxonomies
AMPHIPODA GAMMARIDEA
AMPHIPODA HYPERIIDEA
APPENDICULARIA
CHAETOGNATHA
COPEPODA (by size, 0.5-9 mm, not identified by species)
COPEPODA CALANOID AETIDEIDAE
COPEPODA CALANOID CALANIDAE

Genus/Species

All genera and species

Phase/Stage

All phases and stages

Size (mm)

to

□ Include abundances with questionable taxonomic identifications

For information about taxa enumerated on each cruise see the Southern California Coverage Table and Central California Coverage Table.

Calculations

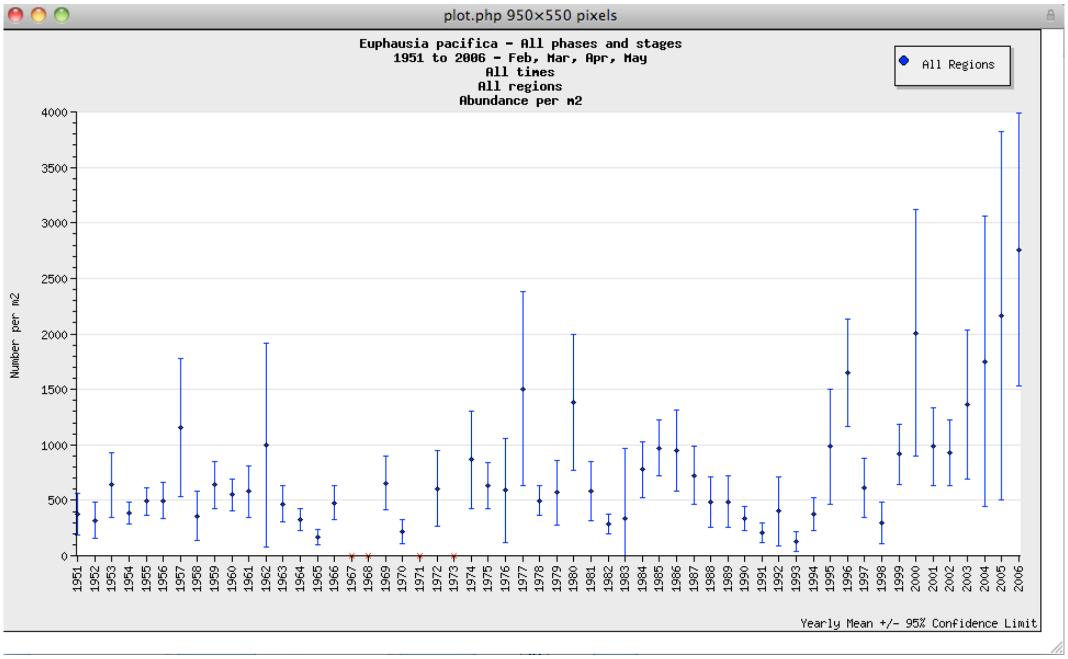
● Combined Data
○ Unpooled Data
○ Pooled Data

● Individual Points
○ Monthly Means ±95% C.L.
○ Yearly Means ±95% C.L.

● per m²
○ per 1000m³
□ Apply Net Correction Factor
□ Log10 Scale

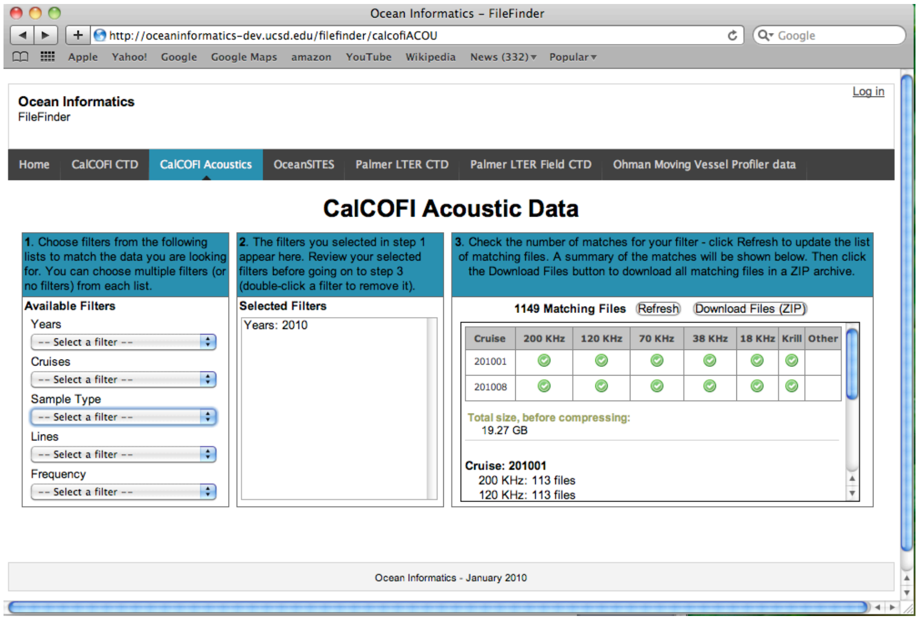
Display Data Plot Data Save Data

Last updated: 2009-04-28 10:22:09



8.3.3 Very Large: Filefinder

The Ocean Informatics FileFinder is service-based application that indexes large, complicated directory structures and maps them to user-friendly search forms. Users are then able to subset a large directory archive based on parameters derived from indexing. FileFinder interfaces with data at a collection level, rather than a record level, supporting the ability to maintain sets of heterogeneous files and data types within a single collection.


Ocean Informatics – FileFinder
http://oceaninformatics-dev.ucsd.edu/filefinder/calcofiACOU
Google
Apple Yahoo! Google Google Maps amazon YouTube Wikipedia News (332) Popular
Log in
Ocean Informatics
FileFinder
Home CalCOFI CTD CalCOFI Acoustics OceanSITES Palmer LTER CTD Palmer LTER Field CTD Ohman Moving Vessel Profiler data
Refresh
Download Files (ZIP)
Ocean Informatics - January 2010

CalCOFI Acoustic Data

1. Choose filters from the following lists to match the data you are looking for. You can choose multiple filters (or no filters) from each list.

2. The filters you selected in step 1 appear here. Review your selected filters before going on to step 3 (double-click a filter to remove it).

3. Check the number of matches for your filter - click Refresh to update the list of matching files. A summary of the matches will be shown below. Then click the Download Files button to download all matching files in a ZIP archive.

Available Filters

Years: -- Select a filter --

Cruises: -- Select a filter --

Sample Type: -- Select a filter --

Lines: -- Select a filter --

Frequency: -- Select a filter --

Selected Filters

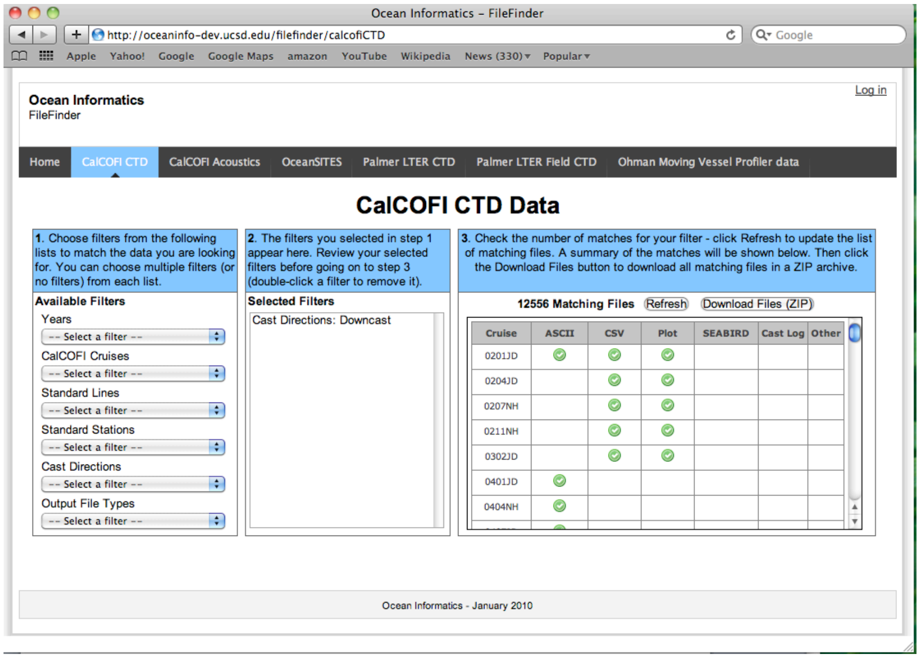
Years: 2010

1149 Matching Files

Cruise	200 KHz	120 KHz	70 KHz	38 KHz	18 KHz	Krill	Other
201001	✓	✓	✓	✓	✓	✓	
201008	✓	✓	✓	✓	✓	✓	

Total size, before compressing:
19.27 GB

Cruise: 201001
200 KHz: 113 files
120 KHz: 113 files


Ocean Informatics – FileFinder
http://oceaninfo-dev.ucsd.edu/filefinder/calcofiCTD
Google
Apple Yahoo! Google Google Maps amazon YouTube Wikipedia News (330) Popular
Log in
Ocean Informatics
FileFinder
Home CalCOFI CTD CalCOFI Acoustics OceanSITES Palmer LTER CTD Palmer LTER Field CTD Ohman Moving Vessel Profiler data
Refresh
Download Files (ZIP)
Ocean Informatics - January 2010

CalCOFI CTD Data

1. Choose filters from the following lists to match the data you are looking for. You can choose multiple filters (or no filters) from each list.

2. The filters you selected in step 1 appear here. Review your selected filters before going on to step 3 (double-click a filter to remove it).

3. Check the number of matches for your filter - click Refresh to update the list of matching files. A summary of the matches will be shown below. Then click the Download Files button to download all matching files in a ZIP archive.

Available Filters

Years: -- Select a filter --

CalCOFI Cruises: -- Select a filter --

Standard Lines: -- Select a filter --

Standard Stations: -- Select a filter --

Cast Directions: -- Select a filter --

Output File Types: -- Select a filter --

Selected Filters

Cast Directions: Downcast

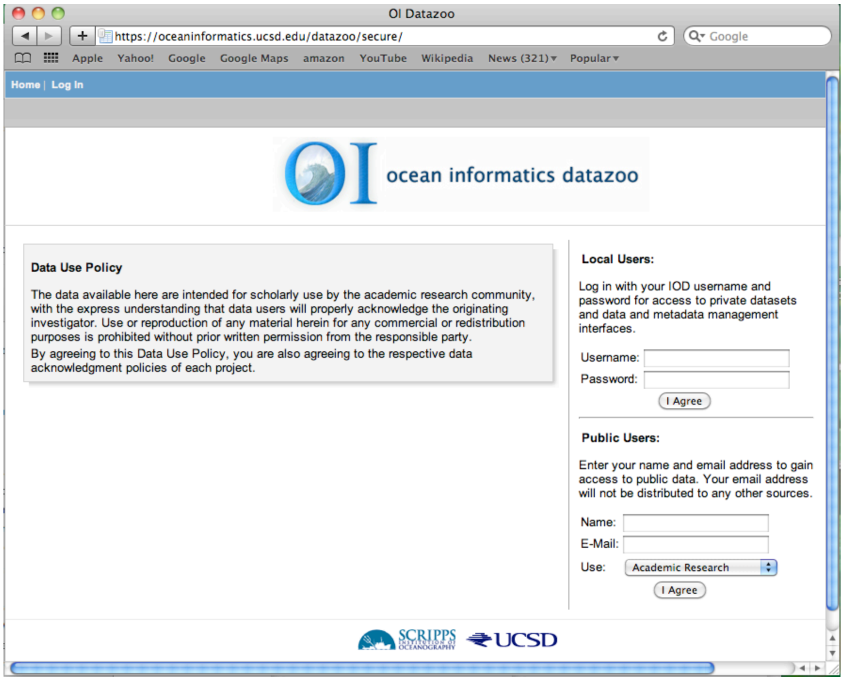
12556 Matching Files

Cruise	ASCII	CSV	Plot	SEABIRD	Cast Log	Other
0201JD	✓	✓	✓			
0204JD		✓	✓			
0207NH		✓	✓			
0211NH		✓	✓			
0302JD		✓	✓			
0401JD	✓					
0404NH	✓					

8.4 Login Registration and Data Policies

Before accessing data from DataZoo or other Ocean Informatics data systems, users must log in. By logging in, users not only identify themselves, but also accept the Data Use and Data Acknowledgement policies associated with the dataset they are accessing. Users may log in either as members of the public or as local users. The latter option allows users to access data and features, such as management options, that are not available to the public. Information on what datasets are being accessed and what type of access (previews, plots, downloads, etc) are being used, is collected in access logs, which can be analyzed to reveal trends in how our data systems are being used.

8.4.1 DataZoo Login


OI Datazoo
https://oceaninformatics.ucsd.edu/datazoo/secure/
Google
Apple Yahoo! Google Google Maps amazon YouTube Wikipedia News (321) Popular
Home | Log In
OI ocean informatics datazoo
SCRIPPS INSTITUTION OF OCEANOGRAPHY
UCSD

Data Use Policy

The data available here are intended for scholarly use by the academic research community, with the express understanding that data users will properly acknowledge the originating investigator. Use or reproduction of any material herein for any commercial or redistribution purposes is prohibited without prior written permission from the responsible party.

By agreeing to this Data Use Policy, you are also agreeing to the respective data acknowledgment policies of each project.

Local Users:

Log in with your IOD username and password for access to private datasets and data and metadata management interfaces.

Username:

Password:

I Agree

Public Users:

Enter your name and email address to gain access to public data. Your email address will not be distributed to any other sources.

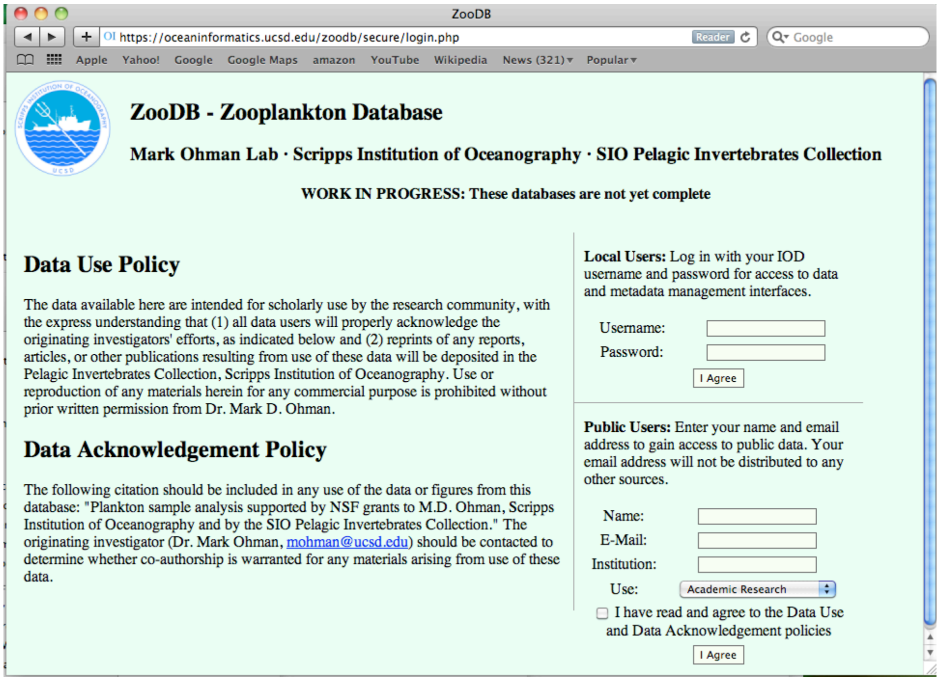
Name:

E-Mail:

Use: Academic Research

I Agree

8.4.2 ZooDB Login


ZooDB
https://oceaninformatics.ucsd.edu/zoodb/secure/login.php
Reader
Google
Apple Yahoo! Google Google Maps amazon YouTube Wikipedia News (321) Popular

ZooDB - Zooplankton Database

Mark Ohman Lab · Scripps Institution of Oceanography · SIO Pelagic Invertebrates Collection

WORK IN PROGRESS: These databases are not yet complete

Data Use Policy

The data available here are intended for scholarly use by the research community, with the express understanding that (1) all data users will properly acknowledge the originating investigators' efforts, as indicated below and (2) reprints of any reports, articles, or other publications resulting from use of these data will be deposited in the Pelagic Invertebrates Collection, Scripps Institution of Oceanography. Use or reproduction of any materials herein for any commercial purpose is prohibited without prior written permission from Dr. Mark D. Ohman.

