Baker, KS; Chandler, CL. 2008. Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics. Deep Sea Research Part II. 55 (18-19): 2132-2142. 10.1016/j.dsr2.2008.05.009

Final Draft

Enabling Long-Term Oceanographic Research: Changing Data Practices, Information Management Strategies and Informatics

Karen S. Baker[1] and Cynthia L. Chandler[2]

1 Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA 92093-0218, USA

2 Woods Hole Oceanographic Institution, MS 43, Woods Hole, MA 02543, USA

Abstract
Interdisciplinary global ocean science requires new ways of thinking about data and data management. With new data policies and growing technological capabilities, datasets of increasing variety and complexity are being made available digitally and data management is coming to be recognized as an integral part of scientific research. To meet the changing expectations of scientists collecting data and of data reuse by others, collaborative strategies involving diverse teams of information professionals are developing. These changes are stimulating the growth of information infrastructures that support multi-scale sampling, data repositories, and data integration. Two examples of oceanographic projects incorporating data management in partnership with science programs are discussed: the Palmer Station Long-Term Ecological Research program (Palmer LTER) and the United States Joint Global Ocean Flux Study (US JGOFS). Lessons learned from a decade of data management within these communities provide an experience base from which to develop information management strategies – short-term and long-term. Ocean Informatics provides one example of a conceptual framework for managing the complexities inherent to sharing oceanographic data. Elements are introduced that address the economies-of-scale *and* the complexities-of-scale pertinent to a broader vision of information management and scientific research.

Keywords: Data collections, Data management, Informatics, Information centers, Information systems, Oceanographic data

1.0 Introduction

Interdisciplinary global ocean science requires new ways of thinking about data and data management. This paper is about informatics and information environments providing an organizational structure for information management in collaboration with scientific research. The experience of two oceanographic projects integrating data management with their respective science programs is described below. With data systems and partnerships evolving rapidly, the goal of this paper is to review current approaches and issues at hand in order to open up discussion on the future of data arrangements: sustainable repositories and networked systems, information management strategies and the role of local information environments. This lays a foundation for imagining an information model large enough to encompass a whole earth ecosystem – an infrastructure greater than the sum of its parts, incorporating the dynamics of environmental, human and information systems.

Data and data practices are central to scientific research. Gold (2007a, 2007b) summarized recently: “To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself.” Taking a step back from the local laboratory, field programs, and data collections, we catch a glimpse of a complex system with multiple components including a web of communities intertwined with networks of data systems. This system co-evolves with a variety of partnerships to become an ecology of information (Kling and Scacchi, 1982; Davenport, 1997; Baker and Bowker, 2005). Nardi and O’Day (1999) define an *ecology of information* simply as “an interdependent system of people, practices, values, and technologies in a particular local environment”.

Data management supports field capture, analysis and publication of data. These data processes have become interleaved with issues of digital data preservation, access and exchange. Data previously available to researchers only through journal publications and informal personal exchange can now be made available by submission to data repositories. Digital data collections increase availability beyond a project's original plan or an individual investigator's career. Changes in data access effect changes in expectations by a variety of stakeholders - scientists, educators, technologists, policy-makers and the public to name a few. These changes lead to expanded responsibilities associated with information management. Ideally, information management blends the anchoring of data and data management practices with the theoretical foundations of informatics that draw in contributions of expertise from complementary fields (see Section 4.1).

The following sections describe data use issues (Section 1), case studies (Section 2) and information management strategies (Section 3). Table 1 provides a summary of projects and infrastructure programs mentioned throughout the text.

Table 1
Full names and associated links of acronyms appearing in the text

Acronym	Name	Link
BCO-DMO	Biological and Chemical Oceanography Data Management Office	http://www.bco-dmo.org
CalCOFI	California Cooperative Oceanic Fisheries Investigations	http://calcofi.org
CCE LTER	California Current Ecosystem LTER	http://cce.lternet.edu
EcoInformatics	EcoInformatics.org	http://www.ecoinformatics.org/
EcoTrends	Ecological Trends	http://www.ecotrends.info
ESSI	Earth and Space Sciences Informatics Group	http://www.agu.org/focus_group/essi
FGDC	Federal Geographic Data Committee	http://www.fgdc.gov
GALEON IE	Geo-interface for Atmosphere, Land, Earth, and Ocean netCDF Interoperability Experiment	http://www.opengeospatial.org/projects/initiatives/galeonie
GEON	Geosciences Network	http://www.geongrid.org
GLOBEC	Global Ocean Ecosystem Dynamics	http://www.globec.org
IBP	International Biological Program	http://www7.nationalacademies.org/archives/International_Biological_Program.ht
ISO	International Standards Organization	http://www.iso.org
IOOS	Integrated Ocean Observatory System	
LTER	Long-Term Ecological Research	http://lternet.edu
MMI	Marine Metadata Initiative/Interoperability	http://marinemetadata.org
NCDDC	National Coastal Data Development Center	http://portal.ncddc.noaa.gov
NCEAS	National Center for Ecosystem Analysis and Synthesis	http://nceas.ucsb.edu
NDBC	National Data Buoy Center	http://ndbc.noaa.gov
NEON	National Enviornment Observatory Network	http://www.neoninc.org
NODC	National Oceanographic Data Center	http://www.nodc.noaa.gov
OGC	Open Geospatial Consortium	http://www.opengeospatial.org
OPeNDAP	Open-source Project for a Network Data Access Protocol	http://www.opendap.org
Palmer LTER	Palmer LTER	http://pal.lternet.edu
QARTOD	Quality Assurance of Real-Time Oceanographic Data	http://www.qartod.org
SCOR	Scientific Committee on Oceanic Research	http://www.scor-int.org
THREDDS	Thematic Realtime Environmental Distributed Data Services	http://www.unidata.ucar.edu
US JGOFS	US Joint Global Ocean Flux Study	http://usjgofs.whoi.edu

1.1 In transition: data use and reuse

While data reuse is not a new concept, the scale of reuse has increased. The decision to serve a wider community requires careful data description and organization. Considerable effort may be required to capture complete information about sampling rationale, conditions and methodologies at the data collection stage. And yet, as data travel from those most knowledgeable about their origins and are shared electronically in the absence of customary data exchange methods such as direct personal conversations and scientific peer review, there is an associated increase in the amount and types of description required to explain their context and meaning.

Approaches to studying the oceans are evolving to be more interdisciplinary and global (NRC, 1992, 1993, 1999, 2003). The scope of data management practices is similarly changing to involve both local and global communities as well as to respond to broader scientific questions. Traditional responsibilities for data capture and project-related data use have broadened to Web-based digital data delivery systems. Fig. 1 illustrates the transition from a scenario of local use of data to an augmented arrangement involving additional audiences that constitute reuse communities. This transition necessitates a shift from individual data management to socially complex and highly mediated information management (Star and Ruhleder, 1996; Birholtz and Bietz, 2003; Zimmerman, 2003). New challenges related to local practices emerge when considering larger-scale and longer-term contexts, e.g. organizational behaviors, semantic arrangements, and long-lived collections (e.g. Kling and Jewett, 1994; Sheth, 1999; NSB, 2005).

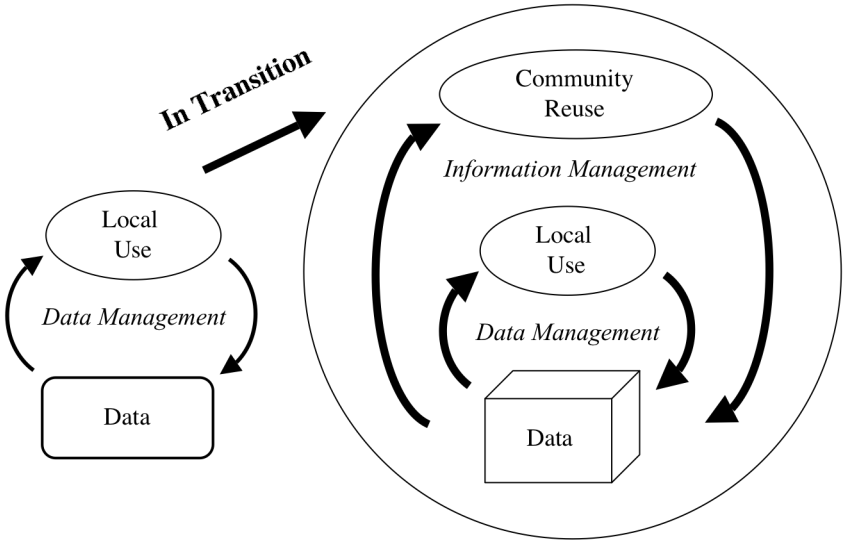

Fig. 1. Scientific data practices are in transition, expanding to include both local data use and community reuse. In this example, data management grows to a community model addressing both local and global information management responsibilities.

1.2 In development: repositories and systems

Digital data systems are designed to improve accessibility to digital collections in data repositories (e.g. local databases), to enable exchange and to ensure data preservation. Information systems to support the ocean sciences have developed over time (Thorley and Trathan, 1994; Baker *et al.*, 2000; Brunt *et al.,* 2002; Chandler, 2004; Glover *et al.*, 2006). The formation of a digital data collection, defined as the product of systematically assembling digital

data from one or more sources for a particular purpose, faces difficulties such as fluidity of digital representations, differences of purpose and diversity or lack of collection membership criteria (Lynch, 2002; Palmer *et al.*, 2006). For example, should a dataset related to a collection in time be included if collected from nearby but outside the designated study area for that collection? Today, informatics promotes partnerships and comparative studies that in turn contribute to development of communities that are 'information aware', that is, cognizant of the significant epistemological and ontological issues associated with interdisciplinary, long-term data efforts (Gold, 2007a, 2007b; Gruber, 1993; Guarino and Welty, 2000; Ribes and Bowker, 2008; Smith, 2003; Smith and Welty, 2001). Information awareness enables community discussion and decision-making with regard to digital collections, data repositories, information system requirements and data policies.

Early data systems developed initially as single package solutions for a specified set of arrangements and a particular audience. Data exchange and analysis were enabled by development of format-specific application standardizations (e.g. netCDF, HDF). With the advent of computer networking, new approaches to data system architecture and to data systems as components of larger systems developed to accommodate a range of situations. New types of exchange mechanisms developed. Table 1 provides some examples that together form a growing information infrastructure: techniques for data exchange (e.g. OPeNDAP, THREDDS, OGC, GALEON IE), discipline-specific national data repositories for data access and availability (e.g. NODC, NDBC, NCDDC), community-specific organizations for data use and data quality (e.g. NCEAS, MMI, QARTOD, Ecoinformatics.org, EcoTrends, ESSI) and international arrangements for developing standards (e.g. ISO).

1.3 In formation: networking and federation

Data flow is often perceived as linear, i.e. data moving from acquisition to repositories to final archives. Fig. 2 shows a traditional hierarchical view of a data source nested within layers of projects, repositories and archives. Data access and reuse exist at points all along 'the line'. In contrast, Fig. 3 portrays an information network as a non-linear, complex system of frequently ill-defined relationships between local repositories and a larger-scale community web of institutional repositories, discipline-specific centers and national archives.

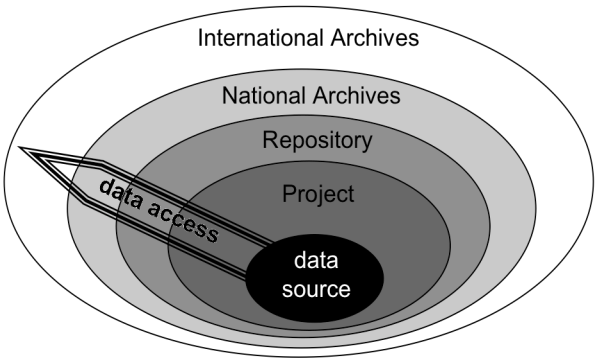

Fig. 2. A nested view of data availability is shown where access occurs at multiple points.

A federation may be defined loosely as a structure that joins together independent entities. Data federation involves collection, system, and network federation. The process of federation involves networking techniques as well as vocabularies and conventions that scale for use across a variety of collections and delivery systems. The proliferation of data collection sites and the desire for their interface highlights the need to define and negotiate their relations. There is a further need to ask the questions: "How are systems federated?", "Who federates the networks?" and "What is required to sustain the federation?".

1.4 In translation: metadata and interoperability

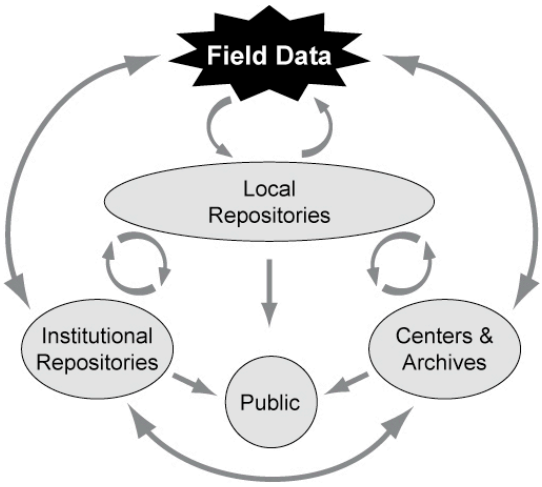

Fig. 3. A local perspective of field data and data repositories is shown in the context of community centers, institutional repositories, national

Heterogeneity is inherent to many types of scientific field data and demands robust metadata description to enable exchange (Goodchild, 1999; Cornillon *et al.*, 2003). Data heterogeneity encompasses a wide range of variations: data sampled according to a variety of criteria in terms of methods and scales; data stored with differing formats, structures and relations; and data processed with differing analytic methodologies and control procedures that have uncertainties commensurate with expected levels of accuracy associated with each step. Thus, even datasets measuring the same physical phenomenon can be disparate. Data similarities *and* data differences are important aspects of scientific work; therefore, accounting for them must be reflected in the corresponding system of information management. Community activities that support data reuse through mitigation of heterogeneous data - creation and refinement of best practices, protocols, dictionaries, ontologies and standards - are gaining recognition.

Data description through metadata (tagged elements describing the data and their context) enables use beyond the originally planned purpose (Michener and Brunt, 2000; Cook *et al.*, 2001). Metadata in a standardized format reduce semantic ambiguities and further enable accurate comparisons. The Federal Geographic Data Committee (FGDC) approved a metadata content standard for geospatial data in 1998. A biological data profile was presented subsequently, but specific guidelines for documenting methods in great detail are lacking in standards. Further, data modifications and the names of those responsible must be captured in metadata. Two metadata concepts capture these aspects of data management: data governance is concerned with documenting who is responsible for data at various stages, and data provenance or lineage is concerned with documenting what has been done to the data and by whom (Greenwood *et al.*, 2003; Simmhan *et al.*, 2005). In studies of complex biotic-abiotic environmental systems, sufficient description to enable accurate data reuse is a metadata grand challenge.

Once data are accessible and well described, they become available for integration, synthesis and interoperability. *Data integration* is a key concept and is frequently used to designate the process of bringing together disparate data through the merging, joining and appending of datasets (Poore, 2003). *Data synthesis* describes the creation of new knowledge achieved through the process of higher-level abstraction. There are cases where the distinction between data synthesis and data integration is arbitrary because there is overlap. Related to the notion of data integration as an activity or process is the concept of interoperability as a state or ability. The IEEE Standard Computer Dictionary (1990) defines interoperability as the ability of two or more systems or components to exchange information and to use the information that has been exchanged.

Recognizing this as a definition of system interoperability, *data interoperability* can be defined as the state of two or more data files being comparable and therefore ready for data integration.

Data interoperability involves a complex matrix of several different types of interoperability. *Semantic interoperability* is a broad term referring to a host of discipline specific issues related to the capture of metadata that are pertinent to data search and data use (Ouksel and Sheth, 1999). Semantics refers to the meaning embedded in the words that comprise the metadata. Interoperability refers to a system's ability to accurately interpret these meanings for purposes of exchange and integration (Ouksel and Sheth, 1999; Sheth, 1999; Friesen, 2002; Cornillon *et al.*, 2003; Cornillon, 2005). *Syntactic* and *structural interoperability* are concerned with the technical aspects of data representation and exchange, such as the organization and format of data and metadata (Visser *et al.*, 2000; Veltman, 2001). Progress towards interoperability has been made in syntactic and structural categories, but semantic interoperability is hindered by differing interpretations of the meaning of words. Fox *et al.* (2007) have demonstrated implementation of semantic web techniques to integrate data from different fields. Data interoperability is often perceived as binary: data either are or are not interoperable (Cornillon *et al.*, 2003). In practice, a continuum exists including cases of data that are almost the same. For instance, data may have the same format and names but may have been acquired using different measurement methods (e.g. two different techniques for measuring biomass or ocean currents).

Although attention and resources have been devoted specifically to the issues of data integration and interoperability, an NRC report (1995) states: "little guidance has been provided on overcoming the barriers frequently encountered in the interfacing of disparate data sets. And although there is a wealth of relevant experience at the working level in the research community, this experience generally has not been analyzed and organized to make it more readily available to researchers."

2.0 Oceanography: science and data

The International Geophysical Year (IGY 1957-1958) was the first of a variety of multi-year and multi-sited global ocean science research projects that have faced the challenges of coordinating data to serve diverse approaches to science. Table 1 provides examples of subsequent projects (e.g. GLOBEC, IBP) and ongoing efforts (SCOR, LTER, NEON, GEON, IOOS). Interdisciplinary research and data synthesis depend upon data organization and data integration as well as the effective use of information technology to facilitate data management and scientific collaboration (NRC, 1993; NSF/AC-ERE, 2003).