Data Acknowledgement Policy

The following citation should be included in any use of the data or figures from this database: "Plankton sample analysis supported by NSF grants to M.D. Ohman, Scripps Institution of Oceanography and by the SIO Pelagic Invertebrates Collection." The originating investigator (Dr. Mark Ohman, mohman@ucsd.edu) should be contacted to determine whether co-authorship is warranted for any materials arising from use of these data.

Local Users: Log in with your IOD username and password for access to data and metadata management interfaces.

Username:

Password:

I Agree

Public Users: Enter your name and email address to gain access to public data. Your email address will not be distributed to any other sources.

Name:

E-Mail:

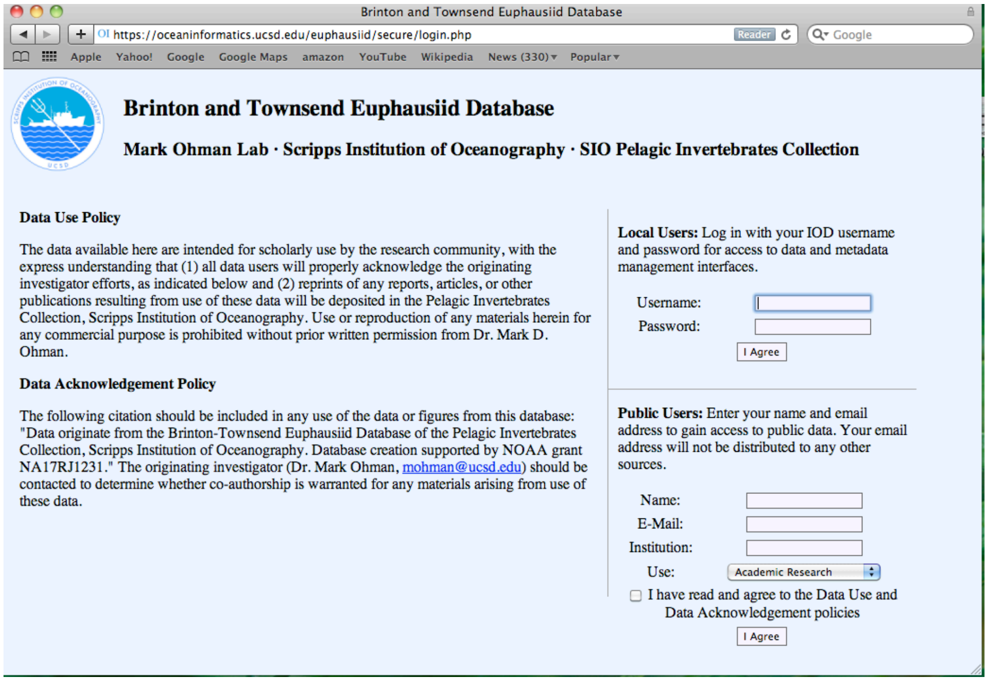
Institution:

Use: Academic Research

□ I have read and agree to the Data Use and Data Acknowledgement policies

I Agree

8.4.3 EuphausiidDB Login


Brinton and Townsend Euphausiid Database
https://oceaninformatics.ucsd.edu/euphausiid/secure/login.php
Reader
Google
Apple Yahoo! Google Google Maps amazon YouTube Wikipedia News (330) Popular

Brinton and Townsend Euphausiid Database

Mark Ohman Lab · Scripps Institution of Oceanography · SIO Pelagic Invertebrates Collection

Data Use Policy

The data available here are intended for scholarly use by the research community, with the express understanding that (1) all data users will properly acknowledge the originating investigator efforts, as indicated below and (2) reprints of any reports, articles, or other publications resulting from use of these data will be deposited in the Pelagic Invertebrates Collection, Scripps Institution of Oceanography. Use or reproduction of any materials herein for any commercial purpose is prohibited without prior written permission from Dr. Mark D. Ohman.

Data Acknowledgement Policy

The following citation should be included in any use of the data or figures from this database: "Data originate from the Brinton-Townsend Euphausiid Database of the Pelagic Invertebrates Collection, Scripps Institution of Oceanography. Database creation supported by NOAA grant NA17RJ1231." The originating investigator (Dr. Mark Ohman, mohman@ucsd.edu) should be contacted to determine whether co-authorship is warranted for any materials arising from use of these data.

Local Users: Log in with your IOD username and password for access to data and metadata management interfaces.

Username:

Password:

I Agree

Public Users: Enter your name and email address to gain access to public data. Your email address will not be distributed to any other sources.

Name:

E-Mail:

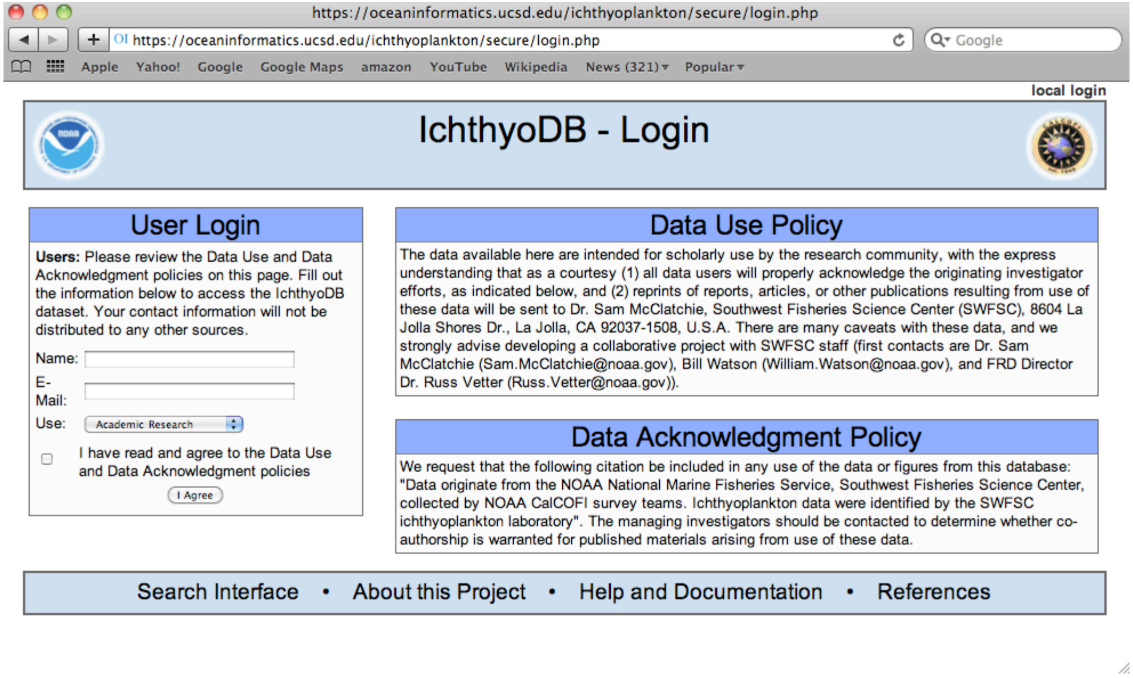
Institution:

Use: Academic Research

□ I have read and agree to the Data Use and Data Acknowledgement policies

I Agree

8.4.4 IchthyoDB Login



https://oceaninformatics.ucsd.edu/ichthyoplankton/secure/login.php
Google
Apple
Yahoo!
Google
Google Maps
amazon
YouTube
Wikipedia
News (321)
Popular
local login
Academic Research
I Agree
Search Interface • About this Project • Help and Documentation • References

IchthyoDB - Login

User Login

Users: Please review the Data Use and Data Acknowledgment policies on this page. Fill out the information below to access the IchthyoDB dataset. Your contact information will not be distributed to any other sources.

Name:

E-Mail:

Use: Academic Research

□ I have read and agree to the Data Use and Data Acknowledgment policies

Data Use Policy

The data available here are intended for scholarly use by the research community, with the express understanding that as a courtesy (1) all data users will properly acknowledge the originating investigator efforts, as indicated below, and (2) reprints of reports, articles, or other publications resulting from use of these data will be sent to Dr. Sam McClatchie, Southwest Fisheries Science Center (SWFSC), 8604 La Jolla Shores Dr., La Jolla, CA 92037-1508, U.S.A. There are many caveats with these data, and we strongly advise developing a collaborative project with SWFSC staff (first contacts are Dr. Sam McClatchie (Sam.McClatchie@noaa.gov), Bill Watson (William.Watson@noaa.gov), and FRD Director Dr. Russ Vetter (Russ.Vetter@noaa.gov)).

Data Acknowledgment Policy

We request that the following citation be included in any use of the data or figures from this database: "Data originate from the NOAA National Marine Fisheries Service, Southwest Fisheries Science Center, collected by NOAA CalCOFI survey teams. Ichthyoplankton data were identified by the SWFSC ichthyoplankton laboratory". The managing investigators should be contacted to determine whether co-authorship is warranted for published materials arising from use of these data.

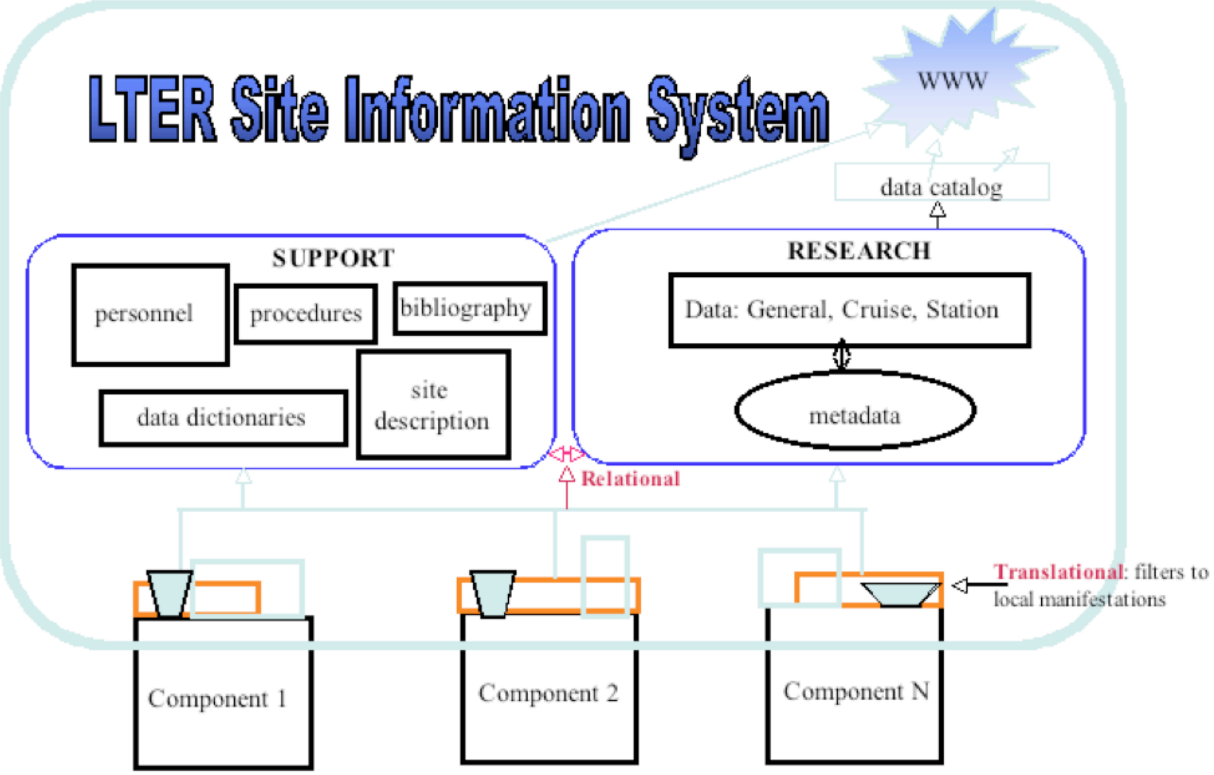
8.5 DataZoo

We use schemas and web interface views to describe the multi-faceted, multi-component DataZoo information system in terms of its development over time and of its functionality.

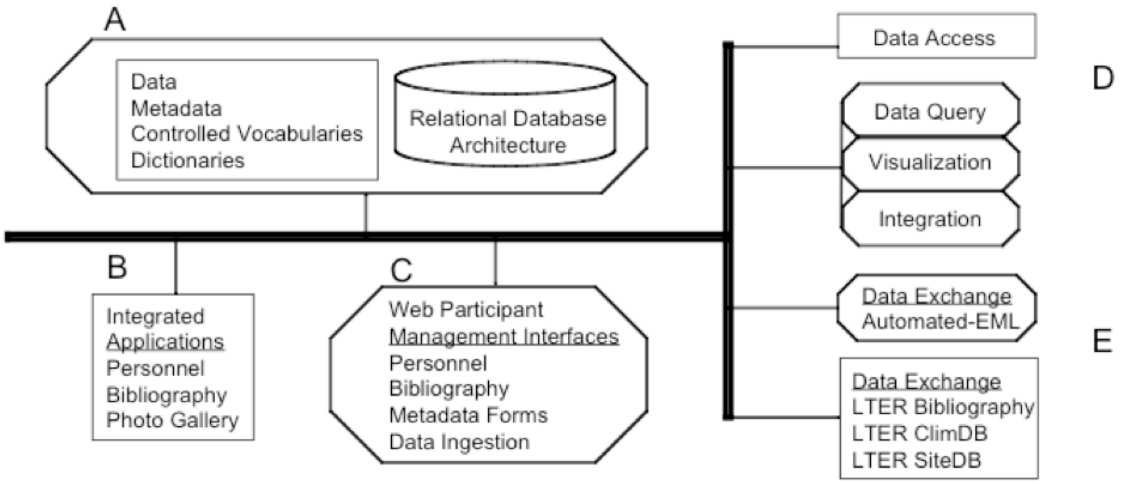
8.5.1 Schemas

The first schema shows how a site information system was envisioned in 2001. Subsequent schemas portray the functionality of existing modules in an integrated system.