The Palmer Station Long-Term Ecological Research program (Palmer LTER) and the United States Joint Global Ocean Flux Study (US JGOFS) provide two examples of oceanographic research programs where data management practices developed in close partnership with a scientific community. An overview of their respective data management efforts highlights experiences from more than a decade of work within a multi-investigator, interdisciplinary culture. Both programs conducted research cruises that featured largely manually-sampled biological and chemical data taken in close coordination with physical oceanographic measurements. Though the two programs progressed independently, common data practices developed.

2.1 Palmer LTER information management

The concept of the LTER Network grew out of the IBP Program (Smith, 1968; Golley, 1993) as a community organization that could address ecological events occurring over multi-decadal timeframes across a variety of ecosystems in a coordinated manner (Hobbie *et al.*, 2003). A national network of study areas was established in 1981 and now includes 26 sites plus a Network Office with each site studying a designated biome. Focusing initially on long-term data and then on regionalization studies, the LTER scientific community designated 2000-2010 as the decade of synthesis. The LTER Information Management Committee (IMC) focus on issues of data management, description and access culminated in 2001 with formal endorsement and adoption of the Ecological Metadata Language (EML) (Jones *et al.*, 2001, 2006). The process of EML implementation has played an important role in providing the LTER community with the ability to conceptualize and address data description (Karasti et al., 2006; Millerand and Bowker, in press). Though the variety and meaning of standards is frequently under-appreciated, the adoption of EML provided experience with standards and the process of standards-making as coordination mechanisms (Star and Lampland, in press; Millerand and Bowker, in press).

Each LTER site has an Information Manager who is a member of the IMC. The IMC is an important forum for communications addressing local as well as cross-site issues (Baker et al, 2000; Karasti and Baker, 2004). It is a Community-of-Practice, a group that meets regularly to discuss issues and to participate in joint activities as a central mechanism for developing common understandings (Lave and Wenger, 1991). Other communication mechanisms include publication of a community information management newsletter with a rotating editorship and the development of conference-style meetings.

Palmer Station, established in 1990 as the first oceanographic LTER site, studies the pelagic marine ecosystem in the Antarctic and the ecological processes that link the extent of annual pack ice to the biological dynamics of different trophic levels (Smith *et al.*, 1995; Ducklow *et al.*, 2006). With the advent of the Internet, data in the form of static text files were posted online (Baker, 1998). A decade later, to meet requests for data queriability and requirements for networking, a new generation information system was designed. The recently launched Palmer information system, DataZoo, features online data access, strategic integration and visualization. Data and metadata management is offered through web interfaces with tiered permissions that enable data provider participation in making their data accessible. The new system is built upon a relational database with an object-oriented API layer that supports Web-based data query. Interdependent sets of dictionaries describe datasets to the column level while databases of term sets and personnel provide a flexible mechanism to capture and make visible information associated with datasets and with the information system itself.

Palmer initiated an informatics focus in 2003 to draw together information theory with practice and developed an information management strategy in partnership with the California Current Ecosystem (CCE) LTER site in 2004 and the California Cooperative Oceanic Fisheries Investigations (CalCOFI) program in 2006. This approach includes design sessions, informatics events and collaboration with science studies partners (Jackson and Baker, 2004; Baker *et al.*, 2005).

2.2 US JGOFS data management

US JGOFS was initiated as a program to understand the global carbon cycle and associated elements in an interdisciplinary view of how the oceans exchanged these elements with the atmosphere, sea floor and continental boundaries (SCOR, 1987; US JGOFS Steering Committee, 1990; NRC, 1999; Buesseler, 2001; Fasham et al, 2001). US JGOFS Scientific Steering Committee members and US NSF Ocean Sciences Division program managers recognized early on that a coordinated, multi-disciplinary, long-term research program would also require a data management strategy that addressed the needs of participating investigators as well as those of the overall program (NAS, 1984; US GOFS Steering Committee, 1986). A US JGOFS data manager was identified in 1988, and a Data Management Office (DMO) with a technical staff was created in 1994. From the beginning, DMO staff members worked together with investigators funded to conduct US JGOFS related projects. The DMO staff coordinated with investigators to define data parameter names that included sampling and analytic methodology described in a UNESCO report (1994). Much of the collaboration focused on issues related to quality control and the collection and subsequent publication of complete metadata for contributed data sets.

All process study data were ingested into an object-oriented, relational database (Flierl *et al.*, 1992; Glover, 2001) and made available via the World Wide Web. Using a standard Web browser client, users of the US JGOFS data system can generate custom data sets that match their research interests by combining multiple data sources 'on-the-fly'. Persistent merged products were created from US JGOFS data by combining all data records from a similar sampling device deployed during all of the cruises. Thus, single integrated products were created for each type of sampling device for each basin studied. The DMO also took responsibility for final contribution of data to NODC as well as for publication of the final data report (United States JGOFS Final Data Report, 2003).

As the US JGOFS research program transitioned from process-oriented field studies to modeling (Sarmiento and Armstrong, 1997), the data system was extended to include a customized Live Access Server (Hankin *et al.*, 1998). Synthesis and model results, larger in volume and often global in scope as opposed to basin-specific (Doney *et al.*, 2002), required a more graphically oriented user interface and extended visualization capabilities (Glover and Chandler, 2001). DMO staff worked closely with investigators to provide timely availability of data during the active research phase and to ensure preservation of the completed data collection as an important part of the JGOFS legacy.

2.3 Data practices in common

Though data management for Palmer LTER and US JGOFS developed separately, common practices can be identified. For both programs, data management was part of the planning process and was recognized as integral to the scientific research process and as requiring close partnership with investigators. Both established centralized local data repositories at the project start and subsequently developed data policies addressing agency, project and institutional

concerns (Data Policy LTER CC; Data Policy US JGOFS). Data catalogs and sampling protocol summaries played an early part in efforts to create centralized data access points.

Sampling grids, event logs and local dictionaries are three coordinating mechanisms that represent best practices common to the two independent research programs. Cooperative planning of cruise sampling strategies initiates cross-component discussions within the community, creates a shared understanding of measurements and informs subsequent data organization. Another product of cooperative planning was a sampling event log with unique sequential identifiers to identify sampling activities during a research cruise. In the absence of an event log, seemingly small differences in how data are gathered in the field (e.g. unsynchronized clocks and differing station-naming conventions) become progressively difficult to reconcile over time. Finally, the complex interdisciplinary investigations that are the hallmark of Palmer LTER and US JGOFS are facilitated by the availability of term dictionaries (see Section 3.2.2). In both programs, custom dictionaries were constructed in order to provide dataset columns with unique, well-defined names and a flexibility that accommodates local naming traditions.

3.0 Information management

With changing data practices as described above, new conceptual frameworks are needed that take into account the heterogeneity of data, complexities of data description and sustainability of community efforts over time. An overarching vision and strategies for information management are presented below. Each framework and strategy contributes in concert with the others to the configuration of information environments described in Section 4.0.

3.1 Data stewardship

Data stewardship – a concern for creation and preservation of data and all the intermediate stages - focuses holistically on the management of data over the long term. It takes into account data flow and transformation, which in turn depend upon choices with respect to data organization, presentation and integrity. Within the stewardship framework, recognition that data are frequently being prepared for a next step influences prioritization with respect to quality, analysis and accountability. Data flow among an assortment of individual repositories within a web of repositories. From a long-term perspective, stewardship involves a suite of interwoven tasks and evolving processes that enable data use and reuse (NSF AC-ERE, 2003; ARL 2006). LTER has been presented as one example of addressing the long-term challenges of data stewardship (Karasti *et al.*, 2006, 2007).

3.2 Information management strategies

Data management experience garnered during Palmer LTER and US JGOFS catalyzed development of methods that represent information management strategies. Twelve strategies have been selected for discussion from past lessons learned (NRC, 1995; Stonebraker, 1994; Strebel *et al.*, 1998; Benson and Olson, 2002; Fugmann, 2004; Glover *et al.*, 2006; Spencer *et al*., 2006). The strategies below are presented in two loose groups based on their implementation (Table 2). Shorter-term strategies may be initiated technically and, at least initially, by a smaller community subgroup. In contrast, long-term strategies frequently involve changes that require initiation within organizational structures or community data practices. Both groups of strategies have long-term timeframes and ramifications.

Table 2
A selection of information management strategies are presented. All strategies have long-term ramifications.

Short-Term Implementation

1. *Local data repository development and maintenance*
2. *Metadata conventions and dictionaries development*
3. *Data access via Web interface to queriable data structure*
4. *Deliberate documentation, articulation and synthesis*
5. *Data quality procedures development*
6. *Online management of data by community members*

Long-Term Implementation

7. *Data policy implementation*
8. *Role development for information mediation*
9. *Collaborative structures and process development*
10. *Design process development for analysis and research*
11. *Reciprocal learning environment development*
12. *Long-term infrastructuring*

Short-term implementation

3.2.1 Local data repository development and maintenance: The role of local repositories is to facilitate data contribution and to start the data description process early on, close to the source of the original data. The local repository focus on targeted scientific research concerns can manifest as local knowledge-building that over time improves the integration of data management techniques into the local research program. Local repositories provide participants the flexibility to consider data in the context of local sampling practices, which may lead to suggested system modifications. Proximity of repository staff to data originators enables dialogue and development of trust through joint planning, shared experiences and collaborative decision-making. Recent database community work broadens the repository concept from databases to dataspaces. Data collections are brought together in loose association through a variety of applications and with the understanding that integration takes time and is rarely accomplished through a single concerted effort. According to Franklin *et al.* (2005), "Dataspaces are not a data integration approach; rather, they are more of a data co-existence approach. … One

of the key properties of dataspaces is that semantic integration evolves over time and only where needed. The most scarce resource available for semantic integration is human attention."

3.2.2 Metadata conventions and dictionaries development: Long-term data use and reuse depend upon complete metadata records for data description and access. Metadata records become more accessible and thorough when tied to controlled vocabularies, shared dictionaries and registered ontologies. The process of fully describing data necessitates development and use of dictionaries, which provide structure for translation of local information into community-wide language. Dictionaries organize metadata, for example local names, associated measurement types and sampling specifics involving methods and units of measurement. Interdependent sets of dictionaries - unit, attribute or parameter and measurement qualifiers - define data to the column level. The goal is to provide sufficient information at the column and dataset levels to allow investigators to assess the value of the data to their research and to incorporate data accurately into customized, integrated products. The stabilization of metadata elements and formats establishes a local foundation for data sharing. Development of local, community, national and international metadata standards is a relatively recent undertaking and involves what sometimes appears to be a dichotomy of efforts: a universal set of standards to coordinate across multiple communities and a local set of conventions familiar to local investigators and labs. These two efforts progress at different rates, the latter more rapidly responsive to local requirements and the former requiring broader coordination and negotiation. The Marine Metadata Interoperability (MMI) project is an example of an organization that hosts community-wide forums, workshops and tutorials (MMI, 2005, 2008) aimed at fostering communication and collaboration.

3.2.3 Data access via Web interface to queriable data structure: Though the Internet permits data access via Web presentation of hierarchical directories of files, a relational information system provides an architecture that allows separation of storage and display and supports queriable interfaces using the cross-community Structured Query Language (SQL). Such architectures allow data requests by cruise, region, dataset or attribute. Further, the combination of unique event numbers and robust metadata records enables generation of merged and integrated data products. The aim of Web-enabled data integration capabilities is to replace labor-intensive manual data integration carried out separately by individual groups.

3.2.4 Deliberate documentation, articulation and synthesis: Documentation is used to convey knowledge about methods and systems as well as goals and strategies. Articulation may be summarized as "bringing awareness of language differences, ramifications of definition and use of categories as well as other coordination mechanisms…[It] is characterized as the interrelating of parts or the alignment of work elements, often involving a range of planning, coordinating and negotiating efforts" (Baker and Millerand, 2007a). In moving from a how-to form of documentation to providing rationale for schema and synthetic materials, data and information are transformed into knowledge that represents something more complex and/or more coherent. Documentation involves names, definitions and categories that constitute classification systems that benefit from local dialogue as well as community exposure. Meta-level insight accompanies the synthetic work of summarizing and assessing that accompanies preparations for oral presentations, newsletters and scholarly forums (Simone *et al*., 1999; see Section 3.2.11). Special informatics events and publication efforts, informal and formal, provide important opportunities to share and record what might otherwise be only tacit and implicit local knowledge.

3.2.5 Data quality procedures development: Data quality assurance (QA) and data quality control (QC) refer to arrangements made prior to or during data acquisition and those made after collection, respectively. The focus of data quality is development, establishment and maintenance of procedures that stabilize data gathering techniques, making note of changes in methods as well as errors in recording. An understanding of data quality exists in explicit, implicit and tacit forms, so locating and recording this information is frequently time consuming. The creation of integrated data products can serve as an important diagnostic tool and a mechanism for reviewing data quality since relations with other datasets can highlight anomalies.

3.2.6 Online management of data by community members: A well-crafted information system with user-friendly interfaces can shift some responsibility for data and metadata management tasks to participants outside the immediate information management team. The aim is to avoid data office staff becoming an obligatory gateway for the flow of data into an information system. Management interfaces are required for data upload and editing. Tiered permission systems allow for data management by defined participant groups, with access granted so that project logistics coordinators manage personnel and bibliographic lists, field team coordinators manage cruise participant lists and event logs and data providers manage data and metadata.

Long-term implementation

3.2.7 Data policy implementation: Policy development represents an important opportunity for scientists and data managers to consider the implications of data reuse and to develop plans for meeting data management goals. Contemporary data policies have been described as representing a 'shift in culture' (Glover *et al.*, 2006). A published data policy that details data contribution requirements, data use and acknowledgement of use serves to align expectations of all members within a community. The data policy gains added significance as funding agencies begin to recognize data access and data sharing as essential to the advancement of science (Arzberger *et al.*, 2004).

3.2.8 Role development for information mediation: Expectations of data access have created shifts in organizational arrangements including the responsibilities, roles and resources relating to management of data. Long-term information infrastructure-building requires a team of information specialists to perform the increased number of liaison and translation functions associated with new interfaces and audiences (Abbott, 1988; Spanner, 2001; Baker and Bowker, 2007). Information mediation includes liaison and translation work associated with the data, project science and technology. Information Managers facilitate communications that bridge data practices and informatics and are central to developing community information management procedures. A few examples of information management liaison work include exploration of information system functionality with respect to participant needs, creation of naming conventions understandable by technical staff and science participants alike, and participation in cross-project metadata and dictionary endeavors. When informatics is an integral component of long-term, data-intensive projects, an information management team with design skills combined with local knowledge can facilitate the selection of new technologies.

3.2.9 Collaborative structures and process development: In progressing beyond *ad hoc* collaborations, there are research fields that address the theory and practice of cooperative work, e.g. participatory design, computer supported cooperative work and infrastructure studies (Schuler and Namioka, 1993; Grudin, 1994; Sandusky, 2003; Bowker *et al*, in press). Nested interest groups such as information management Communities-of-Practice are components of a structure that support collaboration. Organizational arrangements such as inclusive communication lists, planning meetings, problem solving, budgeting and decision-making also have significant ramifications for collaboration. In a recently formulated set of criteria for LTER site information management, periodic reviews of data management at each site are recommended as a way of ensuring that time is scheduled for interactive planning (LTER IMC, 2005). Engagement of community members in local information management discussions provides the experience required to address local needs as well as larger cross-community efforts related to development of standards (Lampland and Star, in press). Maintaining a standard is an ongoing process of collaboration and renegotiation as local and global understandings of data, scientific issues and semantics change.

3.2.10 Design process development for analysis and research: Information systems design is a creative activity that involves the ability to capture and relate data processes, information systems and infrastructures as well as community standards and coordination mechanisms. The design process begins with problem formulation. After framing, the process continues with identifying objectives, developing a strategy and analyzing results. Each phase of the design process generates products and benefits from involvement of participants (Schon, 1987; Schmidt and Simone, 1996; Bratteteig, 2003, Kanstrup, 2005; Giaccardi and Fischer, 2007). Products may include a unit repository or a media gallery, a Web interface for data query or an application programming interface. The study of information systems design is a mechanism for seeding discovery and enriching scientific work (Friedman, 1989; Khazanchi and Munkvold, 2000; Fischer and Ostwald, 2002; Whitman and Woszczynski, 2004). Data management provides an immediate service in terms of local data capture and analysis, while a design perspective provides information management insight into approaches to data heterogeneity, to local solutions that accommodate data exceptions and to bridging the local with larger-scale data structures. A design process that involves information managers recognizes the heterogeneity and anomalies inherent to ecosystems and hence to ecosystem measurements, not as barriers to data integration but rather as design challenges. These challenges demand innovative formulations to address technical constraints and representational limitations inherent to investigation of a dynamic, living world.

3.2.11 Reciprocal learning environment development: Information professionals working closely with data originators ensure that datasets and information systems meet the immediate needs of a research program. New ways of describing data and changing data practices necessitate an *information readiness* on the part of data collectors for identifying cross-community differences in the meaning of terms and categories. The routine use of information system "demos" with individual use cases presented in the context of community development creates an opportunity for important informal dialogue. These are design sessions that encourage discussion among participants and contribute to the development of shared understandings. Intra-community engagement is critical to the process of adapting to new technologies and to changing research interests. Fox *et al.* (2006) emphasize the importance of "use cases" to encourage partnerships

when designing semantically enabled scientific data repositories. A recognized organizational strategy is to encourage continuing learning by supporting community relationships, participant engagement and on-going local prototyping.