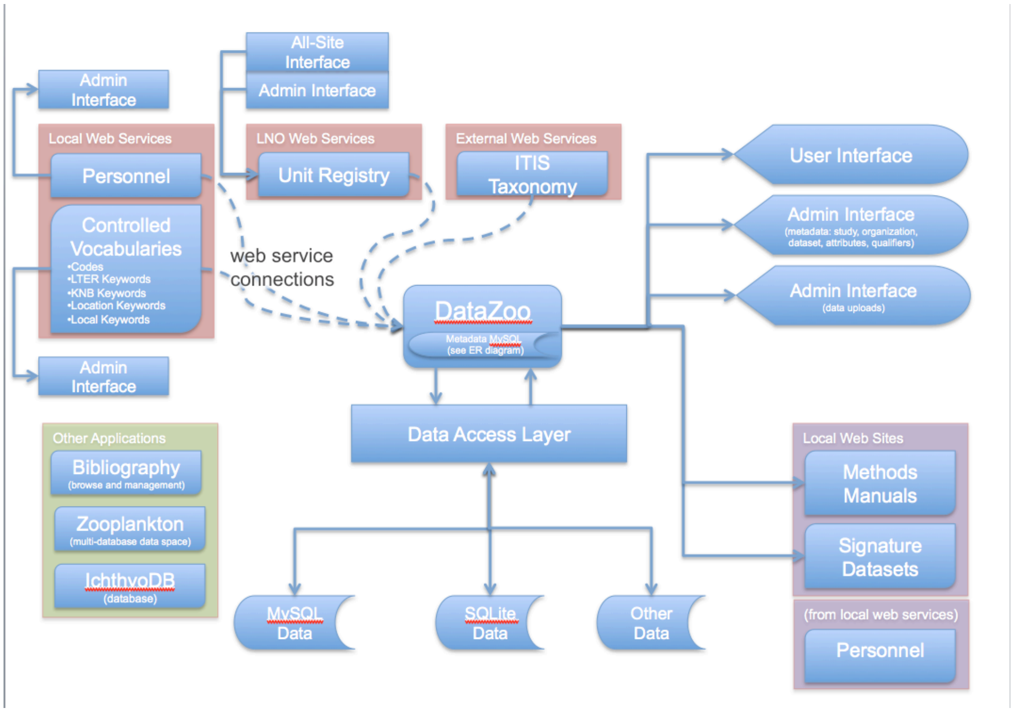
8.5.1.1 Conceptual Framework 2001



8.5.1.2 Schema 2007



8.5.1.3 Schema 2010

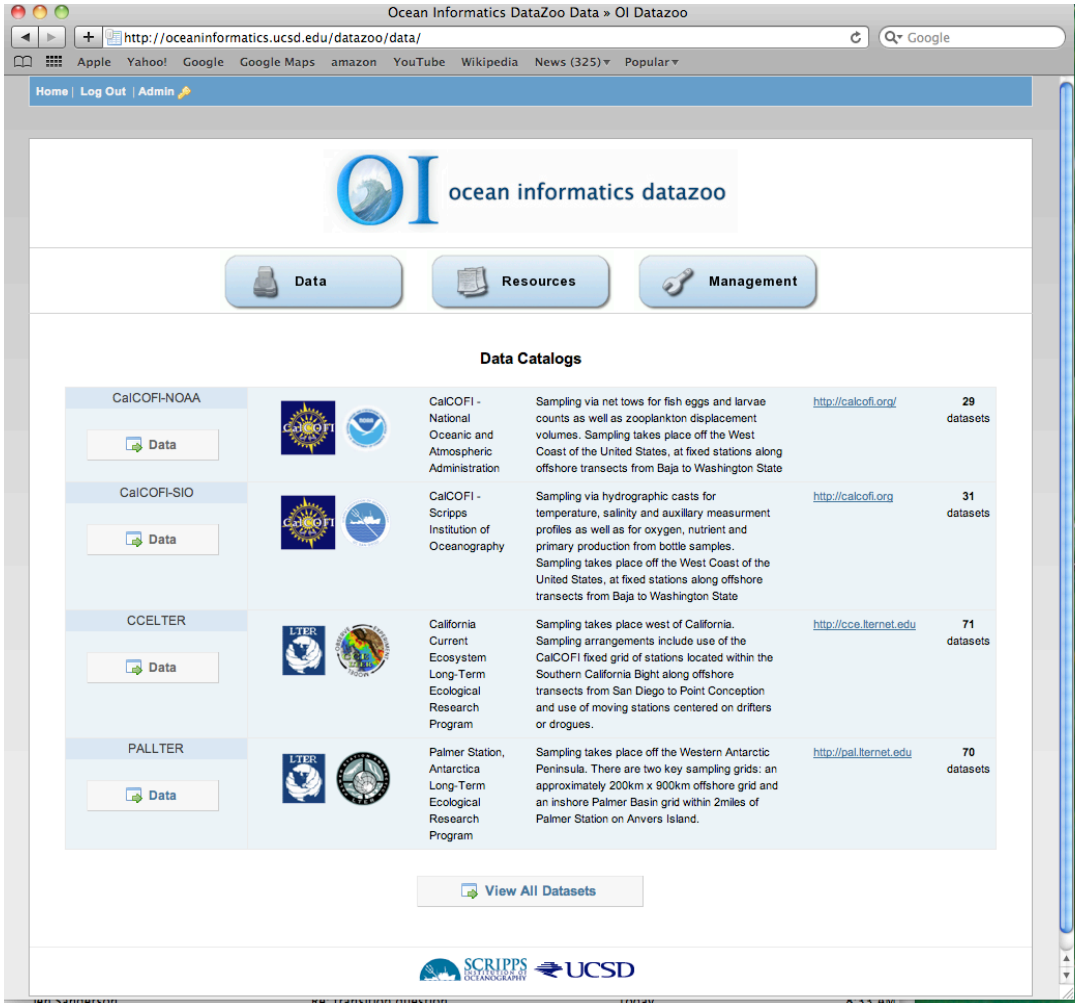


8.5.2 Web Interface Views

The following web interface views provide a brief tour or overview of the DataZoo functionality from an analyst or user perspective. Some views are explained in more detail in the text.

8.5.2.1 DataZoo Project Page

This page provides a list of the current set of data catalogs by project ownership.


Ocean Informatics DataZoo Data » OI Datazoo
http://oceaninformatics.ucsd.edu/datazoo/data/
Google
Apple Yahoo! Google Google Maps amazon YouTube Wikipedia News (325) Popular
Home | Log Out | Admin
ocean informatics datazoo
Data
Resources
Management
SCRIPPS INSTITUTION OF OCEANOGRAPHY
UCSD

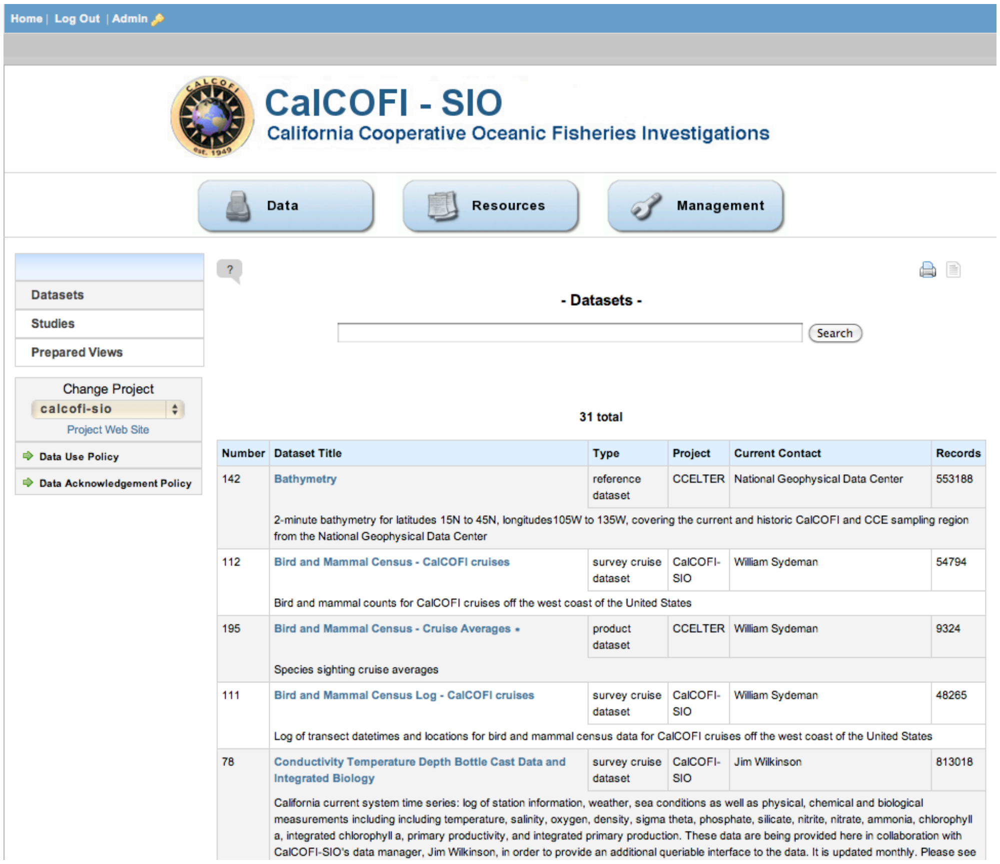
Data Catalogs

CalCOFI-NOAA Data		CalCOFI - National Oceanic and Atmospheric Administration	Sampling via net tows for fish eggs and larvae counts as well as zooplankton displacement volumes. Sampling takes place off the West Coast of the United States, at fixed stations along offshore transects from Baja to Washington State	http://calcofi.org/	29 datasets
CalCOFI-SIO Data		CalCOFI - Scripps Institution of Oceanography	Sampling via hydrographic casts for temperature, salinity and auxillary measurment profiles as well as for oxygen, nutrient and primary production from bottle samples. Sampling takes place off the West Coast of the United States, at fixed stations along offshore transects from Baja to Washington State	http://calcofi.org	31 datasets
CCELTER Data		California Current Ecosystem Long-Term Ecological Research Program	Sampling takes place west of California. Sampling arrangements include use of the CalCOFI fixed grid of stations located within the Southern California Bight along offshore transects from San Diego to Point Conception and use of moving stations centered on drifters or drogues.	http://cce.lternet.edu	71 datasets
PALLTER Data		Palmer Station, Antarctica Long-Term Ecological Research Program	Sampling takes place off the Western Antarctic Peninsula. There are two key sampling grids: an approximately 200km x 900km offshore grid and an inshore Palmer Basin grid within 2miles of Palmer Station on Anvers Island.	http://pal.lternet.edu	70 datasets

View All Datasets

8.5.2.2 Data Catalog

Datasets appear in catalogs through project ownership or affiliation.



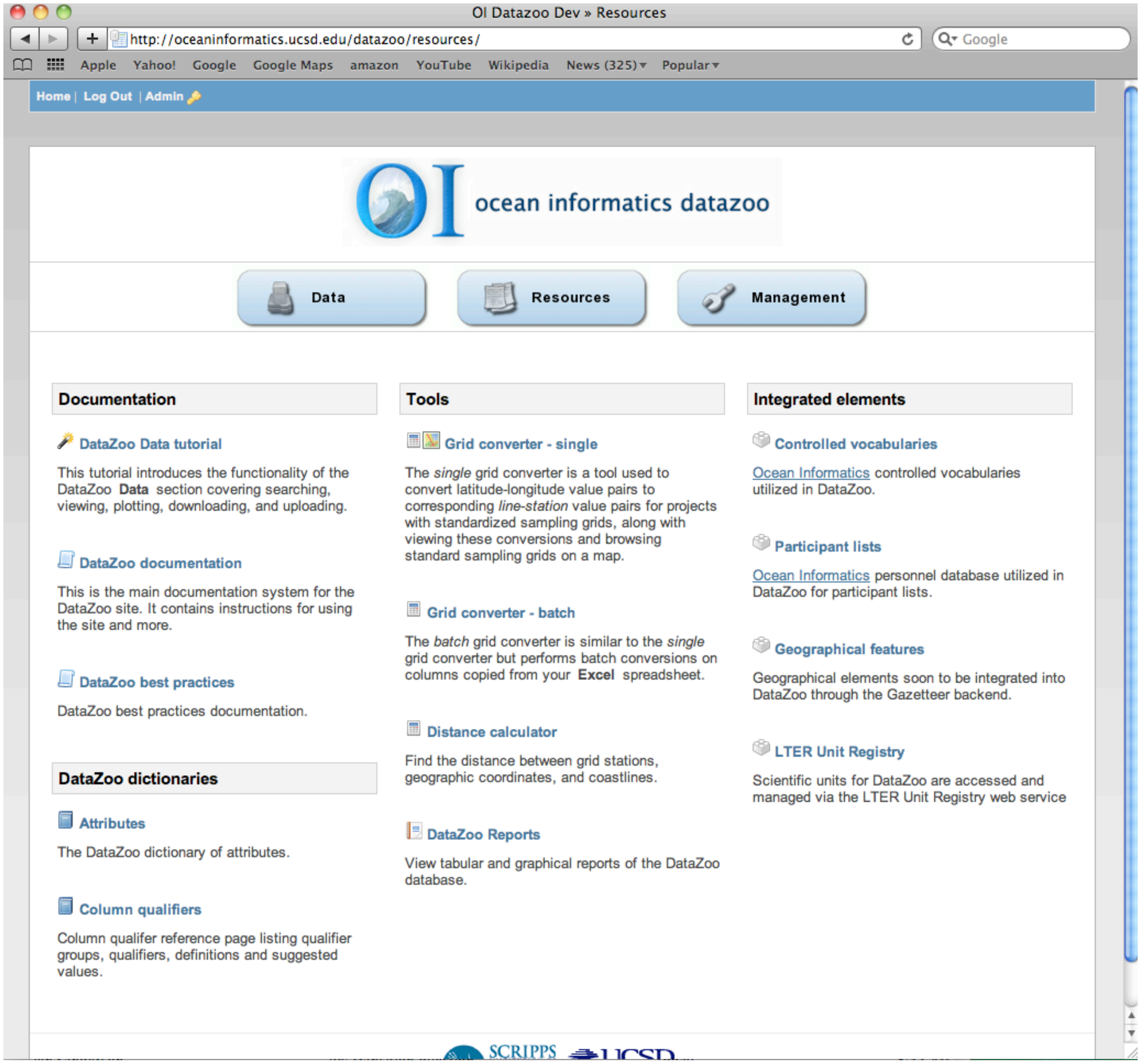
- Datasets -

31 total

Number	Dataset Title	Type	Project	Current Contact	Records
142	Bathymetry	reference dataset	CCELTER	National Geophysical Data Center	553188
	2-minute bathymetry for latitudes 15N to 45N, longitudes105W to 135W, covering the current and historic CalCOFI and CCE sampling region from the National Geophysical Data Center				
112	Bird and Mammal Census - CalCOFI cruises	survey cruise dataset	CalCOFI-SIO	William Sydeman	54794
	Bird and mammal counts for CalCOFI cruises off the west coast of the United States				
195	Bird and Mammal Census - Cruise Averages •	product dataset	CCELTER	William Sydeman	9324
	Species sighting cruise averages				
111	Bird and Mammal Census Log - CalCOFI cruises	survey cruise dataset	CalCOFI-SIO	William Sydeman	48265
	Log of transect datetimes and locations for bird and mammal census data for CalCOFI cruises off the west coast of the United States				
78	Conductivity Temperature Depth Bottle Cast Data and Integrated Biology	survey cruise dataset	CalCOFI-SIO	Jim Wilkinson	813018
	California current system time series: log of station information, weather, sea conditions as well as physical, chemical and biological measurements including including temperature, salinity, oxygen, density, sigma theta, phosphate, silicate, nitrite, nitrate, ammonia, chlorophyll a, integrated chlorophyll a, primary productivity, and integrated primary production. These data are being provided here in collaboration with CalCOFI-SIO's data manager, Jim Wilkinson, in order to provide an additional queriable interface to the data. It is updated monthly. Please see				

8.5.2.3 DataZoo Resources

DataZoo resources present a space for data system elements that do not directly support data discovery or access. Included here are documentation local working dictionaries, tools including grid converters and distance calculators, and integrative elements including controlled vocabularies, personnel, and the unit registry.



Documentation

DataZoo Data tutorial

This tutorial introduces the functionality of the DataZoo **Data** section covering searching, viewing, plotting, downloading, and uploading.

DataZoo documentation

This is the main documentation system for the DataZoo site. It contains instructions for using the site and more.

DataZoo best practices

DataZoo best practices documentation.

DataZoo dictionaries

Attributes

The DataZoo dictionary of attributes.

Column qualifiers

Column qualifer reference page listing qualifier groups, qualifiers, definitions and suggested values.

Tools

Grid converter - single

The *single* grid converter is a tool used to convert latitude-longitude value pairs to corresponding *line-station* value pairs for projects with standardized sampling grids, along with viewing these conversions and browsing standard sampling grids on a map.

Grid converter - batch

The *batch* grid converter is similar to the *single* grid converter but performs batch conversions on columns copied from your **Excel** spreadsheet.

Distance calculator

Find the distance between grid stations, geographic coordinates, and coastlines.

DataZoo Reports

View tabular and graphical reports of the DataZoo database.

Integrated elements

Controlled vocabularies

Ocean Informatics controlled vocabularies utilized in DataZoo.

Participant lists

Ocean Informatics personnel database utilized in DataZoo for participant lists.

Geographical features

Geographical elements soon to be integrated into DataZoo through the Gazetteer backend.