3.2.12 Long-term infrastructuring: Information infrastructure refers to the facilities, the services and resources that support digital work, while infrastructuring refers to the activities involved in the creation and maintenance of an infrastructure. Infrastructure may be recognized as having interdependent technical, organizational and social components intertwined with temporal aspects. It includes individuals and communities designing, building, using, maintaining and redesigning the elements associated with data, human and information systems together with their interfaces (Atkins, 2003; Ribes, 2006; Bowker *et al*., in press). On reflecting upon the first three years of a multi-year, interdisciplinary earth science-computer science project, Stonebraker (1994) described infrastructure as necessary, time-consuming and very expensive. Recently, cyberinfrastructure, the infrastructure associated with large-scale digital endeavors, has been described as a process with a history, a workforce and a unique place in the information landscape (Jackson *et al*., 2007; Edwards *et al.*, 2006). Science, data and infrastructure have been presented as 'growing' together, and local information infrastructure has been described as 'thick infrastructure' (Jackson and Baker, 2004) when the human and technical are recognized as co-constituting each other (Bijker *et al*, 1987; Fischer and Ostwald, 2002; Star and Bowker, 2002). And while there is increasing focus on cyberinfrastructure for large-scale endeavors, the question of local information infrastructures remains under-explored.

4.0 Information environments

Central to scientific environments are member agreements about overarching goals coupled with community planning and shared core activities. An information environment is a structure that provides continuity for data practices and establishes an information management strategy that fulfills the vision of data stewardship. A local scientific environment today requires support from both local and global information environments, each supported by local and global infrastructures. Local participants benefit from an information environment's resources including project bibliographies, shared dictionaries, integrated datasets, communication forums and accumulated expertise.

A local information environment acts as an arena for ongoing design and continued mutual learning. The challenge and intellectual excitement of representing the natural world in digital form and of developing and maintaining that representation over time requires new types of information arrangements that are simultaneously being utilized, redesigned and modified. Though technological advances frequently drive change, an effective information environment provides a critical mass of personnel with community insight who are able to investigate, evaluate and incorporate appropriate technology-related options while providing local continuity through informed decision-making. Traditional training includes classes and technical conferences, but there are a host of additional options such as cross-project design sessions, partnerships with science study programs or information schools and mentoring of design projects. Some information environments offer opportunities for submission of posters, papers and proposals aimed at addressing local information issues. Participant training is needed to sustain a design-oriented information environment but equally important are opportunities for

undertaking 'inquiry-based' or 'research-based' learning. Communities-of-Practice provide a point of educational engagement for information professionals, an informal substrate stimulating professional growth and leadership as well as reconceptualization and innovation.

An information environment fosters a collective mindfulness about the continuity of information management elements within a scientific community; it ensures that organization of data and design of information systems are situated as part of the scientific process. An information environment is characterized by openness, an environment organized for self-assessment and self-reporting of flaws and errors. A local environment provides participants a safe harbor for open discussions about difficult issues including failures in interface design, system architecture and data integration (Lyytinen and Hirschheim, 1987; Weick *et al.*, 1999). A fully functioning environment creates a venue for engagement of scientists in partnership with information professionals. Participants are engaged in the decision-making process about data, informatics and infrastructure issues as part of the everyday scientific environment. Finally, an information environment provides a long-term framework in terms of readiness: the readiness of participants to co-design and to use community systems as well as the readiness of data for integrative and synthetic activities.

4.1 Informatics

Informatics occurs at the intersection of information science, social science and a particular research field such as ocean science. It brings together the theory and practice of information management in meeting the needs of a particular research community. One goal of informatics is to generate data products in order to make data available for scientific use according to mutually agreed upon requirements and to initiate the community processes that build capacity for data interoperability and system federation. Another goal is to generate information infrastructure –technical and collaborative.

In the United States, "informatics" is used in a variety of senses often associated in general with an ecology of information. It includes elements of information systems science, library science, computer science and technology as well as societal interactions with each. As a research field, informatics strives to observe the processes inherent to its application to a particular scientific field. Design and articulation are research undertakings as well as techniques central to an informatics approach (Jackson and Baker, 2004; Baker and Millerand, 2007a). An informatics approach is also concerned with human factors associated with differences in rates of community conceptual readiness (Kaplan and Seebeck, 2001) and change

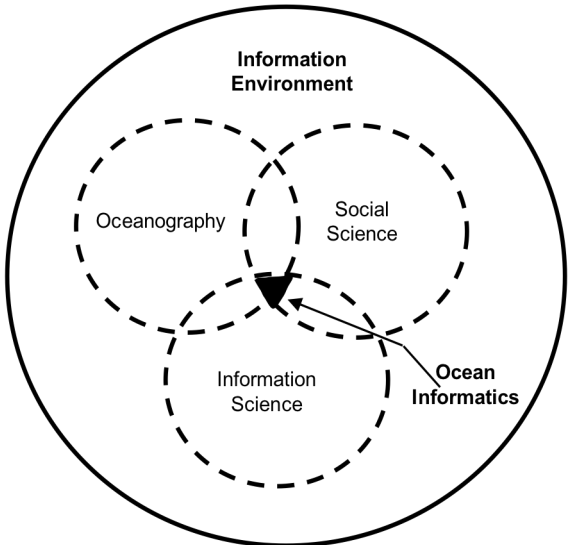

Fig. 4. Ocean Informatics Environment: at the union of oceanography, information sciences, and social sciences is shown a triangular join (solid fill) representing the arena where the work of ocean informatics occurs.

factors such as those associated with management of the unexpected (Weick, 2001).

4.2 Ocean Informatics

Ocean Informatics is the application of informatics to ocean science (Fig. 4) (Baker et al, 2005). The goal of Ocean Informatics is to create local information environments that support the partnership of science and informatics. Ocean Informatics provides a framework within which the concepts introduced above - federation, data stewardship, information management strategies, information environments, and informatics - are assembled in support of oceanographic research over the long term. Another goal is to create an infrastructure that offers collaborative solutions and engages members of the community in co-design.

Ocean Informatics provides an approach that enables learning and communication through establishment of a local information environment close to the source of the data. The work of building repositories prepares data and people, ensuring robust data collections and facilitating interdisciplinary research through increased awareness of data practices and information issues. The local work complements other efforts such as institutional repositories. The variety of repository types are all synergistic but focus on different aspects of the data: local information environments associated with field programs, institutional repositories supported by universities and professional discipline-specific associations, and data archives representing national and international efforts. The concept of data stewardship provides a long-term understanding of data organization across all aspects of the network. There is need for deep scholarly and interdisciplinary research to address the potential ambiguities of both language and methods associated with heterogeneous data, especially when aiming to develop comprehensive and automated approaches to data processing, delivery and preservation through networks.

5.0 Concluding remarks

Palmer LTER and US JGOFS evolved independently as programs but developed data management practices in common that include development of data management systems, dictionaries and metadata conventions. Both programs have continued to evolve in response to changing long-term visions of information management and the needs of interdisciplinary global science. In 2004, Palmer LTER information management began partnering with other projects and programs starting with the CCE LTER. In late 2006, members of the formerly independent US JGOFS DMO and US GLOBEC DMO were funded jointly to form the Biological and Chemical Oceanography Data Management Office (BCO-DMO) to offer data management support for individual investigators as well as investigators associated with larger projects. These initiatives represent contemporary approaches to information management that incorporate informatics concepts and benefit from the efforts of groups representative of larger communities (see Table 1).

Data exchange methods, data integration and metadata standards are all under active development as are the concepts of data federation and data stewardship. Responsible project management must respect the need to develop flexible information systems but must also recognize the necessity for broader frameworks supporting long-term oceanographic research. Ocean Informatics is an information environment that provides such a framework. The field of informatics incorporates a design approach that includes infrastructuring within the context of

local and global information environments and thereby supports ongoing maintenance, implementation and dynamic redesign of information systems that meet both local and global needs. The twelve strategies for information management (Section 3.2) represent mechanisms that within the framework of local information environments support the processes required to address the complexities of data federation and data stewardship.

Local environments exist within a growing web of communities, data system networks and diverse partnerships. In contrast to the notion of *economies-of-scale* for pipelines of data in linear systems with reduced cost of output related to an increased volume of output, an ecology of information is characterized as having *complexities-of-scale* due to data heterogeneity, semantic relations and interdisciplinary collaboration. An informatics approach within an information environment aims to create a well-designed information system architecture buttressed by metadata to help investigators reduce ambiguity in constructing digital records that approximate the natural world.