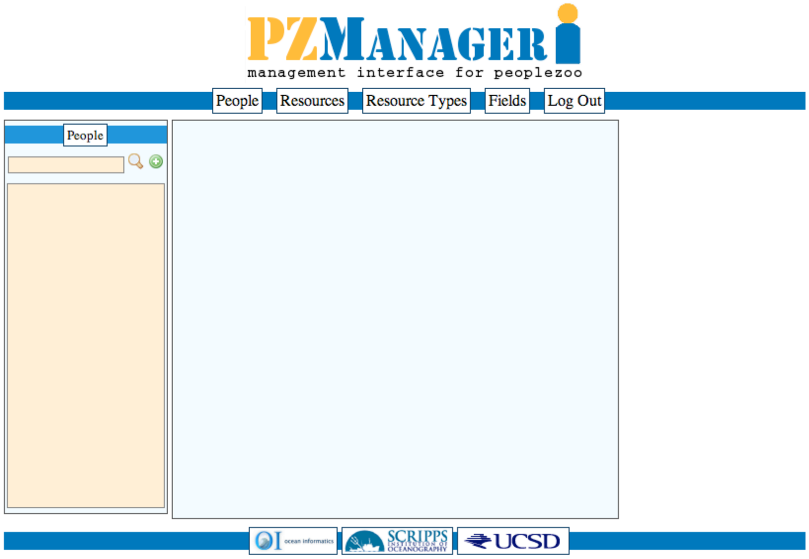
LTER Unit Registry

Scientific units for DataZoo are accessed and managed via the LTER Unit Registry web service

8.5.2.3.1 PeopleZoo Personnel Manager

http://oceaninformatics.ucsd.edu/pzmanager

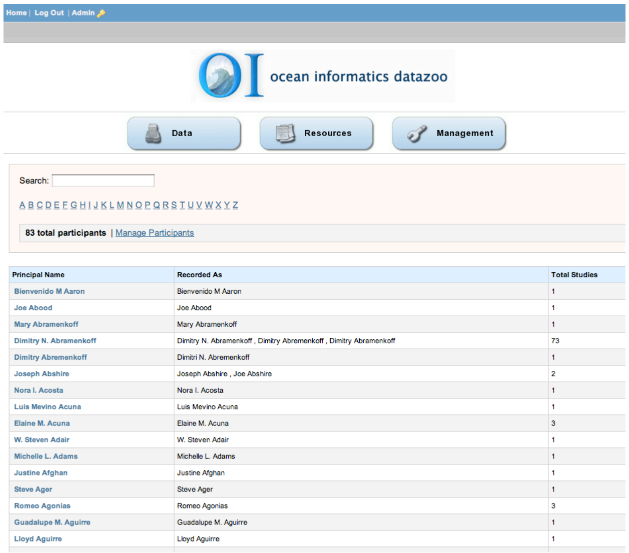
Personnel are documented using a stand-alone personnel management application called PeopleZoo. This application provides support for managing personnel and their relations to defined resources.



8.5.2.3.2 Participant lists

http://oceaninformatics.ucsd.edu/datazoo/resources/participants

A participant application exists to query about an individual's participation in studies over time.



Home | Log Out | Admin
ocean informatics datazoo
Data
Resources
Management
Search:
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

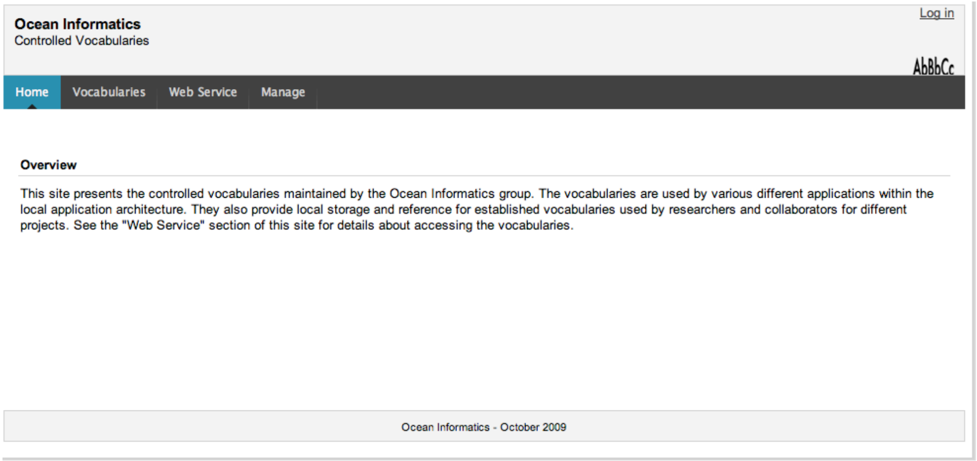
83 total participants | Manage Participants

Principal Name	Recorded As	Total Studies
Bienvenido M Aaron	Bienvenido M Aaron	1
Joe Abood	Joe Abood	1
Mary Abramenkoff	Mary Abramenkoff	1
Dimitry N. Abramenkoff	Dimitry N. Abramenkoff , Dimitry Abremenkoff , Dimitry Abramenkoff	73
Dimitry Abremenkoff	Dimitri N. Abremenkoff	1
Joseph Abshire	Joseph Abshire , Joe Abshire	2
Nora I. Acosta	Nora I. Acosta	1
Luis Mevino Acuna	Luis Mevino Acuna	1
Elaine M. Acuna	Elaine M. Acuna	3
W. Steven Adair	W. Steven Adair	1
Michelle L. Adams	Michelle L. Adams	1
Justine Afghan	Justine Afghan	1
Steve Ager	Steve Ager	1
Romeo Agonias	Romeo Agonias	3
Guadalupe M. Aguirre	Guadalupe M. Aguirre	1
Lloyd Aguirre	Lloyd Aguirre	1

8.5.2.3.3 Controlled Vocabularies

http://oceaninformatics.ucsd.edu/datazoo/resources/controllevocabularies

Controlled vocabularies used throughout the DataZoo environment are gathered together for management through this application.



Overview

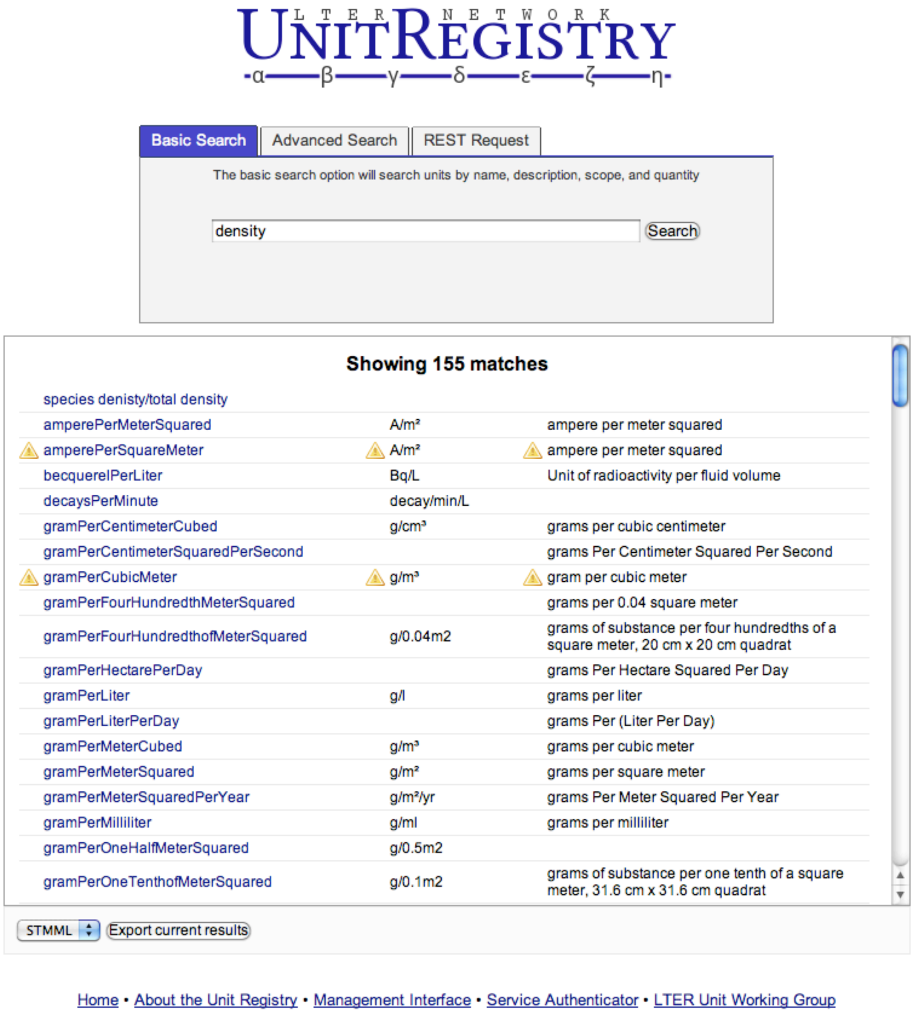
This site presents the controlled vocabularies maintained by the Ocean Informatics group. The vocabularies are used by various different applications within the local application architecture. They also provide local storage and reference for established vocabularies used by researchers and collaborators for different projects. See the "Web Service" section of this site for details about accessing the vocabularies.

Ocean Informatics - October 2009

8.5.2.3.4 Unit Registry

http://unit.lternet.edu/unitregistry/about.php

The LTER Network Unit Registry is a web service-enabled database of standardized scientific units in use at different levels of the LTER research community. The goal of the Unit Registry project is to provide a central tool that is useful to distributed end users as well as application developers. The search interface you are currently using is a client built over the Unit Registry web service - one of many ways that users and applications can interact with the Unit Registry. By creating a database that can be accessed by users and also incorporated into site- and network-level data systems, we aim to support current and future scientific projects through data comparability, integration, and synthesis. Ocean Informatics participants led first an LTER unit Dctionary Working Group starting in 2005, a Unit Metadata group, a Unit Registry Working Group, and then a Web Services Working Group.



The basic search option will search units by name, description, scope, and quantity

density

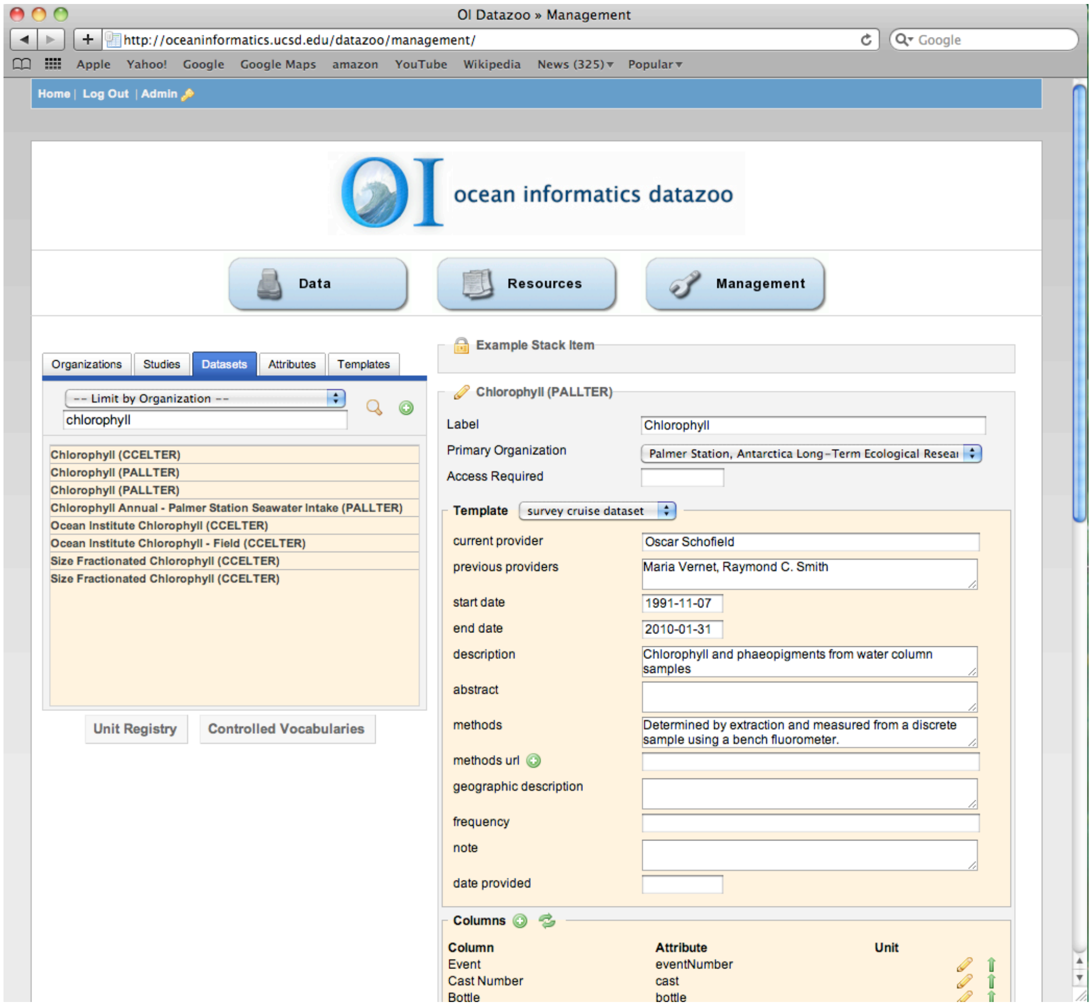
Showing 155 matches

species denisty/total density		
amperePerMeterSquared	A/m^2	ampere per meter squared
⚠ amperePerSquareMeter	⚠ A/m^2	⚠ ampere per meter squared
becquerelPerLiter	Bq/L	Unit of radioactivity per fluid volume
decaysPerMinute	decay/min/L	
gramPerCentimeterCubed	g/cm^3	grams per cubic centimeter
gramPerCentimeterSquaredPerSecond		grams Per Centimeter Squared Per Second
⚠ gramPerCubicMeter	⚠ g/m^3	⚠ gram per cubic meter
gramPerFourHundredthMeterSquared		grams per 0.04 square meter
gramPerFourHundredthofMeterSquared	g/0.04m2	grams of substance per four hundredths of a square meter, 20 cm x 20 cm quadrat
gramPerHectarePerDay		grams Per Hectare Squared Per Day
gramPerLiter	g/l	grams per liter
gramPerLiterPerDay		grams Per (Liter Per Day)
gramPerMeterCubed	g/m^3	grams per cubic meter
gramPerMeterSquared	g/m^2	grams per square meter
gramPerMeterSquaredPerYear	$g/m^2/yr$	grams Per Meter Squared Per Year
gramPerMilliliter	g/ml	grams per milliliter
gramPerOneHalfMeterSquared	g/0.5m2	
gramPerOneTenthofMeterSquared	g/0.1m2	grams of substance per one tenth of a square meter, 31.6 cm x 31.6 cm quadrat

STMML
Export current results
Home • About the Unit Registry • Management Interface • Service Authenticator • LTER Unit Working Group

8.5.2.4 DataZoo Management Interface

DataZoo's management interface provides graphical access for administrative personnel to edit metadata records and upload data.


OI Datazoo » Management
http://oceaninformatics.ucsd.edu/datazoo/management/
Google
Apple Yahoo! Google Google Maps amazon YouTube Wikipedia News (325) Popular
Home | Log Out | Admin
ocean informatics datazoo
Data
Resources
Management
Organizations Studies Datasets Attributes Templates
-- Limit by Organization --
Unit Registry
Controlled Vocabularies

chlorophyll

Chlorophyll (CCELTER)
Chlorophyll (PALLTER)
Chlorophyll (PALLTER)
Chlorophyll Annual - Palmer Station Seawater Intake (PALLTER)
Ocean Institute Chlorophyll (CCELTER)
Ocean Institute Chlorophyll - Field (CCELTER)
Size Fractionated Chlorophyll (CCELTER)
Size Fractionated Chlorophyll (CCELTER)

Example Stack Item

Chlorophyll (PALLTER)

Label: Chlorophyll
Primary Organization: Palmer Station, Antarctica Long-Term Ecological Resea
Access Required:

Template: survey cruise dataset

current provider: Oscar Schofield
previous providers: Maria Vernet, Raymond C. Smith
start date: 1991-11-07
end date: 2010-01-31
description: Chlorophyll and phaeopigments from water column samples
abstract:
methods: Determined by extraction and measured from a discrete sample using a bench fluorometer.
methods url:
geographic description:
frequency:
note:
date provided:

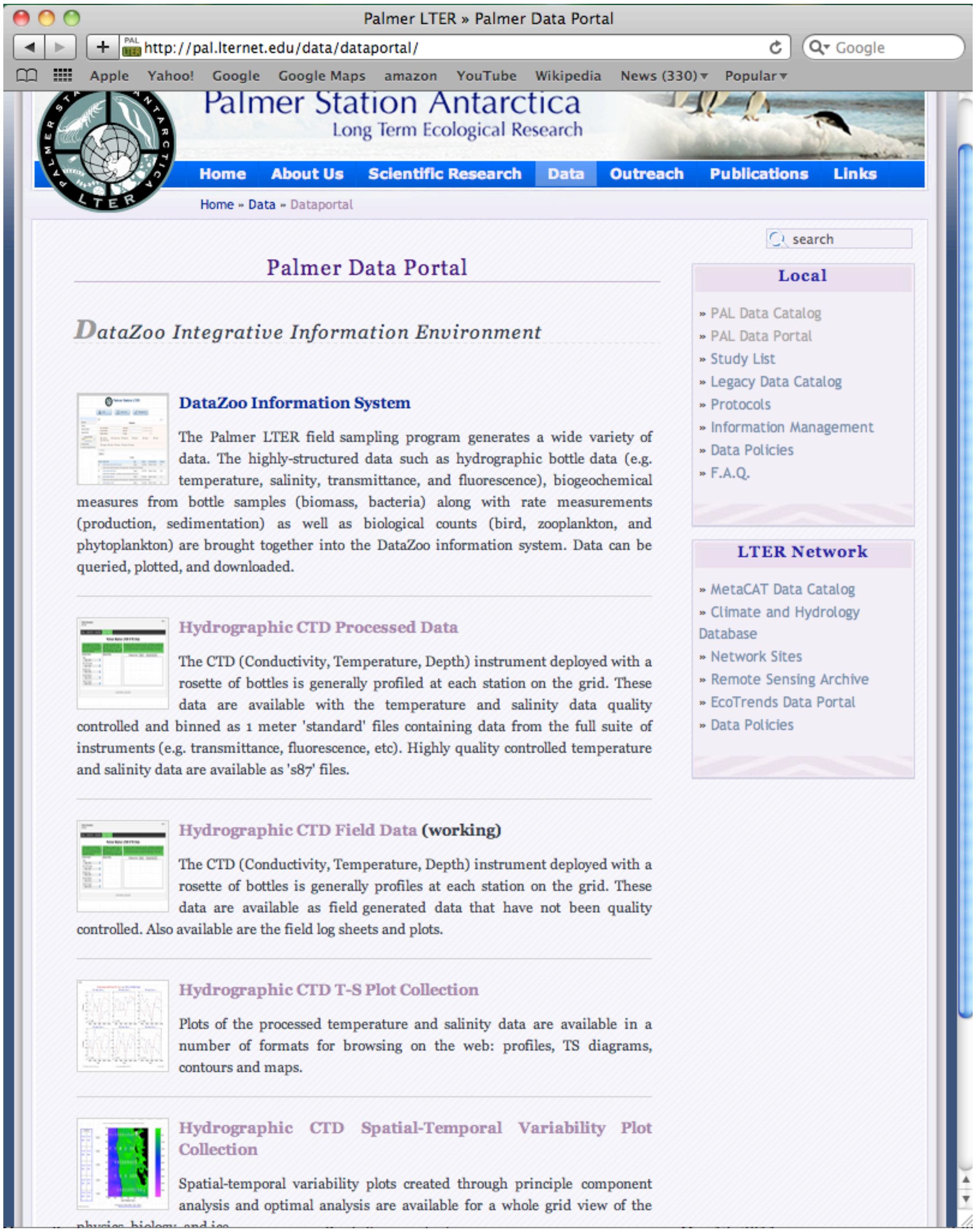
Columns

Column	Attribute	Unit
Event	eventNumber	
Cast Number	cast	
Bottle	bottle	

8.6 *Data Portal*

All of the components of the Ocean Informatics system are accessed via a the Data Portal. Although DataZoo is our core system, very large collections presented online using the FileFinder module and a variety of specialized interfaces for complex data.

8.6.1 Palmer LTER



Palmer Data Portal

DataZoo Integrative Information Environment

DataZoo Information System

The Palmer LTER field sampling program generates a wide variety of data. The highly-structured data such as hydrographic bottle data (e.g. temperature, salinity, transmittance, and fluorescence), biogeochemical measures from bottle samples (biomass, bacteria) along with rate measurements (production, sedimentation) as well as biological counts (bird, zooplankton, and phytoplankton) are brought together into the DataZoo information system. Data can be queried, plotted, and downloaded.

Hydrographic CTD Processed Data

The CTD (Conductivity, Temperature, Depth) instrument deployed with a rosette of bottles is generally profiled at each station on the grid. These data are available with the temperature and salinity data quality controlled and binned as 1 meter 'standard' files containing data from the full suite of instruments (e.g. transmittance, fluorescence, etc). Highly quality controlled temperature and salinity data are available as 's87' files.

Hydrographic CTD Field Data (working)

The CTD (Conductivity, Temperature, Depth) instrument deployed with a rosette of bottles is generally profiles at each station on the grid. These data are available as field generated data that have not been quality controlled. Also available are the field log sheets and plots.

Hydrographic CTD T-S Plot Collection

Plots of the processed temperature and salinity data are available in a number of formats for browsing on the web: profiles, TS diagrams, contours and maps.

Hydrographic CTD Spatial-Temporal Variability Plot Collection

Spatial-temporal variability plots created through principle component analysis and optimal analysis are available for a whole grid view of the physics, biology, and ice.

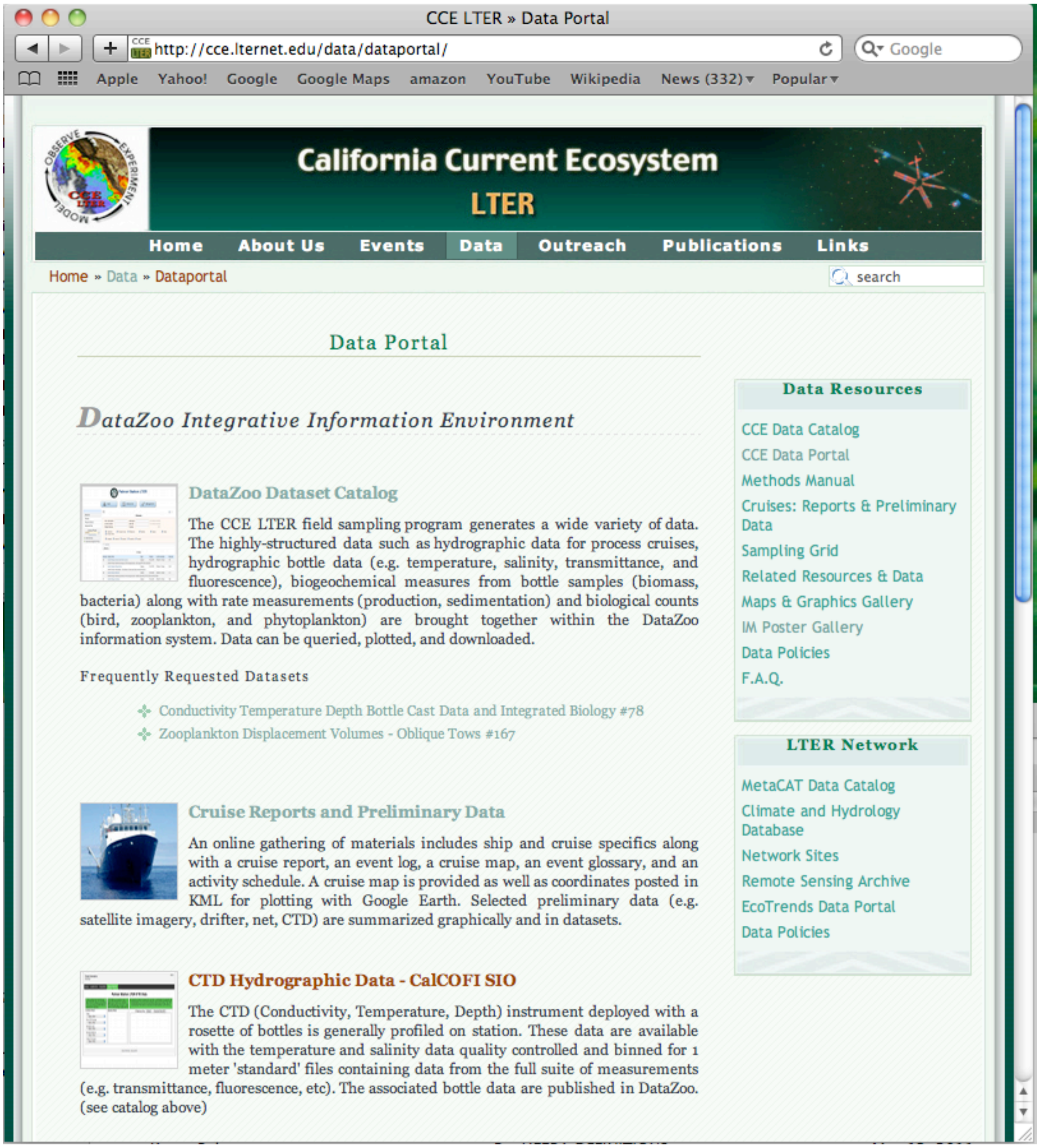
Local

- PAL Data Catalog
- PAL Data Portal
- Study List
- Legacy Data Catalog
- Protocols
- Information Management
- Data Policies
- F.A.Q.

LTER Network

- MetaCAT Data Catalog
- Climate and Hydrology Database
- Network Sites
- Remote Sensing Archive
- EcoTrends Data Portal
- Data Policies

8.6.2 CCELTER



California Current Ecosystem LTER

Home » Data » Dataportal

Data Portal

DataZoo Integrative Information Environment

DataZoo Dataset Catalog

The CCE LTER field sampling program generates a wide variety of data. The highly-structured data such as hydrographic data for process cruises, hydrographic bottle data (e.g. temperature, salinity, transmittance, and fluorescence), biogeochemical measures from bottle samples (biomass, bacteria) along with rate measurements (production, sedimentation) and biological counts (bird, zooplankton, and phytoplankton) are brought together within the DataZoo information system. Data can be queried, plotted, and downloaded.

Frequently Requested Datasets

- Conductivity Temperature Depth Bottle Cast Data and Integrated Biology #78
- Zooplankton Displacement Volumes - Oblique Tows #167

Cruise Reports and Preliminary Data

An online gathering of materials includes ship and cruise specifics along with a cruise report, an event log, a cruise map, an event glossary, and an activity schedule. A cruise map is provided as well as coordinates posted in KML for plotting with Google Earth. Selected preliminary data (e.g. satellite imagery, drifter, net, CTD) are summarized graphically and in datasets.

CTD Hydrographic Data - CalCOFI SIO

The CTD (Conductivity, Temperature, Depth) instrument deployed with a rosette of bottles is generally profiled on station. These data are available with the temperature and salinity data quality controlled and binned for 1 meter 'standard' files containing data from the full suite of measurements (e.g. transmittance, fluorescence, etc). The associated bottle data are published in DataZoo. (see catalog above)

Data Resources

- CCE Data Catalog
- CCE Data Portal
- Methods Manual
- Cruises: Reports & Preliminary Data
- Sampling Grid
- Related Resources & Data
- Maps & Graphics Gallery
- IM Poster Gallery
- Data Policies
- F.A.Q.

LTER Network

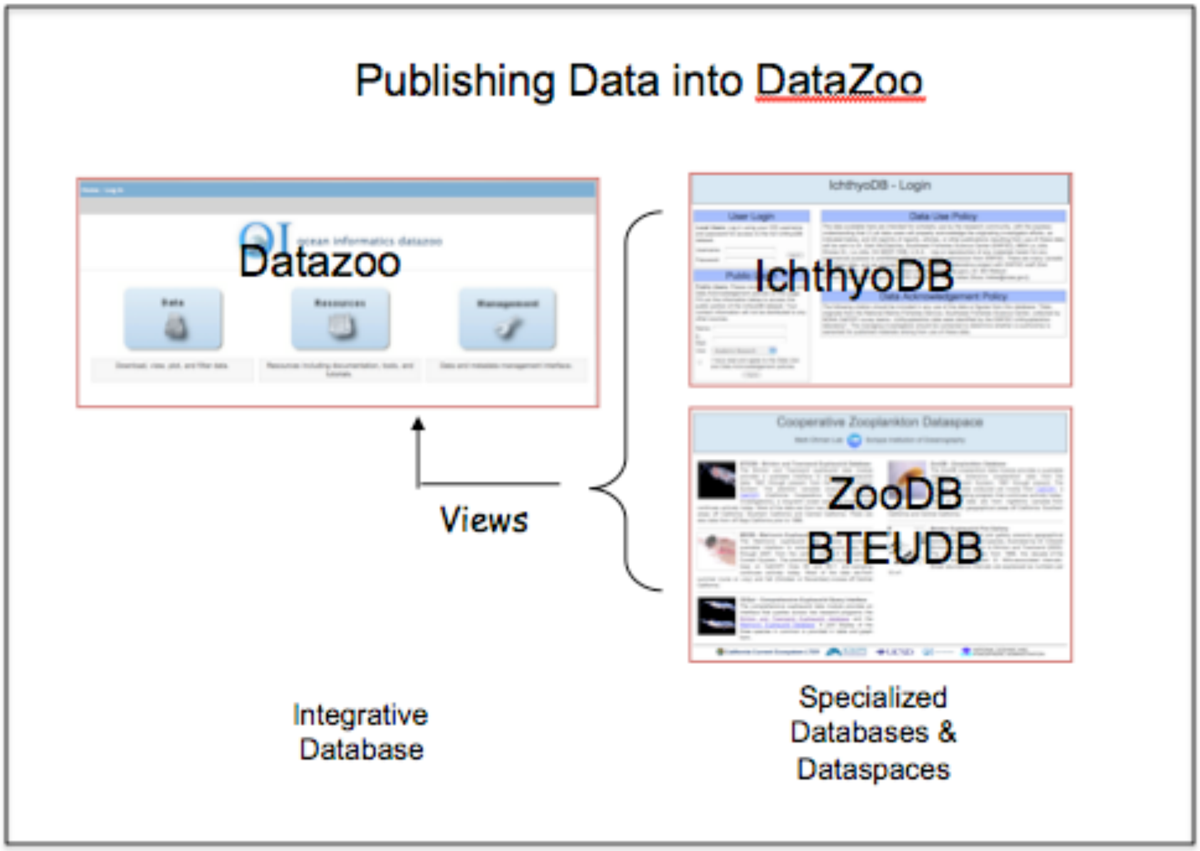
- MetaCAT Data Catalog
- Climate and Hydrology Database
- Network Sites
- Remote Sensing Archive
- EcoTrends Data Portal
- Data Policies

8.7 *Additional Elements*

Additional elements of the DataZoo system and environment include the staging of data in specialized applications prior to publishing in DataZoo and the assembly of a number of collections into a Dataspace.

8.7.1 Publishing Data into DataZoo

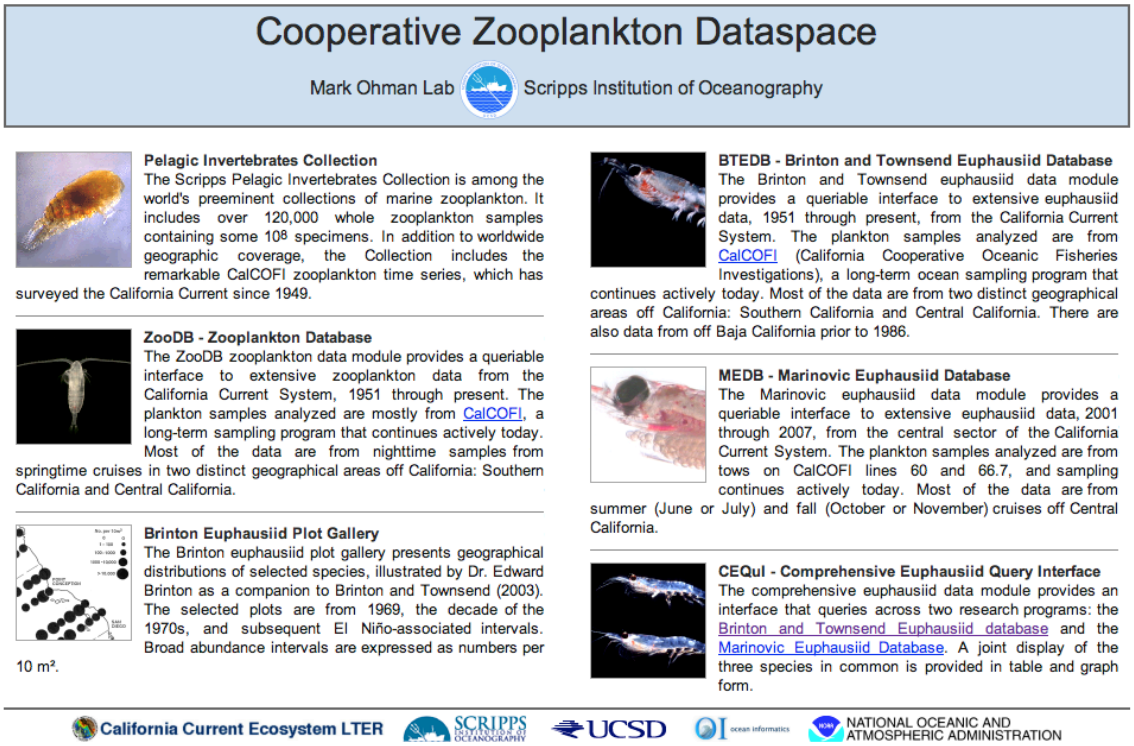
Ocean Informatics has developed several specialized analysis systems that provide access to complex datasets with a level of detail that is not possible in DataZoo. Because DataZoo is the main data catalog for several projects, users may not be aware of these separate, detailed data applications. In order to raise the visibility of these datasets, as well as leverage the capabilities of DataZoo, simplified versions of these complex datasets are published into DataZoo. These simplified datasets are selected by the data provider as useful and representative views into the complete dataset. The metadata for the simplified datasets in DataZoo contain references back to the specialized data systems, providing users with a clear link back to the original source.