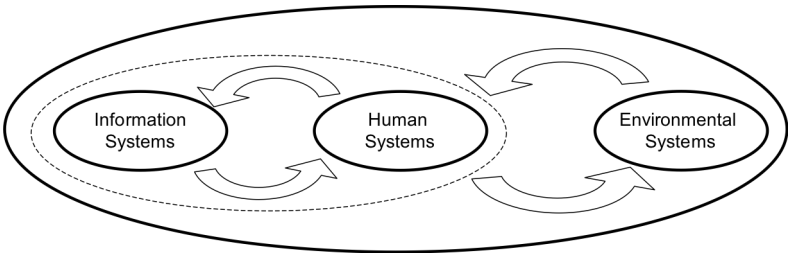

Fig. 5. Representation of the whole earth as an ecosystem, a system of complex systems, taking into account information, human, and environmental dimensions.

As long-term, interdisciplinary researchers recognize and incorporate interconnections between human and environmental systems, informatics assists the transition from what has been called the 'Machine Age' into the 'Systems Age' (Ackoff, 1974). We expand the systems concept to include a federation of distributed repositories and larger scale information systems. Drawing on long-term views of the community (NSF AC-ERE, 2003; Waltner-Toews *et al.*, 2003; LTER CC, 2007), an ecosystem model is presented as inclusive of both natural and human dimensions. Reconceptualizing the system to include information systems explicitly creates a third component to the whole earth ecosystem model (Fig. 5). Modeling an environmental ecosystem as a closed system with defined inputs and outputs is a complex scientific enterprise; modeling a whole earth ecosystem with three components as an open system with emergent characteristics promises to be even more challenging. However, through responsible stewardship of well-described data, the effort to represent the whole earth system - including all its human, environmental and information component systems - opens up endless possibilities for understanding our world.

Acknowledgements

Support is provided by NSF OPP-0217282, OCE-0405069, HSD-0433369 and Scripps Institution of Oceanography (K.S.Baker) and by NSF OCE-8814310, OCE-0097291, OCE-0510046 and OCE-0646353 (C.Chandler). Thanks are given to L.Yarmey for valuable contributions. The manuscript was improved by reviewer input, and the authors acknowledge their significant contributions.

References

Abbott, Andrew, 1988. The System of Professions: An Essay on the Division of Expert Labor, The University of Chicago Press, Chicago.

Ackoff, R., 1974. Redefining the Future. Wiley, New York.

ARL, 2006, To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships, September 26-27, Arlington, VA. The Role of Academic Libraries in the Digital Data Universe. ARL, Washington, DC.

Arzberger, P., Schroeder, P., Beaulieu, A, Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. and Wouters, P., 2004. An international framework to promote access to data. Science 303(5665), 1777-1778.

Atkins, D., 2003. Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Technical Report.

Baker, K. S., 1998. Palmer LTER information management. In: Michener, W., Porter, J. and Stafford, S. (Eds.), Data and Information Management in the Ecological Sciences: A Resource Guide. University of New Mexico, pp. 105-110.

Baker, K.S., Benson, B., Henshaw, D.L., Blodgett, D., Porter, J. and Stafford, S.G., 2000. Evolution of a multi-site network information system: the LTER information management paradigm. BioScience 50(11), 963-983.

Baker, K.S. and Bowker, G., 2007. Information ecology: open system environment for data, memories and knowing. Journal of Intelligent Information Systems 29(1), 127-144.

Baker, K.S., Jackson, S.J. and Wanetick, J.R., 2005. Strategies supporting heterogeneous data and interdisciplinary collaboration: Towards an Ocean Informatics Environment. Proceedings of the 38th Hawaii Annual International Conference on System Sciences (HICSS'05), 03-06 Jan. 2005, IEEE Computer Society, Washington, DC, 219.2.

Baker, K.S. and Millerand, F., 2007a. Articulation work supporting information infrastructure design: coordination, categorization, and assessment in practice. Proceedings of the 40th Annual

Hawaii International Conference on System Sciences (HICSS'07), January 03-06, 2007, IEEE Computer Society, Washington, DC, 242a.

Baker, K.S. and Millerand, F., 2007b. Scientific infrastructure design: information environments and knowledge provinces. Conference Proceedings of the 70th Annual Meeting of the American Society of Information Science and Technology (ASIS&T), 19-24 October, Milwaukee, WI, Vol. 44.

Benson, B.J. and Olson, R.J., 2002. Conducting cross-site studies: lessons learned from partnerships between scientists and information managers. Bulletin of the Ecological Society of America 83(3), 198-200.

Bijker, W.E., Hughes, T.P. and Pinch, T.J., 1987. The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. MIT Press, Cambridge.

Buesseler, K.O., 2001. Ocean biogeochemistry and the global carbon cycle; an introduction to the US Joint Global Ocean Flux Study, Oceanography **14** (4), 5.

Birnholtz, J.P. and Bietz, M.J., 2003. Data at work: supporting sharing in science and engineering. Conference on Supporting Group Work. Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work, Sanibel Island, Florida, November 09-12. ACM, New York, NY, pp. 339-348.

Bowker, G., Baker, K.S., Millerand, F. and Ribes, D., in press. Towards information infrastructure studies: ways of knowing in a networked environment. In: Hunsinger, J., Allen, M. and Klasrup, L. (Eds.), International Handbook of Internet Research, Springer.

Bratteteig, T., 2003. Making change: Dealing with relations between design and use. PhD Thesis, University of Oslo, Faculty of Mathematics and Natural Sciences, Department of Informatics, Oslo.

Brunt, J.W., McCartney, P., Baker, K.S. and Stafford, S., 2002. The future of Ecoinformatics. Long Term Ecological Research. Proceedings of the 6th World Multi-Conference on Systematics, Cybernetics and Informatics, 14-18 July 2002, Orlando, FL. edited by N. Callaos, J. Porter and N. Rishe. IIS 7: 367-372.

Chandler, C.L., 2004. US JGOFS Data Management: Then and Now. US JGOFS Newsletter 12(4), November.

Cook, R.B., Olson, R.J., Kancruk, P., and Hook, L.A., 2001. Best practices for preparing ecological data sets to share and archive. ESA Bulletin 82, 138-141.

Cornillon, P., 2005. What is a Data System, Anyway? Educause Review 40(2), 10-11.

Cornillon, P., Gallagher, J. and Sgouros, T., 2003. OPENDAP: Accessing data in a distributed, heterogeneous environment. Data Science Journal 2, 164-174.

Data Policy, LTER CC (Coordinating Committee), 2005. LTER Network Data Access Policy, Data Access Requirements, and General Data Use Agreement. http://www.lternet.edu/data/netpolicy.html

Data Policy, US JGOFS Data Policy, online. http://usjgofs.whoi.edu/jgofs-data-policy.html

Davenport, T.H., 1997. Information Ecology. New York, Oxford University Press.

Doney, S. C., Kleypas, J.A., Sarmiento, J.L. and Falkowski, P.G., 2002. The US JGOFS synthesis and modeling project - an introduction. Deep Sea Research Part II: Topical Studies in Oceanography 49(1-3), pp. 1-20.

Ducklow, H. W., Baker, K.S., Martinson, D.G., Quetin, L.B., Ross, R.M., Smith, R.C., Stammerjohn, S.E., Vernet, M. and Fraser, W., 2006. Marine pelagic ecosystems: the West Antarctic Peninsula. Philosophical Transactions of the Royal Society of London, (Special Theme Issue, "Antarctic Ecology: From Genes to Ecosystems") 362, 67-94. [doi:10.1098/rstb.2006.1955]

Edwards, P.N., Jackson, S.J., Bowker, G.C. and Knobel, C., 2006. Understanding infrastructure: dynamics, tensions, and design. Report of an NSF Workshop on History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures. http://www.si.umich.edu/InfrastructureWorkshop/documents/UnderstandingInfrastructure2007.pdf

Fasham, M.J.R., Baliño, B.M. and Bowles, M.C., Anderson, R., Archer, D., Bathmann, U., Boyd, P., Buesseler, K., Burkill, P., Bychkov, A., Carlson, C., Chen, C.T.A., Doney, S., Ducklow, H., Emerson, S., Feely, R., Feldman, G., Garcon, V., Hansell, D., Hanson, R., Harrison, P., Honjo, S., Jeandel, C., Karl, D., Le Borgne, R., Liu, K.K., Lochte, K., Louanchi, F., Lowry, R., Michaels, A., Monfray, P., Murray, J., Oschlies, A., Platt, T., Priddle, J., Quinones, R., Ruiz-Pino, D., Saino, T., Sakshaug, E., Shimmield, G., Smith, S., Smith, W., Takahashi, T., Treguer, P., Wallace, D., Wanninkhof, R., Watson, A., Willebrand, J. and Wong, C.S., 2001. A

new vision of ocean biogeochemistry after a decade of the Joint Global Ocean Flux Study (JGOFS). Ambio (Special issue) 10, pp. 4-31.

Finholt, T.A., 2002. Collaboratories. In: Cronin, E.B. (Ed.), Annual Review of Information Science and Technology, Volume 36, Information Today, Medford, NJ, pp. 73-107.

Fischer, Gerald and Ostwald, Jonathan, 2002. Seeding, Evolutionary Growth, and Reseeding: Enriching Participatory Design with Informed Participation. In: Binder,T., Gregory, J. and Wagner, I. (Eds.), PDC'02. Participatory Design Conference, Malmö, Sweden, June 23-25, 2002. Palo Alto, CA: CPSR, pp. 135-143.

Flierl G., Bishop J.K.B., Glover, D.M. and Paranjpe, S., 1992. Data management for JGOFS: theory and design. Proceedings of Ocean Climate Data Workshop. February18–21 1992, Goddard Space Flight Center, Greenbelt, MD. U.S. NOAA and NASA publ. pp. 229-249.

Fox, P., McGuinness, D.L., Middleton, D., Cinquini, L., Darnell, J.A., Garcia, J., West, P., Benedict, J., Solomon, S, 2006. Semantically-Enabled Large-Scale Science Data Repositories. the 5th International Semantic Web Conference (ISWC06), LNCS, ed. Cruz et al., vol. 4273, pp. 792-805, Springer-Verlag, Berlin.

Fox, P., McGuinness, D.L., Raskin, R., Sinha, K, 2007. A Volcano Erupts: Semantically Mediated Integration of Heterogeneous Volcanic and Atmospheric Data. Proceedings of the First Workshop on Cyberinfrastructure: Information Management in eScience, co-located with the ACM Conference on Information and Knowledge Management, Lisbon, Portugal, November 9, 2007.

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

Friedman, A. L., 1989. Computer Systems Development: History, Organization and Implementation. Wiley, Chichester.

Friesen, N., 2002. Semantic interoperability, communities of practice and the cancore learning object metadata profile. WWW2002 (W3C). The Eleventh International World Wide Web Conference, Honolulu, HI. May, 2002.

Fugmann, R. 2004. Learning the lessons of the past. In: Rayward, W.B. and Bowden, M.E. (Eds.), The History and Heritage of Scientific and Technological Information Systems: Proceedings of the 2002 Conference. Information Today, Philadelphia.

Giaccardi, E., and Fischer, G., 2005. Creativity and evolution: a metadesign perspective. Proceedings of the European Academy of Design (EAD-6) Conference, Bremen, Germany, pp. 29-31.

Glover, D.M., 2001. Taking care of the legacy: data management in US JGOFS. Oceanography 14(4), 106-107.

Glover, D. M. and Chandler, C.L., 2001. An update on data management in US JGOFS. US JGOFS News. US JGOFS Planning Office, Woods Hole, MA. 11(3), 16.

Glover, D.M., Chandler, C.L., Doney, S.C., Buesseler, K.O., Heimerdinger, G., Bishop, J.K.B. and Flierl, G.R., 2006. The US JGOFS data management experience. Deep-Sea Research II 53(5-7), 793-802.

Gold, A., 2007a. Cyberinfrastructure, data, and libraries, Part 1: A Cyberinfrastructure primer for librarians. D-Lib Magazine 13(9/10). http://www.dlib.org/dlib/september07/gold/09gold-pt1.html

Gold, A., 2007b. Cyberinfrastructure, data and libraries, Part 2: Libraries and the data challenge: roles and actions for libraries. D-Lib Magazine 13(9/10). http://www.dlib.org/dlib/september07/gold/09gold-pt2.html

Golley, F.B., 1993. A History of the Ecosystem Concept in Ecology: More Than the Sum of the Parts. Yale University Press, New Haven.

Goodchild, M.F., Egenhofer, M.J., Fegeas, R. and Kottman, C., 1999. Interoperating Geographic Information Systems. Kluwer Academic Publishers, Boston.

Greenwood, M., Goble, C.A., Stevens, R., Zhao, J., Addis, M., Marvin, D., Moreau L. and Oinn, T., 2003. Provenance of e-science experiments - experience from bioinformatics, Proceedings of UK e-Science All Hands Meeting 2003. Cox, S.J. (Ed.), East Midlands Conference Centre, Nottingham, pp. 223-226.

Gruber, T.R., 1993. A translation approach to portable ontology specification. Knowledge Acquisition 5, 199-220.

Grudin, J., 1994. Computer-supported cooperative work: its history and participation". IEEE Computer 27 (4), 19-26

Guarino, N and Welty, C, 2000. A formal ontology of properties. knowledge engineering and knowledge management: method, models and tools. 12th International Conference (EKAW 2000). R. Dieng and O. Corby. Berlin/New York, Springer, pp. 97-113.

Hankin, S., Davison, J., Callahan, J., Harrison, D.E. and O'Brien, K., 1998. A configurable web server for gridded data: a framework for collaboration. 14th International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology. AMS Providence, RI, pp. 417–418.

Hobbie, J.E., Carpenter, S.R., Grimm, N.B., Gosz, J.R. and Seastedt, T.R., 2003. The US Long Term Ecological Research Program. BioScience 53(2), 21-32.

Institute of Electrical and Electronics Engineers, 1990. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. New York, NY.

Jackson, S.J. and Baker, K.S., 2004. Ecological design, collaborative care and Ocean Informatics. Proceedings of the Participatory Design Conference 2: 64-67. July 27-31, 2004, Toronto, Canada.

Jackson, S.J., Edwards, P.N., Bowker, G.C. and Knobel, C.P., 2007. Understanding Infrastructure: History, Heuristics, and Cyberinfrastructure Policy, First Monday 12(6). http://www.firstmonday.org/issues/issue12_6/jackson/index.html

Jones, M.B., Berkley, C., Bojilova, J. and Schildhauer, M., 2001. Managing scientific metadata. IEEE Internet Computing 5(5), 59-68.

Jones, M.B., Schildhauer, M.P., Reichman, O.J. and Bowkers, S., 2006. The new bioinformatics: Integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics 37, 519-544.

Kanstrup, A.M., 2005. Local Design. PhD Thesis, Aalborg University, Aalborg, Denmark.

Kaplan, S. and Seebeck, L., 2001. Harnessing complexity in CSCW. Proceedings of the Seventh European Conference on Computer Supported Cooperative Work, 16-20 September 2001, pp. 359-397.

Karasti, H. and Baker, K.S., 2004. Infrastructuring for the long-term: ecological information management. Hawaii International Conference for System Science Proceedings, 2004.

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

Karasti H., Baker, K.S. and Schleidt, K., 2007. Digital Data Practices and the Long Term Ecological Research Program, Proceedings of the Digital Curation Conference, DCC2007, Washington, D.C.

Khazanchi, D. and Munkvold, B.E., 2000. Is information systems a science? An Inquiry into the nature of the information systems discipline. The Database for Advances in Information System 31(3), 24-42.

Kling, R. and Jewett, T., 1994. The social design of worklife with computers and networks: a natural systems perspective. Advances in Computers 39, 239-293.

Kling, R. and Scacchi, W., 1982. The Web of Computing: Computer Technology and Social Organization. Advances in Computers 21, 1-90.

Lave, J. and Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge.

LTER IMC, Information Management Committee and LTER Coordinating Committee, 2005. Review Criteria for LTER Information Management. http://intranet.lternet.edu/im/im_requirements/im_review_criteria

LTER CC, LTER Planning Committee and the Cyberinfrastructure Core Team, 2007. Integrative Science for Society and Environment: A Strategic Research Initiative. LTER Network Office Publication #23. University of New Mexico, Albuquerque, New Mexico.

Lynch, C., 2002. Digital collections, digital libraries, and the digitization of cultural heritage information. First Monday 7(5). http://www.firstmonday.org/issues/issue7_5/lynch/index.html

Lyytinen, K. and Hirschheim, R., 1987. Information Systems Failures – A Survey and Classification of the Empirical Literature. Oxford. Surveys in Information Technology 4, 257-309.

Michener, W.K. and Brunt, J.W., 2000. Ecological Data: Design, Management and Processing. Oxford, Blackwell Science.

Millerand, F. and Bowker, G.C., in press. Metadata standards. trajectories and enactment in the life of an ontology. In: Star, S.L. and Lampland, M. (Eds.), Formalizing Practices: Reckoning with Standards, Numbers and Models in Science and Everyday Life.

MMI, 2005. MMI Workshop: Advancing Domain Vocabularies. Boulder, Colorado. August 2005. http://marinemetadata.org/examples/mmihostedwork/ontologieswork/mmiworkshop05/

MMI, 2008. MMI Metadata Tutorials (Introduction to Metadata, Introduction to Controlled Vocabularies, Submitting to a Metadata Clearinghouse, and Metadata Best Practices) online. http://marinemetadata.org/oceansciences08

Nardi, B.A. and O'Day, V.L., 1999. Information ecologies: Using Technology with Heart. MIT Press, Cambridge, MA.

NAS, 1984. Global ocean flux study: proceedings of a workshop. National Academy of Sciences. Washington, D.C., National Academy Press 360.

NRC, Committee on Exploration of the Seas, 2003. Exploration of the Seas: Voyage into the Unknown, National Academy Press, Washington, D.C., pp. 1-213.