Publishing Data into DataZoo
Datazoo
ocean informatics datazoo
IchthyoDB
IchthyoDB - Login
Cooperative Zooplankton Dataspace
ZooDB
BTEUDB
Views
Integrative Database
Specialized Databases & Dataspaces

8.7.2 Zooplankton Dataspace

http://oceaninformatics.ucsd.edu/zooplankton/

Zooplankton links, plots, and databases are gathered together into a data commons known as a dataspace (Franklin, 1995). In the cooperative zooplankton dataspace, three databases are presented: the Ohman Zooplankton database, the Brinton and Townsend Euphausiid database, and the Marinovic Euphausiid database. A Comprehensive Euphausiid Query Interface queries across the two Euphausiid databases. Links to the Pelagic Invertebrates Collection at SIO and a Brinton Euphausiid plot gallery are also posted. Note, in addition, a subset of the data are published in DataZoo (see Appendix 8.7.1 on data publishing).



Cooperative Zooplankton Dataspace

Mark Ohman Lab Scripps Institution of Oceanography

Pelagic Invertebrates Collection
The Scripps Pelagic Invertebrates Collection is among the world's preeminent collections of marine zooplankton. It includes over 120,000 whole zooplankton samples containing some 10^8 specimens. In addition to worldwide geographic coverage, the Collection includes the remarkable CalCOFI zooplankton time series, which has surveyed the California Current since 1949.

ZooDB - Zooplankton Database
The ZooDB zooplankton data module provides a queriable interface to extensive zooplankton data from the California Current System, 1951 through present. The plankton samples analyzed are mostly from CalCOFI, a long-term sampling program that continues actively today. Most of the data are from nighttime samples from springtime cruises in two distinct geographical areas off California: Southern California and Central California.

Brinton Euphausiid Plot Gallery
The Brinton euphausiid plot gallery presents geographical distributions of selected species, illustrated by Dr. Edward Brinton as a companion to Brinton and Townsend (2003). The selected plots are from 1969, the decade of the 1970s, and subsequent El Niño-associated intervals. Broad abundance intervals are expressed as numbers per 10 m^2.

BTEDB - Brinton and Townsend Euphausiid Database
The Brinton and Townsend euphausiid data module provides a queriable interface to extensive euphausiid data, 1951 through present, from the California Current System. The plankton samples analyzed are from CalCOFI (California Cooperative Oceanic Fisheries Investigations), a long-term ocean sampling program that continues actively today. Most of the data are from two distinct geographical areas off California: Southern California and Central California. There are also data from off Baja California prior to 1986.

MEDB - Marinovic Euphausiid Database
The Marinovic euphausiid data module provides a queriable interface to extensive euphausiid data, 2001 through 2007, from the central sector of the California Current System. The plankton samples analyzed are from tows on CalCOFI lines 60 and 66.7, and sampling continues actively today. Most of the data are from summer (June or July) and fall (October or November) cruises off Central California.

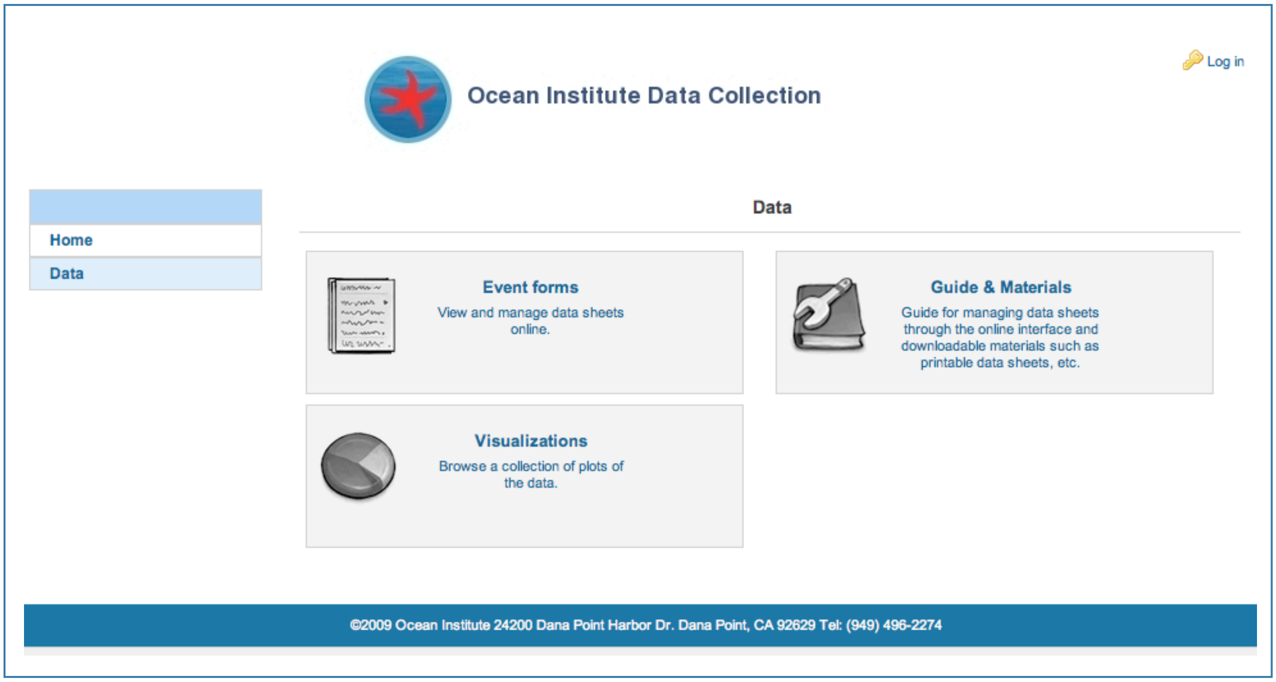
CEQuI - Comprehensive Euphausiid Query Interface
The comprehensive euphausiid data module provides an interface that queries across two research programs: the Brinton and Townsend Euphausiid database and the Marinovic Euphausiid Database. A joint display of the three species in common is provided in table and graph form.

California Current Ecosystem LTER
SCRIPPS INSTITUTION OF OCEANOGRAPHY
UCSD
ocean informatics
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

8.7.3 Ocean Institute

http://oceaninformatics.ucsd.edu/oceaninstitute

This site provides an application for creating and managing online copies of Event data sheets that result from the educational data-collection cruises performed by the Ocean Institute. It is a collaboration between the California Current Ecosystem LTER program and Ocean Institute.



Event forms
View and manage data sheets online.

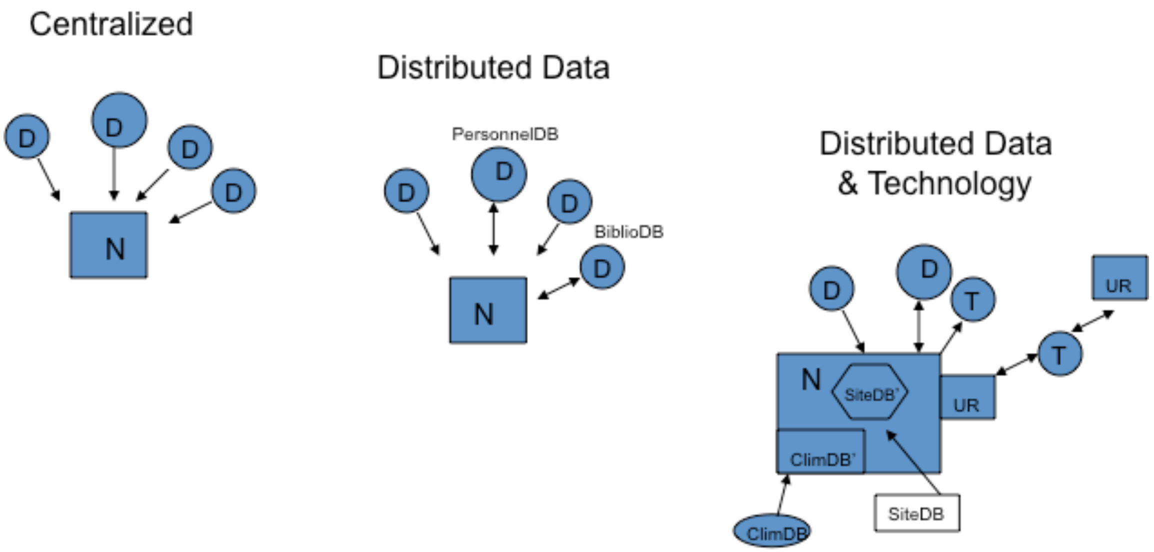
Guide & Materials
Guide for managing data sheets through the online interface and downloadable materials such as printable data sheets, etc.

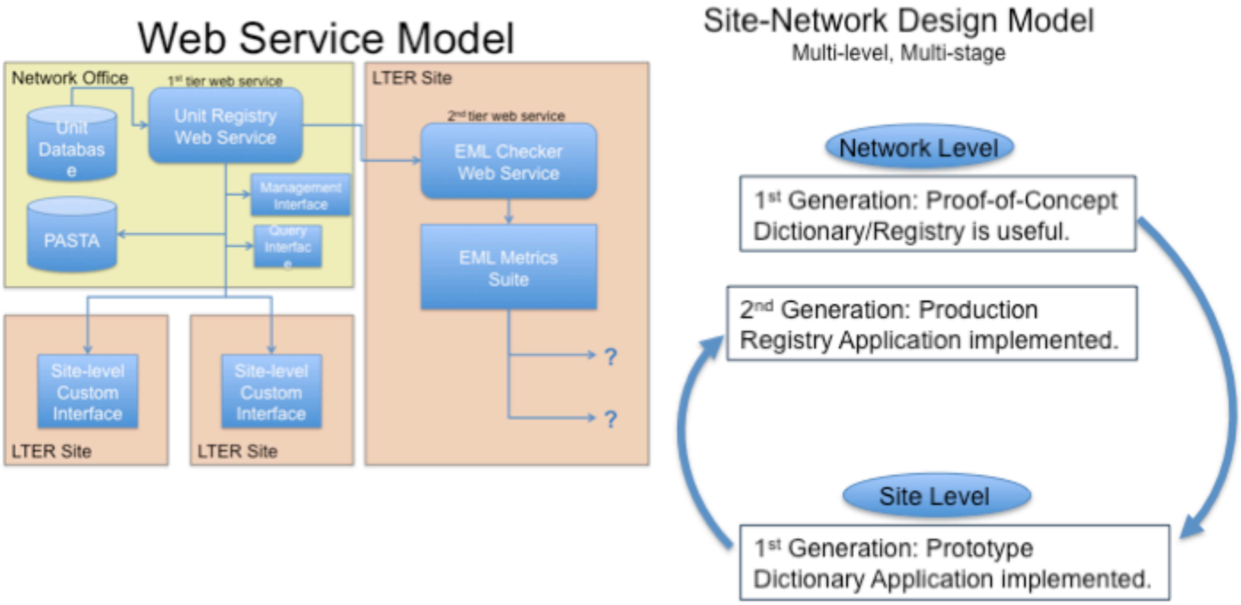
Visualizations
Browse a collection of plots of the data.

©2009 Ocean Institute 24200 Dana Point Harbor Dr. Dana Point, CA 92629 Tel: (949) 496-2274

8.7.4 Site-Network Distributed Model

Our understanding of the site-network distributed model is changing with time. The following two figures attempt to capture the notions of distributed data, distributed data, and the development over time of the DataZoo web to resource-oriented architecture.




Web Service Model
Network Office
1st tier web service
Unit Databas e
Unit Registry Web Service
Management Interface
PASTA
Query Interfac e
LTER Site
2nd tier web service
EML Checker Web Service
EML Metrics Suite
?
?
Site-level Custom Interface
LTER Site
Site-level Custom Interface
LTER Site
Site-Network Design Model
Multi-level, Multi-stage
Network Level
1st Generation: Proof-of-Concept Dictionary/Registry is useful.
2nd Generation: Production Registry Application implemented.
Site Level
1st Generation: Prototype Dictionary Application implemented.

8.8 System Inventories

Maintaining a computational overview of systems, software, and upgrades makes visible the underpinnings of our work environment and also makes evident the work investment required to create and maintain the foundation necessary for information management to be carried out in a professional, production-level manner.

8.8.1 Ocean Informatics Systems Profile 2008

I. System Elements

Type	Element	Current Version	Installation History	Notes
File Sharing Server OS	iOcean	10.5 (Leopard)	From 10.4 (Tiger) Nov08	
Web/Collaboration Server OS	iSurf	10.5 (Leopard)	From 10.4 (Tiger) Dec08	
Database	mySQL	5.0	From 4.1 Dec08	
Language	perl	5.8.8	From 5.8.6 Dec08	
Language	python	2.5	From 2.3 Dec08	
Web server	apache	1.3 – 2.0??		Upgrade to 2.2 Jan08
Language	PHP	4.4.9	From 4.4.7 Dec08	Upgrade to 5.2.6 Jan08
Virtual Machine	VMWare/ XEN	NA	For drupal	

II. Applications

<u>Active</u>

Type	Element	Current Version	Installation History	Notes
Versioning	Subversion	1.4.3	Installed 2006	
DB Admin	CocoaMySQL	varies	varies	
DB Model	MySQL Workbench	varies	varies	Schema visualizations
Wiki	MediaWiki	1.6.10	Installed 2006	Security issue 2008
Wiki	DokuWiki	2007-06-26b	Installed 2007	
CMS	Drupal	6.8	Installed Dec08	Reconfig for common codebase winter/spring09

<u>Deprecated</u>

Type	Element	Current Version	Installation History	Notes
DB Admin	phpMyAdmin	-	Installed 2005, removed Dec08	
Versioning	CVS	-	Removed 2005	
CMS	PostNuke	-	Removed 2006	

III. Protocols

Type	Element	Current Version	Installation History	Notes
Web Service	SOAP	-	Began support Jul08	
Web Service	REST	-	Began support Aug08	

IV. Libraries – modules/packages

Type	Element	Current Version	Installation History	Notes
Plotting	JPGraph	1.26	Installed 2006	Datazoo
Plotting	Matplotlib	0.98	Installed 2008	
Wiki	Text_Wiki	1.2.0	Installed 2007	
Interface	YUI	2.6.0	Installed 2007	
Mapping	Google Maps API		2008	
AJAX	AJAXSLT		Installed 2007	
PHP	PEAR			Datazoo backend

8.8.2 Ocean Informatics Systems Profile 2009

I. System Elements

Type	Element	Current Version	Installation History	Notes
File Sharing Server OS	iOcean	10.5 (Leopard) -08	From 10.4 (Tiger) Nov08	October 2008 October 2009
Web/Collaboration Server OS	iSurf	10.5 (Leopard) -08 10.6 (SnoLeop) -09	From 10.4 (Tiger) Dec08	to OSX 10.6 snow leopard with 64 bit python, saving a lot of recompiling & there's PHP 5.3, which has a lot of nice features.
Database	mySQL	5.0 -08; 5.084-09	From 4.1 Dec08	
Language	perl	5.8.8 -08; 5.10-09	From 5.8.6 Dec08	
Language	python	2.5 -08; 2.6-09	From 2.3 Dec08	
Web server	apache	1.3 – 2.0 – 08 2.2 – oct 09		
Language	PHP	4.4.9 - 08 5.2.6 –jan08; 5.3 – oct 09	From 4.4.7 Dec08	5.3 Oct09
Virtual Machine	VMWare/XEN	NA	For drupal	

II. Applications

<u>Active</u>

Type	Element	Current Version	Installation History	Notes
Versioning	Subversion	1.4.3	Installed 2006	
DB Admin	CocoaMySQL	varies	varies	
DB Model	MySQL Workbench	varies	varies	Schema visualizations
Wiki	MediaWiki	1.6.10	Installed 2006	Security issue 2008
Wiki	DokuWiki	2007-06-26b	Installed 2007	
CMS	Drupal	6.8	Installed Dec08	Reconfig for common codebase winter/spring09
Plotting	GMT			

<u>Deprecated</u>

Type	Element	Current Version	Installation History	Notes
DB Admin	phpMyAdmin	-	Installed 2005, removed Dec08	
Versioning	CVS	-	Removed 2005	
CMS	PostNuke	-	Removed 2006	

III. Protocols

Type	Element	Current Version	Installation History	Notes
Web Service	SOAP	-	Began support Jul08	
Web Service	REST	-	Began support Aug08	

IV. Libraries – modules/packages

Type	Element	Current Version	Installation History	Notes
Plotting	JPGraph	1.26	Installed 2006	Datazoo
Plotting	Matplotlib	0.98	Installed 2008	
Wiki	Text_Wiki	1.2.0	Installed 2007	
Interface	YUI	2.6.0	Installed 2007	
Mapping	Google Maps API		2008	
AJAX	AJAXSLT		Installed 2007	
PHP	PEAR			Datazoo backend

8.8.3 Ocean Informatics Systems Profile 2011

I. Servers

Server	Role	Operating System	Installation History	Notes
iOcean	File sharing server	10.5 (Leopard)	launched 10.3 (2005) 10.4 to 10.5 Nov 08	
vSurf	VM server			
vSurfDev	Development web server	Redhat EL5	launched RHEL5 (2010)	Virtual machine
vSurfWeb	Production web server	Redhat EL5	launched RHEL5 (2010)	Virtual machine
vSurfData	Data server	Redhat EL5	launched RHEL5 (2010)	Virtual machine
iSurf	Web server	10.6 (Snow Leopard)	launched 10.3 (2005) 10.4 to 10.5 (Dec 08) 10.5 to 10.6 (Nov 09)	Retired

II. Programming Languages

Language	Role	Version	Installation History	Notes
PHP	Server-side application/utility	5.3.5	4.4 installed (2005) 4.4 to 5.2 (Dec 08) 5.2 to 5.3 (Oct 09)	
Python	Server-side application/utility	2.4.3	2.3 installed (2007) 2.3 to 2.5 (Dec 08) 2.5 to 2.6 (Nov 09) 2.6 to 2.4 (Dec 10)	Rolled back to 2.3 with iSurf/vSurf migration
Perl	Server-side application/utility	5.8.8	5.8 installed (2005) 5.8 to 5.10 (Nov 09) 5.10 to 5.8 (Dec 10)	Rolled back to 5.8 with iSurf/vSurf migration
JavaScript	Web interfaces	varies	N/A	

III. Server Applications

Application	Role	Version	Installation History	Notes
Apache	Web server	2.2.3	1.3 installed (2005) 1.3 to 2.0 (Dec 08) 2.0 to 2.2 (Oct 09)	
MySQL	DBMS	5.0.77	4.x installed (2005) 4.1 to 5.0 (Dec 08)	
Subversion	Version control	1.6.13	1.3 installed (2006) 1.3 to 1.4 (Dec 08) 1.4 to 1.6 (Dec 10)	
GMT	Geospatial plotting	4.5.3	4.3 installed (2007) 4.3 to 4.5 (Nov 09)	
CVS	Version control	-	Removed 2005	

IV. Client Applications

Application	Role	Version	Installation History	Notes
Sequel Pro	Database administration client	0.9.8.1	N/A	Replaces CocoaMySQL

MySQLWorkbench	Database design	5.2.31a	N/A	
Microsoft Office	Papers/presentations/posters	12.2.8	N/A	

V. Web Applications

Application	Role	Version	Installation History	Notes
DokuWiki	Code documentation	2010-11-07	2007-06-26b installed (2007) 2007-06-26b to 2009-02-14b (Nov 09) 2009-02-14b to 2010-11-07 (Jan 11)	
Wordpress	Blogs	3.0.4	2.x installed (2008) 2.x to 3.0 (Dec 10)	
MediaWiki	Sys admin documentation	1.16.1	1.6 installed 2006) 1.6 to 1.10 (2007) 1.10 to 1.13 (Dec 08) 1.13 to 1.15 (Nov 09) 1.15 to 1.16 (Jan 10)	
phpBB	Sys admin forum Hydro lab forum	3.0.8	2.0 installed (2006) 2.0 to 3.0 (Jan 11)	
Drupal	CMS	-	Removed 2010	Testing phase
phpMyAdmin	Database administration	-	Removed 2008	
PostNuke	CMS	-	Removed 2006	

VI. Libraries

Library	Role	Version	Installation History	Notes
JPGraph	Plotting	3.0.3	1.2 installed (2006) 1.2 to 3.0 (Sep 09)	
Matplotlib	Plotting	0.99.1.2	0.98 installed (2008) 0.98 to 0.99 (Dec 10)	
Yahoo YUI	User interface tools	2.8.2	N/A	
Google Maps	Geospatial plotting	3	N/A	
AJAXSLT	XLST transformation	0.5	N/A	
Perl (various)	varies	varies	varies	
PHP (various)	varies	varies	varies	
Python (various)	varies	varies	varies	

VII. Protocols and Standards

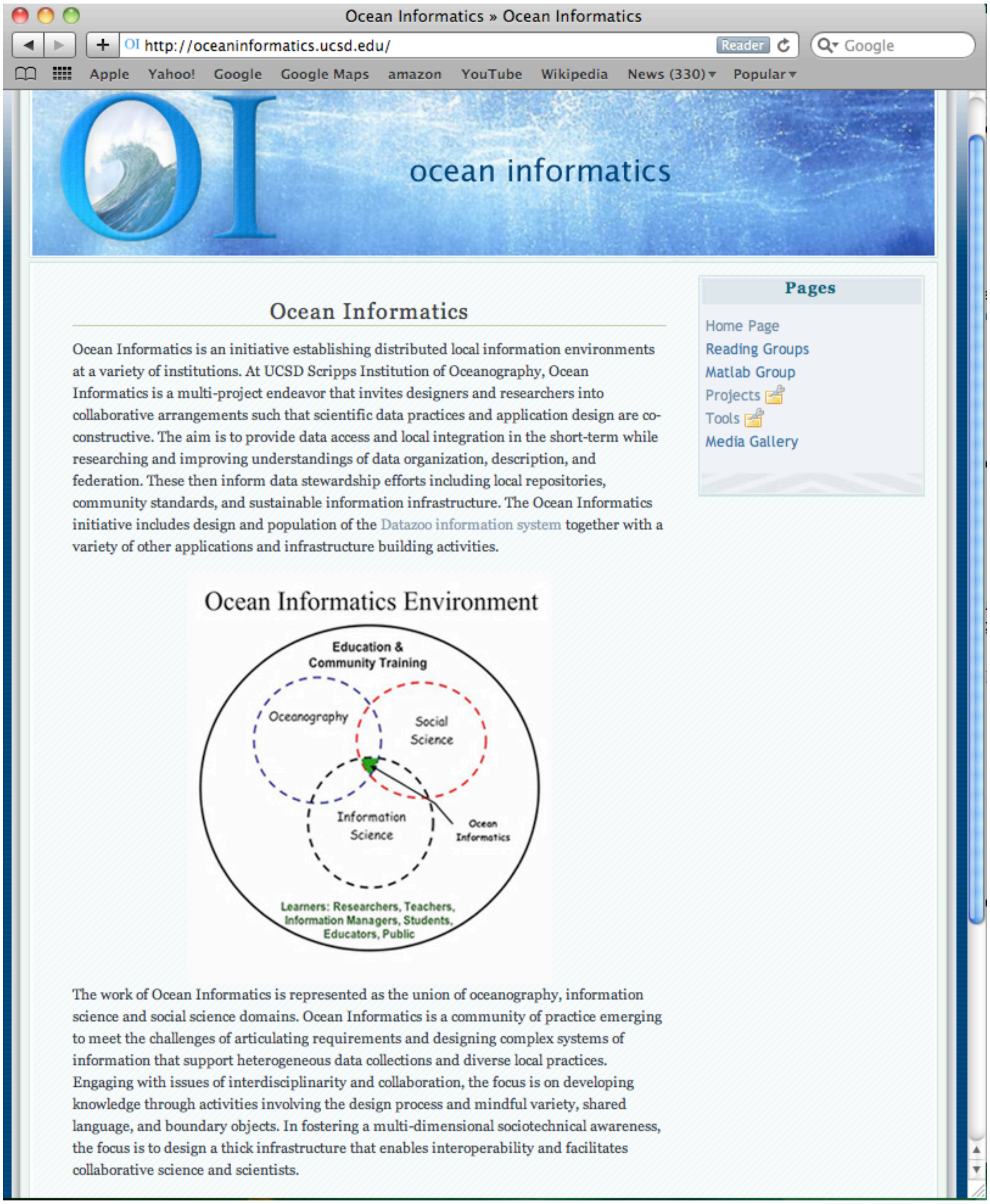
Library	Role	Version	Installation History	Notes
LDAP	Directory service	OpenDirectory 10.6	N/A	Provided by CIS
EML	Metadata exchange standard	2.1.0	2.0.1 to 2.1.0 (Jun 10)	Standardized by KNB
SOAP	Web service standard	N/A	N/A	
REST	Web service standard	N/A	N/A	Adopted Aug 08

8.9 Websites

As discussed in Section 5.6, websites for projects and individuals associated with the Ocean Informatics efforts have been created and supported. Below are some examples.

8.9.1 Ocean Informatics

http://oceaninformatics.ucsd.edu



Ocean Informatics

Ocean Informatics is an initiative establishing distributed local information environments at a variety of institutions. At UCSD Scripps Institution of Oceanography, Ocean Informatics is a multi-project endeavor that invites designers and researchers into collaborative arrangements such that scientific data practices and application design are co-constructive. The aim is to provide data access and local integration in the short-term while researching and improving understandings of data organization, description, and federation. These then inform data stewardship efforts including local repositories, community standards, and sustainable information infrastructure. The Ocean Informatics initiative includes design and population of the Datazoo information system together with a variety of other applications and infrastructure building activities.

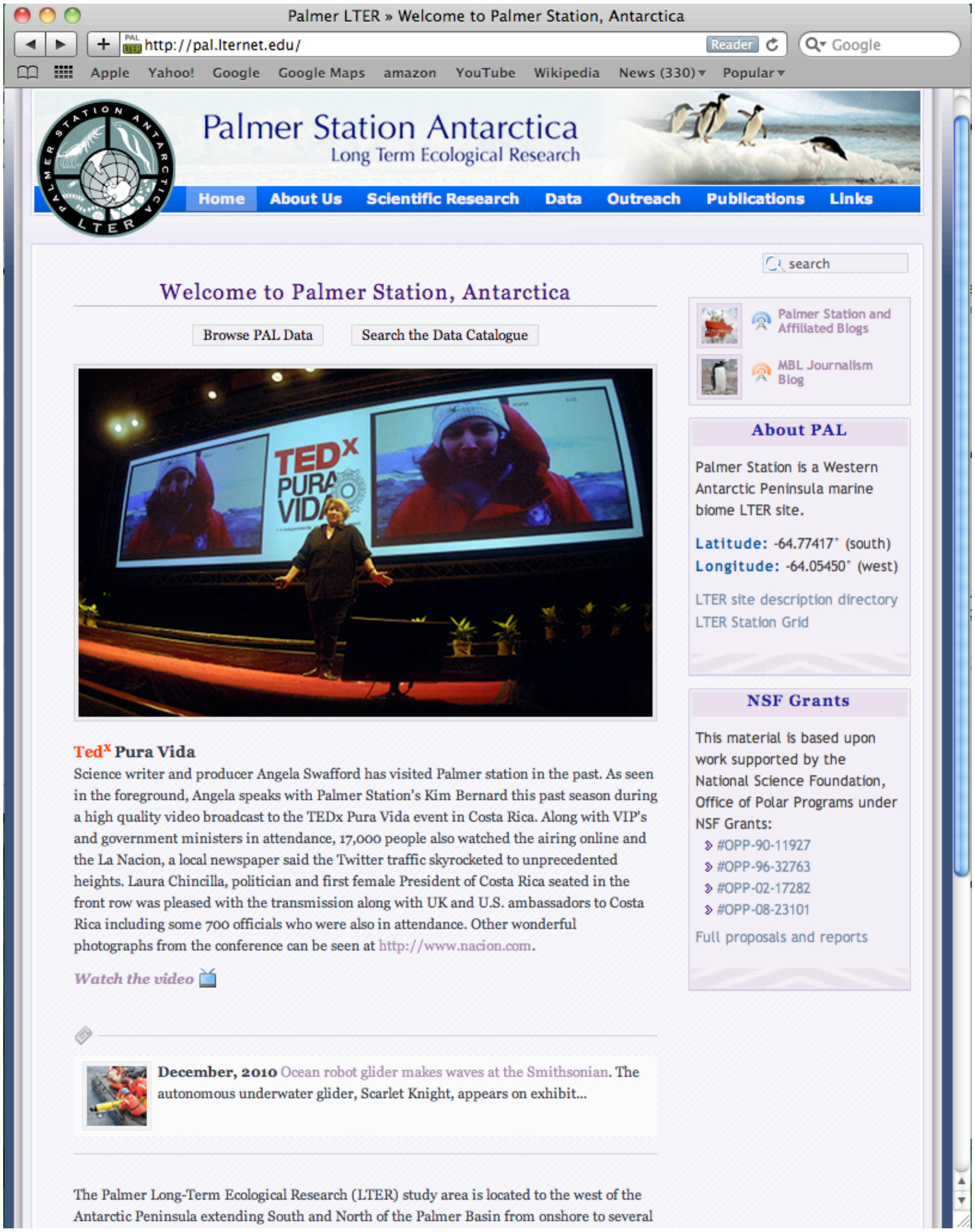
Ocean Informatics Environment
Education & Community Training
Oceanography
Social Science
Information Science
Ocean Informatics
Learners: Researchers, Teachers, Information Managers, Students, Educators, Public

The work of Ocean Informatics is represented as the union of oceanography, information science and social science domains. Ocean Informatics is a community of practice emerging to meet the challenges of articulating requirements and designing complex systems of information that support heterogeneous data collections and diverse local practices. Engaging with issues of interdisciplinarity and collaboration, the focus is on developing knowledge through activities involving the design process and mindful variety, shared language, and boundary objects. In fostering a multi-dimensional sociotechnical awareness, the focus is to design a thick infrastructure that enables interoperability and facilitates collaborative science and scientists.

8.9.2 PAL LTER

http://pal.lternet.edu

Palmer Station LTER site web page including access to data via the 'data' tab. From there, one may navigate to the data catalog that is an entry into DataZoo or via a data portal to DataZoo and the other system components including FileFinder and specialized interfaces.



Welcome to Palmer Station, Antarctica

Ted^{x} Pura Vida

Science writer and producer Angela Swafford has visited Palmer station in the past. As seen in the foreground, Angela speaks with Palmer Station's Kim Bernard this past season during a high quality video broadcast to the TEDx Pura Vida event in Costa Rica. Along with VIP's and government ministers in attendance, 17,000 people also watched the airing online and the La Nacion, a local newspaper said the Twitter traffic skyrocketed to unprecedented heights. Laura Chincilla, politician and first female President of Costa Rica seated in the front row was pleased with the transmission along with UK and U.S. ambassadors to Costa Rica including some 700 officials who were also in attendance. Other wonderful photographs from the conference can be seen at http://www.nacion.com.

December, 2010 Ocean robot glider makes waves at the Smithsonian. The autonomous underwater glider, Scarlet Knight, appears on exhibit...

The Palmer Long-Term Ecological Research (LTER) study area is located to the west of the Antarctic Peninsula extending South and North of the Palmer Basin from onshore to several

About PAL

Palmer Station is a Western Antarctic Peninsula marine biome LTER site.