NRC, Committee on Major US Oceanographic Research Programs, 1999. Global Ocean Science: Toward an Integrated Approach, National Academy Press, Washington, D.C., pp. 1-184.

NRC, Committee toward a national collaboratory: establishing the user-developer partnership, 1993. National Collaboratories: Applying Information Technology for Scientific Research, Computer Science and Telecommunications Board, Commission on Physical Sciences, Mathematics and Applications, National Research Council, Washington, D.C., National Academy Press.

NRC, National Research Council, 1995. Finding the Forest in the Trees: The Challenge of Combining Diverse Environmental Data. National Academy Press. Washington D.C.

NRC, National Research Council, 1992. Oceanography in the Next Decade: Building New Partnerships, National Academy Press, Washington, D.C., pp. 1-216.

NSB, 2005. Long Lived Digital Data Collections: Enabling Research and Education in the 21st Century, National Science Board (NSB-05-40, Revised May 23, 2005).

NSF AC-ERE, 2003. Complex Environmental Systems: Synthesis for Earth, Life, and Society in the 21st Century. Pfirman, S. (Ed.). NSF Advisory Committee for Environmental Research and Education.

Ouksel, A.M., Sheth, Amit P., 1999. Semantic Interoperability in Global In- formation Systems: A Brief Introduction to the Research Area and the Special Section. SIG-MOD Record 28(1), 5-12.

Palmer, C.L., Knutson, E.M., Twidate, M., and Zavalina, O., 2006. Collection Definition in Federated Digital Resource development. American Society of Information Science and Technology, Presented at ASIS&T, November 7, 2006, Austin, Texas, pp. 1-12.

Poore, B., 2003. The open black box: the role of the end-use in GIS integration. The Canadian Geographer 47(1), 62-74

Ribes, D., 2006. Universal informatics: building cyberinfrastructure, interoperating the geosciences. Department of Sociology (Science Studies). San Diego, University of California. Unpublished Ph.D. Dissertation.

Ribes, D. and Bowker, G., 2008. Organizing for multidisciplinary collaboration: the case of the Geosciences Network'. In G.M.Olson, J.S.Olson and A.Zimmerman (eds), Scientific Collaboration on the Internet, Cambridge MIT Press.

Sandusky, R.J., 2003. Infrastructure management as cooperative work: implications for systems design. International Journal of Computer Supported Cooperative Work 12, 97-122.

Sarmiento, J. L. and Armstrong, R.A., 1997. US JGOFS Synthesis and Modeling Project Implementation Plan: The Role of Oceanic Processes in the Global Carbon Cycle. Woods Hole, MA, US JGOFS Planning and Coordination Office, Woods Hole Oceanographic Institution pp. 1-73.

Schmidt, K. and Simone, C., 1996. Coordination mechanisms: towards a conceptual foundation of CSCW systems design. Computer Supported Cooperative Work: The Journal of Collaborative Computing 5, 155-200.

Schon, D., 1987. Educating the Reflective Practioner. San Francisco, Jossey-Bass Publisher.

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

SCOR, 1987. The joint global ocean flux study: background, goals, organization and next steps. Report of the International Scientific Planning and Coordination Meeting for Global Ocean Flux Studies, Paris, 2/17—19/87, Available from SCOR Secretariat, Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J1, pp. 1-42.

Sheth, A.P., 1999. Changing focus on interoperability in information systems: from system, Syntax, structure to semantics. In: Goodchild, M.,Egenhofer, M, Fergeas, R., Kottman, C. (Eds.), Interoperating Geographic Information Systems. Kluwer Academic Publishers, Boston.

Simmhan, Y.L., Plale, B., and Gannon, D., 2005. A survey of data provenance in e-science, SIGMOD Record, 34(3), 32-36.

Simone, M., Mark, G. and Giubbilei, D., 1999. Interoperability as a means of articulation work. Proceedings of the ACM Conference on Work Activities, Coordination and Collaboration, Feb. 22-25, 1999. ACM Press, San Francisco.

Smith, B., 2003. Ontology. In: Floridi, L. (Ed.), Blackwell Guide to the Philosophy of Computing and Information. Oxford, Blackwell, pp. 155-166.

Smith B., and Welty, C., 2001. FOIS introduction: Ontology---towards a new synthesis. Proceedings of the international conference on Formal Ontology in Information Systems, pp. 3-.9, October 17-19, 2001, Ogunquit, Maine, USA. [doi:10.1145/505168.505201]

Smith, F.E., 1968. Proceedings of the National Academy of Sciences of the United States of America 60(1), 5-11.

Smith, R. C., Baker, K.S., Fraser, W.R., Hofmann, E.E., Karl, D.M., Klinck, J.M., Quetin, L.B., Prezelin, B.B., Ross, R.M., Trivelpiece, W.Z. and Vernet, M, 1995. The Palmer LTER: a long-term ecological research program at Palmer Station, Antarctica. Oceanography 8 (3), 77-86.

Spanner, D., 2001. Border Crossings: Understanding the cultural and informational dilemmas of interdisciplinary scholars. The Journal of Academic Librarianship 27(5), 352-360.

Spencer, B.F, Butler, R., Ricker, K., Marcusiu, D., Finholt, T., Foster, I. and Kesselman, C., 2006. Cyberenvironment project management: lessons learned http://neesgrid.ncsa.uiuc.edu/documents/CPMLL.pdf

Star, S.L. and Bowker, G.C., 2002. How to infrastructure. Lievrouw, L.A. and Livingston, S. (Eds.), The Handbook of New Media. SAGE Publications, London.

Star, S.L. and Lampland, M. (Eds.), in press. Formalizing Practices: Reckoning with Standards, Numbers and Models in Science and Everyday Life.

Star, S.L. and Ruhleder, K., 1996. Steps toward an ecology of infrastructure: design and access for large information spaces. Information Systems Research 7(1), 111-134.

Stonebraker, M., 1994. Sequoia 2000: A reflection on the first three years. In: French, J. and Hinterberger, H. (Eds.), Seventh International Working Conference on Scientific and Statistical Database management, 28-30 Sep 1992, Charlottesville, VA. Los Alamitos, CA, IEEE Computer Society Press, pp. 108-116.

Strebel, D.E., Landis, D.R., Huemmrich, K.F., Newcomer, J.A. and Meeson, B.W., 1998. The FIFE Data Publication Experiment. American Meterological Society X, 1277-1283.

Thorley, M.R. and Trathan, P.N., 1994. Biomass data set documentation. Cambridge, CB3 OET, United Kingdom British Antarctic Survey, Natural Environment Research Council. UNESCO, 1994. Protocols for the Joint Global Ocean Flux Study (JGOFS) Core Measurements. IOC Manuals and Guides 29, pp. 1-170.

UNESCO, 1994. Protocols for the Joint Global Ocean Flux Study (JGOFS) Core Measurements. IOC Manuals and Guides No. 29.

US GOFS Steering Committee, 1986. US Global Ocean Flux Study Steering Committee Minutes. 26-27 June 1986, URI GSO, Narragansett, RI. US JGOFS Planning Office, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, 8 pp.

US JGOFS Steering Committee, 1990. US Joint Global Ocean Flux Study Long Range Plan, The Role of Ocean Biogeochemical Cycles in Climate Change. US JGOFS Planning Report Number

11. US JGOFS Planning Office, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, pp. 1-71.

United States JGOFS Final Data Report, 2003. Woods Hole Oceanographic Institution, USA: U.S. JGOFS Data Management Office. http://usjgofs.whoi.edu/publications/FinalDataRpt.html

Veltman, K. H., 2001. Syntactic and semantic interoperability: new approaches to knowledge and the semantic Web. The New Review of Information Networking 7, 159-183.

Visser, U., Stuckenschmidt, H., Wache, H., Vögele, T., 2000. Enabling technologies for interoperability. In: Visser, U. and Pundt, H. (Eds.), Workshop: Information Sharing: Methods and Applications at the 14th International Symposium of Computer Science for Environmental Protection. TZI, Univ. of Bremen, Bonn, Germany, pp. 35-46.

Waltner-Toews, D., Kay, J.J., Neudoerffer, C. and Gitau, T., 2003. Perspective changes everything: managing ecosystems from the inside out. Frontiers in Ecology and the Environment 1, 23-30.

Weick, K.E., and Sutcliffe, K.M., 2001. Managing the Unexpected, Assuring High Performance in an Age of Complexity. Jossey-Bass, San Francisco.

Weick, K.E., Sutcliffe, K.M. and Obstfeld, D., 1999. Organizing for high reliability: processes of collective mindfulness. Research in Organizational Behavior 21, 81-123.

Whitman, M. and Woszczynski, A., 2004. The Handbook of Information Systems Research. Hershey, PA: Idea Group Publishing.

Zimmerman, A.S., 2003. Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists. PhD Thesis, The University of Michigan, Ann Arbor.