Latitude: -64.77417° (south)
Longitude: -64.05450° (west)

LTER site description directory
LTER Station Grid

NSF Grants

This material is based upon work supported by the National Science Foundation, Office of Polar Programs under NSF Grants:

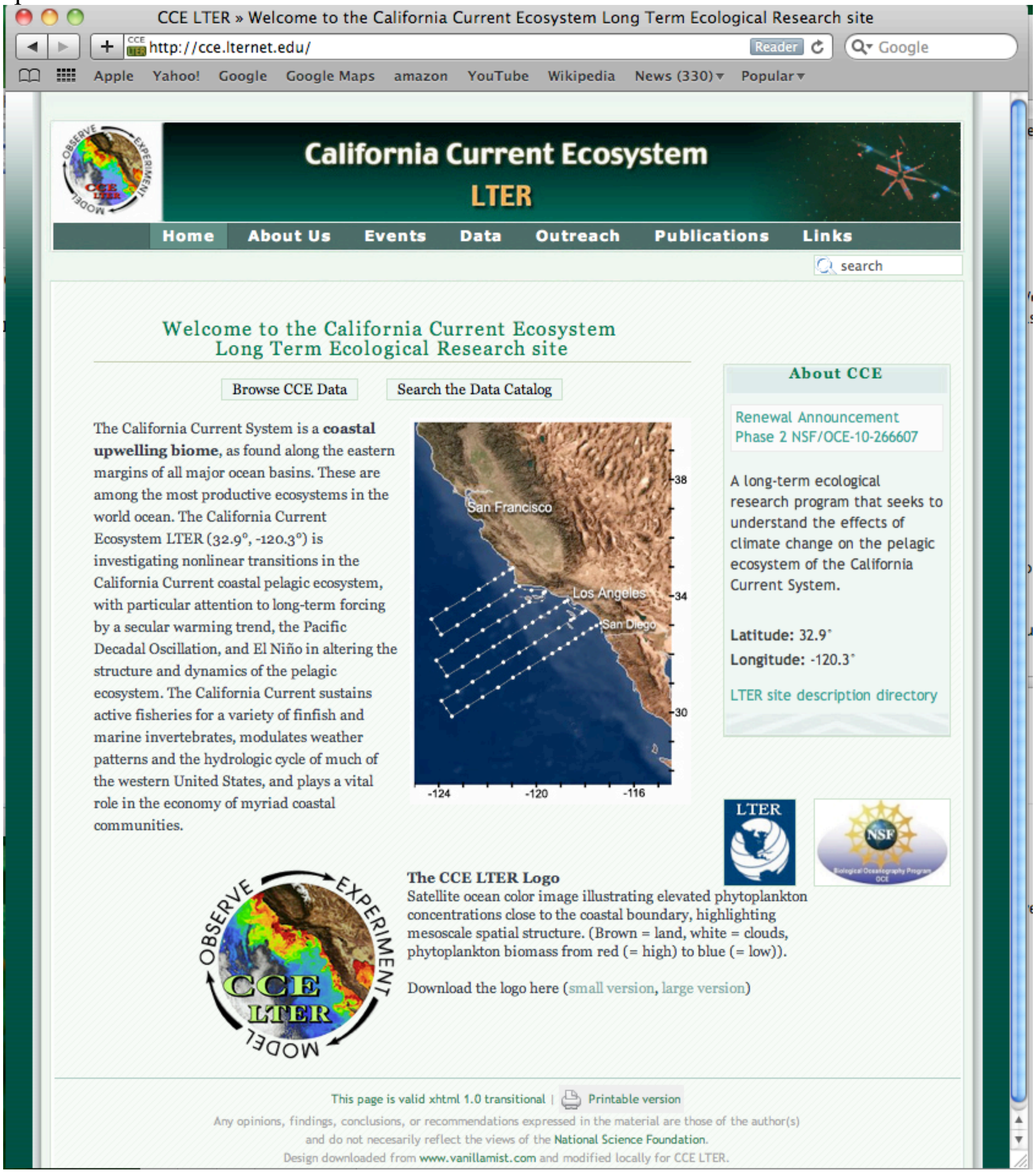
- #OPP-90-11927
- #OPP-96-32763
- #OPP-02-17282
- #OPP-08-23101

Full proposals and reports

8.9.3 CCE LTER

http://cce.lternet.edu

The California Current Ecosystem LTER site web page including access to data via the ‘data’ tab. From there, one may navigate to the data catalog that is an entry into DataZoo or via a data portal to DataZoo and the other system components including FileFinder and specialized interfaces.



Welcome to the California Current Ecosystem Long Term Ecological Research site

The California Current System is a **coastal upwelling biome**, as found along the eastern margins of all major ocean basins. These are among the most productive ecosystems in the world ocean. The California Current Ecosystem LTER (32.9°, -120.3°) is investigating nonlinear transitions in the California Current coastal pelagic ecosystem, with particular attention to long-term forcing by a secular warming trend, the Pacific Decadal Oscillation, and El Niño in altering the structure and dynamics of the pelagic ecosystem. The California Current sustains active fisheries for a variety of finfish and marine invertebrates, modulates weather patterns and the hydrologic cycle of much of the western United States, and plays a vital role in the economy of myriad coastal communities.

About CCE

Renewal Announcement Phase 2 NSF/OCE-10-266607

A long-term ecological research program that seeks to understand the effects of climate change on the pelagic ecosystem of the California Current System.

Latitude: 32.9°
Longitude: -120.3°

LTER site description directory

The CCE LTER Logo
Satellite ocean color image illustrating elevated phytoplankton concentrations close to the coastal boundary, highlighting mesoscale spatial structure. (Brown = land, white = clouds, phytoplankton biomass from red (= high) to blue (= low)).

Download the logo here (small version, large version)

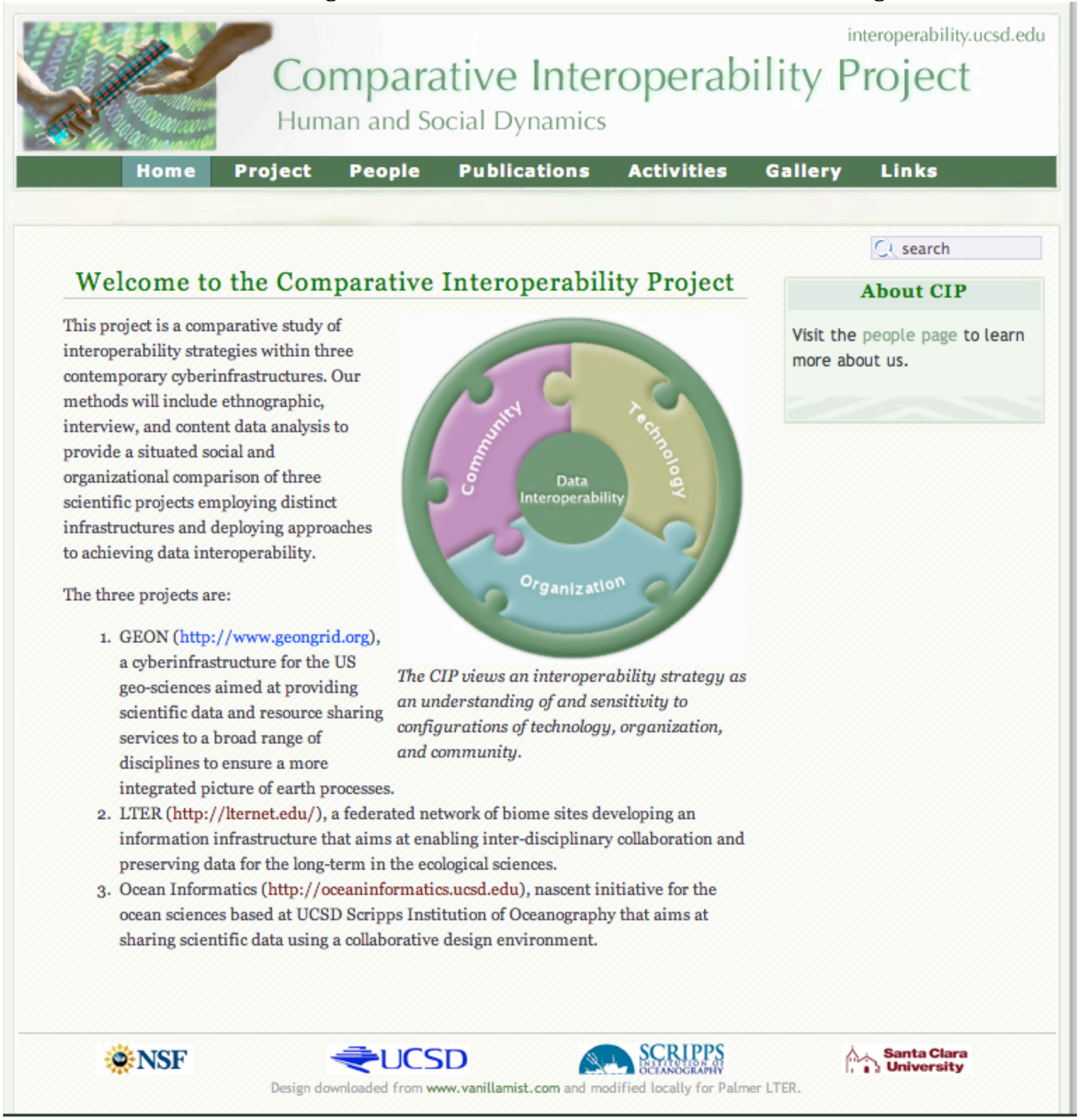
This page is valid xhtml 1.0 transitional | Printable version

Any opinions, findings, conclusions, or recommendations expressed in the material are those of the author(s) and do not necesarily reflect the views of the National Science Foundation.
Design downloaded from www.vanillamist.com and modified locally for CCE LTER.

8.9.4 Interoperability

http://interoperability.ucsd.edu

The interoperability website presents the results of long-term ethnographic studies carried out in collaboration with the LTER and Ocean Informatics. These studies informed our understanding of the role and the work of information management.



Welcome to the Comparative Interoperability Project

This project is a comparative study of interoperability strategies within three contemporary cyberinfrastructures. Our methods will include ethnographic, interview, and content data analysis to provide a situated social and organizational comparison of three scientific projects employing distinct infrastructures and deploying approaches to achieving data interoperability.

The CIP views an interoperability strategy as an understanding of and sensitivity to configurations of technology, organization, and community.

The three projects are:

1. GEON (http://www.geongrid.org), a cyberinfrastructure for the US geo-sciences aimed at providing scientific data and resource sharing services to a broad range of disciplines to ensure a more integrated picture of earth processes.
2. LTER (http://lternet.edu/), a federated network of biome sites developing an information infrastructure that aims at enabling inter-disciplinary collaboration and preserving data for the long-term in the ecological sciences.
3. Ocean Informatics (http://oceaninformatics.ucsd.edu), nascent initiative for the ocean sciences based at UCSD Scripps Institution of Oceanography that aims at sharing scientific data using a collaborative design environment.

About CIP

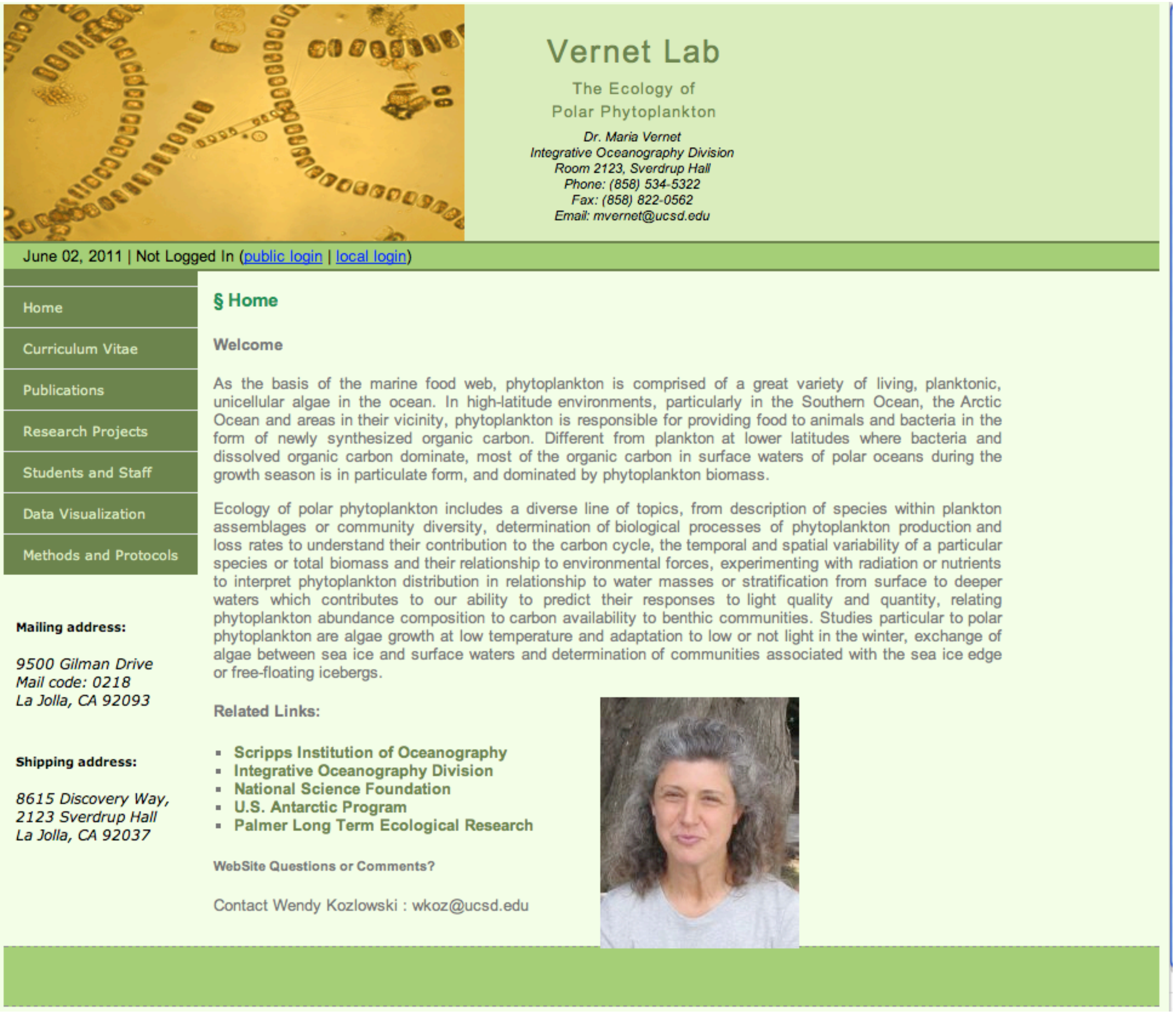
Visit the people page to learn more about us.

Design downloaded from www.vanillamist.com and modified locally for Palmer LTER.

8.9.5 Polar Phytoplankton

http://polarphytophytoplankton.ucsd.edu

This website presents the work of an individual investigator's lab and includes a dynamic data visualization module created in order to provide order and access to hundreds of pre-generated plots.



Vernet Lab

The Ecology of
Polar Phytoplankton

Dr. Maria Vernet
Integrative Oceanography Division
Room 2123, Sverdrup Hall
Phone: (858) 534-5322
Fax: (858) 822-0562
Email: mvernet@ucsd.edu

June 02, 2011 | Not Logged In (public login | local login)

§ Home

Welcome

As the basis of the marine food web, phytoplankton is comprised of a great variety of living, planktonic, unicellular algae in the ocean. In high-latitude environments, particularly in the Southern Ocean, the Arctic Ocean and areas in their vicinity, phytoplankton is responsible for providing food to animals and bacteria in the form of newly synthesized organic carbon. Different from plankton at lower latitudes where bacteria and dissolved organic carbon dominate, most of the organic carbon in surface waters of polar oceans during the growth season is in particulate form, and dominated by phytoplankton biomass.

Ecology of polar phytoplankton includes a diverse line of topics, from description of species within plankton assemblages or community diversity, determination of biological processes of phytoplankton production and loss rates to understand their contribution to the carbon cycle, the temporal and spatial variability of a particular species or total biomass and their relationship to environmental forces, experimenting with radiation or nutrients to interpret phytoplankton distribution in relationship to water masses or stratification from surface to deeper waters which contributes to our ability to predict their responses to light quality and quantity, relating phytoplankton abundance composition to carbon availability to benthic communities. Studies particular to polar phytoplankton are algae growth at low temperature and adaptation to low or not light in the winter, exchange of algae between sea ice and surface waters and determination of communities associated with the sea ice edge or free-floating icebergs.

Related Links:

- **Scripps Institution of Oceanography**
- **Integrative Oceanography Division**
- **National Science Foundation**
- **U.S. Antarctic Program**
- **Palmer Long Term Ecological Research**

WebSite Questions or Comments?

Contact Wendy Kozlowski : wkoz@ucsd.edu

Mailing address:

9500 Gilman Drive
Mail code: 0218
La Jolla, CA 92093

Shipping address:

8615 Discovery Way,
2123 Sverdrup Hall
La Jolla, CA 92037